Data Statements: From Technical Concept to Community Practice

ANGELINA MCMILLAN-MAJOR*, Department of Linguistics, University of Washington, USA
EMILY M. BENDER*, Department of Linguistics, University of Washington, USA
BATYA FRIEDMAN*, Information School, University of Washington, USA

Responsible computing ultimately requires that technical communities develop and adopt tools, processes, and practices that mitigate harms and support human flourishing. Prior efforts toward the responsible development and use of datasets, machine learning models, and other technical systems have led to the creation of documentation toolkits to facilitate transparency, diagnosis, and inclusion. This work takes the next step: to catalyze community uptake, alongside toolkit improvement. Specifically, starting from one such proposed toolkit specialized for language datasets, data statements for natural language processing (NLP), we explore how to improve the toolkit in three senses: (1) the content of the toolkit itself, (2) engagement with professional practice, and (3) moving from a conceptual proposal to a tested schema that the intended community of use may readily adopt. To achieve these goals, we first conducted a workshop with NLP practitioners in order to identify gaps and limitations of the toolkit as well as to develop best practices for writing data statements, yielding an interim improved toolkit. Then we conducted an analytic comparison between the interim toolkit and another documentation toolkit, datasheets for datasets. Based on these two integrated processes, we present our revised Version 2 schema and best practices in a guide for writing data statements. Our findings more generally provide integrated processes for co-evolving both technology and practice to address ethical concerns within situated technical communities.

CCS Concepts: • **Social and professional topics** → *Socio-technical systems*; *Software selection and adaptation*; • **Computing methodologies** → *Language resources.*

Additional Key Words and Phrases: dataset documentation toolkits, data statements, professional practice, responsible innovation, value sensitive design

1 INTRODUCTION

Responsible computing entails, in part, proactively addressing harms (often rooted in social structure and inequities) and supporting human flourishing through our design, development, and use of computing technology. In this vein, scientific and technical communities have a special role to play in developing new tools and processes to mitigate harms by foregrounding human values. However, these tools and processes will only be beneficial if they are used and adapted by the relevant technical communities, ultimately changing practice. We consider here how to bring about such a change in practice, presenting our process for taking one such technology — the data statements toolkit for documenting language datasets used in natural language processing (NLP) systems — from envisioned concept and prototype to a practice that is both adapted to and adopted by the NLP community. We believe that technical community uptake requires interaction with the community, with two-way knowledge sharing: improving the technology based on community insight while training community members in its use.

*Authors contributed equally to this research.

Authors' addresses: Angelina McMillan-Major, aymm@uw.edu, Department of Linguistics, University of Washington, Box 352425, Seattle, Washington, USA, 98195-2425; Emily M. Bender, Department of Linguistics, University of Washington, Box 352425, Seattle, Washington, USA, 98195-2425; Batya Friedman, Information School, University of Washington, Box 352840, Seattle, Washington, USA, 98195-2840.

2832-0565/2023/5-ART
https://doi.org/10.1145/3594737

By co-evolving both our technical tools and our social structure and practice, we are better positioned to arrive at integrated technology and practice that respond meaningfully to ethical concerns.

Data statements were initially proposed by Bender and Friedman [2] in response to growing awareness of the fact that machine learning (ML) approaches to language technology bring various risks of harm to both system users and others affected by system use [5, 25, 30]. ML approaches to any problem involving data created by or about humans have similar risks, but both the particular risks and the ways in which they connect to data collection practices differ by data type. Data statements, which are honed to their data type, are part of a wave of convergent dataset and model documentation proposals (see §2) that seek to position technologists, those who procure and deploy technology, and community members to mitigate potential harms by providing transparency into the data used for training and testing such systems.[1]

Toolkits refer to physical and digital materials that support people in carrying out methods and processes [16]. Considering documentation toolkits and their purposes, no toolkit can produce any benefit if people don't use it to create documentation. Furthermore, the benefit of the documentation will be limited if it is not sufficiently detailed, nor accessible. So we asked two intertwined questions: How do we adapt our proposed toolkit and practice so that it is feasible for the practitioners we hope will take it up to do so? And: How do we facilitate community uptake? In this paper we present both the ways in which we engaged and learned from the community and the resulting improved toolkit, including a revised schema and distilled best practices. We present the revised schema and best practices in a guide which we developed to support data statement authors in creating documentation accessible both to technologists and to third parties who need or want to understand data used to construct technology.[2] As is evident from our characterization of these outcomes, we view the toolkit and associated practices as a single intertwined system. Most broadly, our contributions speak to how to evolve both technology and practice to address ethical concerns within situated technical communities.

We structure the paper as follows: In §2 we present an overview of recent documentation proposals for datasets, models, and systems and situate data statements within this ecosystem, as well as a review of the methodologies that we draw from value sensitive design [9]. We lay out our researcher stance, research questions and specific methods in §3, §4 and §5, respectively. §6 gives an overview of the revisions to the toolkit and in §7 we provide reflections on both our methodology and what we learned about how data statements fit into the landscape of data documentation practice and into practitioners' activities. Finally, in §8, we provide an outlook onto future work, including engaging with a broader set of stakeholders, further study of the uptake and use of data statements and generalizations to other data types.

2 BACKGROUND

2.1 Documentation Toolkits

In response to a wide range of potential harms from applying pattern recognition ("AI") at scale, in 2017-2019 several research groups, mostly from the United States by affiliation, began to develop documentation toolkits to support transparency in AI systems. As shown in Table 1, each of these documentation toolkits was developed with inspiration from a particular non-digital documentation format and with particular users, harms and use cases in mind.

More recently, as documentation toolkits gain traction, we are seeing two trends. First, documentation toolkits are being integrated into standard practice and early-stage standards to mitigate and manage bias in AI systems [28]. Second, initial documentation toolkits are being revised as part of iterative design processes, leading to more formalized and complete versions. For example, based on feedback from legal scholars and user studies, the

[1] Data statements also support the use and understanding of NLP systems built and evaluated with small datasets in resource-constrained scenarios, where ML may not be applicable.

