Accountable Data: The Politics and Pragmatics of Disclosure Datasets

Lindsay Poirier
lpoirier@smith.edu
Smith College
Northampton, MA, USA

ABSTRACT

This paper attends specifically to what I call "disclosure datasets" - tabular datasets produced in accordance with laws requiring various kinds of disclosure. For the purposes of this paper, the most significant defining feature of disclosure datasets is that they aggregate information produced and reported by the same institutions they are meant to hold accountable. Through a series of case studies of disclosure datasets in the United States, I specifically draw attention to two concerns with disclosure datasets: First, for disclosure datasets, there is often political and social mobilization around the definitions that determine reporting thresholds, which in turn implicates what observations end up in the dataset. Changes in reporting thresholds can be traced along changes in political party power as the aims to promote accountability through mandated disclosure often get pitted against the aims to reduce regulatory burden. Second, for disclosure datasets, the observational unit – what is ultimately being counted in the data – is often not a person, institution, or action but instead a form that the reporting institution is required by law to fill out. Forms infrastructure the information that ends up in the dataset in notable ways. This work contributes to recent calls to promote the transparency and accountability of data science work through improved inquiry into and documentation of the social lineages of source datasets. The analysis of disclosure datasets presented in this paper poses important questions regarding what ultimately gets documented in the data, along with the representativeness and usefulness of these accountability mechanisms.

CCS CONCEPTS

• **Theory of computation** → *Data provenance*; **Incomplete, inconsistent, and uncertain databases**; • **Information systems** → *Data dictionaries.*

KEYWORDS

disclosure, accountability, infrastructure, data provenance

ACM Reference Format:
Lindsay Poirier. 2022. Accountable Data: The Politics and Pragmatics of Disclosure Datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3531146.3533201

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533201

1 INTRODUCTION

The publication of open datasets on the World Wide Web has become a key strategy in efforts to promote accountability in corporate and government business activities. With growth in public concern over discrimination and other forms of corporate and government wrong-doing, calls for regulation that enable the public to bear witness to the decisions and actions of various institutions have been amplified. As a result, a growing number of laws have been passed requiring both government and corporate institutions to disclose information regarding their business activities. Sometimes referred to as "sunshine laws," these regulations aim to prevent corruption, highlight malfeasance, and promote public trust in various institutions by making information accessible for public scrutiny.

While information disclosed by such laws can take many different forms (e.g. meeting transcripts, government records), this paper attends specifically to what I call "disclosure datasets" - tabular datasets produced in accordance with laws requiring various kinds of disclosure. For the purposes of this paper, the most significant defining feature of disclosure datasets is that they aggregate information produced and reported by the same institutions they are meant to hold accountable. Further, the values reported in disclosure datasets can lead to adverse actions - either formal or informal - taken against the reporting institutions. Combined, these issues institutionally incentivize misreporting and creative accounting. Examples of disclosure datasets in the United States include the Toxic Release Inventory (a dataset documenting the amounts of toxic chemicals industrial facilities emit in a given year) and police stop count data (municipal datasets documenting police encounters with citizens in order to identify potential racial profiling in policing). In this paper, I analyze dimensions of the provenance of the underlying infrastructures shaping disclosure datasets. Through a series of case studies, I specifically draw attention to two concerns: First, for disclosure datasets, there is often political and social mobilization around the definitions that determine reporting thresholds, which in turn implicates what observations end up in the dataset. Disclosure data definitions are malleable: changes in reporting thresholds can be traced along changes in political party power as the aims to promote accountability through mandated disclosure often get pitted against the aims to reduce regulatory burden. Second, for disclosure datasets, the observational unit – what is ultimately being counted in the data – is often not a person, institution, or action but instead a form that the reporting institution is required by law to fill out. Forms infrastructure the information that ends up in the dataset in notable ways.

This work contributes to recent calls to promote the transparency and accountability of data science work through improved inquiry into and documentation of the social lineages of source datasets. The analysis of disclosure datasets presented in this paper poses important questions regarding what ultimately gets documented in the data, along with the representativeness and usefulness of these accountability mechanisms. This is of heightened concern as advocacy groups call for expanding disclosure programs, and as disclosure datasets become inputs for statistical models and machine learning algorithms. Despite the widely recognized conflicts of interest interwoven through disclosure datasets, advocacy groups have nonetheless consistently mobilized the information they document to build evidence against reporting institutions: presenting the data to document histories of discrimination, environmental pollution, or corruption, while acknowledging potential sources of bias. Disclosure datasets have historically been and continue to be an important civic resource for institutional oversight and accountability. Yet, the lack of attention to their socio-cultural provenance enables more powerful stakeholders to flexibly frame narratives around disclosure datasets - celebrating their quantitative approach to accountability when the values reported in the data serve certain vested interests and highlighting their limitations when they don't. These issues underscore the need for further ethnographic research into the semantics, infrastructures, rituals, and institutional incentives that underlie disclosure dataset collection.

This paper begins by providing some historical background on disclosure datasets in the United States and their relationship to efforts to advance government transparency and accountability. I then briefly review literature on the harms and risks that can emerge from problematic biases in datasets and the need for improved documentary practices. In introducing a series of example disclosure datasets in the United States, I then extrapolate the notable concerns regarding the provenance of disclosure dataset infrastructure. I conclude with a call for further ethnographic research into disclosure dataset collection.

2 ACCOUNTABILITY IN/OF DISCLOSURE DATASETS

While public administration theorists have debated the definition of the term "accountability" and shown how it has taken on a "chameleon-like" character, there is general agreement that it is associated with the process of being called "into account" for one's actions. As Bovens [15] indicates, the term accountability is an Anglo-Norman term that is "closely related to accounting, in its literal sense of bookkeeping." Citing Dubnik [27], he notes how the roots of accountability can be traced back centuries to the reign of William I of England, who required that all property holders report a count of their possessions to royal agents, who would aggregate the information into Domesday Books in order to establish a comprehensive accounting of the king's realm. Today, references to "accountability" tend to reverse the direction of scrutiny - with citizens calling their authorities into account. Bovens [15] goes on to define accountability as:

> ...a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences.

According to this definition, the act of accounting (or, in other words, the disclosure of information) on its own is not sufficient for establishing accountability. Transparency, when defined as a means of governing characterized by openness, is often treated synonymously with accountability, yet does not necessarily involve the social relations and exchanges of holding others to account. Thus, scholars have pointed out the limitations of information disclosure, transparency, and "open government" initiatives alone in identifying corruption or promoting public trust in institutions [29, 39, 65]. Despite these notable distinctions, the reporting of information has long been and continues to remain a key factor in the pursuit of accountability. Along these lines, Mary Graham [37] characterizes governing by disclosure as "the third wave of modern risk regulation."

In the public sector, information disclosure as a means of pursuing accountability predates the open government movements that gained traction in many countries around the world in the late 2000s and early 2010s. Freedom of Information Laws, or "sunshine laws" that mandate openness in public decision-making, often in the form of open meetings and records, have been in place in certain countries around the world for centuries and in certain U.S. states since the early 1900s [7]. The passing of the federal Freedom of Information Act (FOIA) in 1967 further bolstered support for this form of legislation, and following the U.S. Watergate scandal, public support for governmental transparency prompted the passing of the Government in the Sunshine Act of 1976, which amended FOIA to require the opening of all government meetings, except for those that might result in the disclosure of sensitive or protected information. Successive amendments to FOIA have responded to changes in the social and judicial environments.

