The Application of Archival Concepts to a Data-Intensive Environment: Working with Scientists to Understand Data Management and Preservation Needs

Dharma Akmon, Ann Zimmerman, Morgan Daniels, & Margaret Hedstrom

[Final publication is available at http://www.springerlink.com/content/m1n06056870u68u2/]

Dharma Akmon
School of Information
University of Michigan
1075 Beal Avenue
Ann Arbor, Michigan 48109-2112

dharmrae@umich.edu
734-395-0790

Introduction

We live in the age of digital data-- a time in which our ability to create data seems to greatly outpace our capacity to manage and make sense of them. A recent report claims that 1250 billion gigabytes of data were generated in 2010 alone and estimates that we can expect 58 percent growth per year (Gantz & Reinsel 2010). The resulting "data deluge" is a growing concern in many sectors, including government, business, and science, as people and organizations struggle to determine who should be responsible for preserving data, which data should be preserved, and how to protect privacy, provide access to data, and ensure they are described in such a way that they can be used by others[1]. To put it simply, many of our existing tools and methods of information access, storage, and preservation are inadequate for the needs and scale of digital data (Berman 2008).

The data deluge is receiving considerable attention in science. For scientists, data curation comprises a necessary component of the emerging infrastructure for what has been coined "data-intensive science" or "e-science." The conduct of e-science depends on massive repositories of shared scientific data, new data analysis and visualization tools, and open access to scientific publications (Hey, Tansley, & Tolle 2009). Federal funding agencies, scholarly journal publishers, and professional societies increasingly encourage or require scientists to share their data openly and to make provisions for the data's long-term storage (e.g., U.S. National Science Foundation 2010; U.S. National Institutes of Health 2003). As a result, scientists can no longer assume that the data they generate in the course of their research are valuable only to them or their close colleagues or that sharing data through scholarly publication is sufficient. Instead, they must anticipate future uses of data and preserve and describe them to facilitate this use. Scientists, in other words, are increasingly tasked with carrying out archival work, even though they often lack the knowledge, skills, interest, and time to do so.

The collection, organization, and long-term preservation of resources, including those in digital form, are the raisons d'être of archives and archivists. The archival community, however, has largely neglected data. Scientific data were, at least until recently, seen as the responsibility of the scientists who generated them. If they were to be preserved long-term, archivists generally assumed that "the scientists themselves would see to it that they [were] preserved" (Warnow-Blewett, Genuth, & Weart 2001, p. 90). Scientists, on the other hand, increasingly recognize that they lack the skills and expertise needed to meet the demands being placed on them with regard to data curation and are seeking the help of "data archivists" and "data curators" (Curry 2011; Feijen 2011). Some scientists mention archivists explicitly when describing the expertise they need:

---

[1] The data deluge is a topic of many articles in the press and popular scientific publications. Recent examples include *Nature's* special issue, "Data Sharing" (Sept. 2009, Vol. 461(145)); *The Economist's* special report, "Data, Data Everywhere" (Feb. 2010); and *Science Magazine's* special issue, "Dealing with Data" (Feb. 2011, Vol 331(6018)).

> …data sharing involves more than putting the data on a Web site. Scientists and editors of scholarly journals are not professional archivists, and many homegrown one-off solutions do not last long. Data formats have been changing so fast that archiving standards require special preservation formatting, using internationally agreed-upon metadata protocols and appropriate data citation standards (King 2011, p. 720).

We believe that the pressing needs of science in the area of data curation represent a significant opportunity for both archivists and archival scholars to contribute to a key challenge of our time. Admittedly, there are many obstacles to managing, preserving, and providing access to scientific data. It has been difficult, for example, to compel scientists to share their data given that they are rewarded primarily for publishing manuscripts (Borgman, Wallis, & Enyedy 2007). The amount of effort required to describe data for reuse has also limited scientists' willingness to share data (Birnholtz & Bietz 2003; Van House 2003). Additionally, many scientific fields lack a common integrated data infrastructure, which often results in non-standardized, local data management practices (Borgman, Wallis, & Enyedy 2007). This problem is especially prevalent in "small-science" fields. Small-science fields typically require minimal management structure, and research is accomplished by a single investigator or small teams of researchers. Further, "progress and reward are contingent on [individual scientists], generating and analysing [their] own data" (Cragin, Palmer, Carlson, & Witt 2010, p. 4024).

Archivists and archival scholars can offer a valuable perspective to these and other data curation issues, particularly in small-science contexts where "data management systems tend to be ad hoc" (Cragin, Palmer, Carlson, & Witt 2010, p. 4024). Experience with selection, concern with preserving context in order to maintain meaning, and the recognition that preservation practices for digital materials must begin early in the records life cycle would be of particular value to science data curation. However, to begin to move closer to capitalizing on the new opportunities presented by data and understand the role that archivists might play in data curation, archivists must gain insight into data management from the scientists' perspective. This paper represents an effort to provide such insight.

In this paper we present findings from a case study of a group of materials scientists working in a university research lab. Our primary goal in this research was to understand the data that were generated, the practices that the scientists employed to manage their data, and the challenges they faced. We were especially interested in discovering how scientists managed data for their own anticipated needs and how these practices influenced the use of the data over time. Related to this, we sought to identify differences and commonalities in individuals' data practices as well as their perceptions about data management. Lastly we wanted to understand the relationship between data and associated forms of documentation.

We found that most of the concern over data management came from the lab head, but that all of the scientists experienced difficulty at some point in using another scientists', or even their own, data. Even with knowledge about what data were available in the lab's shared computer, the scientists were challenged to make meaningful use of those data without face-to-face contact with data creators. The lab head was eager to make data more amenable to reuse but was also reluctant to impose any particular system because she did not want to monitor adherence to a set of rules; she thought that data management practices appropriately reflected the needs and work styles of individual scientists; and she felt she lacked the expertise to tell the scientists how to manage data. Without guidance on data management practices, scientists devised, by trial and error, particular data management and documentation practices that fit their specific needs and research questions. The resulting multiple systems served individuals well but hindered data reuse between lab members.

Before describing the specific lab we studied and detailing our findings, we provide a brief background on archival involvement with science records and some of the more recent interest by scholars in science data, and then follow that with a description of our methodology.