[2] The guide [3] is available at http://techpolicylab.uw.edu/data-statements/

Toolkit	Inspiration	Focus	Ref
Datasheets for Datasets	Electronics documentation for components, etc.	Datasets: detailed documentation on key dataset design issues; intended for experts	Gebru et al. [13, 14]
Data Nutrition Project	Standardized nutrition labels for prepared food	Datasets: brief standardized format for details on the construction and contents of a dataset; intended for experts and non-experts	Holland et al. [17], Chmielinski et al. [6]
Data Statements for NLP	Description of participants in social and medical research	Datasets: highlights the design, the people represented, and considerations that arise from use of language data types	Bender and Friedman [2]
Nutrition Labels for Data and Models	Standardized nutrition labels for prepared food	Datasets and models: automatically calculated information about data and models to inform on production processes behind ML models	Stoyanovich and Howe [29]
Model Cards for Model Reporting	TRIPOD statement proposal in medicine	ML Models: model characteristics including type, use case, performance variance and performance measures; complement to datasheets	Mitchell et al. [22]
FactSheets	Suppliers Declaration of Conformity (e.g. telecom, transportation)	AI model or service: Purpose and criticality of a model; measures of a dataset, model or service; creation and deployment process	Arnold et al. [1]

Table 1. Documentation Toolkits: Inspiration and Focus

categories and questions employed in datasheets have been refined [13]; the Data Nutrition Project updated their Data Nutrition Label tool to include intended use cases [6]; and IBM expanded their FactSheet to include specialized template development for project teams [26]. In this second stage of documentation toolkit development, the field is moving beyond initial toolkit formulation to explore the needs of documentation writers, including addressing gaps and lack of clarity in the initial toolkit directions and support for skill development in writing, reading and using the toolkits. The work reported on here contributes to these second stage efforts.

2.2 Data Statements

A data statement consists of schema elements and is defined by Bender and Friedman as "a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software" [2, p.587]. The Version 1 schema consists of two parts: a long form and a short form. The long form contains nine schema elements, which each correspond to a set of questions or suggested descriptions about an aspect of the dataset, such as the curation rationale, language variety or demographics of the speakers in the dataset (§5 of [2]). The short form is a summary of the long form designed to be used in publications that reference the dataset. Practitioners are encouraged to use both forms in coordination with papers introducing datasets, as part of reports of experiments that used a dataset, and alongside documentation for a model trained on a dataset. Data statements have been used in dataset cataloging efforts to explore the gaps in existing data collections [33], and recent work with datasheets points to documentation's ability to support developers' awareness of ethical issues in ML technology [4]. For an illustration of both Version 1 and Version 2 of the data statements schema as well as a sense of how they differ, see Figure 1 on page 10, discussed further in §6.

2.3 Value Sensitive Design

Value sensitive design is an established approach for foregrounding human values and well-being in the technical design process [9]. Value sensitive design takes a broad stance in defining technology as a combination of tools, technologies, and infrastructure that shape human activity, encompassing both physical and digital artifacts [9]. Mok and Hyysalo [23] used them to integrate a new solar energy system into the architecture of a historic building, while Millett et al. [21] employed value sensitive design methodologies to improve informed consent features in internet browsers. At the core of value sensitive design is the tripartite methodology of iterative and integrative conceptual, technical and empirical investigations, as well as the practical strategy of co-evolving technology and social structure (including community practice). This methodology allows for extended inquiry into the interaction between technology and society through iterative investigation and evaluation over time. For example, the retrospective analyses that Millett et al. [21] conducted were built on the conceptual investigation described in Friedman et al. [8] and themselves were the foundation for technical interventions that were empirically evaluated in Friedman et al. [10]. Similarly, two of the authors of this paper, Bender and Friedman, employed this approach in the initial development of data statements [2]. We continue to draw on value sensitive design for the subsequent work presented here.

In their initial work, Bender and Friedman began with a conceptual investigation, drawing on the definition of bias presented in [11] as "systematic" and "unfair discrimination." They paid particular attention to how bias in computing systems could reflect preexisting social conditions or emerge over time when computing systems developed for a specific set of circumstances and populations were used in other circumstances and with other populations. As a proof-of-concept and technical investigation, Bender and Friedman then applied the data statements toolkit to two actual datasets, one of English Twitter data and one of English and French video interview data. In addition, they employed value scenarios [9, 24] as a conceptual method to explore how an at-the-time imagined documentation toolkit could provide benefit both in terms of mitigating bias and contributing to better science. Value scenarios provided a structured way of envisioning futures, bringing forward both potential positive and negative impacts of a not-yet-built-and-deployed technology on individuals, communities, fields and societies. One of their value scenarios, concerning the potential for data statements to become a force for exclusion if standardized too quickly, led Bender and Friedman to call for empirical investigations exploring how data statements as a practice would work for a diverse range of practitioners.

In the work reported here, we follow up on this call. In doing so we leaned further into value sensitive design's tripartite methodology. With the goal of improving the 2018 data statement schema from a community-of-use perspective, we first conducted an empirical investigation with one direct stakeholder group,[3] NLP dataset creators, to gather their perspectives and insights for how data statements and the surrounding practice could be improved by clarifying existing schema elements, identifying gaps where additional schema elements were needed and collecting best practices. Our empirical work was followed by two sequential technical investigations to revise the data statement schema. In the first we used the empirical workshop results to guide reformulation of the schema and identification of best practices; in the second, we compared datasheets for datasets to the reformulated schema to identify and fill any additional gaps.

3 RESEARCHER STANCE

Our research team based in the United States is comprised of computational linguists facile with NLP and ML systems and an information scientist skilled in the application of value sensitive design, particularly around mitigating bias in computing systems. All team members previously participated in developing documentation toolkits for datasets used in ML systems.

[3]Here we distinguish between *direct* stakeholders who interact directly with the documentation toolkit either by writing or reading documentation and *indirect* stakeholders who may never see the resulting documentation but nonetheless are affected by others' use of it [9].

4 RESEARCH QUESTIONS

In moving from an envisioned documentation toolkit to one positioned to be taken up by a research community, we sought to make the data statements toolkit more robust with respect to institutional contexts, researcher backgrounds and research goals. This motivated two broad research questions:

(1) How should the data statements for NLP schema be updated to better support the range of projects it might be used for in the international NLP community?
(2) How could we support practitioners in a wide range of institutional contexts in writing data statements and facilitate community uptake of this practice?

