Disentangling knowledge production and data production

KAREN S. BAKER1,† AND MATTHEW S. MAYERNIK2

1School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, Illinois, USA
2National Center for Atmospheric Research (NCAR), University Corporation for Atmospheric Research (UCAR), Boulder, Colorado, USA

Citation: Baker, K. S., and M. S. Mayernik. 2020. Disentangling knowledge production and data production. Ecosphere 11(7):e03191. 10.1002/ecs2.3191

Abstract. With today’s increasing attention to open science and open data, knowledge production and data production have significant impacts on both research and policy arenas. We scrutinize the concept of data production in relation to knowledge production in the earth and environmental sciences. We present three empirical cases that illustrate issues arising when knowledge production and data production as inter-dependent but distinct processes are not clear in the mind of researchers and policymakers. A two-stream model is developed that highlights their interplay yet avoids the conflation of knowledge and data production. This approach highlights knowledge and data in terms of processes rather than stable objects or assets. Further, we suggest that considering knowledge and data production in relation to their development by communities rather than as commodities helps in understanding the debates and issues that arise in contemporary research practice.

Key words: data access; data management; data policy; data production; knowledge production.

Received 6 January 2020; revised 24 April 2020; accepted 29 April 2020; final version received 26 May 2020. Corresponding Editor: Debra P. C. Peters.

Copyright © 2020 The Authors. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

† E-mail: karensbaker@gmail.com

INTRODUCTION

In mid-November 2019, the New York Times reported that the U.S. Environmental Protection Agency (EPA) was preparing new internal rules that would “significantly limit the scientific and medical research that the government can use to determine public health regulations, overriding protests from scientists and physicians…” (Friedman 2019). According to the New York Times, the new rules would require that “scientists disclose all of their raw data, including confidential medical records, before the agency could consider an academic study’s conclusions.” Many scientists and other stakeholders within scientific institutions responded strongly in public venues with negative reactions to the proposed rules (Thorp et al. 2019). On the surface, this presents somewhat of a conundrum, given the otherwise strong trend within the scientific community toward open science and open data. Why would open science and open data advocates oppose a rule to require that government policies be based on open data? Leaving aside political affiliations and priorities (which are significant to this debate), this conundrum involves a conflation of two related but distinct concepts: knowledge production and data production.

Both the EPA proposed rule changes and open science advocates point to the concept of transparency as being a prime motivation (Nosek et al. 2015, Friedman 2019). But the debate about the EPA rules begs the following questions: (1) What is being made transparent? And (2) how is that transparency being achieved (Mayernik 2017)? At the risk of oversimplifying,
it might be possible to say that the scientific community’s pushback against this proposed EPA rule centers on the debate about whether transparent data production is necessary or sufficient for transparent knowledge production to occur. Critics of the proposed EPA rules argue that, at least in some cases, transparent data production is not necessary to the production of transparent science, because the EPA has had in place for years effective processes for producing robust knowledge of the environment based on data that cannot or should not be shared due to privacy and other concerns (Goldman 2018, Nosek 2019, Virginia Tech 2019). And it can also be argued that transparent data production is not sufficient for the creation of transparent knowledge because simply making data accessible does not mean that those are the best or most appropriate data for use in the production of reliable and valid scientific knowledge on the topic at hand.

Researchers use data as evidence in the production of knowledge (Borgman 2015, Mayernik 2019, Leonelli 2019). Contemporary discussions of open science, data sharing, open data, reproducibility, and research policy like those going on in this EPA case are all problematic, however, if they conflated the production of knowledge and data. Given the rapid changes science is undergoing and the interdependence of data and knowledge, data will continue to be difficult to find and share until the distinction between knowledge production and data production becomes apparent across the many scientific data-generating and policymaking arenas.

Data, information, and knowledge are interrelated concepts, each with their own historical and current understandings. A comprehensive analysis of the differences and connections between these terms is beyond the scope of our study (see Zins 2007 for such an analysis). We provide here an illustration of the ways in which the conflation of data production and knowledge production can lead to misplaced expectations, policy confusion, and funding allocation debates, among other potential issues. In particular, through the study of data-generating projects in the field sciences, we provide examples of how data work is evolving as researchers collaborate on large-scale, long-term interdisciplinary projects that involve collective management and use of data in order to generate knowledge. In many research areas, new approaches are emerging that enable investigation of natural ecosystems in the face of rapid changes taking place in technical, methodological, and organizational arrangements as well as in the natural system itself. Despite their interdependence, by considering data production and knowledge production as two distinct concepts, we can deepen our understanding of modern scientific work.

