A Guide for Creating and Documenting Language Datasets with *Data Statements Schema Version 3*

2024

Angelina McMillan-Major & Emily M. Bender

ABOUT THE TECH POLICY LAB

The Tech Policy Lab is a unique, interdisciplinary collaboration at the University of Washington that aims to enhance technology policy through research, education, and thought leadership. Founded in 2013 by faculty from the Paul G. Allen School of Computer Science & Engineering, the Information School, and the School of Law, the Tech Policy Lab aims to bridge the gap between technologists and policymakers and help to generate wiser, more inclusive tech policy.

ABOUT THE VALUE SENSITIVE DESIGN LAB

The Value Sensitive Design Lab brings together designers and researchers to engage human values in the design of tools and technologies that support human flourishing. Founded in 1999 and housed in the Information School at the University of Washington, the Lab has taken on projects across levels of human experience, from security and privacy of web browsing to human dignity and reconciliation in international justice. The Value Sensitive Design Lab aims to develop toolkits and methods, as well as theory and practice, to support the moral and technical imaginations of designers of all kinds.

LICENSE

ACKNOWLEDGMENTS

From their conception, the development of data statements has been a collaborative endeavor. The project began in an exciting discussion over coffee between Emily and Batya Friedman, after Batya had given a guest lecture on value sensitive design in Emily's graduate seminar. The development of the initial schema, along with the sample data statements published in Bender and Friedman 2018 benefitted from interaction with Zeerak Talat and Mitch Marcus. The development of Version 2 of the schema, along with the associated Guide for Writing Data Statements for Natural Language Processing (Bender, Friedman and McMillan-Major 2021), reflects the fruits of collaboration with the participants in our 2020 LREC workshop. We are grateful to Zeerak Talat, Leon Derczynski, Luciana Benotti, Bonaventure F. P. Dossou, Chris Emezue, Itziar Gonzalez-Dios, Amy Isard, Neelam Pirbhai-Jetha, Surangika Ranathunga, Beatrice Savoldi, Marc Schulder, and the other participants for their generosity in sharing their experiences.

Angelina joined the data statements project to support Emily and Batya in organizing the workshop and was inspired by the workshop's enthusiastic discussions to continue the development of data statements as part of her dissertation work. Version 3 of the schema reflects that work, which benefitted from collaborative discussions with Benjamin Frey, Julie A. Hochgesang, Rose Stamp, Santiago Esteban, and Merve Ünlü Menevşe. The visual design of this guide (as well as the previous one) reflect the talent and dedication of Elias Greendorfer. We gratefully acknowledge the thoughtful input of all of these scholars. Most of all, we are grateful to Batya for encouraging our interdisciplinary work together to be patient and methodical. The UW Tech Policy Lab and the Frances and Howard Nostrand Endowed Professorship financially supported the data statements research as well as the development and design of the Version 2 and Version 3 guides.

DATA STATEMENTS LINEAGE

Schema Version 1
Dataset documentation

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science (Bender & Friedman, 2018)

Schema Version 2
Dataset documentation refined by scientific community engagement

A Guide for Writing Data Statements for Natural Language Processing (Bender, Friedman, & McMillan-Major, 2021)

Data Statements: From Technical Concept to Community Practice (McMillan-Major, Bender, & Friedman, 2024)

Schema Version 3
Dataset documentation and creation with best practices for language community dataset development

Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities (McMillan-Major, 2023)

This document:

A Guide for Creating and Documenting Language Datasets with Data Statements (McMillan-Major and Bender, 2024)

A Guide for Creating and Documenting Language Datasets with *Data Statements Schema Version 3*

Angelina McMillan-Major & Emily M. Bender

TABLE OF CONTENTS

INTRODUCTION

Data statements provide essential information about the characteristics of datasets, including but not limited to the curation rationale and data sources. Data statements are intended to help with (1) the conceptualization of and planning for datasets, in order to create datasets that reflect community needs, (2) the mitigation of the harms caused by bias in the dataset (such as a mismatch between training datasets and contexts where systems are deployed) and (3) the creation of a more inclusive data catalog, through identifying gaps. While first developed with language data types, data statements could be produced for a wide range of data types with adjustments to the schema to account for the unique characteristics of the specific data type.

Data statements in the context of language data types were first conceptualized in 2017 by Emily M. Bender and Batya Friedman at the University of Washington. The concept of data statements and first version of the schema was published in 2018 in the *Transactions of the Association for Computational Linguistics* and presented at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). The next two years saw significant interest and uptake. In 2020 Bender, Friedman, and Angelina McMillan-Major organized a workshop at the 12th Language Resources and Evaluation Conference, with the goals of supporting broader uptake and learning how to make data statements a suitable practice across different research and institutional contexts. The results of this workshop led to an updated schema (Version 2), a set of best practices, and a guide to writing data statements, all released in 2021. Data statements schema Version 2 and Bender, Friedman, and McMillan-Major's reflections on the documentation development process were accepted in 2022 and published in 2024 in the first issue of the Association for Computing Machinery (ACM) *Journal of Responsible Computing*. McMillan-Major continued to develop data statements in her dissertation work by shifting the perspective of data statements to include prospective dataset documentation and incorporating language communities as direct stakeholders[1] of the dataset curation and documentation process. McMillan-Major and Bender then refined McMillan-Major's dissertation work into the current version of the schema, Version 3.

This guide contains information about data statements for language datasets used in natural language processing systems. The schema elements have been honed to the particular characteristics of language datasets, including linguistic context, language user demographic, and annotator demographic. This guide for writing data statements provides the rationale, definitions, and suggestions for each of the elements as well as general best practices.

[1] While the term stakeholder is widely used in practice, we acknowledge that the term is inappropriate in some contexts, particularly due to how it continues the dispossession of indigenous communities in the Americas of their rights. For more information, see Bruijn and Whiteman 2010.

HOW TO USE THIS GUIDE

This document is intended to be used both as a guide for writing data statements and as an aid for planning data collection. We recommend familiarizing yourself with the content of the guide as early in your data collection process as possible. It will provide ideas about what design decisions may need to be considered at different points in the dataset curation process and about what information you'll need to collect as you create your dataset in order to complete your data statement. In addition, drafting your data statement early in the data collection process may help reduce the time and resources necessary for producing documentation compared to drafting it after the dataset has been published and rationales for design decisions have been forgotten. Data statements should be revised often throughout the dataset curation process to ensure that they are accurate and up to date and that they continue to guide the conceptualization and construction of the dataset. In drafting and revising your data statement, we also recommend soliciting input from as many people involved in the dataset creation as possible. For projects led by or co-designed with local language communities, consider using the C3DAR guide, which includes the content of this guide as well as additional best practices and considerations for collaboratively developing community language datasets.