5 METHODS

To gain traction on these research questions, we took a two-phased approach, drawing on a similar methodological approach from Friedman et al. [12]. In Phase 1, to understand how NLP dataset creators would make sense of and utilize the existing schema (Version 1) we organized an empirical investigation in the form of an international community-based workshop with NLP practitioners (described in §5.1). Based on the Phase 1 workshop results, we developed an interim revised schema in a technical investigation. Then in Phase 2, to learn from others' efforts developing documentation proposals, we conducted a second technical investigation in which we carried out a close, analytical comparison between the schema and a related documentation toolkit (§5.2). Throughout, we paid particular attention to (1) how NLP dataset creators could effectively collect the information required for data statements; (2) identifying and developing heuristics for writing data statements; (3) managing privacy and ethical considerations, particularly those tied to small or vulnerable populations; (4) how data statements relate to other existing practices in the NLP community; and (5) how to document legacy datasets.

5.1 Phase 1: NLP community-based workshop

To uncover the strengths, gaps, confusions and limitations of the Version 1 schema elements (as published in [2]) as well as to generate best practices for writing data statements, we held an international workshop with members of the NLP community. The workshop was accepted as part of the 12th Language Resources and Evaluation Conference (LREC); due to COVID-19 and the eventual cancellation of the conference, the workshop was held virtually over three days, May 11-13, 2020. In this empirical investigation, we sought feedback from NLP dataset developers in order to evaluate the data statement schema in practice.

Participants and their datasets. We recruited participants through an open invitation over standard workshop announcement channels for the NLP community. Specifically, we invited NLP community members to a working meeting where they would engage in writing data statements. We recruited as broadly as NLP workshop distribution channels would allow, in the hopes of getting a very broad range of perspectives, and succeeded in attracting participants from around the globe, though some regions (Europe, the US) were more represented than others. In total, 38 practitioners from 16 countries participated, including practitioners from Argentina, Mauritius, Sri Lanka as well as the US and Europe. Half (50%) of the participants identified as senior researchers, while 36.8% identified as junior researchers and 13.2% did not provide a response. The workshop was designed around training language technology practitioners. Though we had one participant who came from a different research community (legal scholarship), for the most part, there was considerable shared common ground in the academic training of our participants. This both facilitated productive working sessions and shaped the range of ideas elucidated in those sessions.

Most participants brought datasets to document; where multiple participants represented the same dataset, we considered them part of the same participant team. In total, there were 29 datasets, reflecting the collective geographical diversity both in terms of the language and content of interest. Just over half of the datasets were

collections of varieties of English; other languages represented include Arabic (a mix of Arabic language varieties), Argentinian Spanish, Basque, Javanese and Yoruba, to name a few. The genre of data ranged from Twitter posts to biomedical data to proverbs.

Workshop structure and procedures. The design of the workshop was driven both by our goal of eliciting formative feedback on the data statements schema as well as our goals of providing a useful training and networking experience for the workshop participants. It was also shaped by the fact that it took place over Zoom, early in the global experience of the COVID-19 pandemic. In this context, we sought to balance in-depth paired participant interactions with larger group work. We intended for participants to experience the process of writing and evaluating data statements within a peer review process, and then reflect upon and discuss those experiences with others. Towards the goal of providing a networking opportunity for participants across this international community, we designed workshop activities that we expected to provide opportunities for relationships to form, assigning new participant pairings over the course of the workshop.

The virtual workshop met synchronously in Zoom for six hours total, in 2-hour sessions across three contiguous days. In addition to these synchronous meetings participants completed some work asynchronously between sessions, as preparation for the next meeting. On Day 1, participant teams were introduced to each other and informed of the workshop's twofold goals: (1) for each participant team that brought a dataset to leave the workshop with a solid, if not complete, draft of a data statement for their dataset; and (2) for the workshop participants as a whole to identify improvements to the Version 1 schema elements and generate best practices for writing data statements.

To achieve these ends, we formed small groups of participants around the datasets they brought, with 1–2 datasets per group. In addition, the data statement construction process was supported with a shared digital worksheet presenting the Version 1 schema elements. For each element, the worksheet provided the element explanation (from data statements Version 1, as specified in [2]) and allowed for (a) notes; (b) draft text; (c) feedback; and (d) advice for future data statement authors.

The workshop flow was as follows. On Day 1, after the introductions, we put the participant teams into small groups to develop the first four schema elements using the worksheets. During this writing process, participants took on one of two roles: data statement "author" or "interviewer". The data statement author role entailed writing the actual schema elements for a particular dataset. The interviewer role entailed asking the data statement author questions about the dataset, to bring forward aspects which might need clarification, greater specification or were deemed unnecessary or redundant. In this sense, the schema elements functioned as questions to be asked by the interviewer and answered by the data statement author. Notes from this interview process were recorded on the worksheet. As "homework", participants finished drafting these schema elements. On Day 2, participants worked in small groups to review the schema elements drafted the day before and then in a second small group session repeated the drafting process for the remaining five schema elements, again finishing the drafting as homework. On Day 3, a final small group session allowed for peer review of the second set of elements. Finally, four breakout groups comprised of 8-9 participants with one facilitator met to reflect on the specific workshop activities and on data statements more generally. In these groups, participants were asked about topics such as what advice they would give to future data statements writers, what improvements they would like to see to the schema elements, potential uses as well as harms and misuses of data statements, and suggested best practices. Participants were therefore asked for their suggested best practices having just experienced the process of iteratively improving their own data statements and also providing feedback on others' drafts.

The materials from the workshop that served as the empirical basis for our analysis included recordings of the final breakout sessions and the short full-group debriefing sessions at the end of Days 1 and 2, as well as the data statements produced by the participants, the notes they included in their worksheets, and the notes they provided in the discussion questions worksheet for the breakout sessions on Day 3. We did not create Zoom

recordings of the small-group work on actual data statement development, as we believed that might have been perceived as intrusive and counter to the goal of building relationships among participants.

Data analysis. Using an inductive process [7], we systematically reviewed the recorded material on participants' worksheets and the group discussion transcripts to identify and consolidate potential improvements to the schema and best practices. Specifically, two members of our research team with deep knowledge of language datatypes and NLP systems annotated the worksheets for tips and suggestions as well as for strengths and weaknesses in the participant-written data statements, paying particular attention to where difficulties occurred as a result of the schema definitions and scope. The lens that we used to examine the participant-written data statements was how well and completely they addressed the schema element questions, with an eye toward potential sources of bias. We also attended to overshoot: material that went beyond describing the dataset itself to include background information which would be better placed elsewhere. In evaluating the strengths and weaknesses of participant-written data statements, we found patterns that led us to develop best practices (either as practiced by data statement authors or that would have helped data statement authors). We also observed instances in which the Version 1 schema was ill-suited to certain kinds of language data, as in the case of translation data where participants needed to describe the characteristics of two (or more) languages. For the group discussion transcripts we annotated ideas around best practices. We excluded as out of scope participant comments about creating datasets (rather than documenting them) and automatic generation of data statements. Based on the analysis of these two data sources, we revised the Version 1 schema and we created general and element-specific best practices.

