Slides: bit.ly/EMB-LAW-23

# Data Statements: Empowering Ethical Practice and Accountability through Dataset Documentation

Emily M. Bender - with Batya Friedman and Angelina McMillan-Major
University of Washington

*LAW XVII @ ACL 2023*
*Toronto July 13, 2023*

# Key points

- Dataset documentation is key to enabling ethical practice
- Dataset documentation toolkits exist!
- Data statements (now in v2) are one such toolkit, specialized for natural language datasets
  - With a how to guide + templates!
- Developing effective toolkits requires community engagement

Slides: bit.ly/EMB-LAW-23

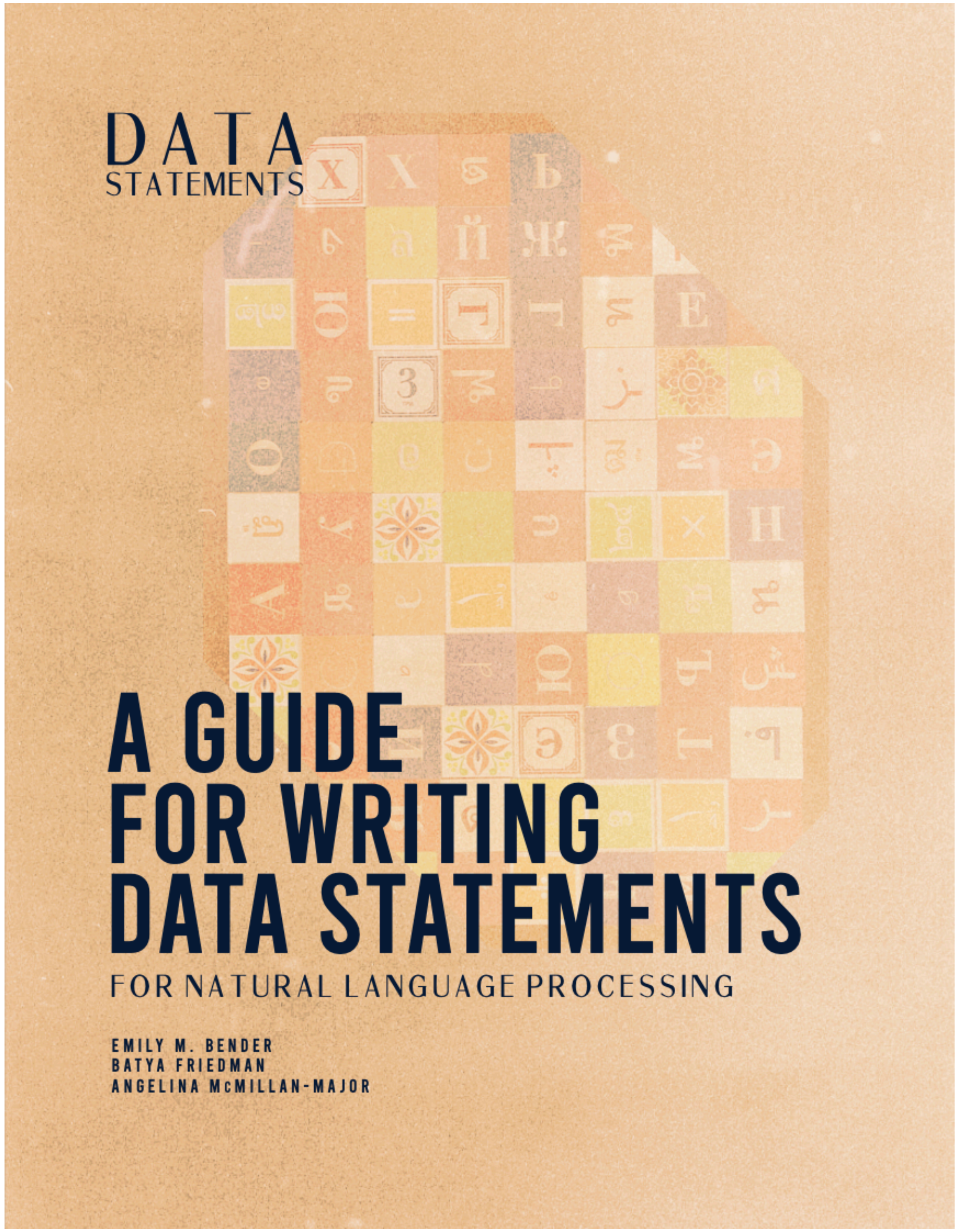
# This talk draws on

- Bender, Emily M. and Batya Friedman. 2018. Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics 6:587-604.

- Bender, Emily M., Batya Friedman, and Angelina McMillan-Major. 2021. A Guide for Writing Data Statements for Natural Language Processing.

- McMillan-Major, Angelina, Emily M. Bender and Batya Friedman. 2023. Data Statements: From Technical Concept to Community Practice, ACM Journal on Responsible Computing.