Background

Archival interest in scientific records has focused primarily on preserving the "historical documentation" of science (Warnow-Blewett, Capitos, Genuth & Weart 1995, p. 9). The most extensive research on archival management of the records of science started in the 1960s and continued into the 1990s under the auspices of the Center for the History of Physics at the American Institute of Physics (AIP) (e.g. King 1964; Hackman & Warnow-Blewett 1987; Warnow-Blewett, Capitos, Genuth, & Weart 1995; Warnow-Blewett, Genuth, & Weart 2001). The series of studies conducted by the AIP were motivated primarily by changes in the scale and methods of science. AIP archivists, historians, and scientists analyzed the records and recordkeeping practices of individual scientists and multi-institutional collaborations in a range of scientific fields and in a variety of organizational contexts, including academic, corporate, and government research laboratories. The main focus of the AIP research and similar documentation analyses in other scientific fields (e.g. Elliot 1983; Haas, Samuels, & Simmons 1985) was to understand the process of scientific research, from the original conception of a research problem to the publication of the final results, in order to identify records of important projects and key individuals that contributed to advances in science and technology. The practical goal of this research was to help archivists select records that would meet the needs of administrators as well as future historians and other scholars.

Preserving scientific *data* has been a secondary concern for archivists until recently, for several key reasons. First, archivists have viewed data as "not useful for historical research" and therefore have placed it outside the bounds of their professional concerns (Warnow-Blewett, Genuth & Weart 2001, p. 90). If data were anticipated to

have value in the future, archivists generally assumed that scientists were in the best position to understand their own needs and make judgments regarding how long data should be retained (Warnow-Blewett, Genuth & Weart 2001, p. 90). Archivists (and scientists) also frequently made a distinction between observational and experimental data, pointing out that observational data might have long-term value if they were irreplaceable or very costly to collect, but that scientists were much more likely to generate or collect new experimental data rather than to reuse "old" data (Elliot 1974; Haas, Samuels, & Simmons 1985; Warnow-Blewett, Genuth & Weart 2001). In cases where data might have long-term value, archivists favored preserving the data in discipline-specific data repositories based in government scientific agencies or large labs: places that many archivists presumed they did not have a role (Elliot 1974; Haas, Samuels, & Simmons 1985; Warnow-Blewett, Genuth & Weart 2001). These assumptions formed the basis for archival guidance on scientific data through the 1990s and positioned issues related to scientific data outside the purview of most archivists' activities and concerns.

More recently, in response to widespread interest in data curation in science, some archivists and archival researchers have begun to pay more attention to scientific data. Shankar (2007), in a notable departure from previous archival studies, conducted an ethnographic study of the recordkeeping practices of scientists in an academic animal neuroscience lab. She examined the connection between science data and the associated records produced with the data in order to understand how "raw" data are transformed "into a reliable trace of scientific activity" (p. 1461). However, since the lab she studied relied mostly on paper documents, her findings may have limited usefulness in understanding the data practices of scientists working primarily with digital data. Additionally, it was outside the scope of Shankar's study to examine the impact that the scientists' records management practices might have on the potential of the data to support new scientific inquiry.

A paper by Lauriault, Craig, Taylor, & Pulsifer (2007), on the other hand, emphasized the value of preserving scientific data to help identify trends and serve as inputs in models and simulations. Lauriault et al. asserted that archivists "need to play a key role" in preserving scientific data, but that to do so they "must understand the scientific context" (p. 127). Additionally, they argued that selection of data for preservation should be based, not on business activities or corporate memory needs (as traditional archival practice has been), but on the "needs of the research community," which, for science data, is comprised of scientists, not historians (p. 137). This reorientation to scientists' needs means that archivists must be more familiar with how scientists work with data and what they are trying to achieve with their data management practices.

Much of the scholarly work on data curation has thus far taken place outside of the archival science literature and has focused on topics such as the degree to which scientists adhere to journals' data publication policies (Piwowar & Chapman 2009); the challenges of reusing another scientist's data (Zimmerman 2008); and the difficulties of creating shareable scientific data (Borgman, Wallis, & Enyedy 2007). While this work has revealed

significant obstacles to creating data that can be used over time and has also highlighted the importance of early involvement of archivists and data curators in the life cycle of science data (Wallis et al. 2008), we still know little about how scientists manage, analyze, and otherwise work with their own data. Further, we do not know much about the key data management characteristics of these data intensive environments and how they fit within the records managment landscape of the lab. The scant amount of research on the subject in the archival literature indicates that archivists might be ill-prepared to assist scientists with data curation, even as scientists increasingly seek out the help of data management and preservation specialists (Curry 2011; Feijen 2011; King 2011).

In the interest of beginning to fill these gaps and further elucidate the scientific context for archivists, we studied a group of materials scientists who are trying to cope with data management and curation on their own. This study was motivated by our larger research agenda in which we are studying data practices in multiple domains as well by more practical concerns. In regard to the latter, we were approached by the head of the laboratory, who knew us as experts in archiving practices, to help her improve management of the data in her lab. We used this as an opportunity to investigate data practices in the lab with the end goals of sharing our analysis with archival researchers and making some initial recommendations for improved data management in the lab.

Methodology

The challenges of data documentation and management call for a research approach that captures and explores scientists' data practices (by which we mean data use and management activities) in all of their complexity. Because of its effectiveness in studying phenomena in depth, we utilized a case study design (Yin 2008) to examine the data practices of one materials science lab group at a large U.S. research university. Materials science is a field of engineering and applied science that examines the properties and characteristics of matter. The group that we studied carried out research on semiconductor materials, which are commonly used in electronic devices. Case studies are bounded by time and employ a variety of data collection procedures. In this case, over a four-month period in summer 2009, we conducted one-on-one interviews with five of the graduate students and the faculty member who directs the lab (Table 1). We asked individuals about their work, the data they created, and the kinds of things they did with the data after collecting them. We observed four of the participants at work in the lab and examined the data they produced.

