Science friction: Data, metadata, and collaboration

Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker and Christine L. Borgman

Social Studies of Science 2011 41: 667 originally published online 15 August 2011
DOI: 10.1177/0306312711413314

S|S|S

Social Studies of Science
41(5) 667–690

DOI: 10.1177/0306312711413314
sss.sagepub.com
SAGE

Science friction: Data, metadata, and collaboration

Paul N. Edwards
School of Information, University of Michigan, Ann Arbor, MI, USA

Matthew S. Mayernik
Graduate School of Education and Information Studies, UCLA, CA, USA

Archer L. Batcheller
School of Information, University of Michigan, Ann Arbor, MI, USA

Geoffrey C. Bowker
School of Information Sciences, University of Pittsburgh, PA, USA

Christine L. Borgman
Graduate School of Education and Information Studies, UCLA, CA, USA

Abstract

When scientists from two or more disciplines work together on related problems, they often face what we call 'science friction'. As science becomes more data-driven, collaborative, and interdisciplinary, demand increases for interoperability among data, tools, and services. Metadata – usually viewed simply as 'data about data', describing objects such as books, journal articles, or datasets – serve key roles in interoperability. Yet we find that metadata may be a source of friction between scientific collaborators, impeding data sharing. We propose an alternative view of metadata, focusing on its role in an ephemeral process of scientific communication, rather than as an enduring outcome or product. We report examples of highly useful, yet ad hoc, incomplete, loosely structured, and mutable, descriptions of data found in our ethnographic studies of several large projects in the environmental sciences. Based on this

Corresponding author:
Paul N. Edwards, School of Information, University of Michigan, 3439 North Quad, 105 S. State St., Ann Arbor, MI 48109-1285, USA.
Email: pne@umich.edu

evidence, we argue that while metadata products can be powerful resources, usually they must be supplemented with metadata processes. Metadata-as-process suggests the very large role of the ad hoc, the incomplete, and the unfinished in everyday scientific work.

Keywords
collaboration, communication, data, metadata

Humanity is now in the business of managing the planet (Elichirigoity, 1999; Serres, 1995, 2007). As the world population has soared over the last 150 years, people have commandeered an ever larger percentage of the incoming solar energy, whether directly by converting it to electricity, or indirectly by harnessing it through biofuels, agriculture, forestry, and use of ecosystem services. According to recent estimates, human beings appropriate about 24 percent of Earth's potential net primary productivity (a measure of the biomass available in terrestrial ecosystems) each year, and approximately 83 percent of the world's land surface is directly influenced by human activity (Haberl et al., 2007; Sanderson et al., 2002). Meanwhile, humanity is provoking very rapid climatic change as well as one of the largest extinction events in the history of life on Earth, even while seeking ways to mitigate the most dramatic of these effects.

Monitoring and managing all this – to the extent that we can – requires vast amounts of observational data, coordinated across a bewildering multitude of so-called scientific disciplines. Meanwhile, the explosion of computer processing power and storage capacity has transformed the sciences' ability to find, use, coordinate, and re-use these data. This paper explores issues arising from this new environment, which some go so far as to call a 'fourth paradigm' of scientific work driven by the availability of large datasets, wherein patterns may be sought directly rather than through more traditional hypothetico-deductive methods (Hey et al., 2009).

Science studies has probed many kinds of data problems within particular scientific disciplines, such as contested interpretations of data, relations between database structures and data collecting practices, questions about when and why certain instrument readings count as data, the 'experimenter's regress', and boundaries between documents and data (Bowker, 2000, 2005; Bowker and Star, 1999; Buckland, 1991, 1997; Collins, 1985; Collins and Pinch, 1993; Zimmerman, 2007). Yet our field has rarely considered how data travel *among* diverse disciplines; as sociologists of science, we have tended to look under the lamppost of whatever field we happen to know. It's interesting (and hard) enough to explicate memory practices within one discipline – why learn five?

Science studies has developed useful ideas about how theories, concepts, specimens, maps, instruments, and other elements of scientific practice travel across various divides: from theoretical to experimental subfields, from professionals to amateurs, from scientists to engineers, and so on. Keystone STS phrases such as 'boundary objects', 'immutable mobiles', 'virtual witnessing', and 'trading zones' help make sense of these transitions (Galison, 1996; Latour, 1987; Shapin and Schaffer, 1985; Star and Griesemer, 1989; Strathern, 2004). There is also, of course, a large literature on the unpacking of data during episodes of scientific or technical controversy (Collins, 1985; Collins and Pinch, 1993; Kevles, 1998; Vaughan, 1996). Yet most of this work

has focused either on higher-level results, products and artifacts, or on mutable interpretations of evidence, rather than on the travels of data per se: data function as an actors' category, as in the cases of collections of instrument readings, field observations, model outputs, and so on, which represent the daily work of science. As datasets become increasingly commoditized, 'mined', and exchanged among distant disciplines, this area needs much closer scrutiny.

Our traditional STS approach to data in science resembles the traditional approach of historians to history. They write national histories because the archives (data) are national; no matter that many real historical processes stubbornly exceed national boundaries (Braudel, 1975; Michelet, 1930; Wallerstein, 1976). And no matter, in our own case, that much of today's most interesting and important science operates between domains. The Comtean hierarchy of physics, chemistry, and biology as driving disciplines is long gone, replaced by a massive proliferation of interdisciplines. Nowhere is this more true than in the Earth and environmental sciences – sciences upon which humanity relies for its overweening yet unavoidable goal of planetary management. Unlike previous macro-paradigms of scientific work, in which data were treated as the private (and closely held) property of individuals or laboratories, in these interdisciplinary domains data need to travel far and wide. It is time for science studies to investigate how data traverse personal, institutional, and disciplinary divides.

Science friction

Friction resists and impedes. At every interface between two surfaces, friction consumes energy, produces heat, and wears down moving parts. Edwards' metaphor of *data friction* describes what happens at the interfaces between data 'surfaces': the points where data move between people, substrates, organizations, or machines – from one lab to another, from one discipline to another, from a sensor to a computer, or from one data format (such as Excel spreadsheets) to another (such as a custom-designed scientific database) (Edwards, 2010). Every movement of data across an interface comes at some cost in time, energy, and human attention. Every interface between groups and organizations, as well as between machines, represents a point of resistance where data can be garbled, misinterpreted, or lost. In social systems, data friction consumes energy and produces turbulence and heat – that is, conflicts, disagreements, and inexact, unruly processes.

Data friction leads inevitably to what we call 'science friction': the difficulties encountered when two scientific disciplines working on related problems try to interoperate. To take a prominent example, consider the tension between weather forecasting and climatology, separate fields within the disciplinary landscape of meteorology. Weather forecasters have been collecting daily observations since the 1850s. In service of their chief goal – accurate near-term forecasting – their priority is swift communication and constant improvement of observing and forecasting systems. Even week-old data have little value for tomorrow's forecast, so until recent decades forecasters placed a low priority on storing, cataloguing, and accessing historical weather data.

Meanwhile, climatologists average daily weather data to create long-term climate statistics. To do this, they need data from the whole world over periods of many decades.

Some climate data come from instruments and observing stations specifically designed for climate studies. But the majority of data used by climatologists come from the weather forecast system. Weather stations frequently change instruments, locations, and observing techniques; over time, they may operate intermittently, change their procedures, or even change nationality after political upheavals. Throughout the history of meteorology, weather data from different parts of the Earth encountered friction at political borders, institutional boundaries, and technical interfaces between national observing systems. Many data either never reached central collectors, or reached them only in processed forms that turned out to be riddled with errors. Therefore, climatologists regard data from the forecast system as unstable. To incorporate these sources in 'climate quality' datasets, climate scientists recover their histories and adjust, analyze, and reanalyze the observations, often down to the level of individual instrument readings. Similarly, data from satellite instruments designed specifically for weather observation have been commandeered to measure the temperature of the lower troposphere (through complex data modeling), creating intense controversy over how such data should be processed and understood (Edwards, 2010).

