Research Article

Inna Kouper*, Anjanette H Raymond, Stacey Giroux

An Exploratory Study of Research Data Governance in the U.S.

https://doi.org/10.1515/opis-2020-0010
Received November 1, 2019; accepted June 30, 2020

Abstract: Making decisions regarding data and the overall credibility of research constitutes research data governance. In this paper, we present results of an exploratory study of the stakeholders of research data governance. The study was conducted among individuals who work in academic and research institutions in the US, with the goal of understanding what entities are perceived as making decisions regarding data and who researchers believe should be responsible for governing research data. Our results show that there is considerable diversity and complexity across stakeholders, both in terms of who they are and their ideas about data governance. To account for this diversity, we propose to frame research data governance in the context of polycentric governance of a knowledge commons. We argue that approaching research data from the commons perspective will allow for a governance framework that can balance the goals of science and society, allow us to shift the discussion toward protection from enclosure and knowledge resilience, and help to ensure that multiple voices are included in all levels of decision-making.

Keywords: data governance, knowledge commons, research community

1 Introduction

Making data part of the published and openly available record of research is seen as key to improving science and furthering knowledge (Nosek et al., 2015; Uhlir & Schröder, 2007). The heterogeneous and fluid nature of research data, however, raises considerations regarding not only how to make data available, but also how to protect the sustainability and resilience of research data. Making decisions regarding the quality of data, access to it, ethical and security implications, as well as the overall credibility of research constitutes research data governance (Leonelli, 2019; Rosenbaum, 2010). Recognizing that the term "governance" has its own history, this paper uses it as an umbrella term that combines aspects of coordination, regulation, curation, and management of research data and highlights the agency-structure interactions in the context of norms, rules, decisions, and institutions (Hufty, 2011).

Who governs or makes decisions regarding research data? On one hand, individual researchers make many decisions drawing on the norms, practices, and guidelines of their disciplinary communities. While this freedom to make decisions allows disciplines to grow, it also leads to great variability across individual practices that, whether deliberately or inadvertently, can lead to questionable practices or even misconduct, including data falsification or fabrication (Fang, Steen, & Casadevall, 2012). Misalignments in career advancement incentives also often discourage researchers from putting effort into archiving datasets and addressing data sustainability issues (Nosek et al., 2015; Vines et al., 2014). The growing involvement in research data governance by entities from outside the traditional academy, including the government

***Corresponding author, Inna Kouper,** Indiana University, Informatics, United States, E-mail: inkouper@indiana.edu
Anjanette H Raymond, Stacey Giroux, Indiana University, Informatics, United States

and the publishing industry, also points to the urgent need for a long-term and proactive approach to the governance of research data (Harmon, 2017; Lamdan, 2018; Larivière, Haustein, & Mongeon, 2015; Sample, 2012).

Current approaches to research data governance include initiatives that emphasize preservation, infrastructures, and open access (Foster & Deardorff, 2017; Vardigan & Whiteman, 2007; Wilkinson et al., 2016). While the open access movement and repositories go a long way toward facilitating sharing of data and research findings, especially for researchers in developing nations (Chan & Costa, 2005), it remains only one aspect of a data governance ecosystem. To broaden such an ecosystem, researchers and policy makers must grapple with the growing and increasingly diverse landscape of organizations and stakeholders involved in the production and use of research data, seeking to understand the relationships both among these entities and between these entities and individuals who carry out research. Governance strategies will require moving beyond sharing tools and compliance, and approaching data as a collective transdisciplinary object that enables knowledge work in multiple domains and over time.

In this paper, we present results of an exploratory study conducted with individuals from academic and research institutions to understand what entities are perceived as making decisions regarding data, to what degree those entities affect how individuals work with data, and who should be responsible for making key decisions in data governance. Our results show that there is considerable diversity and complexity across stakeholders, both in terms of who they are and their ideas about data governance. To make sense of this diversity and to establish a future research agenda, we propose to frame research data governance in the context of the governance of new commons: a set of resources, such as the Internet or digital culture, that has been identified as sustaining multiple stakeholders and being vulnerable to failure, conflicts, and power imbalances (Hess, 2008). More specifically, we discuss research data as part of a knowledge commons, a shared resource that combines the properties of private, public, and common goods and does not yet have stable rules or institutional arrangements in place. As such, we envision our study to be the starting point of an analysis using the theories and empirical frameworks applied to other commons (Hess & Ostrom, 2007; Ostrom, 2010), with the ultimate goal of making recommendations for governance of research data.

2 Background

The concept of data governance is rapidly gaining traction. Mostly grounded in the frameworks of information technology and corporate asset management, it has been defined as decision-making about the effective use of information assets (Ladley, 2012; Marco, 2006). Using "data" and "information" interchangeably, the enterprise asset perspective smoothes out the differences and uncertainties in definitions and concerns itself with the value of data for business, thus seeking to identify types of data / information, roles and responsibilities, costs, and risks associated with storage and management of data (ECAR Working Group, 2015; Hagmann, 2013). While such interchangeable use of terms may work in a corporate setting, its casualness and practicality are not applicable in the context of a research enterprise, where the differences between data, information, and knowledge are important. The development of more precise frameworks of research data governance is still at its early stages, and many approaches focus more on concepts and definitions rather than practices, decision-making, and outcomes (Alhassan, Sammon, & Daly, 2016).

Data governance models also come from the research on organizations, as more and more of them invest in data governing initiatives (Panian, 2010). Approaches range from generalized models that stem from information technology governance and define areas that any company can use in their data governance strategies, such as data quality, metadata, and access, to contingency approaches that argue that each organization requires a specific data governance configuration (Khatri & Brown, 2010; Wende & Otto, 2007). Existing models discuss centralized and decentralized approaches that determine who stores and gathers data assets within the organization, and hierarchical versus cooperative decision-making that assigns responsibilities within organizations (Wende, 2007). Case studies find that models of data governance adopted within organizations help to define roles and responsibilities with regard to data processes and requirements (Cheong & Chang, 2007).

As entities with clearly defined boundaries and stakeholders, individual organizations can adopt existing models and develop their own governance approaches for whatever their needs are. Those models, however, cannot be simply transferred to governing research data. Several factors pose challenges to adopting organizational or corporate models of access and governance for research data, including the changing nature of data, the complexity of research networks that include both data producers and consumers, and the insufficiency of open access models (Hilgartner & Brandt-Rauf, 1994). The nature of data is shifting toward more heterogeneity as well as fusion with other components of research such as physical samples, laboratory techniques and protocols, algorithms, documents, and many other inputs of scientific work. These inputs converge into research products that allow not only for answering research questions, but also supporting independent verification and future reuse (Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010). At the same time, such research products complicate the issues of "asset" identification and sharing in the governance context.

Research networks in which data are generated and shared are increasingly complex. As discussed above, sharing involves many stakeholders beyond the primary researcher and audience. The decision-making in such networks belongs to many actors who may have differing goals and claims to parts of the data products. Moreover, open access models do not necessarily fit everywhere. For example, data are being exchanged privately, published with embargos, used in training prior to publications, released with non-disclosure agreements, and so on. Finally, the legal system, particularly in terms of regulations around intellectual property and commercialization, affects decision-making with regard to data. Boundaries between what is public domain or public good and what is private and patentable shift all the time.

