

# Metadata accounts: Achieving data and evidence in scientific research

Social Studies of Science  
2019, Vol. 49(5) 732–757  
© The Author(s) 2019



Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/0306312719863494  
journals.sagepub.com/home/sss



**Matthew S Mayernik** 

National Center for Atmospheric Research, Boulder, USA

## Abstract

'Metadata' has received a fraction of the attention that 'data' has received in sociological studies of scientific research. A neglect of 'metadata' reduces the attention on a number of critical aspects of scientific work processes, including documentary work, accountability relations, and collaboration routines. Metadata processes and products are essential components of the work needed to practically accomplish day-to-day scientific research tasks, and are central to ensuring that research findings and products meet externally driven standards or requirements. This article is an attempt to open up the discussion on and conceptualization of metadata within the sociology of science and the sociology of data. It presents ethnographic research of metadata creation within everyday scientific practice, focusing on how researchers document, describe, annotate, organize and manage their data, both for their own use and the use of researchers outside of their project. In particular, this article argues that the role and significance of metadata within scientific research contexts are intimately tied to the nature of evidence and accountability within particular social situations. Studying metadata can (1) provide insight into the production of evidence, that is, how something we might call 'data' becomes able to serve an evidentiary role, and (2) provide a mechanism for revealing what people in research contexts are held *accountable for*, and what they achieve *accountability with*.

## Keywords

accountability, data, evidence, metadata

## Introduction

In the context of scientific research, metadata work is canonical infrastructural work: essential yet mundane, and ubiquitous yet often invisible (Borgman, 2003; Edwards, 2010). 'Metadata', as a concept, is more elusive than its close relative 'data'. From an

---

### Correspondence to:

Matthew S Mayernik, NCAR Library, National Center for Atmospheric Research, Boulder, CO 80307-3000, USA.  
Email: mayernik@ucar.edu

analytical view point, it is easy to consider ‘metadata’ as a (systematic or arbitrary) subcategory of ‘data’, or for metadata work to be subsumed as a component under some other topic of investigation, such as inscription, representation, standardization, or collaboration. Calling something ‘metadata’ is a social, situational, and even political designation (Boellstorff, 2013; Eve, 2016), and what are ‘metadata’ to one can be ‘data’ to another (Borgman et al., 2012; Mayernik and Acker, 2018).

Metadata-focused research by Science and Technology Studies (STS) and cognate scholars notes that metadata in the narrow sense – meaning structured information that describes data – are not in and of themselves sufficient to support most uses of data, or most data management and archiving tasks (Birnholz and Bietz, 2003; Leonelli, 2016a; Zimmerman, 2007). What these studies make clear, however, is that metadata work is intimately part of scientific work, while also often being ignored or separated out as specialized labor for specialized staff. Data archiving and curation professions and institutions have developed in fits and starts over the past 50 years and more, becoming central to the conduct of research in some areas, such as climate science, survey-based social sciences, astronomy, and seismology. Metadata, and their related difficulties, are a constant focus of people working in data professions (c.f. Feagraus et al., 2005; Gray et al., 2005; Michener et al., 1997; Vardigan et al., 2016). But even in the context of daily research work, where such data professionals or institutions are not part of the story, metadata processes and products are critical components of collaboration and data work (Edwards et al., 2011).

Historical studies show that metadata are integral to the production of scientific knowledge and infrastructure, both as an enabler and a problematic (e.g., Bowker, 2005; Edwards, 2010). Likewise, any scientific research based in the digital world will inevitably be interacting with metadata, again as both product and process (Mayernik and Acker, 2018). Recent phenomena such as ‘big data’ and ‘data science’ are exemplars for this characterization, being shot through with metadata-related phenomena (Greenberg, 2017).

Subsuming ‘metadata’ under the ‘data’ behemoth is thus an act of ‘ontic occlusion’ (Knobel, 2010), closing off discussion and attention related to particular kinds of work, and often particular people, within scientific institutions. This article is an attempt to open these discussions up. I present ethnographic research of metadata creation within everyday scientific practice, focusing on how researchers document, describe, annotate, organize, and manage their data, both for their own use and the use of researchers outside of their project. In particular, I argue that the role and significance of metadata within scientific research contexts are intimately tied to the nature of evidence and accountability within particular social situations. Focusing study on metadata can (1) provide insight into the production of evidence, that is, how something we might call ‘data’ becomes able to serve an evidentiary role, and (2) provide a mechanism for revealing what people in research contexts are held *accountable for*, and what they achieve *accountability with*.

### *Data and evidence*

I will start by outlining recent conceptual work related to the ontological and epistemological status of data. This discussion will provide a baseline with which to discuss ‘metadata’ as a distinct yet intimately related concept. Multiple lines of recent work have

developed conceptual definitions of ‘data’ that are built on the notion of evidence. Borgman (2015), synthesizing a series of ethnographic studies of science, social science, and humanities research, conceptualizes data as ‘entities used as evidence of phenomena for the purposes of research or scholarship’ (p. 29). Similarly, Leonelli (2015) defines data ‘as a relational category applied to research outputs that are taken, at specific moments of inquiry, to provide evidence for knowledge claims of interest to the researchers involved’ (p. 811; see also Leonelli, 2016a). On this view, researchers do not generate or collect ‘data’. They instead generate or collect entities, which might include physical objects, measurements, or other inscriptions, that can be used *as* data in relation to specific research goals. Being data, on this interpretation, is a rhetorical and sociological role that entities are made to play in particular situations (Rosenberg, 2013), not an inherent property of those entities that can be divorced from circumstances of their use.

Cumulatively, these works provide a consistent message. While it is certainly possible to find definitions of ‘data’ that present domain or technology-centric perspectives (see Borgman, 2015; Furner, 2016), the notion of ‘data’ in the context of scholarly work is intimately tied to the evidentiary value of whatever entities are being marshaled in support of an argument or claim.

This view aligns with a range of scholars from STS and cognate disciplines that view data, and therefore evidence, as achievements. Multiple scholars have pointed out that data are perhaps better conceived as ‘capta’, namely, entities ‘taken not given, constructed as an interpretation of the phenomenal world, not inherent in it’ (Drucker, 2011: 8; see also Kitchin, 2014: 2). As Collins (2013: 20) notes, ‘There are subtleties upon subtleties involved in deciding whether something is sound data.’ Collins’ decades-long study of gravitational wave physics provides numerous examples of how scientific evidence emerges out of configurations of work practices, social practices and documentary practices (see also Collins, 1998, 2017). ‘What counts as an acceptable claim in a science or any other realm of activity is a matter of tradition and the context within which the actors live and work’ (Collins, 2013: 201). If data practices are not aligned in a way that meets the expectations of the situation and institutions at hand, evidence is not achieved. This is a fundamental aspect of scientific investigation. In a study of the production of scientific evidence in the mid-19th century, McCook (1996: 183) states ‘The transformation of an object collected in the field to an object that appeared in a scientific paper was a long and often tenuous process of intellectual legitimation.’ This perspective is perhaps most simply encapsulated by Latour’s (1999) injunction that ‘[o]ne should never speak of “data” – what is given – but rather of *sublata*, that is, of “achievements”’ (p. 42, italics in original).

Given these relational and evidence-focused definitions of ‘data’, how should we conceive of ‘metadata’? My goal is not to provide a single all-encompassing definition of metadata. Instead, the goal is to advance beyond singular definitions of data and metadata toward ‘investigations of the production of the activities glossed by such concepts’ (Lynch, 1993: 201). Defining data as ‘entities used as evidence in support of a knowledge claim’ does not, for example, explain how those entities take on evidentiary roles. ‘[T]here is nothing self-evident about evidence. Evidence is evidence only within contexts of what may be considered to be evidential’ (Day, 2014: 6). The argument developed in this article is that metadata, however instantiated in local situated activities of scientific research, are central to enabling something to serve an evidentiary role, that is,

to *serve as* data. In particular, if data are entities used as evidence, then metadata are the processes and products that enable those entities to be accountable as evidence.