Table 1: Study Participant Details

Name[2]	Position	Research Area	Year of Study
Prof. Alexandra Bennett	Lab Head	The effect of the surface structure of semiconductor materials on the growth of materials	n/a
Susan	Doctoral Student	The effect of modifications of semiconductor surfaces on the photoluminescence of materials	6th
Keith	Doctoral Student	Characterization of multiple, coexisting surface reconstructions on indium arsenide surfaces	6th
Rachel	Doctoral Student	The effect of strain on surface structures of semiconducting films	5th
Thomas	Doctoral Student	Methods of eliminating stress on surfaces comprised of different crystal structures	2nd
Bill	Doctoral Student	The effect of bismuth deposition on the surface shape and structure of gallium arsenide	1st

With the consent of the participants, we recorded and later transcribed the interviews. We identified themes in the interview data and used NVivo, a qualitative data management and analysis tool, to assign codes based on those themes. As a follow-up to the interviews, we scheduled one- to two-hour focused observations with four of the graduate students during which each of them guided us through some part of his or her data collection, management, or retrieval processes. We documented these sessions in field notes and took photographs of relevant visual and textual material. While lab members told us much about their data practices in the interviews, the observations gave us a deeper understanding of what they did and why.

The remainder of the paper presents our findings from this study and discusses implications for scientific data management. In the next section, we describe the scientific work of the laboratory since this is fundamentally important to understanding the scientists' data practices. Next, we outline the data challenges the lab faced from the perspective of the scientists, in particular Professor Bennett. Following our observation of the challenges in this lab, we analyze the data practices of the materials scientists and show what they were trying to accomplish with these practices. Lastly, in the Discussion section, we highlight important characteristics of the science data management landscape for archivists to consider as they work on data curation issues.

Work in the Bennett Lab

The Bennett Lab is led by Professor Alexandra Bennett, a tenured professor who, at the time of our study, had been a materials science faculty member at the university for over a decade. Professor Bennett formed the lab group to study the surface characteristics

[2] Pseudonyms are used to protect participants' identities.

and structures of different “III-V” semiconductors[3]. The results of research on the surface structures of these materials provide insight that might be used by industry developers to leverage or alter the properties of certain types of semiconductors to build better electronics devices, such as lasers and light-emitting diodes (LEDs).

The lab group consisted of nine members: Professor Bennett, seven doctoral students, and a faculty member from a nearby satellite campus of the university. This lab was typical of many "small-science" endeavors in that each lab member worked on research that was related to a common area, but had his/her own specific area of focus and spent little to no time collaborating with other lab members. The doctoral students conducted their own experiments and analyzed their own data, the output of which was intended to lead to scholarly publications, often with Professor Bennett as a co-author, and a completed dissertation.

A significant capital investment is required to build a materials science laboratory, and, as a result, the research of a lab is partially characterized by the equipment utilized in that specific lab. Professor Bennett's lab included a Molecular Beam Epitaxy (MBE) chamber, which the scientists used to deposit very thin layers (hundreds of nanometers thick) of single crystals of semiconductor materials onto a substrate material. The scientists referred to the process of depositing these layers as “growing a sample,” and when they talked about carrying out experiments this was usually the first step.

A growing session typically took a scientist several hours to an entire day to complete. S/he would carry out some analysis during this stage, primarily in order to ensure that growth was proceeding successfully. Subsequent steps varied depending on the research goals of the individual. For some lab members the next step was to alter the surface of their sample. Susan, for example, used a Focused Ion Beam (FIB) to make cuts in the surface of many of her samples since she was interested in looking at how this modification affected the photoluminescence of her material. Whether modification of samples was part of a lab member's experiments or not, each carried out different characterization techniques (observations and measurements of the materials) on his/her samples, which mainly resulted in image data (Figure 1). The scientist would then analyze these data and frequently transform them into other types of data, such as graphs, to facilitate the identification of relationships between variables.

[3] “III-V” (pronounced “three five”) refers to periodic grouping in the periodic table of elements. III-V semiconductors, comprised of a group III element and a group V element, are commonly used in optical electronic devices such as lasers. Examples of III-V semiconductors include gallium arsenide, gallium antimonide, and indium arsenide.

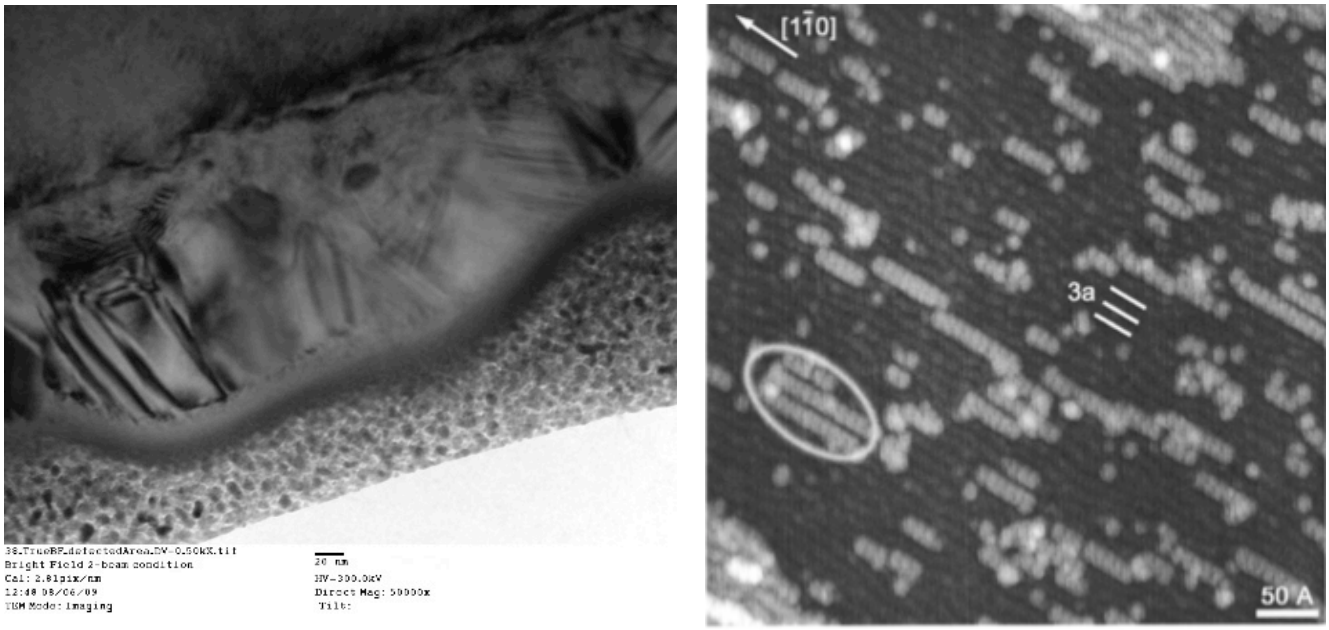

Figure 1: (L) Transmission Electron Microscope (TEM) Image of a Sample; (R) Scanning Tunneling Microscope (STM) Image of a Sample