**Knowledge production**

Knowledge is a complex phenomenon. Social studies of science have made clear that knowledge exists in the context of social settings and dynamics (Edwards et al. 2013, Longino 2019). Whether something exists as knowledge or not is often dependent on the persons who generate and distribute it and the reception of those persons by other members of a social setting (Shapin 1994). There is a history of scholarship that focuses on knowledge as being constructed socially (Bijker et al. 1989). Knowledge is also often distributed among people within a particular social setting or community, often using established practices. People work together, building on each other’s knowledge and skills, to achieve things that they could not achieve alone (Hutchins 1995).

Starting in the 17th century when scientific findings were published in journals, knowledge sharing became a recognized scientific practice. This codification of knowledge into written forms also exists within situated networks of people (Latour 1988). Scientific findings written into journals do not become knowledge unless accepted broadly by others within the scientific community (Collins 2017). In addition, scientific papers are filtered and formalized descriptions of highly situated activities. Recreating the same knowledge outcomes often requires significant effort and relies on the presence of embodied and practice-based knowledge of how the research study was done (Collins 1985, Schmidt 2012).

We refer to knowledge-making activities under the broadly conceived term “knowledge production.” Policy motivations for many open movements directly invoke the idea of facilitating knowledge production, such as the Berlin
Declaration (2003), which defined and promoted the concept of open access:

Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

The data used as evidence underlying knowledge production have shifted over time. In early papers, data involved presentations such as photographs, the description of an experimental device, a summary of observations, or perhaps a small table of numbers. Today, knowledge production can include terabytes of data or use of previously generated data for new research. This expanding role of data propels the need to better understand data production and access.

Data production

In contemporary science, data has burst the bounds of tables, reports, and databases. Increasing volumes and varieties of data together with the increased availability of platforms, computers and digital devices including their software and applications, have created new kinds of issues relating to work with data. We use the term “data production” to refer to producing data with the intention that they may be used by others, where data production can involve a number of activities from generation to sharing of data including its description, management, packaging, archive, and access. The notion of projects and programs establishing data policies in order to share data widely began in earnest in the latter part of the 20th century as the Internet became ubiquitous and remote disk access enabled data exchange in response to the growth of scientific collaboration.

As data have become more visible in policy circles, they have become products in their own right (Leonelli 2019). Data production has become an important and desirable outcome of research funding. The U.S. national data access policy in 2013 deliberately sought to increase data production and sharing. Likewise, internationally agreed-upon data sharing principles provide guidance for data management and stewardship to ensure findable, accessible, interoperable, and reusable (FAIR) data (Wilkinson 2016). With concurrent rapid growth of technologies and data work, there are new roles, practices, and activities emerging to support data initiatives.

The issue of burgeoning data is further complicated by the increasing collaboration in research. In addressing the complexity and scopes associated with today’s grand challenges and big ideas in the earth and environmental sciences, the management of collective efforts and their data is foregrounded. In laboratories and programs, new roles are developing—first informally and recently more formally—with titles such as data manager, data scientist, and data professional. Expertise has been developing in data processing, analysis, and visualization for the immediate use of data as well as in data description, packaging, and curating for its reuse. Data specialists are emerging within collaborative research efforts capable of assembling data within a domain, an institution, a program, a project, or a laboratory as reports aim to provide an overview of existing as well as new roles and training (Pryor and Donnelly 2009, NRC 2015).

Today data are still hard to find

Numerous recent studies have shown that it can be very difficult to find the data that underpin published research (Andreoli-Versbach and Mueller-Langer 2014, Van Tuyl and Whitmire 2016). Descriptions of the efforts undertaken to track down such data often read like archaeological or law enforcement investigations, with would-be data discoverers unearthing layers of detritus, inconsistent documentation, and repeated communication dead ends (Wolins 1962, White 1982, Vines et al. 2014, Van Noorden 2015).

To the scientist, these challenges are speed bumps or dead ends in the pursuit of new or old findings. To the policymaker, these challenges represent inefficiencies in the use of public dollars. To the open science and/or reproducibility evangelist, because there are new ways of working, new institutional forms, and new tools that can enable much broader distribution, use, and collaboration related to data (Goodman et al. 2014, Perez and Granger 2015, Cutcher-Gershenfeld et al. 2017), these challenges demonstrate the insufficiencies of past practices, institutions, and technologies that still live strong in too many parts of the scientific enterprise.
Fig. 1. The two-stream model shows two branches: (1) knowledge production using data optimized for local use with the final form optimized for publication of papers; and (2) data production creates data intended for release to a data repository that makes data accessible for reuse by others.

