From Reflection to Repair: A Scoping Review of Dataset Documentation Tools

A Preprint

Pedro Reynolds-Cuéllar[*]
Robotics and AI Institute
Cambridge, MA 02142
pcuellar@rai-inst.com

Marisol Wong-Villacres[*]
Escuela Superior Politécnica del Litoral
Guayaquil - Ecuador
lvillacr@espol.edu.ec

Adriana Alvarado Garcia
IBM Research
USA
adriana.ag@ibm.com

Heila Precel
Robotics and AI Institute
Cambridge, MA 02142
hprecel@rai-inst.com

February 19, 2026

Abstract

Dataset documentation is widely recognized as essential for the responsible development of automated systems. Despite growing efforts to support documentation through different kinds of artifacts, little is known about the motivations shaping documentation tool design or the factors hindering their adoption. We present a systematic review supported by mixed-methods analysis of 59 dataset documentation publications to examine the motivations behind building documentation tools, how authors conceptualize documentation practices, and how these tools connect to existing systems, regulations, and cultural norms. Our analysis shows four persistent patterns in dataset documentation conceptualization that potentially impede adoption and standardization: unclear operationalizations of documentation's value, decontextualized designs, unaddressed labor demands, and a tendency to treat integration as future work. Building on these findings, we propose a shift in Responsible AI tool design toward institutional rather than individual solutions, and outline actions the HCI community can take to enable sustainable documentation practices.

Keywords Dataset · Documentation · HCI · Data Work · Data Practices

1 Introduction

Increasingly, Human-Computer Interaction (HCI) research in Responsible AI (RAI) has stressed that supporting the careful interrogation and understanding of datasets can facilitate the ethical construction of automated predictive technologies across high-stakes domains (e.g., healthcare, criminal justice, employment, education). To help data practitioners engage with much needed ethical data management practices, Responsible AI and HCI scholars—among others—have increasingly championed data documentation as a preferred mechanism [1, 2, 3]. One of the main reasons is the traceability that documentation provides in interpreting decisions made by automated systems. A rapid rise in tools to support this practice has followed, including frameworks, toolkits, and applications. Studies exploring practitioners' perspectives on documentation, however, suggest that despite the rapid growth of these tools, standardization and adoption of documentation practices remains stagnant [4, 5]. Scholars such as Hutchinson et al. argue that misunderstanding dataset development cultures and overlooking AI dataset expertise are among the factors at play in this phenomenon [6]. However, the role of dataset documentation tools in feeding into adoption barriers remains unexplored.

[*] Authors contributed equally to the paper.

In this paper, we use this gap as an opportunity to explore how creators conceptualize documentation tools, and how those conceptualizations might help or hinder mainstream adoption and standardization. Specifically, our work presents a systematic exploration of a corpus of fifty-nine academic contributions across several computing, machine learning, and Responsible AI venues. Each contribution discusses the use of tools or frameworks for dataset documentation. We investigate three aspects of documentation tools that could critically shape users' adoption: first, the goals and motivations behind tool creation; second, how tool creators understand the dataset documentation process itself; and third, these tools' efforts to connect with larger systems, including regulatory frameworks. Our findings highlight how creators often motivate their work via transparency and accountability, but take radically different routes towards these goals based on their own definition of what productive documentation practices entail. Further, creators tend to offer overly general tools that stem from their own needs rather than those of users. As a result, all documentation tools, even those that rely on automation to minimize implementation effort, end up imposing diverse forms of labor on individuals while overlooking the organizational, collective, and infrastructural support required to sustain effective practices. Finally, our findings highlight that without offering concrete guidance and evidence on how to integrate these tools, integration efforts are mostly aspirational.

Our work makes three primary contributions to the literature. First, we provide a scoping review of the dataset documentation landscape, analyzing tools to identify gaps and common threads to help inform adoption efforts. Second, we offer a mixed-method characterization of how the diversity of goals, conceptualizations, and integration efforts behind these tools fragments the value of documenting datasets, critically impacting their adoption and standardization. Lastly, we discuss how documentation fragmentation stems from a systemic gap between Responsible AI research—specifically around tools—and practice, and outline a research agenda for HCI research to support an institutional—rather than individual—design of documentation tools. Together, these contributions aim to advance the design and integration of Responsible AI tools.

2 Related Work

2.1 Responsible AI Tool Design in HCI

The increasing prominence of LLMs and other automated systems has brought about both massive transformation and disruption to our technological ecosystems [7]. With it, concerns about datasets used in model development and evaluation have vaulted to the forefront of public discourse [8]. Critical scholarship across ML and HCI has demonstrated that datasets can be prone to issues across numerous axes [9, 10], including the over-or-under-representation of subjects across protected characteristics [11, 12, 13, 14]; difficult-to-remove hate speech, sexually exploitative material, personally identifiable information, and intellectual property [15, 16, 17, 18, 19]; and poor working conditions and mental health issues for data labelers, to mention a few [20, 21]. Once data is used to train automated systems, other detrimental downstream effects come into view, including allocation and representational harms [22, 23], hostile model responses that are resistant to finetuning [24], hallucinations and security vulnerabilities [25, 26], and labor displacement [27].

The HCI community has responded by providing insight into how practitioners engage and interact with data, as well as by advancing approaches to designing adequate RAI tools. In addition to paving the way on the critical analysis of automated systems and providing a theoretical backing for analyzing resulting harms [28, 10, 23, 22], HCI researchers have also advanced user studies about the design and efficacy of Responsible AI frameworks and toolkits in practice [29, 30, 31, 32, 33, 34]. For example, findings point out that while toolkits are helpful for identifying and mitigating fairness issues, they also have limitations. Balayn et al. note that toolkits *"create a gateway to a narrow view on distributive justice"* [31], shrinking the perceived problems and solutions to those supported by the tool. These kinds of studies help the community interrogate documentation tools and establish ways in which they can be re-designed or adopted. Different kinds of tools across multiple stages of automated system development have been proposed, including model documentation frameworks, fairness and interpretability toolkits, and educational materials [35]. These tools cover a variety of processes, including e.g., data curation and bias detection [36, 37]. Studies report that while these tools can help practitioners navigate questions around RAI, they often require contextual guidance and can lead to steep learning curves making their adoption and integration difficult [33].

2.2 Dataset Documentation as a Solution for Responsible AI

In recent years, dataset documentation has emerged as a salient complement to fairness toolkits and model design frameworks, but with a specific focus on dataset-driven AI harms. Researchers have argued that "documentation has emerged as an essential component of AI transparency and a foundation for Responsible AI development" [38], and the World Economic Forum has recommended that all entities "develop standards to track the provenance, development, and use of training data sets throughout their life cycle" [39]. Specifically, as suggested by the computing truism "garbage in, garbage out", metadata that describes a dataset's sourcing, provenance, and precise contents can enable model developers to assess and mitigate bias down the line. Maintaining thorough documentation makes it possible to assess and resolve IP-related harms or develop strategies for compensating data creators when their data is used by a model contingent on identifying original creators [40, 18, 41]. Researchers have also argued that the reflection triggered by the documentation process is itself a forcing function for dataset creators to predict and stem potential sources of bias [1].

In the United States, the push for dataset documentation sits within a fragmented and developing regulatory landscape. At the federal level, two consecutive administrations have incorporated data transparency into their AI priorities; the now-repealed Executive Order 14110 calls for "authenticating content and tracking its provenance" [42] while America's AI Action Plan emphasizes increasing data quality standards across the board and developing open-source, open-weight, and interpretable models and datasets [43]. In the National Institute of Standards and Technology (NIST)'s guide to AI risk management, dataset documentation is highlighted as a key strategic element for addressing security, transparency, and accountability concerns [3]. However, despite identifying documentation as a priority, there are neither concrete federal guidelines about the right frameworks to follow nor are there oversight mechanisms to ensure compliance.

In response, states have started to implement their own AI-related legislation. California's Generative AI training data transparency act [44], for example, requires that upon model release, all AI developers also release documentation describing the general composition, provenance, and processing of all training data used. Colorado's AI Act [45] identifies Data Cards [46] specifically, requiring them to be released alongside high-risk models. Meanwhile, intellectual property issues are currently being litigated via numerous court cases representing the interests of—among others—artists, writers, journalists, and musicians [47, 48, 49, 50]. Without clarity around national requirements, a number of independent organizations have proposed their own governance solutions. Key efforts here include Stanford University's Foundation Model Transparency index, which independently assesses publicly available models on transparency (including of training data), [51] and Behavioral Use Licensing, proposing a voluntary licensing model for datasets to enforce stated use cases [52]. Academic institutions and some industrial products have also made efforts to self-regulate: NeurIPS' ethics requirements [53], for example, require data and model documentation as part of submission, and Hugging Face's Data Cards [54] allow for documentation to any datasets published via the site.

2.3 Design and Adoption of Documentation Tools

As dataset documentation garnered interest across academic, corporate, and regulatory spheres, researchers in the HCI community have proposed numerous general-purpose documentation tools, understood as structured artifacts (e.g., software applications, batteries of questions, schemas) that facilitate the recording of key information across different stages of a dataset's lifecycle. Examples include, but are by not limited to, Datasheets for Datasets [1], Data Cards [46], the Data Requirements section of Microsoft's RAI Impact Assessments [55], Aether's Dataset Documentation Template [56], FactSheets [57], Dataset Nutrition Labels [58], and Data Statements [59, 60].

These frameworks generally take the form of structured questions for dataset creators to answer before and during dataset development. In addition to providing downstream users with information about data quality, provenance, and known issues, they are typically also designed "to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use" [1]. These efforts reflect a desire across industry and academia for mechanisms to foreground key information associated with dataset production and further use. However, as compared to the fairness tooling ecosystem we discussed in Section 2.1, not much research has directly investigated the motivation, design choices, or subsequent (lack of) adoption of documentation tools by stakeholders. An empirical basis for understanding both how these tools come to be—and what elements contribute to their adoption—is crucial to inform the HCI community's next steps towards effective dataset documentation.

Existing research suggests that dataset documentation is difficult to achieve in practice. First, datasets are difficult if not impossible to retract once published, and formal retraction is limited for preventing continued distribution [61]. In one dramatic but not atypical example described by Schneider et al., falsified clinical trial data "continues to be cited positively and uncritically... eleven years after its retraction" [62]. Documenting data after release can be just as difficult, with potential issues ranging from multiple competing versions of a single dataset to unrecorded procurement practices [63]. Meanwhile, web-scale data collected via internet crawls can include petabytes of data; one popular dataset, the Common Crawl, includes text from over 1 billion pages [64]. Documenting each page is infeasibly resource-intensive.

Even when technically possible, diligent data documentation is rarely aligned with engineers' interests. Orr and Crawford highlight "the messy and contingent realities of dataset preparation," with a focus on four competing elements:

dataset scale, limited access to resources, reliance on shortcuts, and ambivalence regarding accountability for the final product [28]. Alvarado Garcia et al. found that "informed by LLM's need for scale and nascent community-based recommendations, practitioners...[made] decisions that primarily ensured scale and cost-effectiveness," with the slow process of developing high quality documentation falling by the wayside. Bhardwaj et al. and Yang et al. conducted quantitative analyses investigating "dataset cards" on Hugging Face [4] and documentation practices in the NeurIPS Datasets and Benchmarks track [66], respectively. Both studies found that documentation comprehensiveness varied wildly, with Bhardwaj et al. noting specifically that although the most popular datasets generally include populated dataset cards, the vast majority of datasets overall do not. When documentation is provided, fields related to the data structure and function are significantly more detailed than those related to social, ethical, or contextual considerations (which are either minimally filled in or entirely absent). Despite this, Holstein et al. find that "[industry practitioners] typically look to their training datasets, not their ML models, as the most important place to intervene to improve fairness in their products" [29].

These results point to a critical need for empirical analysis designed to understand the motivations driving documentation tool development, and how those motivations do (or do not) lead to real-world adoption. Our is the first study that, to our knowledge, systematically reviews all available dataset documentation tools, assesses the implicit and explicit values they advance, and synthesizes findings into recommendations for the design and future integration of data documentation tools into widespread community practice.

3 Methods

To comprehensively study how researchers approached the design of dataset documentation tools, specifically in the fields of Computer Science (CS), Machine Learning (ML), HCI, and Ethics of Technology, we conducted a scoping review of written literature across several databases of interest. Our focus was on exploring the goals, motivations, concepts, and efforts towards integration behind these tools. We chose to use a scoping review for this study—including for identifying our research questions—given that, as opposed to a systematic review, we do not evaluate studies on the basis of quality nor use them to respond pre-registered hypotheses [67]. Instead, our research is exploratory: we aim to report on practices related to conceptualizing and designing dataset documentation tools; uncover themes across their use; and identify their labor and integration costs. To strengthen the robustness, transparency, and replicability of our procedures, we followed the PRISMA guidelines for scoping reviews [68]. Lastly, we report our findings using reflexive thematic analysis [69].

3.1 Research Questions

Following Arksey and O'Malley, we began the study by stating our research questions. The first author then engaged in an initial review of a set of papers introducing dataset documentation tools. Subsequently, three authors read a small sample of key papers. Following this second round, we refined our initial set to generate the following three research questions:

- **RQ1:** What were the goals and motivations behind the building of dataset documentation tools? (e.g., different tools may advance different values, different politics, potentially driving different goals)
- **RQ2:** How did creators of documentation tools conceptualize the dataset documentation process? (e.g., what are the different perspectives and approaches to documentation as reported by authors)
- **RQ3:** How do these tools connect and integrate with existing systems, regulation, or cultural norms?

3.2 Data Collection

The entire data collection process took place in four stages: database and repository searches, paper screening, eligibility checks, and data extraction and analysis.

3.2.1 Information Sources and Search Strategy

In an effort to capture a wide net of resources in our review, and to specifically target key venues associated with the work of interest, we queried three established databases (ACM Digital Library, IEEE Xplore, and Science Direct), two domain specific repositories (ArXiv and ACL Anthology) and two major conferences (AAAI and NeurIPS). We included these conferences because of their standing in fields relevant for this study, and because they do not systematically deposit their contributions into any of the previously mentioned databases and repositories.

We used a variety of search terms and term combinations to find tools (e.g. "dataset", "documentation"), and nuanced those searches to match contexts of interest with descriptors like "transparency," "accountability," and "provenance." These words emerged from our initial review of papers. We used these words in formal queries across abstracts, titles, author keywords, and sometimes over full texts depending on the affordances of each resource's search engine. Additionally, following De Angelis and Lonetti, we explored grey literature using Google's advanced search. To determine saturation, we used a five-page-noise rule, where we declared the search exhausted after five pages of results with no relevant hits. This generated 1,491 results across all sources databases, and 118 results across Google's services, all collected in February 2025. This set was complemented with a round of snowballing, resulting in a total of 1,582 materials after removing duplicates. Our search only included English materials. The full queries for each resource are provided in Table C in the appendix.

3.2.2 Screening and Elegibility

While scoping our research questions, we developed a set of inclusion/exclusion criteria. We included items that:

- Proposed or advanced a dataset documentation tool
- Proposed or advanced a documentation framework (to qualify, frameworks needed to include at least one dimension related to data documentation)
- Extended an existing dataset documentation tool
- Provided an empirical evaluation of a tool/framework in practice

Conversely, we excluded (1) items that only offered design recommendations and (2) case studies of the application of a dataset documentation tool or framework that did not include any evaluation or reflexive analysis.

