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

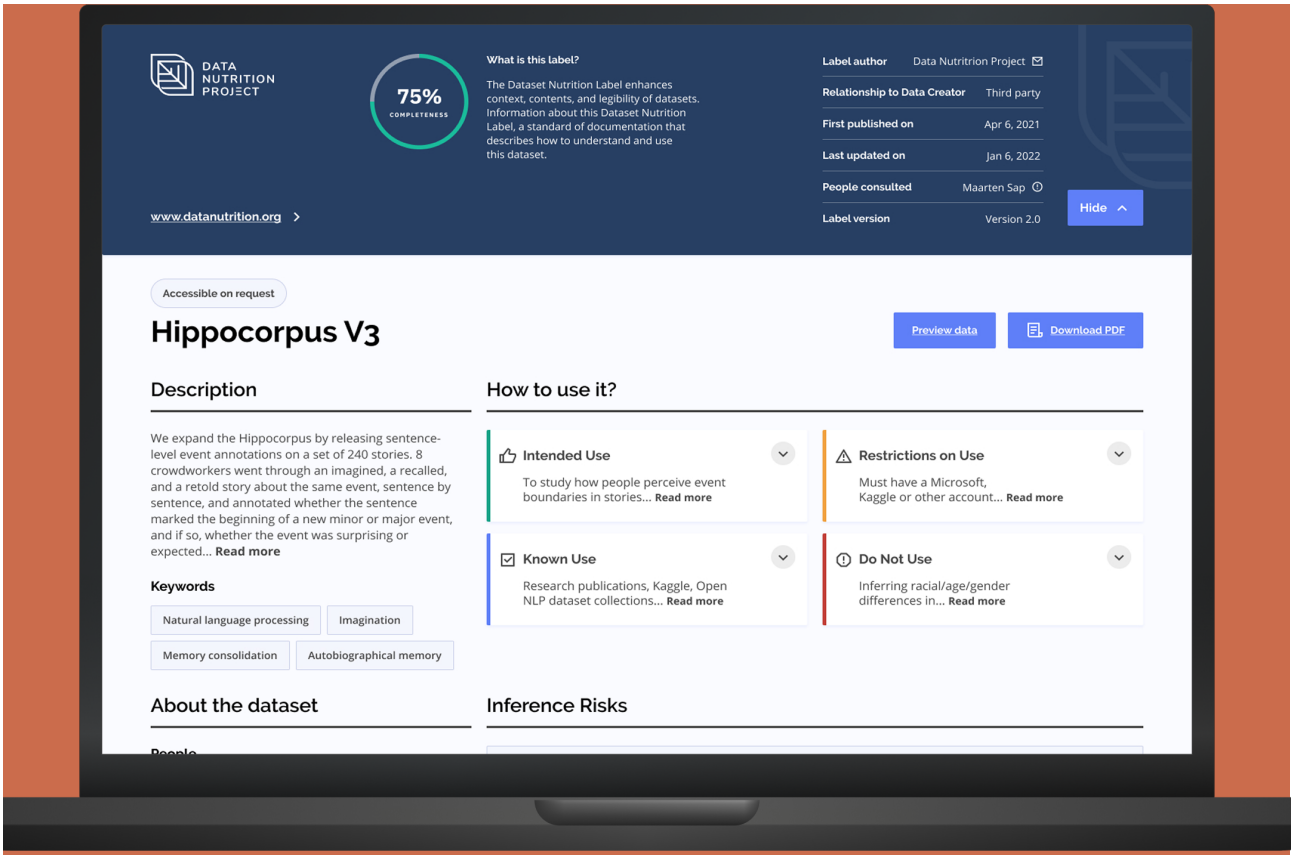
Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Abstract

achine learning community has no stan-
d way to document how and why a dataset
eated, what information it contains, what
should and should not be used for, and
r it might raise any ethical or legal con-
To address this gap, we propose the con-

We therefore propose the concept of datasheets for dataset
In the electronics industry, every component is accompanie
by a datasheet describing standard operating characteristic
test results, and recommended usage. By analogy, we rec
ommend that every dataset be accompanied with a datashee
documenting its motivation, creation, composition, intende
uses, distribution, maintenance, and other information. W
anticipate that such datasheets will increase transparenc

