Making Data Work: A Systematic Mapping of Collaborative Data Curation Practices

Nana Kwame Amagyei[1], Elena Parmiggiani[1], Babak A. Farshchian[1], and Jostein Engesmo[1]

[1] Norwegian University of Science and Technology, NTNU, NO-7491
Trondheim, Norway

Abstract. A growing body of literature in Information Systems focuses on the collaborative data curation practices that support the use of novel technologies in the ongoing datafication of work and organizing. In this study, we map the practices and processes that help make data useful and meaningful so that organizations can take advantage of these technologies. We examine 54 empirical studies and focus on the individuals and groups that collaborate to make data useful and meaningful. We identify the following collaborative data curation practices: (i) engaging multiple users in cooperation, (ii) involving higher-level stakeholders, and (iii) using shared resources. We contribute to the IS literature by broadening the view of data curation as an organizational practice that requires the collective, situated, and ongoing engagement of multiple actors making flexible and interpretive decisions to identify and resolve challenges related to working with data.

Keywords: Data Curation, Systematic Mapping, Practices, Collaboration, Data work

1 Introduction

The ongoing datafication is currently perceived as an opportunity for innovation and more effective decision-making [1]. Despite enthusiastic calls for the potential of Big Data, researchers in Information Systems (IS) have shown that business problems are situated in local contexts [2] and much work is involved into discovering, setting up, preparing, and sharing the data before they can be used to inform business decisions [3, 4]. As a consequence, extracting value from data is a sociotechnical endeavor involving technical experts (e.g., database managers, data scientists, software engineers), domain experts (e.g., business leaders, environmental scientists, medical practitioners) [5, 6], as well as new emerging professionals tasked with working with the data (e.g., messengers, interpreters) [7, 8].

Organizations typically address this trend by introducing overarching data governance frameworks that are aimed at ensuring and tracing the accessibility, consistency, and usability of data throughout their lifecycle [9, 10]. Such frameworks tend to be prescriptive and are often portrayed as technocentric – performed using technological infrastructures without the engagement of other domain experts as they collaborate with technical experts. This tendency is accompanied by warning signs by researchers calling for a better understanding of actual work practices and users' involvement along the value chain [11, 12].

Similar initiatives are taken in the realm of science, where funding bodies introduce funding policies that require scientists to share their data openly across organizational boundaries to receive funding [13] in line with the Open Science agenda and the FAIR data principles [14, 15]. The case of science is useful in this respect, because it reminds us that, as data are used across multiple sites, significant effort is needed to ensure that data remain meaningful and useful over time [16] [17].

Data curation is an increasingly important work practice to support data-related activities within and across organizational boundaries. While definitions of data curation vary, researchers agree that it involves the ongoing effort to select, organize and manage data as an organizational resource [18, 19]. A notable MIS Quarterly research curation defines data curation as involving categorizing and organizing data so they can be easily shared, and emphasize the physical and logical infrastructures that make it feasible to collect, index, and store data, and facilitate data access for subsequent analysis [19].

Data curation is considered an important organizational practice for at least two reasons. First, data governance frameworks and open data sharing policies require that data be findable, accessible, interoperable, and reusable [14], thus challenging organizations to know exactly what data to share, how to share data, where shared data are stored, how their quality is maintained, how they are organized and used, who can access them and for which purposes. Second, the availability of very large datasets challenges the work to handle their variety while ensuring sufficient trustworthiness for further reuse [20].

Unfortunately, dominant characterizations of digitalization tend to overlook the nuanced collective multidisciplinary efforts to collect, index, and store data, and facilitate data access for subsequent analysis. This warrants closer examination of how interdisciplinary experts collaborate to organize data as an organizational resource. This motivates us to undertake a systematic mapping of existing studies that present collaborative data curation practices in IS and the neighboring academic fields. Our aim here is to understand *what is currently known about collaborative data curation practices.*

We find that data curation is a complex phenomenon and is punctuated by practices of domain experts and other involved stakeholders to identify and resolve data issues over time. We contribute to the IS literature by extending the view of data curation as collective, situated, and ongoing engagement of different stakeholders who flexibly and interpretively make decisions to identify challenges and resolve them to reach an optimal outcome for all stakeholders involved.

2 Theoretical Background

In the organizational and IS literature, data have often been viewed as raw material that can be abstracted from the world [20, 21] to unlock the inherent potential of emerging datafication technologies such as Big Data, artificial intelligence, machine learning and related data analytics tools. For example, IS researchers have shown that Big Data enable more accurate insights that lead to better information and better decision making at operational levels [22, 23]. Others demonstrate that such technologies enable new strategic positioning and competitive advantage [24]. Researchers have also shown that these technologies create a foundation for radical innovation in organizations and industries [25, 26]. A common theme running through these studies is "optimistic assumptions" about a new and progressive digital era – where the primary objective for all organizations is to unlock the potential of datafication technologies.