In the 1960s and 1970s, with growing public consciousness around corporate responsibility for environmental pollution and other social consequences for business decisions, a series of laws were passed in the U.S. requiring corporations to disclose information to regulators and concerned publics about their practices. For example, the Home Mortgage Disclosure Act (HMDA) was enacted in 1975 out of concern that unfair lending practices were contributing to the decline of urban communities, and in 1989, the Financial Institutions Reform, Recovery, and Enforcement Act amended the HMDA to require that data on applicant and borrower characteristics be collected and disclosed in order to monitor discriminatory lending. The U.S. Emergency Planning and Community Right-to-Know Act, passed in response to major chemical disasters such as the 1984 Union Carbide gas leak incident in Bhopal, India, mandates the disclosure of information about hazardous materials and toxic emissions present in communities in order to inform residents of risks [45]. A significant body of research has analyzed the effectiveness of information disclosure as a form of environmental regulation, suggesting a multitude of rationales for why such programs work [13, 43, 44, 71].

While these laws set the precedent for transparency through information disclosure, open data movements that sprang up throughout the U.S. in the 2000s set the agenda for publishing government

information as structured datasets. Backed by advocacy groups such as the Sunlight Foundation and the Open Knowledge Foundation, calls to publish government datasets in machine-readable formats and with descriptive metadata gained traction. From his first day in office in 2009, successive executive orders signed by former U.S. President Barack Obama regarding the opening of government datasets culminated in his signing of the "Open Data Policy" - an Executive Order to make machine-readable open data the default for all government information, with exceptions for ensuring privacy and security. The Open Data Policy was eventually codified into U.S. law with the passing of the Open, Public, Electronic, and Necessary (OPEN) Government Data Act in 2018.

With such laws in place, many U.S. disclosure datasets are accessible for download in standardized, tabular formats from open government data portals and agency websites. Based on requirements to supplement open government datasets with structured metadata, many are also accompanied by documents that define key terms in the data, characterize the data's scope, explain aspects of the data's collection, and identify how the data has been and will continue to be managed going forward. While such documents provide important contextual detail about the dataset, they fall short of articulating institutional incentives that can bias data collection, the assumptions built into data infrastructures, and the history and politics of data definitions. Recent scholarship in information studies and critical data studies has demonstrated the significance of strengthening existing data documentation standards [12, 33], and bringing such components of data provenance to the fore.

2.1 Studying Datasets

As computer and data scientists increasingly rely on existing data sources to train machine learning models, researchers have shown how inequitable representations of minoritized groups in datasets have the potential to algorithmically amplify societal biases [58]. In domains such as policing [49], facial recognition [18], and healthcare [56], the reliance on unrepresentative data in predictive capacities has been shown to further entrench historic forms of discrimination. While many studies have focused on analyzing inequities in the contents of datasets to address these issues (for example, [80]), more recent scholarship has called for critically studying the provenance or "genealogies" of datasets in order to extrapolate the assumptions and commitments that informed their creation [25, 68].

Buttressed by the technical character of dataset documentation and their austere form, canonical ideologies tend to position datasets as neutral representational tools, when these structured collections of data may be more aptly characterized as heterogeneous and power-laden systems for signification. In centering certain meanings, datasets are informed by (and in turn animate) a cultural rhetoric that produces certain forms of insight while inevitably creating externalities. Thus, when situated as resources for pursuing accountability, the need for critical analysis of the cultural and political underpinnings of data is imperative.

For the past few years, emerging academic sub-fields have been applying critical hermeneutics to the interpretation of datasets. Frameworks for critically "reading" datasets and studying their provenance have been posed by a number of scholars. For example, Loukissas [48] discusses how to perform "local readings" of datasets; Feinberg [30] describes a method for reading databases "slowly" to encourage critical awareness and reflection, and Bates [10] highlights methods for tracing "data journeys." Engaging modes of data critique, Beaton [11] proposes a form of "data criticism" that attends to the history, genre and form of datasets, and Poirier [60] presents a framework for studying the politics of data signification through critical readings of dataset semiotics.

Further, work ethnographically studying communities generating data and infrastructures supporting data practice demonstrates how datasets emerge as cultural artifacts. Much of this work critiques the notion of data's original "rawness" [16, 24, 34, 78], highlighting the human assumptions, infrastructures, and practices that generate and shape data into various forms. For example, through extended ethnographic research into an AIDS and HIV survey data collection programme in Malawi, Biruk [14] demonstrates how data collectors' beliefs regarding the meaning of local knowledge, along with on-the-ground surveying practices and infrastructures, "cook" survey data in particular ways. Ribes and Jackson [62] show how data collection involves establishing "rituals" to render data as comprehensible. Ethnographic research into data infrastructures [17] and labor [38, 41, 59] show how the cultural underpinnings of data objects and data work are often eclipsed through mechanisms that "delete the social" [32, 70].

Disclosure datasets aggregate data self-reported by institutions the data is meant to hold accountable. Thus, as cultural artifacts, disclosure datasets are riddled with conflicts of interest. Pursuing accountability through the publication of disclosure datasets demands attention not only to disclosures of business activities recorded in the datasets, but also to disclosures regarding data production and reporting practices. Information pertinent to holding data producers and data stakeholders "to account" often goes unrecorded in data documentation, which tends to emphasize technical considerations pertaining to the data without extrapolating the social and political underpinnings of those considerations. In the following case studies, I trace aspects of the provenance of disclosure dataset infrastructure, highlighting certain common dimensions and their role in delimiting the insights that can be garnered from these data sources.

3 CASE STUDIES

While there a number of examples of disclosure datasets maintained by U.S. government agencies, this paper specifically considers three datasets: the Environmental Protection Agency's (EPA) Toxic Release Inventory (TRI), the New York Police Department's (NYPD) Stop, Question and Frisk database, and the Consumer Financial Protection Bureau's (CFPB) Home Mortgage Disclosure Act (HMDA) dataset. I selected these three examples to represent a diversity of data collection domains. In each of these datasets, data gets self-reported by institutions the data is meant to hold accountable.

For each, I first studied data documentation published by the agencies that steward the data, seeking to "invert" [16] the data infrastructure by examining the standards underlying the data's reporting and classification. I then studied the genealogies of these standards by reviewing gray literature documenting their creation and evolution, along with public sentiment towards them. This

included reviewing legislation where standards were encoded into law, judicial cases where the standards were disputed and arbitrated, and news articles/commentaries where the standards were critiqued and debated. For the two case studies involving federal datasets, I reviewed hundreds of public comments submitted to the U.S. Regulations.gov - a website supporting public participation in government rule-making - when notices of proposed changes to the data's collection were posted. It's worth noting that this data source presents its own biases; recent attention has been drawn to swaths of fraudulent and malattributed comments submitted to Regulations.gov in response to particularly controversial issues such as net neutrality and various environmental regulations [9]. While the concern of unrepresentative comments was considered throughout the analysis, the goal of this leg of the work was less to measure public sentiments towards the data reporting programs, and more to document themes in controversies emerging around the programs. These themes were particularly prevalent in cases when coordinated postings by groups with vested interests in the outcome of the process were suspected.