Much of the work on research data governance focuses on public access and sharing practices rather than broader issues of decision-making throughout the lifecycle of research (Perrier et al., 2017). For decades the U.S. government and funding agencies have been trying to encourage researchers to share their data (Fienberg, Martin, & Straf, 1985; OSTP (US Office of Science and Technology Policy), 2013; Shelby, 2000). Encouraged and sometimes challenged by the open access movement, journals and professional societies joined in and began to establish guidelines for publishing data, although approaches to publishing and sharing recommendations and compliance vary (Pitt & Tang, 2013; Stodden, Guo, & Ma, 2013; Van Noorden, 2013; Vasilevsky, Minnier, Haendel, & Champieux, 2017). Institutions of higher education have also begun to develop policies to establish control over data produced by their employees, with marked differences in types and contents of policies across universities (Briney, Goben, & Zilinski, 2015). Additionally, libraries and data repositories develop policies and guidelines that target data management inefficiencies, including lack of documentation, proprietary formats, duplications and inconsistencies, and so on (Borer, Seabloom, Jones, & Schildhauer, 2009).

Research data governance currently exists in many forms and covers such efforts as sharing, openness, curation, management, and compliance. Research data management, in particular, has gained visibility as many research organizations face challenges of developing tools, policies, and services in support of working with data and encouraging its dissemination and archiving (Pinfield, Cox, and Smith, 2014). Discussing differences and similarities between the terms "data management" and "data governance" and their varying academic and professional roots is beyond the scope of this paper. As will be shown later, we consider "governance" to be the broader term and a nexus between several traditions of research and practice.

While the sheer amount of governance-related effort is encouraging, policies are often siloed, do not promote consistency, and may even provoke contradictory behaviors. Thus, data sharing behavior continues to vary across disciplines, work areas, and geographic regions (Tenopir et al., 2011). In addition to disciplinary cultures and regional differences, sharing depends on many factors, including individual researcher characteristics, desired degree of control, available resources, and institutional pressures (Fecher, Friesike, & Hebing, 2015; Kim & Adler, 2015; Kim & Stanton, 2016).

Individuals responsible for data end up acting on their own, increasing the risks of non-compliance, data hoarding, and data loss (Gormley and Gormley, 2012). The complexities of research products and networks illustrated above, as well as the variety of stakeholders, access strategies, and legal contexts, raise questions of who should coordinate and regulate such complexities as well as who owns and takes care of data at various stages of its lifecycle. As various groups attempt to make and enact data policies, understanding settings where they make decisions and the different groups vying for control are crucial for effective governance (Marshall, 1984).

3 Methodology

The impetus behind this exploratory study is the need to understand how various stakeholders in research data come together and make decisions with regard to data. However, before we can examine the dynamics of decision-making and the roles groups or individuals play in the creation of data-related standards and policies, we need to identify those groups. This exploratory study is the first step in examining norms and behaviors associated with the governance of research data, and as such it aims to identify entities that, according to various stakeholders and research communities, are **perceived** to contribute to the governance of research data or **should** be contributing to it. Thus, the study focused on two essential questions:

1. What entities affect one's data use; and
2. Who should be responsible for making decisions with regard to research data.

The study used a **s**tructured survey methodology. An anonymous web survey was open for responses from August 1, 2018 to October 31, 2018. The survey contained approximately 25 questions split into three sections: background and demographics, data sharing and governance, and organizations. The background section included questions about respondents' age, gender, education, main responsibilities at work, and disciplinary orientation. The data sharing and governance section contained questions about data sharing experiences, who should be responsible for governing data, and entities that affect researchers' work with data. The section about organizations asked respondents to identify entities that are involved in data governance in their research areas and describe their own involvement with those organizations. We include the text of our survey instrument as Appendix A.

We designed our questions to be broad to avoid leading the respondents toward a specific understanding and to allow them to provide clarifications if needed. At the same, we provided examples to illustrate what we mean by such terms as compliance and authorization agency, commercial entity or government entity. Most of the questions were multiple choice format and included an option of "Other". The choices were derived from the information science literature and tested in a separate pilot with sixteen researchers. After the feedback from the pilot, we clarified wording, modified response options, and made choice options consistent across the questions about perceived and desired impact.

The study attempted to reach several disciplines with strong histories of data management but with different attitudes to data sharing that range from full and open sharing to embargoes to no sharing, including earth sciences, social sciences, and library and information science. We restricted our region to the United States, asking the survey respondents to acknowledge that they work primarily in the U.S. The survey was disseminated through several academic and professional listservs, including the Federation of Earth Science Information Partners (ESIP), the Research Data Access and Preservation Association (RDAP), the International Association for Social Science Services and Technology (IASSIST), multiple sections of the American Statistical Association, social media groups of the American Sociological Association, and a list of individuals culled by the authors from multiple sources (approximately 1,000 earth scientists and 800 social scientists). Given the overlap in the audiences of many listservs, we estimate that the combined audience that was notified of the survey consisted of at least 3,000 researchers and data professionals. We received 129 responses; therefore our response rate can be estimated to be around 5%.

4 Results

4.1 Demographics

Responses to our survey came from a diverse pool of participants across age, gender, and disciplinary affiliation. In terms of age, the largest proportion of respondents were 31-40 years old (31%; Table 1). With regard to gender, respondents were skewed toward female (53% female, 33% male, 14% prefer not to say or no answer).

Table 1. *Respondents' Age. Respondents were asked, "What is your age?"*

Age group	**Frequency**	**Percent**
Less than 20	2	2%
21-30	14	11%
31-40	40	31%
41-50	16	12%
51-60	18	14%
61-70	17	13%
More than 70	5	4%
Prefer not to say or no answer	17	13%
Total	**129**	**100%**

Respondents also came from a variety of disciplines that define both their areas of degree received and current work (Table 2). The largest proportion of respondents came from the social sciences (49% and 32% for degree taken and work area respectively), followed by the library and information sciences (20% and 19%) and people who work in areas categorized as "other," i.e., biostatistics, public health, and business (12% and 14%). Fewer respondents identified their disciplinary affiliations as earth sciences, computer science, life sciences, and physical sciences. A number of respondents moved out of the traditional sciences, such as geology, chemistry, biology, or statistics to become computer scientists or professionals in information technology, biostatistics, and other areas.

Table 2. *Degrees and Current Area of Work. Respondents were asked, "In what discipline/area did you earn your highest degree?" and "In what discipline/area would you say you currently do most of your work?"*

Discipline / Domain	**Percent Earned Degree**	**Percent Current Area of Work**
Social Sciences	49%	32%
Library and Information Science	20%	19%
Earth Sciences	11%	6%
Life Sciences	5%	2%
Physical Sciences	4%	2%
Humanities	2%	1%
Computer Science	0%	4%
Other (Biostatistics, Health, Business)	12%	14%
No answer	20%	20%

In terms of organizational affiliation, 72% of respondents came from colleges and universities. The remaining were distributed across government (9%), non-profit (6%) and for-profit organizations (7%).

Individuals who selected “Other” as a category added explanations that they were either retired or worked in places that can be considered more than one type, e.g., both government and non-profit.

Rather than asking the respondents for their professional title, e.g., professor, librarian, lecturer, and so on, which carries certain assumptions about what people do at work, we asked questions about the types of responsibilities and data-related activities performed at work, encouraging respondents to think about their actual work tasks. The types of responsibilities included administrative work (e.g., office or grant support), research (tenure or non-tenure track), teaching (tenure or non-tenure track), professional (e.g., library, IT managers), and leadership (e.g., chair, upper management, supervisor). Data-related activities included collecting and analyzing original (one’s own) data, using data collected by other researchers, using data provided by government agencies (e.g., NASA, Bureau of Labor Statistics), using existing archival and library materials, and assisting others in collecting, managing or analyzing their data.

Many participants divided their responsibilities across most if not all types of data-related responsibilities. We grouped the responses into position orientations based on the percentage majority rule. For example, if a respondent indicated that 50% or more of their time was spent on research, his or her position would be categorized as “research”. If research was less than 50% but more than 40%, the position would be “mostly research”. Similar rules were applied to all types of responsibilities. Respondents who did not have at least 40% in any of the five responsibility types were categorized as “distributed”.