One author used these criteria during a first "desk-reject" filtering, removing 1,408 items. We retained articles with ambiguous titles, abstracts, and keywords, leaving 115 items. Using random samples of 20 items, three authors calibrated decisions and reviewed papers for inclusion across three consecutive screening rounds with increasing Cohen Kappa agreement results ($\kappa = 0.29$, $\kappa = 0.54$, and $\kappa = 1.0$, respectively). Following this last round, one of the authors screened the remaining items in the sample resulting in a final set of 59 items for data analysis. Figure 7 in the appendix outlines the process from search to final corpus.

3.2.3 Data Extraction

We retrieved quantitative information from our dataset to power several descriptive analyses. For each item that advanced a dataset documentation tool or framework, we noted (1) if the study was based on a needs assessment/user research study; (2) if a study was used, what type of study; (3) the stated audiences for the tool/framework; (4) if the paper mentioned a *need* to involve stakeholders in tool development; (5) if stakeholders (e.g., users, policymakers) were *actually* involved in the process; (6) the intended use of the tool; (7) if an evaluation of the proposals were included as part of implementation; (8) the degree of automation of each tool (manual, hybrid, fully automated)[2]; (9) any integration with ecosystems of practice, other tools, or regulation; and (10) the terms used by authors to present their proposals (e.g., applications, datasheets, frameworks). Definitions for this categorization of different types of tools are included in Table 1. We also kept track of authors' affiliations based on the information found in the paper ('academia', 'industry', 'government', 'non-profit', and 'other', which included independent researchers or members of the public), and used that classification to determine the distribution of creators across these sectors.

3.3 Data Analysis

We analyzed our corpus using by Braun and Clarke's reflexive thematic analysis approach [69]. We chose this method since it let us recognize and reflect on the influence of our experiences as we navigated the data [116]. Three authors from our team led the analysis, with each assigned a set of roughly 20 items from our final sample. This group of authors independently reviewed each article in their set and inductively developed and maintained codes definitions using the ATLAS.ti qualitative coding software. We maintained an open coding approach for the first round, developing 133 codes grouped into 26 clusters that we later mapped onto our research questions. We used this as a springboard to conceptualize, discuss, and refine themes iteratively across subsequent rounds. Themes were developed together with memos from each of us involved in the coding process, and with reference to representative quotes. We want to

[2] *Manual* tools were almost exclusively based on natural language provided by a human; *hybrid* tools used natural language input from a human and to produce documentation artifacts; and *automated* tools used machine readable information with almost no natural language input to produce documentation.

Tools Definitions		
Type of tool	**Definition**	**Items**
Application	Programs, web interfaces or digital tools provided to users as support to create dataset documentation	[71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82]
Datasheet	Items named "datasheets", as well as proposals that provided structured batteries of questions that once answered became the documentation artifact	[83, 56, 84, 85, 86, 87, 1, 88, 57, 89, 90, 91, 92, 93]
Framework	Structured, reflexive guidelines (sometimes in the form of questions) aimed at helping users define key information to include in a documentation artifact	[94, 95, 46, 96, 59, 97, 98, 99, 100, 101, 102, 38, 58, 2, 6]
Markup format	Machine-readable format, usually implemented as an application in the form of a plugin or package	[103, 104, 105, 106, 107]
Schema	Proposals that combined metadata information along with structured questions meant to become documentation, similar to datasheets	[108, 109]
Study	Items describing an evaluation or study of a dataset documentation tool in practice	[110, 111, 112, 66, 113, 114, 4, 5]
Toolkit	Items providing reusable components such as templates or exercises that practitioners can tailor and use to plan and execute documentation over datasets	[60, 115]

Table 1: Definitions and items for each type of dataset documentation tool identified in the systematic review. The table categorizes seven tool types: Application (13 items), Datasheet (14 items), Framework (15 items), Markup format (5 items), Schema (2 items), Study (8 items), and Toolkit (2 items). Classification was based on the language used by authors in each paper when introducing their proposals.

acknowledge that, in presenting these themes, we often directly quote authors from papers included in the sample and sometimes paraphrase them, often to benefit the flow of our arguments, and always with great attention not to add our own interpretation.

3.4 Positionality of the Research

One of the guiding questions for our review is if and how the Responsible AI community is advancing standards and norms around dataset documentation in light of its commitment to fairness, accountability, and transparency. While our community signals that these goals are important, the way individual researchers hold to them might differ and, as we mention in Section 6, there is no clear path for transforming these values into collective norms. We want to stress that neither our review nor our research intends this as a critique to the researchers or works we review, but rather as a reflection on how we can collectively move towards these goals. This reflection is a result of our experiences and intellectual commitments as authors. All four authors have formal training in HCI. The first author is an HCI and robotics researcher committed to critically studying sociotechnical systems by centering the experiences of groups greatly affected by automated technologies, yet with little influence or power over them. The second author explores the design of data-driven technologies that support historically marginalized groups. The third and fourth authors investigate the data work practices of practitioners developing LLMs. We also want to note that the vast majority of work included in our review is representative of a Western approach to dataset documentation. Therefore, it can only be representative of the way in which that epistemological position understands data production, consumption and dissemination, as well as its corresponding governing processes. As authors, we acknowledge that, while often holding critical views around how these processes and governance take place, we also partake in them. Consequently, this subjective tension is reflected in our analytic approach.

4 Findings: Descriptive Statistics

In this section, we provide descriptive statistics of our corpus regarding the types of documentation tools we found, how tools' prevalence and level of automation has progressed over time, and the audiences they consider.

4.1 Diversity and Evolution

We found that the total number of dataset documentation tools has increased over time, with a noticeable concentration of new tools between 2022 and 2024 (31 out of 51 proposals). The types of tools and terms used describe them varied; we ultimately identified seven: toolkits, frameworks, applications, datasheets, markup formats, and schemas. Definitions for each tool can be found in Table 1. *Toolkits*, *datasheets* [1, 88], *frameworks* [96, 38], and *schemas* [60] offered structured guidance to help users document. *Datasheets*, for example, entailed batteries of questions, while *schemas* combined metadata with questions. *Applications* and machine-readable *markup formats* facilitate the automation of the documentation process. Frameworks (n=16), applications (n=13) and datasheets (n=13) are the most popular types, amounting to 82% of our total sample. Figure 2 shows these distributions.

We also note a prevalence of tools that require manual input to produce documentation (n=25). However, recent years show an increase in hybrid tools that use natural language to automatically produce documentation (n=17) and automated tools that use machine readable information to generate documentation (n=9), a trend that matches advances in large language models. Figure 1 visualizes these trends.

The increase in both volume and types of tools over the past decade or so appears to signal a high diversity in creators' conceptualizations of documentation structure and role within larger processes. Meanwhile, an increase in automated tools may simply coincide with the advent of mainstream, more capable LLMs.

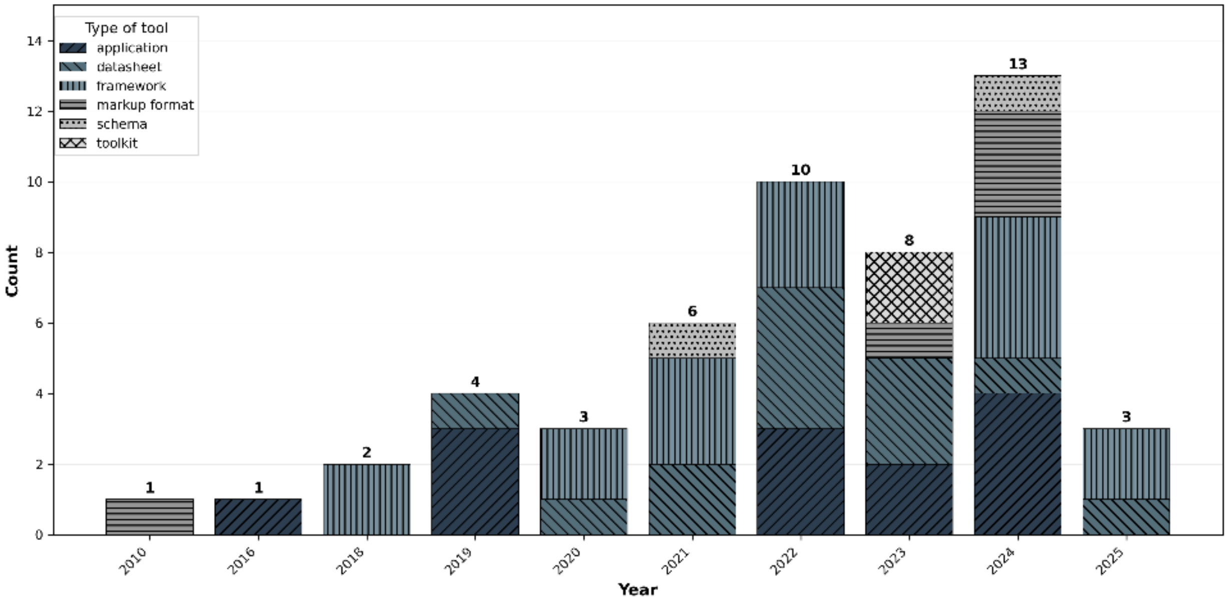

Figure 1: Distribution of different types of tools over time based on the sample in our corpus. It is worth noting that 2025 is an outlier year given our sample was finalized in March of that year.[3]

.

4.2 Creators and considered Audiences

To explore our first research question (goals and motivations of tool development), we first looked at author's affiliations. We found that half of the papers in our sample were authored solely by researchers in academia (31 papers, 52.54%), and a majority of papers included at least one author from academia (46 papers, 77.97%) . Researchers in industry contributed to 27 papers (45.76%), and participated in 12 papers (20.34%) with members from other groups. Participation in papers in our sample from other groups included government (7 papers, 11.86%), non-profits (6 papers, 10.17%), and other (1 paper, 1.69%). We found a similar distribution when looking at authors per affiliation across all papers (with total authors as the denominator). In order: academia (180 authors, 51%), industry (139 authors, 39.38%), government (20 authors, 5.67%), non-profits (13 authors, 3.68%), and others (1 author, 0.28%). Our analysis also highlighted that documentation tools' creators used diverse terms (e.g., dataset creators, dataset experts, and data curators) for

[3] Distribution of different types of dataset documentation tools over time. Data shows a noticeable increase in the production of these tools between 2021 and 2024

defining their target user profile. Dataset creators emerged as the term that creators most commonly used to identify their audience. While some tool types (e.g., datasheets) tended to identify a wide variety of audiences, most failed to engage users in the design process. Visualizations for this and other distributions can be found in appendix A.

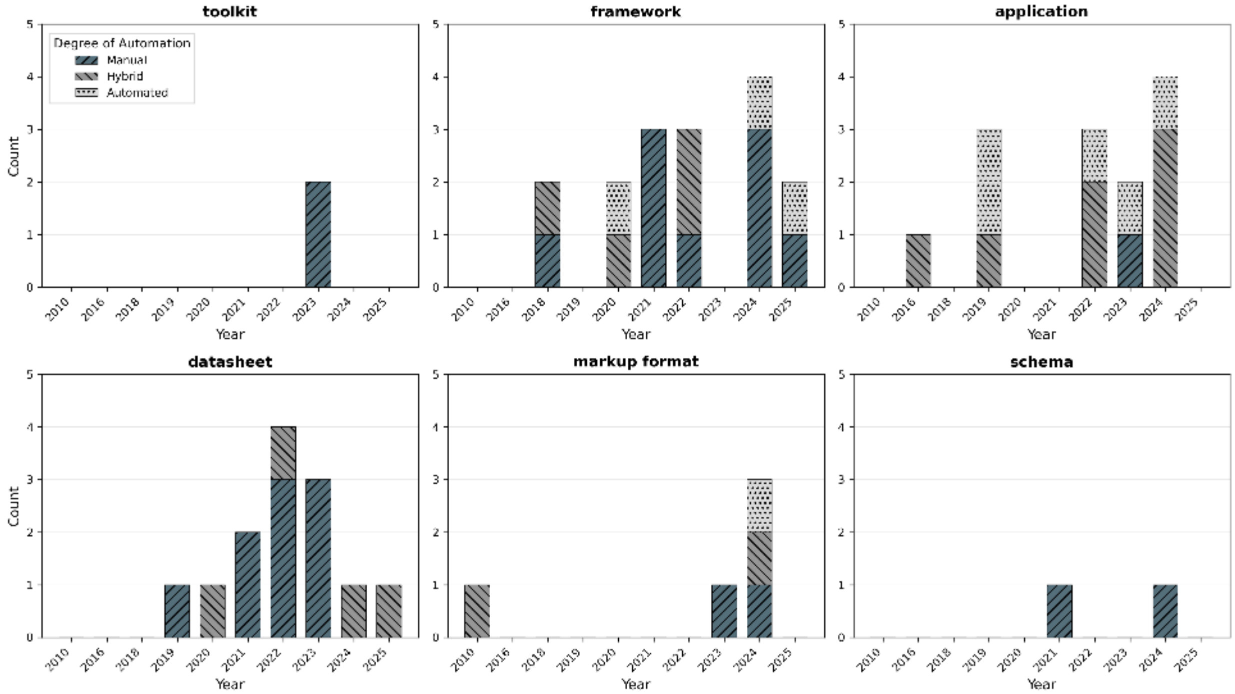

Figure 2: Distribution over time of different approaches to the production of dataset documentation across the six categories of tools in our sample

We further explored this gap between designers and stakeholders, and found that more than half of the tools in our dataset (n=30) did not mention engaging relevant stakeholders during the design process. Of the remaining tools, most only described eliciting views without explicitly integrating them into tool design. Only two items in our sample described eliciting **and** integrating stakeholder input (See Figure 3).

5 Findings: Thematic Analysis

The qualitative analysis of our corpus foregrounded four critical aspects of how tools' creators conceptualize dataset documentation tools: a disagreement on how to operationalize the value of documentation; a tendency to design decontextualized tools disconnected from user experiences; an increase of undiscussed labor demands, even when moving towards automation; and a view of tools integration as an aspirational future endeavor. Each of these aspects illustrates how creators assumptions about productive documentation practices impact the adoption of dataset documentation tools.

