Chapter 6

Infrastructuring Ecology: Challenges in Achieving Data Sharing

Karen S. Baker and Florence Millerand

Introduction

Information infrastructure initiatives have developed in recent years, particularly in the natural sciences, with the goal of enabling large-scale scientific collaborations. These initiatives are called cyberinfrastructure when considered suitable for addressing global-scale scientific challenges. Also referred to as e-Science, e-Research and e-Infrastructure, the initiatives promise profound transformations in scientific collaboration: 'They can serve individuals, teams and organizations in ways that revolutionize what they can do, how they do it and who participates' (Atkins et al. 2003: 2). Among the possibilities provided by new research environments based upon cyberinfrastructure, are new opportunities for sharing research data. Working with a variety of data collected in multiple, heterogeneous settings and asking questions requiring synthesis of these data are activities central to field intensive research domains such as ecology. Concurrent with cyberinfrastructure initiatives, diverse smaller-scale information infrastructure developments are producing new understandings of data and data related issues. While large-scale data initiatives support access to highly-structured data, infrastructure efforts of varied sizes are highlighting significant differences in data practices, methods and purposes as well as data types, sampling and analysis.

Funding agencies today are directing attention to infrastructure elements such as massive data collections, comprehensive data catalogs and high-volume data streams located at data archives though scientific data efforts are not solely large-scale. Researchers in ecology make use of smaller-scale databases and personal data arrangements to organize and handle research data on a daily basis. Ecological data remain closely tied to traditional disciplinary knowledge production for which scientists generate and make use of primary data (Leonelli 2007, Michener and Brunt 2000). Yet data initiatives today frequently focus on data *re*use by multiple 'others', that is, by researchers outside the project, outside the domain and outside the sphere of science itself. These new arrangements are being debated in several ways including as new forms of knowledge production (e.g. Gibbons et al. 1994) or as 'data-driven science' (Arms and Larsen 2007). Many expectations for data sharing are fed by success stories that in the life sciences may be unique and field-specific, e.g. genomic databases like the Protein Data Bank (Berman, Bourne and

Westbrook 2004) rather than being generally representative. The question arises as to whether models of data sharing can be borrowed and imported with equal success in all scientific arenas, notably the environmental sciences characterized by research data and data practices that are highly heterogeneous and complex.

Often envisioned as a 'little science' research arena, ecology is currently undergoing significant changes, particularly in terms of its data practices. New instrumentation such as embedded and autonomous sensor networks (Borgman, Wallis and Enyedy 2008, Hart and Martinez 2006) provides both unprecedented amounts of data and new types of data, thereby requiring development of new methods and analysis techniques. Research projects and programs large and small are planning synthesis activities assuming integration of disparate data, thus confronting scientists with new issues of data interoperability (Baker et al. 2005, Ribes et al. 2005). Also, there are new data access policies from funding agencies such as the Division of Environmental Biology (DEB) as defined within Grant General Conditions of the National Science Foundation (NSF 2009, GAO 2009). The policies promote open access to publicly funded research data by requiring public data publishing 'within a reasonable time' and involve review criteria that require data providers to document their data in new ways.

This chapter gives an account of on-going changes in the data practices in ecology. Though ecological methods include experimentation and manipulation, we focus here on field observations. Our first objective is to highlight scientists' daily work with research data as an important but under-explored aspect of contemporary scientific practice. With the development of technology and of the digital realm, there are new instruments for making measurements, new approaches to organizing data and new methods for sharing and publishing data. Such developments impact everyday scientific work and data practices. Our second objective is to engage with debates about development of large-scale information infrastructure projects to support and enhance scientific collaboration. More specifically, we ask: How are data practices changing in ecology? What are the main challenges in terms of roles and expertise? How can infrastructure development support current and emergent data practices as well as enable new modes of scientific collaboration? Based on a longitudinal qualitative analysis of data practices in one of the largest ecological research communities in the US, the Long-Term Ecological Research Network (LTER), we address scientific work with complex biological data. From field collection to analysis and curation, we consider what is called the 'lifecycle of data'. Within a network science model, we investigate situated infrastructures and local expertise as critical factors relating to data access, data quality and larger-scale cyberinfrastructures.

We begin by drawing upon work on scientific collaboration and data sharing that suggests key issues for ethnographic inquiry. This framework will be used to explore data practices in contemporary scientific practice as well as the role of information infrastructure and associated expertise in enabling scientific collaboration (Collins and Evans 2002, 2007). Next we present our research setting and research approach followed by a description of the complexity of ecological data, the data analysis subcycle and current changes in data practices. Discussion then turns to the challenges of 'infrastructuring' for supporting scientific collaboration. The conclusion explores implications of researching the role of infrastructure in supporting scientific collaboration for science and technology policy. We consider diverse types of information infrastructure and end by suggesting that *cyberinfrastructure entails not only large-scale, cross-program efforts but also involves development of local-scale, cross-project efforts.*

Collaboration, data sharing and infrastructure

It is now almost a platitude to say that information and communication technologies are transforming the world of research. However, it is also well known that the relationship between changes in science and changes in technology is less than straightforward (Galison 1997). In this chapter, we are interested in the systems and technologies by which scientific collaboration[1] is achieved and knowledge produced. More specifically we are interested in research data, data practices and data publication in the context of information technology developments for the support of field science.

Data – primary research data from field observations and measurements – are a fundamental component of scientific work.[2] Data sharing has historically been regarded as a distinguishing, collaborative feature of scientific practice in providing confirmation of research findings through replication and knowledge production, that is, by building on the work of others (Merton 1968). Considered an empirical foundation for knowledge production, data play multiple roles in science work. One of the most obvious is their direct contribution to the production of scientific fact; that is, data are used to confirm scientific expectations of various kinds and to build new explanatory frameworks. Data are then regarded as truthful representations of the physical world, 'immutable mobiles' to be transportable and combinable in Latour's terms (1987, 1990) and as evidence to support scientific claims. Less obvious is the social role that data play in contributing to the formation

[1] We begin with Hackett's definition of scientific collaboration: 'Collaboration is a family of purposeful working relationships between two or more people, groups, or organizations. Collaborations form to share expertise, credibility, material and technical resources, symbolic and social capital' (Hackett 2005). Further, we recognize that a variety of purposes create a variety of relationship types (Briscoe 2008).

[2] We take Hacking's broad definition of data as any 'marks' produced by a 'data generator' (1992: 48) into the e-science arena using the NSB (2005) focus on data as referring 'to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc.'. A simple technical definition is as follows: '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' (CCSDS 2002: 1–9).

and development of scientific communities (Birnholtz and Bietz 2003). Data mean different things to different research communities, are used in different ways and are assessed unevenly in terms of their 'value'. Data may define boundaries between communities, serve as a gateway when access to data becomes a point of entry into a community or act as a status indicator.

Contemporary cyberinfrastructure initiatives are throwing light on data and data practices in the sciences in two principal ways: first, in promoting larger-scale scientific collaboration and second, in making new arrangements for data sharing and more formal digital data publication. Shared data and the data repositories within which the data reside bring a new dimension to scientific collaboration in allowing multiple researchers, laboratories and institutions to collaborate on the creation, use and reuse of very large datasets (NSF 2007a). Funding agencies require grant recipients to make their data public according to a new data access regime where 'publicly funded data are intended to be openly available to the maximum extent possible' (Arzberger et al. 2004). These initiatives, echoing current global trends in academic research where transformations in science organization are framed in terms of growth, are referred to as 'supersizing science' (Vermeulen 2009) or more commonly as 'big science' (Price 1963, Weinberg 1967, Furner 2003, Borgman, Wallis and Enyedy 2008). Does the focus on data sharing and data policy to date over-simplify problems inherent to working with data? Developing understandings of data types, data packaging and data circulation across multiple contexts and epistemic cultures point to a need for further investigation (Wouters and Schroder 2003, Carleson and Anderson 2007, Knorr Cetina 1999, Hilgartner 1995, Leonelli 2008, RIN 2008).

Ecology is a data-rich research domain with a long history of collaboration and interdisciplinarity (Pickett, Kolasa and Jones 2007, Bocking this volume). Indeed, many ecological research activities are observationally oriented and rely on the analysis and integration of many kinds of disciplinary data. It has been traditional for individual ecological researchers or small research groups to collect data in short term projects (typically the length of a funding cycle) over small areas (even as small as one square metre) (Lewontin 2000, Michener and Brunt 2000). But this is no longer sufficient to the task (Bowker 2006). Ecosystems change over larger chunks of time and space than traditionally conceived (O'Neill 2001, Powell and Steele 1995). Ecological researchers now need to be able to use datasets constructed by others for different purposes; they need to be able not only to reach some kind of ontological accord between the disciplines about kinds and classifications of data to be shared (Bowker and Star 2000) but also to be able to trust data produced by others. Ongoing efforts to foster data sharing practices within the domain involve identification of barriers, development of incentives and design of coordination mechanisms (ESA 2009, NOAA/NESDIS 2008).