Below I provide brief descriptions of each dataset, indicating how their self-reported nature has resulted in data quality issues, before unpacking the significance of some common properties of disclosure datasets.

3.1 Environmental Protection Agency's (EPA) Toxic Release Inventory (TRI)

The EPA's Emergency Planning and Community Right to Know Act (EPCRA) of 1986 established the Toxic Release Inventory as a mechanism to monitor and inform the public of toxic emissions released in their communities [75]. Every year, certain U.S. industrial facilities are required to report to the EPA the amounts of certain chemical on-site and off-site releases in pounds. Facilities required to report include those that employ more than 10 individuals, release more than a certain threshold of a TRI-regulated chemical, and are classified by a specified set of Standard Industrial Codes, including mining, utilities, manufacturing, publishing, and hazardous waste.

Notably, while the EPCRA mandates reporting of emissions, it does not mandate monitoring of emissions. While other environmental regulations do set certain monitoring standards for specific TRI chemicals and pollution activities, for all other chemicals and activities, facilities are required to report based on a "reasonable estimate" of releases and other waste management quantities. Studies into TRI data quality have uncovered considerable resulting issues - often attributing them to the self-reported nature of the data [23]. For example, a 1990 report by the National Wildlife Federation entitled *Phantom Reductions: Tracking Toxic Trends*, found that reductions in releases at certain large facilities often had more to do with changes in estimation methods and interpretations of the law, rather than actual reductions in emissions [61]. Despite these concerns, advocacy groups regularly leverage the data to campaign for improvements to environmental policies and to support litigation against certain polluting facilities.

3.2 New York Police Department's (NYPD) Stop, Question and Frisk Database

Every time an NYPD officer stops an individual based on "reasonable suspicion" that they committed or were about to commit a crime, the officer is required to fill out a form documenting information about the stop, including the reason for the stop, the demographics of the individual stopped, any actions taken during the stop, and any contraband found on the individual stopped. These reports get aggregated into a database that became available for public download in 2008 as a result of considerable advocacy efforts by the New York Civil Liberties Union in the wake of high-profile police shootings (see, for example, [54]).

Richardson et al [64] refer to stop and frisk data as "dirty data," or data that is inaccurate, corrupt, and systemically biased as a result of "dirty policing." In the 1990s, crime reduction strategies implemented in major cities across the country demanded the production of statistics to generate evidence of policing effectiveness. With certain consequences tied to failures to demonstrate reductions in crime, the policies institutionally incentivized data manipulation - an issue colloquially referred to as "juking the stats." Despite these data quality issues, the publication of the data in 2008 showed an incredible increase in the number of police stops over a 5-year period, and by 2011, the number of stops had increased 700% from when data collection began in 2002. 88% of the time the individuals stopped were found innocent. In the following years, the data became integral in the New York District Court case Floyd, et al. v. City of New York, et al., which ultimately ruled that stop and frisk was being carried out unconstitutionally in New York City and led to a considerable scaling back of the practice [67]. The New York Civil Liberties Union continues to publish annual reports leveraging the data to assess the current state of discriminatory policing in NYC.

3.3 Consumer Financial Protection Bureau (CFPB) Home Mortgage Disclosure Act (HMDA) dataset

In order to ensure that financial institutions are in compliance with fair lending laws in the U.S. (such as the Equal Credit Opportunity Act and the Fair Housing Act), lenders are required to collect and report data on an applicant's ethnicity, race, gender, and income when they apply for a mortgage. With the passing of the Home Mortgage Disclosure Act (HMDA) in 1975, financial institutions were at first required to report demographic information about applicants, aggregated by census tracts. The reporting of this data was largely prompted by concerns that banks were contributing to the decline of certain urban neighborhoods by denying qualified borrowers loans, as well as concerns that financial institutions were engaging in discriminatory lending practices and contributing to the redlining of neighborhoods. Growing concerns about individual-level discrimination in lending prompted the passing of the Financial Institutions Reform, Recovery, and Enforcement Act (FIRREA) of 1989, which required institutions to report demographic data (or what they call "government monitoring information") for every applicant regardless of whether the loan was approved or denied. Determinations regarding which institutions are required to report HMDA data are made based on the institution's total assets, whether

it is located in a Metropolitan Statistical Area (MSA), the number and types of loans it has originated, and whether it is federally insured [3].

When reporting data relating to race, ethnicity, and gender, lenders are legally required to submit the information that applicants self-report when filling out a loan application to the Consumer Financial Protection Bureau (CFPB). However, in cases where an applicant elects not to provide their demographic data, lenders are required to record race, ethnicity, and gender based on visual observation of the applicant or the applicant's surname. In one of the most notable cases of intentional HMDA misreporting, a CFPB investigation found that, for over three years, loan officers at Freedom Mortgage (one of the top ten lending institutions in the U.S.) were instructed to list "non-Hispanic White" as the race and ethnicity for every applicant that elected not to provide demographic data [46]. In general, similar smaller-scale cases of intentional misreporting are inordinately challenging to audit. Still, public officials, non-profit research organizations such as the Urban Institute (see, [73]), and investigative reporting groups (see, [36]) rely on the data to monitor potential discriminatory lending patterns.

There are a number of other examples of disclosure datasets in the U.S. For instance, the Center for Medicare and Medicaid's Open Payments dataset tracks payments made by drug and medical device companies to medical practitioners in order to document medical conflicts of interest. The Federal Election Commission's campaign finance dataset aggregates data reported from political candidates and committees regarding how they raise and spend money. While I don't specifically reference these other datasets in the following analysis, they share critical dimensions I go on to outline below.

4 MALLEABILITY OF DISCLOSURE DATASET DEFINITIONS

The adage "what gets counted counts" has become increasingly prominent in communities calling for critical inquiry into data [26, 72]. To understand what is being counted in datasets, analysts often refer to data dictionaries - documents that encode definitions for key observations and variables in a dataset. When well-documented (which can notably be rare for open government data in the U.S. despite requirements that all data be accompanied with descriptive metadata), data dictionaries give context to the values recorded in a dataset by communicating what the data collectors considered in scope versus out of scope in the process of tallying, categorizing, and measuring. What often goes unrepresented in data dictionaries, however, are the social, political, and historical conditions from which data definitions emerge. Decades of scholarship in information studies and science and technology studies has shown that the setting of standards and classifications is often animated by various cultural commitments and forms of political contestation [17]. This section demonstrates how social advocacy renders the definitions underpinning disclosure datasets as malleable - subject to continual "ontological change" [63] in response to shifting cultural values and political power.

In his book *Defining Reality: Definitions and the Politics of Meaning*, Edward Schiappa [69] argues that definitions can be understood more productively as involving claims of what a word *should* mean rather than what that word means *in essence*. He writes: "Except for periods of what I call definition ruptures, we normally get by just fine assuming that definitions are 'out there,' specifically in dictionaries, and that dictionaries are reliable guides to the nature of the things they define." He goes on to characterize a definition rupture as a dispute over a definition in which parties take issue with the act of defining itself. Such definition ruptures are pervasive in popular discourse: arguments over what counts as a "person" emerge in debates around abortion, and arguments over what counts as a "terrorist" emerge in debates around mass shootings.

