AETHER DATA DOCUMENTATION TEMPLATE (DRAFT 08/25/2022)

INTRODUCTION

Data is central to the development and evaluation of machine learning models. Many responsible AI harms can be traced back to characteristics of datasets. For example, lack of appropriate representation of different groups of people can lead to models that exhibit performance disparities. Spurious correlations and other unanticipated anomalies in training datasets can result in models that fail to generalize. Subjectivity in dataset labels and inaccurate notions of ground truth can result in models with misleading outputs.

Documenting datasets helps promote more deliberate reflection and transparency about how these datasets might affect machine learning models. For dataset creators, documenting your data can help you think through underlying assumptions, potential risks, and implications of use. It can also help dataset consumers—those who will use a dataset to develop or evaluate their models—make informed decisions about whether specific datasets meet their needs, and what limitations they need to consider. For these reasons, good data documentation practices are an essential component of responsible AI.

This template includes questions that dataset creators should think through and document the answers to. However, it is not meant to be prescriptive. There is no one-size-fits-all format for data documentation. You are encouraged to adapt this template to suit your team's needs and the type of datasets you use. If some questions do not apply to your situation or dataset, it's fine to drop them. You may also choose to integrate these questions into your existing tools and workflows rather than answering them in the template itself.

What kinds of datasets need to be documented?

This template has been developed specifically for datasets that are used to train or evaluate machine learning models. It may be used for datasets intended for both internal and external distribution.

Which version of my data should I document?

This depends on the use case. If you are releasing a dataset publicly or sharing it with another team, you should document the version of the dataset that you share. If you are using a dataset to train or test a model, you should document the version that is used. Most commonly, this means documenting a dataset that has already gone through some amount of cleaning and pre-processing.

Determining which version of your dataset to document becomes trickier if you rely on telemetry data. Here again we invite you to carefully consider the use case. Note that even when documenting fully processed datasets, some questions ask you to reflect on the raw data that went into them.

Who is this documentation for?

When creating your data documentation, it's useful to keep in mind its future readers and ask yourself whether the information provided is enough to meet their needs. These future readers might include:

- People who are considering using this dataset to train or evaluate models
- People who are auditing a model or AI system

- Stakeholders impacted by model(s) trained or evaluated on this dataset
- Future you or your own teammates

All these parties may need to understand what's in the dataset, how it was created and pre-processed, and what uses it's appropriate for.

What level of detail should I include in my answers?

You should provide enough detail that someone completely unfamiliar with the dataset would be able to make an informed decision about whether and how to use this dataset responsibly.

When should I document a dataset?

Data documentation is an iterative process and should not be postponed until the dataset is complete. You can start documenting your planned dataset as early as when you're first designing specs for the data. This will help you reflect on and potentially improve your data collection practices, leading to higher quality data.

Continue iterating on your documentation at key points in the dataset's lifecycle: planning; collection; pre-processing, cleaning, and labeling; distribution; and maintenance.

You can think of your data documentation as a living document. Update it when needed. If applicable, you may want to maintain a separate version to accompany each version of your dataset. This is especially important when older versions of the dataset may still be in circulation or in use.

Who should be involved in dataset documentation?

Like many aspects of responsible AI, data documentation is a collaborative process. Often no one person knows the answers to all questions. Work with your team to determine who is responsible for answering each question. If you don't know the answer to a question, find someone who does.

How was this template created?

This version of the template was created by Microsoft's Aether Transparency Working Group. It evolved from years of research and pilot studies within and outside of Microsoft. We started with the template from Datasheets for Datasets [1] and edited it to reflect the needs of industry practitioners [2]. This template is a work in progress, and we expect it will continue to evolve. Your feedback can help improve it, so don't hesitate to reach out! Contact datadoc@microsoft.com with feedback.

[1] Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Communications of the ACM, 64(12):86–92, December 2021.

[2] Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Proceedings of the ACM on Human-Computer Interaction, 6, CSCW2, Article 340, November 2022.

DATASET OVERVIEW

BASICS: CONTACT, DISTRIBUTION, ACCESS

1. Dataset name
2. Dataset version number or date
3. Dataset owner/manager contact information, including name and email
4. Who can access this dataset (e.g., team only, internal to the company, external to the company)?
5. How can the dataset be accessed?

DATASET CONTENTS

6. What are the contents of this dataset? Please include enough detail that someone unfamiliar with the dataset who might want to use it can understand what is in the dataset.

   Specifically, be sure to include:

   - What does each item/data point represent (e.g., a document, a photo, a person, a country)?
   - How many items are in the dataset?
   - What data is available about each item (e.g., if the item is a person, available data might include age, gender, device usage, etc.)? Is it raw data (e.g., unprocessed text or images) or features (variables)?
   - *For static datasets:* What timeframe does the dataset cover (e.g., tweets from January 2010–December 2020)?

INTENDED & INAPPROPRIATE USES

7. What are the intended purposes for this dataset?
8. What are some tasks/purposes that this dataset is not appropriate for?