Interim products. The data analyses and subsequent revisions resulted in the Version 2 (Phase 1) data statement schema and a draft guide for writing data statements (see §6 for details).

5.2 Phase 2: Analytical Comparison to Datasheets for Datasets

To check for completeness and make the data statement schema and best practices from Phase 1 even more robust, we followed a strategy of leveraging a related model [12], in this instance another documentation toolkit effort. In choosing a documentation toolkit for comparison, we sought one that also engaged with datasets (as opposed to other aspects of systems) in a detailed manner and, ideally, from another organizational and/or institutional context as a means to enrich our development work thus far. Of the documentation toolkits described in §2.1, datasheets for datasets [13] (we used v7 of the paper on arXiv [15]) is the most similar to data statements. As shown in Table 1, only two others pertained solely to datasets. Of these, data nutrition labels were designed to be 'at a glance', where datasheets provided more detail and thus made a better point of comparison to data statements. Datasheets were developed by industry researchers within a large tech company rather than in the academic research community, so we expected that they would capture different contextual and organizational perspectives, aligning with our stated research questions. Datasheets have also seen a high degree of uptake within the community. For example, the Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track recommended that datasets published at their venue be accompanied by documentation following datasheets for datasets, data statements, or data nutrition labels in both 2021 and 2022. This research community interest in datasheets has continued, as evidenced by the datasheets publication having over 1000 citations at the time of writing this paper.

In this technical investigation, we paid particular attention to how each toolkit conceptualizes what data is, who is writing documentation, who is reading documentation, what risks are being mitigated and what other purposes the documentation serves. To situate the comparison in the details of the two toolkits, we sought to account for each of the questions the datasheets schema asks documentation authors to consider, mapping datasheets questions to data statements elements where possible. Where there was no corresponding element in

our Version 2-Phase 1 schema, we either identified a location where the information could be added to the data statements schema, or marked the question as out of scope for data statements. We found information to be out of scope for different reasons, e.g., because it doesn't pertain to language data or because we believe it would be provided in complementary documentation to data statements, such as documentation for important ethical review processes (IRB or otherwise).

6 FINAL PRODUCTS: REVISED SCHEMA, BEST PRACTICES, AND GUIDE

The NLP community-based workshop (Phase 1) and comparison with datasheets for datasets (Phase 2) resulted in three products: (1) a revised schema (Version 2); (2) a list of key terms and best practices (general and element-specific) for writing data statements for NLP; and (3) a guide for writing data statements for NLP that presents (1) and (2) in a cogent manner. As shown in Table 2, the vast majority of the revisions were the result of the community-based workshop.

Revised schema (Version 2). The community-based workshop (Phase 1) resulted in the creation of 7 new schema elements as well as updates to the rationale, description and best practices of the other original 9 schema elements. In addition, the schema elements were reordered and reorganized; in one instance two elements were merged into one, resulting in a total of 15 schema elements in Version 2. These changes emerged from both explicit comments and feedback from the workshop participants as well as our data analysis. For example, 5 of the new schema elements (Preprocessing and Data Formatting, Limitations, Metadata, Disclosures and Ethical Review, and Glossary) were suggested by participants during the group discussions. Our analysis of the participants' data statements resulted in the additional Header and Executive Summary schema elements, as well as merging the Speech Situation and Text Characteristics schema elements into one element. The comparison with datasheets for datasets (Phase 2) yielded five additional revisions; all of these were to element descriptions. To illustrate the substance and depth of changes from Version 1 to Version 2, we present the changes made to two of the schema elements: Curation Rationale and Recording/Capture Quality.

The top part of Figure 1 shows the changes we made to the Curation Rationale schema element. (1) *Element order.* As it was the first element in the Version 1 schema, we observed that workshop participants tended to overload the element with introductory information about the dataset. In response, we made the Curation Rationale the third element, after the new Header and Executive Summary schema elements that allow for more context about the contents of the dataset. (2) *Motivation.* Originally, motivation for how the schema element serves the reader of a data statement came after the description of the content for the element. We moved this motivation to the start of the element in the *Why* section, and included additional motivation for how the Curation Rationale also supports dataset creators. A few other schema elements in Version 1 also included motivation for why the element was included in the schema; we made this consistent across all schema elements in Version 2, including a rationale for both writers and readers in the *Why* section for each element. (3) *Elaboration.* Finally, we drew from the analyses of both phases to add more clarifying questions such that a completed Curation Rationale may better support surfacing sources of societal and/or emergent bias that may be encoded in the dataset.

While the Curation Rationale retained the original conception of the element from Version 1 (with elaborations), the changes made to the Capture Quality schema element (formerly the Version 1 Recording Quality element) illustrate a considerable re-imagining of scope and, hence, name and description of the element. Figure 1 also shows the two changes we made to the Capture Quality schema element. (1) *Scope.* In analyzing workshop participants' data statements, we found that this element — originally designed to capture technical biases related to audiovisual equipment used — was used creatively to document a wider variety of technical considerations. These include systems used for correcting optical character recognition (OCR) output, API reliability when requesting data from online platforms and data degradation stemming from linked data becoming inaccessible. Accordingly, we broadened the element's scope to include these and other possible sources of technical bias when

Revisions	Phase 1: Workshop	Phase 2: Datasheet Comparison
General Best Practices	New	-
Key Terms	New	-
Schema Elements		
1 Header	New	Updated c
2 Executive Summary	New	-
3 Curation Rationale	Updated b, c, d	Updated c
4 Documentation for Source Datasets	Updated a, b, c, d	Updated c
5 Language Varieties	Updated a, b, c, d	-
6 Speaker Demographic	Updated b, c, d	-
7 Annotator Demographic	Updated b, c, d	-
8 Speech Situation and Text Characteristics	Merged and updated a, b, c, d	
9 Preprocessing and Data Formatting	New	Updated c
10 Capture Quality	Updated a, b, c, d	-
11 Limitations	New	-
12 Metadata	New	Updated c
13 Disclosures and Ethical Review	New	-
14 Other	Updated b, c, d	-
15 Glossary	New	-

Table 2. Revisions by source of change. Each element is comprised of a: (a) title, (b) rationale, (c) description, and (d) best practices. "New" refers to the addition of an entirely new element.

capturing observations of language use in the world for use as data in a dataset. (2) *Rationale.* As described above, we added the *Why* section to convey the importance of these considerations to both data statement readers and dataset creators.