Recent work has begun to advocate for examining the collaborative data curation practices through which data realizes its potential value as an organizational resource [27]. These studies demonstrate significant correlations between organizational culture and successful use of emerging technologies [28]. They show organizational data use as an inherently collective action that depends on interdisciplinary domain experts who adopt collaborative strategies, despite their diverging professional or ideological perceptions of data [29]. From this perspective, collaborative data curation practices seek to improve data quality, filter irrelevant data and ensure protection of organizational data [12].

Collaborative data curation practices unpack the conflicts and tensions in working with emerging technologies [30]. The approach leverages human intelligence by involving different professionals to solve the problems of bias, transparency, accountability, and quality [31] associated with working with data. For example, emerging technologies often place high demands on the quality of input data; including, correct labeling, complete data, and detectable noise [32]. Sambasivan and colleagues [33] also show the high prevalence of negative impact for artificial intelligence systems caused by underestimation of data quality [33].

Given the increasing reliance on data, organizations that want to leverage emerging technologies need high quality data [31]. However, achieving such data quality to meet organizational goals requires an understanding of the situated nature of data-related activities [34] – including the methods, infrastructures, technologies, skills, and knowledge developed to handle data.

To address this, we draw on Nicolini's [35] studies on practice theory. Practice theory assumes that issues such as institutions, identity, interests, tensions, conflicts, power, inequalities, or change result from and are mediated by human practices and their aggregates. It foregrounds situated, observable, and meaningful social events performed through language, bodily movements, and the contribution of material artifacts such as data and technologies. From this perspective, team members' practices in collaborative data curation and the ways in which their expertise, skills, and competencies support the resolution of concerns, conflicts, and tensions in using emerging technologies and working toward specific data-related outcomes become imperative.

We draw on these ideas, to analyze existing literature on currently known practices of collaborative data curation to gain a more comprehensive understanding of data curation: not just as a data-related activity that uses technologies and algorithms, but as an inherent organizational practice.

3 Method

We used the systematic mapping method for this study [36]. Systematic mapping is a method for analyzing existing literature in a broad research area. A systematic mapping study (SMS) differs in its objectives from the more familiar systematic literature review. A systematic literature review is an in-depth investigation of a narrow area with specific or narrowly defined research questions [36]. The goal of a systematic literature review (SLR) is to generate new knowledge through a meta-analysis of the existing literature. SLRs use methodological quality as inclusion criteria when searching for and including literature. We realized that SMS does not normally include an in-depth analysis of the papers although in our case we have gone further into analyzing the papers in detail.

3.1 Define

Our research question was inspired by our interest to better understand data curation practices in organizations. After an initial exploratory phase, we eventually defined our research question as: "what is currently known about collaborative data curation practices?". Next, we defined our inclusion/exclusion criteria as follows:

Table 1. Inclusion and Exclusion criteria

Inclusion	Exclusion
• English studies • Scientific journal articles and conference proceedings • Studies that present empirical findings on collaborative data curation practices • Studies published in Information Systems, Computer Science, Computer Supported Cooperative work, Science and Technology Studies, and Human-Computer Interaction	• Studies that do not present empirical findings on collaborative data curation practices • Studies that report on design of new curation tools

3.2 Search

We searched and obtained the relevant articles in the Scopus digital library. This library is among the largest abstract and citation database for peer-reviewed literature with bibliometric tools for tracking, analyzing and visualizing research [38]. The first step in our search phase was to collectively develop a search strategy. This was achieved by developing alternative search keywords through several iterations. All authors examined the results of each iteration and collectively refined the search keywords during weekly meetings to ensure that we had included the relevant keywords used in the literature to capture collaborative data curation practices. For example, we listed the four keywords from our research questions as follows: data, curation, collaborative and practices. We then agreed on synonyms for each of these keywords, based on existing literature and our own experiences.

3.3 Select

In selecting articles, we used two techniques to ensure that our results considered relevant available studies, i.e., selecting from results of the library search, and manually selecting eight relevant articles that did not show up in our library search, but which we considered useful for our study. Figure 1 summarizes the selection process. In total, we obtained 54 articles that were exported to the NVivo computer-assisted coding software for further analysis.

