Title
Who is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility

Authors
Wallis, Jillian C.
Borgman, Christine L.

Publication Date
2011-10-01

DOI
10.1002/meet.2011.14504801188

Peer reviewed

Who is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility

Jillian C. Wallis
Dept of Information Studies
Graduate School of Education & Information Studies, UCLA
00+1+3102060029
jwallisi@ucla.edu

Christine L. Borgman
Dept of Information Studies
Graduate School of Education & Information Studies, UCLA
00+1+3108256164
borgman@gseis.ucla.edu

ABSTRACT
Data repositories rely on the deposit of materials from the communities they serve, forming a chain of stakeholders from the data creator to the repository and data user. Top-down policies that describe the responsibilities of the depositing scientists and other stakeholders are drafted accordingly. But we see very little deposit of scientific data beyond the Big Sciences or communities for whom deposit is required by publications. As part of an ongoing data practices study, we asked scientific researchers about who would be responsible for the data collected. It is clear that researchers are not talking about who is responsible for the data. The results presented here are meant to demonstrate the need for further research into what it means to be responsible for research data and how this responsibility is delegated to members of a research team.

Keywords
Scientific data, data practices, data curation, data deposit, data management plans, IP, responsibility.

INTRODUCTION
Increasingly, research data will be deposited in data repositories, either alone or in conjunction with deposit of publications. Deposit presumes that someone is responsible for those data. Responsibility may or may not cease once data are deposited. Responsibility for data management and contribution is distributed between many stakeholders, including the researcher who collects or generates data, the funding body that supported data creation, the library who wishes to provide data curation services, etc, but the chain starts at the researcher who generated the data or under those whose direction the data were generated. If there is any confusion as to who is the individual responsible at this point, there will potentially be no flow of data downstream to the other stakeholders.

In establishing baseline information about the research practices of researchers within a large collaboration, we had expected the researchers to know who was the data author, data owner, or otherwise responsible for data. This was not the case. We received answers to these questions that were hesitant, ambiguous, and in some cases contradicted answers from their collaborators. The answers provided also did not align with current policy recommendations [7, 16,18]. Results presented here are meant to demonstrate that the question of "who is responsible for data" is not well understood by the researchers at the headwaters, as well as establish new research questions and methods that merit further study in order to bridge the gap between the expectations of those downstream with those at the data source.

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

BACKGROUND
Enabling reuse of scientific data can be of tremendous future value as such data are often expensive to produce or impossible to reproduce. Data associated with specific times and places, such as ecological observations, are irreproducible. They are valuable to multiple communities of scientists, to students, and for applications that extend beyond the sciences, such as drafting public policy. Research on scientific data practices has concentrated on big science such as physics [4-5] or on large collaborations in areas such as biodiversity [6-8]. Equally important in understanding scientific data practices is to study small teams that produce observations of long-term, multi-disciplinary, and international value, such as those in the environmental sciences. The variety of practices associated with data management and range of understanding of what constitutes "data," which are well known issues in social studies of science [9], present practical problems for data curation.

With the NSF requirement for all new proposals to include 2-page data management plans recently put into effect (January 18, 2011), researchers are faced with making some

specific plans about how to collect, organize, and perhaps curate their data. Now more than ever, researchers need to be aware of their own relationship to their data, whether data author, data owner, or otherwise responsible for data, and the responsibilities entailed with their relationship.

Data Authorship

The NSF *Long-Lived Data Report* [1] describes a series of research data stakeholders, including "data authors", i.e., those who create the data. This term is not used in other literature, but makes certain sense. Data are created as a part of research, by the researchers or through the equipment they employ. Foucault [10] describes an author as someone who helps to define the form of the authored work and functionally the author, "points to the existence of certain groups of discourse and refers to the status of this discourse within a society and culture." He goes on to describe how not all texts can be authored, and that there are four important characteristics of the author-function:

1. The author-function is tied to the legal and institutional systems that circumscribe, determine, and articulate the realm of discourses
2. It does not operate in a uniform manner in all discourses, at all times, and in any given culture
3. It is not defined by the spontaneous attribution of a text of a text to its creator, but through a series of precise and complex procedures
4. It does not refer, purely and simply, to an actual individual insofar as it simultaneously gives rise to a variety of egos and to a series of subjective positions that individuals of any class may come to occupy

Foucault restricts his analysis of the "author-function" to the domain of discourse, but he admits that this is not the only domain where the author-function exists and that, "…it is obvious that even within the realm of discourse a person can be the author of much more than a book – of a theory, for instance, of a tradition or a discipline within which new books and authors can proliferate."