For disclosure datasets, definition ruptures tend to emerge around what counts as a reporting entity, along with what counts as a reportable activity. Typically, not every institution that engages in a certain business activity monitored by government entities is required to disclose information through programs that generate disclosure datasets; instead, only those institutions that meet certain criteria or meet a certain threshold of activity are required to report. Notably, disclosure datasets are often embroiled in big government versus small government debates - with small government proponents arguing that mandated reporting places undue regulatory burdens and costs on businesses. Because of this, political debates often emerge over the definitions establishing reporting criteria and thresholds.

Take, for example, the Home Mortgage Disclosure Act (HMDA) data. According to the 2015 HMDA Rule, financial institutions that "originated no fewer than 25 closed-end mortgage loans in each of the two preceding calendar years and meet other reporting criteria such as asset and location tests report their closed-end mortgage activities" [21]. In May 2020, the closed-end reporting threshold was raised from 25 loan originations per year to 100 by the Consumer Financial Protection Bureau, which exempted thousands of institutions from being required to report. The rationale for this change was to reduce the operational challenges of reporting:

> The Bureau recognizes the operational challenges confronted by institutions due to the current COVID-19 pandemic. The Bureau anticipates that this final rule, once effective, will reduce regulatory burden on smaller institutions to help those institutions to focus on responding to consumers in need now and in the longer term. [21]

When first proposed, the change was lauded by real estate finance advocates such as the Mortgage Bankers Association, and hundreds of managers at small community banks submitted public comments in support of the proposed rule [6]. However, the rule also faced considerable opposition from local fair housing organizations, national consumer advocacy organizations, state offices (see, for example, letter from New York Attorney General's Office [8]), and U.S. senators (see [5]). In June 2019, the National Community Reinvestment Coalition (NCRC) composed a comment letter, signed by 158 local and national housing and consumer organizations, in response to CFPB's Notice of Proposed Rulemaking regarding the change to reporting thresholds. In it, they argued that the reduction to the number of reporting institutions would:

> lead to another round of abusive and discriminatory lending. A sizable segment of lenders are more likely to engage in unfair and deceptive practices when

> data is concealed on loan term and conditions and their overall lending patterns to borrowers of different races, genders, and income levels. Unscrupulous lenders will calculate that without publicly available data, members of the public and agencies will have a harder time detecting predatory lending. [52]

When the changes were implemented in 2020, the NCRC, along with several other organizations filed a lawsuit against the CFPB, citing that the changes violated the U.S. Administrative Procedure Act [53]. As of January 2022, the changes remain in effect, and a further scaling back of reporting requirements is scheduled to take effect this year. This evolution in definition not only renders "what counts" in more recent HMDA data incommensurate with the data recorded in previous years; it also signals how external political forces can continuously alter the configuration of disclosure dataset infrastructure, creating new "domains of imperceptibility" [51, 63]. Tracing such ruptures highlights the significance of considering data definitions rhetorically, when often they are only considered referentially.

Similar definition ruptures over reporting thresholds have emerged in regards to which industrial facilities are required to report to the Toxic Release Inventory. In 2006, under the George W. Bush Administration, the TRI Burden Reduction Rule was put into effect, aiming to reduce "the total time, effort, or financial resources expended by persons to generate, maintain, retain, disclose, or provide information to or for a Federal agency" [76]. This rule changed the definition of a TRI reporting facility by raising the threshold at which facilities had to report managed waste of any non-Persistent Bioaccumulative and Toxic (non-PBT) chemicals from 500 to 5,000 pounds, given that no more than 2,000 pounds resulted in releases into the air, water, or soil. Facilities releasing a non-PBT chemical under this amount were permitted to fill out a simplified form that did not require a calculation of total releases. The Rule was implemented despite extraordinary public comment opposing the changes; a report from OMB Watch documented that, of the 122,386 public comments received, more than 99.9% opposed the changes due to potential detrimental impacts on environmental health as a result of unmonitored pollution [57]. Comments opposing the changes were submitted by state agencies or attorney generals in 23 states, along with the EPA Science Advisory Board. The implementation of the TRI Burden Reduction Rule created a situation in which the total releases of chemicals reported in these years was lower than in previous years – not because fewer chemicals were being released but because fewer releases were included in the total calculation. In 2009, the Obama Administration's Omnibus Appropriations Act restored the thresholds for reporting back to their pre-2006 levels [74].

Definition ruptures also emerge over what categories of institutions are required to disclose information, along with what categories of activities they are required to disclose information on. While industrial facilities in mining, manufacturing, and hazardous waste (among others) are required to report emissions to the TRI, as of January 2022, the oil and gas industry has consistently remained exempt from TRI reporting, despite notable pushback from environmentalists. Definitional boundaries have supported oil and gas industries in remaining exempt. In 2012, nine environmental organizations petitioned the EPA to require public reporting of emissions from the oil and gas industry [28]. Responding to the petition, then EPA Administrator Gina McCarthy cited the Environmental Planning and Community Right-to-Know Act's definition of a "facility":

> . . . all buildings, equipment, structures, and other stationary items which are located on a single site or on contiguous or adjacent sites and which are owned or operated by the same person (or by any person which controls, is controlled by, or under common control with, such person). [2]

Because wells are not located on single or adjacent sites, McCarthy argued, dispersed oil wells owned by the same entity should not collectively be considered a single TRI-reporting facility [50]. McCarthy also denied the claim that individual wells should each be considered an individual facility since each well typically employs fewer than 10 individuals, and thus doesn't meet the definitional criteria of a TRI-reporting facility. The definitions underlying disclosure data, while malleable in the face of controversy, can also serve as powerful resources for policing the boundaries of what counts.

In these cases, we see how data definitions are more than stable, factual propositions regarding what the values in a dataset encompass. Data definitions emerge rhetorically and evolve in the wake of political mobilization around what can and should count, along with institutional power to ultimately set the boundaries of what will count. Notably, the malleability of disclosure dataset definitions is often a double-edged sword for advocacy groups. In one sense, the disputability of data definitions establishes the conditions of possibility for extending and/or revising reporting requirements in the wake of changes to social, political, and environmental landscapes. In this sense, the definitions' malleability can be viewed as a feature that enables disclosure data infrastructure to adapt towards strengthening institutional oversight and accountability. Yet, when the definitions underlying disclosure datasets change in response to changes in political commitments and power, it becomes increasingly challenging to perform longitudinal analysis with the available data due to inconsistencies in what is being reported on over time. This is consequential for holding institutions accountable as it poses obstacles to tracking whether issues warranting public concern are being addressed over time, along with the effects of policy on business decisions and activities. Predominant practices in data documentation fail to account for the biases in representation that can emerge from definition ruptures, hiding underlying cultural stories pertinent to interpreting the data.

5 DISCLOSURE FORMS AS INFRASTRUCTURE

In all examples of disclosure datasets presented in this paper, reporters are required to submit their data by filling in a series of structured blanks in a form. In this sense, forms serve as primary data collection infrastructures for disclosure datasets. In the published datasets, the unit observation - or what distinguishes one row of data from the next - is not a person, institution, or activity, but a document representing information about those people, institutions, and activities. If an institution fails to fill out a form related

to something they are supposed to disclose information on, it will not appear in the dataset.