# Data statements schema (v2) preview



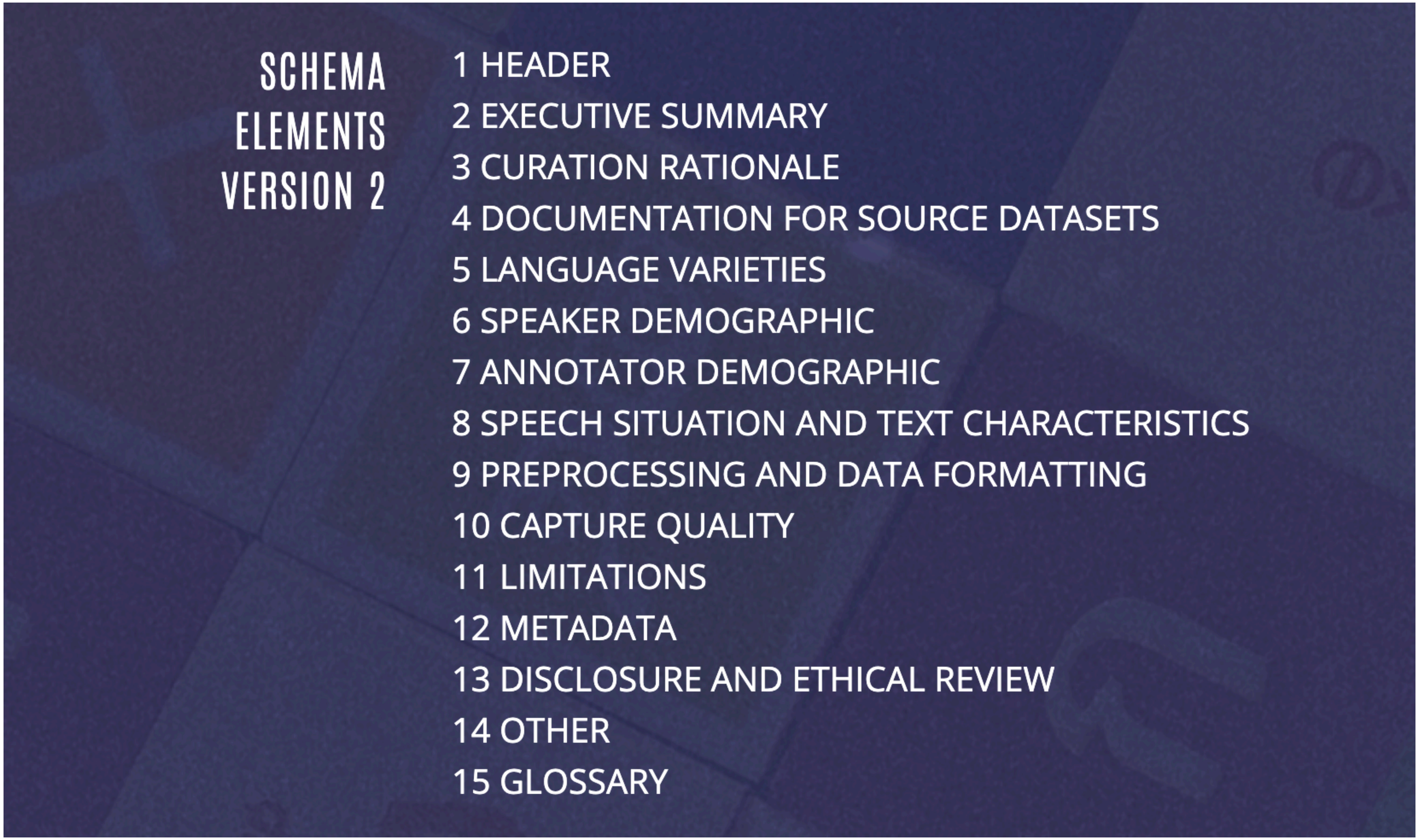
DATA
STATEMENTS

A GUIDE
FOR WRITING
DATA STATEMENTS

FOR NATURAL LANGUAGE PROCESSING

EMILY M. BENDER
BATYA FRIEDMAN
ANGELINA McMILLAN-MAJOR

# Data statements schema (v2) preview

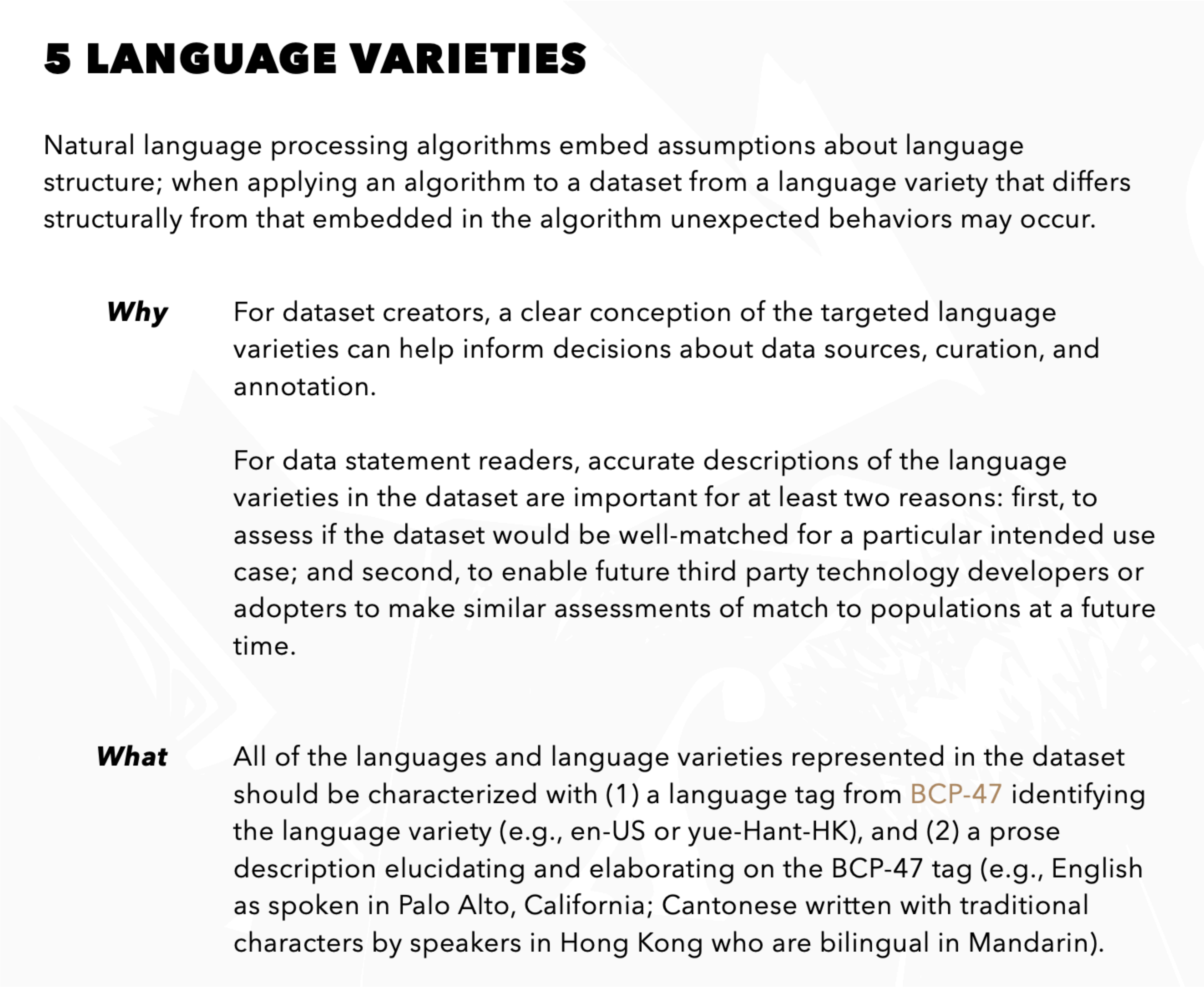


SCHEMA ELEMENTS VERSION 2

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

https://techpolicylab.uw.edu/data-statements/

# Data statements schema (v2) preview



## 5 LANGUAGE VARIETIES

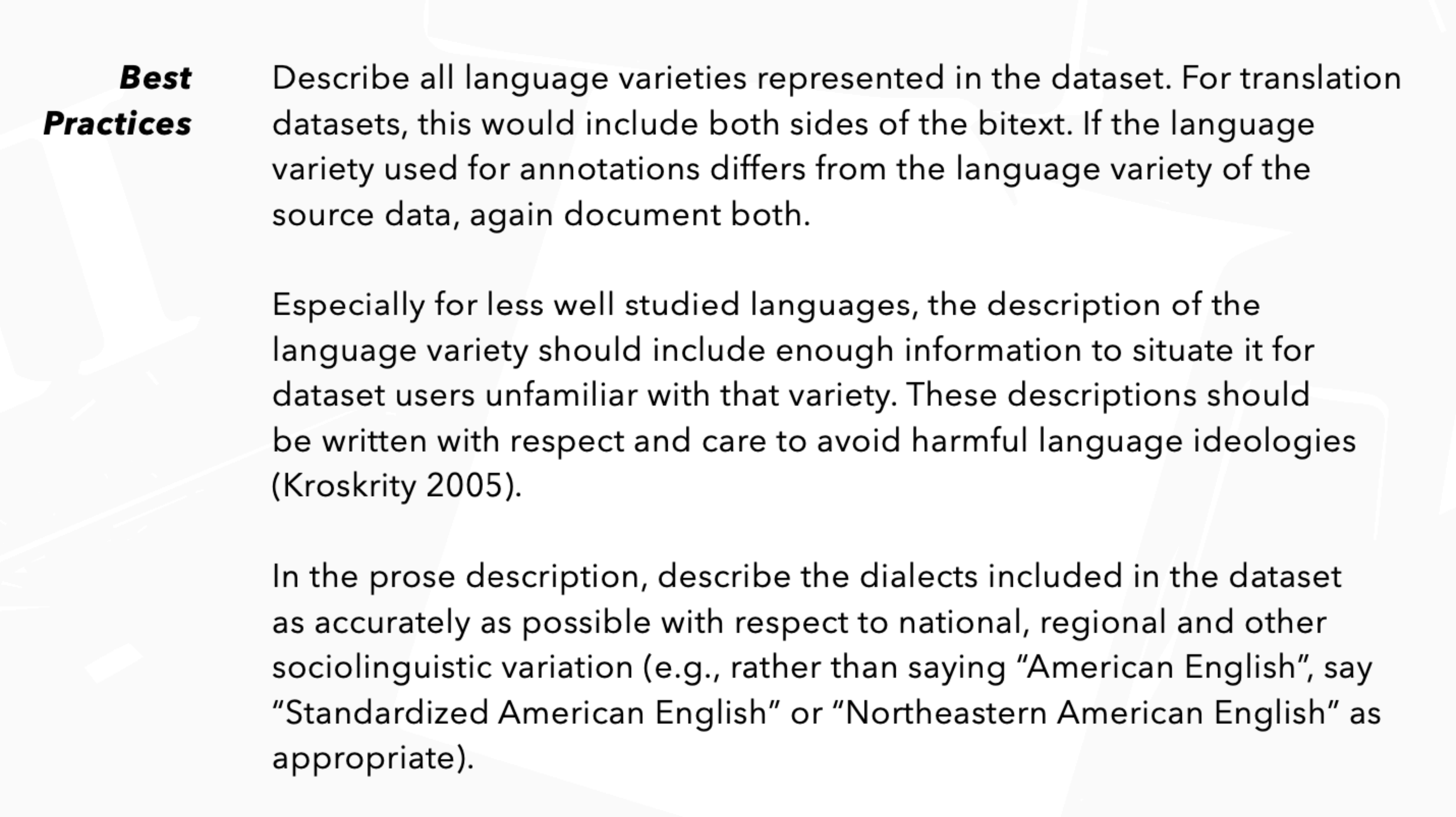
Natural language processing algorithms embed assumptions about language structure; when applying an algorithm to a dataset from a language variety that differs structurally from that embedded in the algorithm unexpected behaviors may occur.

