

2026-03-03

## Data in the Humanities

Wk. 06: Scraping & Curating Structured Data

### Happenings This Week

#### *Workshops*

- Fri 3/6 Research Data Stewardship series: Introduction to Data Management (PUL)

#### *Events*

- Mon 3/23 Philosophical Foundations of AI (NAM), 2-part series
- Mon 3/9 - Wed 3/11 Cultural AI: An Emerging Field at NYU, registration required

## Mid-Semester Survey

Feedback for me, informal self-eval for you.

<https://forms.gle/jFuPZ9oujP6f9oqZ9>

## Readings

- Katie Rawson and Trevor Muñoz, "Against Cleaning," in *Debates in the Digital Humanities 2019* (Minneapolis: University Of Minnesota Press, 2019).
- Henk Alkemade et al., "Datasheets for Digital Cultural Heritage Datasets," *Journal of Open Humanities Data*, 9:2 (October 30, 2023).
- C. Thi Nguyen, "The Limits of Data," *Issues in Science and Technology*, 40:2 (2024).

# Datasheets for Datasets

Gebru<sup>1</sup> Jamie Morgenstern<sup>2</sup> Briana Vecchione<sup>3</sup> Jennifer Wortman Vaughan<sup>1</sup> Hanna Wallach<sup>1</sup>  
Hal Daumé III<sup>1,4</sup> Kate Crawford<sup>1,5</sup>

## Abstract

The machine learning community has no standard way to document how and why a dataset is created, what information it contains, what it should and should not be used for, and what it might raise any ethical or legal concerns. To address this gap, we propose the concept of a datasheet for datasets.

We therefore propose the concept of datasheets for datasets. In the electronics industry, every component is accompanied by a datasheet describing standard operating characteristics, test results, and recommended usage. By analogy, we recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information. We anticipate that such datasheets will increase transparency

Gebru et al., "Datasheets for Datasets" (2018)  
contra [The Pile](#) &c.

The screenshot displays the Data Nutrition Project website interface for the Hippocorpus V3 dataset. The header includes the project logo, a '75% COMPLETENESS' badge, and a 'What is this label?' section explaining the Dataset Nutrition Label. A metadata table on the right lists details such as 'Label author: Data Nutrition Project', 'Relationship to Data Creator: Third party', 'First published on: Apr 6, 2021', 'Last updated on: Jan 6, 2022', 'People consulted: Maarten Sap', and 'Label version: Version 2.0'. The main content area features a 'Description' section, 'Keywords' (Natural language processing, Imagination, Memory consolidation, Autobiographical memory), and 'How to use it?' with expandable sections for 'Intended Use', 'Known Use', 'Restrictions on Use', and 'Do Not Use'. A 'Preview data' button and a 'Download PDF' button are also visible.

<https://datanutrition.org>

Dataset Facts	
<b>Dataset</b> BookCorpus	
<b>Instances Per Dataset</b> 7,185 unique books, 11,038 total	
Motivation	
<b>Original Authors</b>	Zhu and Kiros et al. (2015) [39]
<b>Original Use Case</b>	Sentence embedding
<b>Funding</b>	Google, Samsung, NSERC, CIFAR, ONR
Composition	
<b>Sample or Complete</b>	Sample, ≈2% of smashwords.com in 2014
<b>Missing Data</b>	98 empty files, ≤655 truncated files
<b>Sensitive Information</b>	Author email addresses
Collection	
<b>Sampling Strategy</b>	Free books with ≥20,000 words
<b>Ethical Review</b>	None stated
<b>Author Consent</b>	None
Cleaning and Labeling	
<b>Cleaning Done</b>	None stated, some implicit
<b>Labeling Done</b>	None stated, genres by smashwords.com
Uses and Distribution	
<b>Notable Uses</b>	Language models (e.g. GPT [29], BERT [9])
<b>Other Uses</b>	List available on HuggingFace [12]
<b>Original Distribution</b>	Author website (now defunct) [39]
<b>Replicate Distribution</b>	BookCorpusOpen [13]
Maintenance and Evolution	
<b>Corrections or Erratum</b>	None
<b>Methods to Extend</b>	"Homemade BookCorpus" [21]
<b>Replicate Maintainers</b>	Shawn Presser [12]
Genres	
	% of BookCorpus*
<b>Romance</b> 2,881 books	26.1%
<b>Fantasy</b> 1,502 books	13.6%
<b>Vampires</b> 600 books	5.4%
Horror 4.1%	• Teen 3.9%
Adventure 3.5%	• Literature 3.0%
Historical Fiction 1.6%	
Not a significant source of nonfiction.	
* Percentages based on directories in books_txt_full. Some books cross-listed.	