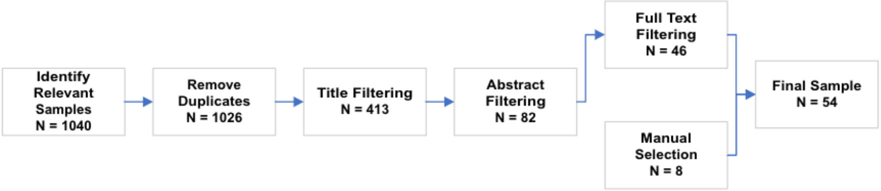
Identify Relevant Samples N = 1040
Remove Duplicates N = 1026
Title Filtering N = 413
Abstract Filtering N = 82
Full Text Filtering N = 46
Manual Selection N = 8
Final Sample N = 54

Fig. 1. Article selection process

3.4 Analyze and Present

We applied an open coding technique [39] to the findings section of the articles. We marked words, sentences, or paragraphs that represented key collaborative practices from the articles and grouped them under the label “extract”. We then grouped the extracts to define concepts. Concepts were further grouped into themes. To increase validity, the concepts were reviewed by all authors to ensure they did not deviate too much from the terminology used in existing literature (*method details in appendix*).

4 Findings

Authors who report on collaborative practices in data-related activities use terms such as data work [40], data curation [12], data science [41], data modeling and visualization [33]. Yet, in our view, these authors report on a consistent theme:

> *"Empirical insights into the broad range of decentralized practices in collecting, generating, producing, cleaning, assembling, setting up, analyzing, modeling, visualizing, and stewarding data toward specific goals; to be used by computational and statistical techniques from areas such as machine learning, data mining, natural language processing, and artificial intelligence."*

Since we are not concerned in this study with establishing a difference in the use of the different terminologies reported in the literature, we use the term "data curation" to refer to all forms of data-related activities as team members collaborate with one another on data to resolve tensions and achieve specific organizational goals, such as sharing [42], advertising [4], teaching, and learning [43].

Collaborative data curation practices relate to the idea of complete decentralization, where there is no central authority that determines and coordinates decision making, but a spontaneous emergence of self-emerging individual decision-making units that are formally independent make mutual arrangements to resolve data issues [44]. For example, Van Den Broek and colleagues show how members of the human resource (HR) department tried to convince employees and their representatives to provide the required data needed by data scientists to train an AI tool for making more effective hiring decisions [45]. The authors show how such practices emerge to negotiate divergent interests between different stakeholders to pave way for artificial intelligence tools to be developed and adopted by the organization. Also, Waardenburg and colleagues distinguish office-based data curation from situated practices and unpack how police officers cope daily with emerging tensions. According to the authors, police officers experience the use of technologies for reporting as constrained by the body, materially rigid and ethereal while they experience situated practice of working collaboratively with humans in crime scenes as embodied, contextual, and lived [46]. Highlighting that no central or coordinating authority determines where local workers direct their attention, but rather ongoing activities inevitably shape local workers' decisions.

Essentially, individuals and groups learn about threats and problems related to data and solve them together through collaborative data curation practices. Collaborative data curation practices, then, means learning in the face of change and uncertainty by progressively defining and adapting rules, plans, and frameworks to fit local conditions, with self-correcting mechanisms for monitoring and compliance. We elaborate on these collaborative data curation practices in more detail.

4.1 Engaging multiple users in cooperation

To engage multiple users in cooperation means the ability of individuals and groups to interact with each other and use their own judgment to carry out data-related activities in complex and uncertain problem areas. Key issues in carrying out data-related tasks require team members to acknowledge the value of adapting previous individual-level actions in favor of new collaborative data arrangements since these ultimately provide higher benefits. Mosconi and colleagues show how doctoral students using a digitally accessible repository needed to understand what others were doing to improve their individual sharing practices and gain benefits from sharing their data [47]. Highlighting the relational nature of data and technologies with the social environment within which they are created and used [34].

When individuals work together with data, they expect to develop a sense of what others are saying and doing in a way that is mutually engaging since "a big part of having teams work more effectively together is to provide more situational awareness of data-related activity workflows and who has done which task" [48]. Suggesting that overcoming problems in working with data requires arrangements that allow individuals the ability to hear or see how others work with data. This can encourage multiple users to engage in creative and cooperative ways to improve the potential value of the data-related activity [43].

Collaborating on data does not only entail knowledge of *what* others are saying or doing but also *when* others say or do things. Some data workers take the approach of sharing different runs of their model *after* tuning and choosing the best run [49]. Since engaging others too soon or too late may compromise the quality outcome of a data-related task. Local workers therefore prefer to engage other team members only after their models are fine-tuned [49].

Because of the interpretive nature of data-related tasks research suggests that individuals may have some bias toward particular users, ideas, or things [49–54]. The literature often uses terms such as bias, need for transparency, and accountability to describe this data-related problem. Ongoing research thus focuses on the subjectivity of data workers to account for such data-related problems and calls for more research to understand how to constrain workers' subjective judgment on data-related tasks and work outcomes [55].