Schema. The schema elements can be written in any order and we recommend drafting them as much as possible early in the dataset design process, even prior to any data collection. Any changes to the dataset design should then lead to revisions of the relevant schema element drafts. Each schema element contains information on *Why* the element is included, *What* to include in that element for your data statements, and *Best Practices* for writing that specific element. Before drafting a schema element for your data statement, read the *Why*, *What*, and *Best Practices* for that element in order to understand what should go in that section of your data statement and how to present that information to your intended audiences. These sections provide specific motivations, criteria, and examples for their respective elements. The questions listed in the *What* section of each schema element are written so that they may be interpreted both as future-looking questions during dataset design and as questions about the current state of the dataset when users are accessing a completed dataset.

Best Practices. You will find best practices for the overall writing of data statements, that we refer to as General Best Practices, at the beginning of the guide. In addition, the description for each schema element contains best practices that pertain to that element specifically. The best practices use particular phrases to communicate the strength of the recommendation: (1) imperatives such as "Write the data statement…" or "Make use of…" indicate practices that *must be followed*; (2) statements using "should" are *strongly advised*; and (3) statements using "We recommend…" are *one good way to proceed*, but not necessarily the only way.

GENERAL BEST PRACTICES

1. Remember that a broad range of people may be consulting data statements including but not limited to researchers within natural language processing, researchers in other fields (e.g., linguistics, law, or digital humanities), regulators, procurers, and members of and advocates for affected communities.

2. For datasets containing sensitive or proprietary information, whenever possible, write the data statement so that it can be made publicly accessible (e.g., avoid including non-anonymized sensitive information).

3. Some of the elements concern information that may require advance 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. Consult with communities early about appropriate demographic categories.

4. For refining your documentation, we recommend using an interview format with an external partner (e.g., someone not involved in the project). This is both fun and instructive. To conduct the mock interview, the external partner treats each element as a question to be posed to a project member. In engaging with someone not involved in the construction of the dataset to discuss and clarify answers, you can get a good sense of what information and how much detail is needed in the documentation.

5. When using technical terms, make use of 17 Glossary.

6. When information is not known or unavailable, state this explicitly. It is valuable for readers to know, for example, that demographic information or information about specific language varieties is unavailable. Missing information is not a reason to forgo creating documentation; clearly indicate what is missing and provide what information you can.

7. For datasets with extensive documentation outside this document (e.g., annotation guides), provide short summaries with pointers to the longer documents. It should be possible to know which key questions are answered in the other document(s).

GENERAL BEST PRACTICES CONTINUED

8. Writing clear, concise documentation takes time and thought. We recommend iterating on the text of the data statement.

9. For datasets containing materials that could be a trigger for trauma, we recommend making a note of this in either 3 Curation Rationale or 16 Other.

10. If you reference papers and resources (aside from the dataset citation provided in 1 Header), include a reference list at the end of the documentation with full citations.

11. Once drafted, review your documentation for words or phrases used to describe language users or their language varieties that might be experienced as diminishing and make revisions as appropriate.

12. Consider accessibility. When possible, use state of the art tools to check for accessibility, for example, for blind and low-vision readers.

13. Publish the documentation in the language(s) of the dataset in addition to any relevant languages of broader communication.

14. Provide the documentation together with the dataset. This is the canonical location for the most up to date version of the documentation. 2 Executive Summary along with a link to the documentation should be included in (1) any paper discussing the dataset or its uses and (2) the documentation for any system trained on the dataset. In publications presenting datasets, we recommend including the documentation as an appendix along with a pointer to where updated versions of the documentation may be found.

15. For datasets that will not be publicly available (e.g., those containing non-anonymized health information or proprietary data), whenever possible make the documentation publicly accessible. See also General Best Practice 2 above.

KEY TERMS

Annotator refers to someone who assigns annotations to the raw language data, including transcribers of spoken or signed data.

Community refers to a larger social structure represented by language users within the dataset, defined in terms of relationships and by members of the community itself.

Data instance refers to the data, metadata, and annotations associated together as one unit within the dataset.

Data subject refers to an individual person represented in the dataset, either as a language user or as a referent in the text.

Disordered speech or sign refers to speech or sign that has been affected by physiological conditions that affect a person's ability to produce speech sounds or signs.

Elicited data refers to text that language users were prompted to produce specifically for the purposes of constructing the dataset.

Found data refers to text that was produced by language users for their own communicative purposes and collected after the fact for a dataset.

Language data refers to spoken, written or signed utterances.

Language user refers to someone who is competent in at least one modality for a language, meaning they are able to speak, sign and/or write in the language as well as perceive and understand speech, sign and/or text in it.

Language variety refers to a manifestation of a given language (e.g., dialect); it does so without privileging one manifestation of the language as primary over others.

Stakeholder refers to a group or individual who can affect and/or is affected by the creation or publication of the dataset.

Synthetic text refers to text produced by an algorithm rather than a person.

Text refers to a sequence of language data, whether written, spoken, or signed.

1 HEADER

Why

For dataset creators and data statement authors, this information ensures that credit and responsibility for the various documents are allocated appropriately.

For data statement readers, this information clarifies the source, authorship, and contributions for the various documents pertaining to a dataset. Such information is particularly important when the sources, authors, and contributors of the data statement differ from those of the dataset, or when different versions of the data statement have different sources, authors, and contributors.

What

The header should include the following:

- Dataset Title
- Dataset Curator(s) [name, affiliation, role]
- Dataset Version [version, date]
- Dataset Citation and, if available, DOI
- Data Statement Author(s) [name, affiliation, role]
- Data Statement Version [version, date]
- Data Statement Citation and, if available, DOI
- Links to versions of this data statement in other languages

Best Practices

In order to manage updates over time, both datasets and their associated data statements should be versioned. That is, each updated dataset version should have its own updated data statement version. The data statement version number should be included in the data statement citation and is requested above. (Note that "Data Statement Version" refers to the version of the data statement, not the version of the data statement schema that is being used.)