### *Accountability*

The concept of ‘accountability’ is central to understanding the contingent nature of evidence. Investigating accounts of scientific work provides a window into how particular readings of data are accomplished (Woolgar, 1976). Hoeppe (2014) depicts this intertwining of evidence and accountability in a study of the production of astronomical data: ‘researchers arrived at a consistent data set by sequentially and reflexively engaging diverse evidential contexts as contexts of accountability’ (p. 264). This view is nicely encapsulated by Woolgar and Neyland’s (2013) analysis of the informal and formal governance of everyday mundane activities like garbage collection and airport operations: ‘Through complex and co-ordinated interactions, evidence is accountably, demonstrably accepted as evidence for all practical purposes of the matter at hand. In our investigations of governance, this suggests treating the evidence of governance as an accountable accomplishment’ (ch. 4, p. 3). Their analysis hinges on two different conceptualizations of accountability: ‘First, there is an understanding of accountability as a mutual, constitutive sense-making action. This contrasts with organizational accountability whereby accountable entities are taken as the basis for assessment’ (ch. 2, p. 1).

Accountability relations of both kinds are central to understanding the production of evidence in particular situations. Accountability in the first sense, coming from the ethnomethodology tradition in sociology (Garfinkel, 1967), encompasses the ability of people to meet the practical and moral expectations of competency in social situations, and give accounts of routine and anomalous events (Lynch and Sharrock, 2003). In the second sense it refers to the things and activities for which people in research settings must be responsible and answerable (Fox, 2007; Mayernik, 2017). These are accountabilities tied to institutional constraints, like the applicable legal or regulatory frameworks in which people work and official organizational rules and procedures.

Evidence is produced and interpreted within particular settings, as appropriate, expected and accountable within those settings, whether they be courts of law, scientific research, or medical diagnoses (c.f. Pollner, 1987; Saunders, 2008). The dynamics of how evidence is interpreted involves navigating articulations between both kinds of accountabilities. Academic research likewise involves negotiating and navigating both immediate and institutional expectations for how evidence, and therefore data and metadata, is to be produced and interpreted. Livingston (1987: 142) describes this dynamic in the context of doing sociological research: ‘The sociologist arranges her work practices in such a way that they can be accountably justified by reference to and manipulation of [sociology’s edifice of proper methodological procedures]. The ways that she does this are, in fact, her practical methods of successfully analyzing her “data”.’

### *Metadata*

The notion of ‘metadata as processes and products’ comes from Edwards et al.’s (2011) analysis of metadata within scientific research. That article was written as a corrective to

discussions of metadata that tended to focus on metadata as product, for example formalized documents, written descriptions, and textual annotations created and used to manage, discover, access, use, share, and preserve informational resources. Standardized metadata products help increase the precision of research interactions by facilitating interoperability, machine readability and resource discoverability. Key for Edwards et al., however, is that informal metadata processes, such as personal emails, face-to-face discussions, and ad hoc document creation, are critical as lubrication for communication and collaboration related to data. Edwards et al. describe a number of cases where formal metadata products – standardized and structured descriptions of data – were absent, or essentially irrelevant to the production of scientific knowledge. Scientists in these cases were able to (or had to) use metadata processes to productively move forward with their research.

This ‘metadata as process’ perspective dovetails with findings from other metadata-focused studies of scientific work. Zimmerman (2007, 2008), for example, shows that metadata are only one factor among many that determine which and whether scientists can reuse data generated by somebody else. Data reuse, in Zimmerman’s characterization, is strongly tied to whether data users can project their own experience in doing field-based data collection onto someone else’s data. In other words, embodied knowledge gained through lived experiences in performing research and data collection is critical to assessing data and metadata produced by others. Shankar (2007, 2009) also emphasizes the essential embodiment of data and metadata work in a study of data management and recordkeeping in an academic laboratory: ‘[B]ecoming an active research scientist requires that the individual mesh his/her personal ways of working with the modes of work demanded of his profession – work that is rich, embodied, often tacit, and as such often anxiety-producing’ (Shankar, 2009: 163).

For data to ‘journey’ among different research settings, this embodied knowledge must be codified via some metadata-based mechanism, whether database structures, textual descriptions or non-textual media such as podcasts or online videos (Leonelli, 2016a). Leonelli emphasizes the critical role of data professionals in enabling such ‘data journeys’. Where data and metadata professionals exist, they often operate as intermediaries, performing the articulation and liaison work necessary to represent complex and dynamic scientific research within standardized metadata products (Karasti et al., 2006; Mayernik et al., 2014).

In many scientific research situations, however, data do not journey beyond the boundaries of a lab, project, or team, and no data professionals exist (Mayernik, 2016). This study specifically focuses on metadata creation in such situations, examining how scientists create metadata themselves in the context of day-to-day research practice.

## Case studies and methods

This article is based on ethnographic research conducted through two cases. Both settings have been described in more detail elsewhere (Borgman et al., 2012; Mayernik, 2011, 2016; Mayernik et al., 2013).

The first case was a study of data and metadata practices within the Center for Embedded Networked Sensing (CENS). From 2002–2012, CENS was a National

Science Foundation (NSF) Science and Technology Center with five partnering universities in southern and central California, and about 300 participants at its peak. CENS supported the research and development of sensing systems for scientific and social applications through interdisciplinary collaborations between scientists and engineers (see Gabrys, 2016, for another perspective on CENS).

Within CENS, I conducted a multi-sited ethnography focusing on the data practices of field-based science and engineering collaborations. The ethnographic work described here took place from 2007 to 2011, and consisted of participant observation, semi-structured interviews, and document analysis. My CENS participant observation consisted of sixteen trips to lab or field settings, encompassing approximately 200 hours of observations, along with regular informal interactions with CENS researchers. I conducted semi-structured interviews with fourteen CENS researchers, averaging 43 minutes in length. Questions focused on the interviewee's research questions and background, work and data flows, metadata practices and processes, software tools, data and metadata formats and structures, and long-term plans for data and metadata. The CENS analysis included examining published papers, data sets, documents, web sites, and emails created and used by my research subjects.

The second case is a study of research data and metadata practices within the University Corporation for Atmospheric Research (UCAR)/National Center for Atmospheric Research (NCAR). NCAR is an NSF Federally Funded Research and Development Center that conducts research in the atmospheric and related sciences. UCAR is a non-profit consortium that manages NCAR on behalf of NSF and over 100 universities members. Together, NCAR and UCAR employ about 1300 staff members, including nearly 500 scientists and engineers. UCAR staff draw from a number of scientific specializations, including climate science, meteorology, solar and space physics, and oceanography, as well as myriad engineering and technical domains. For simplicity, I will use the term 'UCAR' when referring to both UCAR and NCAR.

The UCAR study derives from participant observation research and professional work from 2011 to 2017. As a member of UCAR staff, I have regular interactions with scientists, engineers, and software developers, and data managers from across the organization. The interview quotes presented in the following draw from a set of 19 interviews with UCAR scientific and technical staff which took place during 2011–2013. One interview was excluded from analysis by request of the interviewee, and two others were excluded because they focused on non-research data, for example administrative or educational data. The sixteen analyzed interviews averaged 44 minutes in length. Interview protocols were adjusted from the CENS interview instruments. The UCAR analyses also draw on published literature and public web sites where applicable.

The CENS research projects centered around passive or active deployments of digital sensing technologies in field-based settings. In passive sensor deployments, researchers installed sensors in specific locations for long periods of time, with the goal of recording phenomena as they occurred. In active deployments, sensors were deployed for short periods of time, with the data collection procedures adjusted iteratively in the field. Depending on the deployment approach, researchers spent varying amounts of time in the field, from half-day trips to multi-month excursions. These field deployments of novel sensing and wireless communication systems were the distinctive aspect of CENS research from the point of view of the participating scientists.

UCAR-based research spans from small-scale project-based field studies to large-scale collaborations centered on observational facilities or computational models. Many of the UCAR researchers I interviewed conducted research via simulation and computational modeling of weather or climate phenomena. In some cases, researchers were actively developing the models themselves. In other cases, researchers used computational models to study atmospheric, oceanic, hydrologic, or solar phenomena.

Across both groups, metadata practices varied widely. Details about projects, data, instruments, and computer programs were spread among notes in Excel spreadsheets, idiosyncratic textual documents, notes in field notebooks, headers in data files, computer file names, labels on physical samples, project web sites, emails, and online wikis. As detailed below, all of these documentary forms were enacted in concert with informal communication (in-person and virtual) among collaborators.

