1-1-2023

Attending to the Cultures of Data Science Work

Lindsay Poirier
Smith College, lpoirier@smith.edu

Attending to the Cultures of Data Science Work

COLLECTION:
20 YEARS OF DATA SCIENCE

ESSAY

LINDSAY POIRIER

]u[ubiquity press

ABSTRACT

This essay reflects on the shifting attention to the "social" and the "cultural" in data science communities. While recently the "social" and the "cultural" have been prioritized in data science discourse, social and cultural concerns that get raised in data science are almost always outwardly focused – applying to the communities that data scientists seek to support more so than more computationally-focused data science communities. I argue that data science communities have a responsibility to attend not only to the cultures that orient the work of domain communities, but also to the cultures that orient their own work. I describe how ethnographic frameworks such as thick description can be enlisted to encourage more reflexive data science work, and I conclude with recommendations for documenting the cultural provenance of data policy and infrastructure.

CORRESPONDING AUTHOR:

Lindsay Poirier

Smith College, US

lpoirier@smith.edu

KEYWORDS:

data culture; data infrastructure; science and technology studies; ethnography; data documentation

TO CITE THIS ARTICLE:

Poirier, L. 2023. Attending to the Cultures of Data Science Work. *Data Science Journal*, 22: 6, pp. 1–7. DOI: https://doi.org/10.5334/dsj-2023-006

Studying an artificial intelligence lab focused on developing tools for knowledge representation in the early 1990s, anthropologist Diana Forsythe (1993) argued that computer scientists tend to 'delete the cultural' when rendering knowledge machine readable. In their work to translate knowledge into computer code, the scientists would seek to render visible the complex bodies of information that orient diverse communities, while expunging their own roles as knowledge curators and interpreters from consideration. The argument was an extension of a similarly framed argument from Susan Leigh Star (1991) that computer scientists tend to 'delete the social,' privileging technological concerns over social ones in their design practice. This essay will reflect on how discourse on the 'social' and the 'cultural' has evolved in the data science community, arguing that despite increased attention to sociocultural issues, data scientists tend to overlook the cultures orienting their own work. I then contrast these deletions with the methods that we engage in a course I teach at Smith College called Data Ethnography—a course that aims to foster underprioritized reflexive sensibilities in data science work. I conclude with recommendations for instituting protocols that would prompt researchers and practitioners to document the cultural provenance of their data science policies and infrastructure.

WHERE ARE THE 'SOCIAL' AND THE 'CULTURAL' IN DATA SCIENCE DISCOURSE?

Fast forward to the early 2020s and 'the social' has gained airtime in data science discourse. Perhaps the most common refrain from technologists that I have heard at data science conferences and meetings has been that the key challenges data science communities face 'are not technical but social' and that, as a community, we need to be focused on building social infrastructures in addition to technical ones. Admirably, over the past 20 years, data science organizations and journals have drawn attention to social barriers to data sharing, such as differing incentive structures and levels of training within and across diverse disciplinary communities, scholarly generations, and geographic boundaries (Borgman 2012; Gownaris et al. 2022; Laine 2017; van Panhuis et al. 2014). Research examining how to support the uptake of best practices, such as FAIR (**f**indable, **a**ccessible, **i**nteroperable, and **r**eusable) guidelines and data management planning, has shown that adoption not only depends on clear guidance and well-designed infrastructure but also social advocacy, relationship building, and an amenable financial, legal, and policy landscape (Wong et al. 2022). Movements to prioritize the rights and interests of Indigenous peoples in the knowledge economy have highlighted the need for alternative data governance models that privilege self-determination and collective benefits (Carroll et al. 2020).

Similarly, 'the cultural' has earned a place in data science discourse. I have been in countless data science meetings and conferences where I have heard communities reference the 'culture problem' data science faces. How do we bring about the 'culture change' necessary to facilitate data sharing (Barker et al. 2019; Klump 2017)? How do we get everyone to adopt the same standards or speak the same language? Or, alternatively, how do we develop the translational tools to map common meanings across different languages? It is notable that these concerns are often raised in the name of fairness, equity, and inclusion and have resulted in commendable educational efforts (Bezuidenhout et al. 2021; Dominik et al. 2022; Kouper et al. 2021).