In creating a standard citation for your data statement, we recommend including the following information about the data statement: authors, date, title, version, institution, and URL or DOI. Here are two examples:

> Gonzalez-Dios, Itziar. 2024. Data Statement for the Corpus of Basque Simplified Texts. Version 3. University of the Basque Country (UPV/EHU). http://www.ixa.eus/node/13302
>
> Schulder, Marc, Dolly Blanck, Thomas Hanke, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Lutz König, Susanne König, Reiner Konrad, Gabriele Langer, Rie Nishio and Christian Rathmann. 2024. Data Statement for the Public DGS Corpus. Project Note AP06-2020-01. Version 3. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University. doi (latest version): 10.25592/uhhfdm.1745

1 *HEADER*

Best Practices Continued

Consider web accessibility and the longevity of data statement location (e.g., university archives, ACM digital library, or a community-owned repository). See 15 Maintenance for further considerations.

Discuss with community partners how they would prefer to be acknowledged for their contributions. For some communities, coauthorship is appropriate, while others may have another preferred method. Consider also how to acknowledge contributions such as consultations on local knowledge, reviewing materials, and other efforts supporting the development of the project.

2 EXECUTIVE SUMMARY

Why

For dataset creators, the executive summary provides the project team with a concise description of the dataset that can serve as a guiding statement of purpose throughout dataset development. It can also be used in documents relating to the project, such as grant proposals, dissertation prospectuses, emails to potential collaborators, and project reports. A summary drafted before the data collection will need to be updated to reflect the final version.

For data statement readers, the executive summary provides a concise description of the dataset that can be used to make an initial determination about the appropriateness of the dataset for a specific purpose. The executive summary along with a pointer to the full data statement should be included in any publication using the dataset for training, tuning, or testing a system, and, as appropriate, for certain kinds of system documentation.

What

The executive summary is a short (60–100 word) summary of the data statement that at a minimum should include: (1) a one-sentence description of the curation rationale, (2) the language varieties, and (3) an overview of relevant quantitative information such as the dataset size.

Best Practices

We recommend finalizing the executive summary after the other elements have been drafted as that will help to clarify what level of detail is appropriate for this element and which details are best included in other elements.

We recommend limiting the executive summary to descriptive facts about the dataset in and of itself (e.g., do not make comparisons to or assume familiarity with other datasets). Doing so will enable reuse over longer time periods (e.g., 20+ years).

3 CURATION RATIONALE

Why

For dataset creators, a curation rationale can help to promote intentionality in data selection and ensure representativeness. In addition, as difficult decisions arise, an explicit rationale can help to structure and resolve discussions about the data collection process and select pathways going forward.

For data statement readers, an explicit statement of why and how the dataset was curated can help with inferences about the domain of generalizability of systems trained on the dataset. Knowing which texts were included, and what the goals were in selecting texts, can be especially important in datasets too large to thoroughly inspect by hand.

What

The curation rationale should provide answers to questions including the following, to be interpreted both as future-looking prompts for dataset design and informational questions from users of completed datasets: What is the intended purpose of this dataset? What is the task or research question the dataset is intended to address? What texts are included and what are the goals for selecting them? What is the internal organization of the dataset? What constitutes a data instance?

Best Practices

If the dataset includes different categories of data (e.g., radio news and talk shows), include additional qualitative information describing the rationale for including different categories and their distribution within the larger dataset. Further data statement elements below should address each subcategory.

If the dataset involves a subselection from a larger collection, specify topics, keywords, or other filters used and the reasons for choosing each. Technical details can be provided in 9 Preprocessing and Data Formatting.

We recommend finalizing the curation rationale after the other elements have been drafted. This will help to clarify what level of detail is appropriate for the curation rationale as well as which details are best included in other elements, thereby reducing repetition.

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 LANGUAGE USER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 DISTRIBUTION
15 MAINTENANCE
16 OTHER
17 GLOSSARY

4 DOCUMENTATION FOR SOURCE DATASETS

Why

For dataset creators, the source dataset design and documentation can provide examples and language to draw from or reference when drafting the current dataset design and documentation.

For data statement readers, the source dataset documentation can help with understanding how the current dataset builds upon and differs from the original task and data collection. Links to the source datasets show the user where to go look for further information, especially for the curation rationale of the source dataset(s).

What

For datasets built out of other pre-existing datasets, a link to a data statement for each source dataset should be included. If a data statement is not available, provide a link to a publication or other documentation. Provide links to licenses, copyright, or terms of use for source datasets, where applicable.

Best Practices

Include only immediate sources. For the situation where a chain of datasets have been built (e.g., A was the original source dataset; B was built from A; C was built from B), then the data statement for the most current dataset (e.g., C) should only refer to the immediate source (e.g., B).

Include enough detail in the body of the data statement so that, should the links between the data statement and the immediate source break, the data statement could function reasonably well as a stand-alone document.

If the source dataset was collected under specific consent conditions, ensure that those conditions allow for further reuse and distribution as needed by the current dataset. If not, consider contacting the source dataset manager about developing a opt-in consent procedure for the new use and dissemination of the data.

5 LANGUAGE VARIETIES

Natural language processing algorithms embed assumptions about language structure; when applying an algorithm to a dataset from a language variety that differs structurally from that embedded in the algorithm, unexpected behaviors may occur.

Why

For dataset creators, a clear conception of the targeted language varieties can help inform decisions about data sources, curation, and annotation.

For data statement readers, accurate descriptions of the language varieties in the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third-party technology developers or adopters to make similar assessments of match to populations at a future time.

What

All of the languages and language varieties represented in the dataset should be characterized with (1) a language tag from BCP-47 identifying the language variety (e.g., en-US or yue-Hant-HK), and (2) a prose description elucidating and elaborating on the BCP-47 tag (e.g., English as spoken in Palo Alto, California; Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin; French Sign Language as used in Marseille, France).

Best Practices

Describe all language varieties represented in the dataset (data, annotations, and other metadata). For translation datasets, this would include both sides of the bitext.

Especially for less well studied languages, the description of the language variety should include enough information to situate it for dataset users unfamiliar with that variety. These descriptions should be written with respect and care for how the community would like their language to be known in order to avoid harmful language ideologies (Kroskrity, 2005).

In the prose description, describe the dialects included in the dataset as accurately as possible with respect to national, regional, and other sociolinguistic variation (e.g., rather than saying “American English”, say “Standardized American English” or “Northeastern American English” as appropriate).

