

RESEARCH ARTICLE

Studying Data Communities: Analytical Dimensions from and for Empirical Research

Kathleen Gregory

Leiden University

Sarah R. Davies

University of Vienna

How can data communities be studied? In this paper, we combine existing literature with our own research spanning more than a decade to propose five dimensions and a series of guiding questions that can be used to analyze data communities. These dimensions—disciplinary domains, social orderings, material orderings, data uses, and data expertise—illustrate the variability of data communities and offer starting points for future empirical research, infrastructure design, and policy activities. Analyzing data communities through the multidimensional perspectives that we propose encourages critical interrogation of these heterogeneous groups while highlighting the complex ways in which data practices constitute particular communities and how those communities in turn shape data. Studying data communities with the questions we propose provides ways to honor epistemic diversity, experiment with community-centric infrastructures and policies, and create and value different forms of expertise.

Keywords: data communities; data practices; studying communities; data repositories; biocuration; research data

1. Introduction

The notion of communities has long been important in discussions of data production, management, and reuse. The notion is, however, often taken for granted, one-dimensional, or reduced to a disciplinary community (Borgman 2012). As research in science studies has shown, scientific communities are in fact complex networks involving social, technical, and material elements (Latour and Woolgar 1986; Kastenhofer and Molyneux-Hodgson 2021). Data—be they large or small or generated from instruments or from human observation—form a central part of scientific communities and their work (Borgman 2015). Communities that form around shared data do not have to be bound by disciplines or common research agendas; rather, these “data communities” are often heterogeneously structured groupings unified in some way by their collective dealings with shared data (Cavazzoni 2026). This heterogeneity in data communities raises questions: What are the different ways in which data communities can be structured? What dimensions shape how they form and evolve? How can data communities be studied and supported? The rise in data sharing and changes that shared (open) data have stimulated within various forms of research make answering these questions even more pressing.

In this paper, we draw on existing literature (e.g., Birnholtz and Bietz 2003; Leonelli and Ankeny 2015) and our own research (e.g., Davies 2025; Davies and Holmer 2024; Gregory et al. 2019a; Gregory et al. 2020; Gregory et al. 2024; Gregory et al. 2026) to explore five intertwined dimensions that we suggest can inform how data communities are analyzed. We illustrate these dimensions—disciplinary domains, social orderings, material orderings, data uses, and data expertise—and their variability using examples from our research with data communities. As will become clear, the majority of our work has examined data communities related to academic research. This previous work has informed the scope and framing of the dimensions that we propose. Nonetheless, we believe that several of these dimensions could also be applied to other communities outside of academia, as in the case of science journalists, citizen scientists, or even climate change skeptics. We therefore integrate such examples into our discussion as well.

Our intention in introducing these dimensions is not to provide a final or complete account of the nature of data communities, but rather to provide some conceptual tools with which to think about and study them. Each dimension offers insight into the heterogeneity of data communities and thus a starting point for future empirical research, infrastructure design, and policy activities. In proposing these dimensions, we aim to assist empirical exploration of how communities form and work together and of how different dimensions of those communities become (in)visible in data infrastructures and policies. Alongside other contingencies, we recognize that these dimensions are principally grounded in studies of Western science traditions and therefore may not travel well to other contexts. The dimensions thus signal, while not giving a final account of, the ways that data communities (and their associated practices) are not inevitable but could always be organized and instantiated differently.

2. Theoretical Background: Scientific Communities and Their Data Practices

To reflect on the nature of data communities, we draw on work from science studies that discusses scientific communities and suggests that such communities are central to knowledge production (Kastenhofer and Molyneux-Hodgson 2021). In this view, epistemic work is not located within individuals—and particularly not in “lone geniuses” or brilliant solitary inventors—but is reliant on networks of human and non-human actors (Latour and Woolgar 1986). Knowledge work is thus distributed across a community.

Importantly, knowledge-producing collectives are diverse with regard to their assumptions, practices, and approaches (Kastenhofer and Molyneux-Hodgson 2021). Karin Knorr Cetina (1999), for instance, documents the differences between the “epistemic cultures” of high energy physics and molecular biology, while Sharon Traweek (1992) discusses how even within a single discipline—in this case, physics—diverse national cultures may come to shape how knowledge production is carried out. As this example shows, such knowledge communities are not necessarily equivalent to scientific disciplines: There may be different epistemic cultures within disciplines as well as similarities between those working in different fields (Knorr Cetina 1999).

Shared practices, which can be thought of as sets of routine and learned behaviors, are central to membership in particular collectives (Bowker and Star 1999; Lave and Wenger 1991; Wenger 1998). (In the context of this article, we are specifically concerned with data practices—“data handling and dissemination” [Leonelli 2016, 1]—and how these constitute data communities.) These practices—and knowledge-producing collectives themselves—should be understood as *heterogeneous*. John Law uses the term “knowing spaces” to describe knowledge-producing collectives and to capture the various entities and processes that go into knowing:

Theories, methods, the empirical, modes of writing, disciplinary structures, audiences, authorities, and realities—all are staged together . . . knowing and its methods are materially complex and performative webs of practice that imply particular arrays of subjects, objects, expressions or representations, imaginaries, metaphysical assumptions, normativities, and institutions. (Law 2017, 47)

Within this conceptual framework, knowledge-producing collectives are not just different from each other but are internally heterogeneous in their composition, involving a wide array of human and non-human actants, entities, practices, and discourses. We share and build on this view, choosing to frame data communities as comprising not only a particular group of human actors but also the community’s material instantiations and the diverse entities involved in their shared practices. We therefore conceptualize data communities as dynamic collectives involving a range of different actants that are united through a set of shared practices and epistemic norms. Specifically, they can be understood as coalescing around what Sabina Leonelli and Rachel A. Ankeny (2015, 701) term “research repertoires,” where a repertoire is “a distinctive and shared ensemble of elements that make it practically possible for individuals to cooperate, including the norms for what counts as acceptable behaviors and practices together with the infrastructures, procedures, and resources that make it possible to implement such norms.”

Similarly, Jeremy P. Birnholtz and Melissa J. Bietz (2003) propose that data both define boundaries between people using different methodological approaches (e.g., theoretical vs. empirical approaches) and enable access to those communities. Particularly relevant to our discussion is their idea that individuals participate in communities at different levels of depth. Being responsible for managing, analyzing, and documenting data enables some individuals to pursue “inbound trajectories,” where they move from participating on the periphery of a community to becoming core members. In areas where this type of participation is not supported—for example, where individuals have limited responsibility for data—these inward trajectories are not as possible. Data—and the practices and materialities surrounding data—thus enable

different types of participation in scientific communities. Frameworks for data sovereignty, such as the CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, and Ethics) (Carroll et al. 2023), also point to the role that data can play in creating boundaries, highlighting the need to be aware of who is allowed to participate in data governance and particular communities.