To the social analyst, however, these challenges should represent something different, namely, something to be understood and characterized. The production of knowledge has continued apace even with the above difficulties. This is one indication of how knowledge production and data production have complex and uneven interdependencies within the scientific research enterprise (Borgman 2015). Their intent differs: Knowledge production is based on plans to use data generated; data production aims to make data available for reuse by others.

Simply generating more data is not sufficient to realize the vision of data as open, accessible, and reusable resources (Borgman 2012, Mayernik 2017). In this paper, we address the frequent conflation of knowledge production and data production that makes discussions of open science problematic by conceptualizing the work of data production in relation to knowledge production as shown in Fig. 1. While data are central to the production of scientific knowledge, additional work is required to make earth and environmental science data available as stand-alone, functional research products. This is captured by the branching into two streams of work shown in Fig. 1. Many scientific research projects generate novel and reliable knowledge by following the lower, traditional knowledge production branch shown. To accomplish data production as defined earlier, however, different kinds of data work must take place, as shown in the upper branch.

At present, researchers have a history of managing their data to meet the end goal of knowledge production—the publication of results. The knowledge production process (lower branch) involves using data, subject to a complex mix of data selection, processing, analysis, integration, and presentation strategies driven by scientific inquiry, with the final form optimized for publication of papers. Data production (upper branch) uses a more formalized set of procedures to assemble, describe, and package data for submission to a data repository that makes data accessible for reuse by others using steps and standards described by lifecycle models for preserving and sharing data (Carlson 2014). The data production branch represents an expanded conceptualization of data management. It draws attention to some of the new activities that need to be integrated into the research process for scientists collecting data in...
the field and conducting laboratory experiments. The set of data practices is conceptualized not as data publication but rather as production of data for release for future, unanticipated applications (Parsons and Fox 2013).

Though the two trajectories are shown as distinct, researcher activities may address activities on both trajectories to varying degrees. Further, the bidirectional feedback activities’ arrow indicates that using the data for knowledge production may inform data production and vice versa. C1 refers to a smaller-scale, short-term effort, perhaps an individual’s datasets for a single project with a minimum of metadata adequate for local data use because of the presence of tacit knowledge and experience of those who collect the data. This same researcher may at the same time or subsequently take a more structured approach so performing C2 data assembly which refers to highly structured, multi-project data and/or to diverse kinds of data with rich metadata. The two trajectories in Fig. 1 do not represent an either/or situation. Work may proceed along both trajectories, taking some or all the steps on a particular branch. In practice, researchers exposed to the data production trajectory carried out by data managers may add structure and standardization to their everyday procedures as they evolve their data practices.

Our aim in presenting this two-stream model is to help scientists, policymakers, and other stakeholders in the open science and open data communities appreciate the distinction between knowledge production and data production with new kinds of data work emerging to support the long-term reuse of data.

ILLUSTRATING THE ISSUES

The following three examples in the earth and environmental sciences illustrate some of the issues that arise in practice associated with knowledge production and data production.

The unconfirmed assumption that investing in either knowledge or data production necessarily leads to advances in the other

Institutional arrangements for managing data and metadata can vary significantly within and across scientific organizations and projects (Mayernik 2016). The Center for Embedded Networked Sensing (CENS) from 2002 to 2012 was an NSF-funded Science and Technology Center that supported collaborations between scientists, computer scientists, and engineers with the goal of developing new kinds of sensing systems for scientific use. Scientific and technical research within CENS both involved prototyping sensing instruments in the field. Over time, the CENS collaborations developed effective ways to work together to generate data via humans-in-the-loop-sensing deployments (Mayernik et al. 2013). CENS teams generated many kinds of data, in some cases in high volumes, to support the diverse interests of the teams (Borgman et al. 2012). Engineers tended to be focused on the development, deployment, and performance of instruments, while scientists were eager to use the instruments to study environmental phenomena. Documenting the decision-making occurring throughout the life cycle of data was difficult particularly with the multidisciplinary partners involved (Wallis et al. 2008).

Although CENS teams generated data in large volumes and varieties, in some projects moving one or two steps down the data production trajectory, the goal of CENS clearly resides on the knowledge generation branch of the two-stream model. As an indication of the knowledge produced in CENS, a digital library of CENS products lists 671 documents that were published between 2001 and 2011, including 392 posters, 204 papers, and 70 technical reports (https://escholarship.org/uc/cens). Very few of the datasets that were collected by CENS researchers, however, are easily findable online, as of 2019. Wallis et al. (2010), discussing the interdependence of the scientific and technical data being collected in CENS, asked: How much of the engineering metric data need to be preserved for use in interpreting the scientific data in question? But ten years later, another question arises in light of the two-stream model: How much of the scientific or technical data need to be preserved for use in interpreting the scientific knowledge produced by CENS researchers?

