CA Journal of Cultural Analytics

Data Set Failures and Intersectional Data

Nikki Stevens

06.12.19

Peer-Reviewed By: Anon.

Clusters: Data

Article DOI: 10.22148/16.041

Journal ISSN: 2371-4549

Cite: Nikki Stevens, "Data Set Failures and Intersectional Data," Journal of Cultural Analytics. June 12 , 2019.

In 2016, a software developer named David[12] and I met to discuss creating a quantitative demographic survey of the open source software community to which we were both long-time contributors. David and I did not know each other well, but shared a belief that our open source community (OSC, hereafter) was an unsafe place for anyone who did not identify as white, cisgendered, heterosexual and male. That lack of safety was further complicated by any one individual's distance from privileged modes of contribution. In OSC, developers who were on key Contribution Teams and regularly added code to the codebase were valued more highly than those who contributed documentation, user experience research, or quality assurance work. David asked "What if we made a survey that accounted for all of the ways that people *make* the web? Can we do it intersectionally? Can we do it the OSC way?" Over the following 18 months, David and I worked with a team of OSC community members on the creation, dissemination and analysis of a quantitative demographic survey of OSC. This survey was the first step in a project to create safer spaces within OSC for individuals from marginalized groups. Rather than this survey being an attempt to gather a representative sampling, it was part of a larger political project to increase diversity, inclusion and equity in our community.[1]

[1] We fell prey to what Theodore Porter calls the "insistent quantifrenia" of everything diversity

To further contextualize our motivations, in 2015 Stack Overflow, a popular website for software engineers, released the results of its annual developer survey. They surveyed "over fifty thousand developers" and claimed that the report was "the most comprehensive developer survey ever conducted."[2] Despite contributing their information to the dataset, individuals who took the survey had no access to the resultant data and no input into the questions asked.[3] As engineers steeped in open source ways of working, David and I found Stack Overflow's extractivist approach to gathering data to be both foreign and morally suspect. We viewed our open source community not only as a software production system, but a knowledge production network and believed that knowledge produced *by* the community should belong *to* the community. This belief guided every choice that we made during the survey project.

While our format was quantitative, our approach was unruly and counter to many of the expectations of a detached and objective quantitative survey of a community. We[4] deprioritized data cleanliness, invited individuals to self identify, and were explicitly biased in both our motivations and stewardship of the process - for us, this was a site of intersectional praxis. Intersectionality was about power and the ways that access to power was mediated by individual social location(s). While traditional surveys may be collected in order to obtain an "objective" sampling of community members, this project rejected any ideal of objectivity in favor of a situated and subjective project.

---

and inclusion related. Theodore M. Porter, *Trust in Numbers* (Princeton, N.J: Princeton University Press, 1996), 76. This quantifrenia seems like a way to mediate the unreliable narration of experience performed by those from marginalized groups. My thinking here is influenced by Fricker's construction of epistemic credibility. Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, 1 edition (Oxford: Oxford University Press, 2009). Examining quantification's role in diversity and inclusion work is out of the scope of this paper. The CHAOSS (Community Health Analytics Open Source Software) Project is a prominent group thinking about how to measure every aspect of community "health," including diversity and inclusion. See https://chaoss.community/. For examples of work quantifying the success of diversity projects, see Peggy D Dreachslin PhD; Lee, "Applying Six Sigma and DMAIC to Diversity Initiatives," *Journal of Healthcare Management* 52.6 (2007). For one of the few academic works examining the effects of diversity in open source communities, see S Daniel, Agarwal, and Stewart, "The Effects of Diversity in Global, Distributed Collectives: A Study of Open Source Project Success," *Information Systems Research* 24, no. 2 (2013): 312-33.

[2] Stack Overflow Developer Survey 2016 Results," Stack Overflow, January 15, 2016, https://insights.stackoverflow.com/survey/2016

[3] In July 2016, a few months after the results of the survey were released, Stack Overflow released its "Insight Center" with data sets for all previous years' surveys. When our survey work began in May 2016, there was no way to access the results of the Stack Overflow developer surveys.

[4] Though I am describing my own experience, outside of the clear boundaries of an academic study, I did not act alone, but in consultation and collaboration with others in the community. While I do not speak for them, I also do not wish to represent this work as only my own. I default to the use of the word 'we' when discussing conversations/group actions.