While the 'social' and the 'cultural' have been prioritized in data science discourse, social and cultural concerns that get raised in data science are almost always outwardly focused—applying more to the communities that data scientists seek to support than computationally focused data science communities. This is supported by central organizing principles within data science projects, institutions, and policies. As David Ribes et al. (2019) argue, data science and other computational research communities are often seen as sitting independently from 'domains,' framing the communities as domain agnostic.

We see this guiding principle in countless settings. For instance, Ribes et al. (2019) detail how this logic has guided the funding policies at the US National Science Foundation and National Academy of Science. As data science curriculum builds out at numerous institutions, there is a tendency toward separating computational/analytical data science from more applied data science domains. Work in organizations like the Research Data Alliance tends to organically separate into domain tracks (focused on addressing more disciplinarily specific data infrastructure concerns) and nondomain tracks (focused on developing more domain-agnostic infrastructure and policies in support of domains).

Norms of interaction across 'domain' and 'nondomain' communities further buttress these roles. For instance, to develop data frameworks and infrastructures that can bridge diverse disciplinary cultures, folks in nondomain tracks will draw up use case templates and distribute them to multiple domain communities in an effort to learn more about what makes each community unique—that is, what their assumptions are, what their commitments are, and what unique disciplinary challenges they face. While prioritizing the acquisition of knowledge about other data cultures, we tend not to think about the cultures of the communities authoring the use case templates, interpreting the information collected from domain groups, and translating that information into infrastructure. This domain-agnostic positioning frames data scientists as neutral translators—as responsible for designing the tools to bridge across disparate social systems and cultures, mitigating 'data friction' (Edwards et al. 2011). What would it look like to turn an ethnographic focus back on the data science communities responsible for policy and infrastructure design? What are the stakes?

THICK DESCRIPTION FOR DATA SCIENCE

In his seminal work *The Interpretation of Cultures*, anthropologist Clifford Geertz (1973) presents a case for framing anthropology as a science in search of meaning. As an illustrative example, he asks us to consider how we discern the differences in meaning between a twitch of the eye and a wink. On a surface level, we see a wink as one eyelid closing and reopening; however, on an interpretive level, we categorize the movement as a mode of communication—indicating a joke, affection, or greeting. To ultimately discern the meaning of a wink, we have to take into consideration a number of contextual factors beyond the movement of the eyelid; it requires us to detect the symbolism in the action—to draw out its semiotics. The vehicle through which anthropologists move beyond surface-level observations of behaviors like a wink (and toward their contextualized interpretations) is thick description. Engaging thick description involves documenting detailed descriptions of behaviors or events and enriching those descriptions with interpretations of their symbolic cultural meaning.

Let me provide an example from a recent research project—a cultural analysis of semantic web infrastructure—to demonstrate how this applies to the data science community. Since the start of that project, I had been fascinated by debates around the meaning and flexibility of a property in many ontology languages that serves to indicate that one data point is the 'same as' another data point. I had been following discussions lamenting the misuse of the 'same as' property 'in the wild'—when everyday web users were leveraging the property to mark equivalence between two things that were not 'strictly' identical (Halpin et al. 2010). This issue, one conference paper argued, was leading to a logical 'crisis' of identity that was turning the interconnected web of data into 'the semantic equivalent of mushy peas' (Halpin et al. 2010: 308). Studying these concerns as a data ethnographer provided insight into the language ideologies that guided the design of semantic web infrastructure.

In 2016, while sitting at my desk in Troy, New York, and attempting not to allow email to pull me away from writing, an email with the subject line 'Deprecating owl:sameAs' pinged in my inbox. It was directed to the semantic web World Wide Web Consortium email list—where a great deal of planning for semantic web infrastructure had been occurring. The email read as follows:

> The research that I've done makes me conclude that we need to do a massive sweep of the LOD cloud and adopt owl:sameSameButDifferent.
>
> ...
>
> Such an owl:sameSameButDifferent statement indicates that two URI references actually refer to the same thing but may be different under some circumstances. (Capadisli 2016)

I remember being taken aback initially. Over the next few hours, more responses came flooding in, suggesting additional properties such as 'owl:differentDifferentButSame,' 'owl:isKindaLike,' 'owl:sometimesSameAs,' 'sameAsItEverWas,' and 'owl:actuallySameAsReally.'

It took me reading a few responses to recognize that the date was April 1 and that many (though notably not all) respondents were jumping on a satirical bandwagon to participate in an April Fool's joke.