Out of 95 individuals who responded to the question about responsibility types, 34 were in research-oriented positions, 38 were in professional services positions, six each were in teaching and leadership positions, two stated they do mostly administrative work and nine had responsibilities distributed across four or five categories. Table 3 provides distributions of data-related activities per each position orientation.

Table 3. *Nature of Work Activities per Position Orientation (N = 95, percent per each category, more than one option could be selected). Respondents were presented with a list of activities and asked, “Is the nature of your current work such that you [please select all that apply].”*

Position orientation	Work with original (own) data	Work with others’ data	Work with govt data	Work with archival or library materials	Assist others
Research-oriented positions	82%	65%	47%	35%	62%
Teaching-oriented positions	100%	83%	50%	67%	83%
Professional services	45%	26%	21%	26%	74%
Administrative positions	100%	0%	0%	0%	50%
Leadership positions	67%	67%	17%	0%	50%
Positions with responsibilities distributed across	89%	67%	56%	56%	78%
Percent per Total Respondents	68%	68%	35%	33%	49%

Most of our participants, regardless of their position orientation, collect their own data. Most notably, individuals in teaching, professional, leadership and distributed positions work with the original data they collect. Similarly, many respondents indicated that they assist others in collecting, managing and analyzing their data. Four respondents within professional services position that are not included in the table above selected “Other” as their data-related activity and provided the following activities: data dissemination, metadata documentation, data curation, and collection development for data.

Describing their data sharing experiences, less than half of respondents indicated that they published their own data or shared it privately with researchers outside of their collaborator circle (41% and 38% respectively, see Table 4). A larger proportion (57%) indicated that they assisted others in sharing their data. Nineteen percent of respondents never shared their data. Respondents who selected “Other” added that data was already publicly available or that they share it through their own website rather than an established publishing venue, such as a data journal or a repository.

Table 4. Sharing Experience (N = 94, more than one answer option could be selected). Respondents were asked, “Thinking about data in terms that are applicable to the discipline/area in which you currently work, do you have data sharing experience?”

Sharing Experience	Frequency	Percent
I assisted others in sharing their data	54	57%
I published my own data (as an appendix, a separate publication or in a repository)	39	41%
I sent data privately to another researcher outside of my collaborator circle	36	38%
No, I have never shared data (at all or beyond my immediate collaborators)	18	19%
Other	3	3%

4.2 Involvement in Decision-Making

To learn about entities that are believed to be involved in data governance, we asked several questions about the types of entities that currently make decisions in research data governance and affect participants’ work with data. Respondents were provided with a choice to rate the following entities with respect to how much they affect their work: individual researchers (other than oneself), academic institutions, scientific community as a whole, US government (e.g., Congress, local municipalities), publishers, funding agencies (e.g., NSF or private foundations, compliance and authorization entities such as Institutional Review Boards), commercial entities involved with data (e.g., Microsoft, Facebook), government or non-profit entities involved with data (e.g., World Health Organization). The analysis of responses did not show any statistically significant difference between participants with different disciplinary backgrounds, therefore the results are provided for the whole sample (Figure 1).

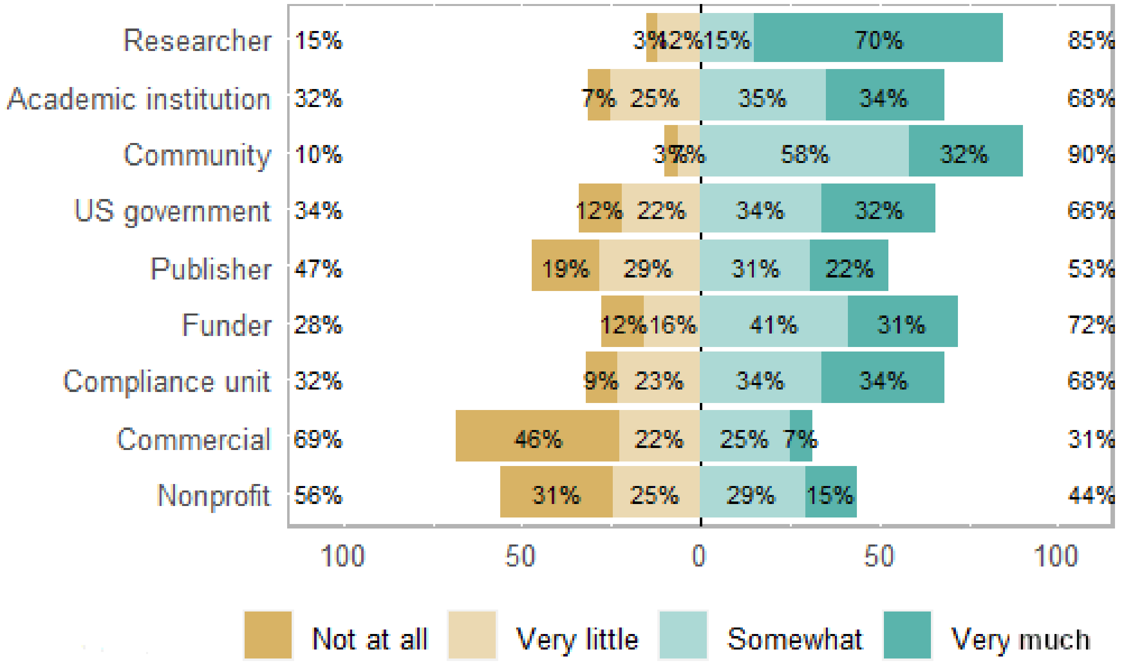


Figure 1. Entities Affecting Data Work (N = 92). Respondents were asked, “To what extent do you know or believe that each of these entities affects your work with data?”

Most respondents identified individual researchers other than themselves as affecting their work very much or somewhat (70% and 15% correspondingly). The scientific community as a whole and funders were the next two groups perceived to affect respondents' work, although the split between "somewhat" and "very much" was more even. Similarly, an even split between "somewhat" and "very much" can be seen in the impact evaluation of academic institutions and compliance units. Publishers received an approximately even split between affecting somewhat / very much and very little / not at all. The least impactful entities include commercial entities and nonprofit organizations (69% and 56% affecting respondents very little or not at all).

We also asked respondents to name specific entities they know or believe make decisions with regard to data, such as decisions about data collection, analysis, documentation, and sharing. Respondents were asked to name up to eight specific organizations and then for each named organization identify what types of decisions those organizations make. Several options for decision-making included adding a data management section to the organization's code of conduct, declaring data sharing as a goal for the profession, requesting journals in the domain to require publishing data, and discussing compliance with the emerging or existing regulations. Respondents were also allowed to add their own types of decisions.

Overall, our respondents provided 185 entities that they believe make decisions regarding data, with about half of the respondents providing more than one entity. Even though the question asked about specific entities, many respondents provided general answers, such as "Funders" or "Publishers". We coded all the responses into categories similar to the categories of entities above. Table 7 below provides frequencies of mention of each type of entity, as well as examples of specific entities.

Table 5. Organizations that Make Decisions about Data. Respondents were asked, "As it pertains to the work that you do, what specific entities do you know or do you believe make decisions regarding data (i.e., decisions regarding collection, analysis, documentation, sharing or any other aspect of data activities)?"