Ecology is often depicted as a field in transition (Jones et al. 2006, Borgman, Wallis and Enyedy 2008, Parker this volume) undergoing both a scientific 'pull' toward global science and a technical 'push' of cyberinfrastructure. The mix of data-intensive (large volumes of data generated by high technology instrumentation) and data-rich (diverse sets of data collected using manual techniques) presents researchers with an array of challenges in organizing, managing, synthesizing and curating data. Yet with all these challenges, still little is understood about the diverse ways scientists actually produce, manage and use data. In addition, the information infrastructure that could facilitate scientific work is also under-studied.

Many facets of scientific practices are typically categorized as tacit knowledge or second-order 'technical' tasks (Polanyi 1967, Whitley 2000, Linde 2001). Wallis et al. (2008) point to the cumulative effect of decisions made within each stage or category of work associated with the data lifecycle; for instance decisions made at the experimental design stage determine what data exist for analysis while decisions regarding calibration are essential to interpreting the data. These authors suggest making the full cycle of data more transparent for the parties involved (scientists, data managers, technologists, data curators, etc.). However, accurately documenting decisions made at each phase of data work is notoriously difficult especially when there are multiple user communities (Parsons and Duerr 2005). Technically called metadata production, this activity is far from being a purely technical task; metadata languages as well as ontologies are sociotechnical artefacts that embed many types of descriptive statements as lasting records (McDonough 2008, Millerand and Bowker 2009). Metadata work requires expertise and is time-consuming, yet support and incentives for effective metadata production are generally lacking in scientific organizations.

Ramifications of changing data practices for ecological data curation are under investigation (e.g. Karasti and Baker 2008). In considering the need for data repositories, Baker and Yarmey (2009) distinguish three organizational arrangements: local repositories, synthetic centres and large-scale archives with their respective 'spheres-of-context' that parallel research-centred, resource-centred and reference-centred data collection categories (National Science Board 2005). Local or research-centred repositories are the products of one or more focused research projects and are intended to serve a specific group, often limited to immediate participants (e.g. Baker and Chandler 2008); data centres or resource-centred repositories serve a science community (e.g. Michener et al. 2001, Romanello 2005); and archives or reference-centred repositories serve many scientific communities (e.g. NRC 2007). Conceptualizing a 'web of repositories' accounts for the interdependent relationships among the multiple organizational arrangements (Baker and Yarmey 2009). Repositories are developed in response to different needs, priorities and cultures and provide distinct views of the data. For instance, a local research-centred system enables community scientists to work with the data, and through use, data are reviewed and validated. A primary goal of the resource-centred data repository is to make data accessible at the broader scope of the domain, thereby enabling data reuse. A reference-centred repository's goal traditionally has been preservation and support of historical investigations.

Though it is clear that there are opportunities to use advanced systems and technologies to facilitate the sharing of research data, there are important issues associated with the publication of data that technical systems will not easily solve (e.g. ethical concerns regarding intellectual property, data quality and fear of data misuse; cultural concerns about sensitive data handling such as with locations of endangered species; organizational issues such as lack of strategic planning for changes in data practices as well as lack of reward or support for a move from knowledge to data production). Birnholtz and Bietz (2003) identify two critical data related issues found across domains, the sharing behaviour among researchers and the diversity of contexts for metadata production.

In ecology, several issues regarding data sharing have been detailed. For instance, the loss of information over time is recognized (Michener 2000) as is the role played by a detailed knowledge of the data setting in the understanding and assessment of the quality of data. Local knowledge is key to recovery of the local details that are so critical to the comprehension of data collected by others. Though there is a developing understanding of standards and standards-making (Hanseth et al. 1996, Star and Lampland 2009), attaining methodological standardization is difficult, if not impossible, in many instances (Zimmerman 2008). An example case is the laborious work associated with the development of the Ecological Metadata Language that included a dictionary for largely physical units (Millerand and Bowker 2009). With biological units of central concern in ecology, implementation of a dictionary demonstrates recognition of a community need; development of a dictionary of physical units could be viewed as a first step in addressing this need (Karasti, Baker and Millerand 2010).

In thinking of technological developments for the support of science, a key idea from infrastructure studies is that an infrastructure neither emerges *ex nihilo* nor builds up in a straightforward way but rather develops in a particular setting and at a particular time, adjusting to, adapting or reshaping elements of the setting (Bowker et al. 2010, Edwards et al. 2007, 2009). Recent work promotes the metaphor of 'growing' an infrastructure in the sense of an organic unfolding rather than construction of a thing according to a plan (Nardi and O'Day 1999, Jackson et al. 2007). Lessons from the history of large-scale scientific projects reveal the value of 'organic' approaches to infrastructure because 'constructed' infrastructures fail to meet the users' needs first, through failure to link successfully with technical, political and/or social systems and second, through inability to adapt to changing circumstances (Edwards et al. 2009). Failure rates with digital configurations are high because innovations in changing environments are hard to plan or anticipate. This is relevant to current cyberinfrastructure initiatives launched within a context of changing data practices, organizational re-arrangements and emergent technological innovations.

The concept of 'infrastructuring' developed by Star and Ruhleder (1996) and Star and Bowker (2002) to account for the complexity of infrastructure design and development, emphasizes the idea of an on-going and active process that the verb 'to infrastructure' aims at capturing. Infrastructuring is used largely as a comprehensive term that encompasses design activities associated with infrastructure development, including work performed by professional designers as well as by users' participation in design and development activities, thus suggesting co-construction and participation as inherent components of infrastructure development (e.g. Karasti and Baker 2004, Pipek and Wulf 2009, Bowker et al. in press). The term 'infrastructuring' connotes a reflective enterprize that challenges common views of infrastructures as being inert, 'already-there' and taken for granted. We use the term infrastructuring in this same line of argument; in this chapter we envision and explore developments for the support of collaboration through data sharing in ecology as an infrastructuring endeavour.

The Long-Term Ecological Research network

The Long-Term Ecological Research (LTER) network offers a rich and unique setting for ethnographic inquiry by providing multiple viewpoints not only on scientific practices that mobilize data uses and reuses (on a site level or network level) but also on collaboration configurations (individual researcher, laboratory, site, community, network). The LTER program is a distributed, heterogeneous network of several thousand research scientists and students. Formed in 1980, the network currently consists of 26 individual sites or research stations, each arranged around study of a particular habitat, for example, a hot desert region, a coastal estuary, a temperate pine forest or a marine ecosystem (Hobbie et al. 2003). A broad mix of researchers exists at each site. Although ecologists share as a unifying theme the largely self-organizing system of 'Nature' and ecosystems, their expertise differs in terms of the object studied be it plankton or penguins, species counts or nutrient flows. The tools and methods used in field sampling, sample analysis and data analysis also differ. The program's mission is to further understanding of environmental change through interdisciplinary, long-term collaboration. Development of the LTER was informed by programs that preceded it such as the International Geophysical Year (IGY) and the International Biological Program (IBP) (Callahan 1984, Golley 1993, Aronova, Baker and Oreskes submitted).

The LTER network represents a collaborative science model that explicitly includes data management at each site. As a result each site in the network manages the research data produced locally via its own data collections, databases and information systems that comprise a site data repository. There is at least one information manager at each site actively involved in work with data. The network is an exemplar of distributed, loosely connected sites with independent information infrastructures that have grown and changed over time. In addition, there is a network office supporting the network. Data sharing and data reuse have a long tradition in the network; data management both within a site and across the sites took place almost from the inception of the network at which time data sharing took place via a simple exchange of handwritten notes. The network initiated an open data sharing policy in 2005 and has developed a 'community approach' to data and information management (Karasti and Baker 2008, Baker et al. 2000) as well as to more recent cyberinfrastructure developments (Brunt et al. 2007).

As LTER scientific focus evolved over time, data practices and information management changed significantly. In 1980 the six LTER research sites were instructed 'to network' (Callahan 1984). Scientifically, there was an initial focus on developing a cross-site, 'community' understanding of 'long-term'. During this first decade, what might be called the 'Decade of Long-Term', site data managers were aggregating data and dealing with legacy as well as continuing datasets so from an informatics perspective, this period was a 'Decade of Time-Series Data' (Table 6.1). By the 1990s, the number of sites tripled and the World Wide Web was providing enhanced functionality for distributed work. Scientists' attention turned to an expanding understanding of spatial scales in a 'Decade of Large-Scale' while data managers addressed the multi-faceted issue of data sharing by adopting a policy of open access to primary research data and development of metadata during what was a 'Decade of Data Sharing'. By the end of this decade, the increased expectations and responsibilities associated with site-based data together with network-level activities, prompted a renaming of the Data Management Committee to Information Management Committee and the creation of an information management vision statement (Baker et al. 2000). Network participants focused as a whole on socioecological research during the first half of the 2000–2010 decade that LTER scientists labelled the 'Decade of Synthesis'. During this period, the information management role expanded to cover a wider range of data-related issues including data access and standards-making. The community is now in the midst of gaining experience with practices that span situated, site-level data use and network-level data reuse, managing multiple data types and contributing to many data partnerships during what can be called a 'Decade of Data Integration'.