Some of these practices and tools reflect the disciplinary affiliations of the researchers, such as the use of specific data formats and analysis tools developed for seismic or meteorological data and metadata. Researchers associated with other academic fields would likely use other discipline-specific tools and practices. Many of the tools used by the researchers observed in this study, however, were general purpose tools that are used ubiquitously across any research area, and even outside of academia, such as Excel, text documents, paper notebooks, and ASCII data formats. As such, my interest is not in the tools used to create and work with data and metadata, but in how any tools fit into the larger process of accounting for research entities as ‘data’. The following cases therefore serve to illustrate particular kinds of metadata accounts, that is, descriptions and explanations coupled with work and collaboration procedures that enable specific things – measurements, samples, observations – to have evidentiary value in the context of a scientific argument.

## Metadata in day-to-day scientific research

This section is organized around the two notions of ‘accountability’, discussed earlier, from Woolgar and Neyland (2013). First, I present findings from the CENS and UCAR studies to illustrate how paying attention to metadata can illustrate institutional and organizational accountabilities related to data, that is, what people are *accountable for*. Second, I discuss how paying attention to metadata, and their presence and absence, is useful to illustrate how people achieve accountability for their work. Metadata can be a useful lens for examining what people are *accountable with*, including the types of documents and associated accounts that are marshaled when problems with data appear.

### *What are people accountable for?*

Consistent with other studies of responsibility and accountability in relation to data work (Leonelli, 2016b; Wallis and Borgman, 2011), responsibilities for data in the research settings I studied were often distributed and vaguely defined. What was clear, however, was that researchers’ primary accountabilities tended to be related to the research questions and goals, not the data per se, as illustrated by this quote from a CENS graduate student:

The data as a goal is only 5% of the entire research. ... I mean, it's a tool in order to do some research, but the data itself is not really the goal. (CENS Int. #3, 2010)

The production of sensor data involves 'an assemblage of practices, objects, spaces and actors' (Calvillo, 2018: 384). Many CENS and UCAR researchers who participated in this study were primarily interested in, and responsible for, the production of technical tools, algorithms or analytical methods. Data-related tasks were often seen as additional work, or tangential to the ultimate goal of writing papers or producing particular findings. As noted in the following interview, not having to deal with questions about data was itself a form of good luck.

Requests for data, ... luckily they're not very frequent because they'd be hard to deal with, because you'd have to package up the data and process it and put it on an ftp site or something like that. We just don't have time or resources to do it. (UCAR Int. #5, 2012)

To be accountable for data, there must be an 'other', that is, some party (real or abstract) to which a person or organization must be accountable (Bovens, 1998). External accountabilities may be enforced informally via norms and social institutions or via explicit legal regimes. For example, some UCAR projects are funded through formal contracts with corporate or government funders that stipulate specific operational products, such as models of particular weather phenomena, forecast systems, or environmental decision support tools. Depending on the contract, data, and/or metadata, in multiple forms, could be critical parts of the deliverables, as noted in the following quote:

So, we typically deliver the code, code documentation, and then some technical documents saying, 'This is what the algorithm does. These are our assumptions. This is what it needs. This is what it does. And here's what the output does and here's how to use it.' So, when we deliver something like that it's the whole package, and it's gone through an approval process with the [contracting organization] and everything. That's a very formal delivery and that's typically what we do when we deliver a system is we give them all the accompanying documents with various levels of formality. Some programs have very specific, 'We need this, this, this, and this', and some just say, 'We need a technical document that explains what the algorithm does. We need code documentation in some form, either comments in the code or [some other way] ....' (UCAR Int. #14, 2013)

In order for metadata products, such as the technical documentation described in the above quote, to exist, somebody needs to be accountable for their creation. For example, a seismic research team within CENS had plans to submit their data to the Incorporated Research Institutions for Seismology (IRIS) data archive. When I asked a Principal Investigator on the project what kind of documentation was necessary when submitting data to IRIS, he described how they would need to create a comprehensive metadata package separate from their data. In the quote below, SEED refers to the Standard for the Exchange of Earthquake Data, a data format established by the seismic community in the 1980s:

Basically, what they [IRIS] would prefer to have is the data, the Mini-SEED file; that's what they deal in and they love that. ... But then you have to send them, what they call, data-less



SEED volume. So it's like a complete SEED volume, it's just that there's no data in it. So that's ... where they learn where the stations are, what are the properties of the station, which instruments are on there, so ... We'll have to spend a little time creating those; we'll create those and send it off. I have a whole staff across the street that's very good at doing that, so I'll get some help on that. (CENS Interview #5, 2010)

This quote is notable for a few reasons. First, the collocation of the seismic data with the metadata about instruments and locations (the 'data-less SEED volume') would not happen until the project was completed and the data were to be submitted to the external data archive. And second, the metadata that document the data station and instrument properties were to be created by 'staff across the street', not the research team. 'Staff across the street' referred to staff of a regional seismology data archive that the Principal Investigator also ran.

Another notable aspect of the above quote is that it completely omits mention of a whole swath of metadata processes used by the seismic team to support the field-based data collection procedures. Among these processes were daily update emails that described what work had taken place in remote field sites each day. The utility of these daily field update emails varied from person to person on the team, with team members who were closer to the field work typically valuing them more. The same Principal Investigator on the project stated:

Frankly we don't do anything with those notes. ... They're used to jog the memories for about a month and [when] something needs to be fixed. (CENS Interview #5, 2010)

In contrast, the primary field engineer on the project used the email notes as a searchable archive of deployment information. He used precise syntax in his emails, specifying exact serial numbers for equipment and giving full names for field sites to enable precise searching later. He was frustrated by emails written by other team members that did not use the same consistent syntax, because it was then difficult for him to fully trace the activities that took place while he was not in the field. Direct personal communication filled any gaps. On one occasion, I observed a two-hour conversation that took place the first morning after the primary engineer returned to the field from a multiple-week absence. During the conversation, the engineer and a staff scientist discussed the entire sensor deployment, 50 different sites in total, with the engineer asking about the current status of each site. Neither individual referred to any documents during this discussion; the entire conversation drew on their memory and embodied knowledge of the field work and sites.

This deeply embodied knowledge is critical for individuals who are accountable for field work. Members of the research teams can face sanctions if problems occur with equipment or field sites for which they are responsible. For example, the CENS aquatic biology team experienced a significant learning curve when it came to getting useful data from a new suite of sensors. Getting over this curve centrally involved developing calibration and documentation methods. For example, a portion of the team's sensor measurements from a prior year had a different water depth reading than the measurements before and after, without obvious reason. The notebooks from the technician who was

doing the data collection at that time were not helpful in answering this question, as a graduate student noted:

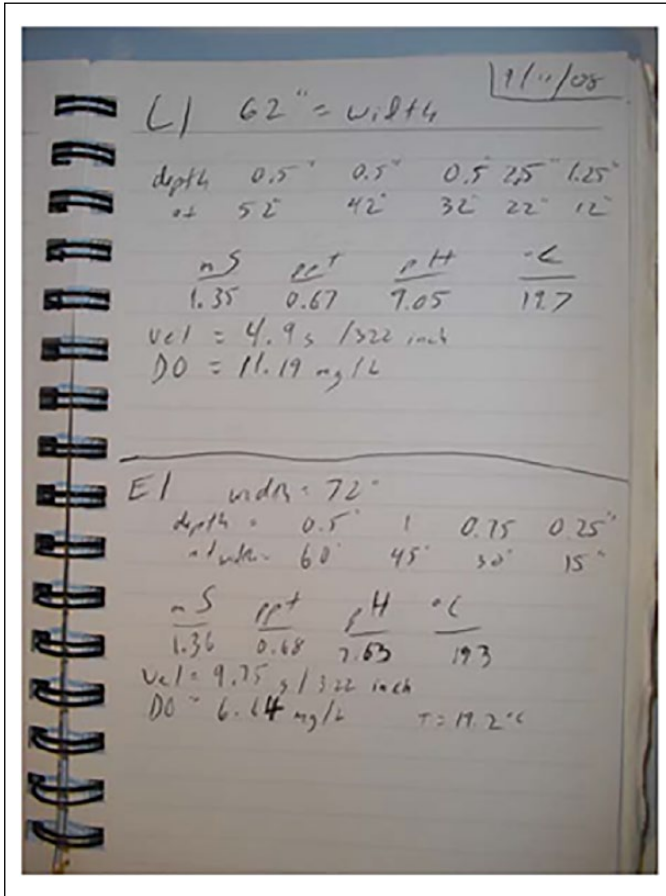
I haven't been able to figure out from the [former technician's] notes if that's a real change in where that sensor was deployed or if it's, you know, if the sensor got sent back and was put back in water without calibrating the depth sensor or pressure sensor. (CENS Interview #9, 2010)