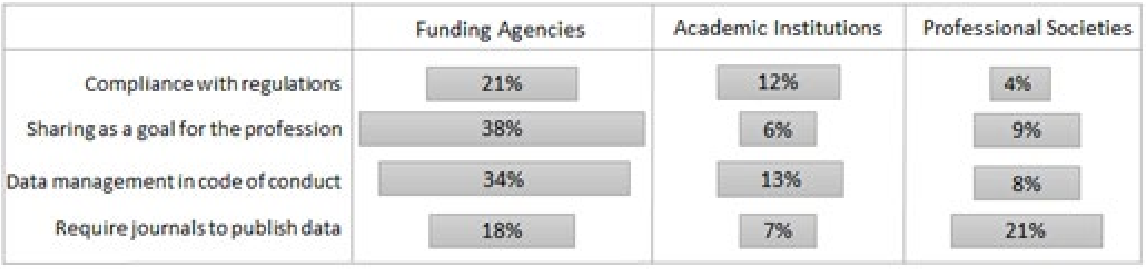
Type of Entity	Frequency	Percent
Funders (e.g., NSF, NIH, DOE)	44	24%
Compliance and authorization (e.g., IRB, HIPAA, IT and security)	23	12%
Data organizations (e.g., NOAA, USGS, ICPSR)	22	12%
Publishers	18	10%
Academic institutions	16	9%
Individual researchers	17	9%
US government (e.g., OSTP, USAID)	12	6%
Professional societies (e.g., AGU)	11	6%
Commercial entities	7	4%
Scientific community as a whole	2	1%
Other (e.g., libraries, IT, data managers)	13	7%
Total	**185**	**100%**

Funding agencies were mentioned most of the time (24% of all organizations mentioned). Among the specific agencies mentioned by the respondents were the National Science Foundation (NSF), the National Institutes of Health (NIH) and its institutes and centers, and the Department of Energy. Data organizations refer to government or non-profit organizations that collect and make available large amounts of data. Such organizations included the US Geological Survey (USGS), the National Oceanographic and Atmospheric Administration (NOAA), the Agency for Healthcare Research and Quality (AHRQ), the Census Bureau, the Bureau of Labor Statistics, and some others, for example, the Inter-university Consortium for Political and Social Research (ICPSR), a repository of social science data. Compliance and authorization entities included both administrative units that oversee compliance with the existing rules (e.g., IRB or research administration) and the rules themselves (e.g., HIPAA or FERPA).

Among the publishers, only two journal publishers were mentioned explicitly - SAGE and PLOS One. The rest included general references to journal policies, journals, publishers, and editors. Academic institutions were mentioned 9% of the time and included several specific universities as well as references such as "my university", "university policies", or "academic institutions". Individual researchers who make decisions about research data included respondents themselves ("myself" or "my team") or others ("peers", "collaborators", "supervisors", etc.)

A new entity that was not mentioned in responses to the previous question was *professional society*. While it was mentioned only 11 times out of 185, it nevertheless adds one more type of organization to the landscape of data governance. The American Geophysical Union was mentioned several times, in addition to the American Meteorological Society and Society for Political Methodology. Commercial entities included specific companies as well as references to "companies" and "data vendors". The category "other" included such entities as libraries, IT departments, and data managers.

For every entity named, respondents were also asked about types of data-related decisions that those entities have made. The decisions of some of the entities were described through their primary function, for example, compliance and authorization units address compliance with emerging or existing regulations, or publishers require journals to publish data. Figure 2 illustrates decisions for the three selected entities that had the most volume and variety of data-related decisions: funders, academic institutions, and professional societies.



	Funding Agencies	Academic Institutions	Professional Societies
Compliance with regulations	21%	12%	4%
Sharing as a goal for the profession	38%	6%	9%
Data management in code of conduct	34%	13%	8%
Require journals to publish data	18%	7%	21%

Figure 2. Decisions Made by Selected Entities (100% is one decision type across all 11 entities; only 3 shown in this figure). Respondents were asked, "What kind of data-related decisions have those entities made? [please select all that apply]"

As can be seen from the illustration above, funding agencies were perceived to be making many decisions regarding research data, including addressing compliance, establishing data sharing as a goal for research as a profession, adding data management to codes of conduct and requiring journals to publish data. Academic institutions were seen to be playing a smaller role and contributing in smaller proportions to each decision. Interestingly, the role of professional societies is perceived to be mostly in requiring professional journals to publish data and not in, say, establishing data sharing as a goal of the profession.

As discussed above, professional societies were mentioned as making decisions about research data, but respondents named only five specific societies that do so. In contrast to this small number, respondents indicated that they belong to **93** unique organizations that support academic and professional communities across a wide variety of areas, including biology, mathematics, environmental sciences, statistics, information science, education, social sciences, and so on. This contrast suggests that there is a considerable lack of decision making on the part of most professional organizations.

4.3 Governing Responsibilities

Finally, respondents were asked to consider which entities **should** be responsible for making decisions regarding research data and rate them on the 3-point scale (should be primarily responsible, should be involved but not primarily responsible, and should not be involved; Figure 3).

The majority of respondents agreed that individual researchers should be primarily responsible for making decisions about data (65%). Many other entities, including academic institutions, the scientific community, US government, funding agencies, compliance units, and nonprofit organizations should be involved but not primarily responsible, with the scientific community also having a larger share of primary responsibilities assigned to it (37%). Overall, all entities except publishers and commercial entities received strong support from our respondents for being involved and/or primarily responsible.

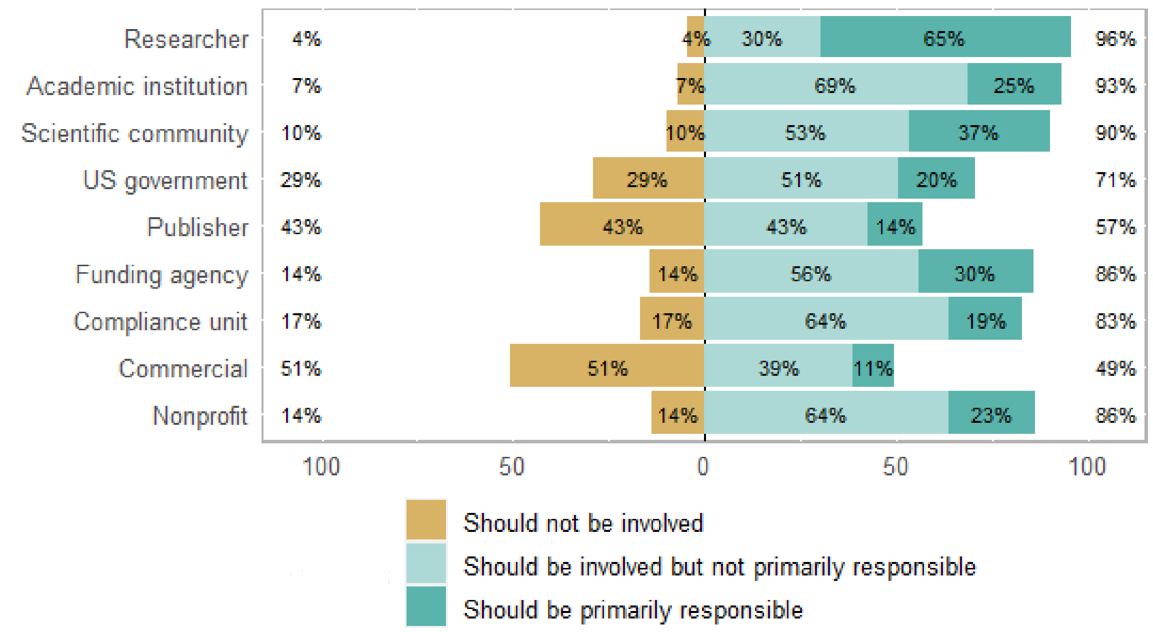


Figure 3. Entities that Should Be Responsible for Decisions Regarding Data. Respondents were asked, “To what extent do you believe each of the following entities SHOULD BE responsible for making key decisions in data governance? [primarily responsible, should be involved but not primarily responsible, should not be involved]”.