Table 6.1 Three decades of the LTER Network

Decade	Science Perspective	Research Focus	Informatics Perspective
1980–1990	Long-Term	Multi-temporal Interdisciplinary	Time-Series Data
1990–2000	Large-Scale	Network-themes Multi-spatial Cross-site	Data Sharing
2000–2010	Synthesis	Communication Socio-ecological Partnering Governance	Data Integration

In terms of research approach, we developed a longitudinal, qualitative analysis framework using ethnographic methods for data collection and a grounded theory lens for data analysis. We have a long-term engagement in the research setting that allowed us to 'follow scientists at work' (Latour and Woolgar 1979) for several years. Baker has a dual role as participant information manager working closely with ecological researchers (since 1990) while also maintaining the stance of an observer. Millerand began fieldwork in 2004 and started community participation as an action science researcher in that same year. Participant observation is the primary source of our ethnographic data with the goal of gathering detailed, in-depth description of everyday work and creating 'thick description' (Geertz 1973). Data collection was carried out for targeted LTER activities and anchored by work at two oceanographic LTER sites, Palmer Station and California Current Ecosystem.

Ecology and ecological data

The biotic environment is diverse, and the earth's web of life is complex. Ecology considers this complex web as a whole, studying the distribution and abundance of organisms, their environment and the relations among them all (Odum and Barret 2005, Pickett, Kolasa and Jones 2007). Ecological research addresses a wide array of spatial and temporal scales: from plots of land to the whole earth, from a moment in time to the long-term that ranges from historical past to predicted futures. In each research endeavour, measurements are made or modelled and data are collected or generated, representing some aspect of the earth as a biosphere.

Ecological data are complex

Scientists aim to discover scientific truth by taking and assembling data; they have similar understandings of how to conduct fieldwork. It is, however, biologists and ecologists who understand the ways in which biological measurements model the realities of Nature's organisms and living systems; they recognize both patterns and anomalies in biological data. They share the proficiency to question, check and compare this data. Discourses are beginning to emerge that capture the complexity and contextualized aspects of data where data are recognized as objects or multi-stage processes and having interpretations and relationships. Today ecological researchers' understandings are frequently discerned and extracted from collections of messy, unruly and irregular data. In comparison, some disciplines are characterized by more homogeneous data and data structures (for instance in molecular biology) or by the predominance of physical data over biological data (for instance in climate science). While physical data typically refer back to one of seven basic unit types (length, mass, time, temperature, current, luminescent intensity and amount of substance), biological data are described by a wide array of units ranging from amount of particular substances per volume and rates of change over varying time intervals to abundances of species, functional groups and size classes grouped in a variety of manners. With living organisms unevenly distributed, their count is complicated by often poorly understood short-term, annual and interannual migrations, reproduction cycles and patch dynamics. Further, sampling of biological entities may lack the ubiquity of physical measurements, i.e. at every point there is a temperature but not necessarily a biological organism. The heterogeneity of data appears, however, not only in terms of sampling issues but also is introduced by collection methods, instrumentation, sampling frequency, field circumstances and analysis procedures.

A description of life on earth involves estimations of the abundances of a large variety of organisms. But estimating the distribution and abundance of the biota across a system as large and diverse as the earth is no simple matter. Biologists use sampling design to take into account the uneven distribution characteristic of the biotic realm. Hand-collection methods have been traditional though time-intensive and cost-prohibitive as well as having irregularities introduced by unanticipated field and analysis conditions. Estimates of abundance such as population abundance may be obtained in two ways: by direct observation of a sample (e.g. counts of number of organisms) and by indirect measures of quantity that are potentially more continuous using instruments configured for in-situ or remote sampling of a proxy (e.g. average biomass). Both methods involve a variety of factors that contribute to sampling error and variance that affects the accuracy and precision of estimates particularly with non-normal distributions. Experience with direct observations and indirect measures suggests that interpreting datasets of diverse types brought together across multiple scales can be a research project in and of itself, a tacit underappreciated part of the scientific process of knowledge building. That is, the mechanics of assembling data in a central location differs from the frequently iterative work of processing and reformatting data in order to be able to interpret *and to evaluate* an integrated result.

Circumstances differ for homogenous and heterogeneous data. When there is regularity such as with data streams from moorings or satellites, the data analysis phase differs significantly from that of hand-collected biological data. Early data-intensive fieldwork initially involved instruments able to measure physical phenomena. In the case of a single, stable instrument or set of procedures, there may be extreme homogeneity in data in contrast to hand-collection methods that produce heterogeneous data typical in ecology. For example, in the case of oceanographic field research, there is a contemporary emphasis on large-scale moored systems that contrast with more traditional ship-based sampling. And although ships are large-scale platforms, they frequently carry a mix of researchers and a bevy of instrumentation that creates heterogeneous data from a myriad of sampling and analysis strategies.

Data analysis and data curation subcycles

The full data lifecycle typically involves a number of tasks and activities that are associated with different categories, phases or stages of the data work (Higgins 2008, Carlson and Anderson 2007). Categories of data work may include data sampling, capture, ingestion, description, appraisal, formatting, storage, transformation, exchange and delivery. Focus on any one category results in a category-specific way of understanding the data. To examine the process of data production more closely, we begin by considering two very broad but distinct data categories: data analysis and data curation.

The data analysis subcycle captures the complexity of data work occurring in hypothesis-driven arenas; this work is closely related to traditional scientific knowledge production. Analysis is a research-centred undertaking focused on making data tractable and understandable as evidence supporting knowledge intended for publication in a journal. The data analysis subcycle includes:

a. Field sampling in terms of quality assurance, sampling design and collection method;
b. Sample analysis as pre-processing in terms of experimental method, sample treatment and analysis method;
c. Data analysis as pre-processing in terms of calculations, calibrations, quality control and data manipulations;
d. Data contextualization as support for data sharing in terms of metadata generation and vocabulary development; and
e. Data processing involving calculations, derived products, visualization and publication or delivery of data. Participation in the data analysis subcycle results in a way of knowing data distinct from that associated with data taking or data curation (Pickstone 2001).

The data curation subcycle may be envisioned as overlapping and co-located with the data analysis subcycle but may also be understood as occurring subsequently at multiple local and/or remote points in the lifecycle of data (Higgins 2008, Baker and Yarmey 2009). Today's cyber-motivated expectation is that data eventually will flow unproblematically between data repositories from the site of origin. When data are transported from their local environment, much of the measurement context and hence the understanding of related factors such as their variability and potential impact, is missing. Metadata is often presented as a solution enabling meaningful data exchange and sharing. The term metadata has emerged in the last decades to become part of the scientific vernacular. It is not unusual today to hear data description recognized as an important issue, a metadata issue. Metadata consists of descriptive statements. RIN (2008) reports 'metadata provides information about an information resource'. A metadata specification defines a structure for categorizing descriptive text; it represents a classification scheme that enables data discovery, exchange and integration. Metadata standards developed initially in support of data catalogues for which a general level of description may be adequate. This level of detail in metadata standards, however, leaves issues of differing resolution of data description unaddressed. For example, the general category 'methods' for capturing text describing the data analysis subcycle activities mentioned above provides no guidance about finer levels of description (Frey 2008).

In formalizing the data analysis and the data curation subcycles, metadata provides an opportunity to make visible and to organize knowledge currently held tacitly. We highlight data analysis and data curation as subcycles in order to capture the individual complexity of their iterative processes comprised of subsets of planned tasks and unexpected activities. Within the data analysis and data curation subcycles, the regularity of a well-defined sequence of steps may be disrupted by unanticipated irregularities. Throughout these subcycles there appear some generative or healthy tensions involving local context-sensitive impulses to accommodate and remote curation-driven impulses to standardize data differences together with the mix of analysis-intensive research impulses to learn from anomalies and data-intensive synthetic efforts to learn from patterns.

Changing data practices

The data analysis subcycle traditionally is hypothesis-driven with a focus on publication of scientific conclusions in print media rather than on sharing of data. With data policies changing to require publication of well-documented data, researchers are faced with developing and adjusting to new data practices. Along with planning and support for the well-established documentation of scientific work via journal publication, planning and support for digital publication of datasets is needed. In this section we consider three cases of scientific publication where the first two are knowledge production processes with traditional scientific journal paper publication as outcomes (one associated with generation of primary data and the other with reuse of existing data) and the third is distinguished as a dataset production process with the goal of publishing reusable data.