Data Management Challenges

In talking to lab members, it was apparent that most of the concern with the state of the lab's data management practices rested with Professor Bennett. This is unsurprising when we consider that much of her body of research depended on the work of the other members of the lab, who were a rotating cast of doctoral students and post-doctoral researchers. The continuity of Professor Bennett's research was jeopardized when she was not able to use data from former lab members, while the lab members' research primarily depended on collecting and analyzing their own data. As a result, we describe most (but not all) of the data management issues in this lab from the perspective of Professor Bennett.

Professor Bennett voiced frustration with the difficulty of accessing and using present and past lab members' data. She characterized her use of current lab members' data as "ad-hoc" and reliant on personal interaction:

> When we're writing a paper maybe I'll have the data, maybe I won't have the data---it depends. And that's another reason why I'm talking to you guys---because I *want* to have access to the data more readily. The only way for me to do it is, you know, it's [through] my e-mail and it's ad hoc...I don't like it at all.

As the only continuous member of the lab, Professor Bennett often recognized (based on her own memory) connections between data being generated by her current lab members and data generated from previous members:

> I'll recollect and I'll say, "Gosh, I remember when Heather, eight years ago, did this set of experiments, and we saw this really weird effect. I wonder. And now we're seeing a similar effect that in my recollection is similar. I'd like to go back and compare this data and see."

Using past lab members' data, however was often impossible or, at the very least, extremely difficult. Professor Bennett partially attributed this situation to suboptimal data organization:

> It scares me how much data was lost, because it wasn't well organized. He [a previous member of the lab] had amazing data, he was very smart, and he wrote just a fantastic thesis. And there was stuff in there that we probably could've pulled out another thesis just on his data alone. I have no idea how to even access it. It's just pffft!

We also learned that the illegibility of paper records and the idiosyncratic methods lab members used to document the experiments sometimes hindered data reuse. One doctoral student, Keith, told us that a previous lab member's data would have been useful to him because they were studying similar phenomena. While the data were technically available to him, he was stymied to decipher the documentation well enough to rely on those data for his own research:

> Given that it's so hard to read his handwriting and things like that, it's almost impossible for me to look at his raw data and go back and try to...I'm kind of limited to what he's published and stuff. Because if not, I don't know what growth conditions and all that other stuff is.

This demonstrates that data reuse in the lab was sometimes made impossible by something quite simple. As Professor Bennett told us, even data that were relatively well-organized and described could prove difficult to use in a meaningful way:

> **Prof. Bennett:** When a student leaves, you know, I've had this happen time and time again. There was one post-doc who left. And, you know, he made a very conscious effort to document his data and to catalog it in a way that was just easy to find, easy to follow, and everything. And it was still incredibly difficult, you know, even under the best intended circumstances.
>
> **I:** What was hard about the guy that did it really well? What made it still difficult? What were the hurdles?
>
> **PB:** Well, because, you know, the way that we label the data…I mean it's all files, filenames. So we have individual tiffs, and in order to figure out, "Okay so what I want is; I want a sample that was grown at 500°C and these, you know, I want these specific conditions, and I want this particular pattern. Okay let's go through the binder, here it is. Let's see, do we have data on

that?" You know, it's just complicated because we don't have a way of tagging all the data all at once so that we can just search it.

I: Right, so there's no good search system? So even if the information is all there it's hard to retrieve it?

PB: Right, exactly.

It was clear that Professor Bennett wanted better access to her lab's data than she had. Despite her frustration, however, she was reluctant to impose data management rules on lab members. She expected lab members to save copies of their data on one of the lab computers, but data management and documentation were primarily left up to each individual scientist to carry out in the way s/he saw fit. One reason for Professor Bennett's hesitation was that she was not trained in data management and therefore did not feel that she was a good source of expertise on best practices:

> I don't really feel like I have the skills or I even know how to tell them "This is how I want you to do it." Nor do I have the patience to be, "Are you doing it this way? You're not doing it this way." You know, so policing that I find that to be very-- I can't do that either. So the way that I've done it is, well I haven't really…As far as organizing their data, I don't have anything that I really tell them they have to do.

Further, she thought that data management practices reflected each scientist's unique needs and work style:

> Everyone works differently. It'd be kind of like me insisting that every one of my colleagues arrange their books in the system that I designate as the proper system. You know, and there are a lot of different ways, and everyone's brain works differently so it feels like--it feels presumptive on my part to do that.

Professor Bennett also believed that there was a prevalent assumption that scientists, as people with technical skills, were naturally skilled at data management, even when they received no training:

> It's just assumed that, "Well, what you mean you want to be trained in how to take data? You're a scientist, that's what you do." But, you know, doing the science is very different than managing data. And I don't know that people have really appreciated that until fairly recently.

We were struck by the tension between Professor Bennett's statements. While she felt better data management practices could make her work easier and facilitate new research, she did not know what improved data practices would look like. Additionally, she was reluctant to introduce a system that required her to micromanage the members of her lab. She appreciated that they were unlikely to adopt a particular practice unless it was part of their workflow, but she wondered if the "right" local infrastructure would provide the necessary "scaffolding" without "extra steps."

With little training in data management and few explicit rules to guide them, what data practices *did* the scientists engage in and to what end? The next section analyzes the data management and documentation landscape in the Bennett Lab to reveal how scientists managed data for their own anticipated needs, what practices were shared among lab members, and how data were connected to other forms of documentation in the lab.