5 Discussion

5.1 Ambivalence toward Responsibilities

Our exploratory study shows that many entities are perceived to be involved with research data governance and many of those same entities should have some level of responsibility for governing research data. These entities are an essential part of creating and supporting communities that produce scientific knowledge. Many individuals also belong to multiple communities, identifying themselves with a specific discipline and with interdisciplinary communities and with communities that are involved in various aspects of the data lifecycle.

The complex and overlapping nature of data, its role in knowledge production, and the associated communities present difficulties in coordinating these various actors. As individuals in these communities carry out a broad range of activities related to data, including data collection, analysis, management, and reuse, they rely upon data produced by or under the control of others, and tend to be exposed to alternative positions, guidance, and influence around data collection and use. For example, government agencies such as NASA provide data for research openly and without restrictions (Murphy, 2019), however, as that data can be used in various domains and combined with other data, researchers may end up working with policies that vary from data sharing as a condition for publication to no guidance about data at all (Vasilevsky et al., 2017). Adding to the complexity of the data sharing environment is consideration that not all data can be shared and researchers have to grapple with the economic, political, and ethical implications of their sharing or non-sharing decisions (Simon et al., 2017).

Almost all our findings point to the complexities of research data networks and an increasing overlap in data responsibilities, which, in turn, increases the ambivalence toward assigning specific responsibilities to specific agents. Our survey respondents did not report clear demarcations of their professional and data orientations, and instead indicated that they are involved in many aspects of data work, including data collection, analysis, management, and assistance to others. They were also ambivalent about who should be primarily responsible for decision-making regarding research data and distributed the responsibilities across many entities. Wallis and Borgman (2011) discussed similar ambivalences in the context of data ownership and accountability. Their exploration indicated that data does not necessarily fit with researchers' interpretations of authorship and responsibility for the products of research. We suggest that as the digital ecosystem of data grows, the complexity of the causes and responses to this ambivalence will have an even greater impact on the resilience and sustainability of research data.

Individuals in our study believed that research data governance is the responsibility of the entire data ecosystem (Table 7, Figure 3). However, the extent to which parties should be involved varied. For example, most of our respondents agreed that individual researchers should be primarily responsible for making decisions about research data. Half of respondents felt that commercial entities should not be involved at all, and close to half felt the same way about publishers. Many concerns have already been raised about publishers and other well-funded entities behaving as rule and norm creators in a data ecosystem that is heavily reliant on external funding and increasingly implicates research data in innovation, collaboration, and return on investment in science (Janicke Hinchliffe, 2018; Maxson Jones, Ankeny, & Cook-Deegan, 2018; McCain, 1995). In addition to economic influence and decision-making models, some academics were equally concerned with the growing influence of politically motivated institutions within the research data ecosystem (Edwards, 1999; Rosen, 2017; Ruppert, Isin, & Bigo, 2017). These questions of who should have a seat at the table, and what the associated costs might be, will be difficult to reconcile as individuals and organizations continue discussing and developing research data governance.

5.2 Public and Market Forces

The reluctance of our respondents to include publishers and commercial entities in decision-making with regard to research data points to the tensions between science as a public institution and commercialization of many aspects of knowledge production, including overreliance on market solutions in information technology, academic administration, and even dissemination of research. Parallels can be drawn between the governance of research data and the governance of the Internet, another complex and diffuse community that faced many sociotechnical dilemmas and can serve as an example of a fragile equilibrium among powerful players (Mueller, 2012). While historically the governance of the Internet belonged to the Internet Engineering Task Force (IETF) and an international community of network designers, researchers, operators, and vendors, which alleviated the fears of one player becoming dominant and preventing open exchange of information and services, over time that governance structure has been challenged as private actors such as network and content providers started performing governance functions (Raymond, 2013). The debates over net neutrality, languages on the Internet, and domain name system (DNS) illustrate how the increasing authority of market-based and technocratic forces and tensions among competing national interests may necessitate stronger regulation and enforcement of rules against preferential treatment based on payments and favoring private or government interests rather than the interests of communities (Abbate, 2000; Electronic Frontier Foundation, n.d.).

This is not intended to suggest that the data ecosystem should not include economic and political considerations or relevant entities. We do, however, suggest that a thoughtfully developed collective action community, one that takes on board the concerns and comments of all community members, is a necessity in situations when economically or politically motivated individuals have the potential to become the primary influencers in the creation of community standards. The results of our exploratory study suggest that those who work with data in academia are already aware of the growth and potential influence of non-academic entities. Half of our respondents were against publishers and commercial entities making decisions about data. This raises a question about forms of governance that would enable collective action and balance the influence of public and market forces in deciding whether research data should be open or closed, where it should reside, and who gets to access and control it.

5.3 Data as Part of the Knowledge Commons

Our study further confirms that the complexity of the data ecosystem is characterized by the following: (1) the varied and substantial number of organizations and institutions involved; (2) individual actors that act both on behalf of their employing organizations and the larger collective communities; (3) the presence of organizations with economic interests; (4) the belief that everyone is responsible for data governance; (5) and the absence of a shared vision and collective action organizations that balance the influences of the various actors. We suggest placing research data within the knowledge commons as one of the approaches that can advance the research and action in data governance.

Knowledge commons refers to making cultural and intellectual resources accessible to all members of society and treating those resources as common pool, i.e., available to all without exclusion (Frischmann, Madison, & Strandburg, 2014; Ostrom, 1994). The proposal to treat data as commons is not new; many commons frameworks have emerged that focus on intangible resources (Hess, 2008; Jimenez, 2019). The most common approach in the context of data, however, is to call data repositories "data commons" and view them as the main governing mechanism that supplies rules for managing data and helps negotiate ownership between individual and institutional actors (Eschenfelder & Johnson, 2014). Many repositories focus on developing the mechanisms of access with a particular emphasis on technological capabilities and policy-making (Anderson, 2017; Grossman, Heath, Murphy, Patterson, & Wells, 2016; Kindling et al., 2017).

The researchers and data professionals in our study believe that researchers, or broadly, data producers, should be responsible for making decisions regarding data, and yet, not many professional organizations and even academic institutions are visibly involved in data governing activities. The discussions about what to do with research data are often framed in the context of funder-compliant data management and repository-driven data curation (Corti, Van den Eynden, Bishop, & Woollard, 2014). With such a strong emphasis on compliance and mandated sharing, and without deeper disciplinary and institutional guidance on navigating data production and consumption, the individual researcher and the community are going to be left with hundreds of decontextualized data sets that become digital graveyards, at best, or lead to an inaccurate scientific record at worst.

While the development of a governance framework is out of the scope of this study, we would like to emphasize that making decisions about storage and licensing is not enough for governing research data. Moreover, according to our study, funding agencies, academic institutions, and repositories are not the main stakeholders in research data decision-making. Therefore, governing initiatives should engage the research communities in wider conversations and discuss the role of data in knowledge production and consumption, its relevance to societal concerns, and the misalignment of incentives that leads to the prioritization of individual private interests over the collective and public interests.

Understanding knowledge commons governance, and research data governance as part of it, is in its nascent state. Given the central role of data in the production of scientific knowledge, ensuring its availability requires explicit thinking about the nature of data as a hybrid collective resource that combines the public, private, and common-pool models of goods. Such thinking therefore necessitates further understanding of research data's boundaries, actors, and outcomes. Strandburg, Frischmann and Madison (2017) proposed the Governing Knowledge Commons (GKC) framework that provides the tools for empirical studies of knowledge resources and their governance. Following the GKC (and other previous research) terminology, we propose to consider research data as an "action arena" – a space in which actors interact with one another and deal with the dilemmas of sharing and sustaining the resource. We have identified many of the actors in this arena, including the researchers and data practitioners, academic institutions, professional organizations, US government, federal agencies, and other organizations.