6 *LANGUAGE USER DEMOGRAPHIC*

Beyond the language variety tied to a community of speakers or signers (see 5 Language Varieties), individual language users bring their own identities to their linguistic patterns. Specifically, sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with the language user's demographic characteristics (Labov, 1966; Kusters and Lucas, 2022), as speakers and signers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). In addition, when individuals speak or sign a second language, properties of their first language affect their production in their second language (Ellis, 1994, Ch. 8; Quinto-Pozos, 2008). A further source of variation can be found in physiological sources such as disordered speech or sign (e.g., dysarthria; see Christensen et al. 2012, Nicolao et al. 2016, Tyrone 2014).

Why

For dataset creators, a clear conception of the demographic categories targeted during the data collection process can help inform decisions about data sources, curation, and annotation. Data statements also enable the discovery of underserved populations across the overall data catalog which, in turn, may influence choices for constructing the new dataset.

For data statement readers, accurate descriptions of the people represented in the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third-party technology developers or adopters to make similar assessments of match to populations at a future time.

6 *LANGUAGE USER DEMOGRAPHIC*

What

All of the language user groups represented in the dataset should be characterized with a prose description. Demographic categories are context- and culture-specific; therefore, locally appropriate categories and definitions should be used as determined by the community. Suggested specifications include:

- Age
- Gender
- Race/ethnicity
- Socioeconomic status
- First language(s)
- Proficiency in the language(s) of the data
- Number of different language users represented
- Presence of disordered speech or sign

Best Practices

Discussions of demographic categories should be informed by current best practice (e.g., as of 2024, for gender see Larson 2017 and Devinney et al. 2022; for race and ethnicity see Squizzero et al. 2021).

Because the definitions and labels of demographic categories can change over time, include the dates when the data were produced and when the data were collected.

If the dataset includes language users with different roles (e.g., interviewers, interviewees, and interpreters), provide demographic information for each role separately.

If the dataset consists entirely of synthetic text, if available, provide demographic information for the language users in the training data for the automatic generation system.

If the dataset contains both found and elicited data, provide separate language user demographics for each.

Be specific when describing demographic information, particularly with respect to category labels (e.g., rather than stating “all races” or “all ages” provide a set of labels or range of values) and source (e.g., self-reported vs. estimated).

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 LANGUAGE USER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 DISTRIBUTION
15 MAINTENANCE
16 OTHER
17 GLOSSARY

6 LANGUAGE USER DEMOGRAPHIC

Best Practices Continued

When self-reported demographic data is not available, we recommend estimating demographic data by referring to studies of relevant larger populations (e.g., surveys of gender identities of Wikipedia editors) rather than trying to infer labels with classification tools (e.g., name-based gender attribution).

Report demographic information at the level of the entire dataset rather than attached to individual language users to help protect their privacy.

Become informed about what demographic information may be safely gathered and shared. For example, when the number of participants and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect participant privacy.

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 LANGUAGE USER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 DISTRIBUTION
15 MAINTENANCE
16 OTHER
17 GLOSSARY

7 ANNOTATOR DEMOGRAPHIC

Linguistic variation correlated with language users' demographics is also relevant for annotators. Specifically, annotators' own life experience influences their knowledge of language and how language is used by others and, thus, their perception of what they are annotating (Derczynski et al 2016, Talat 2016). As people annotate training datasets, they necessarily bring their perspectives to their annotations and, thereby, into the natural language processing models trained on that data.

Why

For dataset creators, an accurate description of annotator demographics can be helpful in hiring annotators whose demographics closely match those of the language users in the dataset or, if that is not feasible, in identifying demographic gaps between annotators and language users in the dataset, and developing annotation guidelines accordingly, sensitive to those gaps.

For data statement readers, accurate descriptions of the annotators' demographics are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third-party technology developers or adopters to make similar assessments of match to populations at a future time.

What

All of the annotator groups represented in the dataset, including those who develop the guidelines, should be characterized with a prose description. Demographic categories are context- and culture-specific; therefore, locally appropriate categories and definitions should be used as determined by the community. Suggested specifications include:

- Age
- Gender
- Race/ethnicity
- Socioeconomic status
- First language(s)
- Proficiency in the language(s) of the data being annotated
- Number of different annotators represented
- Relevant training

7 ANNOTATOR DEMOGRAPHIC

Best Practices

Discussions of demographic categories should be informed by current best practice (e.g., as of 2024, for gender see Larson 2017 and Devinney et al. 2022; for race and ethnicity see Squizzero et al. 2021).

Because the definitions and labels of demographic categories can change over time, include the dates when the annotations were produced.

If the dataset includes annotators with different roles (e.g., translators and labelers), provide demographic information for each role separately.

If the dataset includes automatically produced annotations, to the extent possible provide demographic information for the training data for the automatic annotation system.

If the dataset contains both found and elicited annotations, provide separate annotator demographics for each.

Be specific when describing demographic information, particularly with respect to category labels (e.g., rather than stating “all races” or “all ages” provide a set of labels or range of values) and source (e.g., self-reported vs. estimated).

When self-reported demographic data is not available, we recommend estimating demographic data by referring to studies of relevant larger populations (e.g., surveys of gender identities of Wikipedia editors) rather than trying to infer labels with classification tools (e.g., name-based gender attribution).

Report demographic information at the level of the entire dataset rather than attached to individual annotators to help protect their privacy.

Become informed about what demographic information may be safely gathered and shared. For example, when the number of participants and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect participant privacy.

8 *LINGUISTIC SITUATION AND TEXT CHARACTERISTICS*

Characteristics of the linguistic situation can affect linguistic structure and patterns at many levels. For example, the intended audience of a linguistic performance can affect linguistic choices on the part of speakers, signers, and authors. The time, place, and cultural context allow for deeper understanding of how the language data collected relate to their historical moment. Both genre and topic also influence the vocabulary and structural characteristics of language data (Biber, 1995).

Why

For dataset creators, a clear conception of the targeted linguistic situation can help inform decisions about data sources, curation, and additional information to include through annotation (e.g., the timestamps of turn-taking in an asynchronous conversation).

For data statement readers, accurate descriptions of the linguistic situation in the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third party technology developers or adopters to make similar assessments of match to a target linguistic situation at a future time.

What

A description of the situation in which the linguistic production occurs and/or the relevant text characteristics should be provided. This schema element may also be used to describe the cultural context of the language practices collected. Suggested specifications include:

- Time and place of linguistic activity
- Date(s) of data collection
- Modality (spoken, signed, written)
- Scripted/edited vs. spontaneous
- Synchronous (e.g., in-person or live online chatting) vs. asynchronous (e.g., letters, emails, forums) interaction
- Language users’ intended audience
- Genre (e.g., newswire or social media)
- Topic (e.g., entertainment or natural disaster)
- Non-linguistic context (e.g., photos participants were all looking at; a game participants are playing)
- Any additional details about the cultural context

8 *LINGUISTIC SITUATION AND TEXT CHARACTERISTICS*

Best Practices

We recommend documenting as much of the linguistic situation and text characteristics information as possible before beginning the data collection. As the data is collected, update this information to reflect any changes.