5.1 From Reflection to Repair: A Spectrum of Views on the Value of Documentation

Our analysis suggests that a prevalent goal motivating the design of documentation tools is to counteract the *"undesired consequences and negative downstream effects in the whole machine learning pipeline due to data issues"* [93]. Indeed, most creators agreed that *"the quality of a dataset used to build a model will directly influence the outcomes it produces"* [58] and positioned documentation as the mechanism for ensuring the transparency and accountability of dataset work. In particular, tools' creators argued that only adequately documented data *"can be appropriately used"* [114] given that high-quality documentation *"is a lever that enables accountability"* [38] and *"directly impacts the transparency, reliability, and reproducibility in the field of data-driven research"*. However, authors differed on how, in practice, documentation might achieve these benefits. Views included: (1) that documentation serves as personal reflection for dataset creators, (2) that documentation would allow downstream users to scrutinize datasets, and (3) that documentation would repair broken dataset infrastructure. Given that most tools were designed by two groups—

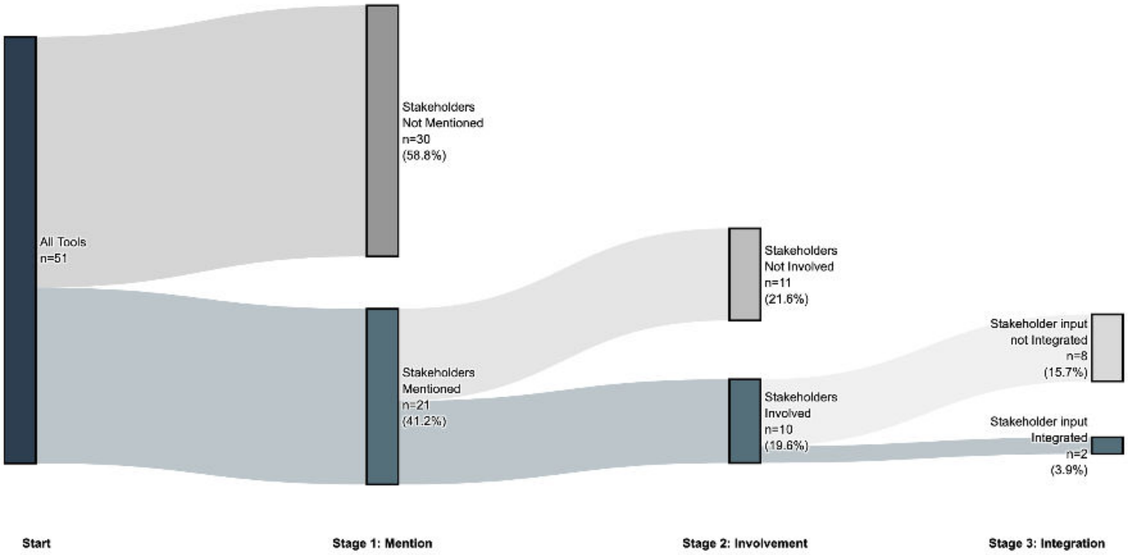

Figure 3: Comparative analysis of stakeholder engagement and integration pipelines including the percentage of proposals that specifically mentioned stakeholders as part of the design/use stages compared to the number of proposals from within that group that included any stakeholder in the process. It also showcases the percentage of proposals that included stakeholders during the design/testing stages compared to the number of proposals that included concrete features or guidance towards integration to stakeholders.

researchers in academia and in industry—one might expect significant overlap around the value of documentation. However, our findings show this assumption does not hold. These perspectives profoundly shaped tools' documentation attributes, target audience, and intended impact. Moreover, they foregrounded the need for evidence supporting the assumption that documentation tools can increase datasets' and AI's transparency and accountability.

5.1.1 Eliciting Dataset Creators' Careful, Personal Reflection

Some proposals described documentation as the means to facilitate personal reflection by dataset creators about dataset-related decisions. As Gebru et al., the creators of 'Datasheets for Datasets' explained, the argument behind this view is that personal, careful reflection on dataset management increases *"transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks"* [1]. Tools advancing this view relied heavily on reflection-eliciting guidance (e.g., questions) particular to the tool's domain or end-goal. For example, in the context of documenting annotators' histories and experiences, Crowdworksheets [95] included considerations for documenting *"how much annotator's identity, lived experience, and prior knowledge of a problem space matters for the task at hand, and how it impacts what the resulting dataset is intended to capture."*

Tools holding this view often struggled to balance between reflection and pragmatism. For example, 'Datasheets for Datasets', which is not domain-specific, prioritizes reflection to the point of advising against automation (as this could *"run counter to our objective of encouraging dataset creators to carefully reflect on the process of creating, distributing, and maintaining a dataset"* [1]). Recognizing the overhead that reflection can create for users, other tools (e.g, Marandi et al. and Luthra and Eskevich) offered a mix of practical, time-efficient sections (e.g., concrete checklists and machine-readable formats) and reflection-eliciting sections (e.g., the dataset creators' positionality [97]).

The reflection viewpoint also saw documentation as an activity that needed to take place *"prior to any data collection"* [1] so as to *"promote a culture of responsible data stewardship from the very beginning of the research lifecycle"* [97]. As such, tools had components for dataset creators to conduct "a thorough structural analysis of the dataset during the planning phase" [97], thereby identifying data ethics issues early on.

Despite the emphasis on reflection-oriented components and early documentation, we found that, in their initial proposals, tools under this view did not provide evidence of their impact. Specifically, they tended to miss evaluations that their reflection-based design choices increased users' reflexivity capacity and, more importantly, improve the transparency, reliability, accountability, and ethical integrity of datasets. In some cases, however, other authors offered evidence for tools with a higher level of adoption (e.g., [46, 111])

5.1.2 Offering Dataset Consumers the Ability to Scrutinize Datasets

The next operationalization frame we identified sought to provide dataset consumers with scalable means for intensely scrutinizing datasets. As the creators of the 'Dataset Nutrition Label' explained, this view relies on the idea that, *"to improve the accuracy and fairness of AI systems, it is imperative that data specialists are able to more quickly assess the viability and fitness of datasets, and more easily find and use better quality data to train their models"* [58]. As such, tools under this perspective included functionality for augmenting the visibility of critical dataset attributes, like those related to fairness or bias concerns [71], and details of *"operations conducted during the data preparation, cleaning, and quality analysis phases in a typical AI life cycle"*[96]. While these tools also included qualitative attributes, their emphasis was on factual rather than reflective information such as metadata, provenance descriptions [58], quality and remediation assessments, [96], and potential bias issues [73]. In a few cases, tools included a comments section where consumers and creators could share concerns [58]. Holland et al. explain that quantitative combined with factual qualitative information can better motivate reflection on the consumer side: *"[the label] enables the data specialist to better understand and ascertain the fitness of a dataset by scanning missing values, summary statistics of the data, correlations or proxies, and other important factors. As a result, the data specialist may discard a problematic dataset or work to improve its viability prior to utilizing it."* [58].

Corpus tools under this view often missed identifying reasons for requiring specific quantitative or qualitative documentation attributes. Further, similarly to the reflective tools discussed above, tools emphasizing consumer scrutiny did not provide evidence that their design choices led to easier, quicker, or more effective dataset analysis. Thus, there is a pressing need to demonstrate how facilitating dataset scrutiny may eventually lead to increased transparency and accuracy.

5.1.3 Facilitating the Repair of Broken Dataset Infrastructure

A third set of tools, championed documentation as the means to repair incomplete and messy data ecosystems. Work from Hutchinson et al., for example, argued that data ecosystems are increasingly breaking down, creating datasets that are *"poorly maintained, lacking in answerability, and have opaque creation processes"*. As such, they proposed to use documentation as *"a deliberative and intentional methodology, rather than the post-hoc justifications that are sometimes observed when datasets are developed hastily or opportunistically."* In contrast with this view of repair as intentional and transversal actions, most other tools under the repair view proposed repair as quick, seamless fixes taking place after dataset creation. For example, to solve the problem of datasets in an infrastructure residing in different storage systems and taking *"a variety of forms, such as structured files, databases, spreadsheets, or even services that provide access to the data"*, Halevy et al. proposed harnessing datasets' metadata and users' comments to organize datasets at scale. The tool's goal was to work *"in the background, in a non-intrusive manner"*. Similarly, the authors of [94] addressed the complexity of identifying multiple versions of the same dataset in a data repository by proposing a Generative AI-powered framework that produced documented benchmarks for data versioning without users' intervention. Other quick fixes entailed creating artifacts (e.g., rubrics [4] and lightweight documentation formats [114]) that could help quickly compensate for poor-quality documentation. As we noted in section 4, the rise in these types of tools coincides with breakthroughs in LLM technology. We discuss the potential ramifications of this move towards automation in the context of this "quick-fix" approach.

Design-wise, tools pursuing repair as quick fixes tended to select documentation attributes that were easy to record and machine-readable (e.g., metadata [76], lineage and transformations [94], and semantic descriptions [107]). However, tools often failed to discuss the implications of focusing on some documentation elements at the expense of others. For example, Fabris et al. missed discussing the impact of compacting the 'Datasheets for Datasets' documentation format, which left behind the reflection-based questions that characterize that Datasheets original design. Furthermore, similarly to tools in the other two perspectives, we found tools did not provide evidence of how different repair strategies managed to repair dataset ecosystems and increased accountability and transparency.

5.2 The Role of De-contextualization in Devaluing the Dataset Infrastructure

As Hutchinson et al. argued, data infrastructure is continuously undergoing a *"systemic devaluation of dataset work"* and of data work more broadly [6]. Our analysis highlights that avoiding such devaluing was critical for many corpus authors such as [56]. As they explained, their tool strives to increase the visibility of existing data work for all possible actors involved: *"someone completely unfamiliar with the dataset would be able to make an informed decision about whether and how to use this dataset responsibly"* [56]. However, our quantitative findings indicated that, as most tools broadened the scope of who can document, they did so without including users and stakeholders in the design process (n=30, 58.8%). Our thematic analysis stressed how this disconnection from users ran the risk of further complicating and devaluing dataset production.

5.2.1 Wide-encompassing tools, narrow-encompassing contexts

In studying how authors reported on the creation of their proposals, we observed a tendency to prioritize domain-agnostic characteristics. This trend matches our quantitative finding that the most open-ended, least constrained tools—datasheets and frameworks—comprised a majority of the total tools in our sample (29 out of 51). For example, 'Data Cards' framed their approach to documentation as *"an underlying framework for transparency reporting for domain and fluency-agnostic readability and scaling in production contexts."* [46] The domain-agnostic characteristic, we found, was also relevant for creators leveraging these particular tools: *"This paper focuses on Datasheets: a technique- and domain-agnostic, lay-language context document for training data. Datasheets are versatile: they can be taught early in ML education to students who will go on to work in diverse domains using a variety of techniques"* [111].

Often, these context-independent design approaches attended to a wide range of goals, overlaid across several stages of the dataset production process, and cutting across domains of application. However, we found that, even when creators recognized that *"... one size does not fit all"*, they often passed the burden of contextualization to their users: *"This work has demonstrated that although FactSheets will contain some common elements, different FactSheets will generally contain different information, at different levels of specificity, depending on domain and model type. They will also contain different information for different industries and the different regulatory schemes within which these industries operate"* [57]. Similarly, the surrounding context was at times obscured by the pursuit of generality with few mentions of specific institutional arrangements, data governance structures, and relevant regional/national regulation of relevance to future users.

Our analysis highlights that more contextualized tools often built upon less contextualized ones. For example, 'Artsheets for Datasets' focuses *"... specifically on the unique considerations (such as social, legal, cultural, historical, and environmental factors) that arise in the development of art datasets."* Along with providing potential users with relevant contextual information to the field of art data, authors provided more detailed implementation guidance to counter the *"common presumption of "one-size-fits-all" ethics checklists in the field by reinforcing the principle that ethical frameworks must be carefully adapted to each use... "* [84]. We found that tools that narrowed the domain and context of application thus presented a clearer picture of use along with responding to specific needs (e.g., terminology, standards, etc.), signaling the usefulness of their approach. However, creators of these domain-specific tools also stressed that a narrower focus came at the cost of reducing the utility of their tools for other domains (e.g., *"focus on healthcare datasets may restrict its broader applicability to other domains"* [93]), and limited the possibility of comparison across datasets (e.g., *"Future work requires a more principled approach for extending and adapting Data Card templates without compromising comparability"* [46]). This highlights an unresolved tension around the right scope for dataset documentation tools.

5.2.2 Scopes of stakeholder engagement

As we mentioned before, few tools engaged stakeholders in their design and integration process. Our thematic analysis, however, stresses that almost half of our corpus tools did discuss the importance of considering stakeholders (e.g., users, consumers and regulatory institutions) at some point of the tools' creation timeline. Further, our analysis highlighted that the high variation of terms to describe stakeholders (e.g., "Data scientists" [57, 2, 112, 58, 115, 80], "data practitioners" [96, 77, 99], "data creators" [82, 100], and "dataset creators" [66, 84, 72, 89, 105]) was accompanied by a lack of definition of such terms and specifications of the date life-cycles where these roles operate. This inconsistency highlights a lack of collective agreement around foundational categories and roles in the dataset production chain, which hinders the identification and engagement of critical stakeholders.

Creators' tendency to draw inspiration from other fields, we found, can also exacerbate issues around stakeholder engagement. For example, Factsheets were *modeled after a supplier's declaration of conformity (SDoC) [...] a document to "show that a product, process or service conforms to a standard or technical regulation, [...] used in many different industries and sectors including telecommunications and transportation* Arnold et al.. Authors, however, do not mention how users of SDoC's may have contributed to translating their use to the realm of datasets nor what limitations this 'modeling-after' might have. Interdisciplinary connections can produce interesting ideas while broadening—and possibly obscuring—the universe of relevant parties to include in the process. Without close understanding of how documentation tools operate in other fields and what stakeholders make that use successful, interdisciplinary connections can prove challenging. We return to the tradeoffs of interdisciplinary work in the next section.

5.3 The Taxing Yet Undiscussed Labor Demands of Documentation

Our analysis stressed how, whether manual, hybrid, or fully automated, documentation tools impose labor demands upon practitioners in the form of time, access to institutionally-bound information, and interpretation. Further, we found that tools tended to place the responsibility of documenting on individuals rather than institutions. Yet, the authors in our corpus rarely discussed the labor demands of documentation and their broader implications. Finally, we observed an increasing shift to automated documentation tools (e.g., automating data extraction, incorporating LLMs into the data documentation pipeline, and hybrid approaches) without recognizing or discussing how it could reshape the nature and distribution of documentation labor.

5.3.1 Information Access and Dataset Interpretation: Undiscussed Forms of Labor

Several authors in our corpus explicitly recognized that documenting can be time-demanding (e.g., it can take between six to twenty-four hours to create a FactSheets template [110]), require diverse forms of human engagement (e.g., "active engagement with regulators and corporations" [103], and "work with experts in other domains such as anthropology, sociology, and science and technology studies" [1]), and needs institutional incentives [95]. However, it was less common for them to discuss these and other demands in terms of labor impositions for users. Our analysis highlighted the intensity and implications of two potentially taxing yet undiscussed labor demands of documentation tools: access to institutionally-regulated information, and interpretation of dataset work.

We found that many of the documentation tools in our corpus assumed that users would have seamless access to institutionally-regulated information. For example, 'Datasheets for Datasets' asks *"Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?"* [1], which assumes users can easily access institutional policies or human resources records. Similarly, in the 'Data Labeling' use case, the Croissant-RAI vocabulary requires detailed information about annotation processes, including annotator demographics, instructions, and devices used. As the authors emphasized, documenting these attributes is essential for assessing label quality, but it also presupposes that the institutions where users operate have in place infrastructures that record and store this information in a systematic way [106].