4.2 Involving higher-level stakeholders

Researchers often locate data-related problems of bias, transparency, and accountability in the use of technical systems: either data, technologies, or algorithms. If one locates data-related problems with the technologies, the algorithms, or the people working with the data, one might propose a simple solution: cleaning and augmenting training datasets with more diverse data [56]. However, research on collaborative data curation practices argues that such approaches could be augmented by a relational view of the power dynamics and economic imperatives driving the use of new technologies [34].

A power-oriented perspective examines technical systems and the relationships that intervene in the data-related task of producing and using data, models, and visualizations. It emphasizes a shift in perspective away from data-related problems arising only from technologies, algorithms, individuals, and groups working in local data contexts to a broader analysis of the influences and power relations associated with data work. For example, accurate facial recognition used for surveillance has been shown to be dangerous in the hands of repressive governments [57]. Consequently, data-related issues are more about power dynamics as much as technology and local workers' practices [58].

Collaborating on data often requires workers to align and adjust their activities in relation to such power relations. Questions for analyzing higher-level stakeholders may include: what specific actions taken during a data-related activity are in line with organizational data policies, how are expectations of data governance frameworks, security policies or data management plans enacted in specific use contexts? For example, a project manager indicated that from his years of experience with discussing models, he had learned that "stakeholders want to see colors and ranges of an aggregate measure like accuracy. Just red, yellow, or green." [49]. This project manager was sharing his frustration about his teams' effort to share more metrics about their work outcomes to help high-level stakeholders, i.e., the client/requester, because the results his team produced (i.e., use their expertise to develop models) was not in line with what the client wanted to see (i.e., use colors to understand model accuracy).

Team members usually organize regular team meetings to jointly deliberate and resolve such data issues. For example, team members attempt to address open data sharing – a requirement by funding agencies for research stations to publicly share data from their science results – by organizing joint sessions to decide whether data is ready to be made openly accessible [59]. Extending data-related tasks with a power-aware perspective could make power asymmetries visible and raise awareness about meaning, subjectivities, impositions, and naturalization [55].

Data-related tasks such as defining computational problems [2], selecting training datasets, measurement interfaces for data collection [34], creating taxonomies for data labeling [60], and designing traceability in AI systems [51] are all choices that are hardly ever made by individual choice or in a vacuum. Instead, they involve power structures and depend on agendas, budgets and revenue plans [56].

Most importantly, power-oriented inquiry might allow researchers to move beyond a simplistic view that assigns responsibility for data-related problems exclusively to data workers and instead interrogate the power relations that inscribe particular forms of knowledge into emerging technologies. Accordingly, data production, design activities, and decisions are influenced not only by data workers but also by data requesters, regulators, funding agencies, and other external stakeholders. A perspective that helps to see such higher-level stakeholders as co-creators of data rather than mere consumers [56]

4.3 Using shared resources

Studies show that individuals and groups leveraging the potential of data do not only require a range of technologies and network resources, but also an understanding of the metrics, theories, and concepts shared by the various interdisciplinary experts during data-related activities [61]. Key issues to using data as an organizational resource include the use of shared resources such as image data, sonographs, theories, instruments, and a wide range of artifacts in ways that support the data-related task.

Collaborative data curation requires a wide range of standards, theories, methods, tools, and technologies [34]. For example, data managers collaborate with oil explorationists and in some cases farmers with adequate local knowledge about where to position sensor devices to adequately record the Artic seabed and make informed decisions on oil exploration [62]. Similarly, sonographers routinely interact with medical practitioners in their use of sonographs to produce information in the form of images and other signs and symbols from which medical practitioners construct diagnoses and prescribe treatment to patients [61]. In both cases, workers use theories and concepts in ways that support reaching an optimal work outcome. These theories and concepts act as a means for workers to adapt to the situated needs of work and reach an optimal work outcome.

In some cases, workers are responsible for ensuring that instrumentation for capturing data is maintained and remains in good working condition. This requires technicians to draw on shared formal procedures, theories, and informal interactions with other domain experts as well as their own experiences [61, 63]. For example, technicians must understand the materials – such as the continuous plankton recorder (CPR) – for sampling phytoplankton if they are to repair malfunctioning CPRs for sampling to support environmentalists in the task of collecting phytoplankton data samples. Also, technicians responsible for ensuring that sensor devices are positioned in the forest to monitor various animal species interact with developers to understand how these sensor devices are to be maintained and kept in good shape over time.

Subsequently, local workers, such as data managers, environmental/medical/laboratory technicians, and research assistants, are guided by the concepts, objectives and directives, methods, tools shared by interdisciplinary domain experts such as climate scientists, medical practitioners, environmental scientists, and software engineers to improve outcomes of a data-related activity.