***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).

# Data statements schema (v2) preview



***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).

# Overview

- Big picture: dataset documentation
- Early history of data statements
- Data statements v2, workshop, writing guide
- Example elements
- Future directions

# Something in the air in 2017…

- Convergent ideas from many groups

- Ethical deployment of pattern matching at scale depends on clear and thorough documentation of source datasets

# Typologizing risks

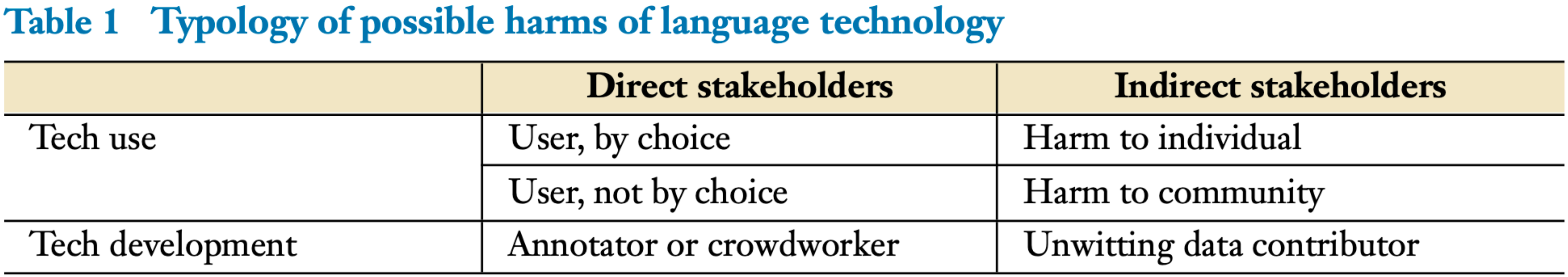
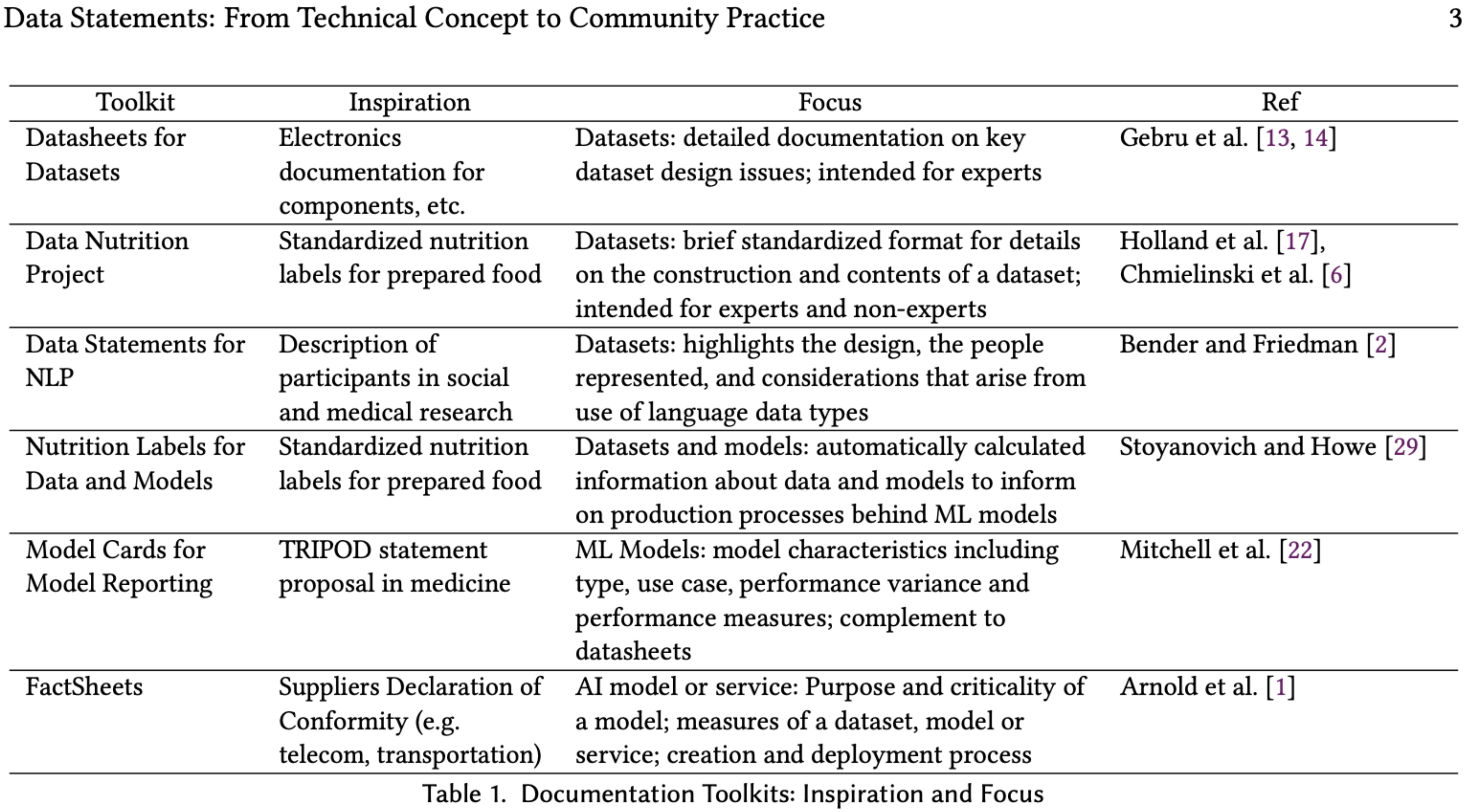


**Table 1 Typology of possible harms of language technology**

| | Direct stakeholders | Indirect stakeholders |
|---|---|---|
| Tech use | User, by choice | Harm to individual |
| | User, not by choice | Harm to community |
| Tech development | Annotator or crowdworker | Unwitting data contributor |

- From D’Arcy & Bender 2023 “Ethics in Linguistics” *Annual Review of Linguistics*

- Documentation is not a panacea, but it can empower people to address harms from *emergent bias* (Friedman & Nissenbaum 2011), *representational harms* (Barocas et al 2017), *data theft*, and *exploitative labor practices* (Fort et al 2011)



Data Statements: From Technical Concept to Community Practice 3

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

Table 1. Documentation Toolkits: Inspiration and Focus

(McMillan-Major et al 2023)

# Dataset documentation enables us to ask

- *Researchers*: Over what domain do I expect my results to generalize?
- *Procurers*: Is this system appropriate for the users I anticipate will interact with it?
- *Policymakers*: Are rights respected in the development and deployment of systems?
- *Community activists*: What patterns are being reproduced which adversely affect my community?