Foucault's criteria alone do not determine whether or not data is author-able, per se, but rather direct us to interrogate the legal and institutional systems that would determine the author-ability of data. According to scientific researchers is data something that can be authored? If so, who identifies themselves as an author, and what conditions determine authorship?

Data Ownership

The NSF has required that researchers who receive NSF funding share their primary data with other researchers "within a reasonable amount of time" [11]. This suggests that the data are owned by researchers for roughly the time it takes them to publish from the data, at which point they become available more broadly to competitors. Although, according to McSherry [12], "data, especially scientific data, are classic public domain material, largely because they are legally treated as synonymous with 'facts.'" This may be how the law views research data, but do researchers believe their data are in the public domain? If so, who has the authority to release the data and when are data released?

Data Responsibility

Many domains of ethics, including that of science look to the philosophical debate and use of "moral responsibility" [13-14]. Even where the term responsibility is used as part of the research integrity literature [15-16], "moral responsibility" is the concept being invoked. "Moral responsibility" emerged from religious philosophy in the debate surrounding causal determinism and became the basis of how a moral society is identified [17]. A moral society is one where individuals have the opportunity to make choices, specifically that they have the opportunity to make moral choices. Thus, moral responsibility is predicated on the person being aware of the responsibility and having the option to choose to undertake the task for which they are being held responsible.

The 2007 RIN Stewardship of Digital Research Data Guidelines [18], provides the following recommendation first and foremost for the establishment of a successful data curation effort. *"**The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible**, and they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfill their responsibilities in accordance with these principles"*. A related report from UKOLN [2] then defines six different roles, rights, and the responsibilities associated with each role. The roles, rights, and responsibilities of these stakeholders as defined by this report are a top-down proscription to ensure that digital libraries of data propagate beyond the life of a grant, an individual, or an institution. The configuration described in the report is not the only possible way, and this is not necessarily how the researchers themselves might envision the stakeholders and their respective roles, rights, and responsibilities. Researchers are not the only ones with a stake in the research data, but they are the source of the data. Their ad hoc data sharing processes affect how those downstream who may wish to maintain and reuse research data. Who is responsible for research data? And what does responsibility entail?

Data Practices Research at CENS

We have been documenting and facilitating the data practices of a distributed, collaborative, and interdisciplinary research center, the Center for Embedded Networked Sensing (CENS), since its inception in 2002. CENS is a National Science Foundation (NSF) Science and Technology Center based at UCLA, with four other partner institutions in California. The mission of CENS is to develop sensing systems for scientific and social applications through collaborations between scientists, computer scientists, engineers, and experts in other domain

areas. CENS initially received five years of NSF funding, from 2002 to 2007; funding was renewed for another five years, from 2007 to 2012. Over 300 faculty members, students, and research staff are now associated with CENS. Technology research partners in CENS include computer scientists, electrical engineers, and statisticians, while application scientists include seismologists, terrestrial ecologists, environmental engineers, and marine biologists. Other members of the Center come from urban planning, design and media arts, and information studies.

Research in the first three years of the Center (2002-2005) was driven more by computer science and engineering requirements than by scientific problems. Initial research focused heavily on the design and deployment of sensing technology. Concerns about equipment reliability, capacity, and battery life, and whether data were being captured at all outweighed considerations of data quality and usefulness. Once the basic technical problems were resolved, the CENS research program became more science-driven, while continuing to explore core computer science and engineering problems in wireless sensing networks. While the initial framework for CENS was based on autonomous networks, early scientific results revealed the difficulty of specifying field requirements in advance well enough to operate systems remotely. Most CENS' research is now based on dynamic "human in the loop" deployments where investigators can adjust monitoring conditions in real time. CENS employs an array of sensor technologies, many of which feed data to one another to perform computation in-network, allowing for real-time feedback. Further discussion of CENS and the data collected at CENS can be found in our previous publications [19].

METHODS

Results presented here are drawn from two rounds of ethnographic interviews, 43 in total, with participants from the CENS community about their data practices. Interviews ranged from 45 minutes to 2 hours, with an average of 60 minutes per interview. During the first round, in 2005-2006, 22 participants were sampled randomly, stratified based by whether their research was in the realm of technology or science. During the second round of interviews, in 2009-2010, 21 participants were sampled randomly, but this time stratified by the magnitude of their coefficient of betweenness centrality, i.e. how connected they were with the rest of a co-authorship network that was constructed using CENS publications. Betweenness centrality was dimension of interest for other topics covered during our data practices interviews and had no bearing on this work per se. There was a 5 person overlap between the two samples that happened by chance. These interview rounds were not intended to be a panel study, thus they are not directly comparable.