This paper begins by reviewing important aspects of open source communities and defining our use of intersectionality. Next, I outline our approach to employing open source production methods to a quantitative demographic survey. I then explore several failures in the survey's lifecycle and offer corrective suggestions where appropriate. Finally, I discuss the concepts underlying our approach and ask "Does a research project designed to be intersectional produce data that itself is intersectional?"

Literature Review

Open source

Open source software communities are those in which individuals work together to produce a technical product. The open source movement[5] is characterized by a few behaviors: working publicly and collaboratively, the flawed notion of meritocracy,[6] and (in its early days) a rejection of capitalist and corporate-controlled modes of production. These behaviors were fundamental parts of OSC.

Open source software (OSS) communities espouse normative modes of knowledge production and participation, subject to many of the same disequilibriums as the societies that contain those communities. I will briefly review two norms relevant here. First, there is an expectation that knowledge made by individuals is accessible to those individuals. While the "open" in "open source" has been described as not open at all,[7] the concept of openness was hugely influential for us. Our adoption of the ethos of open source meant that we began the project with fixed ideas about the way that knowledge should be produced. As evidenced from our initial reaction to the closed-source Stack Overflow survey, we believed data

[5] A review of the history and nuances of open source communities is out of scope for this article. See, among others: Dawn Nafus, " 'Patches Don't Have Gender': What Is Not Open in Open Source Software," *New Media & Society* 14, no. 4 (2012): 669-83; Dany Di Tullio and D. Sandy Staples, "The Governance and Control of Open Source Software Projects," *Journal of Management Information Systems* 30, no. 3 (January 2013): 49-80; Hao-Yun Huang, Qize Le, and Jitesh H. Panchal, "Analysis of the Structure and Evolution of an Open-Source Community," *Journal of Computing and Information Science in Engineering* 11, no. 3 (2011): 031008; Audris Mockus, Roy T Fielding, and James D Herbsleb, "Two Case Studies of Open Source Software Development: Apache and Mozilla," *ACM Transactions on Software Engineering and Methodology (TOSEM)* 11, no. 3 (2002): 309-46; David L. Olson and Kirsten Rosacker, "Crowdsourcing and Open Source Software Participation," *Service Business* 7, no. 4 (2013): 499-511; Eric S. Raymond, *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, 1 edition (Beijing ; Cambridge, Mass: O'Reilly Media, 2001); E. Gabriella Coleman, *Coding Freedom : The Ethics and Aesthetics of Hacking* (Princeton University Press, 2012).

[6] Nafus, " 'Patches Don't Have Gender.'"

[7] Nafus, " 'Patches Don't Have Gender.'"

should be collected and distributed by the people who produced it. Rather than engage in debates about how to get "the public" productively involved in technological knowledge production,[8] we simply invited everyone to participate. Of course, this did not mean that everyone was able to participate and be longitudinally involved. Barriers to all levels of engagement, specifically for minorities, women and LGBTQIA2S+ individuals, can be significant.[9]

Second, OSS communities are structured to expect that contribution and individual value are quantifiable. This expectation results in the erasure of work that is not easily measurable and devalues the individuals producing that work. When software is being written, code writers add code to the codebase. Those additions can be specifically measured in lines of code, number of characters changed, or file size differences. Thus, as a project is developed, contributors can be ranked by the volume (and implied value) of their contribution.[10] This is problematic for several reasons. First, the traditional structure of attributing lines of code to a single individual reinforces the idea that lines of code are the sole product of

[8] Frank Fischer, "Technological Deliberation in a Democratic Society: The Case for Participatory Inquiry," *Science and Public Policy* 26, no. 5 (1999): 294-302; James Wilsdon and Rebecca Willis, *See-Through Science: Why Public Engagement Needs to Move Upstream* (London: Demos, 2004); Andrew D. Zimmerman, "Toward a More Democratic Ethic of Technological Governance," *Science, Technology, & Human Values* 20, no. 1 (1995): 86-107.