When describing the cultural context, use community vocabulary, concepts, and interpretations to convey the cultural significance, when deemed appropriate for public dissemination by the community.

9 *PREPROCESSING AND DATA FORMATTING*

Why

For dataset creators, documenting the preprocessing procedure can help ensure that the procedure is applied consistently, especially when data is drawn from different sources or languages.

For data statement readers, this documentation can help clarify how changes introduced during preprocessing might affect system performance (e.g., replacing personal names with placeholders for anonymization, standardization of spelling, tokenization of sentences into words). Providing information about preprocessing also enables reproducible dataset construction.

What

A description of all preprocessing and data formatting modifications made to the data (except for annotations) should be provided, including information about any anonymization procedures. The description should also specify any tools used to make the modifications and whether the raw data is included in the dataset.

Best Practices

We recommend the description take the form of a list of ordered steps, with a link to external documentation of specific details, as appropriate.

If different preprocessing steps are applied to different parts of the dataset, document each set of steps separately (e.g., adding whitespace only to scripts which do not usually use whitespace).

If the dataset is a filtered version of a larger data collection, we recommend using this schema element to provide technical detail on the specifics of the filters and their applications (e.g., specific search terms or filtering processes). This technical description of the filtering process complements the reasons for filtering provided in 3 Curation Rationale.

To the extent possible, provide software version information, citations, and links to repositories for the tools used in automatic processing.

9 *PREPROCESSING AND DATA FORMATTING*

Best Practices Continued

When anonymizing video or image data, modifications to the data such as blurring faces may remove necessary linguistic context and information, especially for signed languages. If language users in the dataset have not agreed to public dissemination of their video or image data without anonymization, consider all available methods for protecting the language users' privacy, such as access restrictions.

10 CAPTURE QUALITY

Why

For dataset creators, documenting quality issues can help inform decisions about preprocessing.

For data statement readers, accurate descriptions of the capture quality are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case (e.g., a corpus of collected speech may have word level transcription, but may not include disfluencies or mistakes made in the speech); and second, to enable future third-party technology developers or adopters to make similar assessments of match to quality needs at a future time.

What

A description of quality issues in data capture should be provided. This includes all types of quality issues that arise across a broad range of collection methodologies for capturing an otherwise impermanent event.

Best Practices

For data that include audiovisual recordings, describe the quality of the recording equipment and any aspects of the recording situation that could impact recording.

As appropriate, use this element to address other data quality concerns (e.g., image-to-text processing, granularity of transcription, or API reliability).

11 *LIMITATIONS*

Why

For dataset creators, it can be helpful to enumerate issues that have arisen for similar tasks or datasets as well as factors that might hinder the collection of a fully representative dataset. Ideally, this should be done before collecting data, in order to identify mitigation strategies. When setbacks occur in the course of creating a dataset, updating this schema element can help identify practical impacts on the resulting dataset and the extent to which the dataset in its current form meets its stated goal; such assessment can be helpful in guiding further data collection as appropriate.

For data statement readers, accurate descriptions of the challenges encountered in creating the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third-party technology developers or adopters to make similar assessments of match to populations at a future time.

What

For any challenges not fully addressed, a description of those challenges and characterization of the resulting limitations of the dataset should be provided.

Best Practices

We recommend documenting the challenges you encounter in the dataset development as they occur, including both the challenge and your strategy for addressing it.

For identifying possible limitations, we recommend using toolkits, such as Envisioning Cards and the Lifecourse Checklist, which guide practitioners to consider different populations and what representation means, as well as broader impacts.

We recommend noting any further precautions you would like future users of the dataset to be alert to.

12 METADATA

Why

For dataset creators, it is important to be aware of and collect relevant metadata.

For data statement readers, data statements may be the “front door” through which they access the dataset. As such, it is important that the data statement contains pointers to the other metadata.

What

A collection of pointers to relevant metadata should be provided. Suggestions include:

- Annotation Guidelines: Link to the published or online guidelines used by annotators
- Annotation Process: Link to documentation providing metadata about the annotation process, including protections for annotator anonymity, annotator compensation, and any automated processes producing annotation
- Dataset Quality Metrics: Metrics for inter-annotator agreement and/or other numerical scores of dataset quality

Best Practices

Include the most durable citations or links available (e.g., ISBN or DOI).

13 *DISCLOSURE AND ETHICAL REVIEW*

Why

For dataset creators, a clear conception of the terms of the ethical approval can help inform decisions about data sources, curation, and annotation. Awareness of potential conflicts of interest can be helpful with managing or mitigating these. If any community represented in the data has an ethical review process, engagement with this process can help surface and address community-specific concerns with the dataset creation.

For data statement readers, information about funding sources (which may have shaped curation and other decisions at the time of dataset creation) and ethical review (including the conditions of consent) may impact dataset selection.

What

For projects supported by funding, a description of the funding source for the dataset and relevant information (e.g., grant number) should be specified. For projects that went through an ethical approval process, a link to the approving body (e.g., IRB) should be provided. In addition, include: a brief description of any consent processes; if language users in the dataset or annotators are compensated, how compensation rates are determined; and any potential conflicts of interest.

Best Practices

If your data collection process involves a consent procedure, describe this element briefly with phrases such as “written consent”, “oral consent”, or “implied consent”.

If your institution does not have or require an ethical review process, we recommend stating this. Consider using a phrase such as “No institutional ethics review process accessible at the time of dataset creation.”

We recommend stating whether or not the project has engaged with any ethical reviews processes organized by institutions local to the communities represented and describing any results from the engagement.

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 LANGUAGE USER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 DISTRIBUTION
15 MAINTENANCE
16 OTHER
17 GLOSSARY

14 *DISTRIBUTION*

Why

For dataset creators, having a detailed plan for distribution can help inform data curation decisions as it determines whether the team should only collect data that will allow for public distribution or if access to the dataset or parts of the dataset will be restricted. The data collection team should also provide distribution information to the people the data is collected from as part of the consent procedure.