Case 1: Knowledge production process via data use

Traditional knowledge production in environmental sciences may be summarized by the following over-simplified steps. This case involves field work and represents a sampling-based knowledge production process:

(a) field sampling – (b) data analysis – (c) publication of journal article(s)

The process culminates with the sharing of knowledge in a variety of forms including often as a journal publication that may include data in tables and/or graphs. Field sampling that is hypothesis-driven includes a data analysis subcycle *and a consequent use* of the data by a local researcher. With scientific researchers using the data for questions that informed the sampling design, they are fully engaged in assessment of the data and contributing to quality control. Disciplinary researchers carry out the field sampling and data analysis as individuals and in collaboration with other researchers, data scientists, students, data technicians, data analysts, informatics specialists and/or data and information managers.

Case 2: Knowledge production process via data reuse

In this case field sampling is rendered unnecessary by the availability of existing data. This case is a 'data-driven' knowledge production process:

(a) data finding – (b) data analysis – (c) publication of journal article

where the use of existing data is designated 'data reuse', differing from case 1 in terms of how data are obtained and knowledge constructed. In this case researchers typically concerned with disciplinary as well as interdisciplinary data integration and synthesis carry out data analysis, frequently at larger spatial or temporal scales.

Case 3: Data production process for data reuse

In transitions to work with digital data, the possibility of online publication of datasets arises. This change, when conceived as a process mirroring traditional context-sensitive knowledge production based upon field sampling as in case 1, results in a data production process described as follows:

(a) field sampling – (b) data processing – (c) publication of dataset(s)

In this configuration, the data work emphasis is on data processing and culminates with publication of a digital dataset available online. Data processing depends upon a well-documented set of procedures applied in a more or less standardized manner. Data processing involves carrying out both data analysis as well as organization and preparation for publication of data via inclusion in a data repository or submission to one of many developing online services. When a majority of sampling and analysis factors are held constant, case 3 may be referred to as 'dataset production' where the intent is to make data available as a well-documented dataset. Dataset publication depends upon a detailed description of the dataset in the form of metadata.

Publication of datasets facilitates future cycles of reuse among researchers and across disciplines, greatly expanding potential collaborations. However, variations in the data analysis subcycle frequently defeat automated data processing plans due to unanticipated irregularities. Standardized procedures must be adjusted to accommodate underdeveloped or incomplete descriptions of sampling and analysis in addition to data differences requiring differing technical strategies. Further, tacit and implicit knowledge gathered in-situ pertaining to the sampling-specific context as part of the case 1 field experience, requires continuing articulation, documentation and incorporation into explicit vocabularies for case 3. Until there are more mature vocabularies and processes for dealing with the complexities of heterogeneous biological data and recognized coordination mechanisms facilitating information visibility, update and exchange, it will be difficult for a remote user to evaluate the variety of shared datasets complete with all their associated biases and uncertainties.

Challenges of data sharing in ecology

Among the challenges of data sharing is the broadening of our understanding of data sharing from a one-time informal exchange to include planning for a more formal data production process. Such a change depends upon development of a situated infrastructure and of expertise relating to new types of work and redistribution of work. We present examples of transitions from a dyad to a team for data work and of a 'network science' model that incorporates site-based infrastructure and accounts for scaling from local data use to large-scale data reuse. In the final section we draw upon the metaphor of a web of repositories and discuss how it influences data practices.

Situated infrastructure development and expertise

The growth of the LTER network provides an example of the development over time of independent information infrastructure at diverse research sites coordinated loosely as a network. We describe observations of transitions in organizational arrangements for data work in four stages (Figure 6.1).

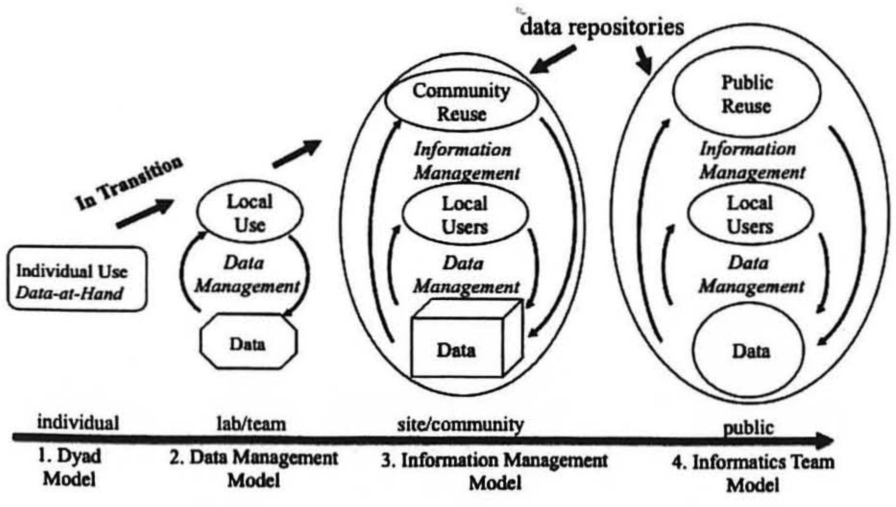

Figure 6.1 Local data and information management development

In light of the changes in expectations and scope of scientific questions that arise with data sharing, it is not surprising to find data practices changing. The variety of arrangements for data and information management that exist at sites constitute local information environments involving both technical infrastructure (data handling procedures and information systems) and associated expertise (information managers, programmers, etc.). Many data efforts begin with a focus on individual data files managed by a single individual who frequently has a history of working with a researcher in what may be referred to as a 'Dyad Model' (Figure 6.1[1]). When this individual takes on the responsibility of data handling for a number of researchers, transition to a 'Data Management Model' occurs (Figure 6.1[2]). With a community of researchers, there are new types of communication and translation work along with multiple types of data. As a site matures, local collaborative activities increase creating new responsibilities such as establishing a data repository for access to multiple data collections and participating in cross-site activities. This wider spectrum of activities can be described as an 'Information Management Model' (Figure 6.1[3]). Some sites add personnel to augment their information management capabilities targeting specialized skills. The next stage of infrastructure development encompasses management of well-documented data, design of information systems, coordination with other scientific networks and new requirements to consider wider audiences (e.g. education, policy, public). An 'Informatics Team Model' may develop in response to a growth in responsibilities. This multi-stage development from dyad to team represents a 'thickening' of a site's research-centred information infrastructure and creates a more complex information environment.

The distribution or redistribution of data work is central to changes in data handling arrangements and to the expertise involved with data decision-making. Traditionally an environmental scientist makes decisions throughout the full data lifecycle. Within a dyad, a researcher may work together with a technician or analyst to whom selected decisions in the data analysis subcycle are delegated. With the growth of a group or community, a broad range of decisions regarding data becomes the responsibility of a local information manager who generally has familiarity with the local data context. For data production, data handling and decision-making responsibilities shift as data work is carried out by information specialists who perform quasi-automated data processing while also checking and adjusting steps in the analysis process to identify and consider irregularities.

When data are no longer available only to experts experienced with that particular data type and no longer associated with testing and validation work carried out in research-centred projects, data publishers are faced either with using existing methods of data and metadata delivery largely suitable for highly-structured, homogeneous data or with developing new methods of data organization and delivery in order to meet the goal of producing well-described data. The emergent role of information specialist brings with it a familiarity with the wide array of data types and an awareness of developments in the field of data description pertinent to avoidance of misconceptions that arise with heterogeneous biological data.

Within a research domain, data are understandable and data work tractable for a scientist with pre-existing knowledge and experience with data collecting, sampling and analysis. The information specialist as somewhat of an outsider – detached from the scientific hypothesis and the field sampling by virtue of focus on the data 'in the digital lab' – is able to facilitate dataset creation by providing quality control and by stabilizing ill-defined characteristics of data as part of the process of transforming a dataset into a stable, publishable data object. In preparing a dataset for publication, data production involves the work of metadata creation as well as structuring for data delivery. Delivery may involve data submission to one or more local or remote repositories. Data access is provided by development of interfaces of various types including file transfer protocols and web access for manual data downloads and/or web service for machine to machine access. When data are published, assembly with other data is possible and typically reveals the need for further work involving data description and interpretation. Data vocabularies frequently require translation in terms of reconsidering data categories, shoehorning irregular data into pre-defined classification schemes, addressing alternative metadata schemas and reformatting to meet data transport specifications (Lowry, Bermudez and Graybeal 2006).

Network science

The term 'network' is used broadly to describe interconnected events, processes, equipment, individuals and organizations; its ambiguity makes it useful in situations where 'relations between and among' are important. It may refer both to limiting and to theoretically unlimited circumstances. In establishing itself as a network, the LTER makes use of this powerful concept. While an early paper announced the focus on 'Long-term ecological research' (Callahan 1984), the title of an overview paper at the end of the first decade of LTER elaborated 'Contributions of the Long-Term Ecological Research Program: An expanded network of scientists, sites and programs can provide crucial comparative analyses' (Franklin, Bledsoe and Callahan 1990), making explicit the participants' view of the strength of the network approach. In growing from an initial set of six to a total of 26 sites today, the LTER network illustrates the network characteristic of expandability.