Literary and media scholars have shown the significance of casting a hermeneutic lens on mundane documents such as various forms of paperwork [42]. Lisa Gitelman traces the history of fill-in-the-blank forms to late nineteenth century job printing, showing how the proliferation of these documents structured knowledge and bureaucratic culture in the United States [35]. While the design of fill-in-the-blank forms served to script responses and establish rules for their form, Gitelman argues that these printed documents did not have authors or readers, but instead users. The textual qualities of forms became eclipsed as a "managerial revolution" rendered the filling out of forms as routine and habitual. Rarely do those filling out forms stop to consider the meaning and cultural rhetoric behind prompts and accompanying blanks.

Yet, forms quite literally "form" the resulting data - shaping them according to the form's flow and structuration. Cal Biruk [14] describes similar material infrastructures for data collection (specifically survey questionnaires) as "framing device(s) whose apparent objectivity hides [their] cultural story and commitments." Citing Latour and Woolgar's [47] work documenting how samples collected from rats in a lab are translated onto pages of paper, Biruk argues that designing these data collection documents involves attempts to transform complex social environments into values that can be plugged into databases, thus rendering them visible and comprehensible. Beyond standardizing the figuration of bureaucratic information through scripted prompts, the presence of blanks on a form also enables improvisation and at times fabrication on the part of the individual filling it out [40]. While forms help establish bureaucracy's "objective character" [35], they carry the potentiality for signifying more than the designers intended.

Studying the provenance of disclosure dataset forms demonstrates how standardized prompts emerge from and evolve in the wake of cultural commitments and political tensions. Consider NYC Stop and Frisk data. Whenever an officer stops an individual under "reasonable suspicion" that they committed or were about to commit a crime, they are supposed to fill out a UF-250 form, documenting information about the stop. The UF-250 form presents officers with a series of fill-in-the-blank textboxes for prompts such as "time of stop," "name of person stopped," "age," "weight," and "build." For most other prompts, such as "race," "sex," "was person frisked," and "was suspect arrested," the officer is presented with a series of checkboxes listing possible choices, along with instructions regarding how many boxes they are allowed to and/or required to check. In published stop and frisk data, each row documents one completed form, and each column documents the officer's response to these questions.

The open-endedness of fill-in-the-blank prompts invites opportunities for data entry issues. For example, in 2011, hundreds of forms reported the stopped individual's age as being between the ages of 100 and 999 [60]. While this might suggest that standardizing data collection in more rigid ways may produce better quality data, tracing the socio-political history of possible checkbox selections suggests otherwise.

In a public 2011 version of the UF-250 form, one question asked "What Were Circumstances Which Led to Stop?" Instructions indicated that the officer "Must check at least one box." Possible responses included:

- Carrying Objects in Plain View Used in Commission of Crime e.g., Slim Jim/Pry Bar, etc.
- Fits Description
- Actions Indicative of "Casing" Victim or Location
- Suspicious Bulge/Object
- Actions Indicative of Engaging in Drug Transaction
- Furtive Movements
- Actions Indicative of Engaging in Violent Crimes
- Wearing Clothes/Disguises Commonly Used in Commission of a Crime
- Other Reasonable Suspicion of Criminal Activity (Specify)

The responses to this question provide critical legal justification for the officer performing the stop. The 1968 U.S. Supreme Court case Terry vs. Ohio for the first time set the precedent for criminal search and seizure without probable cause. To enable officers to rapidly respond to crimes without the encumbrance of having to obtain a warrant, this case permitted officers to stop individuals when there was "reasonable suspicion" that they had committed a crime. Specifically, it was decided that "in justifying the particular intrusion the police officer must be able to point to specific and articulable facts which, taken together with rational inferences from those facts, reasonably warrant that intrusion" [1]. It was further decided that "inarticulate hunches" would not warrant intrusion. The checkboxes available for responding to the question "What Were Circumstances Which Led to Stop?" delineate what the NYPD considers justifiable circumstances for a stop.

When NYC's stop-and-frisk tactics went before the U.S. District Court in 2013, Jeffrey Fagan, a statistician and criminologist, was enlisted as a key witness for the plaintiff. He conducted data analysis showing that, from 2004 to 2009, 42% of submitted UF-250 forms had recorded "furtive movements" as the reason for the stop [66]. Officer testimony delivered throughout the case noted the ambiguousness of this category, which could include movements such as:

> "changing direction," "walking in a certain way," "[a]cting a little suspicious," "making a movement that is not regular," being "very fidgety," "going in and out of his pocket," "going in and out of a location," "looking back and forth constantly," "looking over their shoulder," "adjusting their hip or their belt," "moving in and out of a car too quickly," "[t]urning a part of their body away from you," "[g]rabbing at a certain pocket or something at their waist," "getting a little nervous, maybe shaking," and "stutter[ing]." [66]

None of these movements meet the criteria of reasonable suspicion. In Judge Shira A. Scheindlin's [66] ruling that NYC's stop and frisk tactics were violating the U.S. constitution, she argued, " 'Furtive Movements' is vague and subjective. In fact, an officer's impression of whether a movement was 'furtive' may be affected by unconscious racial biases." Similar concerns were raised regarding the category "Fits Description." The ruling went on to argue that the NYPD had not instituted policies and studies to ensure that officers were not generating "scripts" for filling out the form - reflexively

rechecking the same boxes over and over again in order to "facilitat[e] post-hoc justifications for stops where none may have existed at the time of the stop" [67]. Thus, in addition to structural changes to the stop-and-frisk program, the ruling also required amendments to the UF-250 form, including that the form include 1) a narrative section where the officer would be required to justify the rationale for the stop and any ensuing actions, 2) a tear-off portion of the form that could be handed to the person stopped at the end of the encounter indicating the stop reason, and 3) a simplification of the checkbox system for recording stop reasons. The updated form has not been published publicly, and data journalist Dan Nguyen [55] has sent multiple emails requesting a copy of the form without response.

In structuring available responses for self-categorization, disclosure dataset forms also play a significant role in shaping the narratives regarding who and what is being impacted by the activities institutions are reporting on. Many disclosure dataset forms that collect demographic data adhere to sex and racial categorization standardized through the U.S. Census Bureau. The 2011 UF-250 listed two checkboxes next to Sex (Male and Female), and six checkboxes next to Race (White, Black, White Hispanic, Black Hispanic, Asian/Pacific Islander, American Indian/Alaskan Native). Structured as a series of checkboxes with no "Other" option, the form itself prohibits reporting beyond these available categories, so stopped individuals that do not identify along these lines are either forced into a ill-fitting category, or this section is left blank, registering in the final dataset as "Other" or an empty value.

Similar concerns can be traced through the evolution of forms for collecting government monitoring information for compliance with the HMDA. HMDA requires that financial institutions provide an opportunity for applicants to self-report demographic data on an Application form or another form that makes reference to the Application form. Financial institutions are required to submit the information exactly as the applicant records it to CFPB; they are not permitted to make edits to information that the applicant self-reports. Prior to 2017, HMDA required that data be collected on race and ethnicity in aggregate categories. For race, applicants were given the opportunity to self-identify by checking boxes next to one or more of the following categories on the form: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White. For ethnicity, they were given the options: Not Hispanic or Latino.