Our analysis also stressed that many documentation tools demand ethical interpretive labor. While some interpretation is a routine aspect of knowledge work, the ethical interpretive labor required by documentation tools constitutes a substantially more demanding activity. It requires practitioners—who often come from a technical background and lack ethics training—to make subjective but normative, anticipatory, and context-specific judgments about potential dataset implications. For example, tools such as Crowdworksheets asked users to judge the implications of datasets involving cultural or social contexts they might not fully understand. The tool includes questions such as, *"does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why"* [95], and *"Are there certain perspectives that would be harmful to include? If so, how did you screen these perspectives out?"* [95]. As Heger et al. found, such interpretation could result in invisible and unevenly distributed labor burdens to users: *"practitioners creating dataset documentation had trouble making connections between the questions they were asked to answer and their RAI implications and difficulty providing information that someone unfamiliar with their datasets would need to understand the data"* [80].Recognizing this form of interpretive work is therefore essential for adapting documentation tools so they provide appropriate guidance, distribute responsibilities across institutions rather than individuals, and avoid imposing additional unplanned workloads.

These examples illustrate that documentation requires more than the visible time of writing and coordinating. It also demands the less visible—and thus, less acknowledged and compensated—labor of figuring out how to navigate institutional regulations to access information requested by tools along with interpreting, translating, and judging datasets. These two forms of labor also shed light on the tendency for tools to center individual users as solely responsible for successful documentation, ignoring that the labor of documentation goes beyond the individual. Institutional buy-in and cooperation with other key stakeholders in adopting and enforcing the use of documentation tools is critical to ensure its adoption. If records are scattered or poorly maintained, or if there is a lack of expert support on domain-specific and ethical issues, documentation tools can generate additional workloads, thereby reinforcing users' perceptions of documentation as burdensome.

5.3.2 The Move Towards Automation: Risks Of Disguising Labor

Relatedly, several articles in our corpus proposed hybrid or fully automated approaches to documentation for three purposes: scalability through saving time and resources, compliance with institutional policies, and expanding the scope of documentation analysis. However, we found that the move towards automation could transform and further obscure labor demands to users, engendering risks that are harder and more resource-consuming for users.

Our analysis highlighted how enabling scalable solutions that can cater to an ever-growing scale of current AI technologies was a priority for some tools moving towards automation. As Halevy et al. explained, automation can achieve this goal by preventing effort duplication and reducing manual work during documentation analysis: *"if we can identify natural clusters to which datasets belong, then not only we can provide the users with a useful logical-level abstraction that groups together these different versions but we can also save on metadata extraction."* As the creators of 'DataDoc Analyzer' further argued, automation can reduce documentation processing to fifty and sixty seconds for unseen documents and to up to twenty to twenty-five seconds [73]. Automation, other authors argued, could also relieve users from the need to manually record objective, traceable facts: *"Code repositories adapted for use in AI development can record key facts about training data and model versioning"* [110]. Our analysis showed that, often the time and resource demands of automation did not fully disappear but shifted to other forms or groups. In the case of the 'Data Nutrition Label', for example, creators discussed how their automated analytics module would require community work to prevent failure: *"The veracity and usefulness of the Ground Truth Comparison module depends on the accuracy of the "Ground Truth" dataset [..] a mitigating step is to build Labels for ground truth datasets themselves. If these Labels include community feedback and comment modules, dataset authors can address the issues directly"* [58].

Another group of creators in our corpus framed automation in documentation as a means of ensuring a systematic and sustainable compliance with emerging regulatory and organizational policies. In this sense, automation is positioned both as cost-saving and efficient, as well as a mechanism for embedding policy requirements into documentation workflows. For instance, the authors of 'DataDoc Analyzer' pointed out that *"recent public regulatory initiatives such as the European AI Act and the AI Right of Bills [..] call for this documentation to be easy to understand by non-experts to bridge the gap between technology and end users"* [73]. Similarly, the creators of 'Open Datasheets' emphasized that automation can *"ensure that open datasets align with their Responsible AI policies,"* streamlining much of the evaluation process. Our analysis stressed that such abidance by regulation through automation can also shift forms of labor to other actors. As [80] explain, *"human review will still be necessary to make decisions based on organizational policies."*. Further, relying on automation to abide by regulations could also force users to navigate obscured decisions about datasets and their attributes. For example, the creators of 'Goods' explain how automating dataset organization based on metadata and users' comments depends on an automatic exclusion of *"many types of uninteresting datasets"*, and a normalization of obvious redundancies. These automatic and hidden decisions could generate further unanticipated and hard-to-manage forms of labor for users.

Other authors positioned automation, often through LLMs, as the pathway to extend documentation and dataset analytic possibilities beyond human capacity. Fox et al., for example, highlighted how documenting multi-hop dataset transformations is limited by human capacity and positioned the use of LLMs as a critical tool to navigate this barrier: *"Recently, Large Language Models (LLMs) have been showing promising results in annotation and semantic tasks such as column type detection, entity matching, benchmarking table union search, and more"* [94]. Likewise, Holland et al. suggested introducing more intelligence to next 'Dataset Nutrition Label' iterations to better detect subtle biases in datasets that humans might not recognize: *"Analyses of machine bias indicate that zip codes often proxy for race, but many others proxies still exist, especially as the models themselves approach levels of complexity that are difficult or impossible for humans to comprehend and new or unexpected proxies emerge. Integrating new methods or tools to help identify proxies will be important to the industry ..."* [58]. Despite its promises, the use of AI to expand human analytic capacity can also create new forms of labor, some of which can be institutionalized (e.g., humans in charge of determining the veracity of LLM results in the face of potential hallucinations [94]). However, with the introduction of complex and often obscure LLMs—embedding built-in chains of prompts and assuming categories for dataset attributes [73]—many additional forms of labor might remain hidden, complicating users' work in unanticipated ways.

5.4 The Blurry Views of Dataset Documentation Tools' Broader Integration

Our analysis of tools in our corpus sheds light on a tendency to discuss the work of integrating documentation tools within socio-technical systems in blurry terms, often sustained on assumptions over what users would/could do with the tools, and therefore placed in a future over which authors cannot offer more than recommendations. Exceptions to this include [108, 4], who included concrete examples of integration in connection to the IRB system and the ML conference infrastructure, respectively. We also found that the discourses that creators had about what successful integration might entail and the assumptions that feed these discourses, were not always grounded in empirical data. With this analysis, we echo the results of Heger et al. study with data practitioners that foregrounded the need for "processes by which data documentation is created to be integrated with the databases, cloud platforms, and analysis tools that they use" [5].

5.4.1 Discourses about Integration Work: Relying on Unsupported Assumptions

Several creators in our corpus recognized the importance for documentation tools to be fully integrated into concrete contexts, often expressing that *"[robust documentation's] durability is enabled by sufficient support, integration, and flexibility of documentation practices"* [2]. As such, they discussed plans to help users' and adopters' deploy and use the tools they proposed. Some were thorough in the way use cases could take place and, along with advancing a tool, provided users with a conceptualization of the documentation process. Gebru et al., for example, detailed seven steps for the documentation process all the way from the moment a dataset is motivated to when a dataset is released and requires maintenance [1]. While this level of detail can be helpful for future users, it may fall short in capturing the work required to successfully integrate these tools across a large variety of contexts, some of which may fall outside of this characterization, a challenge acknowledged by some authors [109, 1, 58]. As we reported in section 4, this lack of attention to integration was a common phenomenon across a large portion of tools in our sample.

At the other end of the spectrum were creators who conceptualized integration efforts in terms of their perceived benefits. In many cases this conceptualization led creators to prioritize ease of use, technical features, and connection to technical features they believed would be of value to users without acknowledging existing findings in other fields that stressed other complex factors [87, 106, 74, 94, 73]. The tool that [75] created exemplifies this view of integration: *"One of the main goals when designing our tool has been to keep it simple and easy to use. Therefore, the installation process is as simple as searching the plugin in the VSCode extension tab, download and enable it"* [75]. However, this tool assumed practitioners operated within a straightforward data lifecycle model, something that research has already advised against [117, 32, 30].

Another form of working towards integration in terms of perceived benefit was to provide examples of their proposals in use [1, 60, 80, 46, 2, 96], and ready-to-use templates [6]. These efforts mostly focused on outcomes (e.g., exemplifying the resulting artifact of using a given tool), and provided much needed guidance for potential users. Yet, few proposals took concrete steps to facilitate said integration in detail, often disregarding aspects that are critical for realistic integration such as the data governance structures from where datasets emerged [58, 1, 59, 6].

Authors also conceptualized integration efforts in terms of perceived benefits by seeking to extend existing tools into particular fields of knowledge. A common example of this was the extension of Datasheets into the arts [84], healthcare [93], and energy sectors [88], for example. This strategy, other authors argued, could lead to overlaps and/or duplication, adding to the labor and time concerns we mentioned in previous sections [96]. These instances, while concrete and scoped, did not get into details about practical integration.

Overall, our data suggests that perspectives of integration based on perceived benefits were not grounded on empirical data and rather expressed as assumptions: *"... by harmonising the differing perspectives of data scientists and legal experts, proposed data-envelopes serve as a bridge between technical and legal frameworks, facilitating a more ethical and legally compliant use of historical data"* [97]

5.4.2 Authors' Beliefs on the Integration Process

Our analysis also suggest that authors hold at least two critical beliefs about how tools created as part of research can be integrated into realistic use contexts: integration is an iterative, long-term process, with its final success placed in an unknown future and integration takes place organically. In both cases, creators' discourses suggest a belief that taking on the work of integration is someone else responsibility.

The belief that integration will happen at some point in the future often emerged as a recognition of the challenges related to it: *"Since principles are often compelling in theory but challenging to realize in practice, our hope is that by discussing such tradeoffs, CLeAR can serve as a realistic guide for creating documentation while considering and understanding some of the choices that will need to be made"* [38, 6]. This focus on "future integration" relies in the role of "champions", individuals or groups of users who can take on the labor of integrating these tools, departing from non-prescriptive recommendations offered by authors and transforming them into well-established organizational best practices [58, 1]. This matches a related trend of authors referring to integration as an iterative, long-term process, with its final success placed in an unknown future [97, 75, 6]. Some authors did address integration more specifically, and in reference to current structures that could aid in integrating tools: *"To improve 'findability', we urge NeurIPS to require datasets to have metadata (not just data) assigned a persistent identifier and hosted in a searchable repository (such as Zenodo)"* [4, 106]. This kind of effort then paves the way for studies looking at the use of these integrated tools across dataset repositories, for example. Yang et al. study of the use of Data and Model cards across the Hugging Face repository are one recent example.

As we found, authors also framed integration as an organic process assuming enthusiastic users willing to discuss the ideas advanced by authors and turn them into successful practices: *"Responses to documentation questions are not intended to be prescriptive, nor are they completely comprehensive. Instead, they should be considered as one of many* *valid responses to this line of inquiry, and as a way to provoke further thought and discussion."* [95]. Once again, this adds to the labor of documentation often inadvertently placed in users often complicating the adoption of these tools.

6 Discussion

Our exploration highlighted that, with few exceptions (e.g., [115, 111]), most authors in the corpus conceptualized dataset documentation tools in ways that may hinder their adoption and standardization. These views clustered around four themes: (1) a disagreement on how to operationalize the value of documentation; (2) a tendency to design decontextualized tools disconnected from user experiences; (3) an increase of undiscussed labor demands, even when moving towards automation; and (4) a view of tool integration as an aspirational future endeavor. These patterns echo findings from previous work exploring the design and adoption of other forms of Responsible AI (RAI) tools (e.g.,[118, 30, 31, 32, 119]), suggesting a broader, systemic misalignment between how industry and academics understand RAI work. In what follows, we discuss how our findings shed light on this misalignment. Further, we argue that the HCI community interested in supporting RAI initiatives needs to pause the focus on designing tools; continuing to produce documentation tools that offload responsibilities onto practitioners without addressing the industry-academia misalignment can be counterproductive for sustainable documentation practices. Building on the work of [118, 30, 32, 120, 121, 122] and outlier cases in our corpus [106, 110], we call on the HCI community to lead a shift from a design mindset to a research mindset that centers on systemic issues preventing academia from responding to industry's documentation goals. To that end, we propose an HCI research agenda that supports industry and academia in developing sustainable dataset documentation practices.

6.1 Dataset Documentation Tools and the Systemic Industry-Academia Misalignment: The Need to Rethink the Focus on Design

Our findings highlight that corpus authors, mostly academics (n=46) from CS and ML fields, did not work closely with their end-users or tailor to their specific contexts. Prior research shows that industry environments often operate under fast-paced timelines governed by launches of new software products [119, 122], which can hinder companies' ability to share contextual knowledge with academic researchers [123, 124, 125]. Our findings illuminate three key factors shaping the misalignment between academic conceptualizations of documentation and industry needs: a lack of industry-validated consensus on documentation, academia's disengagement with industry, and academia's disconnection with HCI knowledge. Furthermore, we argue that, given academia's response to these factors, it is crucial to pause the development of additional documentation tools. Instead, we must address the systemic factors that currently hinder the standardization and adoption of documentation practices in the industry.

6.1.1 Non-Standardized, Contrasting Views on Documentation: A Lack of Industry-Validated Consensus

As we observed in our corpus, most authors pursued transparency and accountability through non-standardized and sometimes contrasting interpretations. To articulate their perspectives, many drew inspiration from tools and processes in the fields of Nutrition, Engineering, and Software Engineering, among others. While some corpus authors did study their tools' ability to prompt reflection [46, 111], the majority provided little empirical evidence that their tools could achieve their intended goals. This pattern aligns with prior research showing that RAI tools often lack clear definitions of RAI concepts (e.g., ethical values, social impact, equitable representation) [122, 126]. Operating under multiple, non-standardized understandings of ethics, however, can be problematic for practitioners: it forces them to engage in definitional work with no success criteria and the pressure of organizational leadership to prioritize business discourses and interests [122, 126].

The tendency of tools in our corpus—and RAI tools in general—to offer diverse perspectives on critical ethical concepts (e.g., the value of documentation) suggests that, in the absence of industry-provided clarity, academics resort to connections with theoretical concepts from different disciplines without the obligation to ground them in practice. To address this problem, prior research has recommended designing tools that help practitioners learn and embrace non-technical dimensions of RAI work [118, 31] with the goal of bridging disciplinary and organizational divides [118, 32]. However, our findings indicate that, in the case of documentation tools, such recommendations may inadvertently add to practitioners' struggles. If the industry fails to reach consensus on the value of documentation, academia will continue to produce contrasting perspectives, further complicating practitioners' decision-making processes.

6.1.2 Decontextualized Labor Demands: An Academic Disengagement with Industry

Our findings highlight that corpus authors tended not to engage with end users. Despite some exceptional cases, such as [110, 80], in which authors involved key stakeholders in some part of the creation process (n=10) (see Figure 3), most failed to define an end-user profile, and championed a context-agnostic approach to documentation. As a result, tools in our corpus tended to demand new, undiscussed forms of labor from practitioners such as tools' contextualization and the assessment of potential cultural harms. As Heger et al. and Winecoff and Bogen have emphasized, ethical interpretation of dataset work can be particularly hard for practitioners: in contrast with data interpretation work, ethical interpretations require a type of expertise that most technical practitioners do not have and knowledge of contexts they are often not familiar with [5, 127]. While automated documentation tools purported to reduce labor, they could further obscure ethical interpretative labor by driving practitioners to engage in activities where the potential ethical risks might cause a cascading effect and thus be more complex to identify [127] (e.g., deciding which datasets are uninteresting enough to exclude).