Data Practices in the Bennett Lab

If there were problems of data reuse in the Bennett Lab, it certainly was not due to scientists' disregard for the importance of data management and documentation. In fact, we found that the scientists engaged in a complex set of practices to make sure that they could easily locate their own data when they needed them and that they could understand important contextual information about that data weeks, months, and sometimes even years later. The scientists' concerns about accessing their own data, in other words, were very similar to Professor Bennett's concerns about accessing all the lab members' data. In the following sections, we look at two main areas of these scientists' work where data management played a large role: documenting experiments and characterizing samples.

Documenting Experiments

Documentation began for the scientists in the early stages of their experiments. During a growing session, the scientists recorded information in a shared paper lab notebook that was located on a desk close to the MBE chamber (Figure 2). A lab member typically recorded a new entry for each new sample s/he generated. Each sample was named according to a lab-wide naming convention that conveyed important information about the sample and, hence, its associated data. The alphanumeric string contained a letter that designated a group of samples grown after a particular "vent of the chamber" (when the chamber is opened to replenish semiconducting materials used in growing). Knowing that a sample was grown immediately following a chamber vent could explain possible discrepancies in the samples and the data derived from them.

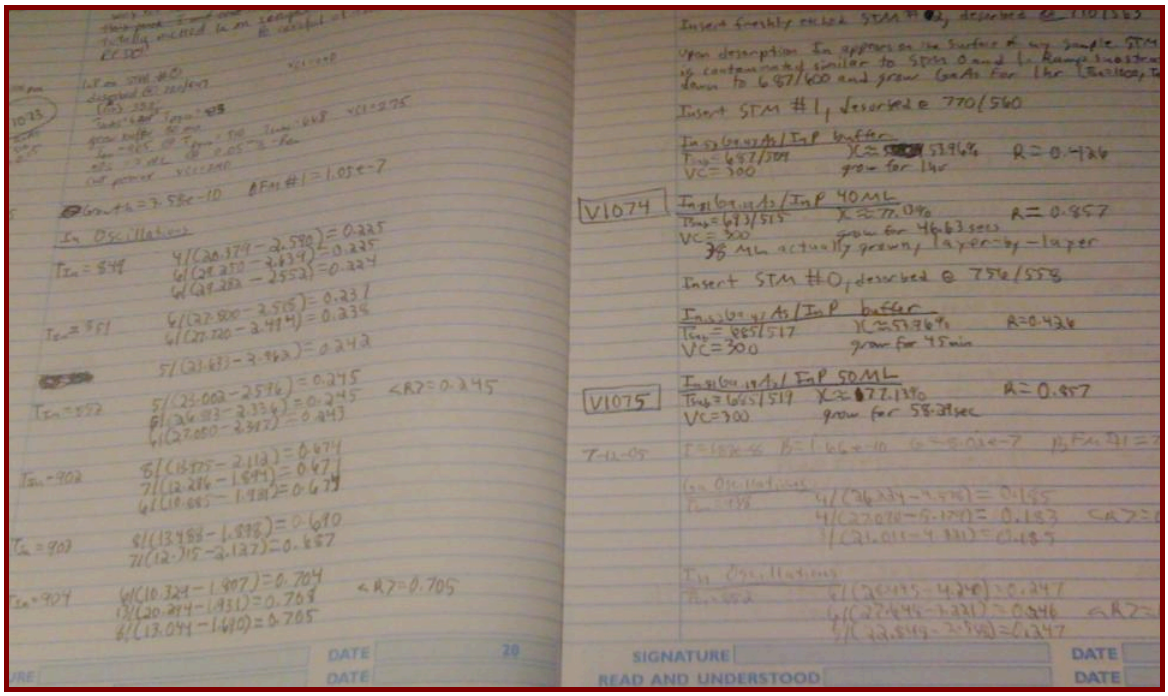

Figure 2: Shared Lab Notebook. Information for samples V1074 and V1075 appear on the right.

Under the entry in the notebook for a given sample, a scientist would record what s/he did to create that particular sample. Information such as the temperature and growth rates was expected in scholarly publications, but was also potentially valuable to anyone in the lab since they all shared the MBE chamber. As Keith reported,

> You write down all of your calibrations at different temperatures. That's very helpful for other people because whenever they come in it gives them a gauge. If you know you need to grow gallium at a half a monolayer a second, and you can look back and see that I did calibrations with gallium, and I did it at .6 and .7 and .3 and .4, it gives you kind of a temperature range between and .4 and .6. I'll probably want to be in between there to start off with.

Several of the scientists emphasized the helpfulness of the other scientists' experiment notes to calibrate the chamber and start their own experiments. As a tool for sharing a very specific kind of information relevant to the operation of an important, shared piece of experimental equipment (the MBE chamber), the paper lab notebook was at least a partial success. However, beyond calibration information, there was significant variation in what scientists recorded, and this variation limited data reuse by others.

For example, all the scientists recorded certain types of information (e.g. pressures in the chamber, cell calibrations, and growth rates), but beyond that, it was "kind of a crapshoot." The scientists we interviewed said that they recorded whatever information they thought they needed and often used phrases like "as much information as possible," "everything," or "all the conditions" to describe what that included. However, it was clear that what was needed or deemed relevant or important varied among the scientists. Keith, for example, needed to know the III-V ratio (the ratio of his group III element to his group

V element) of his samples to address his research questions, but noted that his lab mates did not normally record this information.

While the lab notebook served the scientists well in documenting their experiments, several expressed frustration at their inability to pull up information on samples based on the parameters of the experiments. The scientists frequently wanted to be able to retrieve information about all samples created by a specific lab member or made up of a certain composition, however they could only flip through the paper notebooks, manually looking for entries that matched their needs.

Several years prior to our study, Professor Bennett implemented a database that scientists could use, in addition to the lab notebook, for recording experimental information. The intent of the database was to collect standardized information about data gathered in different studies to make it easy to search for data across individual projects. Unfortunately, the database was not widely adopted. Some of the scientists told us they never used the database, while others reported that they used it at one time, but had since fallen behind on entering information from their experiments. The extra work of copying essentially the same information in two places (the lab notebook and the database) limited the use of the database. A second barrier arose from the fact that the database was limited to a single computer that was not connected to the network, making it accessible only to those physically in the lab. Populating the database would lengthen lab members' already long days at work without providing the benefit of remote access to their experimental information.