This data friction results in enormous expenditures of time, energy, and attention, which can lead to other kinds of science friction as well. Take the so-called 'Climategate' controversy over emails and data stolen from the University of East Anglia's Climatic Research Unit (CRU) in November 2009. The uproar revolved largely around how the CRU adjusted and corrected historical weather and climate records in order to assemble a comprehensive global climate dataset.[1] The controversy reflected divergent understandings of language and methodology between professional climate scientists and the public. Or consider a recent poll showing that virtually all US climate scientists regard global warming as an established fact, while a large minority of weather forecasters remain skeptical – attitudes based largely in the two groups' differential experiences of data and data models (Maibach et al., 2010; Oreskes, 2004).

Throughout the sciences, as computer power and computational methods improve, a rapidly emerging 'fourth paradigm' of data-driven, interdisciplinary research is augmenting the existing paradigms of experimental, theoretical, and computational science (Atkins and National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, 2003; Bell et al., 2009; Hey et al., 2009). The 'fourth paradigm' brings science friction to the foreground. In principle, data collected by widely varying fields can now be assembled and brought to bear upon each other, leading to entirely new perspectives on ecology, Earth system science, medicine, epidemiology, and almost any other area (National Research Council, 1997; O'Brien et al., 2004; Zimmerman, 2003). Many scientists would like to have this ability. Many science funders, supercomputer centers, and institutions such as national science academies would like it even more. They believe that more data sharing will reduce redundancy, improve problem solving, increase research velocity, and cut costs at the same time. And indeed, many important examples of successful data sharing do exist. Yet in practice, science friction can make interdisciplinary data sharing maddeningly difficult.

Science friction is in some respects a generic problem of human communication, known both colloquially and formally as 'common ground' or 'grounding'; of establishing mutual

agreement, and adjusting and confirming shared understanding in collective projects of any kind. Grounding requires a common set of resources ('grounds'), which may be directly present in a shared environment and/or derived from shared vocabulary, ethnic background, or community membership.

A principal finding of discourse analysis is that common ground can never be established once and for all. Even the most ordinary conversation involves frequent moments of 'repair', in which participants re-establish grounding following momentary failure (Sacks et al., 1974). Consider the following exchange:

A: And then I bought two of those.

B: That black is nice.

A: No, those.

B: Oh, the red ones.

A: Yeah, the red.

Here, to repair her ambiguous reference (misunderstood by B), A points to objects in the shared environment. The difficulty of establishing common ground depends precisely on how much participants 'have in common', in many senses. People from very different social worlds typically spend more time grounding their conversations than do people from similar communities (Clark, 1992; Clark and Brennan, 1991; Olson and Olson, 2000).

The existence of ambiguity and the need to establish common ground are givens in ordinary communication. Yet in the sciences they tend to appear as a puzzle. After all, data are data, seemingly little more than spreadsheets full of (for example) instrument readings. Why can't one scientist just take another's dataset and plug it into a model, alongside other data of multiple origins? The answer is that very often he or she will need to know more about the other scientist's data than can be discovered from the cryptic, incomplete descriptions (if any) provided. The additional descriptions necessary to understand data are commonly referred to as 'metadata'. In this paper, we draw from our ethnographic work with several large scientific 'cyberinfrastructure' (or 'e-science') projects that are seeking to build systems that will support scientific work (Atkins and National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, 2003; Borgman et al., 2007; Hey and Trefethen, 2005). Their efforts to organize and share data illustrate some roles of metadata in the environmental sciences.

Metadata friction

In their quest to reduce data friction, scientists who store and use substantial amounts of data have begun to focus intensively on *metadata*. Metadata are often described as 'data about data' (Hey and Trefethen, 2003) or 'information about data' (Michener, 2006). In

most instances, metadata are descriptions of things, whether physical objects (books, items in a warehouse) or information objects (spreadsheets, web pages).

Library catalogs are the paradigmatic case, but mail-order catalogs, inventories, tables of contents, tag clouds, 'folksonomies', and photo captions are everyday examples. Metadata may include information about content (for example, a summary) as well as context (for example, creation dates, instrumentation). In science, the term 'metadata' typically refers to information about a dataset. Such information might include instrument characteristics, table formats, spatial locations, and/or the meanings of variable names. Through metadata, those charged with making data available effectively say to prospective users, 'Here is what you need to know about these data.' In other words, scientific metadata provide the information necessary for investigators separated by space, time, institutions, or disciplinary norms to establish common ground (Hey and Trefethen, 2005; Jones et al., 2001; Lawrence et al., 2009; Michener, 2006; Signell et al., 2008).

Extensive, highly structured metadata often are seen as a holy grail, a magic chalice both necessary and sufficient to render sharing and reusing data seamless, perhaps even automatic. For example, Gray et al. (2005) state that successful scientific analysis of large datasets depends crucially on developing 'extensive metadata and metadata standards that will make it easy to discover what data exist, make it easy for people and programs to understand the data, and make it easy to track data lineage'. In the last two decades, considerable effort has been devoted to defining metadata standards applicable to scientific datasets. Among many others, these include: the Open Archival Information System (OAIS) Reference Model; the Network Common Data Form (NetCDF), a 'self-describing' machine-independent data format; and the Climate and Forecast (CF) metadata conventions, which build on NetCDF for purposes specific to Earth system science (Hankin et al., 2009; Lavoie, 2004; Lawrence et al., 2009). Several metadata standards have been certified by the International Organization for Standardization (ISO). All such schemes approach metadata as explicit information, compiled in standardized categories and tightly controlled vocabularies.

XML schemes are another widely used approach to metadata formalization. XML (Extensible Markup Language) languages are designed to permit machine parsing of 'semantic information', ultimately in service of another chimera: the 'semantic web' of Tim Berners-Lee, a dream that has much in common with earlier (unrealized) visions of artificial intelligence. In principle, a well-defined XML scheme can utilize semantic web standards to offer the possibility of seamless intelligent enquiries across multiple heterogeneous databases (with differing units of analysis and spatio-temporal coordinates). XML schemes are customized to particular applications, and they formalize the language used to describe data and metadata structures.

The Ecological Metadata Language (EML) is a good example of an XML scheme, and one to which we return in our case studies. EML specifies a standardized vocabulary for describing datasets in the environmental sciences, as well as a computer-readable structure in which those descriptions are to be stored, searched, and displayed (Jones et al., 2006). (It is appropriate to use 'sciences' in the plural here, since each domain has its own configuration of classifications, instruments, dates and places.)

In these very typical framings, metadata are seen as *products*: information objects such as sets of descriptors, links, XML tags, catalogs, and so on. They are fixed, highly structured inscriptions, like library catalogs or archival finding aids. Collecting and organizing metadata then becomes a problem of capturing such information in standardized containers.

Just as with data themselves, creating, handling, and managing metadata products always exacts a cost in time, energy, and attention: *metadata friction* (Edwards, 2010). Scientists typically experience this frictional cost as an additional burden on top of their primary work. Research scientists' main interest, after all, is in *using* data, not in describing them for the benefit of invisible, unknown future users, to whom they are not accountable and from whom they receive little if any benefit. Countless discussions of the importance of metadata in the literature reflect the nearly insuperable difficulty of getting research scientists to record even the most basic metadata – let alone the meticulously detailed descriptions needed for long-term, multidisciplinary data sharing. For example, some of our interviewees estimated that up to two full days of work is required to fill in the metadata questionnaire for *each* of the hundreds of model runs (simulation datasets) to be included in the fifth Coupled Model Intercomparison Project. As a result, much attention focuses on such questions as how to adjust incentive structures to encourage recording metadata, or how metadata discovery might be automated (for example, with XML tags that allow a computer to 'understand' the context of a dataset).