5 Discussion

In this study, we set out to systematically map *what is currently known about collaborative data curation practices* in the IS literature and related fields. A common thread running through our findings is the recognition that technological advances have enabled more data to be collected, stored, and processed, requiring organizations to focus on local practices of workers to harness the potential of data for organizational decision making. Such collaborative data curation practices require multiple people and forms of human work [64]. While more scholarly attention has been paid to governance frameworks and the technologies to realize value from data, an increasing body of IS research continue to examine the collaborative data curation practices. These practices have profound impacts on both technology design and human labor [65]. Several recent studies also point to the increased burden of such practices: ranging from physical to emotional burdens, to resolve tensions and engage with data in a datafied environment [40].

While promising, collaborative data curation practices are fraught with challenges related to recruiting appropriate professionals [40], determining appropriate methods for improving data quality [66], and a lack of understanding of the actual work and full scope of curating data [67]. This study aimed at providing an understanding of such collaborative data curation practices that are increasingly gaining IS researchers' attention.

We find that collaborative data curation practices require humans to make many situational and discretionary decisions – sometimes controversial ones at other times straightforward ones – as they use data, technologies, and algorithms. Our findings highlight that harnessing the potential of data cannot be left to traditional technical departments and data scientists [45]. Rather, organizations need an overarching approach to coordinate and organize data-related activities in ways that *engage every employee in cooperation, aware of power dynamics and attentive to invisible forms of work and shared resources*. From this perspective, the actual work of using emerging datafication technologies, such as Big Data and Artificial Intelligence, focuses on resolving concerns related to individual, team and higher-level stakeholders, as well as those related to which algorithms, technologies, methods, expertise and skills are needed [63].

Studies on collaborative data curation practice also draw attention to the wide range of sociotechnical practices of data production and use. Studying collaborative data curation practices therefore requires an examination of the broad range of formal and informal practices employed by the workers who work in the local context of data-related activities [40]. Further, data curation requires both physical and logical infrastructures to manage and facilitate data accessibility [19]. Suggesting that data curation is a critical part of information infrastructure studies [2, 68, 69] and includes the functional elements (i.e., people, practices, policies) as well as actual elements (i.e., technologies, tools, data) that enable organizations to realize value from data in data infrastructures.

Two implications for data governance emerge from our findings, first, given data privacy regulations and data sharing policies organizations are encouraged to not only develop policies to invest in technologies or data management plans for preserving and sharing organizational data respectively, but also to focus on the actual data handling practices to maintain privacy and preserve data; including the methods, capabilities, and knowledge that are developed now and, in the future to handle data [44]. This may provide organizations the chance to adopt a management-oriented approach to organizational data governance and faithfully capture and represent the complex, diverse, and evolving structures, and behaviors within the organization [66].

Second, resources for obtaining value from data, including data management plans, cannot be fully accounted for in prescriptive data governance frameworks, as a result funding agencies are encouraged to create room to address the ongoing physical, emotional, and ethical burdens of individuals. This suggests that in addition to deploying emerging technologies, data governance frameworks and data-savvy managers and staff, organizations are encouraged to empower marginalized roles and capabilities that are typically excluded from technological and socio-economic development (for example, by supporting their daily processes of data curation tasks with the necessary reward structures).

For practical purposes, organizations are encouraged to schedule time and identify local contexts within which data-related activities occur so that they can learn from mistakes and help shape data-related projects and its outcomes. Furthermore, organizations are encouraged to be aware of the different knowledge groups, recognize the importance of mutual respect for these roles, and find opportunities to learn about different knowledge domains.

In summary, if data curation is understood as a situated and ongoing collaborative practice among individuals and groups, then datafication outcomes can be improved by resolving sociopolitical, economic, health, ethical, emotional, cultural, and technological concerns in ways that help organizations and the people who work with the data to achieve more quality outcomes.

6 Conclusions, limitations, and future work

This paper offered a synthesis of empirical insights into the day-to-day realities of people making decisions, pursuing their interests, resolving tensions, and working with data as an organizational resource. Our findings acknowledge that data curation has evolved into a complex phenomenon that requires collective, discretionary, situational, and ongoing engagement of diverse actors to make flexible and interpretive decisions to identify and resolve data-related issues.