Best practices. Our advice to data statement writers takes the form of best practices, identified through analysis of workshop participants' reflections as well as the strengths and weaknesses of the participant-written data statements produced during the event. There are 16 general best practices which are applicable across data statement elements or otherwise pertaining to the data statement as a whole. In addition, there are 47 element-specific best practices, ranging per element from one (for Speech Situation and Text Characteristics, Other and Glossary) to nine (for each of Speaker and Annotator Demographics). The best practices convey three levels of emphasis, distinguished linguistically: (1) Best practices we believe must be followed to create a successful data statement, articulated as imperatives. (However, in many cases, the imperative instruction is to *consider* a course of action.) (2) Best practices we strongly advise, expressed with *should.* (3) Best practices we propose as one good way to proceed, expressed with *recommend.* The determination of which level of emphasis to use for each best practice was decided through deliberation among the three authors on what information we thought would be feasible for data statement authors to provide in most contexts as well as what information data statement readers would need to answer questions relating to possible sources of bias.

As an illustration of the best practices, here is general best practice #4, which reads:

> *Some of the data statement elements concern information that may require advanced planning to collect (e.g., demographic information). We recommend determining what information is to be collected and how at the start of the project, leaving time for ethics review board approval as appropriate.*

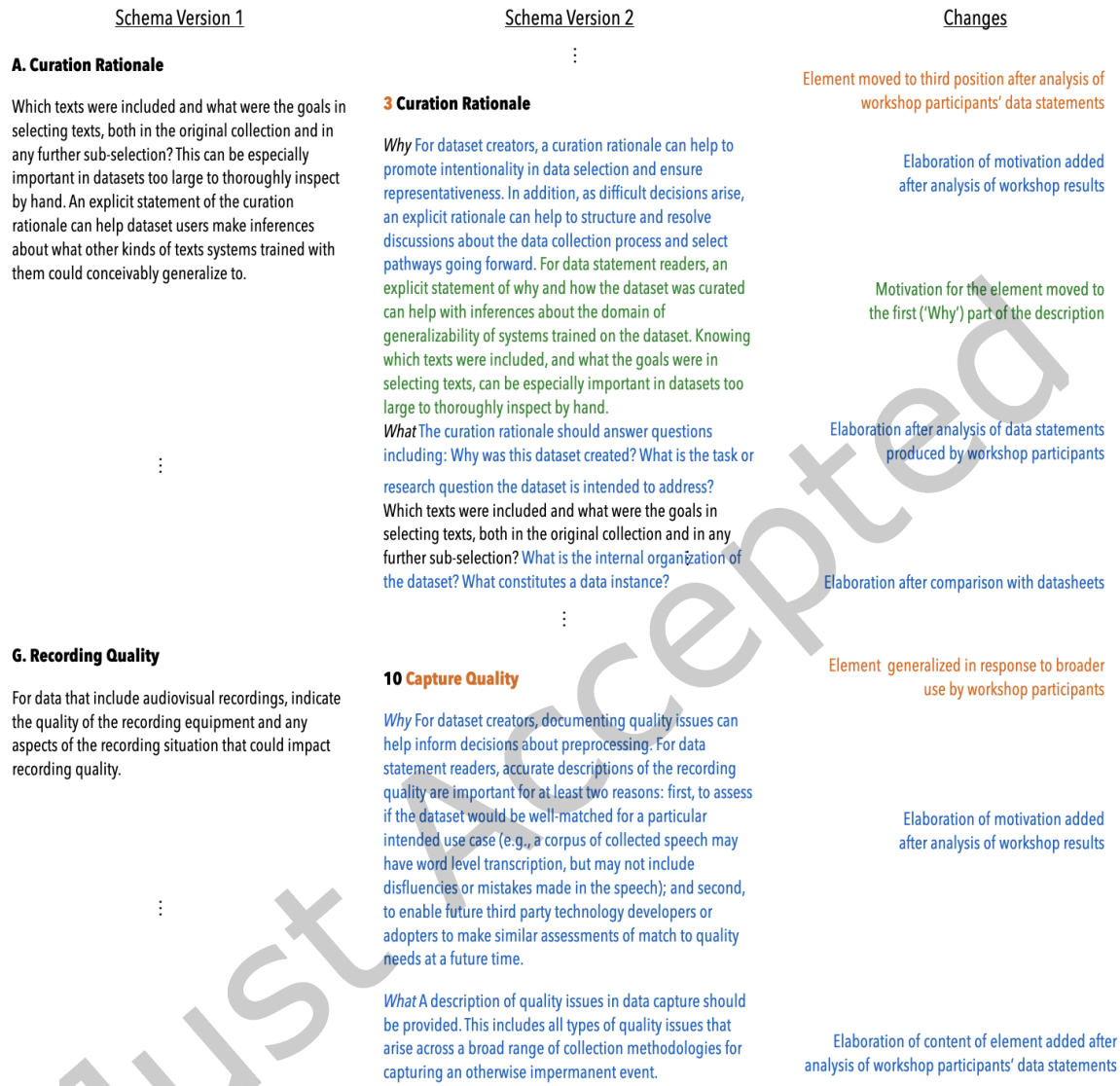

Fig. 1. Sample elements from Version 1 vs. 2 schema. Orange represents change of element order or title; green reorganization within an element; and blue elaborations to content

This best practice is derived from workshop participant comments that advocate working on the data statement early in the dataset development process, e.g., "Recommend drafting the data statement during the data creation process, as some information is more easily available at the time than later." This general best practice also reflects a proactive response to dataset creators who may feel uncomfortable about collecting and handling demographic information, even while understanding the importance of such information for creating representative datasets.