Until that discrepancy was resolved, the measurements could not be used as data in the student's analysis. The technician responsible for these gaps in usable measurements and documentation was ultimately replaced. New procedures, developed with a new technician, involved an Excel file that specifically documented field work that involved anything related to the sensors. The Excel file included calibration parameters that were later used to adjust the values measured by the sensor, as well as comments about the sensors themselves, such as 'Unable to download files. Sensor removed from field.' Other notes indicated when the depth reading on the sensor was reset to zero, when the power was reset on the sensor, and when the sensor gave any error messages.

Metadata processes were even more informal on other CENS projects. The CENS environmental science team regularly visited a small number of field sites to measure stream contamination. During these excursions, sensor readings and associated metadata were written down in a notebook and subsequently transcribed by hand into a computer after the team returned to the lab. Recording the information in the field was seen as a basic bookkeeping task. For example, I was given this bookkeeping role the first time I took part in a field site visit. Figure 1 shows a notebook page from a subsequent trip in which I was again asked to be the data/metadata recorder. My practice for recording and documenting sensor readings and stream measurements emulated notebook pages from previous trips. The top half of the image shows the data collected at site 'C1' on this particular date, starting with the stream width measurement, 62 inches, and the stream depth readings. The stream measurements and depths are followed by the readings from the sensor, 'mS', 'ppt', 'pH', and '°C', which are noted solely via the relevant units.

Below the sensor readings are the stream velocity measurement ('vel') and the dissolved oxygen sensor reading ('DO'). Note that the stream velocity was recorded as '4.9 s/322 inch'. At some point in the past, the team had used seven lengths of a pole that was lying near the stream to measure a distance to be used to make a stream velocity measurement. From that point on, they made all of their stream velocity measurements in units of 'seconds per 7 poles', or 322 inches using conventional units. Looking at the team's main field notebook, I saw numerous measurements in 's/7 poles'. I also saw stream depths and widths recorded in units of 'Samantha's boots' and 'Morgan's shoes'. The Principal Investigator on this project indicated to me that she gave the students complete responsibility for the data collection procedures. She almost never looked at the data directly, and only saw the data in charts and figures after some analysis had been produced by the students.

This section has discussed how paying attention to metadata products and processes can be helpful in understanding what tasks, products, and processes different members of a research team are accountable for. The next section focuses on how metadata processes and products can also illustrate the ways that researchers achieve accountability.



**Figure 1.** Photo of my notebook from a field trip with CENS environmental science researchers.

### What are people accountable with?

With an outsider's point of view, it is easy to see things in researchers' metadata practices that seemed inadequate, incomplete or problematic. This is a core finding from Garfinkel's (1967) foundational text on ethnomethodology, in particular, his study of 'good' and 'bad' medical records. There were numerous examples of ostensibly 'bad' metadata in my studies. As one example, one CENS team analyzed their data using an Excel data file that contained 20 separate time-series analyses of sensor measurements, with each analysis on a different worksheet in the Excel file. Each worksheet had the same structure and embedded formula workflows. The team made notes in the Excel file directly, using flags in particular cells. The fifth worksheet contained a column titled 'Change in Length (mm)' that had a note from one team member, 'DID I DO THE CHANGE OF LENGTH INCORRECTLY ON ALL THE OTHER ANALYSIS (IN

THE WRONG DIRECTION)?!!!!’ The next worksheet, the sixth, had a note from the same person saying, ‘This is the correct way to determine change of length.’ All subsequent analyses, the seventh through the twentieth, retain the ‘DID I DO THE CHANGE OF LENGTH INCORRECTLY ...’ note, indicating that they were copied and pasted from the previous worksheet without the note being changed after the length measurement calculation had been corrected.

Likewise, a UCAR researcher described how the documentation related to a particular data transformation process was being done ‘pretty sloppily’ (UCAR Int. #9, 2012). But, as also suggested by the ethnomethodological view, when questioned further he was able to describe exactly what that process was: he wrote brief text files that described the functions of a particular piece of software. As he stated, the descriptions were not highly detailed, but effectively jogged his memory in situations where he needed a reminder of how the software worked.

The move from saying that documentation was done ‘pretty sloppily’ to describing in detail a specific process is an example of a metadata account. Researchers can typically account for their metadata practices, even if they seem less than ideal from the outside. A CENS seismology graduate student described his need to keep track of equipment fixes in the following manner:

You have to keep track of everything because first, like if you want to do the instrument correction ..., you need to know which sensor came from where, because different sensors have different responses. ... And also, you want to know if there’s some problem that keeps repeating itself at the same site, you want to see, you know, what’s actually happening. So, if we don’t have really substantial sort of notes, but good enough to tell [what’s happening] .... (CENS Interview #3, 2010)

When I followed up this comment by asking him whether they had standard types of documents to keep track of equipment fixes, he said ‘No, it just goes through the email.’ Thus, ‘needing to keep track of everything’ was achieved through emailed notes that were ‘good enough’.

Certainly, creating metadata can take time and effort, and, as noted above, researchers may not be responsible or accountable for such work. It can be hard, or even impossible, to know whether that work will ever show any practical utility. But occasionally, such metadata can prove to be critical:

Q: So if you’re archiving a dataset, or you pull a dataset that’s archived from somewhere else, do you think it’s important for the tools that were used to process this data to be mentioned?

A: It’s icing on the cake. I don’t really like icing, [chuckle] but you know I think metadata is very difficult to create. It’s just a pain in the butt. But every once in a while, you thank whoever lucky stars that they put it in there. Because, oh, well it was maybe MATLAB that you used or maybe IDV and maybe in one of them the byte order is different, that may help you. (UCAR Int. #16, 2013)

The combination of metadata products and processes that scientists use often follows these uncertain payoffs. If formalized documents are known and established to have specific utility, they continue to be created. If not, certain documentary processes might

get dropped. For example, a research staff member within the CENS ecology team showed me a series of files that he considered to be metadata for his sensor data. Among them was an Excel file of sensor calibrations and sensor changes. He noted how his documentation practices had changed over time:

At one point I was keeping track of how we did the calibrations, [and] when we replaced the sensors for example. We kept track of those things. But ... I'm not doing it any more [laughs]. I'm doing it, but not keeping track anymore. (CENS Interview #7, 2010)

Later in the discussion he said that he did write calibration information in his notebooks while in the field, but he had stopped transcribing it to the Excel file because of the work effort involved. He said that it would be difficult to go back and find the calibration information in his notebooks, but if need be, he could find it.

Researchers need to be able to account for metadata products that might get lost over time, whether they get dropped deliberately or accidentally. A student in the CENS environmental science team described the ways that best efforts can go awry:

I'm sure if you think you're being good about like annotating what you need to take notes of, but then somewhere down the line, it's like three months, six months, a year from now, you go back to look at it, 'I swear I made a note of this'. And now, I can't find it or I actually didn't. I think that probably comes up more frequently than we'd like. (CENS Interview #14, 2011)

In this case, the student reduced the impact of losing metadata by running multiple replications of an analysis and by retaining samples so that they could be re-analyzed later if necessary. Samples could be retained 'more or less indefinitely', according to the student, which made re-analysis possible, but also made it necessary for the team to have effective systems for sample labeling.