For data statement readers, a detailed description of the permitted uses of the dataset can help in determining whether the dataset is suitable for a particular use case and whether the dataset can be further redistributed. The distribution explanation can also potentially help the reader find and access the dataset or explain why they are unable to. Data statements that communicate planned revisions or removal of the dataset in advance may also help data statement readers prepare for changes to the dataset.

What

A description of how the dataset is to be distributed should be provided. This includes the method of distribution (e.g., through a data archive, files on website, API, GitHub) and any access restrictions (e.g., sensitive or confidential content, intellectual property (IP) considerations, export controls, or other regulatory restrictions).

If an IP license, copyright, or terms of use (ToU) applies to any portion of the dataset, provide links to or reproduce the licenses, copyright, and/or ToU, and list any fees associated with these restrictions.

Other suggestions for detailing the distribution plan include providing such information as:

- Who has access to the dataset as of the writing of the documentation and who else it is intended to be distributed to
- What conditions, if any, there are for obtaining access to the whole dataset or any subsets of it
- Whether the dataset has a digital object identifier (DOI)
- Date(s) of distribution of the dataset

14 DISTRIBUTION

Best Practices

Whenever possible, data subjects should be made aware of any distribution of their data. Ensure that informed consent processes cover information about the distribution of the dataset, including license, copyright, or ToU.

Ask your working institution(s) and funding institution(s) for guidance around licensing requirements and considerations for your dataset and any associated computer code. Be aware that legal considerations vary by region, e.g., US vs. EU.

When choosing terms for a license, copyright, or ToU, consider uses that will be allowed as well as uses that will be disallowed. Engage stakeholders in decision-making processes when determining whether to allow third-party uses such as research, legal proceedings, technical development, and commercialization.

When choosing terms for a license, copyright, or ToU, also consider whether third parties will be allowed to host copies of the dataset, and if so, under what conditions. If the original dataset has access restrictions, for example to protect from online data scraping, then ensure that the license includes terms that require third parties to also implement those access restrictions. If the dataset will be actively maintained by the dataset creator, consider licenses that don't allow others to host the dataset without active processes for updating the hosted copy from the maintained original dataset.

If the dataset can be openly shared with the public without any restrictions, consider a CC0 license.

If the data is of significant cultural value to the community from which the data is collected, work with the community to determine appropriate licensing terms. Consider using the Traditional Knowledge (TK) Labels and TK Licenses developed by Local Context. Communities and their collaborators may also develop their own ToU or licenses. See examples such as the Kaitiakitanga License developed by Te Hiku Media or the DGS Corpus license conditions developed by the DGS-Korpus team at the University of Hamburg.

14 *DISTRIBUTION*

Best Practices Continued

If the data is not appropriate for open distribution, consider creating a restricted access licensing agreement and ToU. See examples such as the Linguistic Data Consortium's user agreements for members and non-members or the restricted portion of the American National Corpus.

If you are including code along with the distributed dataset, select a license appropriate for your code while taking into account how the license for the code and the license for the data may interact.

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 LANGUAGE USER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 DISTRIBUTION
15 MAINTENANCE
16 OTHER
17 GLOSSARY

15 *MAINTENANCE*

Why

For dataset creators, a maintenance plan for the dataset may help to ensure that the dataset will continue to be usable and accessible to the intended audiences. Developing a maintenance plan may also help in considering archiving options along with their benefits, risks, and costs prior to data collection.

For data statement readers, information about who maintains the dataset may help determine who to contact for questions about the dataset after it has been published. Information about previous updates and planned future updates may help determine which version of the dataset is most applicable to the reader's use case and help the reader plan for integrating dataset updates into their system development.

What

A description of how the dataset is to be maintained should be provided. This description should specify who is supporting, hosting, and maintaining the dataset and how to contact the manager of the dataset. Other suggestions for detailing the maintenance plan include providing such information as:

- Where to find and contribute to information about errors in the dataset
- How often, by whom, and how updates to the dataset are communicated to users
- Applicable limits on data retention and how those limits will be enforced (e.g., respecting agreements with data subjects regarding whether and when their data will be deleted)
- Whether older versions of the dataset are to be supported, hosted, and maintained
- How users are to be notified that the dataset is outdated or no longer available
- Whether others can contribute to the dataset and how; whether and how any contributions are validated and further distributed to other users

15 *MAINTENANCE*

Best Practices

Consider the accessibility and longevity of dataset location (e.g., university archives or a community-owned repository), especially with respect to how data subjects access the data.

We recommend having a process for removing data, including receiving and adjudicating data removal requests and implementing removal when appropriate. Wherever possible, data subjects should be consulted in the development and application of this process.

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 LANGUAGE USER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 DISTRIBUTION
15 MAINTENANCE
16 OTHER
17 GLOSSARY

16 OTHER

Why

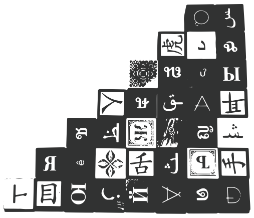
The data statement schema was designed to be broadly applicable to datasets containing language data, however there may be specific situations in which it would be useful to document other aspects of the dataset not covered by the schema.

What

Any further considerations that are relevant for the dataset should be included here.

Best Practices

Avoid blurring the content boundaries of the established schema elements. If you identify a piece of information that does not fit in any of the other schema elements, include it here.



17 GLOSSARY

Why

For data statement authors, using technical terms can make it easier to write efficient and precise documentation. Where relevant, using cultural terminology provided by the community throughout the documentation centers the community's understanding of the data and its cultural significance. In both cases, providing definitions for these terms can make the data statement accessible to a wider variety of audiences.

For data statement readers, definitions of technical terms can be especially important for three purposes: (1) understanding the intended use and limitations of the dataset, (2) conducting diagnostic analyses of system breakdowns, and (3) supporting the ability of impacted individuals, communities and their representatives to seek accountability for potential harms resulting from systems employing the dataset. Definitions of cultural terminology can be important for understanding and interpreting the data in locally appropriate ways and acknowledging community cultural knowledge.

What

A list of terms and associated definitions that may be technical or unfamiliar to non-experts should be provided.

Best Practices

We recommend engaging with someone outside of the project development team in order to determine what terms to include.

APPENDIX: RELATED DOCUMENTATION TOOLKITS

In response to a wide range of potential harms from applying pattern recognition ("AI") at scale, Couillault et al. (2014) identified transparency and systematic documentation as important first steps toward ethical data management and use. Beginning around 2017-2018, several groups developed toolkits for documentation to support transparency in AI systems. Each of these toolkits was developed in a specific research or industry context, with particular authors, users, harms, and use cases in mind. We briefly describe some early documentation efforts:

Data Statements (Bender and Friedman 2018) were inspired by the ways in which participants are described in social science and medical research. As initially conceived, they focused on language datasets and the issues that arise because of the cultural, social, and personal information that is always encoded in language.