In other literature, a common starting point for thinking about data communities is the data repository or the database. Danielle Cooper and Rebecca Springer (Cooper and Springer 2019; Springer and Cooper 2020) suggest that establishing data repositories can stimulate the formation of discipline-spanning data communities that exchange data with each other. Questions of scale quickly become a consideration in such contexts (Cooper and Springer 2019), as it can be difficult to scale up or down services to match the size of particular communities. Christine L. Borgman and Paul Groth (2025) propose a scale of a different sort in their discussion of how the distance between data creators and consumers influences the usability of data. Using the metaphor of distance, the “closer” that data reusers are to data creators in aspects such as disciplinary domain, methodological approaches, and intended tasks, the easier it is to reuse others’ data. This perspective allows us to think about potential communities of data reusers in terms of their similarities and relationships with data creators as well as with other people reusing the same data.

Finally, work on “data cultures” is also of interest to our discussion of data communities. Drawing on the work of Kim Fortun (2009), Lindsay Poirier and Brandon Costello-Kuehn (2019) propose a heuristic for researchers to use to question their own data sharing practices. This heuristic is based on seven different “scales” or frames of reference and proposes questions at the macro level, which attends to financial and legal structures; the meso level, which explores organizational factors; and the micro level, which focuses on research practices and customs. Other work has applied this heuristic to identify salient lenses for studying data cultures, including data-related skills and attitudes, data sharing and data use/reuse practices, data ethics and governance, and Indigenous perspectives (Oliver et al. 2023).

In sum, while scholarship has discussed various forms of collectivity around data, from data cultures to users of particular data resources, the notion of data communities has not been systematized (Cavazzoni 2026). Based on the insights described above, in this article we define these communities as dynamic collectives united through a set of shared practices, epistemic norms, and relationships to data. This notion of data communities thus offers a means of taking seriously the simultaneous collectivity and heterogeneity of such collectives and as such adds to existing concepts such as data cultures or data creators/reusers.

3. Approach: How Can We Analyze Data Communities?

The work discussed above offers a variety of insights into the nature of data communities. Based on the concepts and frameworks that it proposes, we understand data communities as *heterogeneous* and *practice oriented*. Our approach is thus located in a constructivist theoretical framework that understands data and data practices as relational and contingent rather than stable or final (Law 2017; Leonelli 2016; Leonelli and Ankeny 2015). As noted above, scholarship on data communities has only minimally been systematized. Our own experiences with researching data communities suggest that it would be helpful to have a more systematic approach with possible frameworks describing how data communities might be encountered, researched, and (to some extent) categorized.

In this article we attempt to provide such a framework, combining knowledge and theory from the literature regarding the nature of data communities with our own ongoing and previously published research. We outline five dimensions through which to examine the structure of data communities; together, these dimensions offer a potential framework for analysis. These dimensions emerge from a combination of our reading of the research literature and our collective fieldwork experiences and analyses. Much of this work has been published elsewhere¹ (e.g., Davies 2025; Davies and Holmer 2024; Gregory et al. 2019a; Gregory et al. 2020; Gregory et al. 2024; Gregory et al. 2026) but collectively involves qualitative research exploring data reuse, data citation, data visualization practices, and lived experiences of data work. We are thus drawing on a series of empirical interventions that took different forms (as will become clear when we discuss moments from them in more detail) but that all sought to explore the nature of data work and its relation to different forms of collectivity. Specific methods used include digital ethnography, semi-structured interviews, participatory workshops, and surveys and questionnaires. Based on these long-term experiences researching data communities (some fifteen years between us), we share a sense that it would be useful to offer general approaches that might be used in such scholarship—to develop conceptual tools and starting

¹ All studies that we refer to in this paper received ethical approval according to the guidelines of the institutions where we were employed. Participants provided informed consent prior to participation for all interview and survey studies; workshop participants reviewed the material that we incorporate here and have consented to this use.

points for analysis that could be mobilized when exploring data communities of different kinds. Such a framework might also assist in the practical questions that exist in creating and maintaining data infrastructures. In line with our qualitative, constructivist approach, we do not see this framework as final but as a starting point for further research.

The five dimensions that we present in this article as an answer to this need (disciplinary domains; social orderings; material orderings; data uses; and data expertise) were developed through a series of iterative conversations between the authors that focused on (1) key themes from the research literature and (2) central patterns in our empirical fieldwork. Collectively, and with reference to these sets of materials, we sought to examine how data communities are structured and the dimensions along which this can vary. We discussed research findings (including those outlined above), the results of our own studies, and the “ethnographic hunches” (Pink 2021) that emerged from our knowledge of particular data communities in order to cluster key themes and ideas and thus to iteratively develop dimensions that are central to encountering such communities. Our discussions, notes, and eventual concretization of the dimensions were oriented to the ways in which data communities can vary; as such, each dimension offers insight into the heterogeneity of data communities by providing a possible axis around which communities can be oriented. However, these dimensions should be seen as fluid, overlapping dynamics that can be used to ask questions about data communities rather than as a set of rigid variables along which any community can be situated. Communities may, for example, sit at multiple points along a particular dimension when looking at different moments or aspects of their practices. The dimensions that we have identified thus act as a starting point for those interested in analyzing or comparing particular data communities rather than as a final statement regarding their nature.

Because our aim in systematizing literature and empirical findings in this way is to offer a framework for those interested in analyzing data communities, we briefly describe each dimension but also propose a series of questions and foci that can be taken as a starting point for analysis. Due to the nature of our own research, many of these dimensions and questions are oriented to data communities within academia. Mirroring the way in which the dimensions emerged from our conversations and iterative analysis, we combine arguments and findings from the literature with moments and results from our own research as we outline the dimensions. In section 4, we integrate examples and vignettes from across different communities to illustrate both the diversity of data communities and the breadth of the dimensions. We also highlight one specific community we have studied, biocuration, where biocuration is “the translation and integration of information relevant to biology into a database or resource” (see discussion in Davies 2025, 2). The example of biocuration illustrates the reflections informing this framework from the perspective of a particular data community (Table 1). The vignettes, examples, and empirical discussions are not used to answer questions that emerge from the dimensions but to illustrate how they have prompted our call for attentiveness to them. In addition, in what follows we use “we” to refer to the collective scope of our work, although these studies were conducted with different collaborators and both authors were not always involved. We do this for ease of reading and to signal that the analysis presented here—even when based on empirical research we may have carried out separately—is a collective endeavor.

4. Dimensions for Analyzing Data Communities

As described above, we identified five dimensions that our reading and empirical experiences suggest capture central ways in which data communities can vary. We outline these in the sections below. The dimensions and the central questions that they raise are also summarized in Table 1. The final column of this table summarizes how our study of one particular data community, biocurators, relates to each dimension, illustrating how the questions can be used to explore specific empirical cases. In addition, examples from other data communities are integrated into the main text.