[9] For information on LGBTQIA2S+ experiences in STEM: Keith J. Bowman and Lynnette D. Madsen, "Queer Identities in Materials Science and Engineering," *MRS Bulletin* 43, no. 4 (April 2018): 303-7; Erin A. Cech and Michelle V. Pham, "Queer in STEM Organizations: Workplace Disadvantages for LGBT Employees in STEM Related Federal Agencies," *Social Sciences* 6, no. 1 (February 4, 2017): 12; Erin A. Cech and Tom J. Waidzunas, "Navigating the Heteronormativity of Engineering: The Experiences of Lesbian, Gay, and Bisexual Students," *Engineering Studies* 3, no. 1 (April 1, 2011): 1-24. Specific research on women in open source participation: M. Mahmod, S. A. M. Yusof, and Z. M. Dahalin, "Women Contributions to Open Source Software Innovation: A Social Constructivist Perspective," in *2010 International Symposium on Information Technology*, vol. 3, 2010, 1433-8; Nafus, " 'Patches Don't Have Gender".. These barriers can also be attributed to communication styles E. Moon, "Gendered Patterns of Politeness in Free/Libre Open Source Software Development," in *2013 46th Hawaii International Conference on System Sciences*, 2013, 3168-77. For academic work about the demographics and politics of open source participants, see Yuanfeng Cai and Dan Zhu, "Reputation in an Open Source Software Community: Antecedents and Impacts," *Decision Support Systems* 91 (2016): 103-12; Tadeusz Chełkowski, Peter Gloor, and Dariusz Jemielniak, "Inequalities in Open Source Software Development: Analysis of Contributor's Commits in Apache Software Foundation Projects," ed. Christophe Antoniewski, *PLOS ONE* 11, no. 4 (April 20, 2016): e0152976; Daniel, Agarwal, and Stewart, "The Effects of Diversity in Global, Distributed Collectives"; Rajdeep Grewal, Gary L. Lilien, and Girish Mallapragada, "Location, Location, Location: How Network Embeddedness Affects Project Success in Open Source Systems," *Management Science* 52, no. 7 (2006): 1043-56; Mahmod, Yusof, and Dahalin, "Women Contributions to Open Source Software Innovation"; Nafus, " 'Patches Don't Have Gender".

[10] As an example, Docker CLI is a popular open source software project with five years of public code contribution history. At the time of this writing, 607 contributors are listed and ranked by their volume of contribution. See https://github.com/docker/cli/graphs/ contributors for a list of these individuals.

their author, and not the results of extended collaboration with others. Second, ranking contributors by their lines of code erases the other important ways that people contribute to a software product and elides the relationality inherent in creating a project with others. This elision does violence to all of the contributors whose labor is not so easily quantifiable. In order to be successful, software projects need documentation, support forums, outreach individuals, community managers. When a project is large enough to have a physical presence, the project needs people to attend local meetups, to staff booths at conferences, to hold events and coordinate speakers. All of this work directly contributes to the success of the project, but again is not so easily quantifiable.

Intersectionality

At the start of this project, David and I specifically invoked the word intersectionality. Intersectionality can be a synonym for "complexity," and it is often perceived as an essentialist driver of identity politics; however, these are misunderstandings. Rooted in a history of black feminist scholarship, intersectionality is an analytical framework that ultimately is about exposing power structures and systemic inequalities. While Alexander-Floyd[11] argues that for work to be intersectional, it must explicitly center and study black and other women of color, Cooper[12] disagrees. She asserts that we can move intersectionality past the "paradigmatic black woman" as an object of study, and into an era in which black feminist scholarship is powerful enough to act as a foundation upon which to build knowledge about marginalized groups of any identity. It is following Cooper's elucidation of intersectionality that I use the term here and that we used the term as we were developing the survey.

As previously mentioned, the survey team understood intersectionality to be about power structures and social position. We saw the way that individuals who moved further from the "ideal" were given less credibility in OSC and for us, social position included not just one's race, class, gender, sexual orientation and other aspects of "identity", but one's role in the community. To illustrate the layers of marginalization that community members can experience, an example: a few years ago Alex, a genderqueer person of color known in part for their vocal community work, accused Tom, a prominent white male coder, of misogyny and sexually harassing language. The community largely ignored Alex's complaint (despite others coming forward to support and offer similar experiences with

[11] "Disappearing Acts: Reclaiming Intersectionality in the Social Sciences in a Post-Black Feminist Era," *Feminist Formations* 24, no. 1 (2012): 1-25.

[12] "Intersectionality," *The Oxford Handbook of Feminist Theory*, February 1, 2016.

Tom) and continued to encourage Tom's participation. Later, Maggie, a white, cis-gendered woman known for her software engineering, accused Tom of sexual harassment. Maggie's complaint was taken seriously. Individuals in the community gave Maggie's complaint more weight not only because of her race and gender, but because she was seen as a valuable code contributor. In evaluating the validity of the claims against Tom, Maggie's history of quantifiable contribution was actively considered. Without an intersectional framework, we were unable to account for all of the ways the community had done a disservice to Alex and their experience.