Metadata accounts also often illustrate social aspects of the research process. When I asked a CENS graduate student about how data are organized after researchers bring them back to the lab, the student's immediate reaction was to laugh and say, 'not very well'. Later, the student expressed her own opinion of her data organization methods:

I mean it's certainly not a very fancy organizational system, it's a bunch of folders. And it's reasonably doable right now. ... But time-wise, I know it's not huge but it probably should be bigger and then probably make things easier. (CENS Interview #9, 2010)

She then noted that her field technician took on many of the data and metadata management activities:

I mean as far as the organization and the upkeep of it, [the technician] has been really good. ... [The technician] goes out every other Friday, downloads the sensors, basically, just looks through the data, makes sure it's in a good format ....

Thus, the student can account for the limitations in her practices by pointing to the ways that her efforts combine with the field technician's metadata practices to meet the expectations of good practice within their lab.

The last kind of account I illustrate relates to computational simulation-based research. In computational modeling projects, such as weather or climate modeling, researchers rarely collect their own original observations. In this kind of research, it is common for individuals to programmatically access data provided by a collaborator or external data archive. Data analysis scripts and model simulation code are written to directly access the data and pull them into a computational workflow. Once written, these data ingest scripts are often not revisited unless some error occurs. Metadata changes may themselves be the cause of errors in this kind of a workflow, and errors can propagate via software before the knowledge of how to understand or accommodate them reaches the researchers themselves. This is illustrated in the following quote from a UCAR computational scientist:

If there's a change to the software that's in control of the [data] archiving, you sometimes end up with problems on the [data store]. And then, when I go to use it, I'm not informed of those changes and then my software doesn't work anymore because it's expecting a specific, you know, archive path or format that the data are in and then it's no longer like that. ... there's no place where I can go and find out when that change was noted. ... There is a MySQL database available ... and you can kind of access that and view around metadata a little bit about the data, data about the data, but again, if the person in charge of running the archive script doesn't take the time to submit the metadata to the database, it's not useful. (UCAR Int. #19, 2013)

Metadata processes must continue to supplement structured metadata products (a database in the above quote), and researchers must be able to account for metadata errors that they may not have caused, but that came to them.

## **Discussion: Metadata's role in the production of data and evidence**

What role(s) do metadata play in enabling evidence, and therefore data, to be achieved? The two results sections showed how metadata can be viewed in two ways: (1) as something to be (or not to be) *accountable for*, and (2) as something to be *accountable with*. This distinction of metadata as being something people are both (or either) *accountable for* and *accountable with* helps to clarify the work of producing scientific evidence. The arguments made here are focused on research contexts in which knowledge production is the overarching goal. This discussion is not intended, therefore, to characterize use of metadata in such settings as social media, surveillance, and disinformation campaigns (Acker, 2018; Clement, 2014; Mayernik and Acker, 2018), though it may complement previous studies that show how accounting practices are important to understanding people's everyday interactions with information, data, and metadata (Chamberlain and Crabtree, 2016; McKenzie, 2003; Vertesi et al., 2016).

A central element of the story in this article is the work to produce accounts of how scientific evidence, in whatever form, came to be. Numerous scholars have described how the production and circulation of accounts of work are central to the social organization of day-to-day scientific and engineering settings (Bowen and Roth, 2002; Orr, 1996; Roth and Bowen, 2001; Traweek, 1988). Work accounts are likewise central to

the organization and production of data, that is, something to be used as evidence. Becoming a competent researcher involves becoming adept at producing, circulating, and evaluating accounts of data work, including accounting for data problems, whether ‘dirty’, missing, or inconsistent data, as noted in the CENS and UCAR cases. Because day-to-day research so heavily involves mess, contingency, and imperfection, the lack of data problems can be more remarkable, and require more extensive explanation, than the presence of such problems (Helgesson, 2010). Researchers who have experience with data production know to expect to see problems with data produced by others. A significant part of becoming a geologist, meteorologist, or sociologist involves developing an embodied knowledge of what geologic, meteorological, or sociological data look like. Metadata products and processes provide critical gel that enables something to exist as data in whatever form or setting is at hand. For specific entities to be accountable as data, and therefore to be able to be marshaled as evidence, some metadata product and/or process must be in place to reduce the frictions involved in data production and sharing (Edwards et al., 2011).

Metadata support both the ‘ordering from within and without’, to use Suchman’s (2007: 202) phrase – this phrase and associated discussion resonates with the metadata picture I describe earlier, and with Woolgar and Neyland’s dual notion of accountability. Suchman describes how scientific and information management systems serve dual roles in organizing the work they support, as, for example, when ‘systems designed to track planes are simultaneously used by workers as resources for communicating their own activities to co-workers and by management as resources for evaluating how the operation is running’ (p. 203). The following two sections use this framing to discuss the dual roles of metadata as (1) resources for communication, or something to be *accountable with*, and (2) resources for evaluation, or something to be *accountable for*.

### Ordering from within

In supporting ‘ordering from within’, the role of metadata is to help ensure the practical accomplishment of day-to-day work. ‘Rather than replacing users’ expertise and experience in the lab, metadata serve as *prompts* for users to use embodied knowledge to critically assess information about what others have done’ (Leonelli, 2016a: 106, italics in original). Leonelli’s characterization echoes Suchman’s (1987, 2007) description of the role of plans, protocols, and other prescriptive documents in ordering human activity. In Suchman’s characterization, plans serve as orienting devices for people to work toward a common end goal. Likewise, a central role of metadata is to help researchers to orient each other’s work toward particular goals, methodological procedures, and objects (including instruments, samples, measurement sets, and computational models). Metadata are thus situated and emergent phenomena, but often exhibit a stability and routine nature due to the relatively stable relationships researchers have with each other and their environments (Agre, 1997).

Looking across the CENS and UCAR cases, the depicted metadata practices met the ‘expectations of sanctionable performances’ (Garfinkel, 1967: 199), in the sense that apparent gaps, inconsistencies, or mistakes could generally be accounted for, and researchers could provide statements explaining their actions. In cases where metadata

practices were not accountable, as with the CENS aquatic biology technician whose poor sensor calibration documentation resulted in multiple months of unusable data, sanctions occurred – in this case, a job was lost.

In the CENS and UCAR cases, accounts for metadata practices came in multiple permutations. The following list, not intended to be exhaustive, sketches particular kinds of accounts that were heard repeatedly.

Our metadata may not be complete, but we know what we need to document and can do so if necessary.

My metadata practices may not be sufficient individually, but as a team our practices are sufficient.

I do not have my metadata processes and products all available in a displayable form, but I could if I had enough time.

My metadata practices may have been inadequate in one situation, but I can show you many other situations in which they were adequate.

We have established metadata practices that would be effective if everybody followed them all the time.

By having practices that enable these kinds of accounts, researchers can have incomplete, limited, or occasionally problematic metadata or data and still meet the expectations of competency and evidence in their local research situations. As Woolgar (1981: 509) notes, descriptions, whether verbal or written, are ‘only more or less reliable by virtue of their being treated that way for the practical purposes at hand’. If researchers can account for any perceived problems in their data and/or metadata in a situationally satisfactory way, their identity as a researcher within their communities of practice will not be challenged in that regard (Wenger, 1998).

Researchers routinely articulated their metadata practices in terms of whether they were sufficient to achieve their own and their colleagues’ research goals. By using language such as ‘good enough’, ‘my thinking is always basic minimum’, ‘I think that [losing track of notes] probably comes up more frequently than we’d like’, to describe their metadata practices, researchers acknowledged the inherent incompleteness of their efforts. They developed an understanding of the situations in which metadata incompleteness can be tolerated and accounted for as part of their professional and personal development as scientists (Shankar, 2009). Senior graduate students and research staff have experienced the ups and downs of multiple projects. They know that field sites are unpredictable, that equipment and/or software fail unexpectedly, and that data collection processes may need to be adjusted as a project proceeds. These events become part of the social fabric of their research settings and group interactions (Roth and Bowen, 1999). Without metadata processes or products that account for such events, ‘data’, and therefore evidence, cannot exist. Researchers must be able to account for both routine and unanticipated events that occur during the research process, while at the same time fitting metadata practices into the other tasks, both social and individual, that they are



expected to perform (such as manipulating machines, writing papers, and helping team members).