4.1 Disciplinary Domains

Extensive work has documented the importance of disciplinary differences in data practices (e.g., Borgman 2015; Khan, Thelwall, and Kousha 2023). Indeed, discussions of data communities are often framed as discussions of disciplinary groups, be they broad or more specific (Borgman 2012; Gregory et al. 2019a; Cooper and Springer 2019). Despite this, notions of “discipline” are rarely questioned, and the term is often used in a relatively homogenous fashion to refer to groups of people (e.g., biologists) or to data themselves (e.g., biomedical data). Less visible in these framings are the diverse and complex roles that disciplines play in organizing and shaping social groups as well as data and other content (Hammarfelt 2019). With this dimension, we aim to question how we think about disciplinary domains when studying data communities. We encourage thinking about how the framing of disciplinary domains can be used to highlight epistemic

Table 1. Central questions and research foci for each dimension

Dimension	Central research interest	Specific research foci	Reflections from biocuration data communities
Disciplinary domains	How can disciplinary domains be critically studied?	What are the actual practices and friction points of researchers within disciplinary groups? What classifications are used to define disciplinary groupings, and how have these classifications come to be?	Biocuration involves interdisciplinary work and skills from different fields, such as computer and information science, as well as in-depth life science knowledge. The biocuration community includes those with backgrounds as life scientists, computer scientists, and others. Frictions may arise from different disciplinary assumptions, but biocurators come together around a shared vision of rendering biodata accessible.
Social orderings	How are particular communities organized?	What size data communities do people belong to? How strong is their sense of belonging to these communities? How do communities of different sizes and orderings intersect with and shape data practices? At what scale are data communities imagined? How do such imaginings match reality, and what are the effects of such imaginings?	In our fieldwork, it became clear that there are multiple communities at stake within biocuration work. Biocurators often talk about “the community,” meaning not their immediate community of other biocurators but the user community whom their work serves. At the same time, they are also part of intersecting professional communities directly oriented to biocuration, such as collectives around particular model organism databases or the International Society for Biocuration.
Material orderings	How are particular communities constituted and materialized?	How are the data at stake in a particular community materialized? What infrastructures are necessary for the data community to function? What kinds of labor are necessary to the data community, and who or what performs this labor?	Biocuration involves facilitating the sharing of biological data as well as curating and annotating them in order to add value. One aspect of this work is the use of bio-ontologies to organize and categorize datasets. Such ontologies are central to how data are organized and made available. They also codify particular conceptualizations of current knowledge in the biosciences. Bio-ontologies are thus one example of how a particular ordering system can structure data, rendering the information the system captures thinkable and usable in specific ways.

Table 1. (Continued)

Data uses	How are data used, and how do common uses in turn shape data communities?	For what purposes are data communities using data? How are communities interacting and engaging with data? How are data uses (and the involved communities) made visible or invisible?	Biocurators do not often use data to answer new questions or as input for models (as some researchers do). They do, however, use data as they examine datasets in curation workflows, identify missing values or discrepancies, and interact with data through annotation processes. Like other information professionals and research support staff, it is likely that they also use data to train others or in their own education. Like much curation work, these types of data uses are often invisible.
Data expertise	How is expertise instantiated, performed, and contested?	What forms of expertise are present within a particular community, and where are these located? How is expertise assessed and valued? Is expertise questioned and contested within a data community, and if so, how?	The International Society for Biocuration's generic job description notes that "the best biocurators are detail oriented, conscientious, and good communicators. They are adaptable to the needs of the community and/or to the needs of the software systems" (International Society for Biocuration 2026). Subject matter expertise must therefore be combined with other kinds of expert knowledge and practices, including not only understanding of specific digital platforms but also soft skills such as being able to work well in a team. The extent to which individual curators must have both domain expertise and expert knowledge of curation is a subject of discussion.

diversity within fields rather than to obscure it. We propose three possible foci to guide studying data communities through this lens.

First, we suggest that close examinations of *actual practices* can be used to identify epistemic diversity and the plurality of communities to which individuals belong. Questions guiding this line of analysis could be: What are the friction points individuals encounter with more generic or idealized domain policies, data standards, or expectations? Where are there similarities? What variation exists within data communities within a single domain? Such questions do not start with external classifications of disciplinary fields but, rather, are grounded in practice. This approach allows us to understand how individuals themselves think about the role of disciplines in their data practices.

These questions are rooted in our own fieldwork. We have observed many friction points that researchers encounter when working with data, some of which are due to differing norms and standards of different data communities within the same disciplinary domain. For example, one researcher whom we spoke with belonged to multiple data communities in the life sciences at the same time (Gregory et al. 2019b). He worked as a clinician at a hospital in Europe while also belonging to a research group in the United States, where he developed new technologies for fusing spinal cords. Each of the different communities he was a member of used data for different purposes and had different policies and attitudes to making data available

for others. The practices and norms of these communities were at times in conflict, and the researcher needed to adapt his individual data practices according to the particular data community within which he was operating at a given moment. Although both of these communities were situated in the “life sciences,” in one case, he would share data, while in the other, data were a carefully guarded secret. We see in this example that focusing on the actual practices and friction points that researchers encounter within disciplinary groups helps to make the diversity within domains more visible.

Second, we propose to question the classifications used to discuss disciplinary groupings. Disciplinary classifications are critical to organizing information/data; they are also far from neutral, shaped by the world views, political discourses, and social contexts in which they are created and applied (Bowker and Star 1999; Poirier 2023). When studying data communities through a disciplinary lens, we can ask questions such as: What classifications are used to define disciplinary groupings, and how have these classifications come to be? What differences exist in how they are defined and applied? How do these differences influence how data communities are conceptualized?

Disciplinary classifications are perhaps especially important in framing how data repositories and database developers conceptualize the “designated communities” (Borgman 2012) for their services. As researchers interact with repositories—by depositing their data or when searching for data to reuse—they are often confronted with such classifications. When sharing data, for example, researchers choose how to label the “subject” or “discipline” of the data they deposit. It is not always clear, however, what these classifications (should) represent—for example, the aboutness of the data, the expertise of the creators, or imagined future audiences. Disciplinary classifications become associated not just with the data themselves but also with different data communities.

We have seen this association in our own examinations of the data deposit processes at two large data repositories. Acting in the role of data sharers, we attempted to deposit our own data. In one repository, we were confronted with a dropdown menu of forty-two sub-categories of six research fields to indicate the “research domain.” There was no further explanation provided about the meaning of this field, although we assumed that we should use it to indicate what our data were about; in this case, we selected the field “Other social sciences.” In the second repository, we encountered a different, more detailed controlled vocabulary describing disciplinary domains. Here, however, the disciplinary classification appeared in the “Audience” field, and we were instructed to use the classification to describe the anticipated “scientific disciplines to which the data may be relevant.” This instruction made very clear how the classification applied to both the content of the data as well as an imagined community of future data reusers.

4.2 Social Orderings

Communities are ordered according to different social and organizational structures. These structures are influenced in part by the size of a community as well as the sense of belonging that members feel to a particular group (Cohen 1993). The second dimension for analyzing data communities that we propose therefore highlights the ways in which data communities are organized and imagined through social interactions and connections at different scales. It draws in particular on the ideas that members of scientific communities participate to different degrees and intensities (Birnholtz and Bietz 2003) and that research is shaped by broad norms as well as highly local contexts (Knorr Cetina 1999). Rather than using scale to refer to the sociotechnical elements shaping data practices at the macro, meso, and micro levels (as in Poirier and Costello-Kuehn 2019), we focus instead on the different sizes and orderings of data communities themselves. While the size of a data community is an important aspect of this conceptualization, it is equally important to think about the strength of individuals’ feelings of belonging to particular data communities and to consider how these feelings of identity shape practices. In this way, our conceptualization is more similar to the metaphor of “distance” between data creators and data reusers proposed by Borgman and Groth (2025).