Circumstances at CENS have not changed significantly during the interim between rounds, and the same questions were asked during both, so we combined the results from the two rounds and analyzed them as one. When combined, the final sampling is representative of the composition of the larger CENS community, and can be broken down along two compelling categorical dimensions, that of the scientific or technological focus of their research, and their position as faculty, research staff, or student. The distribution or participants along these dimensions is presented in table 1.

Participants	Science	Tech	Totals
Faculty	13	6	***19***
Research Staff	7	3	***10***
Students	5	9	***14***
Totals	***25***	***18***	***43***

Table 1: Participant breakdown

Interview Protocol

The interview questions or interest here were asked very near to the end of the interviews during which data and data practices were discussed throughout, so participants were primed for answering questions about the data they collected as part of their collaborative research. Due to the variability between interviewers, participants, and other circumstances of the interview, not all of the interviews covered all the questions or probes from the protocol and in some cases more probing was used to follow up with answers. Another source of variability is the interview protocol used as the interview protocol changed over the 4 years between interview rounds as the goals of our research evolved. Question 1 below was used during the first interview round and then expanded into a series of three questions, 1-3 below, in the second round. The following three questions are the focus of the research presented here:

1. Would you consider yourself the author of a dataset? Or what are your criteria for determining authorship of data?
2. Who would you consider to be the owner of a dataset?
3. Who is responsible for a dataset?

Initially we were interested in who was considered the creator of the data, and asked participants in the first round of interviews the first question. Of the 43 total interviews 42 of the participants were asked this question, and 41 answered. The second question emerged from the answers to the first question during the first interview round. More than a third of the participants from the first round were uncomfortable with the term 'author' when applied to data and offered 'owner' as an alternative. To more consistently capture who owned the data, the second question was included in the second interview protocol. Of the 43 total interviews, 21 of the participants were asked this question, and 28 participants provided an answer to this question or mentioned ownership during the course of the interview. The third question was added to the second interview protocol when we realized the first two questions did not provide us with an understanding of who is responsible for the data collected at CENS. Of the 43 total interviews, 18

participants, all from the second round of interviews, were asked the third question and 18 participants answered.

Data Analysis

Interviews were coded in Nvivo 8 first by question and uses of the terms 'author', 'owner', and 'responsibility'. The answers to these questions were then coded again by emergent themes. A typology of mutually exclusive answers provided, i.e. who was identified as being author, owner, or responsible, was developed for these three questions, and by this typology answers were categorized and counted. Counts were subdivided by question asked, and dimensions based on two dimensions of interview participant categories: their status as faculty, student, or research staff; and whether their research is considered to be technological or scientific. The answer counts for each question were also used to create state transition tables to see the correlation of a participant providing answer Y to question B given that they provided answer X to question A.

RESULTS

Results are presented in two sections. The first captures the variety of answers provided for each of the three questions, and demonstrates the problems with the terms 'author' and 'responsible'. The second section provides a typology of answers and numerical analyses in the form of answer frequency, frequency broken out by various dimensions, and a state transition table.

Authorship

Interview participants were all asked whether they would consider themselves to be the author of their data, and what criteria determined authorship of data. We thought this question was straightforward and would give us some insight into who would be responsible for the data collected at CENS. Instead we saw that using 'author' with regards to data did not work for many of the participants. Once we were able to get past this block, the participants provided a wide variety of potential data 'authors'.

Author Terminology

Participants from both interview rounds had difficulty understanding what we were trying to ask with the first question, and repeated the question or requested clarification from the interviewer. There were also first-blush responses, such as *"I guess I never thought about it that way,"* or *"That's a good question,"* which were followed by long pauses and then some content-bearing answer. Fourteen participants exhibited the behaviors listed above when asked this question, and this response was unique to this question of all those we asked during the course of the interviews.

Three participants assumed that 'author' implied 'sole-author', and replied negatively. Only through further probing was this sole-authorship assumption made plain. For instance a participant uses the following in her description of criteria that determine authorship, *"Unless I went out and collected it all by myself I wouldn't really consider myself to be the author."* She then goes on to define her criteria for authorship to include herself, and when asked if she would consider herself a co-author of the dataset, she agreed to this statement without hesitation. Another participant made the leap to co-author on her own, *"I don't think I would ever consider myself the author of a data set, just because everything I do is so interdisciplinary. ... I would consider myself a co-author of that."* It is difficult to go back and interpret how many negative answers would have been positive without indicators that were volunteered by the participant, such as the following authorship criteria, "*So I guess if I was doing the experiment alone I would consider myself the author."*