Metadata-as-product is certainly a major and very common 'use case'. Yet, as a generic image of how metadata actually function in the world, it is far better fitted to social worlds where indexing, cataloguing, and searching are explicit and routine – social worlds such as libraries, archives, businesses with inventories to manage and sell, or the Social Security Administration – than it is to research science.

In many, perhaps most, situations of daily scientific practice, highly refined metadata products do not exist. Yet scientists still frequently reuse their own old data, share data within large work groups, and use data created by others. Informal, uncodified knowledge plays a critical role in making this possible (Zimmerman, 2007). How does this informal knowledge spread? In general (we argue), scientists rely on ad hoc, incomplete, poorly structured, mutable descriptions. *Many of these are generated on the fly in communicative exchanges, rather than carefully purpose-built, stored, and ready for use.* The rest of this article proposes an alternative view of metadata, focusing on its role in an ephemeral *process* of scientific communication, rather than an enduring outcome or product.

Metadata as a process

To explore this view, we present cases from our ongoing comparative research on scientific cyberinfrastructure projects. Through interviews, document analysis, and ethnographic studies of laboratory groups, we have been studying the work practices of research scientists, software developers, data managers, and others involved in distributed 'e-science' projects. The quotations in this section are from our interviews; the speakers' names and other identifying details have been changed to protect their identities.

In what follows, we look first at metadata practices in a small, one-off exploratory workshop that tested a portal system for cataloging climate model runs. Next we move to a medium-scale center with several hundred scientists engaged in multiple, overlapping projects. Finally, we look at metadata friction in a very large, continent-scale, multi-decade ecological data collection effort. Together, these cases illustrate the wide variety of ways in which metadata function in practice, illustrating not only friction within and across scientific fields but also the ways in which scientists (sometimes) communicate to overcome that friction.

Dycore workshop

Two atmospheric scientists planned a workshop to compare dynamical cores (or 'dycores'), the components of atmospheric global circulation models (AGCMs) that calculate transfers of energy, mass, and momentum in the atmosphere. AGCMs also include numerous other components that represent physical processes and parameters, such as radiative transfer and cloud formation. AGCMs, in turn, can be coupled to models of ocean circulation, land surface processes, and sea ice formation to represent the entire Earth system.

This workshop focused on dynamical cores in isolation from other AGCM components. The workshop tested all the cores on a common set of conditions in order to compare their performance, resulting in several hundred model 'runs' (datasets). The workshop was structured as an educational outreach, with the goal of training graduate students about dycores, and was hosted at the National Center for Atmospheric Research (NCAR). Some 40 graduate students were selected to attend, along with researchers from 10 different climate modeling groups.

Separately, a project known as Earth System Curator (ESC) had developed a web portal system for cataloging and describing atmospheric model output. A colleague involved in both projects suggested that the Dycore Workshop could use the services provided by the Curator portal, which was still in a prototype phase at the time of the workshop. (More recently, a version of the Curator system has been adopted by the Earth System Grid, a much larger, long-term effort to catalog and systematize climate model runs, largely in support of the climate model intercomparisons that provide a crucial element of the periodic assessment reports of the Intergovernmental Panel on Climate Change (IPCC).)[2]

Comparing model runs requires knowing the basic features of each model as well as the parameter settings used in each run. An example of a basic feature is the type of grid each model uses to represent the atmosphere. Simple latitude–longitude grids were once common, but they have been replaced in recent decades by other types of grids, including triangular, icosahedral, hexagonal, and unstructured grids. An example of a parameter setting is the size of the grid tiles (triangles, icosahedra, and so on).

The portal system would store all the model runs (data) in a repository, along with metadata describing each model and the parameter values used for each run. The portal's main feature was a 'faceted search', allowing all runs for a given parameter or set of parameters to be retrieved and compared. Models' basic features could be readily compared as well. The ESC team was looking for a good case to demonstrate the portal, and

the Dycore Workshop organizers decided they could use its repository and search services. This decision was made ad hoc, late in the workshop planning, and resulted in a hectic rush to set up the system in time.

The Dycore Workshop organizers developed a metadata questionnaire for each of the participating modeling groups to fill out beforehand, detailing characteristics of their models. Prior to the workshop, the organizers, pressed for time, did not confirm that the questionnaires had been completed correctly. It turned out that few modeling groups had devoted much time to the questionnaire in advance of the workshop. As a result, this critical questionnaire was filled out somewhat hastily during the workshop. This process did not go smoothly, because each group's vocabulary, names for variables, and other ways of describing its model differed somewhat from those of the others. Hence, when completing the questionnaire, modelers were often confused by the categories it offered. For example, was their model's diffusion implicit or explicit? Some guessed; some asked the organizers; others simply left fields blank.

Nonetheless, the model runs (datasets) used at the workshop and the metadata from the questionnaires were uploaded to the web portal. As the most complete real-world example of the portal's capabilities, the dycore repository served for over a year as the ESC's major working exemplar. The metadata categories used to describe Dycore Workshop runs included model characteristics such as 'Grid' and 'Conservation type'. Each 'experiment' specified a certain set of parameter values; runs (datasets) generated by different models for the same experiment could then be compared (Dunlap, 2008).

Despite the effort expended to set up the portal and to present it at the workshop, neither the workshop organizers nor the workshop participants made much use of the portal after the workshop. In another twist, when the organizers later examined the metadata captured in the questionnaires, they found numerous problematic imperfections. They began to clean up the metadata—but not on the portal. Instead, two of the scientists (located in different states) entered the cleaned-up metadata into their own, incomplete Excel spreadsheets, which they kept on their office computers and emailed back and forth. Meanwhile, the relatively large collection of model runs (about 1 TB of data) resided on a mainframe computer at NCAR. The following exchange with one scientist highlights the results of this diffusion of the metadata and data:

> Q: Are you making any use of the [portal] website as you analyze the data?
>
> R: Mmmm … not really … Oh, [the ESC team] also actually came back after the workshop and talked to [scientist] and me about the quality of the metadata, the metadata that described the model runs, and we actually started improving it, at least our best guesses at what the modeling groups meant. [Scientist] and I started improving it in the form of an Excel table, started improving the consistency of the metadata. And so we still have actually this improved version. Still, when you look at the [portal] webpage, the improvements that we thought about didn't go into the database after the fact. ... We [the scientists] didn't see this as our high priority at this stage to repair, I guess, the Dycore webpage. And now we are not certain whether it is still worth it … . For NCAR it's probably still worth it. For us, I don't know … I personally would not use [the Dycore webpage], but people could – it's out there. With probably an okay description of the datasets, it's not the perfect one. So our Excel table serves us a little better.

Q: Have the portal project staff seen [the Excel table]? Has [ESC team member] seen that?

R: ... [ESC team member] asked for it a few times and then, it was never quite finished, so we didn't feel comfortable to give it to her. ... And she actually offered that she would be willing to – that she would go in and basically fix the descriptions. The hard part is, I think the database is set up so the metadata of the models is linked to the actual model runs, and my understanding was that there was some … some manual step involved to do that link, or to provide that link between the model run and the metadata. And now when you change the metadata after the fact …

Q: It breaks the link?

R: It breaks the link, or you need to do it again, or something. It's doable, but it sounded – she was willing, very willing to do that. So it's not really that this was an issue. I think for us the issue became, do we as scientists – we didn't see a lot of gain because, again for us we were already insiders and kind of know what's going on. Sure I see that other people would gain from it, I guess.