The first focus we propose starts with two questions: What size data communities do people belong to? How strong is their sense of belonging to these communities? This line of questioning encourages us to move beyond thinking about merely the size of a data community but also about its composition, which can influence everyday practices. We have seen many examples in our work of how the strength of an individual’s feeling of belonging to different communities with different social orderings can influence data practices. In workshops in the life sciences (Gregory and Koesten 2023), for example, researchers commonly shared data scientist-to-scientist within their universities in very local data communities; sharing data with and reusing data from scientists who were not as “local” (e.g., from a different department or a different city) was considered more challenging.

Another example from our work demonstrates the influence of broader communities, where creating something *useful*—be that a code package, a database, or an instrument—for a community of individuals working with a

similar methodology served as a motivation for open science practices (Gregory et al. 2026). One of our interview participants regularly shared data, as well as code, in online repositories. During our conversation, they reflected on why they did this, as they worked in a very small area in which only a handful of people were active, which made it very easy to trade data through personal exchanges. Instead, they were inspired to share code/data packages in repositories as a public service to the larger scientific community as a form of repayment (they had learned to program by looking at open code) and as a way of “drumming up business” for their area of research. Here, this researcher felt a duty to a broader community but was also motivated to strengthen their own local community.

This example also introduces our second question: How do communities of different sizes and orderings intersect with and shape data practices? Communities at different scales always exist in relation to each other. Questioning how they intersect and shape each other allows us not only to explore the plurality of data communities to which individuals belong but also to better understand synergies and tensions that exist between communities. This question allows us also to think about the hierarchies and other forms of organization within which data communities of different scales are situated as well as the role of standards and shared documentation in moving across these scales.

One example of the need for this line of questioning is seen through our work with biocurators (see Davies 2025; Davies and Holmer 2024). In fieldwork, we observed that the notion of “community” is itself not straightforward; rather, it became clear that there were multiple “communities” at stake within biocuration work. Biocurators often talked about “the community,” meaning not their immediate community of other biocurators but the user community whom their work served. At the same time, they were also part of intersecting professional communities directly oriented to biocuration, such as collectives around particular model organism databases or the International Society for Biocuration. The same data or data practices might be at the heart of these different communities, but they are framed in different ways and have different means of rendering data and data work (in)visible. Biocurators were often deeply committed to “their” user community and saw this as the key audience for their work; at the same time, they found that their work was often overlooked or invisible within these communities.

Third, we propose asking: At what scale are data communities imagined? How do such imaginings match reality, and what are the effects of these imaginings? This question is perhaps particularly relevant from a policy or infrastructure perspective, with repercussions for the design and investment in, for example, a data repository or database. We illustrate the effects of scale by contrasting two very different types of data infrastructures and their imagined audiences. [GenBank](#), the genetic sequence database operated by the National Institutes of Health in the United States, “is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information”; it receives data from a variety of data sources, including researchers. The imagined community of users in this case can be thought of as anyone in the broader scientific community who could be interested (and able) to use sequence data. It is also designed to make data mobile so that they may be used in many contexts. This approach contrasts with the imagined communities of smaller, more bespoke infrastructures—for example, institutional repositories designed to support the unique needs of universities (Darragh et al. 2024) or even non-digital infrastructures developed by local communities as forms of resistance to commercialism, which keep information within communities rather than making it globally mobile (Hobbis and Hobbis 2022).

4.3 Material Orderings

The third dimension helps to highlight the materiality of both data and the communities associated with them, encouraging us, as analysts, to explore the material orderings of particular data communities. Based on our understanding of communities as heterogeneous, composed of both human and non-human actants, examining this dimension leads us to ask how particular communities are constituted and materialized. What different forms do data and their communities take, and how do these forms play a role in shaping those data, the journeys that they can take, and the nature of data communities?

We see three foci for explorations of material orderings. First, how are the data at stake in a particular community materialized? This question encourages us to examine the material dimensions of data production and storage, including technical aspects such as file format, size, data type, or the nature of transitions between different types of data. For example, exploring this dimension might lead us to look at where data come from, how they are transformed or worked with, or what is lost or gained as they move between different formats. This consideration is important because the production of data, and their transition from specific material trace to digital signal, is notoriously complex (Metzler, Ferent, and Felt 2023). Often, tacit knowledge of a particular experiment is necessary for making sense of the data produced from it and for any attempts at reproducibility. We observed this firsthand in workshops with researchers centering on data practices (Gregory and Koesten 2023). One researcher explained that, in biochemistry, if they are making a solution, it matters how deeply they put the pipette into the flask and if it was held at eye level or not. Those

standardizing or sharing such data must therefore find ways to account for these specificities in working with data or risk flattening the nuances of experimental evidence.

Second, what infrastructures are necessary for the data community to function? This question encourages us to look at the physical spaces, equipment, technologies, and resources that are put into place in concert with the formation of a particular data community. It is closely linked to the first question in that it is concerned with the materialization of data, but it extends this idea to attend to the infrastructures and resources that are necessary to handle them. For instance, exploring this dimension might lead us to examine what infrastructures (such as energy or IT) are necessary to the data community, what software or other tools are used in working with data, how databases or other “homes” for data are structured, what physical sites are important to the community, or what the geographies and global resources requirements of a data community are.

As an example, biocuration involves not only facilitating the sharing of biological data but also curating and annotating it in order to add value to it (Holinski et al. 2020). One aspect of this work is the use of bio-ontologies to organize and categorize datasets (Bodenreider and Stevens 2006). Bio-ontologies render datasets computable and usable by mobilizing “a model of a portion of (a conceptualization) of reality” (Bodenreider and Stevens 2006, 257). Such ontologies are thus central to how data are organized and made available, but these ontologies also codify particular conceptualizations of current knowledge in the biosciences, relying on discussion and updating of consensus knowledge (Lean 2021). Bio-ontologies are thus one example of how a particular ordering system can structure data, rendering the information it captures thinkable and usable in specific ways.

Furthermore, as with other data communities, biocuration involves different forms of labor—curators might work as paid staff on particular projects or databases, volunteer their curation work or carry it out in their free time, or carry out curation as part of university assignments. Importantly, it seems clear that biocuration involves care work and emotional labor as well as—or entangled with—specific data-related tasks. Ane Møller Gabrielsen (2024, 269) argues that biocuration is frequently framed as care work and that “ideal biocurators are . . . described as caring in the affective and nurturing sense,” while Leonelli has suggested that the field is oriented around a service ethos (2016). Biocuration is understood as requiring careful attention to detail and a desire to offer a “public service.” As such, working within this domain is not only about a particular set of technical skills but also about having the capacity and interest to care for data.

This observation leads us to a third central question: What kinds of labor are necessary to the data community, and who or what performs this labor? This question draws attention to the material conditions of data work and workers and to the ways in which data communities realize particular lived experiences or forms of life. It extends an interest in the material conditions of data to the lives of those who work with them, for instance by asking what actors are involved in data work, what degree of recognition or reward they receive, how they experience and frame the data community, or what different kinds of labor are present. This aspect of material orderings thus draws attention to the way in which data communities—even those that may be run as volunteer groups—involve diverse kinds of work, which may or may not be financially rewarded or widely visible.