Thick description enables an ethnographer to move from seeing this simply as an exchange of suggestions over email to discerning the significance of the humor. It gets us asking the question, Why is this funny, and what prods folks to join in on the joke? Writing on the significance of discerning irony when ethnographically studying computing communities, anthropologist Nick Seaver (2017: 9) notes,

> Only through deep engagement and richly contextual description could the ethnographer distinguish such variety—or, in other words, be in on the joke. Superficial accounts risk taking ironic statements literally or missing the conflicted experience of programmers negotiating between different sets of values.

Drawing on the context of conversations I had heard up to that point, I came to see this satirical exchange as a marker of an emerging collective cynicism, at least among some in this community, around both the proliferation of data standards and the precision of data standards. It marked a recognition of the complexity of 'sameness' and 'identity,' along with a concern over the futility of attempting to nail the concepts down, particularly in a space like the World Wide Web.

Documenting this interaction via thick description helps us see that these shifting beliefs and values inform data infrastructure design work; they become interlaced in the infrastructures we engage as data scientists—and they matter. When digital systems rely on these codified semantics to determine how to generate search results, recommend related content, or make automated decisions, designers' negotiated beliefs, convictions, and hesitations shape how our knowledge systems portray the world to us.

TEACHING THICK DESCRIPTION

Thick description takes center stage in a course I teach at Smith College called Data Ethnography. The aim of the course is to help data science students develop an awareness of and ability to evaluate the cultural logics that orient data science work so that they can recognize and intervene when those logics are out of sync with their own ethics. To develop this awareness, we use thick description to excavate the cultural values and meanings that are often rendered invisible in the data science discipline.

While many principal methods of ethnography could guide this course, we focus on thick description for a few reasons. First, most students entering the course have not had an opportunity to reflect on the symbolic cultural provenance of the data resources, tools, and infrastructures they work with. They are often surprised to learn that a dataset documenting the measurements of different iris species—a dataset leveraged very often in data science courses for its usefulness in introducing machine learning concepts—has ties to the eugenics movement (Horst et al. 2022). Reading seminal thick descriptions of data science infrastructures—for example, of classifications like the International Classification of Diseases (Bowker & Star 1999), database models like NoSQL (Dourish 2014), statistical frameworks like homophily (Chun 2021), and datasets like ImageNet (Denton et al. 2021)—they develop an appreciation for how the designs of data infrastructures are guided by certain belief systems, political commitments, dominant discourses, community rituals, and organizational incentive structures. Turning an ethnographic eye to the communities producing these data infrastructures, we are reminded that their configurations are not given but emerge as data scientists identify priorities and negotiate trade-offs.

Consider the debates between the 'structuralists' and the 'minimalists' that oriented the design of Dublin Core; the debates between the 'neats' and the 'scruffies' that guided the design of the Web Ontology Language (OWL) (Poirier 2019); or how, in its efforts to embody a 'middle' ontology (i.e., to avoid becoming an 'ontology of everything' while still remaining useful) (Ronallo 2012), the collaborators working on schema.org had to make some critical judgments regarding what terms should be enumerated within the core schema and how. For example, does a 'public toilet' deserve a place as a civic infrastructure in schema.org, and if so, should the schema offer subtypes for different genders (danbri 2017)? Scholarship in critical data studies and information studies demonstrates how these design debates and negotiations are not arbitrary; they impact how knowledge forms and disseminates, at times resulting in unjust and discriminatory representations of communities (Chun 2021; D'Ignazio & Klein 2020; Eubanks

2018; Noble 2018; Benjamin 2019). While critical components of the provenance of these infrastructures, references to these debates rarely appear in their documentation.

This erasure of cultural history leads us to the second reason thick description is prioritized in Data Ethnography engaging thick description marks a methodological corollary to the norms and commitments that students have grown accustomed to in data science. The first time I assign students to thickly describe a data environment (such as a classroom of biology students collecting field data or a data science hackathon), I will hear questions like the following: How do I make sure that my own biases don't influence the way I interpret what I'm observing? How do I make sure my presence doesn't influence the way that people behave (and thus the ethnographic data that I collect)? How do I make sure I get a representative sample of observations?

I remind students that ethnographers tend to approach these questions differently than data scientists, assuming that biases will always influence the way we interpret ethnographic data; our presence will always influence the data we collect, and there are no thresholds at which culture becomes 'representative.' In the ethnographic communities that I work within, these personal biases and influences are part of the cultural phenomena that we aim to analyze and document. In providing a methodological corollary, engaging data ethnography encourages students to recognize and interrogate the norms of data science communities while also fostering the reflexive sensibilities that have historically been underrepresented in traditional STEM disciplines. The course asks them to discern their own cultural positioning while analyzing that of others. It is an effort to subvert cultural deletions.