Chances are that much of the data produced in CENS still exist in laboratory or personal computers of the researchers involved. But in light of the two-stream model, without a concerted emphasis on data production, CENS’ legacy is heavily skewed toward the knowledge
production stream. Two important points are highlighted here:

1. The emphasis by CENS on knowledge production has had many positive outcomes. Researchers and graduate students involved in CENS have achieved positions of professional success and influence. And papers written about CENS research are still cited heavily by researchers in all research areas that were involved, indicating the ongoing utility of CENS outcomes within diverse scientific communities.

2. Investing in knowledge production did not lead automatically to success in data production. Within CENS, knowledge was produced and expertise advanced by engineers working in concert with scientists through shared experiences with instruments as they were developed and deployed in the field. Lacking participants focused on data management, there were few intermediaries to establish and carry out data tasks that would ensure the design of routines supporting the assembly, description, and management of the data (Baker and Millerand 2010, Wallis 2012, Mayernik 2016). Though the traditional understanding of computational support was available as needed, the distinction between computational services and data services that was beginning to be recognized at the turn of the century had not yet reached many CENS projects (Mayernik 2016). As a result, projects that were successful with knowledge production often did not advance in terms of data production that would ensure their data were available for interpretation, reuse, and knowledge-making by researchers outside the project.

The awareness of differences in expertise and professional orientation for the work of knowledge vs. data production

If knowledge and data production involve different kinds of work, as shown in the two-stream model, then they also necessarily require different kinds of expertise and skills. Launched in 1980 before the internet was ubiquitous, the U.S. Long-Term Ecological Research (LTER) program launch included the following two significant requirements: first that the geographically distributed sites network together in some way and second that each site designate a data manager to handle their long-term data. Today, the network is comprised of more than twenty-eight member sites, each focusing on a particular biome. Sites coordinate both research and data activities across the network. The LTER is recognized for its contributions to ecology (e.g., Hobbs et al. 2003, Waide and Thomas 2013) as well as its site-based and network-wide management of data (Baker et al. 2000, Benson et al. 2006, Porter 2010, Michener et al. 2011, Servilla et al. 2016).

The initial data manager positions were largely part-time and often made use of existing personnel, perhaps a secretary, a field technician, or a junior researcher. Over time, these roles evolved into full-time information management positions serving as liaisons between project researchers and data management teams. The work of data and information management evolved slowly with intermittent network-wide activities such as focus on data catalogs and data policy as well as metadata standards and their validation. Data managers were at hand to address emergent site-specific digital circumstances as they arose, including design and development of local data practices, identification and alignment with other LTER sites and remote data partners, as well as work with project-related technologies and digital information (Karasti and Baker 2004, 2008, Karasti et al. 2006). In time, as the experience base of the data specialists grew, they were depended upon as data professionals embedded at each science-driven site. LTER became a network focusing on generation and use of data as well as on reuse of long-term data.

Data duties initially focused on assembling data in order to support both planned and unexpected data activities associated with knowledge production, even when practices might result in inadequate metadata, incomplete data checking, or data shared in nonstandard formats. Immersed in the LTER culture of long term, the LTER data managers slowly developed new insights and skills associated with data production, such as an understanding of partnering with highly structured data facilities to preserve and disseminate data given constraints on their flexible but often loosely structured local data systems, of community metadata standards
necessary to avoid cross-site data inconsistencies, and of data packaging to facilitate automated data ingestion (Baker and Millerand 2010). As LTER data managers grew into the role of data management, their initial support of knowledge production broadened to support both branches of the two-stream model while being in a position to convey the value of both to those with whom they worked.

The uncertainty about what kinds of infrastructures and institutions are needed to support knowledge vs. data production

The Ocean Observatories Initiative (OOI) is an NSF Division of Ocean Sciences (OCE) project involving millions of dollars for new large-scale marine infrastructure aimed at addressing major global issues including climate change and ocean acidification. Funded for a 25-yr time span, design and discussion beginning in 2000 were followed by construction starting in 2009 and operations in 2016 that included data online services and a project office. The initiative reflects major advances in observational approaches that exploit new developments in technology enabling longer-term and larger-scale efforts that span oceans from their surface to their floor using new types of instrumentation all tied together as a data-generating marine infrastructure that incorporates contemporary collaboration and communication arrangements (Smith et al. 2018). OOI consists of a distributed set of in situ platforms that generate terabytes of data including 100,000 parameters packaged into 206 unique data products (Smith et al. 2018). With data made accessible via their data portal, this is data production writ large.