Documenting Data

The lab notebook helped the scientists keep track of information about their experiments, but the scientists also needed to record information about the data they created as they characterized the properties of their samples. Additionally, they needed to capture information about how data related to other data. The scientists characterized (or analyzed; they frequently used the terms interchangeably) their samples by using any of a number of high-powered microscopes and specialized measurement tools (Table 2). Some of these tools were connected to the MBE chamber, while others were located in shared campus labs where the scientists had to reserve equipment time. Their data, all of it digital, came primarily in the form of images, but also as numbers and graphs.

Table 2: Common Characterization Methods Employed by the Bennett Lab

Characterization Method	Data Type	Description
AFM (Atomic Force Microscopy)	Image	AFM is a very high-resolution type of microscopy used for creating topographic images of surfaces of materials at the nanoscale. In these types of images, the scientists can see the arrangement of atoms.
FIB (Focused Ion Beam)	Image	With FIB scientists can slice away a section of materials and examine it. This analysis method is inherently destructive to the sample.
MOSS (Multi-beam Optical Stress Sensor)	Numeric	MOSS is used to measure stress in thin films.
RHEED (Reflection High-Energy Electron Diffraction)	Image	RHEED is a technique used to characterize the surface of crystalline materials. The lab members use this technique primarily to monitor the growth of their thin films.
	Graphical	RHEED oscillation graphs depict growth rates, layer thickness, and composition.
SEM (Scanning Electron Microscope)	Image	SEM scans the surface of images. This gives scientists information on a sample's surface topography, composition, and other properties such as electrical conductivity.
SPIP (Scanning Probe Image Processor)	Numeric	SPIP allows scientists to quantitatively analyze images (that they get from microscopy techniques). It outputs numbers that measure things like grain analysis, telling scientists what percentage of the surface was one structure versus another.
STM (Scanning Tunneling Microscope)	Image	STM is used to image surfaces at the atomic level. It shows a three-dimensional image of a sample.
TEM (Transmission Electron Microscopy)	Image	TEM creates images of thin materials as a result of electrons passing through a specimen.

In the interviews, lab members emphasized the importance of understanding how and why they captured data long after they collected them. As Thomas told us, "I took the initial images for the sample back in late September, and I still have to know what it was I was attempting to do and how I took it back in September." For all of the lab members we interviewed, the primary location for recording information about data was in the filenames for the data (Figure 3), though the scientists also documented their data in personal spreadsheets, presentation documents, and personal notebooks. Using filenames to document data not only allowed the scientists to understand how and why they captured the data, but also let them determine, by glancing at a list of files, which data were relevant to their present purpose.

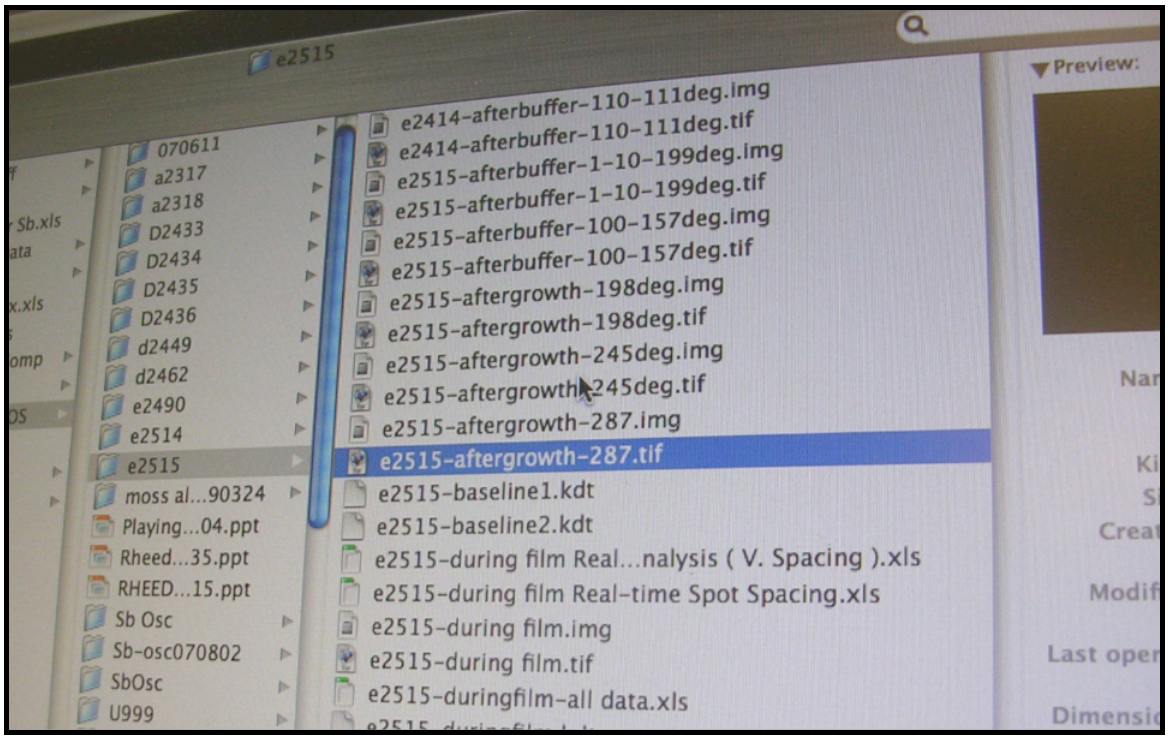

Figure 3: A List of Data Files for Sample e2515

Another important way in which each scientist facilitated his/her own understanding of his/her data over time was by naming and organizing that data in such a way that related data were connected together. This was accomplished in large part by specifying which sample data were derived from, but the scientists also grouped related data together in electronic files and spreadsheets.

All the lab members always recorded the sample number from which data were derived in the filenames for data. However, beyond the sample number, there was significant variation in what lab members captured. What follows is a more detailed description of how scientists documented their data so that they could ensure their own continuing access to them. We found that scientists devised their own methods for data documentation and that these were based, at least partially, on their experience, particular research questions, and research methods. The following examples illustrate these differences.