However, the 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act authorized the CFPB to collect more detailed data regarding race and ethnicity, and in 2015, the CFPB issued a rule amending HMDA to require institutions to revise their application forms in order to allow applicants to report their race and ethnicity according to disaggregated categories [3]. This rule was implemented in 2018. After this, applicants were permitted to select up to five ethnicity categories and five racial categories. In addition to presenting checkboxes next to aggregate ethnicity and racial categories, the form also presented checkboxes next to a series of sub-categories. For instance, under 'Hispanic or Latino' in the 'Ethnicity' category, applicants could further select 'Mexican,' 'Puerto Rican,' 'Cuban,' or 'Other Hispanic or Latino.' When selecting 'Other Hispanic or Latino,' applicants were prompted to fill in a free-form textbox with their "Point of Origin." If an applicant declined to report race or ethnicity information, lenders were expected to report demographic data on the basis of visual observation or surname according to the aggregated categories, and to check a new box on the form indicating that the report was based on these judgment calls.

For some, these changes to the form became a point of political contention. While housing and consumer advocates applauded the changes for considerably expanding possibilities for tracking discriminatory practices against certain subgroups (particularly Asian and Latino subgroups), critics suggested that increasing the number of data points to be reported on would burden financial institutions [31] and introduce data quality issues that could eventually be used as a basis for action against lenders [79]. In public comments on further proposed changes to the form, representatives from financial institutions suggested (with notably little evidence) that lenders confused by the new categories were more likely to check the box declining to provide demographic information and that free-form text fields would be prone to misspellings. Based on these latter concerns, in 2019 CFPB solicited public comments on whether they should ease some of the disaggregated demographic reporting requirements mandated by the 2015 HMDA Rule [20]. While the reporting requirements ultimately stayed in tact, even national housing and consumer advocacy organizations opposing the dilution of the requirements conceded that the CFPB needed to be doing more to ensure consistent data reporting given the changes to the form [4]. As they are forced to grapple with the tradeoffs of promoting flexibility versus standardization in the data infrastructures underpinning disclosure programs, advocacy groups have come to recognize the significance, not only of advocating on behalf of definitions and material forms, but also on behalf of designing scaffolding to support "ontological change" in the data infrastructures [63].

These case studies demonstrate why it is critical to recognize forms as the unit of observation in disclosure datasets. Like data definitions, forms are molded in line with and can evolve in the wake of certain political commitments and interests. They play a critical intermediary role in translating complex social and environmental issues, such as discrimination and environmental injustice, onto paper and eventually into databases. Their enumerability makes measurement of otherwise hard-to-measure phenomena possible, even as their structuration reduces the complexity of what can be reported. Holding institutions to account through disclosure data demands transparency around the development and social evolution of these infrastructures for data collection and reporting.

6 CONCLUSION

For decades, advocacy groups have successfully deployed disclosure data to hold institutions to account. Disclosure datasets have made it possible to track corruption, malfeasance, and discrimination and thus have been integral to investigative reporting, political campaigning, civil liberties advocacy, and legal actions. Mandates to publish the datasets as tabular data, accompanied with metadata descriptions, has enabled stakeholders from various communities to perform their own analyses on the data and their own audits of institutions. Further, government agencies and advocacy groups

have developed tools for making the data more accessible to the public. For instance, the EPA provides access to Toxic Release Inventory data via a "Toxics Tracker" that enables communities to examine pollution and health risks in their own neighborhoods [77]; a team at John Jay College has developed a dashboard for visualizing 14 years of NYPD stop and frisk data [22], and the Connecticut Housing Finance Authority has created a dashboard for visualizing Connecticut's annual HMDA data from 2007-2019 [19]. Calls for expanding disclosure dataset programs (for example, through the mandating of law enforcement offices to report on their use-of-force to a federal database) have been growing in the U.S.

This paper highlighted dimensions of disclosure datasets deserving critical attention from data analysts and critics. The values reported in disclosure datasets and presented in visualizations of the data cannot be separated from key social and political factors shaping the data's infrastructure. For one, disclosure datasets narrate stories about discrimination and other forms of corporate and government wrong-doing almost always solely from the perspective of the institutions they are designed to monitor. The problems with this conflict of interest are demonstrated in examples of institutions deliberately misreporting data or developing creative accounting strategies to "juke the stats." Further, the values recorded in disclosure datasets are not neutral, but instead rhetorically-shaped, as their underlying definitions and data collection tools evolve in the face of political advocacy, contention, and power. In some cases, social pressures have strengthened reporting requirements, and, in other cases, other pressures have weakened the requirements, but in all cases, the data has been biased in favor of assumptions and commitments held by certain social groups. For advocacy groups, the malleability of definitions and material infrastructures has been both a feature and flaw of disclosure data infrastructure - enabling the continual strengthening of the reporting programs, while inhibiting the standardization of the data over time.

While disclosure datasets have historically served an important role in U.S. civic life, the lack of attention to their social and political provenance at times poses challenges to mobilizing the data effectively. The values encoded in disclosure datasets are never "raw"; disclosure dataset infrastructure is always already emergent with various forms of social advocacy. Yet, presented without this context, powerful actors can laud disclosure datasets as advancing a quantitative, evidence-based approach to institutional accountability when it serves their interests, and blame special interest politics for tainting their objectivity when it doesn't. The lack of attention to the provenance of these datasets renders the social actors that have shaped these programs invisible, along with the historical and future role of social advocacy in strengthening them. This demonstrates the need for further ethnographic attention into the creation, maintenance, evolution, and interpretation of disclosure datasets.

More generally, this work speaks to the need for further studies into the social and political provenance of datasets. Cultural dimensions critical to the interpretation of datasets, such as the genealogies of their definitions and reporting forms, are rarely presented in data documentation, contributing to a veneer of objectivity and weakening transparency. Changes to data documentation norms are essential in the pursuit of more accountable data.

ACKNOWLEDGMENTS

Thanks are due to the Spring 2022 cohort of the Smith College Critical Data Analysis Group: Sena Amuzu, Emarie De La Nuez, Juniper Huang, Nicole Tresvalles, and Quinn White.