Two participants were uncomfortable with the term 'author' because of the possible fabrication or artful creation connotation. *"It's an interesting question actually. See for me it's just the wrong way to use that terminology. Because it, to me that must imply fabrication. Like 'authorship' to me implies creation and therefore to say that I authored this dataset to me means, like, I made it up. ... I can't distinguish that question. I guess not. Like, see, to me I, the term I use is collector, like. Or collect, like, I collected this data. That's the verbiage. It's pretty common among other people in my group,... Because that implies that the data is public and out there and all you did was kind of collate."* Another participant had a similar problem with the term, *"I don't like the idea of the word 'author,' because it's not an artful creation. It's a representation of some kind of phenomenon that you're trying to characterize."*

Two participants limited authorship to published works, a rare occurrence for data within this community. One participant cites historical precedent for publishing datasets, *"I mean, I think in smaller studies historically you probably could have said that. But then data sets weren't published as ends into themselves, except for species counts, like identification like Darwin did. They went around the world just identifying things, and then that was a matter of record. Bougainvillea, I think was another one. That's what I would get, a plant named after him. And just botanists and zoologists and stuff, those data sets sort of can be authored by a single person."*

Who Authors Data?

The following are identified as potential authors of data collected at CENS: contributors, intellectual contributors, the equipment, no one, and the institution that supported the data collection. Contribution to the data collection effort is most often offered as the criteria determining authorship, but what counts as contribution varies between participants.

For some participants, it is the collection activity that determines authorship. Three participants said that they would be the author, *"if I was the person who went out and did the collecting,"* or something similar. Another three participants took this notion of collection determining authorship even more literally and ascribed author-status to

the equipment itself that was being used to collect data, with one of these three saying, *"I don't know if I ever consider myself to be the author if the equipment generates the numbers."*

For many of the participants, collection alone is not enough to merit author-status, and some intellectual contribution must be made in order to be considered as an author of a given dataset. The following response is an example of where the participant tries to delineate between cases of intellectual contribution and just data collection. *"So for example, my student [name] is going to be using part of this for his dissertation, so he's clearly an author of the data from my perspective. So what about all of [my co-PI]'s students who go out and help set up? I mean, I don't know. I don't even know their level of involvement. But just because you go out and you set up the ladder and you help tie down the instrument pack to the ladder, that doesn't seem like that would be authorship. And for me and my own data, I hire a lot of technicians. On papers and even on reports, there have sometimes been bad feelings about that, because people have felt like they worked a lot on this project and they thought that they should be an author. But, you know, they didn't have an intellectual contribution. They just did what I wanted them to do - to go out and to collect the data. And so long-term technicians end up having some intellectual contribution, and so they probably deserve to be an author, but short-term people who basically just are bodies to write down the data or something, probably not."*

Participating in the analysis of data, or even writing the code that is used for analysis is considered an intellectual contribution. Experiment design is considered an intellectual contribution, although only in conjunction with either data collection or data analysis. Writing the code that generates a dataset or writing the instrumentation software that helps collect the data can also be considered an intellectual contribution. Processing the data in some way counts, as one participant points out, *"Authorship of data. To me, that would need some sort of processed data, so I wouldn't think of myself as the author of a sequence of images being pulled from a Web cam, but I might think of myself as an author of a highly-processed set of data from a Web cam."* Other intellectual contributions that merit authorship status include being the Principal Investigator of a research group as the PI shapes the research agenda and bring in the funding to support the research, which in one participant's terms makes them a *"de facto author"*.

One participant provided justification for authoring only a subset of a dataset, because his intellectual contribution only extended to the subset, even though the subset makes the whole dataset more usable. *"I guess I can consider myself the author of most of the time corrections. But they are not just the time, they are not only time corrections. Like there are some one-seconds, we have two different sets of instruments in Mexico and [a collaborator] claims they were a second apart. So there is a one second fix on a number of our stations. And then on top of that it's my time corrections. So I am the author of some of the time corrections. I would never claim authorship to the seismic data because so many people have put in so many hours into just making the whole thing work that."*

Contribution was not the only criteria for authorship status, and there were several outliers. Two participants expressed that data cannot be authored, and not just that they could not consider themselves the author of a dataset. One participant offered the hosting institution is the author of the data he works with.

Who Owns Data?

Participants were asked if they would consider themselves the owner of their data, and what criteria would determine ownership. Participants also provided 'owner' as a title they were more comfortable with than 'author' and as a result discussed the criteria that determined ownership even when not asked this question. In contrast with the previous question, there were no ambiguous requests for clarification or hesitation to collect their thoughts on this issue. That said the answers for this question still fell in a wide range, including: contributors, Principal Investigators, the institution, the public, and no one.