# Interlude: Why here (at LAW XVII)?

# Interlude: Why here (at LAW XVII)?

- Data & annotation are at the foundation of our field

# Interlude: Why here (at LAW XVII)?

- Data & annotation are at the foundation of our field
- You are the people who are creating the most carefully curated datasets

# Interlude: Why here (at LAW XVII)?

- Data & annotation are at the foundation of our field
- You are the people who are creating the most carefully curated datasets
- You can set the standards of best practice

# Interlude: Why here (at LAW XVII)?

- Data & annotation are at the foundation of our field
- You are the people who are creating the most carefully curated datasets
- You can set the standards of best practice
- If we don't get this info in at the foundation, we cut off the possibility of ethical practice

# This is still/even more true in the age of LLMs

- Bender, Gebru et al: “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜”

- So, how big is too big?

# This is still/even more true in the age of LLMs

- Bender, Gebru et al: “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜”

- So, how big is too big?

> "Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones. The solution, we propose, is to budget for documentation as part of the planned costs of dataset creation, and only collect as much data as can be thoroughly documented within that budget."

# Data Statements for NLP: Initial steps

- Winter 2017 seminar on “Ethics and NLP” at the University of Washington
- Invited Batya Friedman to do a guest lecture on value sensitive design
- Identified lack of information about dataset contents as a key hurdle to mitigating risks of harm from NLP systems

# Data Statements for NLP: Initial steps

- *Conceptual investigation*: What information would various stakeholders need about datasets (and to what extent is that information collectable)?
    - Data Statements schema version 1
- *Technical investigation*: Create data statements for __ and __
- *Conceptual investigation*: Value scenarios (Nathan et al 2007) imagining positive and negative impacts of data statement

# Data Statements v1 interim conclusions

# Data Statements v1 interim conclusions

As consumers of datasets or products trained with them, NLP researchers, developers and the general public would be well advised to use systems **only** if there is access to the type of information we propose should be included in data statements.

# Data Statements v1 interim conclusions

As consumers of datasets or products trained with them, NLP researchers, developers and the general public would be well advised to use systems **only** if there is access to the type of information we propose should be included in data statements.

An *empirical investigation* is also needed, to explore how data statements would work in practice for a diverse range of practitioners.

# Data Statements workshop “at” LREC 2020

- Virtual event, May 11-13 2020
  - Three two-hour sessions
- 38 participants from 16 countries (including Argentina, Sri Lanka, Mauritius + US and Europe)
- 29 datasets
- 50% senior researchers, 36.8% junior researchers, 13.2% decline to state

# Data Statements workshop “at” LREC 2020

- Goals: Develop data statements, get feedback on schema + best practices
- Day 1 introductions, small group development of first four elements of data statement
    - Homework: finish drafting those sections
- Day 2 small group feedback on drafted sections, development of remaining elements
    - Homework: finish drafting those sections
- Day 3 small group feedback on drafted sections, medium group discussion of schema and best practices

# Analysis of workshop artifacts:
# Data statement worksheets

- Tips & suggestions
  - The worksheet elicited from participants, per element:
    - Feedback/concerns
    - Tips/advice
- Strengths & weaknesses
  - Where did the existing instructions unclear? Unsuited to specific datasets?
  - What creative directions did participants take the schema?

# Analysis of workshop artifacts: Medium-group discussions

- Advice for Developing and/or Writing Good Data Statements
- Additional Elements. What further elements (if any) should a data statement have? Why?
- Uses. What purposes do you see data statements serving? When could they be helpful and for what?
- Possible Harms
- Misuse of Data Statements
- What Content is Hard to Know? (For elicited data, for found data)
- Best Practices
- Anything else?

# Analysis of workshop artifacts: Medium-group discussions

- Advice for Developing and/or Writing Good Data Statements
- Additional Elements. What further elements (if any) should a data statement have? Why?
- Uses. What purposes do you see data statements serving? When could they be helpful and for what?
- Possible Harms
- Misuse of Data Statements
- What Content is Hard to Know? (For elicited data, for found data)
- Best Practices
- Anything else?

=> Draft v2 schema + guide to writing

# Comparison to Datasheets for Datasets (Gebru et al 2018, 2021)

- Another *technical investigation*
- Compared draft v2 schema with datasheets schema
- Mapped datasheets questions to data statements elements

# Comparison to Datasheets for Datasets (Gebru et al 2018, 2021)

- How is data conceptualized?
- Who is writing documentation?
- Who is reading documentation?
- What risks are being mitigated?
- What other purposes are being served?

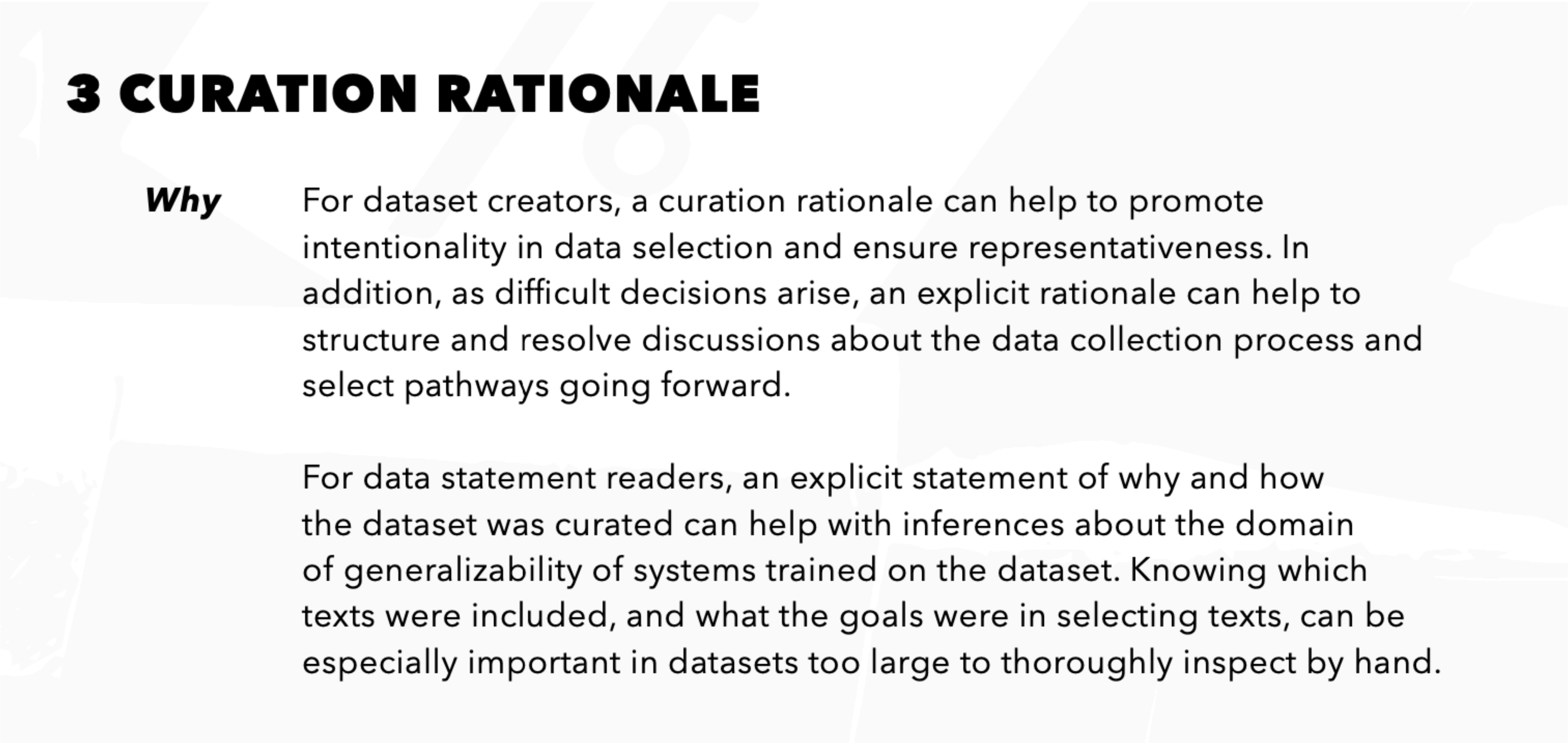
# Comparison to Datasheets for Datasets (Gebru et al 2018, 2021)

- How is data conceptualized?
- Who is writing documentation?
- Who is reading documentation?
- What risks are being mitigated?
- What other purposes are being served?