Intersectional research often deals with complexity, but it also centers those from marginalized groups and addresses social inequality, relationality and social context.[13] Given the reproduction of external oppressions in open source communities, and the violent force of meritocratic ideology,[14] open source communities are sites ripe for intersectional examination. I am dealing only with a small case study, but I further Safiya Noble's assertion that we need intersectionality to examine the racialized and gendered ways that knowledge is produced, not just on the internet, but also within open source communities.[15]

Methodology

David and I began discussing the survey project in May 2016, worked with the community on the design, and opened the survey for data collection in October 2016. During the six months of design work, there were 108 commits to the Github repository made by 8 individuals (here again we see the easy quantification of text commits). Others on the survey team did not commit direct changes but participated in discussions. Those discussions happened both in the repository's issue queue and in specific meetings about questions and goals. An additional 25 people were part of those discussions. David and I agreed that we would involve as many community members as possible and actively seek to involve individuals from different cultural, socioeconomic, and geographic backgrounds. We did not log the race, class or gender of members of the survey team.

I am deeply embedded in this story and this process—removing myself from it to exist as the academic narrator is an impossibility. I was a participant observer

[13] Patricia Hill Collins and Sirma Bilge, *Intersectionality*, 1 edition (Cambridge, UK ; Malden, MA: Polity, 2016).

[14] Nafus, " 'Patches Don't Have Gender.'"

[15] Safiya Umoja Noble, "A Future for Intersectional Black Feminist Technology Studies," *Scholar & Feminist Online* 13, no. 3 (2016): 1-8, http://sfonline.barnard.edu/traversing-technologies/safiya-umoja-noble-a-future-for-intersectional-black-feminist-technology-studies/.

present for every stage of the survey creation and data lifecycle, attended every meeting and approved the majority of the community contributions. Observations about the survey shared below come from my recollections of conversations in meetings and reviewing digital artifacts like Github issue threads.

The Survey Process and the Data lifecycle

The data lifecycle describes a set of phases that data (and those who are responsible for data) can undergo. The lifecycle is cyclical, iterative and, as this project was arranged and managed by engineers, made intuitive and epistemological sense to us. As a result, we approached the dataset resulting from this survey using a similar set of stages. One common lifecycle has eight phases.[16] The eight steps—Plan, Collect, Assure, Describe, Preserve, Discover, Integrate, Analyze—were all present in the creation of this data set. In the following sections, I discuss five major failures during the Planning and Collection stages and the state of the data at the end of the Assuring phase.

Planning

In the planning stage, data stewards create a "description of the data that will be compiled, and how the data will be managed and made accessible throughout its lifetime."[17] In this stage, we designed the survey questions.

Failure #1: By choosing a code-centric platform, despite the goal of validating and encouraging non-code participation, we erected barriers to participation in the planning stage.

In order to facilitate collaboration on the survey, we rendered the questions in plain text, using () to represent radio buttons and [] to represent checkboxes, and we worked in plain text on Github.[18] The final text version of the survey (before it was recreated in the online survey platform we used) is visible in Appendix A. We followed a standard Github pull request workflow[19] and this, despite its

[16] DataOne, "Data Life Cycle | DataONE," December 14, 2015, https://www.dataone.org/data-life-cycle/.

[17] DataOne.

[18] Github.com is a site for people (largely people who write code) to share and collaborate on their projects.

[19] For an explanation of a pull request workflow see Atlassian, "Pull Requests | Atlassian Git Tutorial," Atlassian, March 7, 2014, https://www.atlassian.com/git/tutorials/making-a-pull-request; Github, "Understanding the GitHub Flow ·GitHub Guides," Github, November 30, 2017, https:

publicness, presented an unexpected barrier to entry for individual participation. Our backgrounds as developers and our long experience with collaborative tools like Github blinded us to the fact that for many people working in technology, Github is not a necessary tool. I met many folks who wanted to participate but could not (or did not want to) navigate the Github onboarding process. Despite our explicit conversations about centering marginalized individuals in the survey design, we centered developers and developer-centric tools and did not critically engage with our initial platform selection. For many of our collaborators using a shared document system—like Google Docs—would have been easier.