Two participants designated their research group as the owner of collected datasets, which is then followed by some definition of who constitutes the research group. For instance, *"the group, which is this group and jointly with [the group at a partner institution]"* or *"the lab… I always think of it as, belonging to, say, UCLA, you know, the lab."* These two participants are both the PIs of their respective research groups. A graduate student of the latter PI also identifies her institution and CENS as the owners of this data, and claims *"they will publish those data online."* Two graduate student participants gave similar responses as to the ownership by the research group, one of which is part of the first respondent's research group but from a different community of practice and the other student comes from another group entirely.

The Principal Investigator is identified as the owner by three participants, and another participant conferred ownership more generally on the study designer, which is not always the PI. The PI is identified as the owner by one of these participants because they *"design and guide the data collection".* Another participant identifies the PI as the one with the right to release data because they are *"the PI on the grant that brought in the data",* but also points out that they don't really have a set protocol in his field. The participant who provides a more general option of the study designer as the owner, sees ownership as being the right to publish from the data without contest, *"I do think that if you have an idea and you're trying to study a specific phenomenon and you collect the data to look at that phenomenon, that that person is entitled to carry out their work on what they wanted before a bunch of other people can come steal their idea."*

The public is also named owner, for federally funded research data. The ownership in this case changes over time as the data is owned by the researchers for some period of time and then becomes part of the public record. A participant from the domain of seismology provides the following response. *"It depends on if it was federally funded or not. So, of course we have a requirement to share the data with everybody if the Federal Government has provided funding. And all our experiments have been federally funded. So I consider myself kind of a overall director of the data, but I don't necessarily own it. It is up to me to make sure I've done a good job collecting it and doing some permitted processing. But then once it is in a public archive, you know I can be certainly recognized as the, the one who collected it and the one who knows most about it. But I'm not necessarily the owner of it anymore. Typically in seismology we will have several stages of operating, you know directing the data, actually holding the data and then ownership of the data and what we tend to do is a, for two years, yeah, if we were responsible for raising the money for the experiment and why should we collect the data and have no use of the data. So there is an embargo on actually making it public. After about two years the data needs to be made public. So, at that point we are no longer owners anymore."*

Three other participants identified the public as the owner of the data, and provide varying lengths of embargo period from no time at all to three years. Another participant explained that the data were part of the public domain not because of federal funding, but as a result of trying to capture phenomena on a time-scale longer than any one research career. *"... the point of trying to understand ecological systems, trying to put things together and there is a long run of trying to understand things like global change, basic ecological process, air pollution and so forth. So you know, it's not my data."*

Responsibility

Just under half of the participants were asked who was responsible for their data. Responsibility is assigned to contributors, Principal Investigators, and Institutions, and in some cases researchers did not know who were responsible for their data. Part of the problem with this question is that there are so many ways to be responsible for data, so even if researchers feel responsible, we have no way to tell from this data what these responsibilities entail.

The Multiple Types of Responsibility

When asked who was responsible for the data, the interview participants answered the question based on different types of responsibility that were offered along with their assignment of responsibility. These definitions can be grouped into data quality, data protection, data access, long-term data maintenance, and support for data reuse either through answering questions or providing metadata and documentation.

The quality of the data collected and made available reflects on the researcher who put it out there, and as a result we see participant responses like the following, *"I am responsible for the quality of the data that go out."* One computer science participant mentions consulting with another researcher from Ecology to *"say whether or not [the data is] reasonable"* in dealing with algorithms and models of ecological processes.

A growing portion of CENS researchers are collecting potentially sensitive data from participants, but only one of these researchers commented on being responsible for the protection of this type of data. *"[I] collect very sensitive information from a bunch of participants. So I'm gonna have to like, have the file of like really personal data, I don't want to be responsible for that. ... I think what I have to do is go through legal here. I need a lawyer to bless a project."* Another researcher who collects potentially sensitive data identifies a technician who works with her as being responsible for her data, because he has custody of the data and controls access. *"I guess right now it's [the technician] because he designed the system. So he.. He has all the like passwords, usernames for the database and who works with the database on a daily basis."*

The long-term maintenance and support for reuse responsibility is indicated by a few participants. One participant automatically swaps in the term 'management' when asked about responsibility, *"as far as management, I think I'll be that to the group so. ... I guess I would probably be the point man of that."* One participant equates responsibility to being an authority for answering questions about the data to support reuse, *"I think this kind of practice is just to indicate who generated the data. And the questions would end up going to that person."* Another participant offers that his documentation fulfills this support role for reuse, *"I mean I guess since I lead the data I'm probably the most responsible ... It's documented, you know, you know what the columns are but I'm not going to try and analyze it with you."*

Who is Responsible for Data?