CONCLUSION

As designers of data policy and infrastructure, data scientists play an integral role in shaping what forms of knowledge production are made possible, who can participate in knowledge production, and how cultural meaning is made from collected data. With this in mind, data science communities have a responsibility to attend not only to the cultures that orient the work of domain communities but also to the cultures that orient their own work.

Recent scholarship has pointed to pathways forward: to encourage reflection on the assumptions and motivations that underlie the creation, distribution, or maintenance of datasets, Gebru et al. (2020) recommend that all dataset producers document their practices in 'datasheets.' Further research has shown that the practice of producing these documents has prompted data scientists to recognize and deepen their understanding of ethical issues that emerge in relation to machine learning models (Boyd 2021). There are opportunities to extend and implement these reflexive protocols beyond dataset creation. Organizations responsible for the design and dissemination of data science infrastructure and standards (such as the Research Data Alliance and Committee on Data of the International Science Council) can encourage similar documentative practices whenever a new recommendation or deliverable gets published, prompting designers to not only report on the scope, impact, and use cases for outputs but also on the motivations, assumptions, and debates that guided their design. These organizations can also help network more computationally focused groups to anthropologists, sociologists, and science and technology studies scholars with expertise in detailing this type of cultural provenance. Funders could mandate datasheets and other forms of reflexive documentation in annual project reporting, and the *Data Science Journal* can encourage submissions that position new data science policies, applications, and infrastructures in their historical and cultural context.

Finally, to ensure that the next generation of data scientists is ready to engage in this form of reflection and documentation, it will be important for data science college and university programs to foster skills in critical pedagogical traditions that typically get excluded from STEM (such as ethnography, hermeneutics, and critical analysis). More generally, addressing data science's 'culture problem' will demand widespread recognition that we can never design data infrastructure from a cultureless place.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Lindsay Poirier orcid.org/0000-0001-9307-5834
Smith College, US

REFERENCES

Barker, M, Wilkinson, R and **Treloar, A.** 2019. The Australian Research Data Commons. *Data Science Journal*, 18(1): 44. DOI: https://doi.org/10.5334/dsj-2019-044

Benjamin, R. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity Press. DOI: https://doi.org/10.1093/sf/soz162

Bezuidenhout, L, Drummond-Curtis, S, Walker, B, et al. 2021. A school *and* a network: CODATA-RDA Data Science Summer Schools Alumni Survey. *Data Science Journal*, 20(1): 10. DOI: https://doi.org/10.5334/dsj-2021-010

Borgman, CL. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6): 1059–1078. DOI: https://doi.org/10.1002/asi.22634

Bowker, GC and **Star, SL.** 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/6352.001.0001

Boyd, KL. 2021. Datasheets for datasets help ML engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2), Article 438: 1–27. DOI: https://doi.org/10.1145/3479582

Capadisli, S. 2016. Deprecating owl:sameAs. Available at: https://lists.w3.org/Archives/Public/semantic-web/2016Apr/0002.html (Last accessed 6 February 2023).

Carroll, SR, Garba, I, Figueroa-Rodríguez, OL, et al. 2020. The CARE principles for Indigenous data governance. *Data Science Journal*, 19(1): 43. DOI: https://doi.org/10.5334/dsj-2020-043

Chun, WHK. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/14050.001.0001

danbri. 2017. Add a publictoilet type #1624. schemaorg/schemaorg. Available at: https://github.com/schemaorg/schemaorg/issues/1624 (Last accessed 9 January 2023).

Denton, E, Hanna, A, Amironesei, R, et al. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2). DOI: https://doi.org/10.1177/20539517211035955

D'Ignazio, C and **Klein, LF.** 2020. *Data Feminism*. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/11805.001.0001

Dominik, M, Nzweundji, JG, Ahmed, N, et al. 2022. Open Science—For whom? *Data Science Journal*, 21(1): 1. DOI: https://doi.org/10.5334/dsj-2022-001

Dourish, P. 2014. No SQL: The shifting materialities of database technology. *Computational Culture*, 4. Available at: http://computationalculture.net/article/no-sql-the-shifting-materialities-of-database-technology (Last accessed 25 May 2016).