Previous work analyzing the design of RAI toolkits found a similar tendency: these toolkits often operated under a "decontextualized approach to ethics" [118] that disregarded the various human and organizational factors shaping RAI work, and hindered adoption by placing intensive contextualization labor on practitioners [118, 30, 31, 32]. The recurrent decision from tools' creators to produce decontextualized RAI tools suggests more pathways to engage with industry experiences may be needed. To redress the negative impact that a "decontextualized approach to ethics" has on ethical interpretation demands, existing research has proposed designing for moments of positive ambiguity [30] that prompt practitioners to realize they need to reflect on ethical issues, discuss them with others, and reach a consensus. Our findings indicate, however, that this design recommendation might be dangerous in the context of documentation tools: if the academics creating tools continue to operate with such distance from industry, they will not be able to identify and evaluate productive moments for positive ambiguity and could, instead, end up creating more undiscussed, decontextualized labor demands.

6.1.3 Blurry Integration Possibilities: An Academic Disconnection with HCI Knowledge

Finally, most of the tools in our corpus described possibilities for tool integration in blurry terms (with some notable exceptions, such as [57, 46, 1]). This further highlights two characteristics of how academia is handling the academia-industry misalignment in the context of documentation tools. First, we observed a tendency among corpus authors to release the initial versions of tools without evaluation evidence or detailed integration guidance. Such a tendency suggests that, in the face of a distance from industry, academics resort to a practice of generating knowledge through small iterations that "somebody"—another researcher or industry actor—might eventually use or further extend. However, as previous work on RAI tools has found, *"[RAI] tools and practices that do not align with practitioners' workflows and organizational incentives may not be used as intended or even used at all"* [128, 129, 130, 131].

Second, our findings suggest that a stronger connection with HCI research on RAI practices could have helped corpus authors propose more realistic integration guidance. As our previous discussion sections stressed, the majority of authors disregarded existing HCI research on RAI practices and repeated trends that have been already found as counterproductive for RAI goals [132, 133]. Further, with a few exceptions (e.g., as [110, 111]), tools did not resort to human-centered research knowledge and methods to enhance technology integration. Against this backdrop, expecting tools' authors to follow design alternatives from HCI researchers guiding integration might be unrealistic. For example, HCI-related research has already suggested that a productive integration pathway is to design for challenging linear views of data lifecycles [117, 32, 30]. Many tools in our corpus, however, assumed practitioners operated within a straightforward data lifecycle model [73, 75, 114, 95].

6.2 Towards Sustainable Dataset Documentation Practices: An HCI Research Agenda

Our findings highlight how the misalignment between industry and academic conceptualizations of documentation tools—mainly coming from CS and ML fields—produces documentation tools that may devalue dataset work, thus hindering the standardization and adoption of documentation practices. To reverse the devaluation of dataset work, Hutchinson et al. called for a radical cultural shift involving the "broader systems and ecologies in which datasets are maintained and used" [6]. We draw inspiration from their work and propose that, to change how documentation tools are conceptualized, we need to shift from a design to a research perspective that identifies strategies for addressing the systemic factors that shape the industry-academia misalignment. Furthermore, we argue that the field of HCI, which has developed multidisciplinary and critical approaches for democratizing the design of technologies, including RAI tools [133, 132, 129, 134], is best positioned to lead this shift. Next, we discuss a research agenda for HCI scholars working towards effective and sustainable documentation practices.

6.2.1 Strategies and Methods for an Industry-Consensuated View of Documentation

As our findings highlight, the lack of industry-validated consensus on the value of documentation has led academia to generate numerous, contrasting proposals for how documentation can be valuable. Further, as we observed, the distance between academics who conceptualize tools and actual industry practice prevents proposals from measuring effectiveness in ways that matter to industry. Reaching industry-validated consensus on ethics-related concepts, however, is not seamless. As [31, 121, 122] have stressed, power dynamics complicate this task: organizational leadership often uses its power to prioritize AI perspectives that align with business directives, which fragments and diverts practitioners' efforts away from AI responsibility and ethics [130, 117]. As such, to prevent the creation of non-standardized, contrasting views of documentation that force practitioners to engage in ethics definitional work, it is critical to identify the best ways to align the interests of all actors around documentation.

The field of HCI can help identify consensus strategies that resist organizational leadership's efforts to "tame" ethics [135, 136, 137]. For example, HCI researchers could unpack the difficulties and successes behind the creation of domain-specific documentation tools like the few ones in our corpus that attempted to propose disciplinary consensus on concepts championed by generic tools. Similarly, HCI researchers could study initiatives where professional or non-profit organizations have worked to standardize and disseminate RAI practices and identify aspects such as stakeholders' interests, resources used during the consensus-building process, unavailable resources, and mechanisms to resist subordinating ethical debates to corporate interests. Documenting these aspects could help RAI researchers, practitioners, and organizations define, with the industry, how documentation could be most effective.

Practitioners working within the industry may benefit from strategies to collectively work towards consensus on RAI concepts while navigating internal power dynamics [122, 126, 31, 121]. While individual professionals often employ useful "soft" resistance tactics to work towards more "values-conscious" ends, as Wong and Ali et al. explained, relying on individuals rather than collective action to promote change can be detrimental for workers, especially to those from marginalized backgrounds and positions of relatively low power within their companies. Workers do not always have "the agency and authority to make or contest values and ethics decisions;"[121] their acts of resistance might result in the company firing the practitioner or re-assigning them to a different task. Further, individual resistance strategies are emotionally demanding and vulnerable to subversion from "the dominant discourses and logics of the technology industry" [121, 135, 30]. The field of HCI can intervene by illuminating methods to support practitioners' coming together in communities of practice with other workers. To minimize the production of tools championing contrasting views on the value of documentation and adding undiscussed forms of labor, HCI could explore collaborative and participatory methods for helping practitioners: (1) explore definitions on documentation and desirable documentation practices, (2) develop strategies to document ethical issues, (3) empirically explore the role of automation in documentation, (4) create industry-valued metrics to demonstrate the effectiveness of documentation, and (5) establish effective practices for communicating and negotiating changes with organizational leadership.

6.2.2 Human-Centered Knowledge and Methods for Conceptualizing Documentation Tools in Industry Contexts

Our findings showed how the majority of the documentation tools in our corpus stemmed from academics in the fields of CS and ML. These academics, as we have already discussed, tended to operate in disconnection with HCI knowledge, both in relation to RAI practices and human-centered methodologies for understanding users and their context. As a result, they leveraged the freedom of academia to experiment with theories, producing context-agnostic tools based on academics' own experiences, and with no clarity on who potentially relevant stakeholders would be. Furthermore, they presented initial iterations of their proposals with vague, aspirational integration mechanisms (e.g., ease of use, publication of templates or examples). To ensure that documentation tools do not add decontextualized forms of labor and factor concrete integration strategies, it is critical that academics develop a closer connection with HCI knowledge.

To achieve this connection, the field of HCI can explore with creators the reasons behind the disconnection, including possible communication aspects hindering the application of these findings into their designs. Further, HCI researchers could work with authors who emphasized the importance of generating evidence and metrics about the impact of documentation tools (e.g., [46, 111] in exploring their experiences, and disseminating this knowledge to other tool creators. The overall goal is to identify barriers and opportunities in ensuring that tool creators can utilize evidence-based research on the value of documentation. Finally, HCI researchers can engage in participatory design activities with tool creators to envision better ways to present, disseminate, and operationalize documentation-related findings, including evidence of documentation effectiveness in supporting RAI goals.

The efforts of some authors in our corpus to approach end-users to understand, design with, or gather feedback, emphasizes the difficulties tool creators may face when reaching out to industry practitioners. As our findings show, some of these authors resorted to online methods to gather practitioners' feedback without compromising their identities (e.g., [1].) These efforts, however, were not common in our corpus, which suggests that tool authors need support not only in learning more about possible HCI methods, but also in considering how to tailor them for the specific needs and level of distance from industry actors. To that end, HCI researchers can work with creators to identify the obstacles they foresee in working closely with practitioners and understanding their workflows. Furthermore, HCI researchers can explore appropriate participatory design methodologies to collaborate with creators in developing user study methodologies that help them navigate these obstacles, enabling the analysis of interest alignment and prioritizing collective power.

7 Limitations

In this study, we take an interpretative approach to data analysis that focuses on the motivations, conceptualizations, and execution of data documentation proposals. We recognize that this decision leaves open the possibility of additional findings focused on other themes, especially descriptive comparisons across tools. We also do not elicit practitioners' views first-hand in the context of this study. Instead, we lean on empirical scholarship in HCI, foregrounding prior work and connecting its results to our findings. Our study focuses on a subset of tools related to dataset documentation, specifically those about which information is publicly available in academic databases and on the open web. As such, private resources, tools or frameworks used by organizations in the private sector are outside of our domain of analysis. We argue that increasing focus on producing empirical evidence of how documentation meaningfully contributes to increasing accountability and transparency, can help practitioners better judge the value of tools for their particular applications. We acknowledge that some of this work intersects with studies around AI auditing and recent studies focused on practitioners workflows; this intersection falls outside of the scope of this paper.

8 Conclusion

Data transparency and accountability remain important foundations for the responsible development of automated systems. Appropriate dataset documentation is a key tenet of these pillars. Despite efforts to facilitate dataset documentation through tools including software applications and frameworks, standardized adoption and integration of these tools remains scarce. To reveal barriers and opportunities towards broader use of these tools, we conducted a scoping review examining the motivations and conceptualizations advanced by designers of dataset documentation tools. Our contributions are two-fold. First, we foreground aspects of these tools significantly impacting their adoption, including a spectrum of perspectives on documentation practices, a devaluing of the dataset infrastructure, an unquestioned transition towards automation, and an aspirational view around the integration of these tools. Second, we propose a radical shift in how to design Responsible AI tools focusing on institutions rather than individuals. We discuss critical actions the HCI community can take to support this shift.

Acknowledgements

We would like to thank Jenny Barry, Maximillian Clee, Jessica Hodgins, and the conference referees for their helpful comments.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. URL `http://arxiv.org/abs/1803.09010`. arXiv:1803.09010 [cs].

[2] Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. URL `http://arxiv.org/abs/2201.03954`.

[3] Elham Tabassi. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Gaithersburg, MD, January 2023. doi:10.6028/NIST.AI.100-1. URL `http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf`.

[4] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. The state of data curation at neurips: An assessment of dataset development practices in the datasets and benchmarks track. *Advances in Neural Information Processing Systems*, 37:53626–53648, December 2024.

[5] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2):340:1–340:29, November 2022. doi:10.1145/3555760. URL `https://dl.acm.org/doi/10.1145/3555760`.

[6] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards Accountability for Machine Learning Datasets: Practices from Software

Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 560–575, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi:10.1145/3442188.3445918. URL https://dl.acm.org/doi/10.1145/3442188.3445918.

[7] Laura Manduchi, Clara Meister, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, et al. On the challenges and opportunities in generative ai. *arXiv preprint arXiv:2403.00025*, 2024.

[8] Jasper Roe and Mike Perkins. 'what they're not telling you about chatgpt': exploring the discourse of ai in uk news media headlines. *Humanities and Social Sciences Communications*, 10(1):753, October 2023. ISSN 2662-9992. doi:10.1057/s41599-023-02282-w.

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi:10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

[10] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021. ISSN 26663899. doi:10.1016/j.patter.2021.100336.

[11] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. Thesis, Massachusetts Institute of Technology, 2017. URL https://dspace.mit.edu/handle/1721.1/114068. Accepted: 2018-03-12T19:28:30Z.

[12] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[13] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 1668–1678, Florence, Italy, july 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163/.

[14] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. November 2017. doi:10.48550/arXiv.1711.08536. URL https://ui.adsabs.harvard.edu/abs/2017arXiv171108536S. ADS Bibcode: 2017arXiv171108536S.

[15] Alexandra Luccioni and Joseph Viviano. What's in the box? an analysis of undesirable content in the common crawl corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, page 182–189, Online, August 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-short.24. URL https://aclanthology.org/2021.acl-short.24/.

[16] David Thiel. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory, Cyber Policy Center, December*, 23:3, 2023.

[17] Rachel Hong, Jevan Hutson, William Agnew, Imaad Huda, Tadayoshi Kohno, and Jamie Morgenstern. A common pool of privacy problems: Legal and technical lessons from a large-scale web-scraped machine learning dataset. (arXiv:2506.17185), june 2025. doi:10.48550/arXiv.2506.17185. URL http://arxiv.org/abs/2506.17185. arXiv:2506.17185 [cs].

[18] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8):975–987, August 2024. ISSN 2522-5839. doi:10.1038/s42256-024-00878-8.

[19] Heila Precel, Allison McDonald, Brent Hecht, and Nicholas Vincent. A canary in the ai coal mine: American jews may be disproportionately harmed by intellectual property dispossession in large language model training. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–17, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi:10.1145/3613904.3642749. URL https://dl.acm.org/doi/10.1145/3613904.3642749.

[20] Mary L. Gray and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019. ISBN 978-1-328-56624-9.

[21] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, December 2019. ISBN 978-0-300-23502-9. doi:10.12987/9780300235029. URL `https://www.degruyter.com/document/doi/10.12987/9780300235029/html`.

[22] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Fairness in representation: 19th siam international conference on data mining, sdm 2019. *SIAM International Conference on Data Mining, SDM 2019*, page 801–809, 2019. doi:10.1137/1.9781611975673.90.

[23] Kate Crawford. The trouble with bias, December 2017. URL `https://www.youtube.com/watch?v=fMym_BKWQzk`.

[24] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! (arXiv:2310.03693), October 2023. doi:10.48550/arXiv.2310.03693. URL `http://arxiv.org/abs/2310.03693`.

[25] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? (arXiv:2204.07931), April 2022. doi:10.48550/arXiv.2204.07931. URL `http://arxiv.org/abs/2204.07931`.

[26] Apostol Vassilev. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. Number NIST AI 100-2e2025. Gaithersburg, MD, 2025. doi:10.6028/NIST.AI.100-2e2025. URL `https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf`.

[27] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. (arXiv:2303.10130), August 2023. doi:10.48550/arXiv.2303.10130. URL `http://arxiv.org/abs/2303.10130`.

[28] Will Orr and Kate Crawford. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. 26:4955–4972, sept 2024. ISSN 1461-4448. doi:10.1177/14614448241251797.

[29] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–16, May 2019. doi:10.1145/3290605.3300830. URL `http://arxiv.org/abs/1812.05239`.

[30] Michael A. Madaio, Jingya Chen, Hanna Wallach, and Jennifer Wortman Vaughan. Tinker, tailor, configure, customize: The articulation work of contextualizing an ai fairness checklist. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–20, April 2024. ISSN 2573-0142. doi:10.1145/3653705.

[31] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. "fairness toolkits, a checkbox culture?" on the factors that fragment developer practices in handling algorithmic harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, page 482–495, Montréal QC Canada, August 2023. ACM. ISBN 979-8-4007-0231-0. doi:10.1145/3600211.3604674. URL `https://dl.acm.org/doi/10.1145/3600211.3604674`.

[32] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness Accountability and Transparency*, page 473–484, june 2022. doi:10.1145/3531146.3533113. URL `http://arxiv.org/abs/2205.06922`.