The party responsible for a given dataset seems to tie closely to the criteria for authorship or ownership of the dataset. One participant mentions that the person who will answer reuse questions is the same person who generated the data, which for this participant was also considered the author of the dataset. Another participant sees responsibility as stemming from ownership, *"I think on some level the person who feels a sense of ownership is... who put in effort and who invested the time in collecting it and maintaining it, who will hope to use it, is responsible for that."* Other participants responded with the research group, as being the data authors, data owners, and the party responsible for the data.

Two of the participants who did not think data could be either authored or owned still had a sense of who should be responsible. The first participant is a PI and claims, *"I am*

responsible for the quality of the data that go out." He then goes on to mention that he would pass along any inquiry to the appropriate researcher in his group that should handle the query. The second participant responds, *"I don't own the data or I'm not, it's not data I created per se, it's data I collected. And therefore I'm responsible for how and why it was collected."*

Typology of Answers

The answers discussed above can be grouped into the following seven categories of entity identified as author, owner, or party responsible for data: 'institution', 'Principal Investigator', 'contributor', 'no one', the 'public', 'don't know', and the 'equipment'. Contributor, here is used to include both intellectual contribution as well as lower thresholds of contribution, such as only data collection. While the PI could be considered a special case of a contributor, enough participants mentioned the PI specifically to warrant a separate category. It should be noted that some researchers provided more than one answer per question, when multiples occurred they were both included in the counts, so the numbers across the top are the number of respondents, and the totals at the bottom are the number of responses.

Answer Incidence

The incidence of answer categories by question can be seen in table 2a. We can see from these counts that the 'equipment' is only ever named as an author, and similarly the 'public' is only ever named as an owner. Participants were occasionally unsure of who was author or who was responsible for data, but were sure of who owned the data. The 'institution' is more frequently provided as the owner than the author or responsible for the data. More participants responded with 'no one' could author data than 'no one' could own it. The 'Principal Investigator' is more frequently named responsible for the data than owner or author of the data. By a wide margin, participants confer authorship, ownership, or responsibility on anyone who contributed.

	Author (41)	**Owner (28)**	**Resp (18)**	**Total (87)**
Inst	3	4	1	***8***
PI	4	5	5	***14***
Contributor	20	15	10	***45***
No one	9	2	0	***11***
Public	0	5	0	***5***
Don't know	5	0	2	***7***
Equip	2	0	0	***2***
Totals	***43***	***31***	***18***	***92***

Table 2a: Distribution of Answer Type by Question

In the next two tables, the answers to all three questions were collapsed and then redistributed along the dimensions of the participant. Table 2b presents the distribution of answer type by the role of the interviewed participant: faculty member, research staff, or student. Although not all faculty members at CENS are PIs, the faculty members interviewed were all Primary Investigators. PIs were more apt to name themselves or the 'public' as author, owner, or responsible for data. The research staff members were also more likely to not know or declare that 'no one' can author, own, or be responsible for data. Students were more apt to express that anyone who contributes authors, owns, or is responsible for data. Students less frequently name the PI and more frequently name the institution as author, owner, or responsible for data.

	Faculty (19)	**Staff (10)**	**Student (14)**	**Total (43)**
Inst	2	0	6	***8***
PI	10	2	2	***14***
Contributor	18	6	21	***45***
No one	4	3	4	***11***
Public	5	0	0	***5***
Don't know	2	3	2	***7***
Equip	1	1	0	***2***
Totals	***42***	***15***	***34***	***91***

Table 2b: Distribution of Answer Type by Status

Figure 2c presents the answer frequency by the type of research domain the participant came from. The research at CENS breaks between technology development and application science. While the frequency distributions are quite similar for the first four answer types, we see more divergence with the last three. Technology researchers are more apt than the application science researchers to name 'contributors' as author, owner, or responsible for data. Whereas the application science researchers more apt than the technology researchers to name the 'PI' or 'Institution' as same.

	Application (23)	**Technology (20)**	**Total (43)**
Inst	6	2	***8***
PI	13	1	***14***
Contributor	16	29	***45***
No one	5	6	***11***
Public	3	2	***5***
Don't know	4	3	***7***
Equip	1	1	***2***
Totals	***47***	***42***	***91***

Table 2c: Distribution of Answer Type by Research Focus

State Transition Table

In order to capture the sequence of answers provided by participants, the three state transition tables were constructed, including transitions from question 1 to question 2 (author to owner), question 2 to question 3 (owner to responsible), and question 1 to question 3 (author to responsible). Only the third state transition table is provided, but results from all three are presented. The left column is the initial answer state, and the top row is the following answer state. The values are probabilities based on the answers given by the participants that given some initial answer state X from question A, you would expect a following answer of Y to question B.