4.4 Data Uses

This dimension draws on work proposing that communities—across different social orderings and domains—form among people who are using data for similar purposes and tasks (Gregory et al. 2020; Koesten et al. 2017). Often, these groups make use of similar methodologies, as in disciplines such as life sciences, data science, and digital humanities (Leonelli and Ankeny 2015; Levallois et al. 2013). Exploring data communities through this dimension allows us to investigate how data are *used* within communities and to question how these common uses in turn shape data communities.

As before, we propose three foci to guide analyses of data communities through this lens. First, we can ask: For what purposes are communities using data? How are they mobilizing data in their workflows? This line of questioning helps us to explore common uses, needs, and groups that form around particular ways of working with data. Data uses are multiple and rooted in context; communities may use data for teaching purposes, for validating results, or for more “social” reasons, such as using data as a way of connecting with potential collaborators (Gregory et al. 2019b).

Data uses can also serve as centering points for communities inside and outside of academia; some of these uses can be quite similar to each other, although the communities themselves differ. Both anthropogenic climate change skeptics and climate change researchers, for example, make use of data in order to support and lend credence to their arguments (Wofford and Thomer 2023). Data journalists and researchers also have similar ways of working, from hypothesis-driven approaches—where data are used to either support or counter existing reports—to data-driven approaches, where data are explored to answer individual

research questions (Parasie 2015). We observed these similarities as well in our own study of how data visualizations are created in popular science magazines (Gregory et al. 2024). Particularly in data-driven work-flows, both data journalists and researchers rely on collaborations and communications to access and make sense of data, to check for errors, and to represent and visualize data accurately. Scientists themselves also temporarily become active in journalistic communities as they work with journalists to create visualizations and stories highlighting their scientific research.

The second question we propose to explore is: How are communities interacting and engaging with data? Here, we conceptualize interaction as the more detailed, granular activities that people perform as they work with data: questioning and discussing particular values or methods, annotating sections of text, or working together to document data provenance, for example. These types of interactions are bound up with the material orderings of data and collaborative work. Exploring this question helps us to pay attention to how use, materialities, and communities intersect, and to learn more about the detailed practices grounding data communities. The development of data platforms that support different types of collaborative work make these intersections visible.

[Kaggle](#) is an example of one such platform, where data, code, tools, and communities are brought together. As described by Koesten and colleagues (2025), people interact around shared data on Kaggle as they work together to solve set challenges or competitions. They also hone their data science skills by posting questions and answers in discussion forums. Rather than needing to download data locally, people use and interact with data on the platform, where they can experiment with data spontaneously or discuss data with different people. The structure of Kaggle supports and makes visible certain ways of using and interacting with data, ways which in turn shape how communities form on the platform.

Our third line of questioning involves paying attention to how data uses (and the communities involved) are made visible or invisible. The previous example of Kaggle demonstrates the visibility of certain types of uses, interactions, and communities on a data science platform, but, we argue, there are also data uses and communities that will remain invisible and more difficult to see—for example, “lurkers” or those who choose not to engage with certain data or technologies (Wyatt 2003).

It is also difficult to identify data uses (and communities) in formats such as journal publications or in magazines. Not all purposes for using data will be captured in a publication (Federer 2019), and there is great heterogeneity in terms of how researchers reference data when they use them (Gregory et al. 2023; Robinson-García et al. 2016). While there have been efforts to develop schemas for recognizing co-authorship of datasets or tools for recognizing diverse forms of data work (Wood-Charlson et al. 2022; Zeng et al. 2020), such taxonomies have thus far had limited adoption in practice. Existing forms of recognizing data use therefore often obscure the groups of people who are involved in working with data. Questioning who is visible or invisible in traces of data uses helps us to attend to the composition of the community as well as the labor that is involved in data work.

Our work with data journalists (Gregory et al. 2024) provides an example of how current modes of recognizing data use do not adequately capture the work of a larger group of people. Data visualizations in the popular science magazine we studied are usually accompanied by a citation to a data source—for example, an academic article or a government database—and a byline crediting the work of the person who created the data visualization. Our analysis revealed that such visualizations were rarely, if ever, created by a single person, but were instead the result of many collaborations between fact checkers, other graphic designers, writers, editors, and scientists. One participant reflected that, ideally, they would be able to recognize the diversity of work involved in creating data visualizations by adding a long note explaining who had been responsible for each task.

4.5 Data Expertise

The final dimension we propose for thinking about data communities is expertise. This dimension allows us to explore how specialist knowledge and skills are negotiated within particular communities and who is understood as possessing this expertise. In this regard, there are some commonalities with disciplinary domain, but we view this dimension as highlighting what are understood to be expert practices rather than disciplinary affinities. In discussing expertise, we draw on theorizations of the term that view it not as something static or straightforwardly possessed by individuals with particular credentials but as something relational and performed (Egher 2023; Grundmann 2017; Hilgartner 2000). Data expertise, and who has it, is thus not something to take for granted but a focus for exploring how particular skills or knowledges are framed and valued within a community. In exploring data expertise as a dimension of data communities, we can ask: How is expertise instantiated and performed, and how is it contested?

Again, we propose a number of foci to guide exploration of these overarching questions. First, what forms of expertise are present within a particular community, and where are these located? This question

encourages us to explore the kinds of skills, knowledge, and experience present within a community as well as how these are talked about and framed. What professional or other kinds of expert knowledge and practices are presented as necessary for participation in the community? What kinds of expertise become more or less visible at particular moments? How are different kinds of expert knowledge and practices distributed across a community, and to what extent are these outsourced to other sites? Such questions draw attention to exactly what kinds of expertise are understood as necessary to data communities and to the ways in which this expertise is located and patterned.

For example, biocuration is a highly skilled form of data work, and most biocurators have PhDs in the biosciences and postdoctoral work experience in research laboratories (Burge et al. 2012). As this pattern suggests, domain expertise is often highly valued, with curators using their in-depth knowledge of an area of science to help them as they curate datasets and literature in that area. At the same time, it is clear that biocuration requires additional skills to those used in lab research. The International Society for Biocuration's generic job description notes that "the best biocurators are detail oriented, conscientious, and good communicators. They are adaptable to the needs of the community and/or to the needs of the software systems" (International Society for Biocuration 2026). Subject matter expertise must therefore be combined with other kinds of expert knowledge and practices, including understanding of specific digital platforms, and soft skills like the ability to work well in a team. The extent to which domain expertise and expert knowledge of curation must both be present within individual curators is a subject of discussion. "Community curation" asks lab researchers to become more involved in curating their datasets, but researchers may not always have all the necessary experience and knowledge regarding curation that a professional curator would bring (Arnaboldi et al. 2020).

Second, how is expertise assessed and valued? This question draws attention not only to the different forms and distribution of expertise in a particular data community, but also to the value judgments that are applied to expertise. This question allows us to explore which forms of expertise are more or less appreciated as well as how some forms of expertise may be so devalued that they are rendered invisible. It thus signals the way in which data communities are ultimately evaluative spaces, which are oriented to particular ends and which will apply particular forms of (implicit and explicit) assessment regarding how to meet those ends.