[33] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, page 1–13, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi:10.1145/3411764.3445261. URL `https://dl.acm.org/doi/10.1145/3411764.3445261`.

[34] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. The model card authoring toolkit: Toward community-centered, deliberation-driven ai design. In *2022 ACM Conference on Fairness Accountability and Transparency*, page 440–451, Seoul Republic of Korea, june 2022. ACM. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533110. URL `https://dl.acm.org/doi/10.1145/3531146.3533110`.

[35] Erich Prem. From ethical ai frameworks to tools: a review of approaches. 3:699–716, August 2023. ISSN 2730-5961. doi:10.1007/s43681-023-00258-9.

[36] Robert Cinca, Enrico Costanza, and Mirco Musolesi. Practitioners and Bias in Machine Learning: A Study. *ACM Trans. Interact. Intell. Syst.*, 15(2):12:1–12:28, June 2025. ISSN 2160-6455. doi:10.1145/3733838. URL https://dl.acm.org/doi/10.1145/3733838.

[37] Samia Kabir, Lixiang Li, and Tianyi Zhang. STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–20, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi:10.1145/3613904.3642111. URL https://dl.acm.org/doi/10.1145/3613904.3642111.

[38] Kasia Chmielinski, Sarah Newman, Chris N. Kranzinger, Michael Hind, Jennifer Wortman Vaughan, Margaret Mitchell, Julia Stoyanovich, Angelina McMillan-Major, Emily McReynolds, Kathleen Esfahany, Mary L. Gray, Maui Hudson, and Audrey Chang, May 2024.

[39] Global Future Council on Human Rights. How to prevent discriminatory outcomes in machine learning, March 2018. URL https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.

[40] Tanja Šarčević, Alicja Karlowicz, Rudolf Mayer, Ricardo Baeza-Yates, and Andreas Rauber. U can't gen this? a survey of intellectual property protection methods for data in generative ai. (arXiv:2406.15386), April 2024. doi:10.48550/arXiv.2406.15386. URL http://arxiv.org/abs/2406.15386.

[41] Nikhil Kandpal and Colin Raffel. Position: The most expensive part of an llm should be its training data. (arXiv:2504.12427), April 2025. doi:10.48550/arXiv.2504.12427. URL http://arxiv.org/abs/2504.12427.

[42] Executive Office of the President. Safe, secure, and trustworthy development and use of artificial intelligence. *https://www. federalregister. gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence*, 2023.

[43] The White House. America's ai action plan. july 2025. URL https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

[44] Number AB 2013. sept 2024. URL https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013.

[45] Number SB24-205. May 2024. URL https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf.

[46] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA, june 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533231. URL https://dl.acm.org/doi/10.1145/3531146.3533231.

[47] Andersen v. Stability AI Ltd., 3:23-cv-00201, (N.D. Cal.). URL https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/.

[48] Concord Music Group, Inc. v. Anthropic PBC, 3:23-cv-01092, (M.D. Tenn.). URL https://www.courtlistener.com/docket/67894459/concord-music-group-inc-v-anthropic-pbc/.

[49] Dow Jones & Company, Inc. v. Perplexity AI, Inc., 1:24-cv-07984, (S.D.N.Y.). URL https://www.courtlistener.com/docket/69280523/dow-jones-company-inc-v-perplexity-ai-inc/.

[50] In Re: OpenAI, Inc. Copyright Infringement Litigation, 1:25-md-03143, (S.D.N.Y.). URL https://www.courtlistener.com/docket/69879510/in-re-openai-inc-copyright-infringement-litigation/.

[51] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

[52] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness Accountability and Transparency*, page 778–788, Seoul Republic of Korea, june 2022. ACM. doi:10.1145/3531146.3533143. URL https://dl.acm.org/doi/10.1145/3531146.3533143.

[53] NeurIPS Code of Ethics. URL https://nips.cc/public/EthicsGuidelines.

[54] Hugging Face Dataset Cards. URL https://huggingface.co/docs/hub/datasets-cards.

[55] Microsoft. Microsoft rai impact assessment template, june 2022. URL https://msblogs.thesourcemediaassets.com/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf.

[56] Microsoft. Aether data documentation template, August 2022. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2022/07/aether-datadoc-082522.pdf.

[57] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5): 6:1–6:13, july 2019. ISSN 0018-8646. doi:10.1147/JRD.2019.2942288.

[58] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards, May 2018. URL http://arxiv.org/abs/1805.03677. arXiv:1805.03677 [cs].

[59] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6: 587–604, December 2018. ISSN 2307-387X. doi:10.1162/tacl_a_00041. URL https://doi.org/10.1162/tacl_a_00041.

[60] Angelina McMillan-Major and Emily M. Bender. Data Statements | Tech Policy Lab. Technical report, University of Washington, 2023. URL https://techpolicylab.uw.edu/data-statements/.

[61] Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/077e29b11be80ab57e1a2ecabb7da330-Paper-round2.pdf.

[62] Jodi Schneider, Di Ye, Alison M. Hill, and Ashley S. Whitehorn. Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics*, 125(3): 2877–2913, December 2020. ISSN 1588-2861. doi:10.1007/s11192-020-03631-1.

[63] Jack Bandy and Nicholas Vincent. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. (arXiv:2105.05241), May 2021. doi:10.48550/arXiv.2105.05241. URL http://arxiv.org/abs/2105.05241.

[64] The common crawl. URL https://commoncrawl.org/overview.

[65] Adriana Alvarado Garcia, Heloisa Candello, Karla Badillo-Urquiola, and Marisol Wong-Villacres. Emerging data practices: Data work in the era of large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, page 1–21, Yokohama Japan, April 2025. ACM. ISBN 979-8-4007-1394-1. doi:10.1145/3706598.3714069. URL https://dl.acm.org/doi/10.1145/3706598.3714069.

[66] Xinyu Yang, Weixin Liang, and James Zou. Navigating dataset documentations in ai: A large-scale analysis of dataset cards on huggingface. October 2023. URL https://openreview.net/forum?id=xC8xh2RSs2.

[67] Hilary Arksey and Lisa O'Malley. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1):19–32, February 2005. ISSN 1364-5579, 1464-5300. doi:10.1080/1364557032000119616. URL http://www.tandfonline.com/doi/abs/10.1080/1364557032000119616.

[68] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D.J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, 169(7):467–473, October 2018. ISSN 0003-4819, 1539-3704. doi:10.7326/M18-0850. URL https://www.acpjournals.org/doi/10.7326/M18-0850.

[69] Virginia Braun and Victoria Clarke. *Thematic analysis: a practical guide*. SAGE, London ; Thousand Oaks, California, 2022. ISBN 978-1-4739-5323-9 978-1-4739-5324-6. OCLC: on1247204005.

[70] Guglielmo De Angelis and Francesca Lonetti. About the Assessment of Grey Literature in Software Engineering. In *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*, EASE '21, pages 373–378, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-9053-8. doi:10.1145/3463274.3463362. URL https://dl.acm.org/doi/10.1145/3463274.3463362.

[71] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. A domain-specific language for describing machine learning datasets. *Journal of Computer Languages*, 76:101209, 2023. ISSN 2590-1184. doi:https://doi.org/10.1016/j.cola.2023.101209. URL https://www.sciencedirect.com/science/article/pii/S2590118423000199.

[72] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, FAccT '22, pages 1350–1361, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533192. URL `https://doi.org/10.1145/3531146.3533192`. Number of pages: 12 Place: Seoul, Republic of Korea.

[73] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. DataDoc analyzer: a tool for analyzing the documentation of scientific datasets. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, Cikm '23, pages 5046–5050, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8-4007-0124-5. doi:10.1145/3583780.3614737. URL `https://doi.org/10.1145/3583780.3614737`.

[74] Anne Helby Petersen and Claus Thorn Ekstrøm. dataMaid: Your Assistant for Documenting Supervised Data Quality Screening in R. *Journal of Statistical Software*, 90:1–38, July 2019. ISSN 1548-7660. doi:10.18637/jss.v090.i06.

[75] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. DescribeML: a tool for describing machine learning datasets. In *Proceedings of the 25th international conference on model driven engineering languages and systems: Companion proceedings*, Models '22, pages 22–26, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9467-3. doi:10.1145/3550356.3559087. URL `https://doi.org/10.1145/3550356.3559087`. Number of pages: 5 Place: Montreal, Quebec, Canada.

[76] Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. Goods: Organizing google's datasets. In *Proceedings of the 2016 international conference on management of data*, Sigmod '16, pages 795–806, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-3531-7. doi:10.1145/2882903.2903730. URL `https://doi.org/10.1145/2882903.2903730`.

[77] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. MithraLabel: Flexible Dataset Nutritional Labels for Responsible Data Science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 2893–2896. Association for Computing Machinery, 2019. ISBN 978-1-4503-6976-3. doi:10.1145/3357384.3357853. URL `https://dl.acm.org/doi/10.1145/3357384.3357853`.

[78] Buse Gul Atli Tekgul and N. Asokan. On the effectiveness of dataset watermarking. In *Proceedings of the 2022 ACM on international workshop on security and privacy analytics*, Iwspa '22, pages 93–99, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9230-3. doi:10.1145/3510548.3519376. URL `https://doi.org/10.1145/3510548.3519376`.

[79] Jan Novacek, Ali Ahari, Tobias Müller, Sebastian Reiter, Alexander Viehl, and Oliver Bringmann. Ontology-Supported AI Model and Dataset Management. In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, pages 1–6, August 2024. doi:10.1109/INDIN58382.2024.10774524. URL `https://ieeexplore.ieee.org/abstract/document/10774524`. ISSN: 2378-363X.

[80] Anthony Cintron Roman, Jennifer Wortman Vaughan, Valerie See, Steph Ballard, Jehu Torres, Caleb Robinson, and Juan M. Lavista Ferres. Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments, 2024. URL `http://arxiv.org/abs/2312.06153`.

[81] Victor Alencar, Troy Kohwalter, Vanessa Braganholo, José Ricardo da Silva, and Leonardo Murta. Prov-Dominoes: An approach for knowledge discovery from provenance data. *Expert Systems with Applications*, 245:123030, 2024. ISSN 0957-4174. doi:https://doi.org/10.1016/j.eswa.2023.123030. URL `https://www.sciencedirect.com/science/article/pii/S0957417423035327`.

[82] Ruben C Arslan. How to automatically document data with the codebook package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science*, 2(2):169–187, 2019. doi:10.1177/2515245919838783.

[83] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. A Methodology for Creating AI FactSheets, June 2020. URL `http://arxiv.org/abs/2006.13796`. arXiv:2006.13796 [cs].

[84] Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. Artsheets for Art Datasets. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/9b8619251a19057cff70779273e95aa6-Abstract-round2.html`.

[85] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, FAccT '23, pages 881–904, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8-4007-0192-4.

doi:10.1145/3593013.3594049. URL https://doi.org/10.1145/3593013.3594049. numPages: 24 place: Chicago, IL, USA.

[86] Nabajeet Barman, Yuriy Reznik, and Maria Martini. Datasheet for subjective and objective quality assessment datasets. In *2023 15th international conference on quality of multimedia experience (QoMEX)*, pages 111–114, June 2023. doi:10.1109/QoMEX58391.2023.10178546. ISSN: 2472-7814.

[87] Ramtin Zargari Marandi, Anne Svane Frahm, and Maja Milojevic. Datasheets for AI and medical datasets (DAIMS): a data validation and documentation framework before machine learning analysis in medical research, January 2025. URL http://arxiv.org/abs/2501.14094. arXiv:2501.14094 [cs].

[88] Ilana Heintz. Datasheets for Energy Datasets: An Ethically-Minded Approach to Documentation. In *Companion Proceedings of the 14th ACM International Conference on Future Energy Systems*, e-Energy '23 Companion, pages 40–51, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0227-3. doi:10.1145/3599733.3600249. URL https://dl.acm.org/doi/10.1145/3599733.3600249.

[89] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1943–1961. ACM, 2022. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533239. URL https://dl.acm.org/doi/10.1145/3531146.3533239.

[90] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. Method Cards for Prescriptive Machine-Learning Transparency. In *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, CAIN '22, pages 90–100, New York, NY, USA, May 2022. Association for Computing Machinery. ISBN 978-1-4503-9275-4. doi:10.1145/3522664.3528600. URL https://ieeexplore.ieee.org/document/9796452.

[91] Xinyi Zheng, Ryan A. Rossi, Nesreen K. Ahmed, and Dominik Moritz. Network report: a structured description for network datasets. In *Proceedings of the 31st ACM international conference on information &amp; knowledge management*, Cikm '22, pages 3694–3704, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9236-5. doi:10.1145/3511808.3557115. URL https://doi.org/10.1145/3511808.3557115. Number of pages: 11 Place: Atlanta, GA, USA.

[92] Surbhi Mittal, Kartik Thakral, Richa Singh, Mayank Vatsa, Tamar Glaser, Cristian Canton Ferrer, and Tal Hassner. On Responsible Machine Learning Datasets with Fairness, Privacy, and Regulatory Norms, August 2024. URL http://arxiv.org/abs/2310.15848. arXiv:2310.15848 [cs].

[93] Marjia Siddik and Harshvardhan J. Pandit. Datasheets for Healthcare AI: A Framework for Transparency and Bias Mitigation, January 2025. URL http://arxiv.org/abs/2501.05617. arXiv:2501.05617 [cs].

[94] Daniel C. Fox, Aamod Khatiwada, and Roee Shraga. A generative benchmark creation framework for detecting common data table versions. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, Cikm '24, pages 5365–5369, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0436-9. doi:10.1145/3627673.3679157. URL https://doi.org/10.1145/3627673.3679157. Number of pages: 5 Place: Boise, ID, USA.

[95] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3534647. URL https://dl.acm.org/doi/10.1145/3531146.3534647. tex.ids= diazCrowdWorkSheetsAccountingIndividual2022a arXiv: 2206.08931 [cs].

[96] Shazia Afzal, C Rajmohan, Manish Kesarwani, Sameep Mehta, and Hima Patel. Data Readiness Report. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pages 42–51, September 2021. doi:10.1109/SMDS53860.2021.00016. URL https://ieeexplore.ieee.org/abstract/document/9592479.

[97] Mrinalini Luthra and Maria Eskevich. Data-Envelopes for Cultural Heritage: Going beyond Datasheets. In Ingo Siegert and Khalid Choukri, editors, *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 52–65, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.legal-1.9/.

[98] Graeme Horsman and James R. Lyle. Dataset construction challenges for digital forensics. *Forensic Science International: Digital Investigation*, 38:301264, 2021. ISSN 2666-2817.

doi:https://doi.org/10.1016/j.fsidi.2021.301264. URL https://www.sciencedirect.com/science/article/pii/S2666281721001815.

[99] S. Picard, C. Chapdelaine, C. Cappi, L. Gardes, E. Jenn, B. Lefevre, and T. Soumarmon. Ensuring Dataset Quality for Machine Learning Certification. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 275–282, 2020. doi:10.1109/ISSREW51248.2020.00085. URL https://ieeexplore.ieee.org/abstract/document/9307671.

[100] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. Machine learning data practices through a data curation lens: An evaluation framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1055–1067. Association for Computing Machinery, 2024. ISBN 979-8-4007-0450-5. doi:10.1145/3630106.3658955. URL https://doi.org/10.1145/3630106.3658955.

