DATA
STATEMENTS

A GUIDE FOR WRITING DATA STATEMENTS

FOR NATURAL LANGUAGE PROCESSING

EMILY M. BENDER
BATYA FRIEDMAN
ANGELINA McMILLAN-MAJOR

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.

ACKNOWLEDGMENTS

Data statements were developed at the University of Washington by faculty and students from the Department of Linguistics, Information School, Tech Policy Lab, and Value Sensitive Design Lab. This work was supported financially by the UW Tech Policy Lab and the Frances and Howard Nostrand Endowed Professorship. We gratefully acknowledge the intellectual contributions and support of Zeerak Talat and Leon Derczynski as well as the LREC workshop participants including Luciana Benotti, Bonaventure F. P. Dossou, Chris Emezue, Itziar Gonzalez-Dios, Amy Isard, Neelam Pirbhai-Jetha, Surangika Ranathunga, Beatrice Savoldi, Marc Schulder, and many others.

Visual design: Elias Greendorfer

LICENSE

These materials can be reused for any purpose, including commercially; however, the materials cannot be shared with others in adapted form, and credit must be provided. They are provided under a Creative Commons, Attribution-No Derivatives License. When using these materials, please attribute them to the University of Washington.

DATA STATEMENTS | FOR NATURAL LANGUAGE PROCESSING

TABLE OF CONTENTS

Introduction 4

How to use this Guide 5

General Best Practices 6

Key Terms 8

Schema

- 1 Header 9
- 2 Executive Summary 10
- 3 Curation Rationale 11
- 4 Documentation for Source Datasets 12
- 5 Language Varieties 13
- 6 Speaker Demographic 14
- 7 Annotator Demographic 16
- 8 Speech Situation and Text Characteristics 18
- 9 Preprocessing and Data Formatting 19
- 10 Capture Quality 20
- 11 Limitations 21
- 12 Metadata 22
- 13 Disclosure and Ethical Review 23
- 14 Other 24
- 15 Glossary 25

Appendix A: Conversion Table – Schema Version 1 to Version 2 26

Appendix B: Related Documentation Toolkits 27

Bibliography 28

DATA STATEMENTS | FOR NATURAL LANGUAGE PROCESSING

INTRODUCTION

Data statements provide essential information about the characteristics of datasets, including but not limited to the curation rationale and data sources. The information contained in data statements can be used to help (1) mitigate the harms caused by bias in the dataset (such as a mismatch between training datasets and contexts where systems are deployed) and (2) create a more inclusive data catalog, by identifying gaps. While first developed with language data types, data statements could be produced for a wide range of data types with adjustments 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. With the goals of supporting broader uptake and learning how to make data statements a suitable practice across different research and institutional contexts, in 2020 Emily M. Bender, Batya Friedman, and Angelina McMillan-Major organized a workshop at the 12th Language Resources and Evaluation Conference. The results of this workshop led to an updated schema (Version 2), a set of best practices, and this guide all released in 2021.

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 speech context, speaker 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.

HOW TO USE THIS GUIDE

This document is intended to be used as a guide for writing a data statement. If you are in the process of curating your data, we recommend familiarizing yourself with the content of the guide for ideas about what information you'll need to collect as you create your dataset.

Before starting on any writing, first read the general best practices and key terms. The general best practices include suggestions for how to prepare the information for your dataset to make the writing process easier. The key terms provide definitions for technical vocabulary that are used throughout the guide.

Schema. The schema elements can be written in any order. 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.

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. Consider using the data statement elements as a checklist for dataset design.

4. 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.

5. For crafting your data statement, we recommend using an interview format with an external partner (e.g., someone not involved in the project). This is both fun and instructive. In effect, the external partner treats each data statement 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 data statement.

6. When using technical terms, make use of 15 Glossary.

7. 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 a data statement; clearly indicate what is missing and provide what information you can.

8. For datasets with extensive documentation outside the data statement (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).

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

10. If the content of the dataset contains materials that could be a trigger for trauma, we recommend making a note of this in either 3 Curation Rationale or 14 Other.

GENERAL BEST PRACTICES CONTINUED

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

12. Once drafted, review your data statement for words or phrases used to describe speakers or their language varieties that might be experienced as diminishing and make revisions as appropriate.

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

14. Publish the data statements in the language(s) of the dataset, in addition to any languages of broader communication (such as English).

15. Provide the data statement together with the dataset. This is the canonical location for the most up to date version of the data statement. A link to the data statement along with 2 Executive Summary 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 data statement as an appendix along with a pointer to where updated versions of the data statement may be found.