REFERENCES

[1] 1968. Terry v. Ohio, 392. https://www.law.cornell.edu/supremecourt/text/392/1
[2] 1986. Emergency Planning and Community Right-to-Know Act.
[3] 2015. Home Mortgage Disclosure (Regulation C). https://www.federalregister.gov/documents/2015/10/28/2015-26607/home-mortgage-disclosure-regulation-c
[4] 2019. Comment on Advance Notice of Proposed Rulemaking (ANPR) Concerning HMDA Data Points. https://www.regulations.gov/comment/CFPB-2019-0020-0092
[5] 2019. Letter to the The Honorable Kathleen Kraninger, Director, Consumer Financial Bureau. https://www.banking.senate.gov/imo/media/doc/2019.06.12%20-%20CFPB%20HMDA.pdf
[6] 2019. Proposed Rule: Home Mortgage Disclosure (Regulation C). https://www.regulations.gov/document/CFPB-2019-0021-0369
[7] John M. Ackerman and Irma E. Sandoval-Ballesteros. 2006. The Global Explosion of Freedom of Information Laws Information Regulation: Controlling the Flow of Information to and from Administrative Agencies. *Administrative Law Review* 58, 1 (2006), 85–130. https://heinonline.org/HOL/P?h=hein.journals/admin58&i=104
[8] Jane Azia. 2019. Opposition to Proposed Changes to HMDA Reporting Thresholds Docket No. CFPB-2019-0021/RIN 3170-AA76. https://ag.ny.gov/sites/default/files/hmda_threshold_comments_-_nyag_-_final_-_10.15.19.pdf
[9] Steven J. Balla, Reeve Bull, Bridget C.E. Dooling, Emily Hammond, Michael Herz, Michael Livermore, and Beth Simone Noveck. 2021. *Mass, Computer-Generated, and Fraudulent Comments.* Technical Report. Regulatory Studies Center.
[10] Jo Bates, Yu-Wei Lin, and Paula Goodale. 2016. Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society* 3, 2 (Dec. 2016), 2053951716654502. https://doi.org/10.1177/2053951716654502 Publisher: SAGE Publications Ltd.
[11] Brian Beaton. 2016. How to Respond to Data Science: Early Data Criticism by Lionel Trilling. *Information & Culture* 51, 3 (July 2016), 352–372. https://doi.org/10.7560/IC51303 Publisher: University of Texas Press.
[12] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), 587–604. https://doi.org/10.1162/tacl_a_00041
[13] Lori S. Bennear and Sheila M. Olmstead. 2008. The impacts of the "right to know": Information disclosure and the violation of drinking water standards. *Journal of Environmental Economics and Management* 56, 2 (Sept. 2008), 117–130. https://doi.org/10.1016/j.jeem.2008.03.002
[14] Cal Biruk. 2018. *Cooking Data: Culture and Politics in an African Research World* (illustrated edition ed.). Duke University Press Books, Durham.
[15] Mark Bovens. 2007. Analysing and Assessing Accountability: A Conceptual Framework1. *European Law Journal* 13, 4 (2007), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0386.2007.00378.x.
[16] Geoffrey C. Bowker. 2000. Biodiversity Datadiversity. *Social Studies of Science* 30, 5 (Oct. 2000), 643–683. https://doi.org/10.1177/030631200030005001
[17] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences.* MIT Press, Cambridge, MA.
[18] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.
[19] Connecticut Housing Authority. 2022. Home Mortgage Disclosure Act (HMDA) Dashboard. https://www.chfa.org/about-us/home-mortgage-disclosure-act-hmda-dashboard/
[20] Consumer Financial Protection Bureau. 2019. Home Mortgage Disclosure (Regulation C) Data Points and Coverage. https://www.federalregister.gov/documents/2019/05/08/2019-08979/home-mortgage-disclosure-regulation-c-data-points-and-coverage
[21] Consumer Financial Protection Bureau. 2020. Consumer Financial Protection Bureau Issues Final Rule Raising Data Reporting Thresholds Under the Home Mortgage Disclosure Act. https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-final-rule-raising-data-reporting-thresholds-under-hmda/
[22] Data Collaborative for Justice. [n.d.]. NYPD Dashboard. https://datacollaborativeforjustice.org/dashboard-sqf/
[23] Scott de Marchi and James T. Hamilton. 2006. Assessing the accuracy of self-reported data: an evaluation of the toxics release inventory. *Journal of Risk and Uncertainty* 32, 1 (2006), 57–76. https://www.jstor.org/stable/41761223 Publisher: Springer.

[24] Jérôme Denis and Samuel Goëta. 2017. Rawification and the careful generation of open government data. *Social Studies of Science* 47, 5 (Oct. 2017), 604–629. https://doi.org/10.1177/0306312717712473

[25] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *arXiv:2007.07399 [cs]* (July 2020). http://arxiv.org/abs/2007.07399 arXiv: 2007.07399.

[26] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press, Cambridge, Massachusetts.

[27] Melvin J Dubnick. 2002. Seeking salvation for accountability. In *Annual Meeting of the American Political Science Association*, Vol. 29. 7–9.

[28] Environmental Integrity Project. 2015. Groups Sue EPA to Require Public Reporting of Toxic Chemicals Released During Fracking and Other Oil & Gas Operations. https://www.environmentalintegrity.org/news/groups-sue-epa-to-require-public-reporting-of-toxic-chemicals-released-during-fracking-and-other-oil-gas-operations/

[29] Amitai Etzioni. 2014. *The Limits of Transparency.* SSRN Scholarly Paper ID 2519627. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2519627

[30] Melanie Feinberg. 2017. Reading databases: slow information interactions beyond the retrieval paradigm. *Journal of Documentation* 73, 2 (Feb. 2017), 336–356. https://doi.org/10.1108/JD-03-2016-0030

[31] Michael Flynn and Kimberly Monty Holzel. 2018. The New HMDA Rule's Expanded Ethnicity and Race Categories. *Journal of Taxation & Regulation of Financial Institutions* 31, 2 (2018), 27–29. https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=129377784&site=ehost-live Publisher: Civic Research Institute.

[32] Diana Forsythe. 2001. *Studying Those who Study Us: An Anthropologist in the World of Artificial Intelligence.* Stanford University Press. Google-Books-ID: orNUzuFQeLgC.

[33] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2020). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[34] Lisa Gitelman. 2013. *Raw Data Is an Oxymoron.* MIT Press.

[35] Lisa Gitelman. 2014. *Paper Knowledge: Toward a Media History of Documents.* Duke University Press. https://doi.org/10.2307/j.ctv11smg09

[36] Aaron Glanta and Emmanuel Martinez. 2018. Modern-day redlining: Banks discriminate in lending. http://revealnews.org/article/for-people-of-color-banks-are-shutting-the-door-to-homeownership/

[37] Mary Graham. 2002. *Democracy by Disclosure: The Rise of Technopopulism.* Brookings Institution Press. Google-Books-ID: iGEG3StjgpcC.

[38] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (illustrated edition ed.). Mariner Books, Boston.

[39] Dorathea Greiling and Katharina Spraul. 2010. Accountability and the Challenges of Information Disclosure. *Public Administration Quarterly* 34, 3 (2010), 338–377. https://www.jstor.org/stable/41288352 Publisher: SPAEF.

[40] Barbara Hochman. 2018. Filling in Blanks: Nella Larsen's Application to Library School. *PMLA* 133, 5 (Oct. 2018), 1172–1190. https://doi.org/10.1632/pmla.2018.133.5.1172

[41] Lilly Irani. 2015. Justice for "Data Janitors". https://www.publicbooks.org/justice-for-data-janitors/

[42] Ben Kafka. 2012. *The Demon of Writing: Powers and Failures of Paperwork* (first edition ed.). Zone Books, New York.

[43] Shameek Konar and Mark A. Cohen. 1997. Information As Regulation: The Effect of Community Right to Know Laws on Toxic Emissions. *Journal of Environmental Economics and Management* 32, 1 (Jan. 1997), 109–124. https://doi.org/10.1006/jeem.1996.0955

[44] Michael E. Kraft, Mark Stephan, and Troy D. Abel. 2011. *Coming Clean: Information Disclosure and Environmental Performance.* MIT Press. Google-Books-ID: l83xCwAAQBAJ.

[45] Sarah Lamdan. 2016. Beyond FOIA: Improving Access to Environmental Information in the United States. *Georgetown Environmental Law Review* 29, 3 (2016), 481–512. https://heinonline.org/HOL/P?h=hein.journals/gintenlr29&i=498

[46] Ben Lane. 2019. CFPB finds Freedom Mortgage intentionally reported inaccurate HMDA data. https://www.housingwire.com/articles/49275-cfpb-finds-freedom-mortgage-intentionally-reported-inaccurate-hmda-data/ Section: CFPB / Regulatory, Mortgage.