'Network science' refers to an organizational arrangement that enlarges the context of member sites through association with other sites tied together conceptually by a set of overarching scientific themes (Franklin, Bledsoe and Callahan 1990). The looseness of site associations creates an arena where there is flexibility to change over time. Weick (1979) provides insight into loose coupling as a mechanism that supports sensemaking as well as change by permitting different and even contradictory subsystems to coexist peacefully under the same label or organizational umbrella. Lorelli (2008) discusses network models that are either centralized or decentralized. LTER, along with other fields from historical studies to graph theory, foregrounds local-scales, connections and self-organization, framing discussions in terms of networks rather than focusing on scale (Westfall 2003, Watts 2003). The LTER maintains an emphasis on local-scale efforts including local information environments. Yet, by working together, the ensemble of individual sites represent a large-scale configuration. As such, the LTER case represents a model for growing 'Big Science' or 'Big Biology' from local-scale science (Zimmerman and Nardi this volume, Aronova, Baker and Oreskes submitted).

Change within LTER does not occur abruptly but rather occurs through a continuing mix of activities, reviews, workshops and meetings involving participation from a diversity of members from each site. Trist (1997: 180) describes network structure:

> Networks constitute the basic social form that permits an inter-organizational domain to develop as a system of organizational ecology. Networks are unbounded social systems that are nonhierarchical...They travel through the social ground rather than between institutional figures. They cross-levels and cover the range from private to public. They bring the most unexpected people into relevant contact so that nodes and temporary systems are formed which become levers of change.

Within the LTER network, such a structure enables unplanned encounters with the possibility of intellectual innovation, for instance at events such as LTER All Scientists Meetings and annual Science Council meetings where researchers from all sites engage in working group sessions.

The LTER community has on-going projects that may be recognized as demonstrations of 'network science' activities. One example is the EcoTrends project (Peters 2008), a collaborative effort among state and federal agencies and institutions, at present primarily in the US, to make long-term ecological data easy to access, analyse and compare within and across sites. The project is designed to promote and enable the use and synthesis of long-term data through examination of trends in the Earth's ecosystems. Two specific outcomes involving all sites are planned, a book and a web portal providing data access, exploration and download.

Infrastructuring ecology

With larger research teams and requirements for data publication, comes a formalizing of the data analysis subcycle and infrastructure developments (e.g. metadata-based information systems, standards-based structures and standards-making activities) that enrich local data practices and define information environments. We discuss two main points in this section as key challenges for data sharing in ecology. First, data publication is commonly defined as a problem that can be solved as a technical task though there are important social, organizational and institutional aspects to consider. Further, the scientist and information specialist roles in the publication process are under-estimated even when metadata work is delegated (perhaps to curators at remote locations). And second, organizational arrangements are required that take into account relations between data in the differing locations of a web of repositories. While 'supersized' efforts tend to overshadow other arrangements, we argue that local information infrastructure plays an essential role, particularly in 'network science' configurations.

Data providers sometimes see the issue of data sharing as one that goes to the heart of science and as accompanied by a responsibility to fully describe published results. If data are to become a published result, it is essential that data are 'well-documented' so as to be interpretation ready. The processing, analysis and interpretation of data that are part of the scientific research process itself – requiring skill and insight to validate – must be captured and recorded in some form as metadata. Traditionally, data have been shared by informal communication among peers so articulation work is required to make explicit the contextualization for metadata (Linde 2001, Latour and Woolgar 1979). Resistance from scientists to share their research data is often presented as one of the main challenges to scientific collaboration (Costello 2009, ESA 2009). Reluctance to share has been identified as including a fear of not receiving adequate acknowledgement or benefit from personal efforts (Birnholtz and Beitz 2003). We suggest, however, that the taxonomy of data-sharing behaviour is incomplete; there are additional reasons for resistance that go beyond data ownership and intellectual property issues. There are scientifically salient concerns about the lack of maturity of data classification efforts, the risk of misinterpretation of complex data and inadequate support for local information environments that limit response-ability. A resistance to propagating ill-described data to an audience unfamiliar with the field's data handling issues is an emerging concern. This concern appears subsequent to an early surge of optimism that assumed technology would solve problems in assembling digital data, a period during which non-sharers have been labeled 'uncooperative' and perceived as 'data hoarders'. Description of complex data is the subject of ongoing research as experience grows with rule-based logic, shared vocabularies, domain ontologies and community semantics.

The concept of 'distance-from-data-origin' described by Baker and Yarmey (2009) provides a mechanism for distinguishing two categories of data repositories, both carrying out work associated with the data lifecycle. Data handling at local sites is associated with data collecting by researchers who planned or know of the original intended use of the data. Data handling at sites that are 'remote' focuses on obtaining data collections and preparations for their reuse. Examples of remote sites include disciplinary centers such as the National Ocean Data Center or the National Center for Ecological Analysis Synthesis and the National Oceanographic Data Center. The distance involved is presented as a sociotechnical distance rather than a geometric distance, with differing characterizations of data handling at local sites and remote centres. The data analysis subcycle captures the data work occurring at local repositories (as described by case 1: knowledge production process for data use). In counterpoint, the data curation subcycle dominates at remote data repositories with a focus on data reuse. When data are transported beyond the site of data origin, data decisions generally are made by technologists, curators and data specialists at centers or archives that due to their larger scope consider issues at a higher level of organization and with an emphasis on data similarities. This view provides insights from the perspective of comparative and statistical analysis. Information infrastructuring includes establishing and maintaining or adjusting the arrangements required to facilitate data publication in repositories through support of the full data cycle, that is, both data analysis and data curation subcycles in local and remote repositories, respectively.

A multi-level configuration for data handling, in this case local (by information managers) and remote (by data specialists at centres or archives) depends upon the existence of a fully-formed infrastructure consisting of at least three components: a local information environment with its attendant deep knowledge of local contexts and data practices; a remote centre with its attendant communication procedures and data standards; and finally, some type of negotiated relations between the two. With exchange between local and remote sites, there arises the concept of federation where federation is defined as the act of uniting multiple states or sites where each retains control over its own internal affairs. Federation of data repositories entails extending information infrastructure to provide coordination between repositories.

There has been continued and new funding for remote repositories – disciplinary centres of excellence, national archives and libraries – that has stimulated development of resource-centred repositories both conceptually and technologically. As data curation models have matured, more of the facets and interdependencies of the work have been identified. Data curators, frequently detached from local projects and the data analysis subcycle, carry out mediation and translation; they are often concerned with reuse of the data rather than with the original intended use. While resource-centred data curation focuses on broader cross-comparative views and larger scale data resolutions, research-centred data analysis addresses a wide range of scales, resolutions and irregular cases. Thus, we see advantages to taking into account the support of an active local infrastructure as one component of a multi-component network model. In contrast to a more traditional pipeline model of data travelling linearly from individual collector to a single centralized system (Baker and Millerand 2007), a 'web of repositories' network model recognizes contributions made by multiply-connected, diverse components – local and remote.

Conclusions

By facilitating the move of heterogeneous ecological datasets into the digital realm, information infrastructure growth makes evident a great deal about the complexity of data and of collaboration. Critical elements for scientific collaboration are the recognition of the roles and the sociotechnical dimensions associated with both data and infrastructure.

Understanding and documenting the differences in data requires continuing work by ecological researchers in partnership with information specialists with particular focus on the data analysis subcycle. While expectations for data sharing and data publication are already in place, there is further mediation of the data required in terms of processing, testing and assessment – in addition to innovation in addressing unanticipated data and infrastructure issues. Traditionally there has been little interest and little time for considering data documentation and organization explicitly. As a consequence, distinctions between types of data work and knowledge production processes are under-studied resulting in an under-appreciation of the work involved in online dataset publication.

In making data public, there are questions with respect to what data and what form of the data to make public. For instance, a non-biologist may want to work with data that have been severely quality controlled while a biologist familiar with the data at hand may prefer not to have data anomalies pre-interpreted and filtered out. When thinking about data policies, people traditionally focus on what is being withheld (the desire to keep data private), but another way of thinking about policy is with respect to what is best provided. As this topic develops, so too must data provenance, not just in terms of concern with individual and organizational credit but also more broadly including the full history of decision-making associated with the data. As Bowker puts it: 'What is needed is a record of processes as well as a record of facts' (with facts being data here) in order to have usable data available (Bowker 2001: 664). Needed are methods for describing and documenting data quality, an underdeveloped area of increasing importance if data are to circulate.