The stabilization of identity

Examining the content of *every* question is out of scope for this paper, but I will review two questions as examples of ways that I believe our intersectional goals were in friction with the requirements of quantification.

Failure #2: By asking users to stabilize aspects of their identity, our questions acted as barriers to intersectional representations of an individual.

Because technologists are so often portrayed as cisgendered, heterosexual men, it was important that we include expansive questions about gender and sexual orientation. We explicitly hoped that by framing the questions in a more inclusive way, we would capture a wider variety of gender identities and sexual orientations. The survey team spent a lot of time discussing and debating the "right" choices for these sections.

The sexual orientation question:

Do you identify with any of the following?
[] Asexual
[] Bisexual
[] Gay
[] Lesbian
[] Queer
[] Straight
[] Self Identify: _____
[] Prefer not to answer

The gender identity questions:

//guides.github.com/introduction/flow/.

Do you consider yourself to be transgender/gender non-conforming?
() Yes
() No

What is your primary gender identity today?
() Female
() Genderqueer
() Intersex
() Male
() Trans F-M
() Trans M-F
() Prefer not to answer
[_____] Self-identify

Some dimensions of identity, like sexual orientation and gender, resist stabilization both in their definition and an individual's choice of labels over time.[20] For gender identity, we attempted to make allowances for this by including the word "today" in the question. However, the addition of the word "primary" required respondents to engage with an artificial hierarchy of gender identities. Internal conflict about not having too many questions meant that we removed a separate question asking if the respondent is intersex, and collapsed the trans and intersex identification into the gender identity question. As a result, we excluded trans-nonbinary and other trans* identities. For sexual orientation, we included checkboxes, but still demanded that the respondent explore their definition of the word queer in order to answer the question. Based on feedback from community members, we should have made all questions in this section follow the format of "Check all that apply" + "Self identify."

The normative force of questions

Failure #3: In our focus to perform "open source data collection," we neglected to think critically about the impact of including workplace, education and socioeconomic questions.

Despite the open source rejection of corporate (Stack Overflow, in our case) control of community-sourced data, we were still influenced by corporate interests. We included questions about a respondent's workplace, size of workplace, and

[20] Petra Doan, "To Count or Not to Count: Queering Measurement and the Transgender Community," *Women's Studies Quarterly* 44, no. 3/4 (October 2016): 89-110, http://search.proquest.com/docview/1831356971/.

main business function, invoking capitalist expectations that an individual's work should serve a profit-making enterprise. In contradiction to a seemingly egalitarian ethos about how "everyone" can participate in our open source community, the questions about a respondent's workplace have normative implications for the validity of work produced in and outside of businesses. In the education section, we asked about an individual's level of educational achievement, and we omitted technology bootcamps—a large source of training for non-traditional and later-in-life technologists.[21] Questions about socioeconomic status in childhood compared with adulthood took for granted that people working in technology experience class mobility. As one respondent wrote in the comments box at the end of the survey, "assuming that everyone works for a company or business leaves out those of us working in higher (or other) education, government, or non-profit organizations. It's a common (and annoying) mis-labeling." As an alternative, we could have allowed respondents to share where and how they make technology in a check-all-that-apply style question.

Collecting

During the collection phase of the data lifecycle, "observations are made either by hand or with sensors or other instruments and the data are placed into digital form."[22] During this phase, we created a web-based version of the survey and collected responses.

Failure #4: We allowed the limitations of media to determine the designs of our questions.

We opened the survey for data collection in October 2106 and converted our plain text survey to the format of our chosen survey vendor. In the original question design, we agreed to allow users to check as many boxes as needed, in addition to entering text in a self-identify field. The survey vendor did not support this, and we were bound by the limits of the vendor's plugins. The team wanted to choose a vendor that would not store copies of survey responses after we closed the survey and that would provide a complete data export. In our focus to ensure that we would be able to easily share the data, we made user interface compromises.

[21] G. A. Wilson, "Could a Coding Bootcamp Experience Prepare You for Industry?" *IT Professional* 20, no. 2 (March 2018): 83-87; Sherry Seibel, "Social Motivators and Inhibitors for Women Entering Software Engineering Through Coding Bootcamps Vs. Computer Science Bachelor's Degrees: (Abstract Only)," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18 (New York, NY, USA: ACM, 2018), 274-74.

[22] DataOne, "Data Life Cycle | DataONE."