[101] Joseph E Alderman, Joanne Palmer, Elinor Laws, Melissa D McCradden, Johan Ordish, Marzyeh Ghassemi, Stephen R Pfohl, Negar Rostamzadeh, Heather Cole-Lewis, Ben Glocker, Melanie Calvert, Tom J Pollard, Jaspret Gill, Jacqui Gath, Adewale Adebajo, Jude Beng, Cassandra H Leung, Stephanie Kuku, Lesley-Anne Farmer, Rubeta N Matin, Bilal A Mateen, Francis McKay, Katherine Heller, Alan Karthikesalingam, Darren Treanor, Maxine Mackintosh, Lauren Oakden-Rayner, Russell Pearson, Arjun K Manrai, Puja Myles, Judit Kumuthini, Zoher Kapacee, Neil J Sebire, Lama H Nazer, Jarrel Seah, Ashley Akbari, Lew Berman, Judy W Gichoya, Lorenzo Righetto, Diana Samuel, William Wasswa, Maria Charalambides, Anmol Arora, Sameer Pujari, Charlotte Summers, Elizabeth Sapey, Sharon Wilkinson, Vishal Thakker, Alastair Denniston, and Xiaoxuan Liu. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *The Lancet Digital Health*, 7(1):e64–e88, 2025. ISSN 2589-7500. doi:https://doi.org/10.1016/S2589-7500(24)00224-3. URL https://www.sciencedirect.com/science/article/pii/S2589750024002243.

[102] Laurens A. Castelijns, Yuri Maas, and Joaquin Vanschoren. The ABC of Data: A Classifying Framework for Data Readiness. In Peggy Cellier and Kurt Driessens, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 3–16, Cham, 2020. Springer International Publishing. ISBN 978-3-030-43823-4. doi:10.1007/978-3-030-43823-4_1.

[103] Nitisha Jain, Mubashara Akhtar, Joan Giner-Miguelez, Rajat Shinde, Joaquin Vanschoren, Steffen Vogler, Sujata Goswami, Yuhan Rao, Tim Santos, Luis Oala, Michalis Karamousadakis, Manil Maskey, Pierre Marcenac, Costanza Conforti, Michael Kuchnik, Lora Aroyo, Omar Benjelloun, and Elena Simperl. A Standardized Machine-readable Dataset Documentation Format for Responsible AI, June 2024.

[104] Leon J. Osterweil, Lori A. Clarke, Aaron M. Ellison, Emery Boose, Rodion Podorozhny, and Alexander Wise. Clear and precise specification of ecological data management processes and dataset provenance. *IEEE Transactions on Automation Science and Engineering*, 7(1):189–195, January 2010. ISSN 1558-3783. doi:10.1109/TASE.2009.2021774.

[105] Marco Rondina, Antonio Vetrò, and Juan Carlos De Martin. Completeness of Datasets Documentation on ML/AI Repositories: An Empirical Investigation. In Nuno Moniz, Zita Vale, José Cascalho, Catarina Silva, and Raquel Sebastião, editors, *Progress in Artificial Intelligence*, pages 79–91. Springer Nature Switzerland, 2023. ISBN 978-3-031-49008-8.

[106] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. Croissant: A Metadata Format for ML-Ready Datasets. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, DEEM '24, pages 1–6, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0611-0. doi:10.1145/3650203.3663326.

[107] Raia Abu Ahmad, Jennifer D'Souza, Matthäus Zloch, Wolfgang Otto, Georg Rehm, Allard Oelen, Stefan Dietze, and Sören Auer. Toward FAIR Semantic Publishing of Research Dataset Metadata in the Open Research Knowledge Graph, April 2024. URL http://arxiv.org/abs/2404.08443. arXiv:2404.08443 [cs].

[108] Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. Data Statements: From Technical Concept to Community Practice. *ACM J. Responsib. Comput.*, 1(1):1:1–1:17, March 2024. doi:10.1145/3594737. URL https://dl.acm.org/doi/10.1145/3594737.

[109] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM*

2021), pages 121–135, 2021. doi:10.18653/v1/2021.gem-1.11. URL http://arxiv.org/abs/2108.07374. arXiv:2108.07374 [cs].

[110] David Piorkowski, John Richards, and Michael Hind. A Field Study of a Human-Centered Process for Increasing AI Transparency, 2024. URL http://arxiv.org/abs/2201.13224.

[111] Karen L. Boyd. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):438:1–438:27, October 2021. doi:10.1145/3479582. URL https://dl.acm.org/doi/10.1145/3479582.

[112] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 427–439. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533108. URL https://dl.acm.org/doi/10.1145/3531146.3533108.

[113] Kathy Reid and Elizabeth T. Williams. Right the docs: Characterising voice dataset documentation practices used in machine learning. In Smaranda Muresan, Vivian Chen, Kennington Casey, Vandyke David, Dethlefs Nina, Inoue Koji, Ekstedt Erik, and Ultes Stefan, editors, *Proceedings of the 21st annual workshop of the australasian language technology association*, pages 51–66, Melbourne, Australia, November 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.alta-1.6/.

[114] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 1–13, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9477-2. doi:10.1145/3551624.3555286. URL https://dl.acm.org/doi/10.1145/3551624.3555286.

[115] Angelina Yvonne McMillan-Major. Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities, 2023. URL http://hdl.handle.net/1773/50854.

[116] Virginia Braun and Victoria Clarke. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, July 2021. ISSN 1478-0887, 1478-0895. doi:10.1080/14780887.2020.1769238. URL https://www.tandfonline.com/doi/full/10.1080/14780887.2020.1769238.

[117] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. Investigating practices and opportunities for cross-functional collaboration around ai fairness in industry practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 705–716, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi:10.1145/3593013.3594037. URL https://doi.org/10.1145/3593013.3594037.

[118] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. Seeing like a toolkit: How toolkits envision the work of ai ethics. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–27, April 2023. ISSN 2573-0142. doi:10.1145/3579621.

[119] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), April 2022. doi:10.1145/3512899. URL https://doi.org/10.1145/3512899.

[120] Vera Khovanskaya and Phoebe Sengers. Data rhetoric and uneasy alliances: Data advocacy in us labor history. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 1391–1403, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450358507. doi:10.1145/3322276.3323691. URL https://doi.org/10.1145/3322276.3323691.

[121] Richmond Y. Wong. Tactics of soft resistance in user experience professionals' values work. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021. doi:10.1145/3479499. URL https://doi.org/10.1145/3479499.

[122] Sanna J. Ali, Angèle Christin, Andrew Smart, and Riitta Katila. Walking the walk of ai ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 217–226, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi:10.1145/3593013.3593990. URL https://doi.org/10.1145/3593013.3593990.

[123] Nur Ahmed and Neil C. Thompson. What should be done about the growing influence of industry in ai research?, Dec 2023. URL https://www.brookings.edu/articles/what-should-be-done-about-the-growing-influence-of-industry-in-ai-research.

[124] Brian Eastwood. Study: Industry now dominates ai research, May 2023. URL https://mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research.

[125] Nur Ahmed, Muntasir Wahed, and Neil C. Thompson. The growing influence of industry in ai research. *Science*, 379(6635):884–886, 2023. doi:10.1126/science.ade2420. URL https://www.science.org/doi/abs/10.1126/science.ade2420.

[126] Vyacheslav Polonski. The hard problem of ai ethics—three guidelines for building morality into machines. https://www.oecd-forum.org/posts/30743-the-hard-problem-of-ai-ethics-three-guidelines-for-building-morality-into-machines, 2018. OECD Forum.

[127] Amy Winecoff and Miranda Bogen. Improving governance outcomes through ai documentation: Bridging theory and practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi:10.1145/3706598.3713814. URL https://doi.org/10.1145/3706598.3713814.

[128] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi:10.1145/3411764.3445261. URL https://doi.org/10.1145/3411764.3445261.

[129] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi:10.1145/3313831.3376445. URL https://doi.org/10.1145/3313831.3376445.

[130] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. doi:10.1145/3449081. URL https://doi.org/10.1145/3449081.

[131] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi:10.1145/3411764.3445604. URL https://doi.org/10.1145/3411764.3445604.

[132] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi:10.1145/3544548.3581278. URL https://doi.org/10.1145/3544548.3581278.

[133] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. Guidelines for integrating value sensitive design in responsible ai toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi:10.1145/3613904.3642810. URL https://doi.org/10.1145/3613904.3642810.

[134] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–18, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. doi:10.1145/3544548.3581026. URL https://dl.acm.org/doi/10.1145/3544548.3581026.

[135] Jacob Metcalf, Emanuel Moss, and Danah Boyd. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly*, 86(2):449–476, 2019. doi:10.1353/sor.2019.0022. URL https://muse.jhu.edu/article/732185.

[136] Thao Phan, Jake Goldenfein, Monique Mann, and Declan Kuch. Economies of virtue: The circulation of 'ethics' Big Tech. *Science as Culture*, 31(1):121–135, 2021. doi:10.1080/09505431.2021.1990875. URL https://doi.org/10.1080/09505431.2021.1990875.

[137] Thao Phan, Jake Goldenfein, Declan Kuch, and Monique Mann, editors. *Economies of Virtue: The Circulation of 'Ethics' in AI*, volume 46 of *Theory on Demand*. Institute of Network Cultures, Amsterdam, 2022. ISBN 9789492302960. URL https://networkcultures.org/blog/publication/economies-of-virtue-the-circulation-of-ethics-in-ai/.

[138] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. Using Large Language Models to Enrich the Documentation of Datasets for Machine Learning, May 2024. URL http://arxiv.org/abs/2404.15320. arXiv:2404.15320 [cs].

A Further Descriptive Analyses

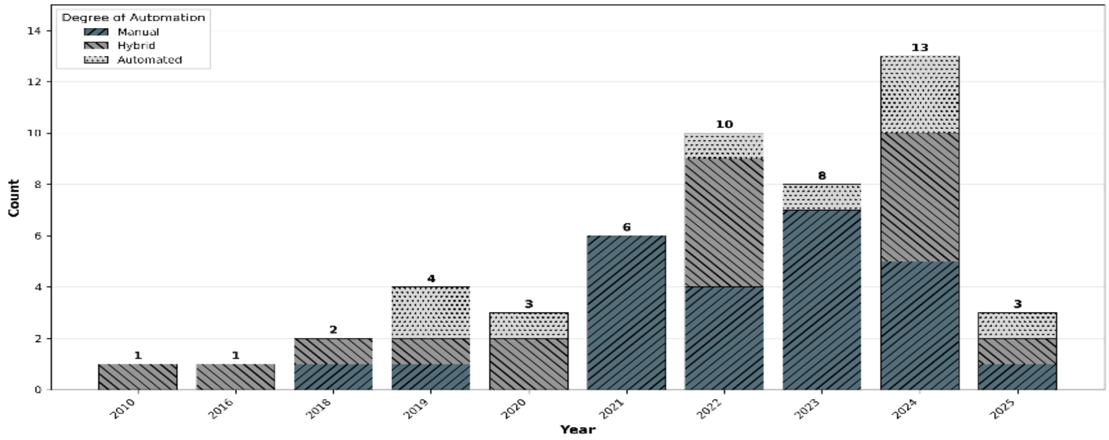

Figure 4: Distribution of approaches to the construction of dataset documentation over the years. Data shows a prevalence of manual approaches with a sustained increase in automated and semi-automated approaches between 2022 and 2025.

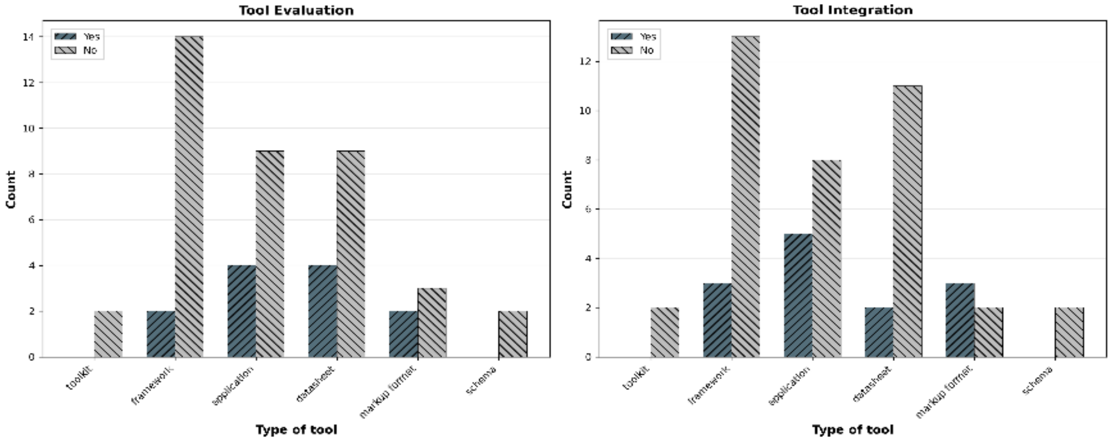

Figure 5: Comparative analysis of different types of tools that included tool evaluation or features towards integration. **Left:** Count of proposals that included an evaluation study at any stage of the design or integration of the tool. **Right:** Percentage of proposals that included stakeholders during the design/testing stages compared to the number of proposals that included concrete features or guidance towards integration to stakeholders.

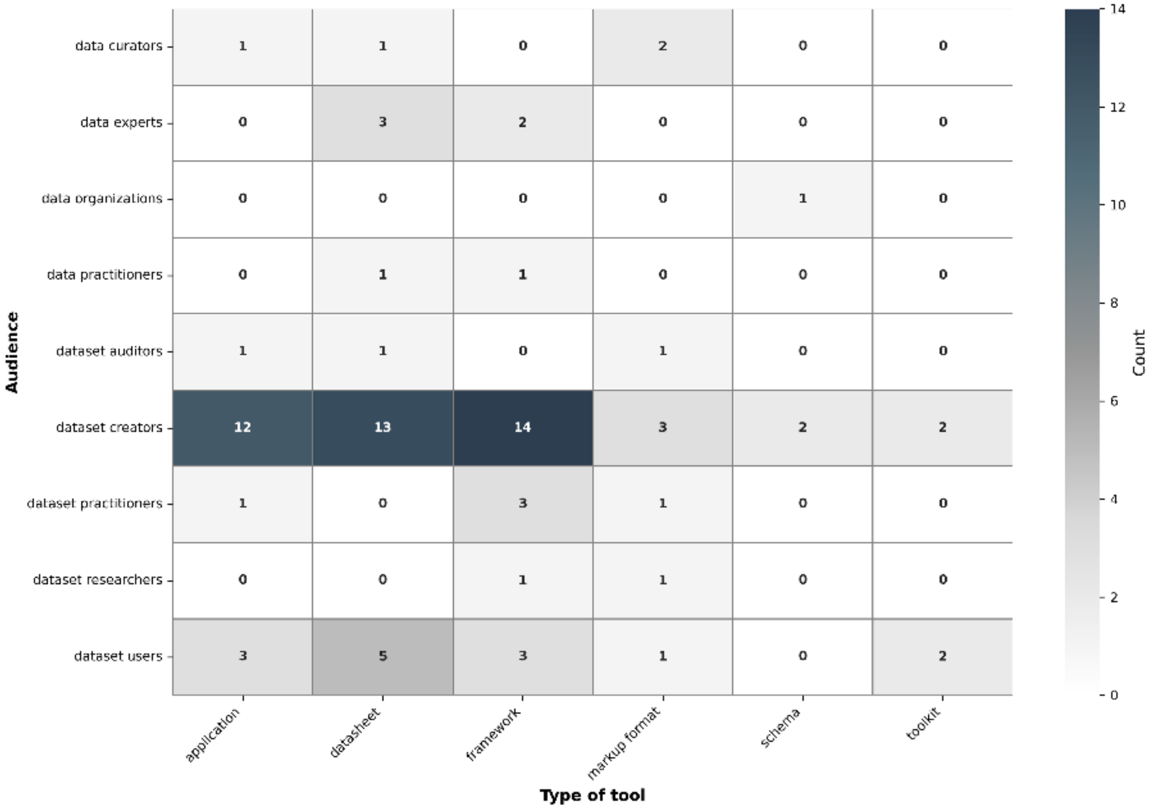

Figure 6: Overlap of different types of tools across different audiences as stated by authors. Data suggests 'dataset creators' as a main category of interest for authors, and 'applications, datasheets and frameworks, as the most preferred categories of tools.