Funding opportunities influence developments in scientific data sharing and infrastructure growth. NSF added a second criterion for evaluation of scientific research in 1997 to the original one of intellectual merit in order to stimulate new types of learning and wider engagement. The second criterion addresses broader impacts and has implications for growth of information infrastructure as well as a growth of scientific networks. It provides a new framework when issues of data and infrastructure are considered central to 'broader impacts' rather than as peripheral. The lack of such a framing is evident in a recent explanation of the criterion that details the 'benefits of the proposed activity to society' and the need to 'analyze, interpret and synthesize research and education results in formats understandable and useful for non-scientists' (NSF 2007b). There is an assumption here of the ability to collaborate with non-science communities at a time when collaboration *within* scientific communities is limited by existing data practices and lack of infrastructure. It is worthwhile to take note that data access is a new practice; it is a new benefit not only to society but to scientific communities as well. Thus in addition to the broader impacts criterion supporting 'training and outreach' focused on society, the second criterion could be expanded in terms of 'training and outreach' focused on 'science and society' or, alternatively, a third criterion could be added specifically pertaining to data and information. Highlighting information literacy together with information dissemination would stimulate and support new types of information environments, effectively expanding the reach of 'broader impacts' to include new types of collaboration, learning and infrastructure within scientific and public communities alike.

There is a need for information infrastructure; there is need for diverse types of information infrastructure. Growth of in-situ, local-scale infrastructure addresses issues relating to data production of heterogeneous ecological data at research sites and supports development of processes for sharing datasets on a larger scale, both within a research community and across a domain. New responsibilities and roles are associated with new types of data, data work and knowledge production. Development within the LTER (Figure 6.1), illustrates scientific arrangements for data that require thinking about data practices and scientific work in new ways. Simple notions of data sharing are developing into an understanding of data publication involving a variety of types of data repositories and of information environments concerned with assembly, exchange and curation of data.

Cyberinfrastructure initiatives carry lots of promises in terms of potentially leading to more global science – given changed methods and research objectives. However, these efforts are nascent and their impacts complex to assess. Our study focuses on situated information infrastructures and information environments as elements that support collaborative scientific research within an ecological research community configured as a scientific network able to address larger-scale science questions. The concept of infrastructuring captures the ongoing, active process of growing and (re)negotiating relations between sites in a network as well as between local and remote data repositories at a time when information management is grappling with classification and description of complex biological data. Failing to take account of complexity in data (the types and structures, the sampling and analysis differences, the biases and errors) and in metadata (the tacit, contextual and procedural) leads one to plan solutions rather than developing continuing, collaborative processes. Participants preparing cyberinfrastructure – grids, hubs and automated systems – today frequently have large-scale, standards-based models in mind that involve archives and data centres for making data available. While large-scale undertakings dominate current infrastructure development planning for science envisioned as 'Big Science', our ethnographic account of local-scale infrastructure development illustrates the variety that exists in the information infrastructure landscape. Rather than an alternative to cyberinfrastructure, development of local-scale, research-centred infrastructures may be understood as complementary efforts supporting the local use of data and the continuing identification of data characteristics and practices critical to publication of data. Diverse, evolutionary infrastructure arrangements and new data practices can facilitate scientific collaborations that depend upon the sharing of high quality data.

References

Arms, W.Y. and Larsen, R.L. 2007. The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship: Report of a NSF workshop in Phoenix, Arizona, April 17–19: National Science Foundation.

Aronova, E., Baker, K.S. and Oreskes, N. submitted. From the International Geophysical Year to the International Biological Program: Big science and big data in biology, 1957–present. *Historical Studies in the Natural Sciences*.

Atkins, D.E. et al. 2003. Revolutionizing Science and Engineering through Cyberinfrastructure. Report of the NSF blue-ribbon Advisory Panel on Cyberinfrastructure: National Science Foundation. Available at: http://www.communitytechnology.org/nsf_ci_report [accessed: 28 December 2009].

Arzberger, P.W., Schroeder, P., Beaulieu, A., Bowker, G.C., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P. 2004. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3(29), 135–52.

Baker, K.S., Benson, B., Henshaw, Blodgett, D., Porter, J. and Stafford, S.G. 2000. Evolution of a multi-site network information system: The LTER information management paradigm. *BioScience*, 50(11), 963–83.

Baker, K.S. and Chandler, C. 2008. Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. *Deep Sea Research II*, 55(18–19), 2132–42.

Baker, K.S. and Millerand, F. 2007. *Scientific Infrastructure Design: Information Environments and Knowledge Provinces*, American Society of Information Science and Technology, Milwaukee, WI, 19–24.

Baker, K.S., Ribes, D., Millerand, F. and Bowker, G.C. 2005. *Interoperability Strategies for Scientific Cyberinfrastructure: Research and Practice*, American Society for Information Systems and Technology Conference, Charlotte, North Carolina, October 28–November 2.

Baker, K.S. and Yarmey, L. 2009. Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation*, 4(2), 12–27.

Berman, H.M., Bourne, P.E. and Westbrook, J. 2004. The protein data bank: A case study in management of community data. *Current Proteomics*, 1, 49–57.

Birnholtz, J. and Bietz, M. 2003. *Data at Work: Supporting Sharing in Science and Engineering*, ACM Conference on Supporting Group Work, Sanibel Island, Florida, 9–12 November, 2003.

Bocking, S. 2010. Organising the Field: Collaboration in the History of Ecology and Environmental Science, in *Collaboration in the New Life Sciences*, edited by J.N. Parker, N. Vermeulen and B. Penders. Farnham: Ashgate.

Borgman, C.L., Wallis, J. and Enyedy, N. 2008. Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal of Digital Libraries*, 17(1), 17–30.

Bowker, G.C. 2001. Biodiversity, data diversity. *Social Studies of Science*, 30, 643–84.

Bowker, G.C. 2006. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, G.C., Baker, K.S., Millerand, F. and Ribes, D. 2010. Towards information infrastructure studies: Ways of knowing in a networked environment, in *International Handbook of Internet Research*, edited by J.Hunsinger et al. New York: Springer, 97–118.

Bowker, G.C. and Star, S.L. 2000. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Briscoe, M.G. 2008. Collaboration in the ocean sciences: Best practices and common pitfalls. *Oceanography*, 21(3), 58–65.

Brunt, J., Benson, B., Vande Castle, J., Henshaw, D. and Porter, J. 2007. LTER Network Cyberinfrastructure Strategic Plan, Version 4.2.

Callahan, J.T. 1984. Long-term ecological research. *BioScience*, 34(6), 363–7.

Carlson, S. and Anderson, B. 2007. What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, 12(2). Available at: http://jcmc.indiana.edu/vol12/issue2/carlson.html [accessed: 28 December 2009].

CCSDS, Consultative Committee for Space Data Systems 2002. Reference Model for an Open Archival Information System (OAIS). Blue Book. Washington, DC: CCSDS. Available at: http://public.ccsds.org/publications/archive/650x0b1.pdf [accessed: 28 December 2009].

Collins, H.M. and Evans, R. 2002. The third wave of science studies: Studies of expertise and experience. *Social Studies of Science*, 32(2), 235–96.

Collins, H.M. and Evans, R. 2007. *Rethinking Expertise*. Chicago, IL: University Of Chicago Press.

Costello, M.J. 2009. Motivating online publication of data. *BioScience*, 59(5), 418–27.

DEB, Divison of Environmental Biology, National Science Foundation. Available at: http://www.nsf.gov/bio/deb/about.jsp [accessed: 28 December 2009].

Edwards, P.N., Jackson, S.J., Bowker, G.C. and Knobel, C.P. 2007. *Understanding Infrastructure: Dynamics, Tensions and Design*. Report of the workshop: History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures. Office of Cyberinfrastructure. National Science Foundation.

Edwards, P.N., Jackson, S.J., Bowker, G.C. and Williams, R. 2009. Introduction: An agenda for infrastructure studies. *Journal of the Association for Information Systems*, 10(5), 364–74.

ESA, Ecological Society of America 2009. Incentives for data sharing, in *Ecology, Evolution, and Organismal Biology*. Workshop Report. Washington, DC: National Science Foundation.

Franklin, J.F., Bledsoe, C.S. and Callahan, J.T. 1990. Contributions of the Long-Term Ecological Research Program: An expanded network of scientists, sites, and programs can provide crucial comparative analyses. *Bioscience*, 40(7), 509–23.

Frey, J. 2008. Curation of laboratory experimental data as part of the overall data lifecycle. *International Journal of Digital Curation*, 3(1), 44–62.

Furner, J. 2003. Little book, big book: Before and after little science, big science, a review article Part 1. *Journal of Librarianship and Information Science*, 35, 115–25.

Galison, P. 1997. Three laboratories. *Social Research*, 64(3), 1127–55.