Datasheets for Datasets (Gebru et al. 2018, 2021) were inspired by the documentation used for electronics to specify components, tolerances and so forth. The questions posed focus dataset developers' attention on key dataset design issues and the resulting documentation is detailed and intended for use by experts.

Dataset Nutrition Labels (Holland et al. 2018, 2020; Chmielinski et al. 2020), inspired by standardized nutrition labels on prepared foods, detail the construction and contents of a dataset in a brief, standardized format intended to be accessible to both experts and non-experts.

FactSheets (Arnold et al. 2019) include descriptions of training and evaluation data for machine learning systems, in addition to algorithmic concerns. They were modeled after a supplier's declaration of conformity (SDoC) as used in industries such as telecommunications and transportation.

Model Cards for Model Reporting (Mitchell et al. 2019) were intended to complement datasheets by providing a holistic description of a system. Inspired by the TRIPOD statement proposal in medicine, they report trained characteristics of a model such as type, intended use cases, performance variance, and performance measures.

Nutritional Labels for Data and Models (Stoyanovich and Howe 2019), also inspired by labels on prepared foods, explore interpretable displays of automatically calculated information about data and models, to provide insight into the production processes behind machine learning models.

Data Cards (McMillan-Major et al. 2021), drawing on both data statements and datasheets, are specialized to situate particular datasets within a data and tool ecosystem, as well as to provide descriptions of the dataset creation, linguistic characteristics, and potential implications for use in models.

BIBLIOGRAPHY

Arnold, Matthew, 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

Bender, Emily M., 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, 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, Emily M., Batya Friedman, and Angelina McMillan-Major. 2021. *A Guide for Writing Data Statements for Natural Language Processing*. https://techpolicylab.uw.edu/wp-content/uploads/2021/11/Data_Statements_Guide_V2.pdf

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge: Cambridge University Press

Bruijn, Eveline, and Gail Whiteman. 2010. "That Which Doesn't Break Us: Identity Work by Local Indigenous 'Stakeholders.'" *Journal of Business Ethics* 96 *(*3): 479–95. http://www.jstor.org/stable/40863837

Chmielinski, Kasia S., Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2020. "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence." In *NeurIPS 2020 Workshop on Dataset Curation and Security.* http://securedata.lol/camera_ready/26.pdf

Christensen, Heidi, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. 2012. "A Comparative Study of Adaptive, Automatic Recognition of Disordered Speech." In *Proc. Interspeech 2012,* 1776-1779. https://doi.org/10.21437/Interspeech.2012-484

Couillault, Alain, Karën Fort, Gilles Adda, and Hugues de Mazancourt. 2014. "Evaluating corpora documentation with regards to the Ethics and Big Data Charter." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 4225–4229, Reykjavik, Iceland. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/424_Paper.pdf

Derczynski, Leon, Kalina Bontcheva, and Ian Roberts. 2016. "Broad Twitter Corpus: A Diverse Named Entity Recognition Resource." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* The COLING 2016 Organizing Committee, Osaka, Japan, 1169–1179. https://aclanthology.org/C16-1111

Devinney, Hannah, Jenny Björklund, and Henrik Björklund. 2022. "Theories of "Gender" in NLP Bias Research." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22).* Association for Computing Machinery, New York, NY, USA, 2083–2102. https://doi.org/10.1145/3531146.3534627

Eckert, Penelope and John R. Rickford, eds. 2001. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press

Ellis, Rod. 1994. *The Study of Second Language Acquisition.* Oxford: Oxford University Press

Friedman, Batya and David Hendry. 2012. "The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12).* Association for Computing Machinery, New York, NY, USA, 1145–1148. https://doi.org/10.1145/2207676.2208562

BIBLIOGRAPHY

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. "Datasheets for Datasets." In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden. https://www.fatml.org/media/documents/datasheets_for_datasets.pdf

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for datasets." *Communications of the ACM* 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards." In *Data Protection and Privacy*, edited by Dara Hallinan, Ronald Leenes, Serge Gutwirth, Paul De Hert, 1-25. Bloomsbury Publishing

Kusters, Annelies and Ceil Lucas. 2022. "Emergence and evolutions: Introducing sign language sociolinguistics." In *Journal of Sociolinguistics*, 26(1):84–98

Kroskrity, Paul V. 2005. "Language Ideologies." In *A Companion to Linguistic Anthropology*, edited by Alessandro Duranti, 496–517. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470996522.ch22

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics

Larson, Brian. 2017. "Gender as a Variable in Natural-Language Processing: Ethical Considerations." In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 1–11. https://doi.org/10.18653/v1/W17-1601

McMillan-Major, Angelina. 2023. "Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities." PhD diss., University of Washington

McMillan-Major, Angelina, Emily M. Bender, and Batya Friedman. 2024. "Data Statements: From Technical Concept to Community Practice." In *ACM Journal on Responsible Computing , 1(1): 1-17*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3594737

McMillan-Major, Angelina, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. "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 (GEM2021)*. Association for Computational Linguistics, Online,121–135. https://doi.org/10.18653/v1/2021.gem-1.11

Mitchell, Margaret, 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

Nicolao, Mauro, Heidi Christensen, Stuart Cunningham, Phil Green, and Thomas Hain. 2016. "A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia,1993–1997. https://aclanthology.org/L16-1315

Quinto-Pozos, David. 2008. "Sign Language Contact and Interference: ASL and LSM." In *Language in Society, 37*(2):161–189. https://doi.org/10.1017/S0047404508080251

BIBLIOGRAPHY

Sim, Kate, Andrew Brown, and Amelia Hassoun. 2021. "Thinking Through and Writing About Research Ethics Beyond 'Broader Impact'." *CoRR* abs/2104.08205 (2021). https://arxiv.org/abs/2104.08205

Squizzero, Robert, Martin Horst, Alicia Beckford Wassink, Alex Panicacci, Monica Jensen, Anna Kristina Moroz, Kirby Conrod, and Emily M. Bender. 2021. "Collecting and using race and ethnicity information in linguistic studies." In *University of Washington Working Papers in Linguistics*. https://digital.lib.washington.edu/researchworks/handle/1773/48570

