

Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources (C3DAR) Toolkit

The C3DAR toolkit is designed to support collaborative dataset curation and documentation between language communities and technical communities. It consists of general best practices, a list of key terms, and 17 schema elements corresponding to key considerations for designing datasets and writing documentation. Each schema element includes the rationale for its inclusion in the schema, its definition, and suggested best practices. By filling out each of the schema elements with a future dataset in mind, the dataset design team can thoroughly discuss plans for the dataset's content, creation process, and publication while also producing an initial draft of the dataset documentation. This process is intended to be iterative, with schema elements being drafted as decisions are made and updated as the project develops.

The C3DAR toolkit was developed by Angelina McMillan-Major, based on the Data Statement Version 2 Schema as prepared by Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. The C3DAR Toolkit and data statements are from the University of Washington. Contact: aymm@uw.edu. This document is licensed as [CC0](https://creativecommons.org/licenses/by/4.0/).

General Best Practices for Collaboration

1. Collaboration requires honesty, respect and care for team members, the community, and the community's history and values (South African San Institute, 2017; Hudson et al., 2010). Be mindful of asymmetrical power relations throughout the project within this historical context and how these interact with community cultural norms (Harris et al., 2009; Ontario Federation of Indigenous Friendship Centres, 2016).
2. Whenever possible, communicate in the language the community prefers. Relying on interpretation services may affect the results of the project and the research team's ability to understand the community's perspective, so it is best if at least one team member is able to communicate in the language of the community.
3. The project should center the needs and understandings of the community. The community should be involved in determining the project goals, methods, and evaluation criteria. The community's knowledge and ways of knowing are valid without reaffirmation via mainstream understandings and analysis (Ontario Federation of Indigenous Friendship Centres, 2016).

4. Be aware of the relevant axes of diversity within the community. Community representatives should reflect the community diversity, which may be uniquely defined depending on what demographic information and personal characteristics are most salient to the community. Be transparent in any recruiting processes. Recruitment of particular community members should be transparent as to why those community members were selected for the role so as not to create mistrust from the community with respect to the project or negatively impact the recruited community member (World Federation of the Deaf Expert Group on Developing Countries, 2016).
5. Allow time for negotiation processes according to community customs as well as for feedback and reviewing processes. While community collaborators may be aware of this difference, communicating about expected time frames of both academic and community processes can help the project members to prepare ahead of time for setbacks or find other ways to make use of time spent waiting (Coeur d'Alene Tribe of Idaho and University of Idaho, 2015).
6. Discuss the benefits that all parties will derive from the dataset, related projects, and the collaboration itself. In particular, the outcomes should include tangible and meaningful benefits to the community that address their self-identified needs. Consider whether commercialization will be allowed on the dataset or products derived from the dataset, and if so, how the benefits and responsibilities of commercialization will be managed (Argumedo et al., 2011).
7. The community and its knowledge should be protected against risks related to the project or resulting from later use of the dataset. The community members may need to be protected from physical and psychological harm, disparagement or disrespect, and confidentiality breaches. Community knowledge may be deemed sensitive and therefore inappropriate to include in any publications or publicly distributed data. Discuss the possible risks and develop mitigation strategies with the community. The implication is not that the collaboration team should be able to foresee all harms, but rather that active measures should be put in place to prevent harms and assess risks.
8. The community should be involved in developing culturally appropriate procedures for ongoing free, prior, informed and educated consent. This may include the language(s) that the procedure will be available in, whether a written version will be available, and how to make the project methods, potential risks and benefits, and confidentiality procedures clear to the community. Avoid assuming that potential benefits are obvious, making exaggerated claims, and understating the potential risks. The community should decide whether this consent is individual or collective.
9. Ownership of the community's knowledge, cultural heritage and data belongs with the community. Copies of the dataset and any other products should be returned to the community physically and/or in an accessible format. Discuss the ownership and management of the project deliverables and document the terms in schema element 14 Distribution.

10. Plan to meet periodically with collaborators to discuss updates and relevant questions. Establish a mediation process for handling disagreements as they arise.
11. Share updates with the community in a way that is transparent and comprehensible to those outside the project.
12. Each member of the project team should receive acknowledgement and due credit for their contributions to the project in a way that is meaningful to the team member.