16. For datasets that are not publicly available (e.g., those containing non-anonymized health information or proprietary data), whenever possible make the data statement 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.

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

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

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

Language data refers to spoken, written or signed utterances.

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.

Speaker 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 or text in it.

Speech refers to linguistic activity (i.e., the production of spoken, signed or written language).

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

Text refers to a sequence of language data.

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 and authorship for the various documents pertaining to a dataset. Such information is particularly important when the author and source of the data statement differs from the author and source of the dataset, or when different versions of the data statement have different authors and sources.

What

The header should include the following:

- Dataset Title
- Dataset Curator(s) [name, affiliation]
- Dataset Version [version, date]
- Dataset Citation and, if available, DOI
- Data Statement Author(s) [name, affiliation]
- Data Statement Version [version, date]
- Data Statement Citation
- 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. The following is an example data statement citation:

Gonzalez-Dios, Itziar. (2021). *Data Statement for the Corpus of Basque Simplified Texts*. Version 2. University of the Basque Country (UPV/EHU). http://www.ixa.eus/node/13302

Consider web accessibility and the longevity of data statement location (e.g., university archives or ACM digital library).

SCHEMA

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

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 the 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(s), 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 executive summary 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).

SCHEMA

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

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 answer questions including: Why was this dataset created? What is the task or research question the dataset is intended to address? Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection? 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 speak to each subcategory.

If the dataset involves 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 writing 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.

SCHEMA

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

4 DOCUMENTATION FOR SOURCE DATASETS

Why

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

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 dataset show the user where to go look for further information, especially for the curation rationale of the source dataset.

What

For datasets built out of 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 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 data set; 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.

SCHEMA

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

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).

Best Practices

Describe all language varieties represented in the dataset. For translation datasets, this would include both sides of the bitext. If the language variety used for annotations differs from the language variety of the source data, again document both.

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 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).

SCHEMA

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

6 SPEAKER DEMOGRAPHIC

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

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 catalogue 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.

What All of the speaker groups represented in the dataset should be characterized with a prose description. Demographic categories are culture-specific and locally appropriate categories and definitions should be used. Suggested specifications include:

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

SCHEMA

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

6 SPEAKER DEMOGRAPHIC CONTINUED

Best Practices

Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).

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 speakers 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 speakers in the training data for the automatic generation system.

If the dataset contains both found and elicited data, provide separate speaker 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 speakers to help protect their privacy. 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.

SCHEMA

1 HEADER
2 EXECUTIVE SUMMARY
3 CURATION RATIONALE
4 DOCUMENTATION FOR SOURCE DATASETS
5 LANGUAGE VARIETIES
6 SPEAKER DEMOGRAPHIC
7 ANNOTATOR DEMOGRAPHIC
8 SPEECH SITUATION AND TEXT CHARACTERISTICS
9 PREPROCESSING AND DATA FORMATTING
10 CAPTURE QUALITY
11 LIMITATIONS
12 METADATA
13 DISCLOSURE AND ETHICAL REVIEW
14 OTHER
15 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 speakers or, if that is not feasible, in identifying demographic gaps between annotators and speakers, 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 developed 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. 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

Best Practices

Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).

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

SCHEMA

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

7 ANNOTATOR DEMOGRAPHIC CONTINUED

Best Practices

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. When the number of annotators and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect annotator privacy.

SCHEMA

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

8 SPEECH SITUATION AND TEXT CHARACTERISTICS

Characteristics of the speech 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. The time, place, and cultural context allow for deeper understanding of how the texts collected relate to their historical moment. Both genre and topic also influence the vocabulary and structural characteristics of texts (Biber, 1995).

Why

For dataset creators, a clear conception of the targeted speech 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 speech 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 speech situation at a future time.

What

A description of the speech situation in which the linguistic production occurred 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. 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
- Speakers' intended audience
- Genre (e.g., newswire vs. social media)
- Topic (e.g., entertainment vs. natural disaster)
- Non-linguistic context (e.g., photos speakers were all looking at; a game participants are playing)
- Additional details about the cultural context (optional)

Best Practices

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

SCHEMA

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

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 which, if any, tools were 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.

SCHEMA

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

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).

SCHEMA

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

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 that could not be 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.

SCHEMA

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

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:

- License: Link to the license/copyright permissions for use or modification of the dataset
- Annotation Guidelines: Link to the published or online guidelines that annotators used to annotate the data
- Annotation Process: Link to documentation providing metadata about the annotation process, including protections for annotator anonymity, how annotators were compensated, and which aspects of the annotation were produced automatically
- Dataset Quality Metrics: Metrics for inter-annotator agreement and/or other numerical scores of dataset quality
- Errata: Link to the list of known errors and how to report additional ones

Best Practices

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

Include a link to the licensing/copyright permissions for both the dataset itself and the data curated to create the dataset.