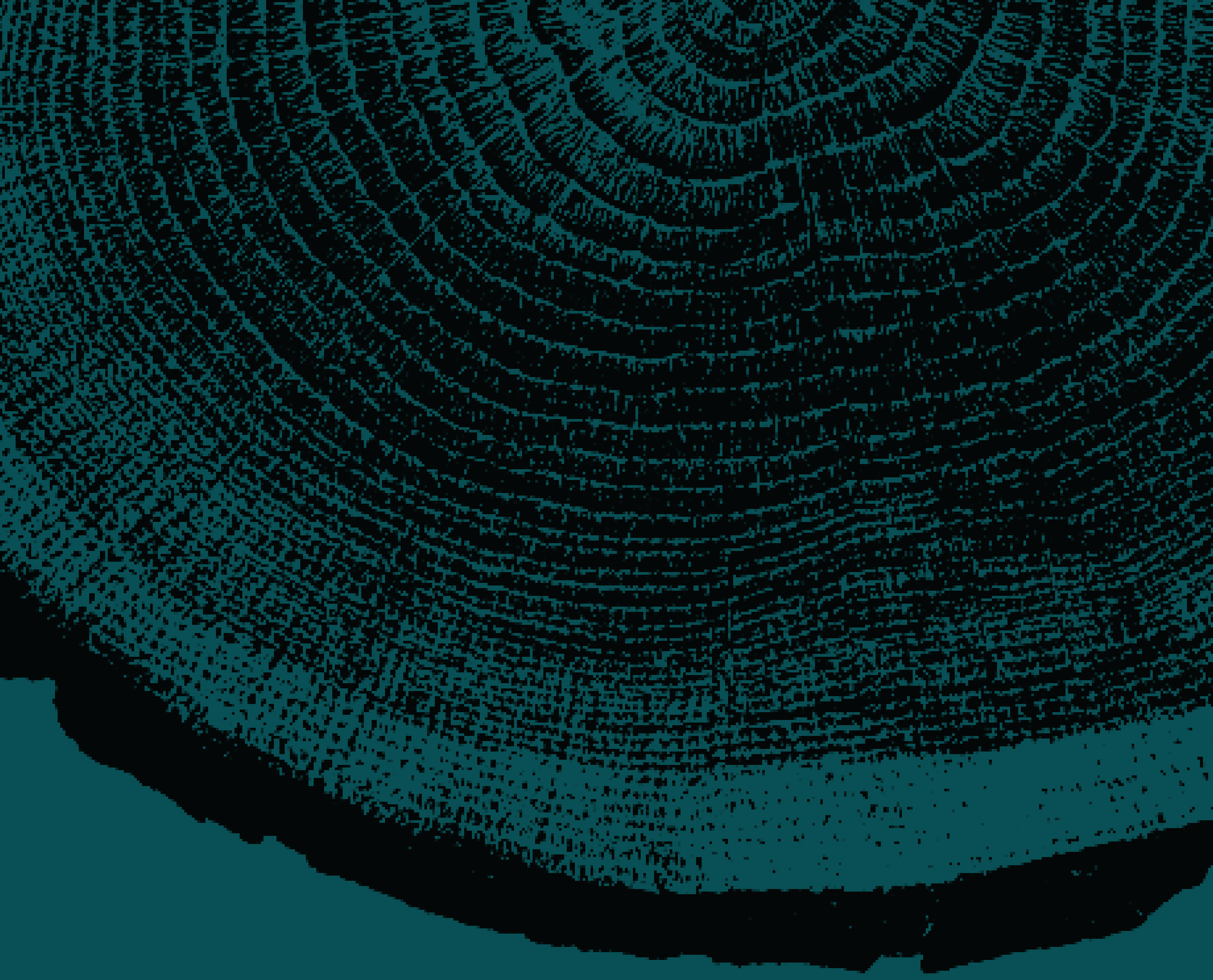
Stoyanovich, Julia and Bill Howe. 2019. "Nutritional labels for data and models." In *A Quarterly Bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 42, 3 (2019). http://sites.computer.org/debull/A19sept/p13.pdf

Talat, Zeerak. 2016. "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618

Tyrone, Martha E. 2014. "7. Sign Dysarthria: A Speech Disorder in Signed Language." In *Multilingual Aspects of Signed Language Communication and Disorder*, edited by David Quinto-Pozos, 161-184. Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781783091317-010

LINKS TO RESOURCES

Page 2	Tech Policy Lab - https://techpolicylab.uw.edu/
	Value Sensitive Design Lab - https://vsdesign.org/
	Creative Commons Attribution No Derivatives License - https://creativecommons.org/licenses/by-nd/4.0
Page 6	Schema Version 1 (Bender and Friedman, 2018) - https://aclanthology.org/Q18-1041/
	2020 LREC workshop - https://techpolicylab.uw.edu/events/event/data-statements-for-nlp/
	Schema Version 2 Guide (Bender, Friedman, and McMillan-Major, 2021) - https://techpolicylab.uw.edu/wp-content/uploads/2021/11/Data_Statements_Guide_V2.pdf
	Schema Version 2 journal paper (McMillan-Major, Bender, and Friedman, 2024) - https://dl.acm.org/doi/10.1145/3594737
	McMillan-Major (2023) - https://digital.lib.washington.edu/researchworks/handle/1773/50854
Page 7	C3DAR guide - https://digital.lib.washington.edu/researchworks/handle/1773/50585
Page 16	BCP-47 - https://www.rfc-editor.org/rfc/bcp/bcp47.txt
Page 27	Envisioning Cards - http://www.envisioningcards.com
	Lifecourse Checklist - https://docs.google.com/document/d/1uODpC40TQbD3VKjaorzSXY9Qc-Z9PB2qBf4SrwNiyOw/edit
Page 31	CC0 license - https://creativecommons.org/public-domain/cc0/
	Traditional Knowledge (TK) Labels - https://localcontexts.org/labels/traditional-knowledge-labels/
	TK Licenses - https://localcontexts.org/licenses/traditional-knowledge-licenses/
	Kaitiakitanga License - https://github.com/TeHikuMedia/Kaitiakitanga-License
	DGS Corpus license conditions - https://www.sign-lang.uni-hamburg.de/meinedgs/ling/license_en.html
Page 32	Linguistic Data Consortium - https://www.ldc.upenn.edu/data-management/using/licensing
	American National Corpus - https://catalog.ldc.upenn.edu/LDC2005T35



TECH POLICY LAB

VALUE SENSITIVE DESIGN LAB

Seattle, WA 2024