We could have addressed this limitation in at least two ways. First, we could have chosen a survey vendor before drafting questions, and subsequently designed questions within that vendor's technological limitations. Second, given that many of us were programmers, we could have modified an existing open source survey tool to accommodate the team's vision for question interfaces. This is another example of the team centering expectations that technology most often acts as a tool and not a barrier.

Failure #5: We ignored the contradictions in involving corporations in a survey designed to reject corporate control of data.

Like the corporate interests that manifested in planning the questions, corporate interests intervened in collection as well. After running the survey for a few months, we still had fewer than 200 respondents. David secured $20,000 from his employer to spend on advertising to recruit survey participants. One of the places we advertised was on Stack Overflow, who matched that $20,000 with another $20,000 of in-kind advertising. Like David's employer, Stack Overflow was interested in gathering more data about people. The donations were given with the understanding that those companies would have access to the data as soon as it was released. I was not present for the negotiations regarding the initial money or the matching money, but the team collectively agreed that involving corporations to get more respondents was a good choice. Ultimately, we were complicit in the gathering of data for corporate interests, despite rationalizing that our involvement with corporate funding was in service of our larger project to benefit the community.

Assuring

In the assuring phase, "the quality of the data are assured through checks and inspections."[23] Implied in data assurance are quality checks and transformations to ensure that data is recorded in a particular way and that it meets certain standards. Data that does not meet standards is often described as "dirty," a counterpoint to "clean" data that is ready to progress to the next phase of the lifecycle. Numerous works[24] assert the importance of having a system for cleaning data,

[23] DataOne.

[24] Jason W Osborne, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data* (Sage Publications, 2012); Erhard Rahm and Hong Hai Do, "Data Cleaning: Problems and Current Approaches," *IEEE Data Eng. Bull.* 23, no. 4 (2000): 3-13; Vijayshankar Raman and Joseph M Hellerstein, "Potter's Wheel: An Interactive Data Cleaning System," in *Proceedings of the 27th VLBD Conference* (Rome, Italy, 2001), 10.

and make a connection between "clean data" and "quality data."[25] Those of us involved—nearly all software engineers because of our earlier choice to work on Github—had extensive training on collecting "clean" data and some anticipatory concerns about whether or not the data will be "useful" if it is not clean.

However, if we consider that data is abstract and conceptual,[26] then as a concept it cannot be "clean" or "dirty". If we consider data as a material object, or collection of objects, it similarly cannot be inherently clean or dirty, but simply in violation of an external standard. I am arguing that a focus on data's cleanliness is a way of controlling which knowledge is "valid" and is directly counter to intersectional aims. This informed our choice not to "clean" the data we obtained. Intersectional data is messy data.[27] We embraced the messiness and all aspects of the data that might resist traditional quantitative analysis. We elected to make no changes to the survey responses or do any preparation before opening the data set. In this way, I believe that our dataset stands as a counterpoint to the shaped, "cleaned," "interpreted" data produced by surveys like those done by Gitlab and Stack Overflow.[28]

Discussion

Failure #6: We engaged in an ethically complicated process without considering the implications.

From its inception, we positioned the survey as counter to the Stack Overflow surveys. Unlike Stack Overflow's closed and restricted approach, we would be doing all of our work in the open. Our intentions were to document the (all-axes) demographics of the community as a starting point for intervention; however none of us had considered the implications of undertaking such a survey. The responsibility of designing and asking demographic questions, the power dynamics inherent in positioning the survey as by and for open source, the epistemological

[25] N. Peng et al., "Finding Interesting Cleaning Rules from Dirty Data," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, 2017, 378-82; Chen-Bo Zhong and Katie Liljenquist, "Washing Away Your Sins: Threatened Morality and Physical Cleansing," *Science* 313, no. 5792 (September 8, 2006): 1451-2.

[26] Lisa Gitelman, *Raw Data Is an Oxymoron* (MIT Press, 2013).

[27] Jacqueline Wernimont, "Notes Toward a Post on Intersectional Data - Jacqueline Wernimont," December 7, 2015, https://jwernimont.com/notes-toward-a-post-on-intersectional-data/.

[28] GitLab, "GitLab 2018 Global Developer Report," GitLab, March 7, 2018, https://about.gitlab.com/developer-survey/2018/; "Stack Overflow Developer Survey 2017," Stack Overflow, March 22, 2017, https://stackoverflow.com/insights/survey/2017/?utm_source=so-owned&utm_medium=social&utm_campaign=dev-survey-2017&utm_content=social-share; "Stack Overflow Developer Survey 2016 Results."

challenges of representing individuals in numbers—all lost on us at the time. I believe that this was our biggest and most foundational failure.