Indeed, there has been much discussion in European science policy about the need to value different types of expertise and research outputs. Some of this discussion focuses on professionalizing certain data-related roles—for example, data stewardship (Jetten et al. 2021)—while other work advocates for data themselves being valued as "first-class research objects" (e.g., Wilkinson et al. 2016). In recent interviews with researchers (Gregory et al. 2026), we found that different types of data work and expertise (as well as data themselves) are valued in different ways. Some participants strongly believed that data should be seen as equal in value to publications in research assessments, while others reported a methodological divide in how data expertise is understood and valued. One computational researcher described how researchers using more qualitative methods simply do not regard the work needed to prepare and curate large datasets (which takes up 75 percent of their time) as being "on the same level" as other types of research in the field.

Finally, we can ask if and how expertise is questioned and contested within a data community. This question draws attention to the (in)stability of particular forms of expertise and to the ways in which it may be challenged. While some data communities may possess a widely shared consensus regarding which kinds of expertise are important and how they should be applied, others may be more fractious or have a less stable set of norms and assumptions. This question thus encourages us to explore moments of tension, disagreement, and difference regarding best practices for the data work at stake and to understand which actors are connected to different positions within such debates. For example, heterogeneous data communities that converge around particular, and sometimes contested, sets of data or findings offer a key example of how contestation and difference can help to identify what is at stake within a particular community. Gwen Ottinger (2010, 2022, 2023) has extensively documented the activities of environmental activists as they engage with data produced by regulatory scientists while simultaneously producing their own datasets. Activists' need to contest government data is urgent, Ottinger writes, because "the epistemic resources that dominate environmental regulation are especially ill-suited to capturing the experiences of the communities most exposed to pollution" (2022, 510). The lived experiences of community members and activists, which might include accounts of changes in odor in their environment or in the taste of their drinking water, are generally not captured by official measurements; activists, then, develop their own measurement tools and compile their own datasets. To activists, regulatory expertise is inadequate because it functions within a narrow epistemic framework that cannot capture their concerns; to regulators, on the other hand, activists are viewed as having inadequate expertise with regard to assessing environmental pollution (as Ottinger writes, "regulators and refinery representatives . . . tended to participate in conversations with community members and activists by instructing them on the 'right' ways to think about pollution and air

quality” [2023, 11]). Activists and regulators thus fundamentally disagree about what kinds of expertise should be applied in data collection and interpretation.

5. Conclusion

The five dimensions described above can be used as starting points and frameworks to guide empirical research into data communities. For instance, we might imagine comparative research that looks at how particular dimensions are instantiated across or within particular data communities; in-depth case studies that seek to describe specific data communities through one of the five dimensions; or examinations of particular data communities that further expand scholarly understanding of one of the dimensions (by developing conceptualizations of data expertise through empirical engagement with its realization and contestation, for example).

Taken together, the dimensions also point to a number of common themes and ideas that span these perspectives: honouring epistemic diversity, infrastructure and policy design, and crediting and valuing different forms of expertise.

5.1 Honouring Epistemic Diversity

Across dimensions, studying data communities with the questions and sensitivities we suggest surfaces epistemic diversity rather than obscures it. Paying attention to the plurality of situations and environments that shape data communities and practices—including but not limited to data infrastructures, ontologies, social hierarchies, and uses—emphasizes this diversity. Our analysis also highlights that data communities are multiple and overlapping, with individuals (be they researchers, journalists, or biocurators) belonging to numerous communities at different times. These intersecting and overlapping boundaries are themselves a key part of understanding the formation of data communities.

5.2 Infrastructure and Policy Design

Our analysis also surfaces issues related to the design of data infrastructures and the creation of policies around open science and data management. We have seen throughout the five dimensions that communities are diverse and that they shape and are shaped by various materialities, from data formats to ontologies to repositories to infrastructures. We have also seen that perceptions of a data community (e.g., its disciplinary composition or its envisioned size) influence design choices when building infrastructure. A central challenge for both infrastructures and policy makers is to balance supporting diversity with the need to identify broad patterns or to navigate the tension between local and global communities and practices. While this tension is challenging (and not unique to the realm of open data), it could also be a chance for experimentation—for example, by developing shared data spaces for collaborative projects of varying sizes within a repository, allowing for different forms of “use” (e.g., interacting with data in community forums), or creating policies according to methodologies rather than disciplinary domains. The idea of data communities also points to the importance of bringing together actors with different expertise (curators, repository managers, policy makers, researchers) around infrastructures and various policy instruments—not only to inform their development, but also to further the creation of other forms of data communities.

5.3 Crediting and Valuing Different Forms of Expertise

Data expertise can serve as an entry point to deeper involvement in certain communities (Birnholtz and Bietz 2003), but, as seen in our examples, expertise is situated and valued differently in different data communities. Our analysis also highlights that much data work involves hidden labor that is not always valued within academia (Davies and Holmer 2024). Questioning the social and material orderings of data communities can help to reveal such labor; it can also lead to understanding the value that different forms of data expertise provide—and the values underlying various forms of data work. Identifying these underlying values could be a starting point for thinking about how to recognize and reward different forms of data expertise within academia (Gregory et al., 2026; UNESCO 2021).

Analyzing data communities through the multidimensional perspectives that we propose serves as a means of avoiding flattening communities into homogenous groups or blithely using the term without reflection. Questioning and studying the heterogeneity of data communities through these perspectives facilitates interrogation and exploration, adding depth and nuance while highlighting the complex ways in which data practices constitute particular communities and those communities simultaneously shape data.

CRedit

Kathleen Gregory and Sarah R. Davies contributed to the conceptualization, formal analysis, investigation, methodology, writing – original draft, and writing – review and editing.

Acknowledgements

We are extremely grateful to the many interlocutors in diverse data communities who have participated in our research over the last years and otherwise offered feedback and collaboration, and whose ideas and experiences have informed this article. In addition, we want to acknowledge our other colleagues, support staff, and students at the Department of Science and Technology Studies, whose work supports ours in multiple ways.