B Data sourcing and screening process

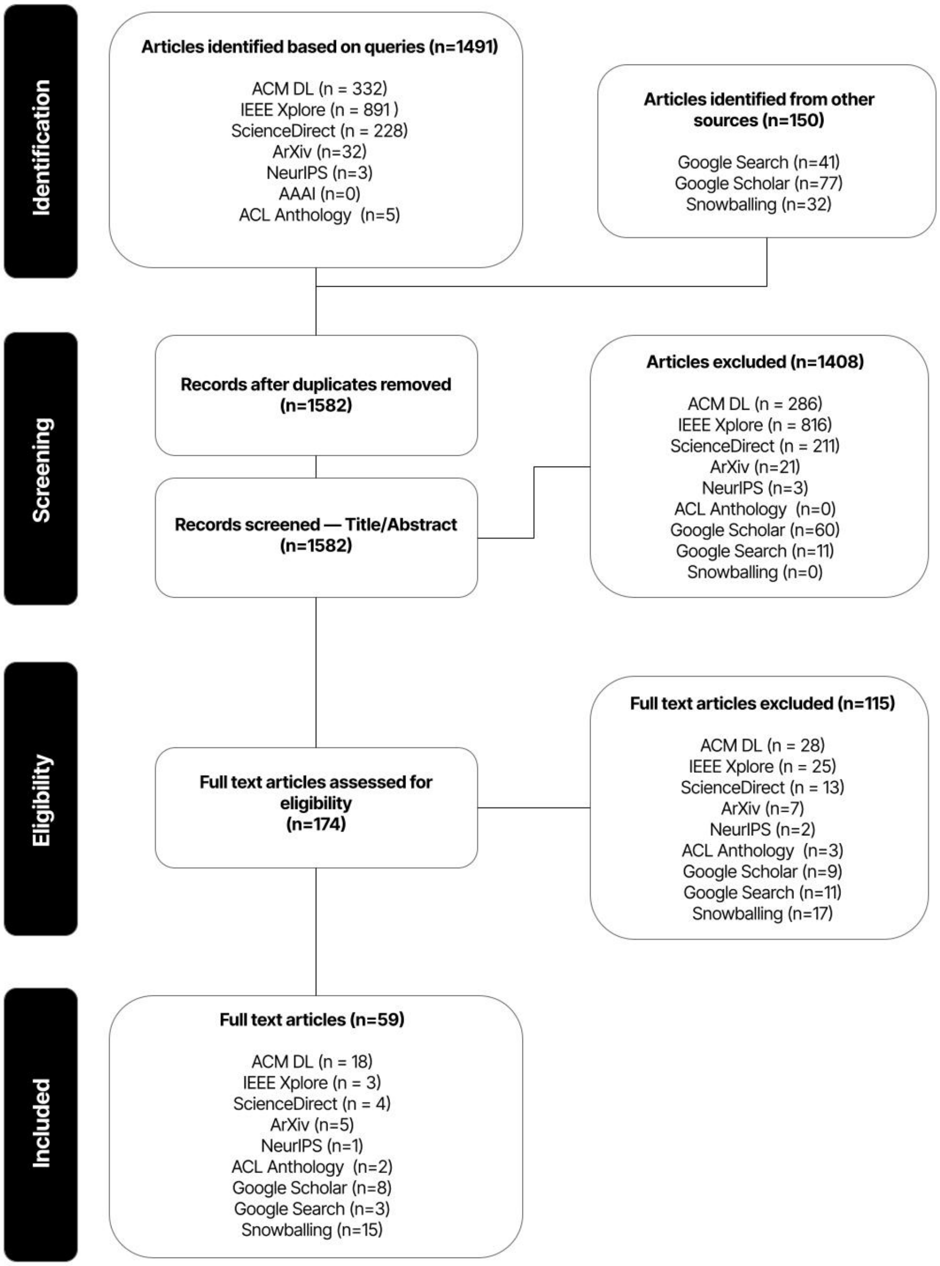

Figure 7: Adapted flowchart from PRISMA outline, describing demonstrating how we searched, screened, and selected items for inclusion in our corpus

C Data collection search queries

Keywords Queries				
Resource	**Query**	**Date**	**Field**	**Notes**
NeurIPS	"dataset documentation" OR "dataset transparency" OR "dataset accountability" OR "dataset provenance" OR "dataset datasheet" OR “dataset management”	Feb 1st	All fields search and manual search over Dataset and Benchmarks repositories	This track has contributions from 2021-2024
ACM	Title:(“dataset” AND (“documentation” OR “transparency” OR “accountability” OR “provenance” OR “datasheet” OR “management”)) OR Abstract:(“dataset” AND (“documentation” OR “transparency” OR “accountability” OR “provenance” OR “datasheet” OR “management”)) OR Keyword:(“dataset” AND (“documentation” OR “transparency” OR “accountability” OR “provenance” OR “datasheet” OR “management”))	Feb 7th	title-abstract-keywords	
IEEE	(“Document Title”:“dataset” OR “Abstract”:“dataset” OR “Author Keywords”:“dataset”) AND (“Document Title”:“documentation” OR “Document Title”:“transparency” OR “Document Title”:“accountability” OR “Document Title”:“provenance” OR “Document Title”:“datasheet” OR “Document Title”:“management” OR “Abstract”:“documentation” OR “Abstract”:“transparency” OR “Abstract”:“accountability” OR “Abstract”:“provenance” OR “Abstract”:“datasheet” OR “Abstract”:“management” OR “Author Keywords”:“documentation” OR “Author Keywords”:“transparency” OR “Author Keywords”:“accountability” OR “Author Keywords”:“provenance” OR “Author Keywords”:“datasheet” OR “Author Keywords”:“management”)	Feb 7th	title-abstract-keywords	
ScienceDirect	(“dataset documentation”) OR (“dataset transparency”) OR (“dataset accountability”) OR (“dataset provenance”) OR (“dataset datasheet”) OR (“dataset management”)	Feb 7th	all fields search	
ArXiv	order: -announced_date_first; size: 200; classification: Computer Science (cs), Economics (econ), Electrical Engineering and Systems Science (eess), Statistics (stat); include_cross_list: True; terms: AND all=“dataset documentation”; OR all=“dataset transparency”; OR all=“dataset accountability”; OR all=“dataset provenance”; OR all=“dataset datasheet” OR all=“dataset management”	Feb 12th	all fields search	
AAAI	“dataset documentation” OR “dataset transparency” OR “dataset accountability” OR “dataset provenance” OR “dataset datasheet” OR “dataset management”	Feb 18th	all fields search	

ACL Anthology	“dataset documentation” OR “dataset transparency” OR “dataset accountability” OR “dataset provenance” OR “dataset datasheet” OR “dataset management”	Feb 21st	all-fields	Search done on Zotero over full ACL catalog BibTeX file available at https://aclanthology.org/
Google Search	“dataset documentation” OR “dataset transparency” OR “dataset accountability” OR “dataset provenance” OR “dataset datasheet” OR “dataset management”	Feb 24th	advanced search with omitted results included	Used the 5-pages-of-noise or data saturation rule. Done from a Cambridge, MA IP address over incognito mode on Brave browser v1.75.180.
Google Scholar	“dataset documentation” OR “dataset transparency” OR “dataset accountability” OR “dataset provenance” OR “dataset datasheet” OR “dataset management”	Feb 24th	advanced search	Used the 5-pages-of-noise or data saturation rule. Done from a Cambridge, MA IP address over incognito mode on Brave browser v1.75.180.

Table 2: Detailed queries as used for each different resource. ‘Field’ denotes the search fields that were leveraged within each resource in order to return results.

D Dataset corpus description

Corpus Description				
Tool/Item	**Year**	**Audiences addressed**	**Degree of Automation**	**Type of Tool**
A domain-specific language for describing machine learning datasets [71]	2023	dataset creators	Manual	Application
A Field Study of a Human-Centered Process for Increasing AI Transparency [110]	2024	N/A	N/A	Study
A generative benchmark creation framework for detecting common data table versions [94]	2024	dataset creators	Automated	Framework
A Methodology for Creating AI FactSheets [83]	2020	dataset creators, data curators	Hybrid	Datasheet
A Standardized Machine-readable Dataset Documentation Format for Responsible AI [103]	2024	dataset creators, dataset users, dataset auditors	Manual	Markup format
Aether Data Documentation Template [56]	2022	dataset creators, dataset auditors	Manual	Datasheet
Artsheets for Art Datasets [84]	2021	dataset creators, dataset users, dataset users	Manual	Datasheet
Augmented datasheets for speech datasets and ethical decision-making [85]	2023	dataset creators	Manual	Datasheet
Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? [72]	2022	dataset creators, data curators	Hybrid	Application
Clear and precise specification of ecological data management processes and dataset provenance [104]	2010	dataset creators, data curators	Hybrid	Markup format
Completeness of Datasets Documentation on ML/AI Repositories: An Empirical Investigation [105]	2023	dataset creators, data curators	Manual	Markup format
Croissant: A Metadata Format for ML-Ready Datasets [106]	2024	dataset practitioners	Automated	Markup format
CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation [95]	2022	dataset creators	Manual	Framework
Data cards: Purposeful and transparent dataset documentation for responsible AI [46]	2022	dataset creators, dataset users	Hybrid	Framework
Data Readiness Report [96]	2021	dataset practitioners, dataset creators	Manual	Framework
Data Statements | Tech Policy Lab [60]	2023	dataset creators, dataset users	Manual	Toolkit
Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science [59]	2018	dataset creators	Manual	Framework
Data statements: From technical concept to community practice [108]	2024	dataset creators	Manual	Schema
Data-envelopes for cultural heritage: Going beyond datasheets [97]	2024	dataset creators	Manual	Framework

DataDoc analyzer: A tool for analyzing the documentation of scientific datasets [73]	2023	dataset creators, dataset users	Automated	Application
dataMaid: Your Assistant for Documenting Supervised Data Quality Screening in R [74]	2019	dataset creators	Hybrid	Application
Dataset construction challenges for digital forensics [98]	2021	dataset creators, dataset users	Manual	Framework
Datasheet for subjective and objective quality assessment datasets [86]	2023	dataset creators, dataset users	Manual	Datasheet
Datasheets for AI and medical datasets (DAIMS): a data validation and documentation framework before machine learning analysis in medical research [87]	2025	dataset creators, data experts	Hybrid	Datasheet
Datasheets for datasets [1]	2021	dataset creators, dataset users	Manual	Datasheet
Datasheets for datasets help ML engineers notice and understand ethical issues in training data [111]	2021	N/A	N/A	Study
Datasheets for Energy Datasets: An Ethically-Minded Approach to Documentation [88]	2023	dataset creators	Manual	Datasheet
Datasheets for Healthcare AI: A Framework for Transparency and Bias Mitigation [93]	2025	dataset creators, data experts	Automated	Framework
DescribeML: a tool for describing machine learning datasets [75]	2022	dataset creators	Hybrid	Application
Ensuring Dataset Quality for Machine Learning Certification [99]	2020	dataset creators, dataset users	Hybrid	Framework
FactSheets: Increasing trust in AI services through supplier's declarations of conformity [57]	2019	dataset creators	Manual	Datasheet
Goods: Organizing google's datasets [76]	2016	dataset creators, dataset users, dataset auditors	Hybrid	Application
Healthsheet: Development of a Transparency Artifact for Health Datasets [89]	2022	dataset creators, data experts, data practitioners	Manual	Datasheet
How to Automatically Document Data With the codebook Package to Facilitate Data Reuse [82]	2019	dataset creators	Automated	Application
Interactive Model Cards: A Human-Centered Approach to Model Documentation [112]	2022	N/A	N/A	study
Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities [115]	2023	dataset creators, dataset users	Manual	Toolkit
Machine learning data practices through a data curation lens: An evaluation framework [100]	2024	dataset practitioners	Manual	Framework
Method cards for prescriptive machine-learning transparency [90]	2022	dataset creators	Manual	Datasheet
MithraLabel: Flexible Dataset Nutritional Labels for Responsible Data Science [77]	2019	dataset creators	Automated	Application
Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face [66]	2024	N/A	N/A	Study

Network report: a structured description for network datasets [91]	2022	dataset creators, dataset users	Hybrid	Datasheet
On Responsible Machine Learning Datasets with Fairness, Privacy, and Regulatory Norms [92]	2024	dataset creators, data experts	Hybrid	Datasheet
On the effectiveness of dataset watermarking [78]	2022	dataset creators	Automated	Application
Ontology-Supported AI Model and Dataset Management [79]	2024	dataset creators, dataset users	Hybrid	Application
Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments [80]	2024	dataset creators	Hybrid	Application
Prov-Dominoes: An approach for knowledge discovery from provenance data [81]	2024	dataset creators	Hybrid	Application
Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards [109]	2021	dataset creators, data organizations	Manual	Schema
Right the docs: Characterising voice dataset documentation practices used in machine learning [113]	2023	N/A	N/A	Study
Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations [101]	2025	dataset creators, data experts, dataset researchers	Manual	Framework
Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets [114]	2022	N/A	N/A	Study
The ABC of Data: A Classifying Framework for Data Readiness [102]	2020	dataset creators	Automated	Framework
The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners and Context for Policymakers [38]	2024	dataset practitioners	Manual	Framework
The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence [2]	2022	dataset creators, data practitioners	Hybrid	Framework
The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards [58]	2018	dataset creators	Hybrid	Framework
The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track [4]	2025	N/A	N/A	Study
Toward FAIR Semantic Publishing of Research Dataset Metadata in the Open Research Knowledge Graph [107]	2024	dataset researchers	Hybrid	Markup format
Towards accountability for machine learning datasets: Practices from software engineering and infrastructure [6]	2021	dataset creators	Manual	Framework
Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata [5]	2022	N/A	N/A	Study

Using Large Language Models to Enrich the Documentation of Datasets for Machine Learning [138]	2024	dataset practitioners	Automated	Application

Table 3: Items included in the review, along with the level of automation for each tool as described in Section 3.2.3. Audiences are based on target audiences identified by authors in their contributions.