The GKC framework also calls for defining the goals and objectives of the commons under examination. Research data is not only evidence in support of scientific claims or an object that is imbued with certain value and demarcates the boundaries of scientific communities (Baker & Millerand, 2012), it is a resource that contributes to the sustainability of humankind and the increased solidarity for the common good that is built on trust between academia and the public (Fitzpatrick, 2019). As such, the goals and the challenges of building and sharing research data and the collective effort to govern it will have to go beyond the development of repositories and formal policies and cross other dimensions, such as a collective mission and action, cultural principles and social norms, design of platforms of participation, self-management of contributions, and conflict resolution systems (Fuster Morell, 2014). Key questions for future research include the complex interplays of various actors and resources involved in research data action arenas, the dilemmas and dependencies that are being created by non-profit, university, and commercial infrastructure and policy provisions in research communities, and how governance plays out in practice – at the levels of an individual, an institution, a professional society, and networks of stakeholders. Larger data collection efforts can also shed light on how various dimensions of research, such as field of study, regional context, types of data, individuals' rank and position and so on, affect data governance models and frameworks.

6 Conclusion

We began this project with a simple goal - to learn about what specific entities are involved in research data governance and who, according to individual researchers, should be involved. What we learned is that the responsibility for decision-making in research data is perceived to be distributed. Furthermore, this distribution appears to be uneven, and researchers' perceptions of decision making across entities is not consistent. Acknowledging the obvious and inevitable limitations of an exploratory study and a convenience sample, we combine our findings with the existing literature on data governance and management and posit that these perceptions point to gaps in research data governance that need to be filled. Namely, those gaps concern issues that go beyond tools and compliance, i.e., current research data governance models do not address the goal and missions of research, social and cultural norms, forms of conflict resolution, and dilemmas in sharing and use.

To begin to address these gaps and move toward a model of research data governance, we propose to use the conceptual frameworks of the commons and view data through the lens of the knowledge commons. This will allow us to expand the conception of data commons beyond the tools for storage, sharing, and compliance to include discussions of the normative aspects of governance in a way that a) ties together all professional practices around research data and associated means and objects of production and b) articulates ethical commitments and responsibilities of multiple stakeholders in research data, including individuals, research communities and government and commercial entities. Without deeper understanding of the norms, rules, subjects, and objects of research data it will be increasingly difficult to create and sustain just and socially relevant environments of knowledge production.

References

Abbate, J. (2000). *Inventing the Internet*. MIT Press.

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: An analysis of the literature. *Journal of Decision Systems*, *25*(1). doi: 10.1080/12460125.2016.1187397

Anderson, W. P. (2017, March 8). Data management: A global coalition to sustain core data. *Nature*(543). doi: 10.1038/543179a

Baker, K., & Millerand, F. (2012). Infrastructuring ecology: Challenges in achieving data sharing. In J. Parker, N. Vermeulen, & B. Penders (Eds.), *Collaboration in the New Life Sciences* (pp. 111–138).

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings.* Retrieved from http://eprints.soton.ac.uk/268555/

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some simple guidelines for effective data management. *Bulletin of the Ecological Society of America*, *90*(2), 205–214. doi: 10.1890/0012-9623-90.2.205

Briney, K., Goben, A., & Zilinski, L. (2015). Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication*, *3*(2), eP1232. doi: 10.7710/2162-3309.1232

Chan, L., & Costa, S. (2005). Participation in the global knowledge commons. *New Library World*, *106*(3/4), 141–163. doi: 10.1108/03074800510587354

Cheong, L. K., & Chang, V. (2007). The need for data governance: A case study. *Australasian Conference on Information System*. Retrieved from http://www.isihome.ir/freearticle/ISIHome.ir-26086.pdf

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: a guide to good practice*. SAGE.

ECAR Working Group. (2015). *The compelling case for data governance*. Retrieved from https://library.educause.edu/-/media/files/library/2015/3/ewg1501-pdf.pdf

Edwards, P. N. (1999). Global climate science, uncertainty and politics: Data-laden models, model-filtered data. *Science as Culture*, *8*(4), 437–472. doi: 10.1080/09505439909526558

Electronic Frontier Foundation. (n.d.). Net Neutrality. Retrieved April 24, 2019, from https://www.eff.org/issues/net-neutrality

Eschenfelder, K. R., & Johnson, A. (2014). Managing the data commons: Controlled sharing of scholarly data. *Journal of the Association for Information Science and Technology*, *65*(9), 1757–1774. doi: 10.1002/asi.23086

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. doi: 10.1073/PNAS.1212247109

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLOS ONE*, *10*(2). doi: 10.1371/journal.pone.0118053

Fienberg, S., Martin, M. E., & Straf, M. (Eds.). (1985). *Sharing research data*. Washington, D.C.: National Academies Press.

Fitzpatrick, K. (2019). *Generous thinking : A radical approach to saving the university*. Johns Hopkins University Press.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2), 203–206. doi: 10.5195/jmla.2017.88

Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (Eds.). (2014). *Governing knowledge commons*. Oxford Scholarship Online.

Fuster Morell, M. (2014). Governance of online creation communities for the building of digital commons. In B. Frischmann, M. J. Madison, & K. J. Strandburg (Eds.), *Governing Knowledge Commons*. doi: 10.2139/ssrn.2842586

Gormley, C. J., & Gormley, S. J. (2012). Data hoarding and information clutter: The impact on cost, life span of data, effectiveness, sharing, productivity and knowledge management culture. *Issues in Information Systems*, *13*(2), 90–95.

Grossman, R. L., Heath, A., Murphy, M., Patterson, M., & Wells, W. (2016). A case for data commons: Toward data science as a service. *Computing in Science & Engineering*, *18*(5), 10–20. doi: 10.1109/MCSE.2016.92

Hagmann, J. (2013). Information governance – beyond the buzz. *Records Management Journal*, *23*(3), 228–240. doi: 10.1108/RMJ-04-2013-0008

Harmon, A. (2017, March). Activists rush to save government science data — if they can find it. *New York Times*. Retrieved from https://www.nytimes.com/2017/03/06/science/donald-trump-data-rescue-science.html

Hess, C. (2008). Mapping the new commons. *The 12th Biennial Conference of the International Association for the Study of the Commons."* doi: 10.2139/ssrn.1356835

Hess, C., & Ostrom, E. (2007). A framework for analyzing the knowledge commons. In E. Ostrom & C. Hess (Eds.), *Understanding knowledge as a commons: From theory to practice*. Retrieved from http://www.wtf.tw/ref/hess_ostrom_2007.pdf

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership, and control toward empirical studies of access practices. *Science Communication*, *15*(4), 355–372. doi: 10.1177/107554709401500401

Hufty, M. (2011). Governance: Exploring four approaches and their relevance to research. In U. Wiesmann & H. Hurni (Eds.), *Research for Sustainable Development: Foundations, Experiences, and Perspectives* (pp. 165–183). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2019013

Janicke Hinchliffe, L. (2018, August). Advancing an integrated vertical stack of publication services? *The Scholarly Kitchen*. Retrieved from https://scholarlykitchen.sspnet.org/2018/08/08/integrated-vertical-stack-of-publication-services/

Jimenez, J. G. (2019, April). The key models for a legal data commons. *Medium: Legal Design and Innovation*. Retrieved from https://medium.com/legal-design-and-innovation/a-data-commons-for-law-a8ca365d10fe

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1). doi: 10.1145/1629175.1629210

Kim, Y., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management, 35*(4), 408-418. doi: 10.1016/j.ijinfomgt.2015.04.007

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology, 67(*4), 776-799. doi: 10.1002/asi.23424

Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., … Scholze, F. (2017). The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine*, *23*(3–4). doi: 10.1045/march2017-kindling

Ladley, J. (2012). *Data governance*. Elsevier.