General Best Practices for Documentation

1. Remember that a broad range of people may be consulting this documentation 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 that will contain sensitive or proprietary information, whenever possible write the documentation 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 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. 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. In effect, 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).
8. Writing clear, concise documentation takes time and thought. We recommend iterating on the text of the documentation development.
9. If the content of the dataset will contain 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. For datasets concerning languages other than English, also publish the documentation in the language(s) of the dataset.
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 for Documentation 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 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 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.

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

Text refers to a sequence of language data.

Schema

1 HEADER

Why

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

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

What

The header should include the following:

- Dataset Title
- Dataset Contributor(s) [name, affiliation, role]
- Dataset Version [version, date]
- Dataset Citation and DOI
- Documentation Contributor(s) [name, affiliation, role]
- Documentation Version [version, date]
- Documentation Citation
- Links to versions of this documentation in other languages

Best Practices

1. In order to manage updates over time, both datasets and their associated documentation should be versioned. That is, each updated dataset version should have its own updated documentation version. The documentation version number should be included in the documentation citation and is requested above. (Note that “Documentation Version” refers to the version of the documentation, not the version of the documentation schema that is being used.)
2. In creating a standard citation for your documentation, we recommend including the following information about the documentation: authors, date, title, version, institution, and URL or DOI.

3. Consider web accessibility and the longevity of documentation location (e.g., university archives or a community-owned repository). See 15 Maintenance for further considerations.
4. 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 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 to the community. A summary drafted before the data collection will need to be updated to reflect the final version.

For documentation 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 documentation 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 documentation that at a minimum should include: (1) a one-sentence description of the curation rationale, (2) the language(s), (3) an overview of relevant quantitative information such as the anticipated dataset size, and (4) a short description of how the community has been involved in the project.

Best Practices

1. 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.
2. 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 documentation 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: What is the intended purpose of this dataset? What is the task or research question the dataset is intended to address? Which texts will be included and what are the goals in selecting texts, both in the original collection and in any further subselection? What will be the internal organization of the dataset? What will constitute a data instance? How will the dataset support community goals?

Best Practices

1. If the dataset will include 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 elements below should speak to each subcategory.
2. If the dataset will involve subselection from a larger collection, specify topics, keywords, or other filters that will be used and the reasons for choosing each. Technical details can be provided in 9 Preprocessing and Data Formatting.
3. 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.

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 documentation readers, the source dataset documentation can help with understanding how the current dataset will build upon and differ 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 that will be built out of pre-existing datasets, a link to the documentation for each source dataset should be included. Provide links to licenses, copyright, or terms of use for source datasets, where applicable.

Best Practices

1. 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 documentation for the most current dataset (e.g., C) should only refer to the immediate source (e.g., B).
2. Include enough detail in the body of the documentation so that should the links between the documentation and the immediate source break, the documentation could function reasonably well as a stand-alone document.
3. 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. When in doubt, contact the source dataset manager and ask about developing a new opt-in consent procedure for the language users who created the source data to agree to the new use and dissemination of their data.

5 LANGUAGE VARIETIES

Why

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.

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

For documentation readers, accurate descriptions of the language varieties in the dataset are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, 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

All of the languages and language varieties that will be 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

1. Describe all language varieties that will be represented in the dataset and the metadata. For translation datasets, this would include both sides of the bitext. If the language variety that will be used for annotations differs from the language variety of the source data, again document both.
2. 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 and avoid harmful language ideologies (Kroskrity, 2005).
3. In the prose description, describe the dialects that will be 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

Why

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) (Christensen et al. 2012, Nicolao et al. 2016; for dysarthria in signed languages see Tyrone 2014).

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. Documentation can also enable the discovery of underserved populations across the overall data catalog which, in turn, may influence choices for constructing the new dataset.