<https://github.com/jackbandy/bookcorpus-datasheet>

**DAY LINE**

HUDSON RIVER DAY LINE  
STR. "NEW YORK"

**DINNER À LA CARTE**

**SOUPS**

Chicken Broth 25    Consommé 25    Mock Turtle 30    Clam Broth 25

**FISH**

Salmon Steak 50    Broiled Bluefish 40    Boiled Codfish, anchovy sauce 40  
Broiled Spanish mackerel, parsley sauce 40    Baked Bluefish, wine sauce 40

**BOILED**

Leg of Mutton, caper sauce 50    Corned Beef and cabbage 40

**ROAST**

Spring Chicken (half) 60    Mutton 40    Lamb, mint sauce 50  
Ribs of Beef 50    Veal, brown sauce 50

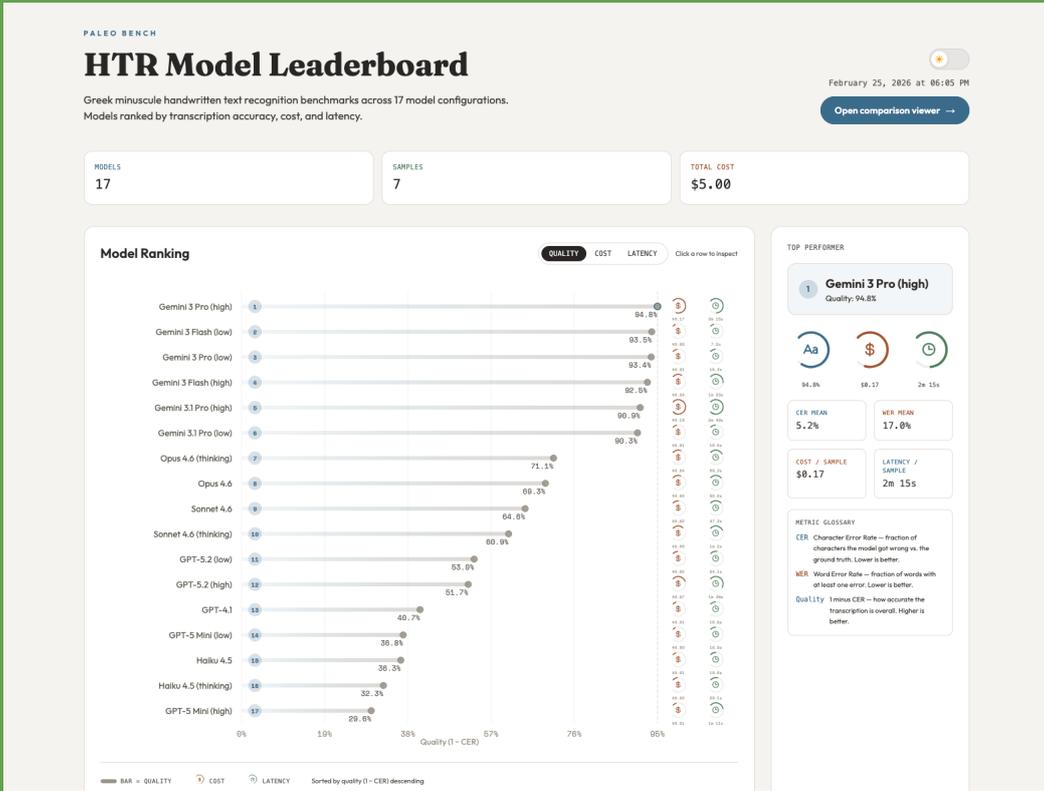
**ENTREES**

Porterhouse Steak, plain 1 25    with tomatoes 1 40    with mushrooms 1 75  
Tenderloin Steak, plain 75    with mushrooms 1 00    Veal Cutlet, breaded 50  
Lamb Chops, plain 60    with peas 75    Mutton Chops, plain 50  
Half Broiled Chicken 60    Soft Shell Crabs, tartar sauce 60

<https://menus.nyp1.org/>

🔍 +

Dish
Clam Broth
Soup, Mock Turtle
Consomme
Chicken Broth
Boiled Codfish, Anchovy Sat
Broiled Bluefish
Salmon Steak
Baked Bluefish, Wine Sauce
Broiled Spanish Mackerel, Parsley Sauce
Boiled, Corned Beef And Cabbage
Boiled, Leg Of Mutton, Cape Sauce
Roast, Lamb, Mint Sauce
Roast, Mutton
Roast, Spring Chicken (Half
Roast, Veal, Brown Sauce
Roast, Ribs Of Beef
Porterhouse Steak With Mushrooms



Paleo Bench Leaderboard, ranking Greek minuscule HTR performance

Fonts:  
 Klima  
 Katwijk Mono  
 both by Matthew Hinders-Anderson