Guide for writing data statements. To assist technologists, scholars and others with writing data statements using the revised schema (Version 2) and drawing on the best practices, we created a third product which took the form of "A Guide to Writing Data Statements: For Natural Language Processing" [3]. The guide brings together the Version 2 schema elements and best practices into one integrated document that is organized to support the data statement writing process. General best practices (a total of 16) that cut across all aspects of the data statement writing process appear first, followed by key terms germane for language data types: annotator, disordered speech, elicited data, found data, language data, language variety, speaker, speech synthetic text and text. Next come the 15 schema elements, each on its own page. For each element, we provide a rationale (the *Why*), a description (the *What*) and element-specific best practices. Most pages have ample white space for note-taking and the user's annotations. A sidebar "schema-map" acts as a memory aid and facilitates flipping among related elements. The guide concludes with two appendices, the first for converting schema Version 1 to Version 2; the second for situating data statements within other documentation toolkits.

7 REFLECTIONS ON PROCESS AND PRODUCTS

On methodology. We took a two-phased approach, engaging first with NLP practitioners directly in the context of their own work writing data statements and then conducting a comparative analysis with a closely related documentation toolkit. In reflecting on our methodological strategy, we can make several observations. First, following value sensitive design's tripartite methodology, the two approaches we employed represented different types of investigations. Specifically, the workshop was an empirical investigation which positioned participants to directly engage with the data statement writing process and share their insights and advice in addition to the data statement artifacts they generated for their own datasets. As such, this empirical method invited participant creativity and allowed participants to express themselves in whatever ways they wished. We then followed this with two technical investigations: the Phase 1 reformulation of the schema and development of best practices and the comparative analysis with a closely related documentation toolkit. The comparative analysis focused on the technical structure and details of the two toolkits. This technical method afforded systematic and comprehensive surface-level comparison and was well positioned to shine a light on omissions in the interim Version 2 schema. Second, employing empirical and technical approaches in tandem yielded a broader set of improvements than either approach would have in isolation. Others wishing to improve similar toolkits might wish to employ a similar strategy: engaging a combination of empirical and technical investigations.

Considering the workshop further, we next call out two aspects of special interest: one following from participant make-up, the other from process. In terms of participant make-up, engaging directly with NLP practitioners from different countries and different institutional research contexts provided us with access to their collective wisdom and creativity. As a group, their depth and breadth helped us understand where the Version 1 data statement schema could be improved to better meet a wide range of needs and backgrounds, how data statement writers could be better supported with a structured and detailed guide to writing data statements and which key insights and best practices to share with others. This was particularly valuable given our goal of creating a documentation toolkit which would be accessible to researchers from institutional contexts different from US academia. Researchers based in different cultures helped us learn about different ways in which particular kinds of data about speakers and annotators might be considered sensitive as well as different levels of institutional support around ethics review. These lessons informed both the design of the schema and the best practices we articulated. In terms of process, the practice of interviewing a dataset developer as a means to elicit meaningful documentation led us to a more general observation: namely, that interviewing by an outsider serves as an effective method for eliciting content from dataset developers at a meaningful level of granularity. Where the term "documentation" usually evokes a dry asynchronous practice — the documenter writes, later

others read — we found the interview technique made data statement writing interactive and as a result more rewarding for both the data statement authors and (future) readers.

On automation. Among the suggestions from our workshop participants (and other members of the NLP community who we have spoken with about data statements) were those concerning automation and data statements. There are two variants: We might ask to what extent the production of data statements can be automated and to what extent data statements might be rendered automatically processable. In both respects, we see value in keeping this process manual. For the former, we believe that writing a thorough and beneficial data statement requires engaging thoughtfully with the data being documented, whereas automation tends to produce distance between author and dataset. For the latter, it is very important that data statements remain designed to be accessible to human readers, from a wide range of stakeholder groups. Designing them for automatic processing would likely render them less readable. In this sense, we see data statements are very much complementary to other kinds of metadata, such as the Dublin Core metadata standards [31, 32]. Such standards support discoverability of datasets; a data statement provides the reader who has discovered a dataset of interest with information about its *content* and *context*. That said, data statement authors are encouraged to use BCP-47 language codes which would allow for automation to determine which languages are represented in the data catalog and, importantly, which are not yet represented. As envisioned in the original data statements paper [2], that information would position the field as a collective to systematically fill gaps for underrepresented languages. Consistent with the sentiments above, this particular automated task would not interfere with the benefits of a primarily manual cataloging process.

On unanticipated use cases. From our perspective, one of the more interesting outcomes concerned use cases for data statements. Recall that data statements were envisioned to mitigate the harms of exclusion and bias in language technology and support transparency in the future applications of that technology through informed dataset selection, more thorough dataset analysis and bringing the ethical considerations of NLP data to the foreground for all NLP practitioners [2]. That said, the workshop participants identified several other use cases including functioning as an analogy to code README documents in increasing the accessibility of datasets, increasing the accessibility of NLP research to other fields, contributing to data repository metadata and serving as a planning tool for careful dataset development. These unanticipated uses point to the need for more general support of dataset development, integration and communication and increased valorization of the work that goes into data creation and dataset maintenance [27].

On situatedness of documentation practices. Our comparison to the datasheets schema allowed us to see some of the ways in which the initial development context of data statements shaped the resulting toolkit. Two key features of that context is that data statements (both Version 1 and Version 2) were developed from the perspective of academia and with a concrete focus on language datasets. We see the impact of the academic context in the way that data statements seek to complement rather than encompass work done by institutional review boards (IRBs), incorporating a place for a pointer to any IRB documentation in the Disclosures and Ethical Review element.

We find that our specific focus on language data enabled several key features of our toolkit. First, we are able to provide prompts in the schema for particular kinds of information that are relevant to issues of emergent bias with language datasets (e.g., dialect, genre). Second, we have a clear distinction between data (language produced by language users) and annotations (any additional labels added to that language data), and we prompt for information about the people involved in each process. Separating these out, we argue, will position dataset and technology users to better diagnose the source of problems as they arise. Third, and possibly most importantly, by grounding our toolkit in a specific data type, we are able to make our recommendations more concrete, which in turn makes data statements easier for dataset producers to write and for data statement readers of all backgrounds to understand.