In OSC, it was emphasized that there was no reason to ever wait for an invitation to do things and that leadership and opportunity were rarely given, but taken by those with the time and inclination to do the work. While we may have realized that we were in a uniquely privileged position to be able to start the survey project, we did not have a formal conversation about the nuances of collecting data about people and the responsibilities we might have for the data at every phase of the lifecycle. We did not draw a discrete line between the documenters and the documented and as a result, did not understand the complexities of counting humans or that "the classification of individuals is at the heart of [...] social control."[29] We realized after the fact that our envelopment in an open source space and our invocation of intersectional praxis did not protect us from enacting traditional methods of control.

The final question in the survey asked users to add any other aspects of their identity they felt were important. One hundred and seventy-nine respondents entered text, including: Quaker, polyamorous, Russian-speaking Asian, autistic, former fundamentalist, feminist, abuse survivor, liberal, immigrant, depressed. These answers reinforced for us the importance of self-identify boxes and the ways that the identity markers we provided were insufficient. I believe that we absolutely did not offer folks a configuration of questions and answers that would have empowered them to represent themselves.

Conclusions and Implications for Further Research

Before data can be planned, before it exists as a data point, it exists as an imaginary.[30] Our data imaginary contained amorphous representations of individuals and their many identities. Scholars have expanded the criteria by which we can determine if a research method (qualitative or quantitative) is intersectional[31] but make no specific mention of whether or not the artifacts resulting from research can be intersectional themselves. Can data, as a concept and/or as a material object, be intersectional? Regarding the methodological challenges of intersectional research, Lisa Bowleg writes "I question whether the positivistic assumptions implicit in quantification are compatible with intersectionality

[29] Peter J Aspinall, "Answer Formats in British Census and Survey Ethnicity Questions: Does Open Response Better Capture 'Superdiversity'?" *Sociology* 46, no. 2 (April 2012): 354-64.

[30] Gitelman, *Raw Data Is an Oxymoron*.

[31] Nicole M Else-Quest and Janet Shibley Hyde, "Intersectionality in Quantitative Psychological Research," *Psychology of Women Quarterly* 40, no. 3 (September 2016): 319-36.

research.."[32] In the survey, we were limited by the media: the limitations of the form's user interfaces constrained the ways that users were able to represent themselves. Could (or should) we represent the matrix of intersectionality[33] in data entry? Guided by existing user interfaces and data structures, we asked users to compartmentalize their identities for easy entry into forms and storage in static data structures. Can we create and design data structures to mirror the fluidity and fullness of identity? Can data as an object, or as a collection of data-record objects, mirror the matrix perspective of an identity?

These questions arise as I review my assessment of some of the mistakes we made during this survey process. Had we been able to correct all of our mistakes, how, if at all, would that have changed the intrinsic nature of the data collected? It is my hope that by sharing our shortcomings, others can critically engage with their own human quantification practices and the nature of the resulting data.

Appendix A

This is the text of the survey in Github before we launched it.

Survey

Thanks for taking a few minutes to help make the web a better place! The world wide web is perhaps the greatest egalitarian communications platform, ever, and we believe we have a part to play in towards that end. As a community of professional developers, designers, editors, project managers, open-source contributors, we help create the web and help individuals and organizations communicate their message. Who exactly makes up this community of individuals who are contributing to a more free and open communication medium?

Let us know how you identify yourself, so we can get a better vision of who we are as a whole, and thus how we can leverage our diverse identity.

How do you help build the internet?

Position

[32] Lisa Bowleg, "When Black + Lesbian + Woman != Black Lesbian Woman: The Methodological Challenges of Qualitative and Quantitative Intersectionality Research," *Sex Roles* 59, nos. 5-6 (September 1, 2008): 37.

[33] Vivian M. May, *Pursuing Intersectionality, Unsettling Dominant Imaginaries*, 1 edition (New York, NY: Routledge, 2015).

What professional role(s) do you play?
[] Content Developer/Strategist
[] Designer
[] Developer
[] Product Manager/Owner
[] QA/Testing
[] Technical/Solutions Architect
[] Technical Lead
[] UX/UI
[] Other

Management/Leadership

Which most accurately describes your role?