=> Final v2 schema + guide to writing

| 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<br>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 |

# 3. Curation Rationale

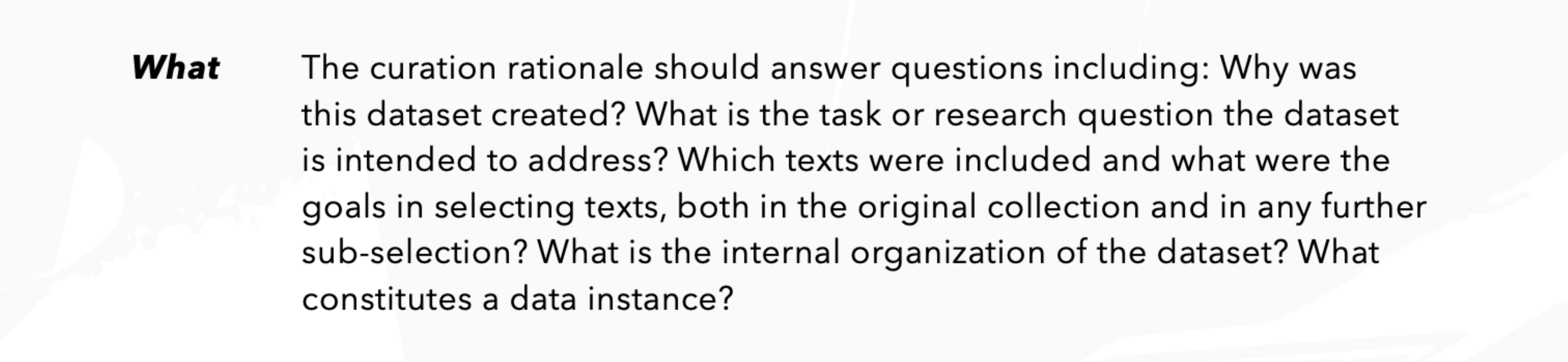


## 3 CURATION RATIONALE

***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.

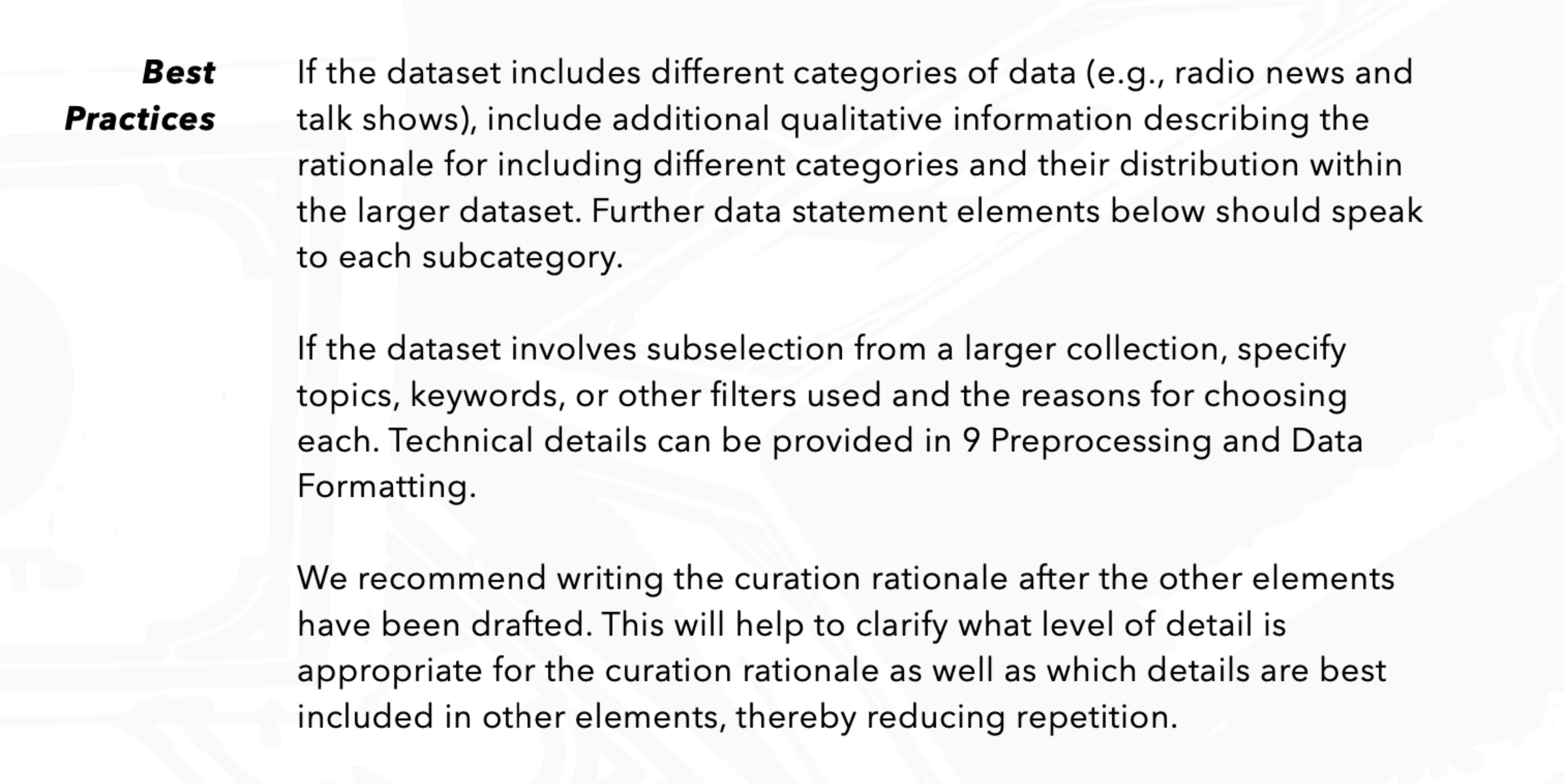
For data statement readers, an explicit statement of why and how the dataset was curated can help with inferences about the domain of generalizability of systems trained on the dataset. Knowing which texts were included, and what the goals were in selecting texts, can be especially important in datasets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