The issue of increasing costs for ocean infrastructure was addressed by NSF in 2015 as OOI was being readied for launch, and as OCE requested, the National Research Council provides guidance on support for the next decade of ocean research. The report stated, “From 2000 through 2014, there has been a shift in investment from the core research programs to the operations and maintenance of infrastructure.” Kintisch (2015) summarized, “In 2000, 62% of NSF’s ocean budget went to research grants; the share is now [in 2015] about 45% and sinking.” The need for course corrections was detailed in the report. The 2015 NRC report also noted a “lack of broad community support for this initiative” though support is difficult to assess in cases where new technology-rich infrastructure is involved. OOI survived its descoping. A view into the reorganization and breaking down that accompanied the development of a large-scale infrastructure like OOI is described by Steinhart (2016). The global-scale ocean sensor network of networks, an infrastructure focused on data production with a design approach that takes years to plan and build before operation begins, is based on the promise of data becoming available to others, that is, to researchers largely focused on knowledge production.

DISCUSSION

\[ \text{Just-right balance that changes over time} \]

With a fixed budget that limits time, resources, and energy, the reciprocal relationship between knowledge production and data production can be a zero-sum game where an increase in one results in a loss for the other. When an investment is made in data production, it might impact the resources and emphasis put toward knowledge production, and vice versa (Fig. 2).

The OOI case provides a clear illustration of this zero-sum game. Funds invested at the agency level in sustainability of the OOI data

![Fig. 2. The dotted line shows a sliding-scale relationship between knowledge production (KP) and data production (DP) within a fixed budget. A position may be changed over time if the distribution of funds between KP and DP changes, for instance from high KP and low DP shown on the left to low KP and high DP on the right.](image-url)
production infrastructure are funds that cannot be allocated to individuals or teams of researchers for basic research supported by traditional science-driven grants. In another example, researchers who use local or distributed computing environments may have computing and storage costs combined in their budget. Any allocation of funds toward data storage may directly reduce the available funds for generating new knowledge via computing.

For each data environment, a just-right balance of knowledge production and data production will be established at any one moment in time. Each case of data management evolves by balancing institutional, political, technical, and human arrangements to support their research. The balance established will likely change as participant aims and capabilities evolve or funding and circumstances change. Feedbacks between data production and knowledge production have the potential to reshape research practices, infrastructures, and the balance of knowledge production and data production.

It is informative to note that each of the field science case examples discussed has developed different approaches to investing in data production. This emphasis can change over time. Table 1 shows where emphasis was placed in general on data production and knowledge production for our empirical cases. OOI—a large budget, continental-scale project planned on a multi-decade timeline with the goal of producing well-defined data products—emphasized data production via digital instruments and platforms early on as well as the movement of data to a centralized hub. In contrast, the U.S. LTER—a long-term program with the goal of knowledge-making both at individual sites and across the sites as a network—invested a small amount in data management initially and then each site in concert with their local logistics moved at their own pace from investment in part-time to full-time data management positions and in some cases to information management teams. Their investment changed over time as researchers modified or replaced data practices and benefited from data access. CENS—a 10-yr center enabling interdisciplinary innovation in instrument development—emphasized knowledge production as a priority.

Changing language for shifting perspectives

Leonelli (2019) underscores that data may be recognized as “relational objects, the very identity of which as sources of evidence—let alone their significance and interpretation—depends on the interests, goals and motives of the people involved, and their institutional and financial context. Extracting knowledge from data is not a neutral act.” She points to changing views of data. First seen as a stable object explained by data generators or data owners, data have evolved in the digital era into something viewed as an asset that can be moved and exchanged via digital infrastructures. When viewed as an asset, contrasts can be drawn between the open availability of data in some communities and the withholding of data as private commodities in other sectors. Until the recent decade, research has been conducted with budgets planned to purpose, for the project at hand, with the use of local data in mind rather than data as a project or institutional asset to share. Packaging data as an asset to share—beyond the purview of the project’s use of the data—requires new data practices and capabilities. As shown by the two-stream model, passing data forward requires augmenting traditional, short-term activities with a long-term view as well as with an understanding of data care (Baker and Karasti 2018). It requires processes for packaging data so that a dataset conforms to standard templates and is well-described using a metadata standard that will pass through metadata validators that check for completeness and accuracy.