For documentation readers, accurate descriptions of the people represented in the dataset are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, 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

All of the language user groups that will be 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
- Proposed number of different speakers or signers represented

- Presence of disordered speech or sign

Best Practices

1. Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).
2. Because the definitions and labels of demographic categories can change over time, include the dates when the data were or will be produced and when the data will be collected.
3. If the dataset will include language users with different roles (e.g., interviewers, interviewees, and interpreters), provide demographic information for each role separately.
4. If the dataset will consist entirely of synthetic text, if available, provide demographic information for the language users in the training data for the automatic generation system.
5. If the dataset will contain both found and elicited data, provide separate language user demographics for each.
6. 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).
7. 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).
8. Report demographic information at the level of the entire dataset rather than attached to individual language users to help protect their privacy.
9. 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. Discuss with community representatives what demographic information may be safely gathered and shared.

7 ANNOTATOR DEMOGRAPHIC

Why

Linguistic variation correlated with the language user's demographics is also relevant for annotators. Specifically, the 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.

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 documentation readers, accurate descriptions of the annotators' demographics are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, 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

All of the annotator groups that will be represented in the dataset, including those who will 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
- Proposed number of different annotators represented
- Relevant training

Best Practices

1. Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).
2. Because the definitions and labels of demographic categories can change over time, include the dates for when the annotations will be produced.
3. If the dataset will include annotators with different roles (e.g., translators and labelers), provide demographic information for each role separately.
4. If the dataset will include automatically produced annotations, if available provide demographic information for the training data for the automatic annotation system.
5. If the dataset will contain both found and elicited annotations, provide separate annotator demographics for each.
6. 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).
7. 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).
8. Report demographic information at the level of the entire dataset rather than attached to individual annotators to help protect their privacy.
9. 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. Discuss with community representatives what demographic information may be safely gathered and shared.

8 LINGUISTIC SITUATION AND TEXT CHARACTERISTICS

Why

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).

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 documentation readers, accurate descriptions of the linguistic situation in the dataset are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, 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 will occur 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 that will be collected. Specifications include:

- Time and place of linguistic activity
- Proposed 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 vs. social media)
- Topic (e.g., entertainment vs. natural disaster)
- Non-linguistic context (e.g., photos participants were all looking at; a game participants are playing)
- Additional details about the cultural context (optional)

Best Practices

1. 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.
2. 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 documentation 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 that will be 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 will be used to make the modifications and whether the raw data will be included in the dataset.

Best Practices

1. We recommend the description take the form of a list of ordered steps, with a link to external documentation of specific details, as appropriate.
2. If different preprocessing steps will be 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).
3. If the dataset will be 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.
4. To the extent possible, provide software version information, citations, and links to repositories for the tools that will be used in automatic processing.
5. When anonymizing video or image data, modifications to the data such as blurring faces may remove necessary linguistics 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, and ensuring the usefulness of the dataset for the community.

10 CAPTURE QUALITY

Why

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

For documentation readers, accurate descriptions of the recording quality are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, 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 third, 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 anticipated 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

1. For data that will include audiovisual recordings, describe the quality of the recording equipment and any aspects of the recording situation that could impact recording.
2. 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 documentation readers, accurate descriptions of the challenges encountered in creating the dataset are important for at least two three reasons: first, to evaluate the degree to which this dataset has contributed towards community goals; second, to assess if the dataset would be well-matched for a particular intended use case; and third, to enable future third party technology developers or adopters to make similar assessments of match to populations at a future time.

What

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

Best Practices

1. We recommend documenting the challenges you encounter in the dataset development as they occur, including both the challenge and your strategy for addressing it.
2. 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.
3. 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 documentation readers, documentation may be the “front door” through which they access the dataset. As such, it is important that the documentation 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 that annotators will use to annotate the data
- Annotation Process: Link to documentation providing metadata about the proposed annotation process, including protections for annotator anonymity, how annotators will be compensated, and which aspects of the annotation will be produced automatically
- Dataset Quality Metrics: Proposed metrics for inter-annotator agreement and/or other numerical scores of dataset quality

Best Practices

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

13 DISCLOSURES 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 a community has an ethical review process, engagement with this process can help surface community-specific concerns with the dataset creation and help guide the dataset creation to support community goals.