DETAILS

DATA COLLECTION PROCEDURES

9. How was the data collected?

   Describe data collection procedures and instruments.

   Describe who collected the data (e.g., contractors).

10. Describe considerations taken for responsible and ethical data collection (e.g., procedures, use of crowd workers, recruitment, compensation).

11. Describe procedures and include language used for getting explicit consent for data collection and use, and/or revoking consent (e.g., for future uses or for certain uses). If explicit consent was not secured, describe procedures and include language used for notifying people about data collection and use.

REPRESENTATIVENESS

12. How representative is this dataset? What population(s), contexts (e.g., scripted vs. conversational speech), conditions (e.g., lighting for images) is it representative of?

    How was representativeness ensured or validated?

    What are known limits to this dataset's representativeness?

13. What demographic groups (e.g., gender, race, age, etc.) are identified in the dataset, if any?

    How were these demographic groups identified (e.g., self-identified, inferred)?

    What is the breakdown of the dataset across demographic groups? Consider also reporting intersectional groups (e.g., race x gender) and including proportions, counts, means or other relevant summary statistics.

    Note: This information can help a user of this dataset understand what groups are represented in the dataset. This has implications for the performance of models trained on the dataset and on its appropriateness for fairness evaluations – e.g., comparisons of performance across groups.

DATA QUALITY

14. Is there any missing information in the dataset? If yes, please explain what information is missing and why (e.g., some people did not report their gender).

    Note: Consider the impact of missing information on appropriate uses of this dataset.

15. What errors, sources of noise, or redundancies are important for dataset users to be aware of?

    Note: Consider how errors, noise, redundancies might impact appropriate uses of this dataset.

16. What data might be out of date or no longer available (e.g., broken links in old tweets)?

17. How was the data validated/verified?

18. What are potential validity issues a user of this dataset needs to be aware of (e.g., survey answers might not be truthful, age was guessed by a model and might be incorrect, GPA was used to quantify intelligence)?

19. What are other potential data quality issues a user of this dataset needs to be aware of?

PRE-PROCESSING, CLEANING, AND LABELING

20. What pre-processing, cleaning, and/or labeling was done on this dataset?

    Include information such as: how labels were obtained, treatment of missing values, grouping data into categories (e.g., was gender treated as a binary variable?), dropping data points.

Who did the pre-processing, cleaning, and/or labeling (e.g., were crowd workers involved in labeling?)

Note: Consider how this might impact appropriate users of this dataset (e.g., binary gender might be insufficient for fairness evaluations; imputing missing values with the mean may create anomalies in models trained on the data).

21. Provide a link to the code used to preprocess/clean/label the data, if available.

22. If there are any recommended data splits (e.g., training, development/validation, testing), please explain.

PRIVACY

23. What are potential data confidentiality issues a user of this dataset needs to be aware of?

    How might a dataset user protect data confidentiality?

24. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

    Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race, sexual orientation, age, ethnicity, disability status, political orientation, religious beliefs, union memberships; location; financial or health data; biometric or genetic data; criminal history)?

    If the answer to either of these questions is yes, please be sure to consult with a privacy expert and receive approvals for storing, using, or distributing this dataset.

25. If an analysis of the potential impact of the dataset and its uses on data subjects (e.g., a data protection impact analysis) exists, please provide a brief description of the analysis and its outcomes here and include a link to any supporting documentation.

26. If the dataset has undergone any other privacy reviews or other relevant reviews (legal, security) please include the determinations of these reviews, including any limits on dataset usage or distribution.

ADDITIONAL DETAILS ON DISTRIBUTION & ACCESS

27. How can dataset users receive information if this dataset is updated (e.g., corrections, additions, removals)?

    Note: Consider creating a distribution list people can subscribe to.

28. *For static datasets:* What will happen to older versions of the dataset? Will they continue to be maintained?

29. *For streaming datasets:* If this dataset pulls telemetry data from other sources, please specify:

    - What sources
    - How frequently the dataset is refreshed

- Who controls access to these sources
- Whether access to these sources will remain available, and for how long
- Any applicable access restrictions to these sources including licenses and fees
- Any other available access points to these sources
- Any relevant information about versioning

Are there any other ways in which these sources might affect this dataset that a dataset user needs to be aware of?

30. If this dataset links to data from other sources (e.g., this dataset includes links to content such as social media posts or, news articles, but not the actual content), please specify:
    - What sources
    - Whether access to these sources will remain available, and for how long
    - Who controls access to these sources
    - Any applicable access restrictions to these sources including licenses and fees
    - *For static datasets:* If an official archival version of the complete dataset exists (i.e., including the content as it was at the time the dataset was created), where it can be accessed

    Are there any other ways in which these sources might affect this dataset that a dataset user needs to be aware of?

31. Describe any applicable intellectual property (IP) licenses, copyright, fees, terms of use, export controls, or other regulatory restrictions that apply to this dataset or individual data points.

    These might include access restrictions related to data subjects' consenting or being notified of data collection and use, as well as revoking consent.

    Provide links to or copies of any such applicable terms.