GAO 2007. United States Government Accountability Office. Agencies Have Data-Sharing Politics but Could Do More to Enhance the Availability of Data from Federally Funded Research. Report GAO 07-1172. Available at: http://www.gao.gov/products/GAO-07-1172 [accessed: 28 December 2009].

Geertz, C. 1973. Thick description: Toward an interpretive theory of culture, in *The Interpretation of Cultures*, edited by C. Geetz. New York, NY: Basic Books, 3–30.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P. and Trow, M. 1994. *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: Sage Publications.

Golley, F.B. 1993. *A History of the Ecosystem Concept in Ecology: More Than the Sum of the Parts*. New Haven, CT: Yale University Press.

Hackett, E.J. 2005. Introduction to the special guest-edited issue on scientific collaboration. *Social Studies of Science*, 35(5), 667–71.

Hacking, I. 1992. The self-vindication of the laboratory sciences, in *Science as Practice and Culture*, edited by A. Pickering. Chicago, IL: University of Chicago Press, 29–64.

Hanseth, O., Monteiro, E. and Hatling, M. 1996. Developing information infrastructure: The tension between standardization and flexibility. *Science, Technology and Human Values*, 21(4), 407–26.

Hart, J.K. and Martinez, K. 2006. Environmental sensor networks: A revolution in the earth system science? *Earth-Science Reviews*, 78(3–4), 177–91.

Higgins, S. 2008. The DCC curation lifecycle model. *International Journal of Digital Curation* 3(1), 134–40. Available at: http://www.ijdc.net/index.php/ijdc/article/view/69/69 [accessed: 28 December 2009].

Hilgartner, S. 1995. Biomolecular databases: New communication regimes for biology? *Science Communication*, 17(2), 240–63.

Hobbie, J.E., Carpenter, S.R., Grimm, N.B., Gosz, J.R. and Seastedt, T.R. 2003. The US Long Term Ecological Research Program. *BioScience*, 53(1), 21–32.

Jackson, S.J., Edwards, P.N., Bowker, G.C. and Knobel, C.P. 2007. Understanding infrastructure: History heuristics, and cyberinfrastructure policy. *First Monday*. Available at: http://www.firstmonday.org/issues/issue12_6/jackson/index.html [accessed: 28 December 2009].

Jones, M.B., Schildhauer, M., Reichman, O.J. and Bowers, S. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37, 519–44.

Karasti, H. and Baker, K.S. 2004. *Infrastructuring for the Long-term: Ecological Information Management,* Hawaii International Conference for System Science, Hawaii, Big Island, January 2004.

Karasti, H. and Baker, K.S. 2008. Digital data practices and the global Long Term Ecological Research Program. *International Journal of Digital Curation*, 3(2), 42–58.

Karasti, H., Baker, K.S. and Millerand, F. submitted. Infrastructure time: Long-term matters in collaborative development. *Journal of Computer Supported Cooperative Work.*

Knorr Centina, K. 1999. *Epistemic Cultures: How the Sciences Make Knowledge.* Cambridge, MA: Harvard University Press.

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge: Harvard University Press.

Latour, B. 1990. Drawing things together, in *Representation in Scientific Practice*, edited by M. Lynch and S. Woolgar. Cambridge, MA: MIT Press, 19–68.

Latour, B. and Woolgar, S. 1979. *Laboratory Life: The Social Construction of Scientific Facts.* Beverly Hills, CA: Sage Publications.

Leonelli, S. 2007. Weed for Thought, Using Arabidopsis thaliana to Understand Plant Biology. PhD Thesis. Amsterdam: Vrije Universiteit.

Leonelli, S. 2008. Circulating evidence across research contexts: The locality of data and claims in model organism research, in *The Nature of Evidence: How Well Do 'Facts' Travel?* Report Number 25. Economic History Department, The London School of Economics and Political Science.

Lewontin, R.C. 2000. *The Triple Helix: Gene, Organism, and Environment.* Cambridge: Harvard University Press.

Linde, C. 2001. Narrative and social tacit knowledge. *Journal of Knowledge Management*, 5(2), 160–70.

Lowry, R., Bermudez, L. and Graybeal, J. 2006. Semantic interoperability: A goal for marine data management. ICES International Council for Exploration of the Sea. CM Environmental fisheries data management, access, and integration.

McDonough, J. 2008. *Structural Metadata and the Social Limitation of Interoperability: A Sociotechnical View of XML and Digital Library Standards Development*, The Markup Conference, Montreal, Canada, 12–15 August 2008. Available at: http://www.balisage.net/Proceedings/html/2008/McDonough01/Balisage2008-McDonough01.html [accessed: 28 December 2009].

Merton, R.K. 1968. *Social Theory and Social Structure*. New York: Free Press.

Michener, W.K. 2000. Metadata, in *Ecological Data: Design, Management and Processing*, edited by W.K. Michener and J.W. Brunt. Malden, MA: Blackwell Science, 92–116.

Michener, W.K. and Brunt, J.W. 2000. *Ecological Data: Design, Management and Processing*. Malden, MA: Blackwell Science.

Michener, W.K., Baerwald, T.J., Firth, P., Palmer, M.A., Rosenberg, J.L., Sandlin, E.A. and Zimmerman, H. 2001. Defining and unraveling biocomplexity. *BioScience*, 51, 1018–23.

Millerand, F. and Bowker, G.C. 2009. Metadata standards: Trajectories and enactment in the life of an ontology, in *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*, edited by S.L. Star and M. Lampland. New York, NY: Cornell University Press, 149–76.

Nardi, B. and O'Day, V. 1999. *Information Ecologies: Using Technology with Heart*. Cambridge, MA: MIT Press.

NOAA/NESDIS 2008. *Comprehensive Large Array-data Stewardship System (CLASS)*, Information Heterogeneity White Paper. November 2008. US Department of Commerce. NOAA/NESDIS CLASS-1215-CLS-WHT-IHET.

NRC, National Research Council 2007. *Environmental Data Management at NOAA: Archiving, Stewardship, and Access*. Washington, DC: National Academies Press.

NSB, National Science Board 2005. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. Arlington, VA: National Science Foundation.

NSF, National Science Foundation 2007a. *Cyberinfrastructure Vision for 21st Century Discovery*. Washington, DC: National Science Foundation Cyberinfrastructure Council.

NSF, National Science Foundation 2007b. *Merit Review Broader Impacts Criterion: Representative Activities (July 2007)*. Available at: http://www.nsf.gov/pubs/gpg/broaderimpacts.pdf.

NSF, National Science Foundation 2009. Grant General Conditions (GC-1). Washington, DC: National Science Foundation.

Odum, E. and Barrett, G.W. 2005. *Fundamentals of Ecology*. Fifth Edition. Belmont, CA: Thomson Brooks/Cole.

O'Neill, R.V. 2001. Is it time to bury the ecosystem concept? (with full military honors, of course!). *Ecology*, 82(12), 3275–84.

Parker, J.N. 2010. Integrating the social into the ecological: Organizational and research group challenges, in *Collaboration in the New Life Sciences*, edited by J.N. Parker, N. Vermeulen and B. Penders. Farnham: Ashgate.

Parsons, M.A. and Duerr, R. 2005. Designating user communities for scientific data: Challenges and solutions. *Data Science Journal*, 4, 31–38.

Peters, D.P.C., Groffman, P.M., Nadelhoffer, K.J., Grimm, N.B., Collins, S.L., Michener, W.K. and Huston, M.A. 2008. The changing landscape: Ecosystem responses to urbanization and pollution across climatic and societal gradients. *Frontiers in Ecology*, 6(5), 229–37.

Pickett, S.T.A., Kolasa, J. and Jones, C.G. 2007. *Ecological Understanding: The Nature of Theory and the Theory of Nature*. Amsterdam: Elsevier.

Pickstone, J. 2001. *Ways of Knowing: A New History of Science, Technology, and Medicine*. Chicago, IL: University of Chicago Press.

Pipek, V. and 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), 447–73.

Polanyi, M. 1967. *The Tacit Dimension*. New York, NY: Anchor Books.

Powell, T.M. and Steele, J.H. 1995. *Ecological Time Series*. New York, NY: Chapman and Hall.

Price, J.D. 1963. *Little Science, Big Science*. New York, NY: Columbia University Press.

Ribes, D., Baker, K.S., Millerand, F. and Bowker, G.C. 2005. *Comparative Interoperability Project: Configurations of Community, Technology, and Organization*, ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, CO, 7–11 June.

RIN, Research Information Network 2008. *To Share or Not to Share: Publication and Quality Assurance of Research Outputs*. Natural Environment Research Council RIN Report.

Romanello, S., Beach, J., Bowkers, S., Jones, M. and Ludascher, B. 2005. *Creating and Providing Data Management Devices for the Biological and Ecological sciences: Science Environment for Ecological Knowledge*, International Conference on Scientific and Statistical Database Management, Santa Barbara, California, 2005.