Becoming a scientist, identifying as a scientist, requires becoming adept at achieving accountable evidence within whatever social milieus one is embedded. As Day (2014: 81) describes, 'one must prove one's self within systems of proof or evidence; this is the political and moral economy of life's meaning – of one's being – in such systems.' This connection between morality and evidence, also noted by Woolgar and Neyland (2013: ch. 4, pp. 24–25) and Yakel (2001), could be seen in researchers' occasional denigration of their own practices. Researchers occasionally described feelings of ambivalence regarding their metadata practices and in some cases laughed at their own practices, calling them 'not very fancy' or 'sloppy'. This despite their many other scientific accomplishments and successes in publication and grant writing. Vertesi et al. (2016) note how these kinds of moral implications are a general feature of personal data management practices.

Looking across the research cases and associated theory, metadata, whether process or product, when looked at as something to be *accountable with*, can be viewed as having the following characteristics:

Metadata are always indexical and selective.

Metadata are enacted according to the particular occurrences of immediate situations in a manner adequate for enabling immediate research tasks.

Metadata encompass negotiated shared meanings. They are imbued with the expectation that readers or users of the descriptions will have knowledge of how to interpret them.

Metadata are 'account-able' by their creators. Scientists or other researchers are able to give 'accounts' for why metadata descriptions are or are not created for their data, as well as for the selectivity of those descriptions.

### *Ordering from without*

On the flip side of Suchman's phrase 'ordering from within and without', another role of metadata is to help ensure that research findings and products meet any externally driven standards or requirements. As noted above, externally driven data sharing was not a specific goal for most of the researchers observed in my study. Researchers were usually willing to share their data with other interested individuals, but rarely documented their projects or data specifically to facilitate such sharing. Research projects did not need to enable widespread data sharing in order to ensure the continued existence and success of their research programs. Data sharing with outside users, if and when it occurred, was considered by different individuals to be either an added bonus or a source of additional work, and sometimes both.

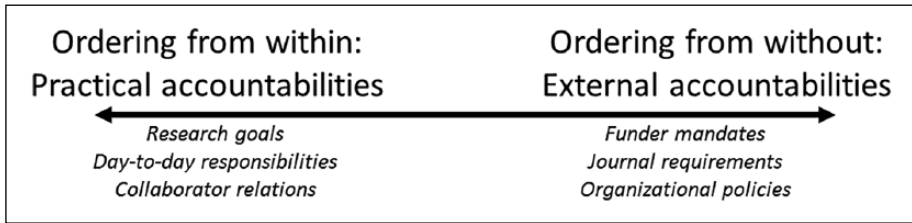
Leonelli's (2016a) concept of 'data journeys' provides a useful framework for discussing how these externally focused accountabilities manifested themselves in the CENS and UCAR cases. First, in the cases in which researchers described external motivations for the creation of metadata, they typically identified a priori a particular target

for those metadata, either a specific data repository, such as the IRIS repository of seismic data, or a particular client in the case of the UCAR weather forecast modeling group. This suggests that when the destination of a ‘data journey’ is known or pre-ordained, externally oriented accountabilities associated with ensuring that data reach the target destination can emerge as a strong factor in driving metadata processes and products. Targeted circulation of data often provides an illustration or map to the relations among stakeholders in a scientific endeavor (Walford, 2012). Once again, however, data and metadata do not need to be perfect to ‘journey’ outside of their originating settings. Data and metadata just need to be ‘just good enough’ (Gabrys et al., 2016) to support the goal of the journey, whether to meet external data-archiving requirements or to enable a specific audience to understand and use the data in a ‘virtual witnessing’ sense (Woolgar and Coopmans, 2006).

The second aspect of Leonelli’s ‘data journey’ framework centers on the importance of data and metadata intermediaries (also noted by Mayernik, 2016; Rood and Edwards, 2014). Such intermediaries serve as ‘invisible technicians’ (Pontille, 2010; Shapin, 1989) who get lost from view when the analytical emphasis is on data production or use. When looking at metadata, however, these intermediaries (or the lack thereof) become central to the narrative. Most research teams observed for this study, with the exception of the CENS seismic team, lacked intermediaries who were specifically tasked with ensuring external accountability for data (or metadata). Many teams did, however, have informal or de facto roles and expectations in which certain people were tasked with metadata work, often field technicians and graduate students. Such data intermediaries feel and respond to accountabilities that are not primary for data creators. Data intermediaries face being answerable for their own work, including the responsibility to be ‘documenting documentation’ (Bearman, 1994: ch. 8). There is a reflexivity in this need for somebody to be accountable for metadata. The distinction drawn by Edwards et al. (2011) between metadata as process and product illustrates how documents themselves are created in particular social environments, often in concert with informal communication processes.

There is no clean boundary between these practical and external-facing notions of accountability. It should also not be taken as a given that the distinction between metadata as process and product map cleanly to either accountability category. In the CENS seismic case, for example, the seismic measurements were compiled using a standard file format, SEED, that provides for the creation of particular data and metadata products. In the same way that other CENS projects produced non-standardized data and metadata, however, the production of these SEED files by the CENS seismic team was an accountable achievement reliant on a wide range of informal metadata processes. Likewise, many UCAR weather and climate modeling teams employ widely used tools to produce standardized data and metadata products. Interpreting how to apply metadata schemas and vocabularies to ensure organizational compliance with community standards, however, is not a straightforward technical process. Scientists making use of standardized data and metadata products rely considerably on informal metadata processes (Rood and Edwards, 2014; Zimmerman, 2008).

Figure 2 depicts how the two forms of accountability flow into each other. In Suchman’s terms, metadata processes and products help order a work environment from



**Figure 2.** Metadata accountability spectrum.

within by supporting day-to-day research tasks and collaborations. Metadata processes and products are also critical in ordering a work environment from without when funders, journals, or data repositories specify data and metadata accountabilities. Researchers deal with both kinds of accountabilities, from within and without, on an ongoing basis. Becoming a successful researcher involves developing ways to meet the accountability requirements across the spectrum.

### *Metadata occlusion*

As this discussion illustrates, STS and related research has been touching on metadata for many years, through studies of scientists' practices of being accountable for data via documents and verbal descriptions. However, metadata phenomena are often occluded in studies that focus narrowly on data. I use a few examples drawn from recent literature to illustrate how the metadata processes and products depicted in this study can be overlooked or glossed when studies focus on phenomena like inscription, interpretation, or representation. The following examples are not meant as critiques of the respective studies, each of which presents important insight into how scientific data and evidence are produced. My intention is to illustrate how an orientation toward metadata would result in different kinds of questions, and different analytical focus.

In the first example, Busch (2016) presents an analysis of what counts as evidence within statistical-based research. Busch lays out five steps needed to construct statistical data: attributing variables, isolating characteristics of interest, assigning values to the variables, identifying a population, and taking a sample. Statistical analysis and interpretation can only take place after these steps have been taken. As Busch notes, each of these steps may present their own challenges, and require specific decisions to be made. For example, '[h]ermeneutical issues ... arise in decisions about which data to include, what to do with outliers, with missing observations, with "cleaning" the data' (p. 667). There is no discussion, however, about how such decisions are made. It is certainly the case that statistical researchers must document and/or account for critical decisions about variables, populations, data cleaning, or data interpretation. They must be able to justify their work according to the practical and external accountabilities noted in Figure 2. In Busch's account, however, these metadata issues are submerged beneath the focus on data construction and analysis.

Similar examples can be found in Levin (2014) and Gabrys et al. (2016). Levin's analysis depicts the kinds of work needed to translate between clinical and laboratory evidence in the biomedical field. Again, interpretation of data is a key element of the story: 'Though technological innovation, through the creation and value of particular types of "data" is posed as a solution to the problem of translation, human interpretation emerges as a fundamental necessity for the alignment of the laboratory and the clinic. Data cannot exist independently of human practices, such that the negotiation of the form and value of data remains one of the main challenges facing translational research' (p. 107). As Levin describes, one significant component of data interpretation involved understanding machine capabilities and uncertainties. Few details are provided, however, about how researchers actually account for any uncertainties related to the data collection machines. This almost certainly involves producing, documenting, and adjusting for calibration curves, as described in the CENS case, among other metadata products and processes.

