Datasheet for Earth Science Datasets Instructional Guide

Released: December 12, 2022
Last updated: March 8, 2024

Charlotte Connolly
Department of Atmospheric Science
Colorado State University
Fort Collins, CO
cconn@rams.colostate.edu

Daniel M. Hueholt
Department of Atmospheric Science
Colorado State University
Fort Collins, CO
daniel.hueholt@colostate.edu

1. Welcome!

This instructional guide documents the beta version of Datasheets for Earth Science Datasets, which we have released for community feedback. We encourage you to reach out to us (email addresses in author block) and provide any comments, feedback, or questions. We greatly appreciate all input!

2. Background

In "Datasheets for Datasets" ([1], original preprint [2]), Gebru et al. propose a structured questionnaire format to provide a practical method of transparently documenting big datasets in software engineering. Answering the questions in the datasheet guides the creator(s) of a dataset through self-reflection to document biases, communicate human impacts, and provide technical information. Datasheets and their derivatives (e.g., Mitchell et al.'s "Model Cards for Model Reporting" [3]) have been widely adopted in machine learning research and industry, and enable both transparency and critical analysis of these applications (e.g., [3], [4], [5]).

As two Earth scientists, we have tailored the datasheet format to Earth science datasets and have made improvements based on community feedback at our institution (Colorado State University [CSU]) and elsewhere. We still consider this a "beta version" and welcome continuing feedback, comments, and questions from the broader Earth science community. We plan to formally publish on this topic after further iteration, similar to the process that led to the establishment of Gebru et al.'s [1] original Datasheets for Datasets in software engineering. We hope this format will prove broadly useful to Earth scientists!

3. Using Datasheets for Earth Science Datasets

We see two major use cases for Datasheets for Earth Science Datasets. The first is a datasheet written by an individual or small team within the scope of a single project or publication. The second is a datasheet written by a large team for data from a large-scale project that may span multiple institutions (e.g., an Earth system modeling experiment, or major field campaign). This beta version of Datasheets for Earth Science Datasets is focused on the first use, but later editions will encompass the second case as we gain more insight from the community. If you would like to provide feedback on Datasheets for Earth Science Datasets, please read and/or complete a datasheet for your data. Ideally, datasheets are completed during the project–alongside the process of generating, collecting, and processing the data–but can easily be completed at the end of a project, as well.

To fill out a datasheet, respond to each of the questions with the information that pertains to your dataset to the best of your ability. In response to previous feedback, we address a few points that commonly arise when filling out a datasheet.

1) **Not all questions may apply to your specific data.** That's okay! Questions that do not apply can be left blank–although we ask that you consider each question carefully.
2) **We are aware some questions may seem repetitive.** In the case where a question is repetitive, you can reference your response elsewhere in the datasheet. If you believe a certain question may be too repetitive we encourage you to contact us; we would be grateful for your input.
3) **Some information requested on the datasheet may be documented elsewhere.** If so, simply link to that document within your datasheet. We want datasheets to complement, not duplicate, existing standards for documentation.
4) **Datasheet questions are not meant to require a literature review!** You should not feel obligated to find every existing citation for a question, or to document every possible error or bias for every possible use of a dataset. Your answers should be limited to the scope of the project the data is used for.
5) **The time commitment for creating a datasheet is designed to be manageable.** We find that it takes users about 2-6 hours to complete a datasheet. Ideally, the datasheet is completed alongside the project as decisions are actively being made.

4. Demonstration Datasheets

We filled out datasheets for two Earth science datasets that we knew well from our graduate research work. Both of these datasheets are for modeling experiments, albeit for very different kinds. Additionally, we track all known completed

datasheets in the Readme of this repository. Please email us if you have a datasheet which you would like us to add!

If you intend to fill out a datasheet, we encourage you to first make an attempt on your own. Our demonstration datasheets do not provide a general "ground truth"–the answers to every datasheet will be unique to the project and to the author(s). We provide the demonstration datasheets as examples or references if you get stuck while creating your own. Note that these demonstration datasheets were created with the initial version (v0.0) of the datasheet template, and do not reflect the current version found on the repository.

Our demonstration datasheets can be found in the `demo_EarthScienceDatasheets` folder in this repository.

5. How to Contribute

Following the methods that Gebru et al. used to develop Datasheets for Datasets [1] in software engineering, we wish to develop Datasheets for Earth Science Datasets with the broader Earth science community. We encourage you to email us at the email addresses found at the top of this document. While email is the best way to get in touch with us, you may also raise an "issue" on the GitHub repository itself. We invite feedback, comments, or questions about all aspects of Datasheets for Earth Science Datasets. Additionally, we invite you to share these materials with anyone who may be interested.

We adapted Gebru et al.'s Datasheets for Datasets [1] by considering what questions would be most useful for Earth science. Some questions (particularly high-level questions about authorship or purpose) have been borrowed or derived from Gebru et al.'s datasheets [1], while many are new questions which have been designed specifically for Earth science applications. We further refined these questions based on feedback from presentations and discussions with other researchers in various branches of Earth science. However, we are still limited by our own background and experiences. We are particularly interested in community insight on the following themes.

1) **The current Datasheets for Earth Science Datasets draft does not comprehensively represent data encountered in all branches of Earth science.** We designed this beta version to be applicable to numerical model output, in situ observations, remote sensing observations, and augmented/synthetic data to the best of our ability. Some data types that we see as particularly notable gaps include paleoclimate proxies and laboratory results. We are very interested in expanding the data types that the datasheets apply to, and would be grateful for any input that readers can provide.
2) **We have not completed a datasheet in a group context, i.e. for a large dataset.** Both of our demonstration datasheets apply to datasets that accompany individual papers and were completed by their respective first authors. If you are completing a datasheet in a group setting, please reach out–we would love to discuss this process! This is a use case which we see as critical to the large-scale adoption of any documentation format in Earth science, and we want to ensure that datasheets work in this setting.
3) **There may be gaps remaining in our coverage of possible human impacts and discriminatory outcomes.** Biases in datasets and their impacts on humans are the primary reason for the development of the original Gebru et al. datasheets standard [1] and drove our initial interest in this topic, as well. We have attempted to keep these issues in mind throughout the development of Datasheets for Earth Science Datasets. However, we recognize there may still be gaps remaining in our approach and we encourage relevant feedback from the community.
4) **Any other comments that we are not aware of!** We would be grateful for feedback, comments, or questions regarding any aspects of our datasheets–not just the ones mentioned above.

6. Acknowledgements

We are grateful for valuable feedback, insights, and constructive criticism from the Barnes Group in September 2021 and January 2023, Prof. Melissa Burt and our classmates in ATS680: Social Responsibility in Atmospheric Science at CSU, attendees of the "Biased Data, Biased Model" workshop at CSU in November 2022, attendees of the "Datasheets for Earth Science Datasets" workshop at CSU in February 2023, Dr. Imme Ebert-Uphoff, Prof. Elizabeth Barnes, and Prof. James Hurrell. The LaTeX formatting for the datasheet is based on an Overleaf template by Olamilekan Wahab.

This work was developed primarily as part of the semester project in ATS680.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, November 2021.

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, March 2018. arXiv:1803.09010 [cs] version: 1.

[3] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA, January 2019. Association for Computing Machinery.

[4] Christian Garbin. Assessing Methods and Tools to Improve Reporting, Increase Transparency, and Reduce Failures in Machine Learning Applications in Healthcare. Master's thesis, Florida Atlantic University, 2020.

[5] Yang Trista Cao and Hal Daumé, III. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661, November 2021.