Table 1. A snapshot of data production and knowledge production emphasis for three cases.

| Case acronym | Case name                          | Data production emphasis | Knowledge production emphasis |
|--------------|------------------------------------|--------------------------|------------------------------|
| CENS         | Center for Embedded Network Sensing| x                        |                              |
| LTER         | Long-Term Ecological Research Network| x                        | x                            |
| OOI          | Ocean Observatories Research Initiative | x                        | x                            |
Finally, here we reflect on this two-stream knowledge and data production model in the context of Agre’s (2000) discussion of the coexistence and complementarity of a community model and a commodity model of the scientific enterprise. From one perspective, science can be seen as principally involving the development of communities. As noted previously, science is a social activity. Scholarly communities are effective in both sharing and producing knowledge via interactive venues such as universities, workshops, conferences, seminars, as well as digitally supported teams and projects. In contrast, when regarded as a commodity, knowledge is something to be bundled and packaged for sale such as a product supporting student services. In the case of data created as a public good, however, it is important to distinguish data packaged for sharing from the economic concept of commodity. Ribes and Jackson’s (2013) presentation of the notion of commodity fiction encourages us to consider data as entities tied to their roots in order not to dissociate them from their origin and the creator’s sphere of context that traditionally includes a knowledge-making community of practice (Baker and Yarmey 2009). No matter the richness or completeness of its metadata, we must recognize that data were initially produced in the context of a community. Packaging data to permit their travel away from the community of origin enables their distribution as a renewable resource, but also the potential for misinterpretation of the data because they are not interchangeable goods, but rather are entities with provenance or history that is critical to the data use in knowledge-making.

CONCLUSION

In this paper, we analyze two terms and their relationship—knowledge production and data production—with the understanding that in the sciences, the notion of production includes both planned, formal procedures and messy, unplanned activities. Scientists trained in collecting and sharing their own data often fail to distinguish the traditional use of data for the production of knowledge from the more recent notion of data production that refers to the management of data for reuse by other communities and larger audiences.

We return here to the EPA case discussed in the introduction. The two-stream model of knowledge and data production shown in Fig. 1 has explanatory power that draws attention in the EPA case to misunderstandings that occur (deliberately or accidentally) when the distinction between knowledge production and data production is not acknowledged or understood. Producing data and/or knowledge is a process. Viewed in the context of the two-stream model, critics of the proposed new U.S. EPA rule argue that transparency of data is not something that is uniformly required or possible for more transparent knowledge. Proponents of the proposed rule, on the other hand, argue that data production and knowledge production can be seen as the same thing, in the sense that without the former, the latter is not possible. The reality is of course more complex. The trustworthiness of any scientific and regulatory activities involves assemblages of people, processes, institutions, infrastructures, technologies, and data (Shapin 1994, Jirotka et al. 2005, Gray et al. 2018,Pasquetto et al. 2019).

The EPA case is but one example of the entanglement of knowledge production and data production. This entanglement, however, should not result in their conflation. Traditional research norms and infrastructures have prioritized knowledge production. In recent years, however, particularly as the volume and heterogeneity of data have grown, research and policy institutions are shifting the relative prioritization of knowledge production vs. data production. In this paper, we suggest that considering knowledge and data in terms of processes vs. objects, and considering knowledge production and data production in relation to the development of communities vs. commodities can be of help in understanding arrangements and investments that impact open science and open data.

ACKNOWLEDGMENTS

We thank the anonymous peer reviewers for comments on an earlier version of this paper. Support by University of Illinois Urbana-Champaign is acknowledged including a School for Information Sciences Graduate Research Assistantship for fieldwork at the National Center for Atmospheric Research (NCAR). This material is based upon work.
supported by NCAR, a major facility sponsored by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of NCAR or the NSF.

LITERATURE CITED

Agre, P. E. 2000. Commodity and community: institutional design for the networked university. Planning for Higher Education 29:5–14.

Andreoli-Versbach, P., and F. Mueller-Langer. 2014. Open access to data: An ideal professed but not practised. Research Policy 43:1621–1633.

Baker, K. S., B. J. Benson, D. L. Henshaw, D. Blodgett, J. H. Porter, and S. G. Stafford. 2000. Evolution of a multisite network information system: the LTER information management paradigm. BioScience 50:963–978.

Baker, K. S., and H. Karasti. 2018. Data care and its politics: designing for a neglected thing. In Proceedings of the 15th Participatory Design Conference (PDC’18): Full Papers - Volume 1, 20–24 August 2018, Hasselt and Genk, Belgium. ACM, New York, New York, USA.