Likewise, Gabrys et al. (2016) discuss how groups of citizen science projects produce data and associated data stories to build evidence about environmental problems. As they describe, data stories were essential to the success of the activist groups seeking environmental interventions on the basis of seemingly intractable data. 'Even with these preliminary forms of evidence [measurements of air pollution by citizen scientists], more work was still required to establish patterns in the data so that stories could be generated, and so that citizens' concerns could be figured into a collective account' (p. 9). The citizen scientists faced significant challenges in collecting and presenting data in ways that were understandable and acceptable to regulators and legal institutions, in particular because they were trying to make the argument that existing regulatory data collections were not measuring sources of air and water pollution sufficiently. In one example, citizen scientists were able to organize a teleconference with environmental and health regulators in which the data stories were presented, leading to more focused monitoring in an area noted as problematic by the citizen scientists. In Gabrys et al.'s analysis, data stories were the glue that held together the disparate data. On the one hand, this is a great example of how data must be coupled with accounts of the data that meet the evidentiary requirements of the situation. On the other hand, Gabrys et al. provide little detail of how the citizen scientists were able to formulate the appropriate accounts to justify a use of data that was out of the norm in the context of environmental regulatory monitoring. What kinds of documents were presented? How were they accounted for? What justifications were provided for instruments, measuring protocols, and data integrations? These kinds of activities inevitably involve metadata products and processes like those discussed above.

In some ways, this discussion is simply recounting foundational insights from Actor-Network Theory and ethnomethodology, namely that scientific research is organized around processes of 'inscription' in which the objects of study are transformed across time and space by machines and people via chains of traces, graphs and texts (Latour and Woolgar, 1979); and that such inscriptions are marshaled and activated (or bypassed) in particular settings via accounting mechanisms appropriate for the situation at hand (Garfinkel, 1967). These findings become particularly salient

when the analytical focus includes metadata, alongside data. Paying attention to metadata within knowledge production settings, the processes and limitations of textual inscription become clear, and the centrality of the corresponding metadata accounts are inescapable.

Achieving data, and therefore evidence, involves chains of inscriptions and chains of accounts that activate and situate those inscriptions. But not every inscription is used as data, and not every account of scientific work should be designated as metadata. This prompts the question, when is something evidence, data, metadata, or just an inscription (Lynch, 2013)? There are no easy answers to this question, though Borgman notes that '[s]ome objects come to the fore as data and others remain in the background as context or simply as noise' (Borgman, 2015: 223). This distinction between 'foreground' and 'background' data, first presented by Wynholds et al. (2012), is useful to move beyond circular statements about the production of scientific evidence, such as the 'evidence becomes evidence when deemed evidential' formulation provided earlier in this paper. It is possible to investigate what researchers foreground as evidence and what they keep in the background as context or points of comparison (Wallis et al., 2013). Even if '[r]arely can a magic moment be established when things become data' (Borgman, 2015: 62), it might be insightful to look closely for times when researchers marshal or 'bring to the fore' things that were previously 'backgrounded'. In the CENS and UCAR cases, these included sensor calibration curves, field notes, and descriptions of software or algorithms. Instances in which such traces and accounts are prominent provide indications that an 'evidential' situation is at hand, that is, a situation in which the boundaries of what counts as 'evidence' are being contested.

Paying attention to metadata-related phenomena can itself provide insight into the accountability relations that apply to researchers in different settings. In particular, metadata provide a useful lens to examine more closely what researchers are *accountable for*, and what they are *accountable with*. Some people are *accountable for* data and metadata, and some are not. These accountabilities vary among research teams, as well as over time within the same team, and are often designated informally or de facto. When metadata are the focus of investigation, these social relations and responsibilities for data work come into focus. Investigating metadata is also a useful way to reveal what researchers are *accountable with*. In particular, researchers' accounts of their metadata and data illustrate the ways in which evidence comes to be. Any researcher with even a limited amount of experience has encountered problems in collecting and analyzing data. But what kinds of data problems can be overcome, and what cannot be overcome? Accounts of how data problems could or could not be overcome are valuable resources for understanding the nature of evidence in different settings.


### Acknowledgments

I am grateful to the CENS and UCAR study participants for their time and insight. Erica Johns and Nancy Hunter contributed as co-interviewers for some of the UCAR interviews. Thank you to Christine Borgman, Peter Darch, Milena Golshan, Irene Pasquetto, Sergio Sismondo, the anonymous reviewers, and attendees of a 2017 seminar at the University of Michigan for comments on earlier versions of this work.

## Funding

The CENS portion of the empirical work was supported by the Center for Embedded Networked Sensing (NSF Cooperative Agreement #CCR-0120778), Microsoft Technical Computing and External Research, and the UCLA Graduate Division. NCAR is sponsored by the US NSF. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of NCAR or the NSF.

## ORCID iD

Matthew S. Mayernik  <https://orcid.org/0000-0002-4122-0910>

## References

- Acker A (2018) *Data Craft: The Manipulation of Social Media Metadata*. New York: Data & Society Research Institute.
- Agre PE (1997) *Computation and Human Experience*. Cambridge: Cambridge University Press.
- Bearman D (1994) *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh: Archives and Museum Informatics.
- Birnholtz JP and Bietz MJ (2003) Data at work: Supporting sharing in science and engineering. In: Tremaine M (ed.) *GROUP '03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*. New York: ACM Press, 330–348.
- Boellstorff T (2013) Making big data, in theory. *First Monday* 18(10).
- Borgman CL (2003) The invisible library: Paradox of the global information infrastructure. *Library Trends* 51(4): 652–674.
- Borgman CL (2015) *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge: MIT Press.
- Borgman CL, Wallis JC and Mayernik MS (2012) Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work* 21(6): 485–523.
- Bovens M (1998) *The Quest for Responsibility: Accountability and Citizenship in Complex Organizations*. Cambridge: Cambridge University Press.
- Bowen GM and Roth WM (2002) The 'socialization' and enculturation of ecologists in formal and informal settings. *Electronic Journal of Science Education* 6(3).
- Bowker GC (2005) *Memory Practices in the Sciences*. Cambridge: MIT Press.
- Busch L (2016) Looking in the wrong (La)place? The promise and perils of becoming big data. *Science, Technology, & Human Values* 42(4): 657–678.
- Calvillo N (2018) Political airs: From monitoring to attuned sensing air pollution. *Social Studies of Science* 48(3): 372–388.
- Chamberlain A and Crabtree A (2016) Searching for music: Understanding the discovery, acquisition, processing and organization of music in a domestic setting for design. *Personal and Ubiquitous Computing* 20(4): 559–571.
- Clement A (2014) NSA surveillance: Exploring the geographies of internet interception. In *iConference 2014 Proceedings*, 412–425. Urbana-Champaign: University of Illinois at Urbana-Champaign.
- Collins H (2013) *Gravity's Ghost and Big Dog: Scientific Discovery and Social Analysis in the Twenty-first Century*. Chicago: University of Chicago Press.
- Collins H (2017) *Gravity's Kiss: The Detection of Gravitational Waves*. Cambridge, MA: MIT Press.
- Collins HM (1998) The meaning of data: Open and closed evidential cultures in the search for gravitational waves. *American Journal of Sociology* 104(2): 293–338.