SCHEMA

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

13 DISCLOSURE AND ETHICAL REVIEW

Why

For dataset creators, a clear conception of the terms of 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.

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 institution (e.g., IRB) should be provided. In addition, include: a brief description of any consent process used; if speakers or annotators were compensated, how compensation rates were determined; any access restrictions to the data; 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 “An institutional ethics review process was not accessible at the time of dataset creation."

SCHEMA

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

14 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.

SCHEMA

1 HEADER

2 EXECUTIVE SUMMARY

3 CURATION RATIONALE

4 DOCUMENTATION FOR SOURCE DATASETS

5 LANGUAGE VARIETIES

6 SPEAKER DEMOGRAPHIC

7 ANNOTATOR DEMOGRAPHIC

8 SPEECH SITUATION AND TEXT CHARACTERISTICS

9 PREPROCESSING AND DATA FORMATTING

10 CAPTURE QUALITY

11 LIMITATIONS

12 METADATA

13 DISCLOSURE AND ETHICAL REVIEW

14 OTHER

15 GLOSSARY

15 GLOSSARY

Why

For data statement authors, using technical terms can make it easier to write efficient and precise documentation. Providing definitions for these technical 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.

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.

SCHEMA

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

APPENDIX A: CONVERSION TABLE – SCHEMA VERSION 1 TO VERSION 2

Data statements were first developed and published in 2018. The schema elements and descriptions from 2018 are Version 1. For details, see Bender and Friedman (2018). In 2021 based on feedback from professional and community sources including a workshop held at Language Resources and Evaluation Conference 2020, May 11-13, 2020, the schema elements were updated and a set of best practices developed. Together, these updated schema elements from 2021 are Version 2.

Version 1 Data Statements are valuable and useful as is. However, if you would like to update a Version 1 Data Statement to Version 2, the table and instructions below provide a road map for how to do so. As you update each schema element, please consult this guide for suggested best practices.

The table below contains instructions for how to convert a Version 1 to a Version 2 Data Statement. A summary of the changes are as follows:

- 7 new schema elements
- Merge 2 previous schema elements into 1 new schema element
- Reorder schema elements to better model the flow of dataset use

Version 1	**Version 2**	**Update Instructions**
	1. Header	Add
	2. Executive Summary	Add
A. Curation Rationale	3. Curation Rationale	Update
I. Provenance Appendix	4. Documentation For Source Datasets	Rename and update
B. Language Variety/Varieties	5. Language Varieties	Rename and update
C. Speaker Demographic	6. Speaker Demographic	Update
D. Annotator Demographic	7. Annotator	Update
E. Speech Situation and F. Text Characteristics	8. Speech Situation and Text Characteristics	Merge, rename, and update
	9. Preprocessing and Data Formatting	Add
G. Recording Quality	10. Capture Quality	Rename and update
	11. Limitations	Add
	12. Metadata	Add
	13. Disclosure and Ethical Review	Add
H. Other	14. Other	Update
	15. Glossary	Add

APPENDIX B: RELATED DOCUMENTATION TOOLKITS

Beginning around 2017-2018, in response to a wide range of potential harms from applying pattern recognition ("AI") at scale, 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. Some early documentation efforts include:

- ***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; Gebru et al. 2020) 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; 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

M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu,D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 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

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

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

Kasia S. Chmielinski, 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*.

Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. 2012. A Comparative Study of Adaptive, Automatic Recognition of Disordered Speech. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Leon Derczynski, 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

Penelope Eckert and John R. Rickford (Eds.). 2001. *Style and Sociolinguistic Variation.* Cambridge University Press, Cambridge.

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

Batya Friedman and David Hendry. 2012. *The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations.* Association for Computing Machinery, New York, NY, USA, 1145–1148. https://doi.org/10.1145/2207676.2208562

Timnit Gebru, 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.

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

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). arXiv:1805.03677 http://arxiv.org/abs/1805.03677

Paul V. Kroskrity. 2005. *Language Ideologies*. John Wiley Sons, Ltd, Chapter 22, 496–517. https://doi.org/10.1002/9780470996522.ch22

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

Brian Larson. 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

BIBLIOGRAPHY CONTINUED

Angelina McMillan-Major, 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

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

Mauro Nicolao, 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

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

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).

Zeerak Talat. 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