# 3. Curation Rationale



***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?

# 3. Curation Rationale



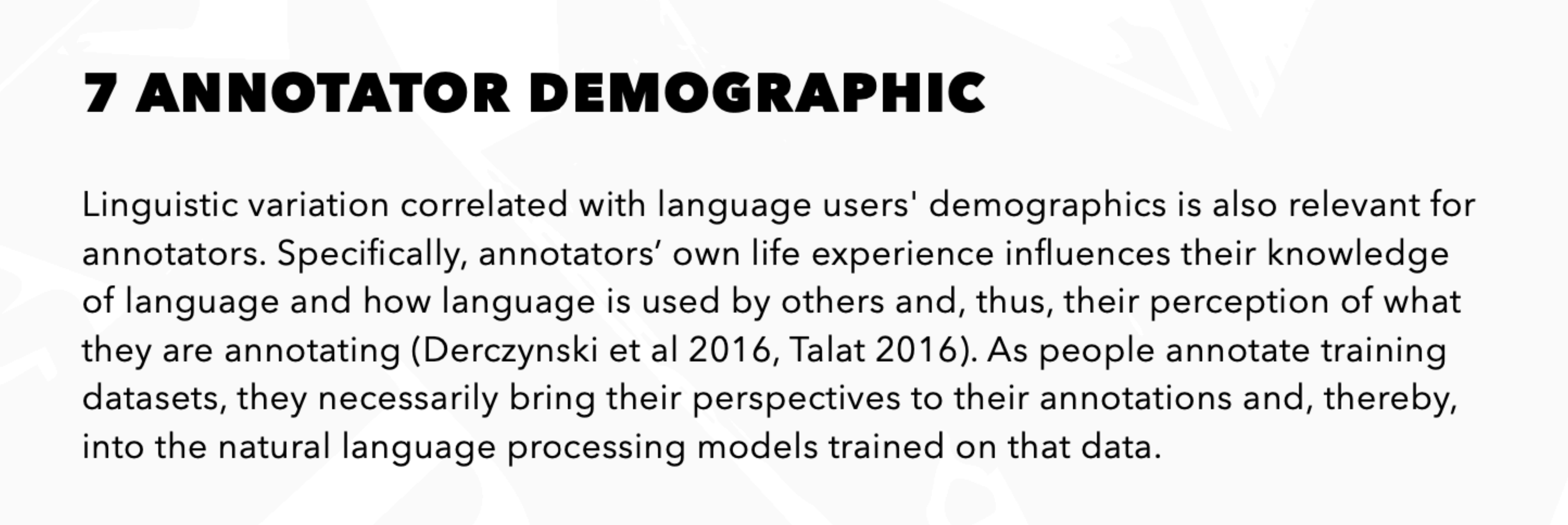
***Best Practices***

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.

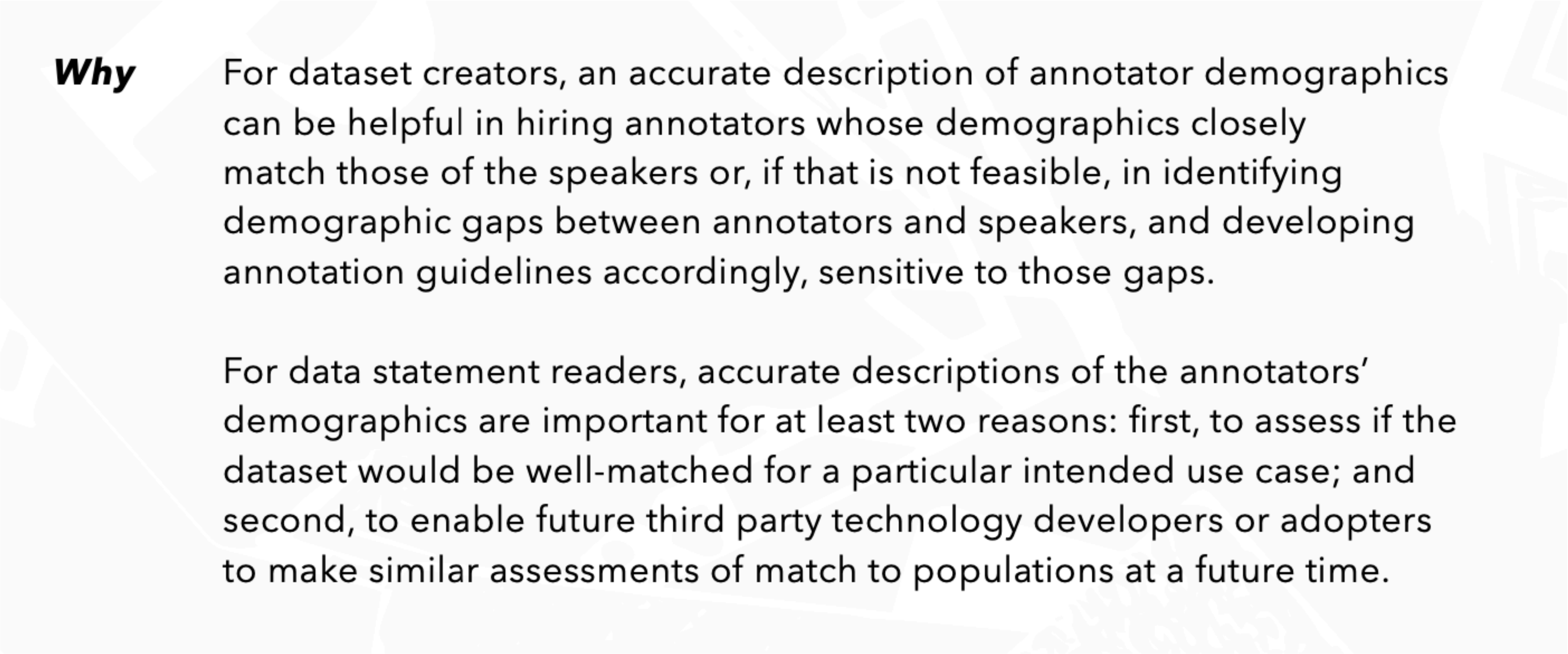
# 7. Annotator Demographic



## 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.

# 7. Annotator Demographic

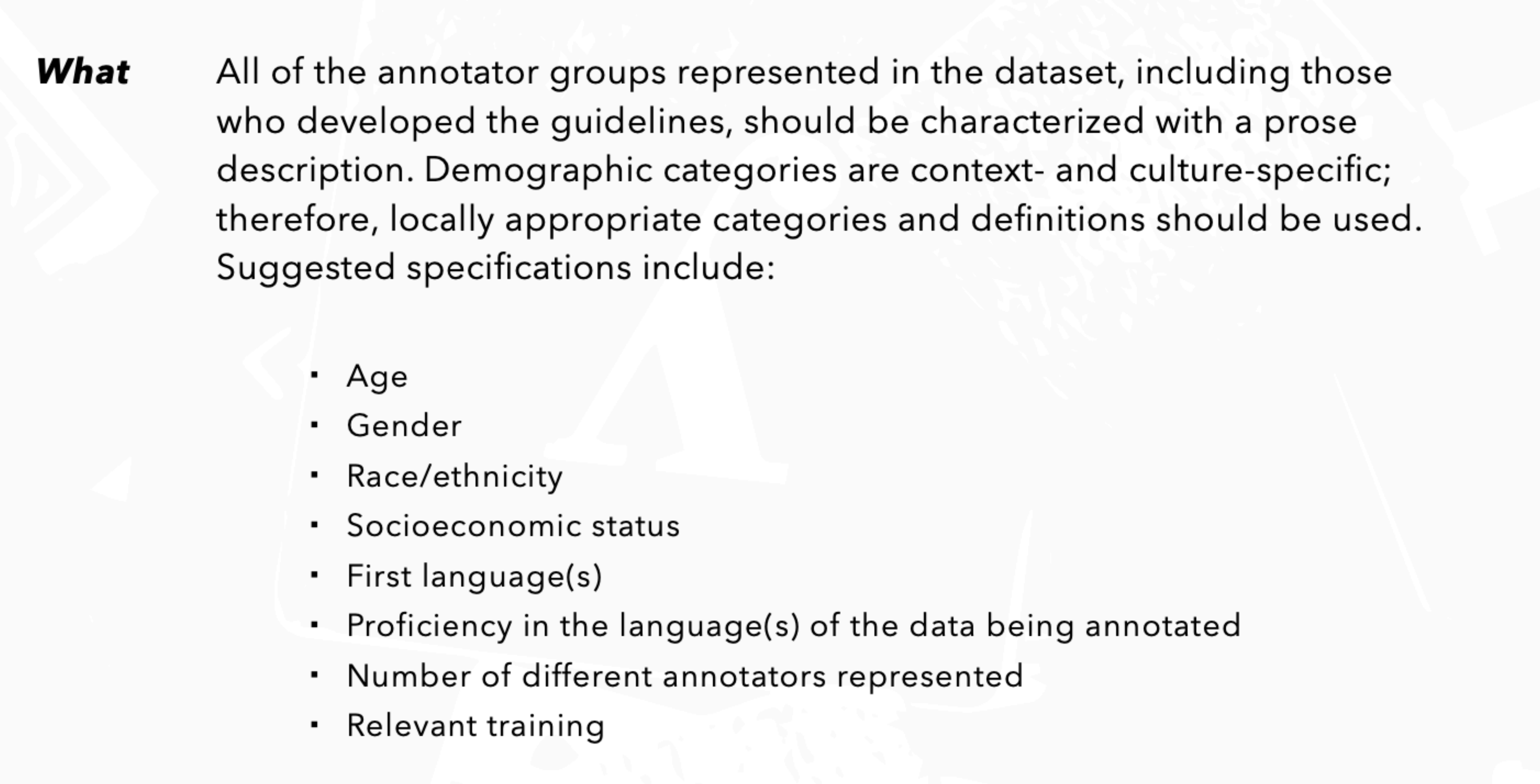


***Why***

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.