In addition to the manual, ad hoc quality of the metadata collection process, we see here a number of features typical of metadata's role in scientific communication.

The process is *fragmented*, with many individuals contributing. These include someone from each modeling group, who answers the questionnaire; an ESC team member who is trying to create a website with consistent descriptors linked to the model runs; and the scientists leading the Dycore project, who assist with the questionnaire but then also create and edit their own Excel spreadsheets, leaving out the links to the model runs carefully created by the ESC team.

The process is *divergent*, with two versions of the metadata (the website and the Excel sheet) appearing almost immediately without ever being reconciled.

Metadata production is *iterative,* with considerable effort devoted to repairing misunderstandings and mistakes. An ESC team member needs to solicit information from the modelers repeatedly. The modelers ask questions to clarify categories and terms as they fill out the metadata questionnaire. The scientist-organizers work together, making a series of small improvements to their private Excel spreadsheets. Retrospectively, the ESC team noted that it would be 'naïve to think that we will ever come to a finished metadata model' (Dunlap et al., 2008). Both the metadata categories and the contents of those categories remain in flux.

For the workshop organizers, *local use* of the metadata dominates over their desire to contribute to the global project.

Fragmentation, divergence, iteration, and local-centeredness all act to necessitate more work in producing metadata. In the interests of the portal project, an ESC team member was willing to devote the time and energy to push for a synthetic, public metadata product, but her effort to overcome metadata friction succeeded only partially. Meanwhile the Dycore Workshop organizers also invested time and effort in metadata cleanup.

The rough process of metadata production thus yielded multiple, rough products. In principle, outsiders could have discovered the public-facing showpiece data portal (now

dismounted); had they done so, they could have used the metadata portal provided to explore the data. Yet for anyone outside the workshop (and perhaps many within it), making serious use of the model runs would require communicating with someone. Prospective users would have needed to ask the workshop organizers how to interpret the metadata; had they done so, they might also have discovered the more accurate and current Excel-based metadata document.

These points came alive for us when we participated in a semi-public web-based demonstration of the portal in 2009, 'attended' via teleconference by more than 25 software developers and scientists, all with considerable experience in climate modeling and/or climate science software. Within minutes, members of this group – who shared a degree of common ground unlikely to be matched by any other conceivable set of users – began expressing confusion over the terminology used to label elements of the models. This continued throughout the presentation, with participants expressing a need for further levels of explanation at every turn.

This case study represents metadata friction *within* a field. Even across two small projects with similar, but not identical goals (the Dycore Workshop and the Curator data portal), metadata could not function purely as a product. Instead, breakdown occurred at numerous points, each requiring repair to re-establish common ground. These were achieved through direct communication among the individuals managing the data and metadata products. As we will see in the following case studies, these challenges will be greater for scientific groups with less in common.

The Center for Embedded Networked Sensing

Scientists at another center we have studied, the Center for Embedded Networked Sensing (CENS), develop sensing systems for real-world scientific and social applications through collaborations between scientists, computer scientists, and engineers. CENS was funded for 5 years by the US National Science Foundation in 2002, and renewed for an additional 5 years in 2007. As of 2010, CENS has over 300 associated faculty members, students, and research staff from a number of disciplines. The majority of CENS' participants are technologists (computer scientists, electrical engineers, and environmental engineers), while others are scientists (seismologists, terrestrial ecologists, and aquatic biologists) whose research employs CENS-developed sensor technology. Still other members of the Center come from urban planning, design and media arts, and information studies. Our research at CENS has focused on data practices in the ecological and environmental sensing collaborations (Borgman, 2007; Borgman et al., 2006, 2007; Mayernik et al., 2008; Wallis, et al., 2007, 2008, 2010).

In a large center dedicated to collaborative research, building common ground for communication is an important task, as the following passages from interviews with CENS researchers illustrate:

> Technologist: That's one thing that we definitely learned, just like working across different fields. We learn that we have different vocabulary and that when I say 'sensor fault', that means something different than maybe when [my science collaborator] says it.

> Scientist: [My engineering collaborator] and I have spent the better part of four years learning each other's vocabulary. … We've spent a fair amount of time saying, 'What exactly do you mean when you say that?'

Common ground is necessary for the frequent collaborations around data that occur within CENS projects. Team members regularly pass data files back and forth by hand, by email, and by using shared lab or project servers, websites, and databases.

Our interviews reveal four interesting aspects of metadata practices at CENS. First, formal metadata schemas are rarely used. When created at all, metadata are typically generated on an ad hoc basis for a specific purpose, as discussed in the following exchange with a faculty engineer:

> Q: What do you do to your data so that you can use it again in both the short term and the long term?

> R: Nothing special. These are plain text files that sit someplace. We don't do anything. We don't annotate them or anything in any major way that's worthy of mention. People will do a little massaging. They'll add a few comments to the file to remind them of what it is, but nothing hugely beyond that.

Second, data are the responsibility of the individual. In any given project, the responsibility for keeping track of data usually rests with graduate students. The following exchange with a faculty environmental engineer illustrates how data are organized by students for their own projects:

> Q: What do you do to your data so you can use it again in the future? We're trying to get a sense of what you're doing about sort of standard data formats or archiving practices.

> R: I think that one would boil down to the dissertation level student and how they archive their modeling runs and their observation datasets so that it's convenient for whatever programs they're using to call it in, and compare it in the guts of the code.

> Q: So – can you get back to the data that your students did after they finish a dissertation?

> R: Yes I can. Do I always get all the way down there? I don't know. But we always make sure that the raw data stays on the server somewhere.

Third, while the absence of formal metadata has not prevented data from being shared between CENS research teams, or with people from outside of CENS, the informal process of creating and streamlining metadata is a substantial source of friction that can impede data sharing:

> Q: So people have asked you to use your data, people from outside the project?

> R: [Another CENS research group] is interested in it. One guy who visited here is interested in trying to use it. He's got some big three-dimensional fluid mechanics models and I would

> actually like to let him have the data because I don't see myself having those kinds of models in the near future.
>
> Q: Okay. And have you released it yet?
>
> R: No I haven't. I haven't been around enough to get it in a nice form for him.

Fourth, different members of the collaboration assign different levels of importance to metadata. This can be another major source of friction. The next passage is from an interview with a senior ecologist in which he describes a meeting with a database developer and another staff ecologist who were starting to build a communal database. The two disagreed over implementing Ecological Metadata Language (EML) in the database:

> Scientist: We had a conversation the other day with [database developer] where [ecologist] basically said 'EML is really important to us, and how soon can we get this into the database?' And [database developer], who's been out of the loop, I mean, he's a new employee and he hasn't been part of any of these conversations, just basically said 'What are you talking about? I'm not doing any of this. Metadata isn't important to me right now.' It's like, I beg to disagree with you. … I mean, I don't know why it would be that difficult to just add a bunch of fields, at least a minimum, be able to associate those fields with the sensor values and say hey, this is the kind of sensor probe we used, and … define some of these things.

In this case, informal communication was not enough to overcome the friction represented by the developer, and EML was not implemented in the database.

These CENS cases illustrate how taking metadata as a formalized representation of data glosses over many nuances of interaction and communication around data and metadata. Formal metadata records that conform to established standards are almost nonexistent in the day-to-day work of CENS researchers, and the different priorities of interdisciplinary collaborators work against the implementation of single-disciplinary standards, such as EML, in communal data systems.