We recognize that the methods used in this study could be improved and do not claim this to be a comprehensive assessment of the literature on collaborative data curation practices. However, we hope that this study provides a new perspective on the emerging role of collaborative data curation practices as a useful organizational practice. Opportunities and key questions for IS literature to answer in the future include: *what data curation practices are employed to support the use of data as an organizational resource? How does data curation support long-term and unknown future data uses? How do different team members coordinate their activities to create sustainable data infrastructures? What skills do different team members need to collaborate on improving the quality of data? What new sociotechnical infrastructures are emerging in the era of Big Data? How does situated data curation practices support technology design? What are the ethical and political concerns associated with curating healthcare data? How does data curation relate to the results of data analytics? What new roles arise from adopting a data curation approach to governance?*

References

1. Günther, W.A., Rezazade Mehrizi, M.H., Huysman, M., Feldberg, F.: Debating big data: A literature review on realizing value from big data. J. Strateg. Inf. Syst. 26, 191–209 (2017). https://doi.org/10.1016/j.jsis.2017.07.003.
2. Monteiro, E., Parmiggiani, E.: Synthetic Knowing: The Politics of the Internet of Things. MIS Q. 43, 167–184 (2019). https://doi.org/10.25300/MISQ/2019/13799.
3. Parmiggiani, E., Østerlie, T., Almklov, P.G.: In the Backrooms of Data Science. J. Assoc. Inf. Syst. 23, 139–164 (2022). https://doi.org/10.17705/1jais.00718.
4. Aaltonen, A., Alaimo, C., Kallinikos, J.: The Making of Data Commodities:

Data Analytics as an Embedded Process. J. Manag. Inf. Syst. 38, 401–429 (2021). https://doi.org/10.1080/07421222.2021.1912928.

5. Mikalef, P., Pappas, I.O., Krogstie, J., Pavlou, P.A.: Big data and business analytics: A research agenda for realizing business value. Inf. Manag. 57, 103237 (2020). https://doi.org/10.1016/j.im.2019.103237.
6. Slota, S.C., Hoffman, A.S., Ribes, D., Bowker, G.C.: Prospecting (in) the data sciences. Big Data Soc. 7, 205395172090684 (2020). https://doi.org/10.1177/2053951720906849.
7. Waardenburg, L., Huysman, M., Sergeeva, A. V.: In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. Organ. Sci. 33, 59–82 (2022). https://doi.org/10.1287/orsc.2021.1544.
8. Bossen, C., Chen, Y., Pine, K.H.: The emergence of new data work occupations in healthcare: The case of medical scribes. Int. J. Med. Inform. 123, 76–83 (2019). https://doi.org/10.1016/j.ijmedinf.2019.01.001.
9. Otto, B.: Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. Commun. Assoc. Inf. Syst. 29, (2011). https://doi.org/10.17705/1CAIS.02903.
10. Tallon, P.P., Ramirez, R. V., Short, J.E.: The Information Artifact in IT Governance: Toward a Theory of Information Governance. J. Manag. Inf. Syst. 30, 141–178 (2013). https://doi.org/10.2753/MIS0742-1222300306.
11. Iivari, J., Isomäki, H., Pekkola, S.: The user - the great unknown of systems development: reasons, forms, challenges, experiences and intellectual contributions of user involvement. Inf. Syst. J. 20, 109–117 (2010). https://doi.org/10.1111/j.1365-2575.2009.00336.x.
12. Parmiggiani, E., Grisot, M.: Data Curation as Governance Practice. Scand. J. Inf. Syst. 32, (2020).
13. ESFRI: Making Science Happen: A new ambition for Research Infrastructures in the European Research Area, https://www.esfri.eu/sites/default/files/White_paper_ESFRI-final.pdf, last accessed 2020/11/08.
14. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.
15. Friesike, S., Widenmayer, B., Gassmann, O., Schildhauer, T.: Opening science: towards an agenda of open science in academia and industry. J. Technol.

Transf. (2015). https://doi.org/10.1007/s10961-014-9375-6.