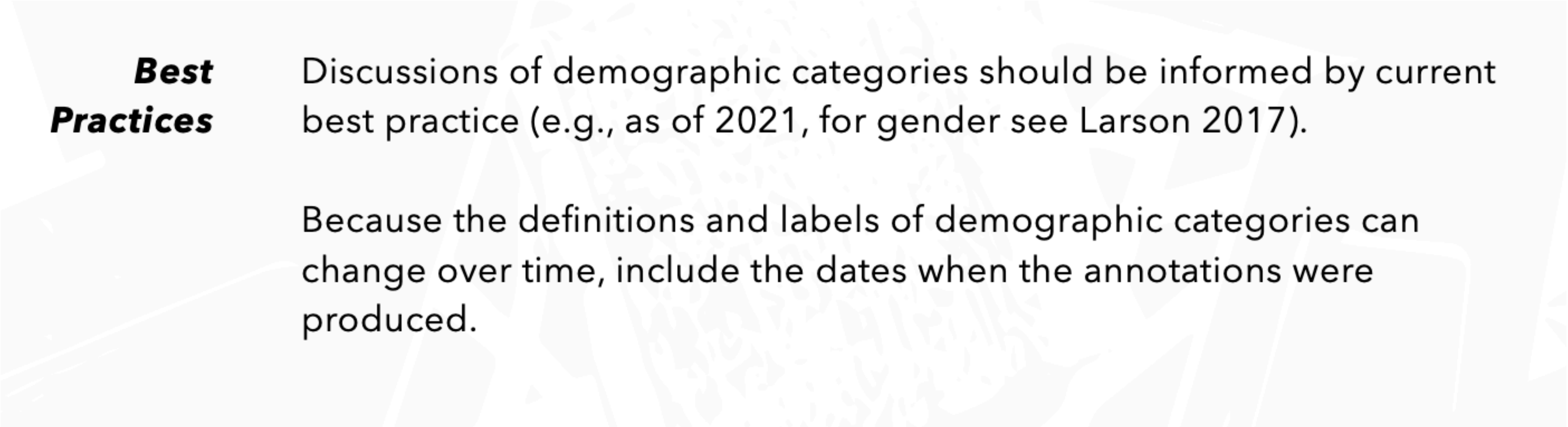
# 7. Annotator Demographic



***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

# 7. Annotator Demographic



***Best Practices***

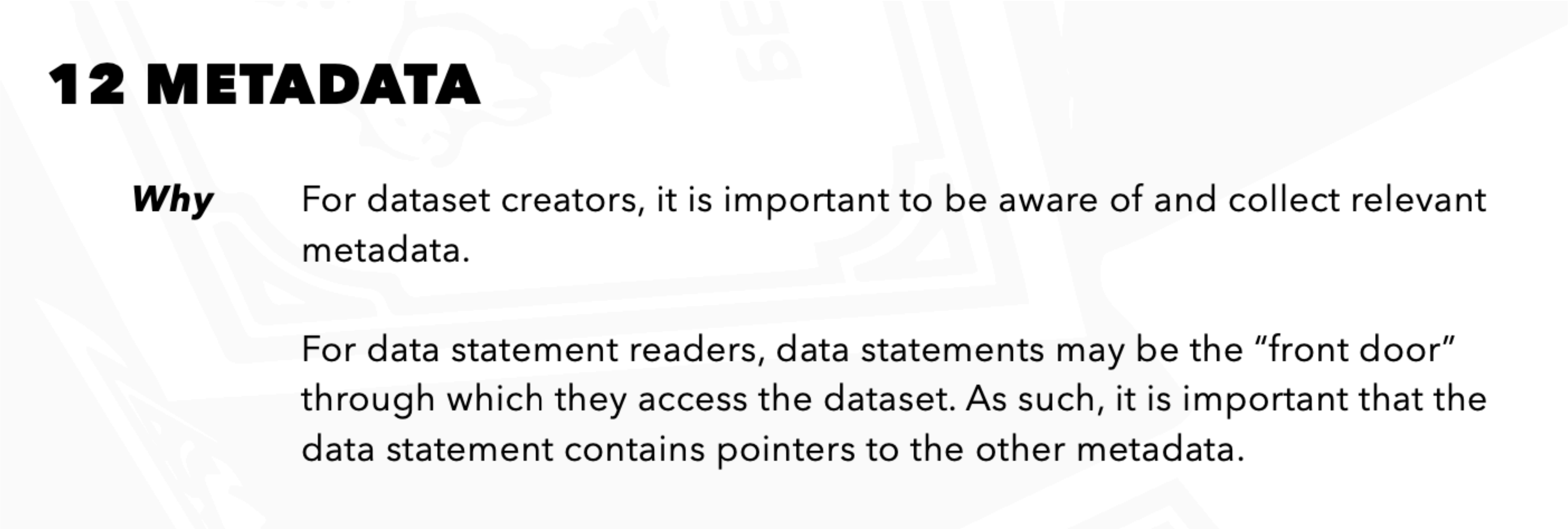
Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).

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

# 7. Annotator Demographic

- Lessons form Tedeschi et al 2023 “What’s the Meaning of Superhuman Performance in Today’s NLU?” (ACL):
- In addition to annotator demographic, we probably also need information about:
    - How annotators were selected
    - Annotator working conditions
- See also: Fort et al 2011 “Amazon Mechanical Turk: Gold mine or coal mine?” (*Computational Linguistics*)