Instead, both inter- and intra-disciplinary collaboration around data takes place, within CENS, through direct communication via phone, email, instant messaging, or the exchange of physical media, such as flash memory cards and CDs. Communication *about* data always occurs alongside communication *with* data. Sharing data involves, as one CENS research told us, 'a lot of hand-holding until people got used to it. Maybe third- or fourth-time users would probably start to get a feel for it, but [with] first-time users you're going to probably be answering two or three e-mails a day ... . ' Data and metadata frictions like those illustrated in this section slow down local collaborations, prevent others from occurring, and impede the drive toward 'fourth paradigm' data-intensive scientific research.

The Long-Term Ecological Research program

The Long-Term Ecological Research (LTER) program constitutes a distributed, heterogeneous network of more than 1800 research scientists and students. Formed in 1980, the network currently consists of 26 sites or research stations (Hobbie et al., 2003). The

program's mission is to further understanding of environmental change through interdisciplinary collaboration and long-term research projects. Each LTER site is arranged around a particular biome – for example a hot desert region, a coastal estuary, a temperate pine forest or an Arctic tundra region – in the continental US and Antarctica. A 27th site is charged with the administration and coordination of the group. Ironically, now that the US LTER recognizes that sites comprising 'pure nature' are the extreme exception, not the rule, a new suite of urban sites is being developed, and some 'long term' sites have already closed.

Over the past decade, attempts have been made to integrate the US national LTER into an International LTER. A chief challenge, both within the US LTER and in the larger international effort, has been to achieve genuine data sharing across a community whose members belong to heterogeneous disciplines. (In future work, we will discuss the particular challenges of international data sharing: to paraphrase Trotsky, you can't have cyberinfrastructure in one country, even though many are attempting to build their own for purposes of national advantage.)

Each of the 26 LTER data sites takes responsibility for managing locally produced research data. In general, each site has its own information system (including its own databases) and its own information manager, who is charged with the development and maintenance of local infrastructures. Across the network, then, data are stored autonomously by individual sites. This fact renders the search for and access to data complex and laborious, militating against the prospect that the network will realize its mission. Accordingly, in 1996 LTER initiated a project for a networked information infrastructure that would federate the local databases and improve data exchange.

The integration project has encountered three major challenges: (1) the heterogeneity of the data that circulate through the LTER research community; (2) the wide dispersal of LTER sites; and (3) the multiple metadata schemes (Jones et al., 2001) required to capture all the details necessary for all possible secondary users of the data (an ideal solution that evokes Spinoza's problem: to know a single fact about the world, you need to know every fact about the world). These schemes include detailed and diverse information such as the names of the researchers who collected the data, the title of the project on which they were working, a project summary, keywords, the type of biome, sampling techniques, and calibration of the measuring tools at the time of data collection. Calibration is a significant issue: while local scientists know the variability and character of their own sensors, they are frequently unaware of new capture methods and new research procedures used elsewhere (Bowker, 2005). By extension, the possibility of complex analyses drawing on physical, chemical, and biological data across the many geographical areas represented by LTER sites depends on the quality of the metadata. Hence this community has taken metadata as central, both organizationally and intellectually.

In 2002, as part of a larger attempt to standardize its data management practices, the LTER research community adopted EML, introduced above. EML is closely associated with the LTER. It originated in 1997 at the National Center for Ecological Analysis and Synthesis (NCEAS), a research center focused on developing tools and techniques for analyzing and synthesizing environmental data. As part of this effort, in 1997 a researcher in ecological informatics working with two doctoral students produced EML version 1.0.

The team drew the standard's content from the main data description types already in use, such as those recognized by the Ecological Society of America, itself a pioneer in preserving datasets alongside the papers written from those data. Between 1997 and 1999, NCEAS developed and tested EML versions 1.0 to 1.4.

Responding to difficulties encountered in use, developers planned a major revision of EML. (Would that one could vary natural languages so simply. Despite legislative efforts, most French refer to 'le walkman' and not 'le balladeur'.) The three-person development team expanded into a collaboratory (Olson et al., 2001) – a collaborative platform based on voluntary participation and open to the whole community of environmental scientists. This open development model was not immediately successful, though the team was able to attract more developers, including, for the first time, a separately but synergistically funded information manager from the LTER. Seventeen versions of EML were produced between 1999 and 2002, at which stage it was adopted by LTER.

For the LTER, results over the 8 years since EML's adoption have been slow in coming. Some sites began the work of implementing the standard relatively quickly, but most of them ran into significant problems. At about 250 pages in length, the standard is complex and difficult to grasp in its entirety (Berkley, et al., 2010). The data management tools intended to facilitate EML implementation proved unusable due to incompatibility with existing local practices and infrastructures. This excerpt from a 2009 interview with an LTER information manager conveys some of the difficulties encountered:

> We made a concerted effort to get people's attention this year, about following through our data management. Because in the past more of my time than I think should have been put into it was spent tracking people down ... to make sure that we were current. Some people are very good about this, others are less good. And for a long time it was falling on me to just follow up with people, badger them in some cases and threaten them. But it's a waste of time to do all that. So we all agreed that we really want to move away from that to a situation where people really buy into this. So we tried the carrot and stick approach ... .
>
> So we sent out this message that you're not gonna work here if you don't do this. So people do it now. So, you know, they both helped. So what they actually do is they will contact me when the time comes and make a submission of data, metadata. And the procedures we use are very simple. ... We have a combination of tabular data and spatial data, most of the datasets are still just two-dimensional arrays of numbers. ... And at this time we don't serve the data, we don't preprocess it, we just make it available, in clickable form, in an archive format so people can download it.
>
> The assumption we made until now is that most scientists would just get the whole thing and put it into their program of choice and subset or analyze it as they wish. ... We have put a lot of time into metadata though, because the LTER network adopted EML, Ecological Metadata Language, as a standard almost six or seven years ago. And this is something that we believe in firmly here at the site. Although it's been a lot of work, I think the advantages ... of having structured metadata and complete metadata, the advantages are enormous. And the problem is that, as so often happens, the network adopted the standard before the software tools were really there to make this an easy task.

> And so the LTER sites fell into a couple of groups that were, sites like ours that had a huge amount of legacy data, legacy metadata, so there was a problem of how do you take the metadata you have and get it into the new format. And then for newer sites, in fact for all sites it's a problem of how do you deal with the new information coming in. And that's still – there's not a good solution to that I think. ... I'm not sure you can remove all of the drudgery from doing metadata, at some level it's always going to be a fairly painstaking process. But I can imagine that the software tools will ease it somewhat in time. Once you have the information in EML, it's now possible to do wonderful things with it.

As this interviewee mentioned, implementing EML was a mostly unfunded mandate for LTER sites, requiring a huge amount of work (on top of the normal workload) with minimal resources. Some sites had to completely restructure their data management practices. In the scramble to implement the standard, numerous ad hoc solutions were brought to bear. For example, some of the information managers shared home-grown tools amongst themselves to facilitate converting local systems into the EML format. Speaking in 2009, a senior LTER ecologist noted the disarray caused by the lack of effective tools:

> I think the LTER network office and their cyberinfrastructure group, as well as the National Center for Ecological Analysis and Synthesis, have frankly done a crappy job making tools available to the broader community to make it easier to share data. It's hard to develop metadata that's, you know, that will work and is searchable and things like that. Because it's not an interesting research project. ... Once you figure out how you should do it, that's the interesting thing. Then you need a bunch of programmers who are going to sit down and implement it. And that's a boring thing to do. So researchers are not going to do that, NSF isn't going to pay for that with, you know, grants, and it's not a commercializable application, so it just languishes. So the technology for helping us to do that has not developed at the same pace. ... It does not help that we all work with Excel and other proprietary software. So posting an Excel spreadsheet is fine if someone is using Excel, right. Those of us working on this in the early nineties are, like, we know the technologies are not perfect, if we could even get people to post Excel spreadsheets that would be a big jump in changing the culture, because as the tools have evolved the culture evolves much more slowly. So if we can change the culture, the tools will follow is the idea.