On productive friction. The work reported here is the product of an interdisciplinary team. Authors McMillan-Major and Bender are computational linguists; author Friedman is a designer and technologist with expertise on human values in technical design. Navigating our interdisciplinary discussions was difficult and time consuming. We found that it was easy to misunderstand, both at the level of vocabulary and at the level of work behind the results from the other field. However, at the same time, we found that the resulting friction was generative, and taking the time to reach understanding led both to valuable new insights and to research products accessible to broader communities. For example, we developed the key terms in the data statements schema both to aid our own mutual understandings as well as to support non-NLP experts in their engagement and work with data statements. Ultimately, we found that the interdisciplinary experience brought value even beyond meeting this necessity: attending to the turbulence rather than trying to push past it and extending grace and respect across the disciplinary differences brought us benefits in the form of learning opportunities and insights that come from having to actively work towards clarity and mutual understanding.

On standardization: why, what and when? Those differing contexts of documentation schema development, varied targeted objects for documentation and disparate experiences of the developers themselves have resulted in a proliferation of diverse documentation schemas. With all of these different formats come challenges for coherent and widespread uptake of documentation. While standardization towards a few documentation schemas offers one way forward, it raises yet another set of questions: Should the schemas themselves or just the content of the documentation be standardized? At what jurisdiction should documentation be standardized, especially within interdisciplinary fields where contexts and data types may vary greatly? In the case of NLP, language data in the form of text is often accompanied by video and image data, which carry their own unique considerations for bias and ethical data management. Is it time to converge and standardize now, or is it better leave time for additional innovation and standardize at some point in the future? What rhythms of the innovation-convergence-uptake life cycle should we consider, which should we avoid? While institutions involved with standardization, such as NIST [28], ISO [19] and IEEE [18], work to provide broad guidance in terms of documentation over technical fields of all kinds, we expect that the answers to these questions and others for localized research communities will require active and inclusive community engagement to encourage uptake and effective documentation processes, practices and products.

On co-evolving technology and social structure. Value sensitive design points us to the need and opportunity to co-evolve technology with social structure [9]. That is, by developing technical tools and toolkits along with the social environments in which they will be used, we have a larger design space with which to engage and greater possibility to ensure that resulting practices will be responsive to the needs of individuals, communities, fields and society writ large. Doing this kind of co-evolution work is complex, nuanced work. Mok and Hyysalo [23] explore such co-evolution in the context of energy transition for a historical building in Finland; Magassa and Friedman [20] for the Washington State Access to Justice Technology Principles. Our work improving the data statements documentation toolkit contributes a focused case study for such co-evolution — one in which we worked directly with the community of practice both to improve the technology and to explicitly identify best practices around the technology's use. Our final products reflect this co-evolution approach, resulting in both a revised documentation toolkit (Data Statements Schema Version 2) and a set of best practices and guide for writing data statements. As the data statement toolkit is integrated into community practice, these methods could be used to understand how the integration process has changed the community and how those community changes necessitate the schema be once again revised. The overall approach we have taken as well as some of our specific methods for simultaneously engaging with a community around the development of the technical artifact will be of use to others who wish to pursue such co-evolution in their own design situations.

8 FUTURE WORK

The approach and methods reported here make progress on the trajectory from technical concept to widespread community practice. Yet more remains to be done. We point to three promising directions for future work.

Engaging with a broader set of stakeholders. Value sensitive design calls for a robust engagement with both direct and indirect key stakeholder groups. A stakeholder analysis for data statements yields many, diverse stakeholder groups, each of whom may interact with data statements in distinct ways. These include but are not limited to those (linguists, data scientists and others) who *create datasets*; those (computer scientists, data scientists and others) who *develop systems trained and tested on datasets* created by others; those (institutional decision-makers and IT personnel in organizations) who *select systems trained on datasets* created by others; those (doctors, human resources personnel, judges, lawyers, loan officers and others) who *use the outputs of systems trained on datasets* created by others; and those (individuals, communities, advocacy organizations and societies) who may *never touch the systems that were trained on the datasets but nonetheless are affected* by how others interpret and act upon the outcomes. All of these stakeholder groups need to be brought into the design process for data statements to ensure that the documentation contains the necessary information to be useful and that information is presented in a readable, comprehensible and usable form and format for each of the stakeholder groups. Our current work primarily addresses only the first stakeholder group above — those who create datasets.

Iteration and integration: use cases and on-going technical refinement. As data statements for NLP systems continue to be taken up, engaged and refined by diverse stakeholders, as a field we will be positioned to study their adoption, adaptation and effectiveness in practice. Open research questions include:

- What use cases emerge for data statements for NLP systems?
- How does the data statements schema for language data types need to be refined so as to be fully general, accommodating all kinds of observational data that may co-occur with or provide context for text or audiovisual language data?
- How does domain of use and organizational context impact the content of data statement schema elements and how those elements are used in practice (e.g., medical texts with patient, disease, and drug information vs. legal texts with case law)?
- How do diverse stakeholders read data statements and how readable are data statements, particularly for non-technical stakeholder groups?
- What evidence is there for the success of data statements for NLP (and related documentation toolkits) in mitigating bias and enabling better science?
- Where and how do data statements as a documentation toolkit come up short?

Generalizing to other data types. A key strength of data statements is their precision in relation to the dataset's data type. That is, the schema elements are honed to the data type that is being documented. The strength comes at the expense of generalizability, that is how readily data statement schema elements that were initially developed for language data types as used in NLP systems could be adapted in conception and structure to other data types. Our intuition was that some elements of the schema would likely carry across to other data types. After all, documentation for any data type will need to address the reasons underlying selection and inclusion (i.e., Curation Rationale) as well as disclosures and information on ethical review processes (i.e., Disclosures and Ethical Review). But elements specific to language data would need to be removed and new elements relevant to the data type being documented would need to be developed. Datasets with mixed data types (e.g., images with captions) present further complexities in documentation.

To explore further, we conducted a thought experiment as follows. Each of the authors chose a different (non-language) data type and considered how the schema elements developed for language data types might apply: vision data used to detect motion; sensor data used to train autonomous vehicles; and electrical signal data

used in brain-machine interaction. We compared our judgments about each schema element. By consensus we identified only 4 out of the 15 schema elements that would not carry over (elements 5–8: Language Varieties, Speaker Demographic, Annotator Demographic, and Speech Situation and Text Characteristics). The remaining 11 elements all carried over to each of the three considered data types, in some cases without modification, in others with minor adaptation to the element description. A development process akin to that of data statements for NLP could build out data statements for additional data types, replacing elements 5–8 with data type specific elements and adapting the details of the others. This thought experiment suggests that the grounding of the data statements toolkit in a specific data type, far from making it inflexibly bound to that data type, produced a resource that would be a beneficial starting point for adaptation to other domains.