# 12. Metadata

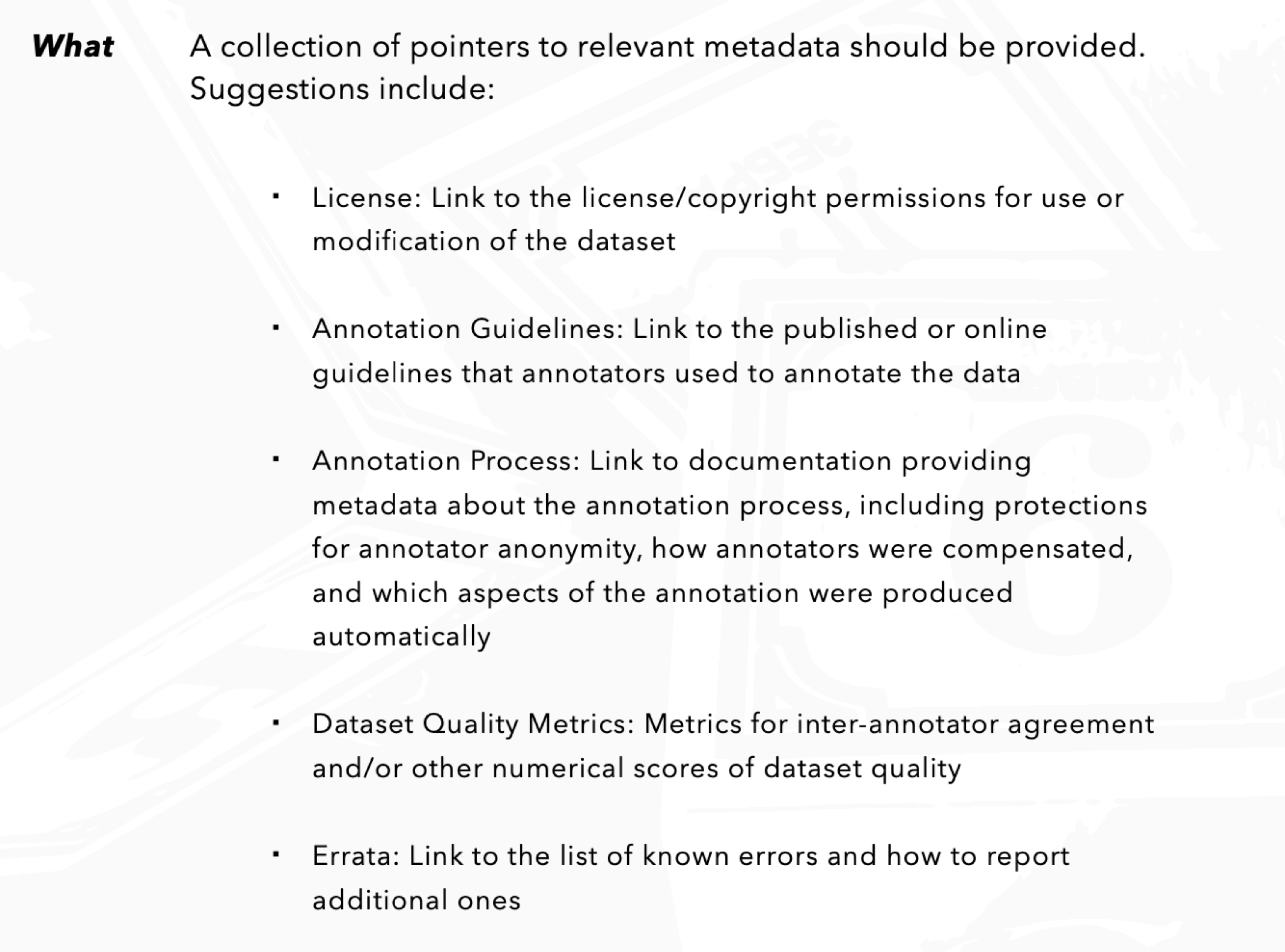


## 12 METADATA

***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.

# 12. 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

# 12. Metadata

**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.

# 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.

# General best practices

- 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.

# General best practices

- 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).

# General best practices

- 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.

- 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.

# General best practices

- 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).

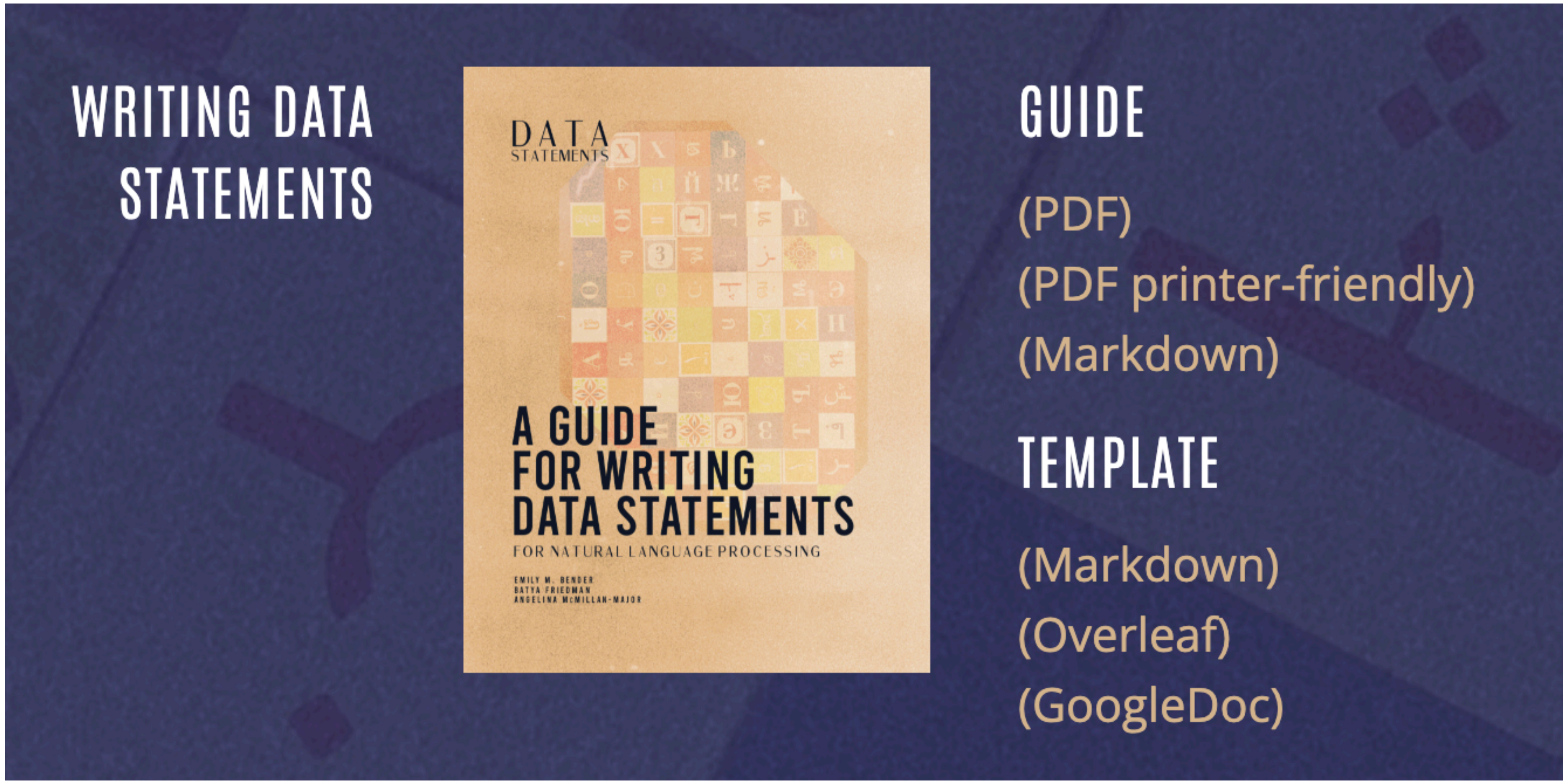
# General best practices

- 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.

# https://techpolicylab.uw.edu/data-statements/

- Guide to writing data statements
- Templates (markdown, overleaf, Google docs)
- Sample data statements



WRITING DATA STATEMENTS

DATA STATEMENTS

A GUIDE FOR WRITING DATA STATEMENTS

FOR NATURAL LANGUAGE PROCESSING

EMILY M. BENDER
BATYA FRIEDMAN
ANGELINA McMILLAN-MAJOR

GUIDE

(PDF)
(PDF printer-friendly)
(Markdown)

TEMPLATE

(Markdown)
(Overleaf)
(GoogleDoc)

# Future directions

- What of this can/should be automated?
  - Process should be of reflective engagement
  - Main audience should be human readers, w/varied relationships to data
- That said, standardization (e.g. BCP-47 codes) supports examination of representativeness of the data catalogue
  - See also Vidgen & Derczynski 2020 “Directions in abusive language training data, a systematic review: Garbage in, garbage out” (*PLOS ONE*)
- Using data statements in planning dataset collection (McMillan-Major, forthcoming)

# Key points

- Dataset documentation is key to enabling ethical practice
- Dataset documentation toolkits exist!
- Data statements (now in v2) are one such toolkit, specialized for natural language datasets
    - With a how to guide + templates!
- Developing effective toolkits requires community engagement

Slides: bit.ly/EMB-LAW-23