LTER scientists (as opposed to information managers) express a variety of feelings about EML and the metadata recording process. Some find the project less useful than it might be. Asked to comment on EML, an LTER geologist articulated frustration with its restricted scope (2009 interview):

> R: EML is such limited metadata. I mean all you have to give it is you know, the dataset, what was collected. ... There's no real requirement to organize it, you just have to say what it is. As far as I have seen ... there is not even very much geospatial information you have to store with it. I mean, it's a very small subset of like the FGDC [Federal Geographic Data Committee] ... I don't find that particularly effective metadata ever.
>
> Q: Do you think it will ever be in the future?
>
> R: If they add fields and enforce it.

Reflecting the notion of metadata as a communication process discussed earlier, the same geologist indicated that he preferred to discuss his data with other potential users rather than provide metadata in written form:

> R: I digitized the geology the first year, or two; I was there and people were wanting that. So that's been sent out. ... I have it listed on the – I have some of my geospatial datasets but they are so, I've been writing the metadata for them because of the nature of the data they are, it's just, I would much rather talk to the person. I guess I'm doing the same. Well I haven't put it up since I got out of grad school, which I need to do, but that's –You know, I should do that.

These views highlight a disjuncture between two moments of metadata projects in general. In the first moment, everyone agrees that a standard set of metadata would be helpful and important. In our studies of LTER, this was clearly the case with most of the actors involved in the standardization process (EML developers, LTER network coordinators, information managers, domain researchers, and so on). All of them have supported – and continue to support – the 'EML project'. Yet at a second moment – the moment of implementing the standard – critical problems emerge and discordant voices proliferate. All recognize metadata's potential value, but when the rubber meets the road, an unfunded mandate to be altruistic (and simultaneously to lose one's own tried-and-true local bricolage with data structures) does not prove highly attractive. Introduced in order to reduce data friction, metadata creates its own kind of friction.

This finding accords with Star and Ruhleder's (1994) observation of the phenomenon of 'almost use' of software (we've got it loaded, we've hit a glitch, some day we'll deal with it). Metadata standards fall into the category of 'almost standards': everyone agrees they are a good idea, most have some such standards, yet few deploy them completely or effectively. In an earlier study, we found such an 'almost standard' at the National Science Foundation, whose program officers were convinced that their data policy ensured publication of publicly funded data, even though the majority of NSF-funded Principal Investigators (including ourselves) displayed near-complete ignorance of this policy.[3] The answer in these and other cases is not stricter standards, but a successful effort to integrate understandings of the working culture and practices of scientists into the design and implementation of those standards – an issue for research-in-progress that some of us raised with climate science software developers.

'Science friction', as we have called it, includes not only the particular problems with metadata we have described here, but also numerous larger issues about how data travel among disciplines. These issues include differences in how graduate students are trained, in the character of data production and analysis, and in the types of software, instrumentation, and other technology used to 'make data' (Edwards, 2010). Such differences are inscribed in the complex web of often overlapping and/or competing national and international standards for data, metadata, and data analysis. All of these issues present fertile topics for future sociological and policy research.

Conclusions

Most discussions of scientific metadata treat them as products: static, definitive descriptions of data characteristics, like library catalogs. Yet, in routine scientific

practice, metadata are called into being much more dynamically, during requests for and conversations about data. Some of these conversations are direct (as in the Dycore Workshop case above, or some of the exchanges mentioned in the CENS and LTER cases). Many others are mediated in various ways.

When ordinary conversations are mediated – through email, telephone, videoconference, and so on – they typically require more repair than do face-to-face conversations, due to the lack of shared physical context and nonverbal cues (Clark and Brennan, 1991; Olson and Olson, 2000). Similarly, mediated conversations about data require more grounding than do face-to-face ones. Metadata products are supposed to substitute for direct contact with data producers – and they *can* do that, to a greater or lesser degree, in many contexts. Yet in very many cases, metadata products remain incomplete, ambiguous, or corrupted (Wayne, 2005). When this happens, the conversation about data cannot continue without repair. Such repair can, and often does, include direct communication with the data creators: metadata-as-process. As with ordinary conversations, the greater the social distance between the disciplines of data creators, the more metadata-as-process is likely to be needed. The examples described in this article represent just a few of the many methods – conversations, emails, annotations, ad hoc tools, and other means, often informal and/or ephemeral – by which scientists overcome metadata friction, freeing data to travel more widely.

To put the point another way, consider the following analogy. Engineers reduce friction with precision – making interacting parts mesh better – and with lubricants. Typical discussions of metadata see them as contributing to *precision*, making it possible to join one part (dataset) more perfectly to another one. This may involve considerable effort at shaping and polishing a part – refining its metadata – to reduce its coefficient of friction. By contrast, the process view we explore here looks primarily at *lubrication*: the practices through which people overcome friction *without* precise solutions or the need to modify components. Does interdisciplinary data sharing work more like a fine Swiss watch, with dozens of gears and jeweled pivots so precisely engineered that they never need lubrication? Or does it work (as we believe) more like a car engine, running fast and hot, bathed constantly in motor oil to keep the parts from burning up?

We have argued, first, that metadata represent a form of scientific communication, and second, that both precision and lubrication have important roles to play in reducing science friction. Well-codified metadata *products* increase the precision with which a dataset can be fitted to purposes for which it was not originally intended, or can be reused by people who did not participate in creating it. At the same time, ephemeral, incomplete, ad hoc metadata *processes* act as lubricants in disjointed, imprecise scientific communication. This latter category of metadata frequently appears alone, in the case of datasets for which no metadata products exist, but it also frequently appears in the actual use of metadata products. This second, complementary form of metadata has typically been brushed aside in the quest to achieve comprehensive, stable, permanent catalogs.

Our line of reasoning here resembles the classic analyses of plans and situated action by Lucy Suchman (1987, 2007), and Phil Agre and David Chapman (1990). Those analyses challenged a conception of plans as programs (in our terms, as fixed products), arguing that it is impossible ever to specify everything about the conditions under which any plan will be carried out. ‘Carrying out a plan necessarily and fundamentally presupposes

improvisation' (L. Suchman, personal communication). The availability of both planning and improvisation, and their interaction, makes human action simultaneously focused and flexible (unlike that of computer programs, whose performance typically degrades precipitously or fails altogether in the presence of unanticipated contingencies). Analogously, we are arguing here for a revisionist view of metadata-as-product. Metadata products can be powerful resources, but very often – perhaps even usually – they work only when metadata processes are *also* available. As with improvisation in action patterns, metadata-as-process suggests the very large role of ad hoc practices, incomplete information, and unfinished agendas in everyday scientific work. In future articles, we will attend to the question of how best to support both the process and product modes simultaneously.

Science friction occurs far beyond laboratories and e-science networks, because in today's world scientists are not the only ones who want to know about other people's data. For example, the release of the Climategate emails and datasets followed a barrage of Freedom of Information Act requests filed by climate skeptic Stephen McIntyre and others; during the summer of 2009, McIntyre alone filed 58 requests for emails and data in a single week. Climate-change skeptics are presently employing a similar strategy to gain access to emails and climate data from scientists at the Goddard Institute for Space Studies, the Lawrence Livermore National Laboratory, and other institutions. The ideas we have developed here show why it is significant that these requests ask not only for data, *but also for emails about the data.* The skeptics don't just want the numbers. They also want to know what was said about the data; what decisions went into the choice of some numbers and not others; how raw instrument readings were adjusted (for example, to account for changes in instrumentation or station siting); who created and managed the datasets; and so on.