Competing Interests

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Arnaboldi, Valerio, Daniela Raciti, Kimberly Van Auken, Juancarlos N. Chan, Hans-Michael Müller, and Paul W. Sternberg. 2020. "Text Mining Meets Community Curation: A Newly Designed Curation Platform to Improve Author Experience and Participation at WormBase." *Database* 2020: 1–16. <https://doi.org/10.1093/database/baaa006>.
- Birnholtz, Jeremy P., and Matthew J. Bietz. 2003. "Data at Work: Supporting Sharing in Science and Engineering." In *GROUP '03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*. Association for Computing Machinery. <https://doi.org/10.1145/958160.958215>.
- Bodenreider, Olivier, and Robert Stevens. 2006. "Bio-ontologies: Current Trends and Future Directions." *Briefings in Bioinformatics* 7 (3): 256–74. <https://doi.org/10.1093/bib/bbl027>.
- Borgman, Christine L. 2012. "The Conundrum of Sharing Research Data." *Journal of the American Society for Information Science and Technology* 63 (6): 1059–78. <https://doi.org/10.1002/asi.22634>.
- Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press. <https://doi.org/10.7551/mitpress/9963.001.0001>.
- Borgman, Christine L., and Paul Groth. 2025. "From Data Creator to Data Reuser: Distance Matters." *Harvard Data Science Review* 7 (2). <https://doi.org/10.1162/99608f92.35d32cfc>.
- Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. The MIT Press. <https://doi.org/10.7551/mitpress/6352.001.0001>.
- Burge, Sarah, Teresa K. Attwood, Alex Bateman, Tanya Z. Berardini, Michael Cherry, Claire O'Donovan, Ioannis Xenarios, and Pascale Gaudet. 2012. "Biocurators and Biocuration: Surveying the 21st Century Challenges." *Database* 2012: bar059. <https://doi.org/10.1093/database/bar059>.
- Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2023. "The CARE Principles for Indigenous Data Governance." In *Open Scholarship Press Curated Volumes: Policy*. <https://openscholarshippress.pubpub.org/pub/xx3kj9rv>.
- Cavazzoni, Emma. 2026. "Data-Technology Communities: Collaboration and Diversity in Data- and Technology-Intensive Multidisciplinary Research." *BioSocieties* (January). <https://doi.org/10.1057/s41292-025-00379-w>.
- Cohen, Anthony P. (1985) 1993. *The Symbolic Construction of Community*. Key Ideas. Routledge.
- Cooper, Danielle Miriam, and Rebecca Springer. 2019. "Data Communities: A New Model for Supporting STEM Data Sharing." *Ithaka S+R*. <https://doi.org/10.18665/sr.311396>.
- Darragh, Jen, Mikala R. Narlock, Halle Burns, Peter A. Cerda, Wind Cowles, Leslie Delserone, Seth Erickson, Joel Herndon, Heidi Imker, Lisa R. Johnston, Sherry Lake, Michael Lenard, Alicia Hofelich Mohr, Jennifer Moore, Jonathan Petters, Brandie Pullen, Shawna Taylor, and Briana Wham. 2024. "Institutional Data Repositories Are Vital." *Science* 385 (6714): 1174. <http://doi.org/10.1126/science.adr0789>.
- Davies, Sarah R. 2025. "Working in Biocuration: Contemporary Experiences and Perspectives." *Database* 2025 (January): baaf003. <https://doi.org/10.1093/database/baaf003>.
- Davies, Sarah R., and Constantin Holmer. 2024. "Care, Collaboration, and Service in Academic Data Work: Biocuration as 'Academia Otherwise.'" *Information, Communication & Society* 4: 683–701. <https://doi.org/10.1080/1369118X.2024.2315285>.
- Egher, Claudia. 2023. *Digital Healthcare and Expertise: Mental Health and New Knowledge Practices*. Palgrave Macmillan. <https://doi.org/10.1007/978-981-16-9178-2>.

- Federer, Lisa M. 2019. "Who, What, When, Where, and Why? Quantifying and Understanding Biomedical Data Reuse." PhD diss., University of Maryland. <https://doi.org/10.13016/60jd-9hux>.
- Fortun, Kim. 2009. "Scaling and Visualizing Multi-sited Ethnography." In *Multi-Sited Ethnography: Theory, Praxis and Locality in Contemporary Research*, edited by Mark-Anthony Falzon. Routledge. <https://doi.org/10.4324/9781315596389>.
- Gabrielsen, Ane Møller. 2024. "Gendering Data Care: Curators, Care, and Computers in Data-Centric Biology." *Science as Culture* 33 (2): 256–80. <https://doi.org/10.1080/09505431.2023.2260830>.
- Gregory, Kathleen, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2019a. "Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines." *Journal of the Association for Information Science and Technology* 70 (5): 419–32. <https://doi.org/10.1002/asi.24165>.
- Gregory, Kathleen M., Helena Cousijn, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2019b. "Understanding Data Search as a Socio-technical Practice." *Journal of Information Science* 46 (4): 459–75. <https://doi.org/10.1177/0165551519837182>.
- Gregory, Kathleen, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2020. "Lost or Found? Discovering Data Needed for Research." *Harvard Data Science Review* (April). <https://doi.org/10.1162/99608f92.e38165eb>.
- Gregory, Kathleen, Anton Ninkov, Chantal Ripp, Emma Roblin, Isabella Peters, and Stefanie Haustein. 2023. "Tracing Data: A Survey Investigating Disciplinary Differences in Data Citation." *Quantitative Science Studies* 4 (3): 622–49. https://doi.org/10.1162/qss_a_00264.
- Gregory, Kathleen, and Laura Koesten. 2023. "Opening Research Data." Private report.
- Gregory, Kathleen, Laura Koesten, Regina Schuster, Torsten Möller, and Sarah Davies. 2024. "Data Journeys in Popular Science: Producing Climate Change and COVID-19 Data Visualizations at Scientific American." *Harvard Data Science Review* 6 (2). <https://doi.org/10.1162/99608f92.141c99cf>.
- Gregory, Kathleen, Stefanie Haustein, Constance Poitras, Emma Roblin, Anton Ninkov, Chantal Ripp, Isabella Peters. 2026. "Digging Deeper into Data Citations: Recognizing and Rewarding Data Work." *Research Evaluation* 35: rvag008, <https://doi.org/10.1093/reseval/rvag008>.
- Grundmann, Reiner. 2017. "The Problem of Expertise in Knowledge Societies." *Minerva* 55 (1): 25–48. <https://doi.org/10.1007/s11024-016-9308-7>.
- Hammarfelt, Björn. 2019. "Discipline." In *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli. <https://www.isko.org/cyclo/discipline>.
- Hilgartner, Stephen. 2000. *Science on Stage: Expert Advice as Public Drama*. Stanford University Press. <https://doi.org/10.1515/9781503618220>.
- Hobbs, Stephanie Ketterer, and Geoffrey Hobbs. 2022. "Non-/Human Infrastructures and Digital Gifts: The Cables, Waves and Brokers of Solomon Islands Internet." *Ethnos* 87 (5): 851–73. <https://doi.org/10.1080/00141844.2020.1828969>.
- Holinski, Alexandra, Melissa L. Burke, Sarah L. Morgan, Peter McQuilton, and Patricia M. Palagi. 2020. "Biocuration - Mapping Resources and Needs." *F1000Research* 9 (ELIXIR): 1094. <https://doi.org/10.12688/f1000research.25413.2>.
- International Society for Biocuration. 2026. "Biocuration Generic Job Description." <https://www.biocuration.org/community/biocuration-generic-job-description/>.
- Jetten, Mijke, Marjan Grootveld, Annemie Mordant, Mascha Jansen, Margreet Bloemers, Margriet Miedema, and Celia W. G. Van Gelder. 2021. "Professionalising Data Stewardship in the Netherlands. Competences, Training and Education. Dutch Roadmap Towards National Implementation of FAIR Data Stewardship." March 19, 2021. Zenodo. <https://doi.org/10.5281/zenodo.4623713>.
- Kastenhofer, Karen, and Susan Molyneux-Hodgson, eds. 2021. *Community and Identity in Contemporary Technosciences*. Sociology of the Sciences Yearbook 31. Springer. <https://doi.org/10.1007/978-3-030-61728-8>.
- Khan, Nushrat, Mike Thelwall, and Kayvan Kousha. 2023. "Data Sharing and Reuse Practices: Disciplinary Differences and Improvements Needed." *Online Information Review* 47 (6): 1036–64. <https://doi.org/10.1108/OIR-08-2021-0423>.
- Knorr Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.
- Koesten, Laura M., Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. "The Trials and Tribulations of Working with Structured Data: A Study on Information Seeking Behaviour." In *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3025453.3025838>.
- Koesten, Laura, Jude Yew, and Kathleen Gregory. 2025. "Human-Data Interaction: Thinking Beyond Individual Datasets." *Interactions* 32 (1): 34–38. <https://doi.org/10.1145/3705537>.