16. Ribes, D., Polk, J.: Flexibility Relative to What? Change to Research Infrastructure. J. Assoc. Inf. Syst. 15, 287–305 (2014). https://doi.org/10.17705/1jais.00360.
17. Zhao, Z., Hellström, M. eds: Towards Interoperable Research Infrastructures for Environmental and Earth Sciences. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-52829-4.
18. Leonelli, S.: Data-Centri Biology: A philosophical study. University of Chicago Press (2016). https://doi.org/https://doi.org/10.7208/9780226416502.
19. Chua, C., Indulska, M., Lukyanenko, R., Wolfgang, M., Storey, V.C.: Data Management, https://www.misqresearchcurations.org/blog/2022/2/11/data-management, last accessed 2022/03/26.
20. Kitchin, R.: The data revolution: Big data, open data, data infrastructures and their consequences. Sage Publications (2014).
21. Gitelman, L., Jackson, V.: Introduction. In: “Raw data” is an oxymoron (2013).
22. H. Davenport, T.: How strategists use “big data” to support internal business decisions, discovery and production. Strateg. Leadersh. 42, 45–50 (2014). https://doi.org/10.1108/SL-05-2014-0034.
23. DalleMule, L., Davenport, T.H.: “What’s Your Data Strategy?”
24. Elia, G., Raguseo, E., Solazzo, G., Pigni, F.: Strategic business value from big data analytics: An empirical analysis of the mediating effects of value creation mechanisms. Inf. Manag. 59, 103701 (2022). https://doi.org/10.1016/j.im.2022.103701.
25. Su, X., Zeng, W., Zheng, M., Jiang, X., Lin, W., Xu, A.: Big data analytics capabilities and organizational performance: the mediating effect of dual innovations. Eur. J. Innov. Manag. 25, 1142–1160 (2022). https://doi.org/10.1108/EJIM-10-2020-0431.
26. Marshall, A., Mueck, S., Shockley, R.: How leading organizations use big data and analytics to innovate. Strateg. Leadersh. 43, 32–39 (2015). https://doi.org/10.1108/SL-06-2015-0054.
27. Mikalef, P., Boura, M., Lekakos, G., Krogstie, J.: The role of information governance in big data analytics driven innovation. Inf. Manag. 57, 103361 (2020). https://doi.org/10.1016/j.im.2020.103361.
28. Eikebrokk, T.R., Nilsen, R.E., Garmann-Johnsen, Frederik, N.: Exploring the role of process orientation in healthcare service innovation: the case of digital night surveillance. In: AMCIS 2017 Proceedings (2017). https://doi.org/http://hdl.handle.net/10125/59749.
29. Karlsen, C., Haraldstad, K., Moe, C.E., Thygesen, E.: Challenges of Mainstreaming Telecare. Exploring Actualization of Telecare Affordances in Home Care Services. In: Scandinavian Journal of Information Systems (2019).
30. Stang Våland, M., Svejenova, S., Clausen, R.T.J.: Renewing creative work for business innovation: Architectural practice in the trading zone. Eur. Manag. Rev. 18, 389–403 (2021). https://doi.org/10.1111/emre.12464.
31. Toussaint, P.J., Melby, L., Hellesø, R., Brattheim, B.J.: Does Information Quality matter? In: CEUR Workshop Proceedings (2017).

32. von Krogh, G.: Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. Acad. Manag. Discov. 4, 404–409 (2018). https://doi.org/10.5465/amd.2018.0084.

33. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L.M.: “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–15. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3411764.3445518.

34. Leonelli, S.: What difference does quantity make? On the epistemology of Big Data in biology. Big Data Soc. (2014). https://doi.org/10.1177/2053951714534395.

35. Nicolini, D.: Practice theory, work, and organization: An introduction. OUP Oxford (2012).

36. Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: An update. Inf. Softw. Technol. 64, 1–18 (2015). https://doi.org/10.1016/j.infsof.2015.03.007.

37. Farshchian, B.A., Dahl, Y.: The role of ICT in addressing the challenges of age-related falls: a research agenda based on a systematic mapping of the literature. Pers. Ubiquitous Comput. 19, 649–666 (2015). https://doi.org/10.1007/s00779-015-0852-1.

38. Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R.: Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quant. Sci. Stud. 1, 377–386 (2020). https://doi.org/10.1162/qss_a_00019.

39. Wolfswinkel, J.F., Furtmueller, E., Wilderom, C.P.M.: Using grounded theory as a method for rigorously reviewing literature. Eur. J. Inf. Syst. 22, 45–55 (2013). https://doi.org/10.1057/ejis.2011.51.

40. Pine, K., Bossen, C., Holten Møller, N., Miceli, M., Lu, A.J., Chen, Y., Horgan, L., Su, Z., Neff, G., Mazmanian, M.: Investigating Data Work Across Domains. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–6. ACM, New York, NY, USA (2022). https://doi.org/10.1145/3491101.3503724.

41. Passi, S., Sengers, P.: Making data science systems work. Big Data Soc. 7, 205395172093960 (2020). https://doi.org/10.1177/2053951720939605.

42. Zimmerman, A.S.: New Knowledge from Old Data. Sci. Technol. Hum. Values. 33, 631–652 (2008). https://doi.org/10.1177/0162243907306704.

43. Passi, S., Jackson, S.J.: Data vision: Learning to see through algorithmic abstraction. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW (2017). https://doi.org/10.1145/2998181.2998331.

44. Leonelli, S.: Data - from objects to assets. Nature. 574, 317–320 (2019). https://doi.org/10.1038/d41586-019-03062-w.

45. Van Den Broek, E., Levina, N., Sergeeva, A.: In Pursuit of Data: Negotiating Data Tensions Between Data Scientists and Users of AI Tools. Acad. Manag. Proc. 2022, (2022). https://doi.org/10.5465/AMBPP.2022.182.