In the state transition from question 1 of authorship to question 2 of ownership. There is a high correlation from the author answer to the owner answer, so if the participant named X as author they were more likely to also name X as owner. The one exception to this is the PI, from the data above we saw that the PI is more likely to name the 'PI' as author, and here we see that those participants who named the 'PI' as author then name the 'public' as owner. In the state transition table from question 2 of ownership to question 3 of responsibility, as in the previous there is some correlation between the initial and following state, but it is not as strong. There is less correlation because no matter who was named owner, almost always the 'contributors' or the 'PI' are named responsible. In the state transition table from question 1 of authorship to question 3 of responsibility, Table 3 below, there is still a correlation between author answer and responsible answer. But, here we also see that the 'PI' is still likely to be named responsible no matter who is identified as owner, unless a 'contributor' is identified as the author, in which case the 'contributor' is also identified as responsible for the data.

Resp > / Author ∨	Equip	Don't know	Public	No one	Contrib	PI	Inst
Equip							
Don't know		1					
Public							
No one						1	
Contrib		.06		.06	.53	.13	.06
PI						1	
Inst						1	

Table 3: Author to Responsibility State Transition Table

DISCUSSION

Problem of Terminology

One of the most compelling problems with data responsibility, including authorship and ownership aspects, is that of the available terminology. These terms all evoke complex connotations that make it difficult to get at the heart of the issue of who is responsible for data. The researchers at CENS have some notion of what it means to be an author, and being a "data author" does not fit within their notion of authorship. Most researchers took author to mean that they were the creator, fabricator, or generator of data and missed the "authority" connotation. Authority would be conferred on these researchers because they designed the data collection method and actually performed the data collection, and are thus the authority on why and how the data were collected. Rarely were we able to get the participants own terms for this concept because the concept did not really exist with authorship, or had multiple interpretations in the case of responsibility.

If there is no person named responsible for the data, due to some confusions over who should be the responsible party or what is even meant by 'responsibility', then it is difficult to be able to establish complete metadata about a dataset. Metadata for this type of object should be able to provide a person who made the data possible through whatever definition of author or creator. The legal metadata should include some ownership or rights information for those who would like to use the data. And finally some contact person is necessary for answering questions for these one-off datasets that do not fit into some larger, established digital library of data.

Conflation of Ownership and Access

For many of these participants the notions of data authorship and ownership are tangled up with granting access. The author of a dataset should not necessarily become the public in order for the public to have access to the data. The author of a dataset should be set in stone regardless of who has access to the data. Ownership is similarly entangled, and can negatively impact the depositing of research data if researchers believe that the institution or the public are the 'owners' of the data. Above a participant mentions that the data she has been party to collecting is owned by her institution and CENS, and that CENS will make the data public. It should be noted that CENS is a collection of individual researchers and does not have the agency to put data in the public domain, individual researchers must do this for themselves for it to happen. With responsibility of data relying so heavily on ownership, we would argue that it is important for researchers to maintain a sense of ownership.

Data as a Product of Research

Other complications occur with data responsibility and data authorship. As we saw above a few of the participants linked authorship with requiring the publication of data, something that very rarely happens in the CENS community, and the research community at large. The book chapter, article, or conference is seen as the product of research and contains a bit of the data, but is more of a vehicle for disseminating data interpretations. If we are

hoping that researchers will be responsible for their data, they need to be given opportunities to author their data through publication.

Collaboration Exacerbates the Problem

If the participant community had been more uniform, from the same discipline or status we may have seen less disparate results. As indicated by the answer incidence tables broken down by science and technology, or by the status of the participant, the participants working together in research groups have quite different notions of who authors, owns, or is otherwise responsible for data. This disparity is a good indicator that very few groups are discussing these issues or we would have seen more uniform results. This is further supported by the participant that remarked on the lack of protocol in their field.

Authority & Accountability

From the answers to the ownership and responsibility questions we can pull out a list of tasks which the person responsible would have the authority to perform and be held accountable for with regards to their data. Participants listed that as the owner they would: have the right to publish their data online, they would have the right to publish from their data without contest, they would be recognized as the authority on their data, they would have the authority to answer any questions about the data, and they have the authority to document their data. The ways that researchers would be accountable for their data include: the quality of the data, the quality of the documentation, protection of sensitive data, provision of data access, long-term maintenance of data, and support for data reuse. Some of these overlap with the rights and responsibilities laid out by the Dealing with Data report, where rights such as "first use" are balanced with responsibilities such as "managing data for the life of the project" [2].