Bill, a first-year student, told us he used his filenames to document variables such as substrate temperature, fluxes of elements, rates of deposition, and what he was "trying to prove" or what kind of experiment the data were for. In Bill's filename for an STM image (Table 3), the following pieces of information appear from left to right: image and sample number; image size (in this case the image is 500 angstoms); the X and Y positions on the sample; the direction of the scan (in this case along the y axis); scan rate; tunneling voltage and current; and "other special conditions," which in this example was the fact that there were 1024 scan lines instead of the usual 512.

Susan, who was in her sixth year of graduate work, aimed to keep file names "somewhat short and just keep the most relevant variables to the filename." She reported that this frequently meant recording "increased dwell time" and "change in temperature." In the example filename for an AFM image Susan recorded the following information: the sample name; the thickness (in monolayers) of indium arsenide; "PL", which meant that the sample was ready for a photoluminescence study; and the thickness of gallium arsenide in the sample.

Thomas, a second year student, described his own filenames as being "long and complicated." In the example, which is for a TEM image, he recorded the following: the date of the imaging; the instrument number; the sample identifier (FIB038), which in this case represented his own sample numbering system, not the group numbering, because he had not grown anything in the chamber; next, the section number of the sample that he imaged; the imaging session number; the image number for the imaging session; the imaging mode (DF indicates that he was using dark-field); the diffracted beams he used; the condition of the objective lens of the microscope; "some extra data" to indicate what part of the sample he was looking at; and the magnification level of the image.

Table 3: Example Data Filenames

Bill	61_F2560_500A_X2463_Y2.09_yscan_150nm-s -2.70V 0.1nA_1024_scan lines.sm4
Susan	e2509_1.2inpl_500nmgaas.tif
Thomas	20090716_3011_FIB038_B-3_TEM04/66_DF_gAnd2g_D+0_mesa1_30kX.tif

From just these three filenames we can see considerable variation in what is important or relevant contextual information for data. Some of the variation is due to the use of different types of analysis tools used to characterize samples. However, the scientists' descriptions of their own practices reveal that their specific research questions and experience also played a large role in the ways they chose to describe their data.

Most of the scientists we interviewed talked about a process of trial and error whereby they learned what information was important for them to document so that they could effectively use their data. In fact, two told us of data collected early in their careers that they could not use because they failed to record what turned out to be important information about the data. This experiential component was also evident when we compared doctoral students at various stages in their careers. The research direction of the two newest members, Bill and Thomas, was not as solidified as those who were nearing graduation. While they had specific research areas, they were still trying to narrow their focus and determine which aspects of the research area were most promising. Both said they were unsure exactly what information they would need later. In an effort to ensure that they would have the necessary data documentation later, Bill and Thomas recorded extensive information in their filenames (relative to their lab mates). Contrast this with Keith, who said, "I'm probably one of the ones who-- I guess after being here six years I just write down whatever I need to and just go for it." Likewise, Susan, who was also in her sixth year, described a minimalist approach to her file naming and tended to use shorter filenames.

Experience and learning were important factors in these scientists' data documentation practices, but we were also struck by the importance of scientists' specific research questions in their resulting data documentation. The scientists we studied were always ultimately interested in the correlations between different variables. Examples include how surface reconstruction varied with changes in temperature and surface thickness; how a particular method of altering the surface affected the photoluminescence of the material; and how cutting different shapes in the materials altered the strain or stress seen in a thin film. While there are likely many more variables related to the data they created, the scientists were interested in the variables that addressed *their* questions. Through practice and trial and error they learned what those were in addition to learning which associated information was needed to present their results. The freedom they had to decide what to document allowed them to tailor their practices to fit their own needs. This freedom also explains why it could be difficult to use others' data even in the same lab—if the scientists did not document the particular variables of value to a given research question, the data had limited use.

Discussion

These scientists engaged in extensive data management and documentation activities, but reuse of data between scientists was still difficult. This problem was particularly acute for the lab head, who felt that her inability to easily access the data of previous lab members affected her research and the kinds of questions her students could explore. The scientists shared common motivations (to understand data later; to have ready access to data) for their data practices, but what they captured and how they did so varied considerably.

This variation was due to two intersecting factors. First, the lab head, while eager to improve data management in her lab, was reluctant to "impose" a system on the group and had never given students guidance on data management. There were several reasons for this reluctance: namely that she felt unqualified as a data management expert, was not interested in policing lab members' data management practices, and thought that data practices needed to fit each scientists' particular research questions and style of work. The last point is a valid one, for we found that the scientists needed to document specific contextual information to study their particular research questions. The challenge of data documentation has been noted elsewhere, particularly as it relates to predicting future needs (e.g. Bowker 2006). The scientists we studied clearly valued data documentation for their own needs, but it is unclear how much we can expect them to document and manage their data for others' potential needs. How can we define those needs given that they are so dependent on research questions and approaches? More importantly, would extensive documentation yield a benefit that outweighs what would be a considerable investment of time? These are questions that remain to be answered, but we think that the receptiveness of the lab head and her students to our insight indicates a possible opening for archivists and archival scholars to sit down with scientists and begin to answer these questions.

Our study also reveals several important characteristics of the data intensive science context that archivists and archival scholars will need to be aware of as they become more involved with data curation. First, we found that, without exception, these scientists took care to devise systems for finding and using their own data into the future. These systems were, like those in many "networked organizations" (Botticelli 2000), idiosynchratic, but they were not necessarily ad-hoc. That is to say that the scientists created personal systems of documentation and management and, once established, they used those systems fairly consistently. This indicates the high value of data management and documentation to scientific work and also serves to signal that any alterations to those systems will have signficant effects on the work that is done in the lab.

Another important characteristic of the scientific context is the degree to which the connection betweeen data and associated records influences data's potential for reuse. Archivists have long emphasized preserving original order to maintain the meaning of records. For traditional archival collections, this maintenance of meaning serves to facilitate scholars' ability to understand how records were used to carry out work. Maintaining a conceptual connection between science data and their documentation, however, is a prerequisite to ensuring that data can be used as inputs to answer new scientific questions. This emphasizes the importance of keeping what archivists may consider to be more typical science records and data together, in one repository. We can no longer assume that textual science records with potential to support historical research belong in a traditional archival repostory, while data belong in a different repository. In the lab we studied, scientists would not have been able to make sense out of their data without access to associated documentation in the lab notebook.