Lamdan, S. (2018). Lessons from DataRescue: The limits of grassroots climate change data preservation and the need for federal records law reform. *University of Pennsylvania Law Review*, *166*(1), Article 12. Retrieved from https://scholarship.law.upenn.edu/penn_law_review_online/vol166/iss1/12

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLoS ONE*, *10*(6). doi: 10.1371/journal.pone.0127502

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*. doi: 10.1162/99608f92.17405bb6

Marco, D. (2006). Understanding data governance and stewardship, Part 1. *DM Review*, *16*(9).

Marshall, C. (1984). Elites, bureaucrats, ostriches, and pussycats: Managing research in policy settings. *Anthropology & Education Quarterly*, *15*(3), 235–251. doi: 10.2307/3216539

Maxson Jones, K., Ankeny, R. A., & Cook-Deegan, R. (2018). The Bermuda Triangle: The pragmatics, policies, and principles for data sharing in the history of the human genome project. *Journal of the History of Biology*, *51*(4), 693–805. doi: 10.1007/s10739-018-9538-7

McCain, K. W. (1995). Mandating sharing: Journal policies in the natural sciences. *Science Communication*, *16*(4), 403–431. doi: 10.1177/1075547095016004003

Mueller, M. L. (2012). Property and commons in internet governance. In E. Brousseau, M. Marzouki, & C. Meadel (Eds.), *Governance, regulation and powers on the Internet* (pp. 39–62). doi: 10.1017/CBO9781139004145.004

Murphy, K. (2019). NASA Earth science data: Yours to use, fully and without restrictions.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015, June 26). Promoting an open research culture. *Science*, Vol. 348, pp. 1422–1425. doi: 10.1126/science.aab2374

OSTP (US Office of Science and Technology Policy). (2013). *Increasing Access to the Results of Federally Funded Scientific Research*. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Ostrom, E. (1994). Neither market nor state: Governance of common-pool resources in the twenty-first century. *International Food Policy Research Institute*. Retrieved from http://hdl.handle.net/10535/891

Ostrom, E. (2010). The institutional analysis and development framework and the commons. *Cornell Law Review*, pp. 807–815.

Panian, Z. (2010). Some practical experiences in data governance. *World Academy of Science, Engineering and Technology*, *62*.

Perrier, L., Blondal, E., Ayala, A. P., Dearborn, D., Kenny, T., Lightfoot, D., … MacDonald, H. (2017). Research data management in academic institutions: A scoping review. *PLOS ONE*, *12*(5), e0178261. doi: 10.1371/JOURNAL.PONE.0178261

Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PloS One*, *9*(12), e114734. https://doi.org/10.1371/journal.pone.0114734

Pitt, M. A., & Tang, Y. (2013). What should be the data sharing policy of cognitive science? *Topics in Cognitive Science*, *5*, 214–221. doi: 10.1111/tops.12006

Raymond, M. (2013). Puncturing the myth of the Internet as a commons. *Georgetown Journal of International Affairs*, 53–64.

Rosen, J. (2017). Politics: Turbulence ahead. *Nature*, *544*(7651), 509–511. doi: 10.1038/nj7651-509a

Rosenbaum, S. (2010). Data governance and stewardship: Designing data stewardship entities and advancing data access. *Health Services Research*, *45*(5p2), 1442–1455. doi: 10.1111/j.1475-6773.2010.01140.x

Ruppert, E., Isin, E., & Bigo, D. (2017). Data politics. *Big Data & Society*, *4*(2). doi: 10.1177/2053951717717749

Sample, I. (2012, April). Harvard University says it can't afford journal publishers' prices. *The Guardian*. Retrieved from https://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices

Shelby, R. (2000). Accountability and transparency: Public access to federally funded research data. *Harvard Journal on Legislation*, *37*(2), 369–390.

Simon, G. E., Coronado, G., Debar, L. L., Dember, L. M., Green, B. B., Huang, S. S., … Platt, R. (2017, November 7). Data sharing and embedded research. *Annals of Internal Medicine*, Vol. 167, pp. 668–670. doi: 10.7326/M17-0863

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PloS One*, *8*(6), e67111. doi: 10.1371/journal.pone.0067111

Strandburg, K. J., Frischmann, B. M., & Madison, M. J. (2017). The Knowledge Commons Framework. In *Governing medical knowledge commons* (pp. 9–18). doi: 10.1017/9781316544587.002

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6). doi: 10.1371/journal.pone.0021101

Uhlir, P. F., & Schröder, P. (2007). Open data for global science. *Data Science Journal*, *6*, OD36–OD53. doi: 10.2481/dsj.6.od36

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature News*, *495*(7442), 426. doi: 10.1038/495426a

Vardigan, M., & Whiteman, C. (2007). ICPSR meets OAIS: Applying the OAIS reference model to the social science archive context. *Archival Science*. doi: 10.1007/s10502-006-9037-z

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ*, *5*. doi: 10.7717/peerj.3208

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. doi: 10.1016/J.CUB.2013.11.014

Wallis, J. C., & Borgman, C. L. (2011). Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the American Society for Information Science and Technology, 48(*1), 1-10. doi: 10.1002/meet.2011.14504801188

Wende, K. (2007). Data governance - defining accountabilities for data quality management. *Proceedings of Swiss - Italian Workshop on Information Systems (SIWIS)*. Retrieved from https://www.alexandria.unisg.ch/214034/

Wende, K., & Otto, B. (2007). A contingency approach to data governance. *Proceedings of 12th International Conference on Information Quality*. Retrieved from http://mitiq.mit.edu/iciq/PDF/A CONTINGENCY APPROACH TO DATA GOVERNANCE.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. doi: 10.1038/sdata.2016.18

Appendix A. Survey Questionnaire

Definitions
In this survey we use the following definitions:
Data is defined broadly as evidence in support of one's findings. Such broad definition covers structured and unstructured data as well as historical and library sources.
Data governance is defined as norms, rules, and decisions that affect how data is collected, stored, analyzed, and shared.

Background / Demographics
1. AGE What is your age?: [≤20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, >80, prefer not to say]

2. GENDER (GENDER_OTHER) What is your gender?: [F, M, Other - specify, Prefer not to say]

3. EDUC What is the highest level of education you have completed?
   a. Associate's degree (AA, AS)
   b. Bachelor's degree (BA, BS)
   c. Master's degree (MA, MS, MEng, MEd, MSW, MBA)
   d. Professional degree (MD, DDS, DVM, LLB, JD)
   e. Doctorate degree (PhD, EdD)
   f. Other (please specify)

4. ORG_TYPE With what type of organization in the US are you affiliated? If you are affiliated with more than one, please select the one you spend the most time working with:
   a. College / University
   b. Non-profit
   c. Private for-profit
   d. Government
   e. Self-employed
   f. Other (please specify)

5. EXPERIENCE How would you describe your level of experience / seniority?
   a. Student (working towards a degree)
   b. Early career (0-5 years out of the last degree)
   c. Mid-level (6-10 years out of the last degree)
   d. Senior (> 10 years out of the last degree)