46. Waardenburg, L., Sergeeva, A., Huysman, M.: Juggling Street Work and Data

Work: An Ethnography of Policing and Reporting Practices. Acad. Manag. Proc. 2022, (2022). https://doi.org/10.5465/AMBPP.2022.215.

47. Mosconi, G., Li, Q., Randall, D., Karasti, H., Tolmie, P., Barutzky, J., Korn, M., Pipek, V.: Three Gaps in Opening Science. Computer Supported Cooperative Work (CSCW) (2019). https://doi.org/10.1007/s10606-019-09354-z.

48. Crisan, A., Fiore-Gartland, B.: Fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–15. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3411764.3445775.

49. Almahmoud, J., DeLine, R., Drucker, S.M.: How Teams Communicate about the Quality of ML Models: A Case Study at an International Technology Company. Proc. ACM Human-Computer Interact. 5, 1–24 (2021). https://doi.org/10.1145/3463934.

50. Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M., Ju, L., Su, H.: Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. pp. 297–307. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3377325.3377501.

51. Kroll, J.A.: Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 758–771. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445937.

52. Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., Hanna, A.: Documenting Computer Vision Datasets. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 161–172. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445880.

53. Passi, S., Jackson, S.J.: Trust in Data Science: Collaboration, translation, and accountability in corporate data science projects. Proc. ACM Human-Computer Interact. 2, 1–28 (2018). https://doi.org/10.1145/3274405.

54. Thornton, L., Knowles, B., Blair, G.: Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 64–76. ACM, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445871.

55. Miceli, M., Schuessler, M., Yang, T.: Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. In: Proceedings of the ACM on Human-Computer Interaction. pp. 1–25 (2020). https://doi.org/10.1145/3415186.

56. Miceli, M., Posada, J., Yang, T.: Studying Up Machine Learning Data. Proc. ACM Human-Computer Interact. 6, 1–14 (2022). https://doi.org/10.1145/3492853.

57. D’Ignazio, C., Klein, L.F.: Data feminism. Strong ideas series. The MIT Press,

Cambridge, Massachusetts (2020).

58. Miceli, M., Schuessler, M., Yang, T.: Between Subjectivity and Imposition. Proc. ACM Human-Computer Interact. 4, 1–25 (2020). https://doi.org/10.1145/3415186.
59. Hoeppe, G.: Encoding Collective Knowledge, Instructing Data Reusers: The Collaborative Fixation of a Digital Scientific Data Set. Comput. Support. Coop. Work. 30, 463–505 (2021). https://doi.org/10.1007/s10606-021-09407-2.
60. Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Vera Liao, Q., Dugan, C., Erickson, T.: How data science workers work with data. In: Conference on Human Factors in Computing Systems - Proceedings (2019). https://doi.org/10.1145/3290605.3300356.
61. Barley, S.R., Bechky, B.A.: In the Backrooms of Science. Work Occup. 21, 85–126 (1994). https://doi.org/10.1177/0730888494021001004.
62. Parmiggiani, E.: This Is Not a Fish: On the Scale and Politics of Infrastructure Design Studies. Comput. Support. Coop. Work. 26, 205–243 (2017). https://doi.org/10.1007/s10606-017-9266-0.
63. Borgman, C.L., Wofford, M.F., Golshan, M.S., Darch, P.T., Scroggins, M.J.: Collaborative ethnography at scale: reflections on 20 years of data integration. Presented at the (2020).
64. Møller, N., Claus, B., Pine, K., Nielsen, T., Neff, G.: Who does the work of the data?, http://interactions.acm.org//archive/view/may-june-2020/who-does-the-work-of-data, last accessed 2020/04/06.
65. Jones, M.: What we talk about when we talk about (big) data. J. Strateg. Inf. Syst. (2019). https://doi.org/10.1016/j.jsis.2018.10.005.
66. Leonelli, S., Tempini, N.: Data Journeys in the Sciences. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-37177-7.
67. Karasti, H., Baker, K.S., Halkola, E.: Enriching the notion of data curation in e-Science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. Comput. Support. Coop. Work. 15, 321–358 (2006). https://doi.org/10.1007/s10606-006-9023-2.
68. Ciborra, C.U., Hanseth, O.: From tool to Gestell: Agendas for managing the information infrastructure. Inf. Technol. People. 11, 305–327 (1998). https://doi.org/10.1108/09593849810246129.
69. Monteiro, E., Pollock, N., Hanseth, O., Williams, R.: From artefacts to infrastructures. Comput. Support. Coop. Work CSCW An Int. J. (2013). https://doi.org/10.1007/s10606-012-9167-1.

Appendix

For a detailed appendix, kindly visit https://doi.org/10.5281/zenodo.7223824