Baker, K. S., and F. Millerand. 2010. Infrastructuring ecology: challenges in achieving data sharing. Pages 111–138 in B. Penders, J. N. Parker, and N. Vermeulen, editors. Collaboration in the new life sciences. Ashgate Publishing, Surrey, UK.

Baker, K. S., and L. Yarmey. 2009. Data stewardship: environmental data curation and a web-of-repositories. International Journal of Digital Curation 4:12–27.

Benson, B. J., P. C. Hanson, J. W. Chipman, and C. J. Bowser. 2006. Breaking the data barrier: research facilitation through information management. Pages 258–279 in J. J. Magnuson, T. K. Kratz, and B. J. Benson, editors. Long-term dynamics of lakes in the landscape: long-term ecological research on North Temperate Lakes. Oxford University Press, Oxford, UK.

Berlin Declaration. 2003. Berlin Declaration on open access to knowledge in the sciences and humanities. Max Planck Society, Berlin, Germany. https://openaccess.mpg.de/Berlin-Declaration

Bijker, W., T. P. Hughes, and T. Pinch. 1989. The social construction of technological systems: new directions in the sociology and history of technology. MIT Press, Cambridge, Massachusetts, USA.

Borgman, C. L. 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63:1059–1078.

Borgman, C. L. 2015. Big data, little data, no data: scholarship in the networked world. MIT Press, Cambridge, Massachusetts, USA.

Borgman, C. L., J. C. Wallis, and M. S. Mayernik. 2012. Who’s got the data? Interdependencies in science and technology collaborations. Computer Supported Cooperative Work (CSCW) 21:485–523.

Carlson, J. 2014. The use of life cycle models in developing and supporting data services. Pages 63–86 in J. M. Ray, editor. Research data management: practical strategies for information professionals. Purdue University Press, West Lafayette, Indiana, USA.

Collins, H. 1985. Changing order: replication and induction in scientific practice. University of Chicago Press, Chicago, Illinois, USA.

Collins, H. 2015. Big data, little data, no data: scholarship in the networked world. MIT Press, Cambridge, Massachusetts, USA.

Cutcher-Gershenfeld, J., et al. 2017. Five ways consortia can catalyse open science. Nature 543:615–617.

Edwards, P. N., et al. 2013. Knowledge infrastructures: intellectual frameworks and research challenges. University of Michigan, Ann Arbor, Michigan, USA. http://hdl.handle.net/2027.42/97552

Friedman, L. 2019. E.P.A. to limit science used to write public health rules. New York Times. https://www.nytimes.com/2019/11/11/climate/epa-science-trump.html

Goldman, G. 2018. Scott Pruitt will restrict the EPA's use of legitimate science [Scientific American Blog]. https://blogs.scientificamerican.com/observations/scott-pruitt-will-restrict-the-epas-use-of-legitimate-science/

Goodman, A., et al. 2014. Ten simple rules for the care and feeding of scientific data. PLoS Computational Biology 10:e1003542.

Gray, J., C. Gerlitz, and L. Bounegru. 2018. Data infrastructures: intellectual frameworks and research challenges. University of Michigan, Ann Arbor, Michigan, USA. http://hdl.handle.net/2027.42/97552

Hutchins, E. 1995. Cognition in the wild. MIT Press, Cambridge, Massachusetts, USA.

Jirotka, M., et al. 2005. Collaboration and trust in healthcare innovation: the EDiaMoND case study. Computer Supported Cooperative Work (CSCW) 21:485–523.

Karasti, H., and K. S. Baker. 2008. Infrastructuring for the long-term: ecological information management. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences. IEEE, Piscataway, New Jersey, USA.