Archivists and archival scholars may wonder how much domain knowledge is required to understand data management in labs and how they will be received by scientists. We found that, while it was necessary to learn a good deal of technical terminology, to know the types of questions the scientists were interested in, and to become familiar with the scientists' methods for conducting research, the requisite domain knowledge was not beyond what we could gather in a brief discipline overview by the lab head and our interviews with the scientists. Additionally, we were pleased to discover that the scientists were not only open and receptive to our study, but were also interested in what we learned about their practices and were eager to hear our suggestions. This, along with reports of scientists' frustration with data management, suggests that the expertise of non-scientist data curators or data archivists may indeed be welcome in the lab.

Conclusion

Our case study represents an effort to understand data management as practiced in a small-science lab and has revealed important characteristics of a data intensive environment. Our observations of the structure of work in this lab are consistent with other studies of small-science lab work (e.g. Borgman, Wallis, & Enyedy 2007; Shankar 2007; Cragin, Palmer, Carlson, & Witt 2010). Additionally, the type of data that Bennett

Lab scientists created are fairly typical of the kind of data generated in other materials science labs (Madnick, Smith, & Clopeck 2009). Further research that examines data practices in other labs would be valuable for deepening knowledge about how data management issues vary across different kinds of contexts and the reasons for those variations.

Data curation presents exciting opportunities for archivists and archival scholars to make an impact on the practice of science. Indeed, scientists increasingly seek out archival perspectives and expertise as they struggle to meet the challenges of the data deluge. In working with and studying data curation issues, archivists will likely find that some of their assumptions built on experience with more bureaucratic environments do not hold in data intensive science contexts. By updating these assumptions, archivists increase the relevance of their skills in the age of digital data.

Acknowledgments

We gratefully acknowledge the materials scientists who shared their experiences with us. We also thank Elizabeth Yakel for her comments on several versions of the manuscript, the members of the University of Michigan Archival Research Group for their suggestions, and the anonymous reviewers for their valuable feedback, which helped to improve the manuscript.

This material is based upon work supported by the National Science Foundation under Grant No. 0724300. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

American Institute of Physics (2001) AIP study of multi-institutional collaborations. College Park, MD: Center for the History of Physics

Berman F (2008) Got data? A guide to data preservation in the information age. Communications of the ACM, 51(12), 50-55

Birnholtz J P, Bietz M J (2003) Data at work: Supporting sharing in science and engineering. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, FL, 339-348.

Borgman C L, Wallis J C, Enyedy N (2007) Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal of Digital Libraries, 7(1-2), 17-30

Botticelli P (2000) Records appraisal in network organizations. Archivaria 49, 161-191

Bowker G (2006) Memory practices in the sciences. Cambridge, MA: MIT Press

Curry A (2011) Reuse of old data offers lesson for particle physicists. Science 331, 694-695

Cragin M H, Palmer C L, Carlson J R, Witt M (2010) Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society A 368, 4023–4038

Elliott C A (1974) Experimental data as a source for the history of science. American Archivist 37 (1), 27-35

Feijen M (2011) What researchers want. Utrecht: SURFfoundation

Gantz J, Reinsel D (2010) The digital universe decade—are you ready? IDC White Paper, May 2010. Retrieved February 27, 2011 from http://idcdocserv.com/925

Haas J, Samuels H, Simmons B (1985) Appraising the records of modern science and technology: a guide. Cambridge, MA: MIT

Hackman L, Warnow-Blewett, J. (1987). The documentation strategy: a model and a case study. American Archivist 50 (1), 12-27.

Hey T, Tansley S, Tolle K (Eds.). (2009) The fourth paradigm: data-intensive scientific discovery. Redmond, Washington: Microsoft Research

Joint Committee on Archives of Science and Technology (1983) Understanding progress as process. Chicago: Society of American Archivists

King G (2011) Ensuring the data-rich future of the social sciences. Science 331, 719-721

King W J (1964) The project on the history of recent physics in the United States. American Archivist 27 (2), 237-243

Lauriault T, Craig B, Taylor D R, Pulsifer P (2007) Today's data are part of tomorrow's research: Archival issues in the sciences. Archivaria 64, 123-178

Manick S, Smith M, Clopeck K (2009) Materials science and engineering at MIT. The scientific data flood: A case study of "how much information?" Retrieved March 16, 2011 from http://hmi.ucsd.edu/pdf/HMI_Case_MaterialsScienEng.pdf

Piwowar H, Chapman W (2009) Public sharing of research datasets: a pilot study of associations. Journal of Informetrics, 4(2), 148-156

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

U.S. National Institutes of Health (2003) Data sharing policy and implementation guidance. Retrieved March 15, 2011 from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

U.S. National Science Foundation (2010). Application and administration guide, Chapter IV.D.4. Retrieved March 15, 2011 from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

Van House N A (2003) Digital libraries and collaborative knowledge construction. In A. P. Bishop, B. Buttenfield, & N. A. Van House, (Eds.), Digital library use: Social practice in design and evaluation. (pp 271-295). Cambridge, MA: MIT Press

Wallis J C, Borgman C L, Mayernik M, 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). Retrieved March 17, 2011 from http://www.ijdc.net/index.php/ijdc/article/view/67/46

Warnow-Blewett J, Capitos A J, Genuth J, Weart S. R. (1995) AIP study of multi-institutional collaborations: Phase I: High energy physics. Report No. 1: Summary of

project activities and findings: Project recommendations. College Park, MD: American Institute of Physics. Retrieved March 15, 2011 from http://www.aip.org/history/pubs/collabs/hep-rp1.htm

Warnow-Blewett J, Genuth J, Weart S R (2001) AIP Study of Multi-Institutional Collaborations: Final Report: Highlights and Project Recommendations. College Park, MD: American Institute of Physics. Retrieved on March 15, 2011 from http://www.aip.org/history/pubs/collabs/highlights.html

Yin R K (2008) Case study research: Design and methods, 4th ed. Thousand Oaks, CA: Sage

Zimmerman A (2008) New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology & Human Values, 33(5), 631-652