Datasheet for Earth Science Datasets Instructional Guide

Released: December 12, 2022
Last updated: April 21, 2025

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 Datasheets for Earth Science Datasets templates, which we have released for community use. We encourage you to use the templates for your own projects, and reach out to us (email addresses in author block) to provide comments, feedback, or questions. We greatly appreciate all input!

A paper about datasheets for Earth science datasets is now published at the *Bulletin of the American Meteorological Society*: doi.org/10.1175/BAMS-D-24-0203.1 [1].

2. Background

In "Datasheets for Datasets" ([2], original preprint [3]), 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 a reflective process 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" [4]) have been widely adopted in machine learning research and industry, and enable both transparency and critical analysis of these applications (e.g., [4], [5], [6], [7]).

As two Earth scientists, we tailored the datasheet format to Earth science datasets and made improvements based on community feedback between fall 2021 and 2024. We encourage you to read the paper (doi.org/10.1175/BAMS-D-24-0203.1) to learn more: [1]. 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). 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 2-7 hours to complete a datasheet. Ideally, the datasheet is completed alongside the project as decisions are actively being made.

We currently track examples of completed datasheets on this GitHub repository in the "Datasheet tracker" section of the Readme. This list is non-comprehensive. 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. The answers to every datasheet will be unique to the project and to the author(s), and the examples tracked here do not provide a general ground truth. However, completed datasheets may be helpful references if you get stuck.

4. ACKNOWLEDGEMENTS

We thank all those who have provided community feedback on Datasheets for Earth Science Datasets since its release online, including: Dr. Marybeth Arcodia, Jamin Rader, and Daniel Veloso Águila, as well as other individuals through private communication.

We are grateful for valuable feedback, insights, and constructive criticism during early development of this work 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.

REFERENCES

[1] Charlotte J. Connolly, Daniel M. Hueholt, and Melissa A. Burt. Datasheets for Earth Science Datasets. *Bulletin of the American Meteorological Society*, 106(4):E642–E648, April 2025. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Karen L. Boyd. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):438:1–438:27, October 2021.