Edwards, P, Mayernik, MS, Batcheller, A, et al. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5): 667–690. DOI: https://doi.org/10.1177/0306312711413314

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Forsythe, DE. 1993. Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science*, 23(3): 445–477. DOI: https://doi.org/10.1177/0306312793023003002

Gebru, T, Morgenstern, J, Vecchione, B, et al. 2020. Datasheets for datasets. arXiv:1803.09010 [cs]. Cornell University. Available at: http://arxiv.org/abs/1803.09010 (Last accessed 24 January 2021).

Geertz, C. 1973. Thick description: Towards an interpretive theory of culture. In: Geertz, C (ed.), *The Interpretation of Cultures*. New York: Basic Books, pp. 3–36.

Gownaris, NJ, Vermeir, K, Bittner, M-I, et al. 2022. Barriers to full participation in the open science life cycle among early career researchers. *Data Science Journal*, 21(1): 2. DOI: https://doi.org/10.5334/dsj-2022-002

Halpin, H, Hayes, PJ, McCusker, JP, et al. 2010. When owl:sameAs isn't the same: An analysis of identity in linked data. In: Patel-Schneider, PF, Pan, Y, Hitzler, P, et al. (eds.), *The Semantic Web—ISWC 2010*. Lecture Notes in Computer Science 6496. Heidelberg, Germany: Springer Berlin Heidelberg, pp. 305–320. DOI: https://doi.org/10.1007/978-3-642-17746-0_20

Horst, AM, Hill, AP and **Gorman, KB.** 2022. Palmer archipelago penguins data in the palmerpenguins R package-an alternative to Anderson's irises. *R Journal*, 14(1): 244–254. DOI: https://doi.org/10.32614/RJ-2022-020

Klump, J. 2017. Data as social capital and the gift culture in research. *Data Science Journal*, 16: 14. DOI: https://doi.org/10.5334/dsj-2017-014

Kouper, I, Scheidt, LA and **Plale, BA.** 2021. Fostering interdisciplinary data cultures through early career development: The RDA/US Data Share Fellowship. *Data Science Journal*, 20(1): 2. DOI: https://doi.org/10.5334/dsj-2021-002

Laine, H. 2017. Afraid of scooping—Case study on researcher strategies against fear of scooping in the context of open science. *Data Science Journal*, 16: 29. DOI: https://doi.org/10.5334/dsj-2017-029

Noble, SU. 2018. *Algorithms of Oppression*. New York: NYU Press. Available at: https://nyupress.org/9781479837243/algorithms-of-oppression (Last accessed 2 May 2019).

Poirier, L. 2019. Classification as catachresis: Double binds of representing difference with semiotic infrastructure. *Canadian Journal of Communication*, 44(3). DOI: https://doi.org/10.22230/cjc.2019v44n3a3455

Ribes, D, Hoffman, AS, Slota, SC, et al. 2019. The logic of domains. *Social Studies of Science*, 49(3): 281–309. DOI: https://doi.org/10.1177/0306312719849709

Ronallo, J. 2012. HTML5 Microdata and Schema.org. *The Code4Lib Journal*, 16. Available at: http://journal.code4lib.org/articles/6400?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+c4lj+(The+Code4Lib+Journal) (Last accessed 15 October 2014).

Seaver, N. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2). DOI: https://doi.org/10.1177/2053951717738104

Star, SL. 1991. The sociology of the invisible: The primacy of work in the writings of Anselm Strauss. In: Strauss, AL and Maines, DR (eds.), *Social Organization and Social Process: Essays in Honor of Anselm Strauss*. Piscataway, NJ: Transaction Publishers, pp. 265–283.

van Panhuis, WG, Paul, P, Emerson, C, et al. 2014. A systematic review of barriers to data sharing in public health. *BMC Public Health*, 14(1): 1144. DOI: https://doi.org/10.1186/1471-2458-14-1144

Wong, M, Levett, K, Lee, A, et al. 2022. Development and governance of FAIR thresholds for a data federation. *Data Science Journal*, 21(1): 13. DOI: https://doi.org/10.5334/dsj-2022-013

TO CITE THIS ARTICLE:
Poirier, L. 2023. Attending to the Cultures of Data Science Work. *Data Science Journal*, 22: 6, pp. 1–7. DOI: https://doi.org/10.5334/dsj-2023-006

Submitted: 10 January 2023
Accepted: 06 February 2023
Published: 03 April 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.