() I am a formal manager
() I am not a manager, though I do hold a leadership role
() My role does not involve leadership or management

Years of Experience

() 0 - 1
() 2 - 5
() 6 - 10
() 11 - 15
() 16 - 20
() 20+

Company Size

() Under 10 employees
() 10 - 20 employees
() 20 - 50 employees
() 50 - 100 employees
() 100 - 200 employees
() 200+ employees

Location

Please choose your country of residence. |v| Dropdown of Countries |

Which word best describes the area where you live and work?

() Rural
() Suburban
() Urban

Type of Company

What type of company do you work for (check all that apply?)
[] Freelancer
[] Web Agency / Development Shop
[] Digital Agency
[] Advertising Agency
[] Non-agency organization
[] Other

Education Level

What is your education level?
() Some high school
() Some college/technical training
() Completed college
() Completed master's degree
() Completed Ph.D

If you have a degree, did you study a technology-related field?
() Yes
() No

How do you identify?

Turns out, diversity is hard to classify. It's personal, contextual, and depends as much on who you are as who you are around. Here are some identifiers we're going to consider.

- Ability - Mental and/or physical
- Age
- Ethnicity
- Gender
- Race
- Religion
- Sexual Orientation
- Socio-Economic Status/Class
- ? / & / Etc.

Disability

Do you identify as having a disability as defined under the [Americans with Disabilities Act]?[34]
[] Yes, Cognitive
[] Yes, Emotional
[] Yes, Hearing
[] Yes, Mental
[] Yes, Physical
[] Yes, Visual
[] Yes, Other
[] No
[] Prefer not to answer

Does your disability affect how you work?
() Yes
() No
() Prefer not to answer

Age

What is your age range?
() 0-15
() 16-24
() 25-34
() 35-44
() 45-54
() 55-64
() 65-74
() 75-84
() 85+
() Prefer not to answer

Racial Identification

Do you identify as a person of color?
() Yes
() No

With which racial background(s) do you identify? (check all that apply)

[34] https://adata.org/faq/what-definition-disability-under-ada

[] Asian
[] Black
[] Latino
[] Native American
[] Pacific Islander
[] White
[] Other
[] Prefer not to answer

Sexual Orientation

Do you identify with any of the following?
[] Asexual
[] Bisexual
[] Gay
[] Lesbian
[] Queer
[] Straight
[] Self Identify: _____
[] Prefer not to answer

Gender Identification

Do you consider yourself to be transgender/gender non-conforming?
() Yes
() No

What is your primary gender identity today?
() Female
() Genderqueer
() Intersex
() Male
() Trans F-M
() Trans M-F
() Prefer not to answer
[_____] Self-identify

Religious Identification

Do you practice, worship, or observe a particular religion (or agnostic/atheist theology)?

() Yes
() No

Do you identify as a minority because of your religion (or lack thereof)?
() Yes
() No
() N/A

Socio-economic class

Thinking about your childhood, which socio-economic class did you identify with?
() Working class
() Lower middle class
() Upper middle class
() Upper class

Thinking about your current situation, which socio-economic class do you identify with?
() Working class
() Lower middle class
() Upper middle class
() Upper class

Language

Choose the language(s) you speak and work with and identify your proficiency
[_____|v] [_____|v]
(list of languages) (fluency levels)

add as many as appropriate.

Other facets

At work, I feel comfortable expressing the aspects of my identity that are important to me.
() Strongly Agree
() Agree
() Neutral
() Disagree
() Strongly Disagree

At work, it's important to me that I feel comfortable expressing my identity.
() Strongly Agree
() Agree
() Neutral
() Disagree
() Strongly Disagree

Encouraging people to express any/all aspects of their identities benefits my company.
() Strongly Agree
() Agree
() Neutral
() Disagree
() Strongly Disagree

? / & / Etc.

What else do you identify with? Call out anything we've missed that makes you, well, you! [ ]

add as many as appropriate.

Post-survey CTAs

- If you'd be willing to participate in a semi-structured interview about your experiences
[Link to new form not tied to participant's data]

- If you're interested in being notified when the data is released
[Link to new form not tied to participant's data]

- If you're interested in being a part of the diversity working group mailing list
[Link to new form not tied to participant's data]

Survey conventions

() radio buttons
[] check boxes
[_____] fill-in-the-blank

[_____|v] drop-down select one
??? combo box/select many