- Day RE (2014) *Indexing It All: The Subject in the Age of Documentation, Information, and Data*. Cambridge: MIT Press.
- Drucker J (2011) Humanities approaches to graphical display. *Digital Humanities Quarterly* 5(1).
- Edwards PN (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge: MIT Press.
- Edwards PN, Mayernik MS, Batcheller A, et al. (2011) Science friction: Data, metadata, and collaboration in the interdisciplinary sciences. *Social Studies of Science* 41(5): 667–690.
- Eve MP (2016) On the political aesthetics of metadata. *Alluvium* 5(1).
- Fegraus EH, Andelman S, Jones MB, et al. (2005) Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America* 86(3): 158–168.
- Fox J (2007) The uncertain relationship between transparency and accountability. *Development in Practice* 17(4–5): 663–671.
- Furner J (2016) ‘Data’: The data. In: Kelly M and Bielby J (eds) *Information Cultures in the Digital Age: A Festschrift in Honor of Rafael Capurro*. Wiesbaden: Springer, 287–306.
- Gabrys J (2016) *Program Earth: Environmental Sensing Technology and the Making of a Computational Planet*. Minneapolis, MN: University of Minnesota Press.
- Gabrys J, Pritchard H and Barratt B (2016) Just good enough data: Figuring data citizenships through air pollution sensing and data stories. *Big Data & Society* 3(2).
- Garfinkel H (1967) *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.
- Gray J, Liu DT, Nieto-Santisteban M, et al. (2005) Scientific data management in the coming decade. *CTWatch Quarterly* 1(1).
- Greenberg J (2017) Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. *Journal of Data and Information Science* 2(3): 19–36.
- Helgesson CF (2010) From dirty data to credible scientific evidence: Some practices used to clean data in large randomised clinical trials. In: Will C and Moreira T (eds) *Medical Proofs, Social Experiments: Clinical Trials in Shifting Contexts*. London: Routledge, 49–63.
- Hoeppe G (2014) Working data together: The accountability and reflexivity of digital astronomical practice. *Social Studies of Science* 44(2): 243–270.
- Karasti H, Baker KS and Halkola E (2006) Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. *Computer Supported Cooperative Work* 15(4): 321–358.
- Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures, & Their Consequences*. Los Angeles: SAGE.
- Knobel C (2010). *Ontic Occlusion and Exposure in Sociotechnical Systems*. PhD Thesis, University of Michigan.
- Latour B (1999) *Pandora’s Hope: Essays on the Reality of Science Studies*. Cambridge, MA: Harvard University Press.
- Latour B and Woolgar S (1979) *Laboratory Life: The Construction of Scientific Facts*. Princeton: Princeton University Press.
- Leonelli S (2015) What counts as scientific data? A relational framework. *Philosophy of Science* 82(5): 810–821.
- Leonelli S (2016a) *Data-centric Biology: A Philosophical Study*. Chicago, IL: University of Chicago Press.
- Leonelli S (2016b) Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374(2083): pii: 20160122.

- Levin N (2014) What's being translated in translational research? Making and making sense of data between the laboratory and the clinic. *Tecnoscienza: Italian Journal of Science & Technology Studies* 5(1): 91–113.
- Livingston E (1987) *Making Sense of Ethnomethodology*. New York: Routledge and Kegan Paul.
- Lynch M (1993) *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge: Cambridge University Press.
- Lynch M (2013) Ontography: Investigating the production of things, deflating ontology. *Social Studies of Science* 43(3): 444–462.
- Lynch M and Sharrock W (2003) Editors' introduction. In: Lynch M and Sharrock W (eds) *Harold Garfinkel*. London: SAGE.
- McCook S (1996) 'It may be truth, but it is not evidence': Paul du Chaillu and the legitimation of evidence in the field sciences. *Osiris* 11: 177–197.
- McKenzie PJ (2003) A model of information practices in accounts of everyday-life information seeking. *Journal of Documentation* 59(1): 19–40.
- Mayernik MS (2011) *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators*. PhD Thesis, University of California, Los Angeles.
- Mayernik MS (2016) Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology* 67(4): 973–993.
- Mayernik MS (2017) Open data: Accountability and transparency. *Big Data & Society* 4(2).
- Mayernik MS and Acker A (2018) Tracing the traces: The critical role of metadata within networked communications. *Journal of the Association for Information Science and Technology* 69(1): 177–180.
- Mayernik MS, Davis L, Kelly K, et al. (2014) Research center insights into data curation education and curriculum. In: Bolikowski Ł, Casarosa V, Goodale P, et al. (eds) *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*. New York: Springer International Publishing, 239–248.
- Mayernik MS, Wallis JC and Borgman CL (2013) Unearthing the infrastructure: Humans and sensors in field-based scientific research. *Computer Supported Cooperative Work* 22(1): 65–101.
- Michener WK, Brunt JW, Helly JJ, et al. (1997) Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7(1): 330–342.
- Orr JE (1996) *Talking About Machines: An Ethnography of a Modern Dob*. Ithaca: Cornell University Press.
- Pollner M (1987) *Mundane Reason: Reality in Everyday and Sociological Discourse*. Cambridge: Cambridge University Press.
- Pontille D (2010) Updating a biomedical database: Writing, reading and invisible contribution. In: Barton D and Papen U (eds) *Anthropology of Writing: Understanding Textually-mediated Worlds*. New York: Continuum, 47–66.
- Rood RB and Edwards PN (2014) Climate informatics: Human experts and the end-to-end system. *Earthzine*. Available at: <https://earthzine.org/climate-informatics-human-experts-and-the-end-to-end-system/> (accessed 23 June 2019).
- Rosenberg D (2013) Data before the fact. In: Gitelman L (ed) *'Raw Data' Is an Oxymoron*. Cambridge: MIT Press, 15–40.
- Roth WM and Bowen GM (1999) Digitizing lizards: The topology of 'vision' in ecological fieldwork. *Social Studies of Science* 29(5): 719–764.
- Roth WM and Bowen GM (2001) 'Creative solutions' and 'fibbing results': Enculturation in field ecology. *Social Studies of Science* 31(4): 533–556.
- Saunders BF (2008) *CT Suite: The Work of Diagnosis in the Age of Noninvasive Cutting*. Durham: Duke University Press.



- Shankar K (2007) Order from chaos: The poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology* 58(10): 1457–1466.
- Shankar K (2009) Ambiguity and legitimate peripheral participation in the creation of scientific documents. *Journal of Documentation* 65(1): 151–165.
- Shapin S (1989) The invisible technician. *American Scientist* 77(6): 554–563.
- Suchman LA (1987) *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge: Cambridge University Press.
- Suchman LA (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*, 2nd edn. Cambridge: Cambridge University Press.
- Traweek S (1988) *Beamtimes and Lifetimes: The World of High Energy Physicists*. Cambridge: Harvard University Press.
- Vardigan M, Granda P and Hoelter L (2016) Documenting survey data across the life cycle. In: Wolf C, Joye D, Smith TW, et al. (eds) *The SAGE Handbook of Survey Methodology*. Los Angeles, CA: SAGE, 443–459.
- Vertesi J, Kaye J, Jarosewski SN, et al. (2016) Data narratives: Uncovering tensions in personal data management. In: *CSCW '16: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. New York: ACM Press, 478–490.
- Wallis JC and Borgman CL (2011) Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the American Society for Information Science and Technology* 48(1): 1–10.
- Wallis JC, Rolando E and Borgman CL (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE* 8(7): e67332.
- Walford A (2012) Data moves: Taking Amazonian climate science seriously. *The Cambridge Journal of Anthropology* 30(2): 101–117.
- Wenger E (1998) *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.
- Woolgar S (1981) Critique and criticism: Two readings of ethnomethodology. *Social Studies of Science* 11(4): 504–514.
- Woolgar S and Coopmans C (2006) Virtual witnessing in a virtual age: A prospectus for social studies of e-science. In: Hine C (ed) *New Infrastructures for Knowledge Production: Understanding E-Science*. Hershey: Information Science Publishing, 1–26.
- Woolgar S and Neyland D (2013) *Mundane Governance: Ontology and Accountability*. Oxford: Oxford University Press.
- Woolgar SW (1976) *Problems and possibilities of the sociological analysis of scientific accounts (or the why and the how in reading scientists' accounts)*. Paper presented at the First Annual Meeting of the Society for Social Studies of Science/ISA Conference on the Sociology of Science, Cornell University, Ithaca, USA.
- Wynholds LA, Wallis JC, Borgman CL, et al. (2012) Data, data use, and scientific inquiry. In: *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries – JCDL '12*. New York: ACM Press, 19–22.
- Yakel E (2001) The social construction of accountability: Radiologists and their recordkeeping practices. *Information Society* 17(4): 233–245.
- Zimmerman AS (2007) Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal of Digital Libraries* 7(1/2): 5–16.
- Zimmerman AS (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): 631–652.

**Author biography**

Matthew S Mayernik is a Project Scientist and Research Data Services Specialist in the NCAR Library. In this role, he leads research projects and operational services related to research data curation. His research interests include metadata practices and standards, data citation and identity, and social and institutional aspects of research data. He received his PhD from the University of California, Los Angeles Department of Information Studies.