9 CONCLUSION

Responsible approaches to machine learning will only gain purchase when the tools and technologies designed to support these outcomes are taken up and integrated into the everyday practices of technical and non-technical communities alike. In the work reported here, we explored how to support uptake of such a toolkit within in one particular technical community: data statements within the NLP community. Along the way, we also demonstrated how engagement with the technical community can be used to improve the toolkit, thus achieving two goals with one intervention. Framed in this manner, our work makes four key contributions. First, we provide a revised version of the data statements schema, together with a set of best practices for writing data statements, both presented together in a guide for writing data statements. Second, we developed a method for engaging a technical research community in uptake and adaptation of a documentation toolkit for machine learning systems, including workshop structure and interaction strategies. Third, with respect to improving the documentation toolkit itself, we provide a method and practice for further developing and improving such toolkits. Finally and most generally, we demonstrate how to move from an early-stage technical concept and innovation informed by value sensitive design to a community practice around a more robust technical artifact.

REFERENCES

[1] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John Richards, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

[2] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[3] Emily M. Bender, Batya Friedman, and Angelina McMillan-Major. 2021. A Guide for Writing Data Statements for Natural Language Processing. Available at http://techpolicylab.uw.edu/data-statements/.

[4] Karen L. Boyd. 2021. Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 438 (oct 2021), 27 pages. https://doi.org/10.1145/3479582

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science* 356, 6334 (2017), 183–186.

[6] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954* (2022).

[7] Juliet M. Corbin and Anselm L. Strauss. 2008. *Basics of qualitative research : techniques and procedures for developing grounded theory* (3e [ed.] / juliet corbin, anselm strauss. ed.). SAGE, Los Angeles [Calif.] ;.

[8] Batya Friedman, Edward Felten, and Lynette I Millett. 2000. Informed consent online: A conceptual model and design principles. *University of Washington Computer Science & Engineering Technical Report 00–12–2* 8 (2000).

[9] Batya Friedman and David G. Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination.* MIT Press, Cambridge.

[10] B. Friedman, D.C. Howe, and E. Felten. 2002. Informed consent in the Mozilla browser: implementing value-sensitive design. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.* 10 pp.–. https://doi.org/10.1109/HICSS.2002.994366

[11] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[12] Batya Friedman, Ian Smith, Peter H. Kahn, Sunny Consolvo, and Jaina Selawski. 2006. Development of a Privacy Addendum for Open Source Licenses: Value Sensitive Design in Industry. In *Proceedings of the 8th International Conference on Ubiquitous Computing* (Orange County, CA) *(UbiComp'06)*. Springer-Verlag, Berlin, Heidelberg, 194–211. https://doi.org/10.1007/11853565_12

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010v1 (2018). arXiv:1803.09010v1 http://arxiv.org/abs/1803.09010v1

[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *CoRR* abs/1803.09010v7 (2020). arXiv:1803.09010v7 http://arxiv.org/abs/1803.09010v7

[16] David G Hendry, Batya Friedman, and Stephanie Ballard. 2021. Value sensitive design as a formative framework. *Ethics and Information Technology* 23 (2021), 39–44. https://doi.org/10.1007/s10676-021-09579-x

[17] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv preprint arXiv:1805.03677* (2018). arXiv:1805.03677 [cs.DB]

[18] Intelligent Transportation Systems Committee of the IEEE Vehicular Technology Society and Standing Committee for Standards or the IEEE Robotics and Automation Society. 2021. *IEEE Standard for Transparency of Autonomous Systems.* Technical Report IEEE Std 7001-2021. Institute of Electrical and Electronics Engineers, New York, NY. https://doi.org/10.1109/IEEESTD.2022.9726144

[19] Joint Technical Committee ISO/IEC JTC 1, Information Technology, Subcommittee SC 38, Cloud Computing and Distributed Platforms. 2020. *Cloud computing and distributed platforms Data flow, data categories and data use Part 1: Fundamentals.* Technical Report ISO/IEC 19944-1:2020(en). International Organization for Standardization and the International Electrotechnical Commission, Geneva. https://www.iso.org/standard/79573.html

[20] Lassana Magassa and Batya Friedman. Under review. Toward inclusive justice: Applying the Diverse Voices design method to improve the Washington State Access to Justice Technology Principles. (Under review).

[21] Lynette I. Millett, Batya Friedman, and Edward Felten. 2001. Cookies and Web Browser Design: Toward Realizing Informed Consent Online. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 46–52. https://doi.org/10.1145/365024.365034

[22] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[23] Luisa Mok and Sampsa Hyysalo. 2018. Designing for energy transition through Value Sensitive Design. *Design Studies* 54 (2018), 162–183. https://doi.org/10.1016/j.destud.2017.09.006

[24] Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value Scenarios: A Technique for Envisioning Systemic Effects of New Technologies. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI EA '07)*. Association for Computing Machinery, New York, NY, USA, 2585–2590. https://doi.org/10.1145/1240866.1241046

[25] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, New York.

[26] John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. A Methodology for Creating AI FactSheets. *arXiv preprint arXiv:2006.13796* (2020). https://doi.org/10.48550/ARXIV.2006.13796

[27] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[28] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* Technical Report NIST Special Publication (SP) 1270, Includes updates as of March 2022. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1270

[29] Julia Stoyanovich and Bill Howe. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 42, 3 (2019).

[30] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Commun. ACM* 56, 5 (May 2013), 44–54. https://doi.org/10.1145/2447976.2447990

[31] Technical Committee ISO/TC 46, Information and documentation, Subcommittee SC 4, Technical interoperability. 2017. *Information and documentation — The Dublin Core metadata element set — Part 1: Core elements.* Technical Report ISO 15836-1:2017. International Organization for Standardization, Geneva. https://www.iso.org/standard/71339.html

[32] Technical Committee ISO/TC 46, Information and documentation, Subcommittee SC 4, Technical interoperability. 2019. *Information and documentation — The Dublin Core metadata element set — Part 2: DCMI Properties and classes.* Technical Report ISO 15836-2:2019.

International Organization for Standardization, Geneva. https://www.iso.org/standard/71341.html
[33] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE* 15, 12 (Dec 2020), e0243300. https://doi.org/10.1371/journal.pone.0243300