[47] Bruno Latour and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts, 2nd Edition* (2nd edition ed.). Princeton University Press, Princeton, N.J.

[48] Yanni Alexander Loukissas. 2017. Taking Big Data apart: local readings of composite media collections. *Information, Communication & Society* 20, 5 (May 2017), 651–664. https://doi.org/10.1080/1369118X.2016.1211722

[49] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[50] Gina McCarthy. 2015. Formal Response to October, 24, 2012, Petition to Add the Oil and Gas Extraction Industry, Standard Industrial Classification Code 13, to the List of Facilities Required to Report under Section 313 of the Emergency Planning and Community Right-to-Know Act. https://www.epa.gov/sites/production/files/2015-10/documents/signed_eip_tri_petition_response_10.22.15.pdf

[51] Michelle Murphy. 2006. *Sick Building Syndrome and the Problem of Uncertainty: Environmental Politics, Technoscience, and Women Workers.* Duke University Press.

[52] NCRC. 2019. Notice of proposed rulemaking, HMDA reporting thresholds. https://ncrc.org/notice-of-proposed-rulemaking-hmda-reporting-thresholds/ Section: Testimony & Regulatory Comments.

[53] NCRC. 2020. LexBlog: NCRC Files Suit Against CFPB over HMDA Reporting Thresholds » NCRC. https://ncrc.org/lexblog-ncrc-files-suit-against-cfpb-over-hmda-reporting-thresholds/ Section: In the News.

[54] New York Civil Liberties Union. 2007. Long-Awaited "Stop-and-Frisk" Data Raises Questions About Racial Profiling and Overly Aggressive Policing, NYCLU Says. https://www.nyclu.org/en/press-releases/long-awaited-stop-and-frisk-data-raises-questions-about-racial-profiling-and-overly

[55] Dan Nguyen. 2020. NYPD Stop, Question, and Frisk Worksheet (UF-250); My attempt at getting the latest version as 'just a researcher'. http://blog.danwin.com/request-nypd-form-uf250/

[56] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453. https://doi.org/10.1126/science.aax2342 Publisher: American Association for the Advancement of Science Section: Research Article.

[57] OMB Watch. 2006. *Against the Public's Will: Summary of Responses to The Environmental Protection Agency's Plans to Cut Toxic Reporting.* Technical Report. https://www.foreffectivegov.org/sites/default/files/info/TRICommentsReport.pdf

[58] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (Nov. 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[59] Jean-Christophe Plantin. 2019. Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science. *Science, Technology, & Human Values* 44, 1 (Jan. 2019), 52–73. https://doi.org/10.1177/0162243918781268 Publisher: SAGE Publications Inc.

[60] Lindsay Poirier. 2021. Reading datasets: Strategies for interpreting the politics of data signification. *Big Data & Society* 8, 2 (July 2021), 20539517211029322. https://doi.org/10.1177/20539517211029322 Publisher: SAGE Publications Ltd.

[61] Gerald V Poje and Daniel M Horowitz. 1990. *Phantom reductions: Tracking toxic trends.* National Wildlife Federation.

[62] David Ribes and Steven J Jackson. 2013. Data bite man: The work of sustaining a long-term study. In *Raw data" is an oxymoron*, Lisa Gitelman (Ed.). MIT Press, Cambridge, MA, 147–166.

[63] David Ribes and Jessica Beth Polk. 2015. Organizing for Ontological Change: The Kernel of an AIDS Research Infrastructure. *Social Studies of Science* 45, 2 (April 2015), 214–241. https://doi.org/10.1177/0306312714558136

[64] R. Richardson, Jason Schultz, and K. Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *NYU Law Review* 94, 192 (2019), 192–233. https://www.semanticscholar.org/paper/Dirty-Data%2C-Bad-Predictions%3A-How-Civil-Rights-Data%2C-Richardson-Schultz/9a43ab4a3d1aab2095bfbba60a1ddb8396d5c084

[65] John Roberts. 2009. No one is perfect: The limits of transparency and an ethic for 'intelligent' accountability. *Accounting, Organizations and Society* 34, 8 (Nov. 2009), 957–970. https://doi.org/10.1016/j.aos.2009.04.005

[66] Shira A. Scheindlin, U.S.D.J. 2013. Liability Opinion: Floyd et al v. City of New York. https://s3.documentcloud.org/documents/750446/stop-and-frisk-memoranda.pdf

[67] Shira A. Scheindlin, U.S.D.J. 2013. Remedies Opinion: Floyd et al v. City of New York. https://s3.documentcloud.org/documents/750446/stop-and-frisk-memoranda.pdf

[68] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 317:1–317:37. https://doi.org/10.1145/3476058

[69] Edward Schiappa. 2003. *Defining Reality: Definitions and the Politics of Meaning* (1st edition ed.). Southern Illinois University Press, Carbondale.

[70] Susan Leigh Star. 1991. The sociology of the invisible: The primacy of work in the writings of Anselm Strauss. In *Social Organization and Social Process: Essays in Honor of Anselm Strauss*, Anselm Leonard Strauss and David R. Maines (Eds.). Transaction Publishers, 265–283.

[71] Mark Stephan. 2002. Environmental Information Disclosure Programs: They Work, but Why? *Social Science Quarterly* 83, 1 (2002), 190–205. https://doi.org/10.1111/1540-6237.00078 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6237.00078.

[72] Deborah Stone. 2020. *Counting: How We Use Numbers to Decide What Matters* (1st edition ed.). Liveright, New York, NY.

[73] Sarah Strochak, Aaron Williams, John Walsh, Ben Chartoff, Jerry Ta, Alice Feng, David Hinson, Alex Tammaro, and Sheryl Pardo. 2020. An interactive view of the housing boom and bust. http://urbn.is/MortgageByRace

[74] US EPA. 2009. Toxics Release Inventory Form A Eligibility Revisions Implementing the 2009 Omnibus Appropriations Act. , 19001–19006 pages. https://www.federalregister.gov/documents/2009/04/27/E9-9530/toxics-release-inventory-form-a-eligibility-revisions-implementing-the-2009-omnibus-appropriations

[75] US EPA. 2015. *Factors to Consider When Using Toxics Release Inventory Data*. Technical Report. https://www.epa.gov/toxics-release-inventory-tri-program/factors-consider-when-using-toxics-release-inventory-data

[76] US EPA. 2015. TRI Burden Reduction Rule. , 76932–76945 pages. https://www.epa.gov/toxics-release-inventory-tri-program/tri-burden-reduction-rule

[77] US EPA. 2022. TRI Toxics Tracker. https://edap.epa.gov/public/extensions/TRIToxicsTracker/TRIToxicsTracker.html#continue

[78] Antonia Walford. 2017. Raw Data: Making Relations Matter. *Social Analysis* 61, 2 (June 2017), 65–80. https://doi.org/10.3167/sa.2017.610205 Publisher: Berghahn Journals Section: Social Analysis.

[79] Sarah Wheeler. 2017. On thin ice: expanded HMDA reporting represents new risks for lenders - HousingWire. https://www.housingwire.com/articles/41930-on-thin-ice-expanded-hmda-reporting-represents-new-risks-for-lenders/

[80] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Jan. 2020), 547–558. https://doi.org/10.1145/3351095.3375709 arXiv: 1912.07726.