CONCLUSION

Responsibility is a multi-faceted concept and one that differs by situation. As a diverse community we would expect a diverse array of answers from CENS researchers to questions of responsibility that break along disciplinary boundaries. But we see diversity within even the same research group, leading us to believe that very few researchers are discussing who authors, owns, or is responsible for their data. We began thinking that it would be easy to identify who was the data creator and who was responsible, but this is not the case. Reports that rely on clearly defined roles around data responsibility will find that there is a great disparity between the top-down recommendations for data stewardship and on-the-ground awareness of data responsibility. Further research is needed to capture responsibility with more nuance and depth.

FUTURE RESEARCH

Further research should capture how responsibility is distributed within research collaborations, as well as who will be responsible for writing data management plans and who will be responsible for fulfilling the data management plan. Rather than rely on the problematic 'author', 'owner', and 'responsibility' language used as part of the research described here, we would opt for a new framework of authority and accountability, where researchers are asked about who would be asked who was accountable for the different aspects of responsibility that emerged from this research, or who had the authority to perform certain tasks with regards to the data.

- Who has the authority to: release the data, answer questions about the data, write data management plans, etc.?
- Who is accountable for: data quality, protection of and access to data, providing metadata or documentation, long-term maintenance, etc.?

By asking about the tasks rather than abstract concepts with ambiguous interpretations we should be able to capture a much high resolution understanding of how roles, rights, and responsibilities for data are distributed according to the researchers themselves.

ACKNOWLEDGEMENTS

Research reported here is supported in part by grants from the National Science Foundation (NSF): (1) The Center for Embedded Networked Sensing (CENS) is funded by NSF Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; (2) CENS Education Infrastructure (CENSEI), under which much of this research was conducted, is funded by National Science Foundation grant #ESI-0352572, William A. Sandoval, Principal Investigator and Christine L. Borgman, co-Principal Investigator. (3) Towards a Virtual Organization for Data Cyberinfrastructure, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI; (4) Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU. The research reported here relied on data collected and analyzed by Matthew S. Mayernik in addition to the authors.

REFERENCES

[1] (2005). Long-Lived Digital Data Collections: Enabling Research and Education for the 21st Century. National Science Board. Retrieved from http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf on 18 July 2005.

[2] (2007). Dealing with Data: Roles, Rights, Responsibilities, and Relationships. UKOLN.

[3] (2003). Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. Meeting organized by the Wellcome Trust, Fort Lauderdale, Florida, Wellcome Trust.

Retrieved on 29 December 2009 from http://www.wellcome.ac.uk/.../groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf.

[4] Traweek, S. (1992). Beamtimes and lifetimes : the world of high energy physicists. Cambridge, Mass., Harvard University Press.

[5] Galison, P. (1997). Image and Logic: a material culture of microphysics. Chicago, Chicago University Press.

[6] Bowker, G. C. (2000). "Biodiversity datadiversity." Social Studies of Science **30**(5): 643-683.

[7] Bowker, G. C. (2000). "Mapping biodiversity." International Journal of Geographical Information Science **14**(8): 739-754.

[8] Bowker, G. C. (2000). Work and information practices in the sciences of biodiversity. VLDB 2000, Proceedings of 26th international conference on very large data bases, Cairo, Egypt, Kaufmann.

[9] Bowker, G. C. (2005). Memory Practices in the Sciences. Cambridge, MA, MIT Press.

[10] Foucault, M. (1977) What is an author? In Language, counter-memory, and practice. Trans. D.Bouchard and S. Simon. Ithaca, Cornell University Press.

[11] NSF Award and Administration Guide, Chapter VI, 4b. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

[12] McSherry, C. (2001) Who owns academic work? Battling for control of intellectual property. Cambridge, MA, Harvard University Press.

[13] Floridi, L. (Ed.). (2010). The Cambridge Handbook of Information and Computer Ethics. Cambridge, UK: Cambridge University Press.

[14] Mitcham, C. (2003). Co-Responsibility for Research Integrity. Science and Engineering Ethics, 9(2).

[15] (1992). Responsible Science, Volume I: Ensuring the Integrity of Research Process. National Academy of Sciences.

[16] Environments, C. o. A. I. i. R. (2002). Integrity in Scientific Research: Creating an environment that promotes responsible conduct. National Research Council.

[17] Eshleman, A. (2009). Moral Responsibility. Retrieved from http://plato.stanford.edu/entries/moral-responsibility/.

[18] (2008). Stewardship of digital research data: a framework of principles and guidelines: responsibilities of research institutions and funders, data managers, learned societies and publishers, Research Information Network.

[19] Borgman, C. L., J. C. Wallis, et al. (2007). "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries." International Journal on Digital Libraries **7**(1-2).

The columns on the last page should be of approximately equal length.