Karasti, H., and K. S. Baker. 2008. Digital data practices and the long term ecological research
program growing global. International Journal of Digital Curation 3:42–58.
Karasti, H., K. S. Baker, and E. Halkola. 2006. Enriching the notion of data curation in e-science: data managing and information infrastructuring in the Long Term Ecological Research (LTER) Network. Computer Supported Cooperative Work (CSCW) 15:321–358.
Kintisch, E. 2015. A moment of truth arrives for U.S. ocean science. Report urges infrastructure cuts to boost research. Science 347:463.
Latour, B. 1988. The pasteurization of France. Harvard University Press, Cambridge, Massachusetts, USA.
Leonelli, S. 2019. Data - from objects to assets. Nature 574:317–320.
Longino, H. 2019. The social dimensions of scientific knowledge. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/scientific-knowledge-social/
Mayernik, M. S. 2016. Research data and metadata curation as institutional issues. Journal of the Association for Information Science and Technology 67:973–993.
Mayernik, M. S. 2017. Open data: accountability and transparency. Big Data & Society 4:1–5.
Mayernik, M. S. 2019. Metadata accounts: achieving data and evidence in scientific research. Social Studies of Science 49:732–757.
Mayernik, M. S., J. C. Wallis, and C. L. Borgman. 2013. Unearthing the infrastructure: humans and sensors in field-based scientific research. Computer Supported Cooperative Work (CSCW) 22:65–101.
Michener, W. K., J. Porter, M. Servilla, and K. Vanderbilt. 2011. Long term ecological research and information management. Ecological Informatics 6:13–24.
Nosek, B. A., et al. 2015. Promoting an open research culture. Science 348:1422–1425.
Nosek, B. A. 2019. Testimony of Brian A. Nosek before the Committee on Science, Space, and Technology, U.S. House of Representatives, November 13, 2019. https://science.house.gov/imo/media/doc/Nosek%20Testimony.pdf
NRC. 2015. National Research Council, Sea change: 2015–2025 decadal survey of ocean sciences. National Academies Press, Washington, D.C., USA.
Parsons, M., and P. Fox. 2013. Is data publication the right metaphor? Data Science Journal 12:WDS32–WDS46.
Pasquetto, I. V., C. L. Borgman, and M. F. Wofford. 2019. Uses and reuses of scientific data: the data creators’ advantage. Harvard Data Science Review 1:1–35.
Perez, F., and B. E. Granger. 2015. Project Jupyter: computational narratives as the engine of collaborative data science [Jupyter.org blog]. https://blog.jupyter.org/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58
Porter, J. H. 2010. A brief history of data sharing in the U.S. Long Term Ecological Research Network. Bulletin of the Ecological Society of America 91:14–20.
Pryor, G., and M. Donnelly. 2009. Skilling up to do data: Whose role, whose responsibility, whose career? International Journal of Digital Curation 4:158–170.
Ribes, D., and S. Jackson. 2013. Data bite man: the work of sustaining a long-term study. Pages 147–166 in L. Gitleman, editor. “Raw data” is an oxymoron. MIT Press, Cambridge, Massachusetts, USA.
Schmidt, K. 2012. The trouble with ‘tacit knowledge’. Computer Supported Cooperative Work (CSCW) 21:163–225.
Servilla, M., J. Brunt, D. Costa, J. McGann, and R. Waide. 2016. The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. Ecological Informatics 36:247–258.
Shapin, S. 1994. A social history of truth: civility and science in seventeenth-century England. University of Chicago Press, Chicago, Illinois, USA.
Smith, L. M., et al. 2018. The ocean observatories initiative. Oceanography 31:16–35.
Steinhardt, S. B. 2016. Breaking down while building up: design and decline in emerging infrastructures. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, New York, New York, USA.
Thorpe, H. H., M. Skipper, V. Kiermer, M. Berenbaum, D. Sweet, and R. Horton. 2019. Joint Statement on EPA proposed rule and public availability of data. Science. https://doi.org/10.1126/science.aba3197
Van Tuyl, S., and A. L. Whitmire. 2016. Water, water, everywhere: defining and assessing data sharing in academia. PLOS ONE 11:e0147942.
Van Noorden, R. 2015. Sluggish data sharing hampers reproducibility effort. Nature News. https://doi.org/10.1038/nature.2015.17694
Vines, T. H., et al. 2014. The availability of research data declines rapidly with article age. Current Biology 24:94–97.
Virginia Tech. 2019. EPA’s proposed rule raises concerns and hinders ability to carry out its mission, says Virginia Tech data management expert. Virginia Tech Daily. https://vtnews.vt.edu/articles/2019/11/epa-expert.html
Waide, R. B., and M. O. Thomas. 2013. Long-Term Ecological Research Network. Pages 233–268 in
Wallis, J. C., C. L. Borgman, M. S. Mayernik, and A. Pepe. 2008. Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research. International Journal of Digital Curation 3:114–126.

Wallis, J. C., M. S. Mayernik, C. L. Borgman, and A. Pepe. 2010. Digital libraries for scientific data discovery and reuse. Proceedings of the 10th annual Joint Conference on Digital Libraries, JCDL ’10. ACM Press, New York, New York, USA.

White, H. D. 1982. Citation analysis of data file use. Library Trends 31:467–477.

Wilkinson, M. D., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3:160018.

Wolins, L. 1962. Responsibility for raw data. American Psychologist 17:657–658.

Zins, C. 2007. Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology 58:479–493.