6. MAIN_RESP (*if EXPERIENCE is a, skip, if b, c, or d - ask this question*): About what percent of your time do you typically spend in a week on each of the following types of responsibilities in your current role or position at your primary organization? Responses must total 100%.
   a. Administrative (e.g., office or grant support)
   b. Research (tenure or non-tenure track)
   c. Teaching (tenure or non-tenure track)
   d. Professional (e.g., library, IT managers)
   e. Leadership (e.g., chair, upper management, supervisor)
   f. Other, please specify

7. AREA_DEGREE In what discipline / area did you earn your highest degree?
   a. Earth Sciences (e.g., geology, meteorology)
   b. Physical Sciences (e.g., physics, chemistry, astronomy)
   c. Social Sciences (e.g., psychology, sociology, economics)

d. Humanities (e.g., history, philosophy)
e. Computer Science
f. Library and Information Science
g. Other (please specify)

8. AREA_CURR In what discipline / area would you say you currently do most of your work? (*carry options from the previous q*)

9. WORK_CURR Is the nature of your current work such that you [select all that apply]:
a. collect and analyze original (your own) data
b. use data collected by other researchers
c. use data provided by government agencies (e.g., NASA, Bureau of Labor Statistics)
d. use existing archival and library materials
e. assist others in collecting, managing or analyzing their data
f. other (please specify)

10. (*skip if a to EXPERIENCE, but If b*): LESS5_YRS In your career since you finished your degree, have you been supported through funding from any of these types of sources external to your organization (either as a PI or a team member)? *If c or d to EXPERIENCE*: 6PLUS_YRS In the last five years of your career, have you been supported through funding from any of these types of external sources (either as a PI or a team member)?
a. no external funding received
b. federal government funding
c. private foundations or non-profit
d. industry or commercial research groups
e. other (please specify)

Data sharing and governance

11. SHARING_EXP Thinking about data in the terms that are applicable to the discipline / area in which you currently work, do you have data sharing experience?
a. No, I have never shared data (at all or beyond my immediate collaborators)
b. I assisted others in sharing their data
c. I sent data privately to another researcher outside of my collaborator circle
d. I published my own data (as an appendix, a separate publication or in a repository)
e. Other (please specify)

12. GOVERN_RESP To what extent do you believe each of the following entities SHOULD BE responsible for making key decisions in data governance? (*primarily responsible, should be involved but not primarily responsible, should not be involved*)
a. Individual researchers
b. Academic institutions
c. Scientific community as a whole
d. US Government (e.g., Congress, local municipalities)
e. Publishers
f. Funding agencies (e.g., NSF, Sloan Foundation)
g. Compliance and authorization entities (e.g., Institutional Review Boards, offices of environmental health and safety)
h. Commercial entities involved with data (e.g., Microsoft, Facebook)
i. Government or non-profit entities involved with data (e.g., World Health Organization)
j. Other (please specify)

13. AFFECT_DATAUSE To what extent do you know or believe that each of these entities affects your work with data? (*very much, somewhat, very little, not at all, do not know*)

a. Individual researchers (other than yourself)
b. Academic institutions
c. Scientific community as a whole
d. US Government (e.g., Congress, local municipalities)
e. Publishers
f. Funding agencies (e.g., NSF, Sloan Foundation)
g. Compliance and authorization entities (e.g., Institutional Review Boards, offices of environmental health and safety Research administrators
h. Commercial entities involved with data (e.g., Microsoft, Facebook)
i. Government or non-profit entities involved with data (e.g., World Health Organization)
j. Other (please specify)

14. CHANGE_DATA Have you ever been approached with requests or pressure to change how you handle your data in the following ways?

a. Collect more data
b. Produce “positive” data or remove “bad” data
c. Better document your data, its collection, or analysis techniques
d. Make data available to others
e. Other (please specify)

15. (carry forward choices from previous q) CHANGE_WHO Who approached you to request or pressure you to change how you handle your data? [select all that apply]

	PI or supervisor	Collaborator	Representative from the funding agency	Journal editor or reviewer	Other peers	Would rather not say	Other
Collect more data							
Produce “positive” data or remove “bad” data							
Better document data and techniques							
Make data available for others							

16. CHANGE_OTHER If you checked “other” to any of the items in the previous question, please elaborate as to who approached you to request or pressure you to change how you handle your data

Organizations

17. ORG_DECISIONS As it pertains to the work that you do, what specific entities do you know or do you believe make decisions regarding data (i.e., decisions regarding collection, analysis, documentation, sharing or any other aspect of data activities)? PLEASE ENTER ONLY ONE ORGANIZATION OR ENTITY PER LINE and spell out its name

a. organization (name)

18. ORG_INTERESTS Thinking about the entities you just named (we can have qualtrics push this info forward from the last q into this one), in whose interest do you think they mostly act and make decisions about data

	In their own interest (to sustain the organization)	In the interest of their members	In the interest of science	In the interest of general public	Other
name 1					
…					
name X					

19. ORG_INTERESTS_OTHER question here to elaborate on options that selected “other” (question placed here for approval of codebook code)

20. ORG_DATADECIS What kind of data-related decisions have those entities made?
   a. Added data management section to their code of conduct
   b. Declared data sharing as a goal for the profession
   c. Requested journals in the domain to require publishing data
   d. Discussed or addressed compliance with the emerging or existing regulations (e.g., human subjects, EU GDPR)
   e. Other
   f. Don’t know

21. ORG_DATADECIS_OTHER question here to elaborate on options that selected “other”

22. ORG_BELONG To which professional organizations do you belong? [Please enter only one organization or entity per line and spell out its name]

23. ORG_INTERACT Currently, do you ever interact professionally with individuals in any of the following organizations (e.g., attend meetings, follow the news, contribute to discussions or work)?
   a. FORCE11
   b. Research Data Alliance (RDA)
   c. The Open Scholarship Initiative (OSI)
   d. The Committee on Data of the International Council for Science (CODATA)
   e. The International Council for Scientific and Technical Information (ICSTI)
   f. The Confederation of Open Access Repositories (COAR)
   g. The Scholarly Publishing and Academic Resources Coalition (SPARC)
   h. The Open Research Funders Group
   i. F1000research
   j. The Center for Open Science (COS)
   k. The World Data System (WDS)
   l. Coalition on Publishing Data in the Earth and Space Sciences
   m. Federation of Earth Science Information Foundation for Earth Science (ESIP)
   n. American Geophysical Union (AGU)

24. ORG_INTERACT_LEAD For those professional organizations to which you said you belong, are you currently or have you been involved in the leadership of those organizations?

25. ORG_BELONG_INV For those professional organizations to which you belong, how would you characterize your current involvement? [check all that apply]
   a. active (e.g., present at events, participate in discussions, part of leadership)
   b. observer (e.g., monitor news, attend events)
   c. not currently involved
   d. other (please specify)

26. ORG_BELONG_INV_OTHER question here to elaborate on options that selected “other”

27. ORG_DISCUSS *If a in ORG_BELONG_INV*, for those organizations how often do you discuss any aspect of data governance with any members of the organization? ORG_DISCUSS_THEY *If b in ORG_BELONG_INV*, how often have you heard or seen anyone discuss any aspects of data governance among any members of the organization?

a. about once a week or more
b. a few times a month
c. about once a month
d. about every other month or every quarter
e. 1-3 times a year
f. never
Other

28. IMPACT Are there other individuals, institutions or organizations in your professional area that impact the decisions you make concerning data? If so, please describe (text box)

29. IMPACT_OTHER What else besides improving how we manage and share data do you believe would help make scientific knowledge production more transparent and trustworthy?

30. COMMENTS Are there any other comments on this topic that you would like to share?

31. FOLLOWUP If you’re willing to participate in a follow-up interview, please leave your name and email (validate for email address)