For documentation readers, information about funding sources (which may shape 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 proposed consent process; if language users in the dataset or annotators will be compensated, how compensation rates will be determined; and any potential conflicts of interest.

Best Practices

1. If your data collection process will involve a consent procedure, describe this element briefly with phrases such as “written consent”, “oral consent”, or “implied consent”.
2. 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 will not be accessible at the time of dataset creation.”
3. If the community has an ethical review process, we recommend stating whether or not the project has engaged with the process and any results from the engagement.

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 will also need to provide distribution information to the people the data is collected from as part of the consent procedure.

For documentation readers, a detailed description of the permitted uses of this dataset can help in determining whether the dataset is suitable for a particular use case and whether the dataset can be further redistributed. If the documentation is the first access point for a reader, the distribution explanation can help the reader find and access the dataset or explain why they are unable to find or access the dataset. Documentation that communicates planned revisions or removal of the dataset in advance may also help documentation readers prepare for changes to the dataset.

What

A description of how the dataset will be distributed should be specified. This includes the method of distribution (e.g., through a data archive, files on website, API, GitHub) and any access restrictions on the dataset or subsets of the dataset (e.g. sensitive or confidential content, intellectual property (IP)-based restrictions, export controls, or other regulatory restrictions). If the dataset or portions of the dataset will be distributed under an IP license, copyright, or terms of use (ToU), describe the licenses, copyright, and/or ToU. Provide links or other access points to, or otherwise reproduce, any relevant licensing terms or ToU, and list any fees associated with these restrictions. Other suggestions for detailing the distribution plan include:

- Who the dataset will be distributed to (e.g. third parties outside of the entity (community, company, institution, or organization) on behalf of which the dataset was created)
- If there are conditions for accessing the dataset or subsets of the dataset, and if so, what the conditions for being granted access are
- If the dataset will have a digital object identifier (DOI)
- When the dataset will be distributed

Best Practices

1. Review the data with community representatives to determine which portions of the dataset will be culturally appropriate to share broadly, which portions should be restricted to relevant groups, and which portions should be accessible to community members only.

2. In addition to the general distribution method, provide the community with a locally accessible copy of the dataset.
3. When choosing terms for a license, copyright, or ToU, consider uses that will be allowed as well as uses that will be disallowed. The community should decide on whether they want to allow third-party uses such as research, use in court, technical development, and commercialization.

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 both the community and other intended audiences. For communities, developing a maintenance plan may help in considering archiving options along with their benefits, risks, and costs prior to data collection.

For documentation readers, information about the dataset's maintenance will help determine who to contact for questions about the dataset after it has been published. Information about previous updates may help determine which version of the dataset will be 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 will be maintained should be specified. This includes who will support, host, and maintain the dataset and what the proposed method for contacting the manager of the dataset will be. Other considerations include:

- If and where a list of errors found after the dataset's publication will be maintained and how to report errors
- How often, by whom, and how updates to the dataset (e.g., to correct labeling errors, add new data, delete data) will be communicated to users (e.g., mailing list, GitHub)
- Applicable limits on the retention of the data associated with the instances (e.g., will individuals in question be told that their data will be retained for a fixed period of time and then deleted) and how those limits will be enforced
- Whether older versions of the dataset will continue to be supported, hosted, and maintained
- How users will be notified that the dataset is outdated or no longer available
- Whether others will be able to extend/augment/build on/contribute to the dataset, and if so, how others will be able to contribute, if and how these contributions will be validated, and whether these contributions will be further communicated and distributed to other users

Best Practices

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

2. We recommend having a process for removing data, in the event that someone would like to have their data or community-sensitive data removed from the dataset.

16 OTHER

Why

This toolkit 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

1. 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 documentation authors, using technical terms can make it easier to write efficient and precise documentation. Using local terminology throughout the documentation centers the community's understanding of the data and its cultural significance. Providing definitions for these technical terms can make the data statement accessible to a wider variety of audiences.

For documentation 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 local vocabulary can be important for understanding and interpreting the data in community-appropriate ways and acknowledging the validity of community knowledge and ways of knowing.

What

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

Best Practices

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