The scientists quite naturally fear that their internal communications will be understood as metadata, as significant information about the datasets and their interpretation. In the Climategate case, the scientists argued that their email shorthand was misunderstood by others with whom they shared little common ground. And, in fact, it *was* misunderstood. Phrases taken from the emails, such as 'Mike's *Nature* trick' and 'to hide the decline', became the object of scrutiny. Originally a metadata process – a means of discussing and settling understandings of data within a small community, accompanied by conversations and other informal communication that left fewer traces – the emails became transformed into a metadata product, stripped of both its community context and its role as communication (rather than fixed product). Once the emails were released into a highly politicized, highly public space, re-establishing common ground through communication – an additional metadata process – became nearly impossible. The Climategate episode thus illustrates at once the fundamental role of metadata processes in data production and the difficulties they pose for the movement of data among highly diverse communities. Metadata as process – as communication – will inevitably both resolve and create misunderstandings.

For sociologists of science, this opens an important research agenda. As planetary management becomes a more complex and urgent problem, better metadata products will be imperative, but they will never eliminate the need for informal, ad hoc, incomplete and contested processes of communicating about data. Those processes – and the

repair they can bring to misunderstandings, as well as the misunderstandings they can create – will be more important, and more fraught, than ever.

Acknowledgements

This work was supported by the National Science Foundation SBE/SES Human Social Dynamics grant #0827322, 'Monitoring Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures', Paul N. Edwards, (University of Michigan), Principal Investigator; Geoffrey C. Bowker (University of Pittsburgh), Christine L. Borgman (UCLA), and Steven J. Jackson (University of Michigan), co-Principal Investigators. Funding for collecting the data used in our case studies comes from multiple sources, including National Science Foundation Cooperative Agreement #CCR-0120778 for the Center for Embedded Network Sensing, Deborah L. Estrin, UCLA, PI; National Science Foundation grant #ESI-0352572 for CENS Education Infrastructure, William A. Sandoval, PI and Christine L. Borgman, co-PI; National Science Foundation Award #OCI-0750529, 'Towards a Virtual Organization for Data Cyberinfrastructure', Christine L. Borgman, PI, Geoffrey C. Bowker and Thomas Finholt, co-PIs; and gifts from Microsoft External Research. Additional data for this paper were contributed by Steven J. Jackson and Jillian C. Wallis. Despite the intensive effort it can involve, data collection remains a poorly recognized and rewarded type of contribution throughout the natural and social sciences. We specifically acknowledge this effort here. In addition, we acknowledge contributions in many forms from all members of our research group, the Monitoring, Modeling and Memory Project (http://monmodmem.org). MMM's membership includes Paul N. Edwards (PI, University of Michigan), Steven J. Jackson (co-PI, University of Michigan), Thomas Finholt (co-PI, University of Michigan); Archer L. Batcheller (graduate student, University of Michigan), Ayse G. Buyuktur (graduate student, University of Michigan); Geoffrey C. Bowker (co-PI, University of Pittsburgh), Susan Leigh Star (co-PI, University of Pittsburgh, deceased); Christine L. Borgman (co-PI, UCLA); Jillian C. Wallis (graduate student, UCLA), Matthew S. Mayernik (graduate student, UCLA), David S. Fearon, Jr (graduate student, UCLA), and David Ribes (co-PI, Georgetown University).

Notes

1. See the 12-part special report on 'Climate wars' in *The Guardian* (Pearce, 2010). In April 2010, an independent House of Commons investigation exonerated the CRU team of any scientific wrongdoing, but this has not extinguished the controversy (Reed, 2010).
2. See the websites of the Program on Climate Model Diagnosis and Intercomparison (PCMDI, n.d.) and the Coupled Model Intercomparison Project (CMPI5, n.d.).
3. See the website of the OECD Follow-up Group on Issues of Access to Publicly Funded Research Data (Organization for Economic Co-operation and Development, n.d.).

References

Agre PE and Chapman D (1990) What are plans for? *Robotics and Autonomous Systems* 6(1/2): 17–34.

Atkins DE and National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure (2003) *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructures*. Arlington, VA: National Science Foundation.

Bell G, Hey T and Szalay A (2009) Beyond the data deluge. *Science* 323: 1297–1298.

Berkley C, Blankman D, Brunt J, Gries C, Jones MB and Jones C, et al. (2010) Ecological Metadata Language (EML) Specification. Available at http://knb.ecoinformatics.org/software/eml/eml-2.1.0/index.html (accessed 13 April 2011).

Borgman CL (2007) *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman CL, Wallis JC and Enyedy N (2006) Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. Unpublished paper presented at the 10th European Conference on Digital Libraries, Alicante, Spain (17–22 September).

Borgman CL, Wallis JC and Enyedy N (2007) Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries* 7(1/2): 17–30.

Bowker GC (2000) Biodiversity datadiversity. *Social Studies of Science* 30(5): 643–683.

Bowker GC (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Braudel F (1975) *The Mediterranean and the Mediterranean World in the Age of Philip II*. New York: Harper & Row.

Buckland MK (1991) Information as thing. *Journal of the American Society for Information Science* 42(5): 351–360.

Buckland MK (1997) What is a 'document'? *Journal of the American Society for Information Science* 48(9): 804–809.

Clark HH (1992) *Arenas of Language Use*. Chicago: University of Chicago Press.

Clark HH and Brennan SE (1991) Grounding in communication. In: Resnick L, Levine J and Teasley S (eds) *Perspectives on Socially Shared Cognition*. Washington, DC: Amerian Psychological Association, 127–149.

CMIP5 (n.d.) *Coupled Model Intercomparison Project Phase 5 – Overview*. Available at http://cmip-pcmdi.llnl.gov/cmip5/ (accessed 16 April 2011).

Collins HM (1985) *Changing Order: Replication and Induction in Scientific Practice*. London and Beverly Hills: Sage Publications.

Collins HM and Pinch T (1993) *The Golem: What Everyone Should Know about Science*. Cambridge: Cambridge University Press.

Dunlap R (2008) The Earth System Curator: Metadata Infrastructure for Climate Modeling. SIParCS Final Presentation, Boulder, CO, 4 August 2008. Available at www.earthsystemcurator.org/presentations/pres_0808_rocky.ppt (accessed 13 April 2011).

Dunlap R, Mark L, Rugaber S, Balaji V, Chastang J and Cinquini L, et al. (2008) Earth system curator: Metadata infrastructure for climate modeling. *Earth Science Informatics* 1(3): 131–149.

Edwards PN (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Elichirigoity F (1999) *Planet Management: Limits to Growth, Computer Simulation, and the Emergence of Global Spaces*. Evanston, IL: Northwestern University Press.

Galison PL (1996) Computer simulations and the trading zone. In: Galison PL and Stump DJ (eds) *The Disunity of Science: Boundaries, Contexts, and Power*. Stanford: Stanford University Press, 118–157.

Gray J, Liu DT, Nieto-Santisteban M, Szalay A, DeWitt D and Heber G (2005) Scientific data management in the coming decade. *CTWatch Quarterly* 1(1). Available at: www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/ (accessed 13 April 2011).

Haberl H, Erb KH, Krausmann F, Gaube V, Bondeau A and Plutzar C, et al. (2007) Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. *Proceedings of the US National Academy of Sciences* 104(31): 12942–12947.

Hankin S, Blower JD, Carval T, Casey KS, Donlon C and Lauret O, et al. (2009) NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes. *Community White Paper for Ocean Observations* 9.

Hey T, Tansley S and Tolle K (eds) (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research. Available at http://fourthparadigm.org (accessed 13 April 2011).

Hey T and Trefethen A (2003) The data deluge: An e-science perspective. In: Berman F, Fox G and Hey AJG (eds) *Grid Computing: Making the Global Infrastructure a Reality*. Chichester: Wiley, 809–824.