- Latour, Bruno, and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press. <https://doi.org/10.2307/j.ctt32bbxc>.
- Lave, Jean, and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.
- Law, John. 2017. "STS as Method." In *The Handbook of Science and Technology Studies*, 4th ed., edited by Ulrike Felt, Rayvon Fouché, Clark A. Miller, and Laurel Smith-Doerr. The MIT Press.
- Lean, Oliver M. 2021. "Are Bio-ontologies Metaphysical Theories?" *Synthese* 199: 11587–11608. <https://doi.org/10.1007/s11229-021-03303-4>.
- Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226416502.001.0001>.
- Leonelli, Sabina, and Rachel A. Ankeny. 2015. "Repertoires: How to Transform a Project into a Research Community." *BioScience* 65 (7): 701–8. <https://doi.org/10.1093/biosci/biv061>.
- Levallois, Clement, Stephanie Steinmetz, and Paul Wouters. 2013. "Sloppy Data Floods or Precise Social Science Methodologies? Dilemmas in the Transition to Data-Intensive Research in Sociology and Economics." In *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*, edited by Paul Wouters, Anne Beaulieu, Andrea Scharnhorst, and Sally Wyatt. The MIT Press. <https://doi.org/10.7551/mitpress/9274.003.0007>.
- Metzler, Ingrid, Lisa-Maria Ferent, and Ulrike Felt. 2023. "On Samples, Data, and Their Mobility in Biobanking: How Imagined Travels Help to Relate Samples and Data." *Big Data & Society* 10 (1): 1–13. <https://doi.org/10.1177/20539517231158635>.
- Oliver, Gillian, Jocelyn Cranefield, Spencer Lilley, and Matthew Lewellen. 2023. "Data Cultures: A Scoping Literature Review." *Information Research* 28 (1): 3–29. <https://doi.org/10.47989/irpaper950>.
- Ottinger, Gwen. 2010. "Buckets of Resistance: Standards and the Effectiveness of Citizen Science." *Science, Technology & Human Values* 35 (2): 244–70. <https://doi.org/10.1177/0162243909337121>.
- Ottinger, Gwen. 2022. "Misunderstanding Citizen Science: Hermeneutic Ignorance in U.S. Environmental Regulation." *Science as Culture* 34 (1): 504–29. <https://doi.org/10.1080/09505431.2022.2035710>.
- Ottinger, Gwen. 2023. "Responsible Epistemic Innovation: How Combatting Epistemic Injustice Advances Responsible Innovation (and Vice Versa)." *Journal of Responsible Innovation* 10 (1): 1–19. <https://doi.org/10.1080/23299460.2022.2054306>.
- Parasie, Sylvain. 2015. "Data-Driven Revelation?: Epistemological Tensions in Investigative Journalism in the Age of 'Big Data.'" *Digital Journalism* 3 (3): 364–80. <https://doi.org/10.1080/21670811.2014.976408>.
- Pink, Sarah. 2021. "The Ethnographic Hunch." In *Experimenting with Ethnography: A Companion to Analysis*, edited by Andrea Ballesterio and Brit Ross Winthereik. Duke University Press. <https://doi.org/10.1515/9781478091691-005>.
- Poirier, Lindsay. 2023. "Attending to the Cultures of Data Science Work." *Data Science Journal* 22 (6): 1–7. <https://doi.org/10.5334/dsj-2023-006>.
- Poirier, Lindsay, and Brandon Costelloe-Kuehn. 2019. "Data Sharing at Scale: A Heuristic for Affirming Data Cultures." *Data Science Journal* 18 (48): 1–7. <https://doi.org/10.5334/dsj-2019-048>.
- Robinson-García, Nicolas, Evaristo Jiménez-Contreras, and Daniel Torres-Salinas. 2016. "Analyzing Data Citation Practices Using the Data Citation Index." *Journal of the Association for Information Science and Technology* 67 (12): 2964–75. <https://doi.org/10.1002/asi.23529>.
- Springer, Rebecca, and Danielle Cooper. 2020. "Data Communities: Empowering Researcher-Driven Data Sharing in the Sciences." *International Journal of Digital Curation* 15 (1): 1–8. <https://doi.org/10.2218/ijdc.v15i1.695>.
- Traweek, Sharon. (1988) 1992. *Beamtimes and Lifetimes: The World of High Energy Physicists*. Harvard University Press.
- UNESCO. 2021. *UNESCO Recommendation on Open Science*. UNESCO Digital Library. <https://doi.org/10.54677/MNMH8546>.
- Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter

- Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (160018). <https://doi.org/10.1038/sdata.2016.18>.
- Wofford, Morgan F., and Andrea K. Thomer. 2023. "Curating for Contrarian Communities: Data Practices of Anthropogenic Climate Change Skeptics." *Proceedings of the Association for Information Science and Technology* 60 (1): 442–55. <https://doi.org/10.1002/pra2.802>.
- Wood-Charlson, Elisha M., Zachary Crockett, Chris Erdmann, Adam P. Arkin, and Carly B. Robinson. 2022. "Ten Simple Rules for Getting and Giving Credit for Data." *PLoS Computational Biology* 18 (9): e1010476. <https://doi.org/10.1371/journal.pcbi.1010476>.
- Wyatt, Sally. 2003. "Non-Users Also Matter: The Construction of Users and Non-Users of the Internet." In *How Users Matter: The Co-Construction of Users and Technology*, edited by Nelly Oudshoorn and Trevor Pinch. The MIT Press. <https://doi.org/10.7551/mitpress/3592.003.0006>.
- Zeng, Tong, Longfeng Wu, Sarah Bratt, and Daniel E. Acuna. 2020. "Assigning Credit to Scientific Datasets Using Article Citation Networks." *Journal of Informetrics* 14 (2): 1–25. <https://doi.org/10.1016/j.joi.2020.101013>.

How to cite this article: Gregory, Kathleen, and Sarah R. Davies. 2026. Studying Data Communities: Analytical Dimensions from and for Empirical Research. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 9(2). <https://doi.org/10.18357/kula.327>

Submitted: 18 August 2025 **Accepted:** 24 April 2026 **Published:** 29 September 2026

Copyright: © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

KULA: Knowledge Creation, Dissemination, and Preservation Studies is a peer-reviewed open access journal published by University of Victoria Libraries.

OPEN ACCESS 