Star, S.L. and Bowker, G.C. 2002. How to infrastructure, in *Handbook of New Media: Social Shaping and Consequences of ICTs*, edited by L.A. Lievouw and S. Livingstone. London: Sage Publications, 151–62.

Star, S.L. and Lampland, M. 2009. *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. New York, NY: Cornell University Press.

Star, S.L. and Ruhleder, K. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7, 111–33.

Trist, E. 1997. Referent organizations and the development of inter-organizational domains, in *The Social Engagement of Social Science: A Tavistock Anthology. Vol III: The Socio-Ecological Perspective*, edited by E. Trist, F. Emery, H. Murray and B. Trist. Philadelphia, PA: The University of Pennsylvania Press.

Vermeulen, N. 2009. *Supersizing Science: On Building Large-Scale Research Projects in Biology*. Maastricht: Universitaire Pers Masstricht.

Wallis, J.C., Borgman, C.L., Mayernik, M.S. and 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), 114–26.

Watts, D.J. 2003. *Six Degrees: The Science of a Connected Age*. New York, NY: W.W. Norton and Company.

Weick, K.E. 1979. *The Social Psychology of Organizing*. Reading, MA: Addison-Wesley.

Weinberg, A.M. 1967. *Reflections on Big Science*. Oxford: Pergamon Press.

Westfall, C. 2003. Rethinking big science: Modest, mezzo, grand science the development of Bevalac, 1971–1993. *The History of Science Society ISIS*, 94(1), 30–56.

Whitley, E.A. 2000. Tacit and Explicit Knowledge: Conceptual Confusion around the Commodification of Knowledge, Knowledge Management: Concepts and Controversies Conference, University of Warwick, 10–11 February, 62.

Wouters, P. and Schroder, P. 2003. Promise and practice in data sharing: Networked research and digital information, in *The Public Domain of Digital Research Data*, edited by P.Wouters and P. Schroder. Amsterdam: Nerdi, NIWI-KNAW.

Zimmerman, A. 2008. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, and Human Values*, 33(5), 631–52.

Zimmerman, A. and Nardi, B. 2010. The competition to be big: An analysis of LTER and NEON, in *Collaboration in the New Life Sciences*, edited by J.N. Parker, N. Vermeulen and B. Penders. Farnham: Ashgate.

For our mentors

Collaboration in the New Life Sciences

Edited by
JOHN N. PARKER
National Center for Ecological Analysis and Synthesis, USA
NIKI VERMEULEN
Department of Social Studies of Science, University of Vienna, Austria
and
BART PENDERS
Maastricht University and Radboud University Nijmegen,
The Netherlands

ASHGATE

Published by
Ashgate Publishing Limited
Wey Court East
Union Road
Farnham
Surrey, GU9 7PT
England

Ashgate Publishing Company
Suite 420
101 Cherry Street
Burlington
VT 05401-4405
USA

www.ashgate.com

British Library Cataloguing in Publication Data
Collaboration in the new life sciences.
1. Life sciences--Research. 2. Intellectual cooperation.
3. Group work in research. 4. Life scientists--
Professional relationships.
I. Parker, John N. II. Vermeulen, Niki. III. Penders, Bart.
570.7'2-dc22

Library of Congress Cataloging-in-Publication Data
Parker, John N.
Collaboration in the new life sciences / [edited] by John N. Parker, Niki Vermeulen, and Bart Penders.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-7546-7870-0 (hbk) -- ISBN 978-0-7546-9717-6 (ebk)
1. Life sciences--Social aspects. 2. Cooperation. I. Vermeulen, Niki. II. Penders, Bart.
III. Title.
QH333.P37 2010
570.72--dc22

2010017061

ISBN 9780754678700 (hbk)
ISBN 9780754697176 (ebk)

Printed and bound in Great Britain by
TJ International Ltd, Padstow, Cornwall

Contents

PART I COLLABORATION ARISES

PART II COLLABORATION IN ECOLOGY AND THE ENVIRONMENT

PART III COLLABORATION IN THE MOLECULAR LIFE SCIENCES

PART IV COLLABORATION IN THE NEW LIFE SCIENCES REVISITED

List of Figures

Foreword

Science studies scholars of the 1960s and 1970s first glimpsed the extent and significance of scientific collaboration through networks of co-authorship derived from the by-lines of articles. In some cases these data were gathered laboriously by undergraduate research assistants who transcribed information from articles, while in other cases the data were obtained through the less visible labor of the staff of the Institute for Scientific Information, which published *Science Citation Index* (in bi-monthly paperback volumes that were aggregated into an annual hardcover). Whatever the means, the effort was justified by the central role collaboration was understood to play in structuring science and shaping the conduct of research. At the time, the driving concern of science studies was to define and depict the fundamental social structures of science: the array of disciplines, specialties, research schools, and such that constituted the community of scientists and patterned their informal communications, publication decisions, career trajectories, choice of research problems and methods, and access to resources. Such studies of patterned interaction and structure, accumulated over time, afforded insight into the dynamics of science. Since publication is the existential act of science, and publications are the artifact that remains as evidence that the existential research act was consummated, scholars of the day reasonably analyzed the co-authorship of articles in order to discover the collaborative patterns that constitute the structure and genealogy of science.

Even with the crude tools of the day we were stunned by what we found. Network techniques applied to bibliometric data revealed durable and dynamic patterns of association that indicated the birth of a specialty, the formation of opposing research schools, the arrival of transdisciplinary pioneers, the departures of the disenchanted, and, on rare occasion, the metamorphosis of a specialty into a fully-fledged scientific discipline. Collaboration, we knew, was the conjugal act that drove such processes; co-authorship was its residue, something one analyzed, almost archaeologically, for residual clues to a process we could only dimly perceive. But that was then...

Ensuing decades witnessed an explosion of perceptive and systematic studies of the process and substance of scientific work: life in the laboratory, replete with inscription and *bricolage*, do-able problems fashioned of epistemic things, scientists wrestling with their ensembles of research technologies and the essential tensions of collaborative work. Driven by a complex of intellectual, technical, and policy forces, collaboration has become the generative act of contemporary science, and in consequence studies of collaboration are moving from the periphery toward the center of analysis in science studies. Progress in this line of inquiry will require an accumulation of studies informed by history and theory and motivated by policy and practice that carefully situate their subjects in the spaces of science, methodology, geography, composition, motivation, and more. To this end, editors John N. Parker, Niki Vermeulen, and Bart Penders have curated some varied and revealing specimens of *Collaboration in the New Life Sciences*, drawn from the dynamic fields of ecology and oceanography, genomics and systems biology, proteomics, toxicology, and bioinformatics. From these perceptive chapters we learn how nature, data, and purpose shape scientific collaboration, how new research technologies and extensive collaborative networks—the growing scale and scope of science—frame the organization of inquiry, how the confluence of diverse intellectual currents change the work arrangements of scientists, and how properties of place, on the globe and in the intellectual firmament, alter the circumstances and conduct of science. In all, these detailed and insightful studies of scientific collaboration will help us embark upon rich and generative pathways of inquiry.

We are now about a decade into what some have called the century of the life sciences. Powerful theories aspire to integrate evolution, development, and environment into a grand synthesis that connects the origins of species with the development of individuals, embedded within dynamic ecosystems driven by complex and coupled human and biogeophysical systems: a Grand Unified Theory of life. Whether the goal is ever reached may matter less than the audacity of its pursuit: in this quest the life sciences will develop research technologies of unparalleled scope and precision, exploring the intimate dynamics of genes, proteins, and their constituents while monitoring ecosystems of unprecedented size and complexity at resolutions and for durations unimaginable just a few decades ago. Cyber-enabled technologies will allow such data to be stored, integrated, analyzed, and modeled by teams of scientists working together at times and apart at other times, sometimes synchronously and sometimes not. Climate change, resource scarcity, and economic demands in the developed and developing world increase the stakes and add urgency and gravity to the endeavor. To effect the interdisciplinary synthesis, to work with massive and complex data, to meet the demands of policy and decision makers, and to engage diverse publics will require modes of collaboration more extensive and involved than any seen to date.

Collaboration will create the new life sciences and, in turn, the quest for these new sciences will create new forms of collaboration, new means of doing science, and new ways of being a scientist. Science studies will be present at the creation of these new sciences and new ways of doing science: no longer is the study of collaboration relegated to the archaeology of publications or an outsider's occasional foray into the laboratory, but it is becoming a welcome and valuable component of the process of innovation in the organization and conduct of research. For those of us who lived and survived the "science wars," this is vindication and victory of the most powerful variety. Our growing engagement with science and its publics promises not only to generate rich empirical material

that will advance knowledge, but also to impose new responsibilities on the science studies community as our colleagues in the sciences, in science policy, among decision makers, and in the wider public look to us for guidance in meeting the challenges ahead.

Professor Edward J. Hackett
Arizona State University