

C3DAR Google Doc Template

About the C3DAR Toolkit and this template

The Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources (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.

This template is for the C3DAR Toolkit and 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 template is licensed as [CC0](https://creativecommons.org/licenses/by/4.0/).

How to use this document

Instructions are given as **gray text**. Fill in each section according to the instructions. For more information about each schema element as well as best practices for completing the C3DAR Toolkit, see the C3DAR Toolkit Guide.

When you're done, delete the text above the black line as well as the instructions in **gray text** embedded in the template. Provide the file containing the documentation produced with the C3DAR Toolkit with your data, for example as "C3DAR.pdf".

C3DAR Toolkit for [Dataset Name]

1. HEADER

The header should include the following:

Dataset Title:

Dataset Curator(s): [name, affiliation, role]

Dataset Version: [version, date]

Dataset Citation and DOI:

Documentation Author(s): [name, affiliation, role]

Documentation Version: [version, date]

Documentation Citation:

Links to versions of this documentation in other languages:

2. EXECUTIVE SUMMARY

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.

3. CURATION RATIONALE

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 sub-selection? What will be the internal organization of the dataset? What will constitute a data instance? How will the dataset support community goals?

4. DOCUMENTATION FOR SOURCE DATASETS

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.

5. LANGUAGE VARIETIES

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).

6. LANGUAGE USER DEMOGRAPHIC

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

7. ANNOTATOR DEMOGRAPHIC

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

8. LINGUISTIC SITUATION AND TEXT CHARACTERISTICS

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)

9. PREPROCESSING AND DATA FORMATTING

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.

10. CAPTURE QUALITY

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.

11. LIMITATIONS

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.

12. METADATA

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)

13. DISCLOSURES AND ETHICAL REVIEW

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.

14. DISTRIBUTION

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

15. MAINTENANCE

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

16. OTHER

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

17. GLOSSARY

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

About this document

Include this information about the document verbatim at the end of your documentation. If you adapt the C3DAR template, include a note about your changes here.

This documentation was written based on the template for the C3DAR Toolkit. The template 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.