Hey T and Trefethen AE (2005) Cyberinfrastructure for e-Science. *Science* 308(5723): 817–821.

Hobbie JE, Carpenter SR, Grimm NB, Gosz JR and Seastedt TR (2003) The US long term ecological research program. *BioScience* 53(1): 21.

Jones MB, Berkley C, Bojilova J and Schildhauer M (2001) Managing scientific metadata. *IEEE Internet Computing* 5(5): 59–68.

Jones MB, Schildhauer MP, Reichman OJ and Bowers S (2006) The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics* 37(1): 519–544.

Kevles DJ (1998) *The Baltimore Case: A Trial of Politics, Science, and Character*. New York: WW Norton.

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

Lavoie BF (2004) The Open Archival Information System reference model: Introductory guide. *Microform and Imaging Review* 33(2): 68–81.

Lawrence BN, Lowry R, Miller P, Snaith H and Woolf A (2009) Information in environmental data grids. *Philosophical Transactions of the Royal Society (A): Mathematical Physical and Engineering Sciences* 367(1890): 1003–1014.

Maibach E, Wilson K and Witte J (2010) *A National Survey of Television Meteorologists about Climate Change: Preliminary Findings*. Fairfax, VA: Center for Climate Change Communication, George Mason University.

Mayernik MS, Wallis JC, Pepe A and Borgman CL (2008) Whose data do you trust? Integrity issues in the preservation of scientific data. Unpublished paper presented at the iConference, Los Angeles, CA (29 February).

Michelet J (1930) *Oeuvres de Michelet 1, Autobiographie, Introduction à l'Histoire Universelle* (Chabot H, trans.). Paris: Larousse.

Michener WK (2006) Meta-information concepts for ecological data management. *Ecological Informatics* 1(1): 3–7.

National Research Council (1997) *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: National Academy Press.

O'Brien K, Hankin S, Callahan J, Balaji V, Schweitzer R and Mclean J, et al. (2004) The GFDL data portal: A doorway to sharing model outputs. Unpublished paper presented at the American Geophysical Union, San Francisco (13–17 December).

Olson GM, Atkins D, Clauer R, Weymouth T, Prakash A and Finholt T, et al. (2001) Technology to support distributed team science: The first phase of the Upper Atmospheric Research Collaboratory (UARC). In: Olson G, Malone T and Smith J (eds) *Coordination Theory and Collaboration Technology*. Hillsdale, NJ: Lawrence Erlbaum Associates, 761–784.

Olson GM and Olson JS (2000) Distance matters. *Human-Computer Interaction* 15: 139–179.

Oreskes N (2004) Beyond the ivory tower: The scientific consensus on climate change. *Science* 306(5702): 1686.

Organization for Economic Co-operation and Development (n.d.) *The Public Domain of Digital Research Data*. Follow-up Group on Issues of Access to Publicly Funded Research Data. Available at http://dataaccess.ucsd.edu (accessed 16 April 2011).

PCMDI (n.d.) *Program for Climate Model Diagnosis and Intercomparison*. Available at www-pcmdi.llnl.gov/ (accessed 16 April 2011).

Pearce F (2010) Climate wars: Guardian special investigation. *The Guardian*. Available at www.guardian.co.uk/environment/2010/feb/09/climate-change-data-request-war (accessed 14 April 2011).

Reed S (2010) Oxburgh report clears controversial climate research unit. *ScienceInsider*. Available at http://news.sciencemag.org/scienceinsider/2010/04/oxburgh-report-clears-controvers.html (accessed 14 April 2011).

Sacks H, Schegloff EA and Jefferson G (1974) A simplest systematics for the organization of turn-taking in conversation. *Language* 50(4): 696–735.

Sanderson EW, Jaiteh M, Levy MA, Redford KH, Wannebo AV and Woolmer G (2002) The human footprint and the last of the wild. *BioScience* 52(10): 891–904.

Serres M (1995) *The Natural Contract* (MacArthur E and Paulson W, trans.). Ann Arbor: University of Michigan Press.

Serres M (2007) A return to the natural contract. In: Bindé J (ed.) *Making Peace with the Earth: What Future for the Human Species and the Planet*. Paris: UNESCO Pub.; Berghahn Books, 129–137.

Shapin S and Schaffer S (1985) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton, NJ: Princeton University Press.

Signell RP, Carniel S, Chiggiato J, Janekovic I, Pullen J and Sherwood CR (2008) Collaboration tools and techniques for large model datasets. *Journal of Marine Systems* 69(1/2): 154–161.

Star SL and Griesemer J (1989) Institutional ecology, 'translations', and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. *Social Studies of Science* 19(3): 387–420.

Star SL and Ruhleder K (1994) Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*. New York: Association for Computing Machinery.

Strathern M (2004) *Commons and Borderlands: Working Papers on Interdisciplinarity, Accountability and the Flow of Knowledge*. Wantage: Sean Kingston Publishing.

Suchman LA (1987) *Plans and Situated Actions: The Problem of Human-Machine Communication*. New York: Cambridge University Press.

Suchman LA (2007) *Human-Machine Reconfigurations: Plans and Situated Actions* (2nd edn). New York: Cambridge University Press.

Vaughan D (1996) *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press.

Wallerstein I (1976) A world-system perspective on the social sciences.*British Journal of Sociology* 27(3): 343–352.

Wallis JC, Borgman CL, Mayernik MS 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): 114–126.

Wallis JC, Borgman CL, Mayernik MS, Pepe A, Ramanathan N and Hansen M (2007) Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. Unpublished paper presented at the 11th European Conference on Digital Libraries, Budapest, Hungary.

Wallis JC, Mayernik MS, Borgman CL and Pepe A (2010) Digital libraries for scientific data discovery and reuse: From vision to practical reality. Paper presented at the Joint Conference on Digital Libraries, Brisbane, Australia (21–25 June).

Wayne L (2005) *Institutionalize metadata before it institutionalizes you*. Reston, VA: Federal Geographic Data Committee. Available at www.fgdc.gov/metadata/metadata-publications-list (accessed 14 April 2011).

Zimmerman AS (2003) *Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists*. Unpublished PhD dissertation. School of Information, University of Michigan, Ann Arbor.

Zimmerman AS (2007) Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7(1): 5–16.

Biographical notes

Paul N. Edwards is Professor of Information and History at the University of Michigan's School of Information. His most recent book, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, 2010), was named a '2010 Book of the Year' by *The Economist*. His research centers on the history, politics, and culture of information infrastructures.

Matthew S. Mayernik recently completed his PhD in Information Studies at UCLA. His dissertation, *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators*, examined everyday metadata practices of small-scale field-based research teams in seismology, ecology, aquatic biology, and environmental science. Mayernik is now a Research Data Services Manager at the University Corporation for Atmospheric Research (UCAR) in Boulder, CO, USA.

Archer L. Batcheller received his PhD from the University of Michigan School of Information in 2011, with a dissertation entitled *Requirements Engineering in Building Climate Science Software.* He is presently a fellow in the Future Technical Leaders program at Northrop Grumman.

Geoffrey C. Bowker is Professor and Senior Researcher in Cyberscholarship at the iSchool, University of Pittsburgh. His most recent book is *Memory Practices in the Sciences* (MIT Press, 2006). He studies emergent teams in cyberinfrastructure and emergent forms of knowledge expression in the sciences and humanities.

Christine L. Borgman is Professor and Presidential Chair in Information Studies at UCLA. Her most recent book, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (MIT Press, 2007), won the Best Information Science Book of the Year award from the American Society for Information Science and Technology. Borgman's research on data practices spans the domains of earth and space sciences, life sciences, computer science, engineering, and the humanities.