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



https://datanutrition.org

Dataset Facts

Dataset BookCorpus
Instances Per Dataset 7,185 unique books, 11,038 total

Motivation	
Original Authors	Zhu and Kiros et al. (2015) [39]
Original Use Case	Sentence embedding
Funding	Google, Samsung, NSERC, CIFAR, ONR
Composition	
Sample or Complete	Sample, ≈2% of smashwords.com in 2014
Missing Data	98 empty files, ≤655 truncated files
Sensitive Information	Author email addresses
Collection	
Sampling Strategy	Free books with ≥20,000 words
Ethical Review	None stated
Author Consent	None
Cleaning and Labeling	
Cleaning Done	None stated, some implicit
Labeling Done	None stated, genres by smashwords.com
Uses and Distribution	
Notable Uses	Language models (e.g. GPT [29], BERT [9])
Other Uses	List available on HuggingFace [12]
Original Distribution	Author website (now defunct) [39]
Replicate Distribution	BookCorpusOpen [13]
Maintenance and Evolution	
Corrections or Erratum	None
Methods to Extend	"Homemade BookCorpus" [21]
Replicate Maintainers	Shawn Presser [12]

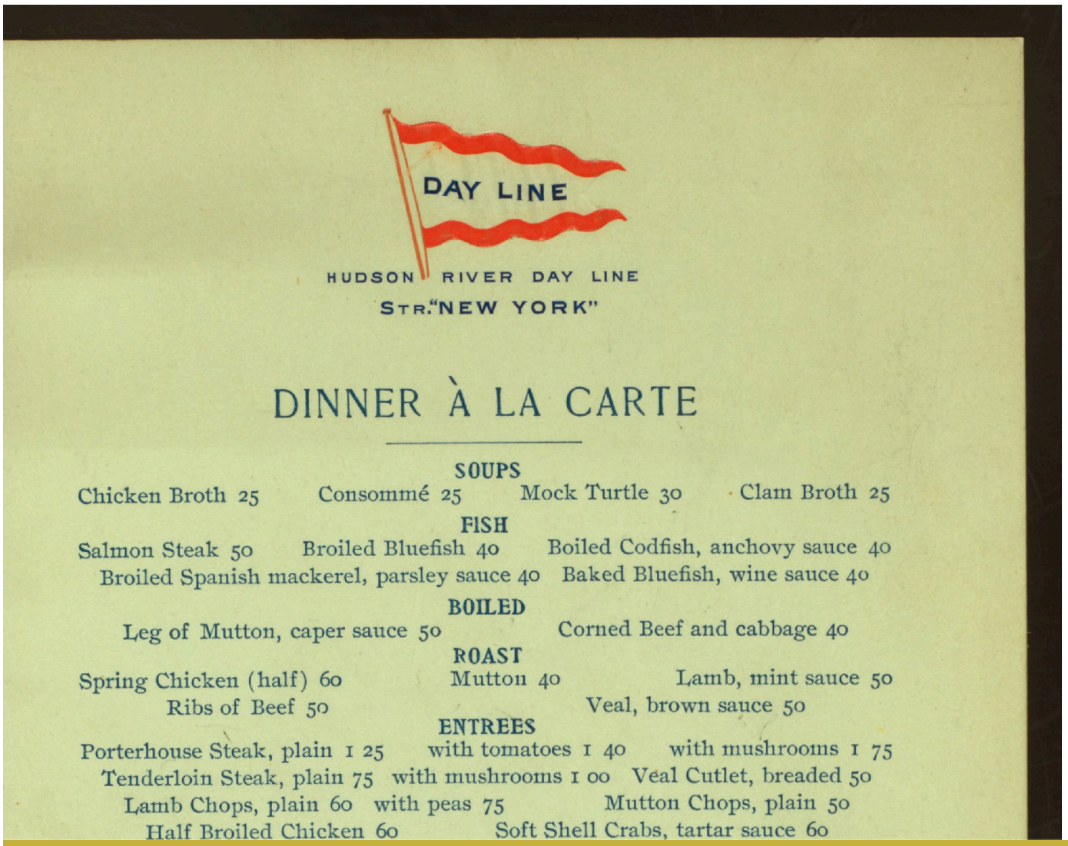
Genres	% of BookCorpus*
Romance 2,881 books	26.1%
Fantasy 1,502 books	13.6%
Vampires 600 books	5.4%

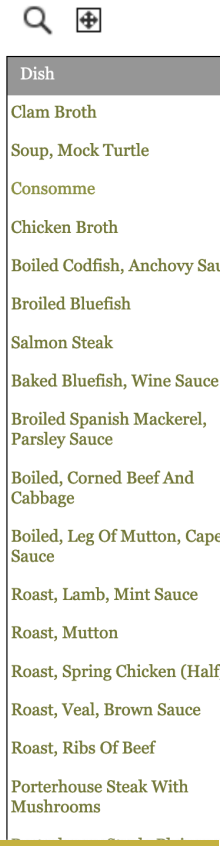
Horror 4.1%
Adventure 3.5%
Historical Fiction 1.6%
• Teen 3.9%
• Literature 3.0%

Not a significant source of nonfiction.

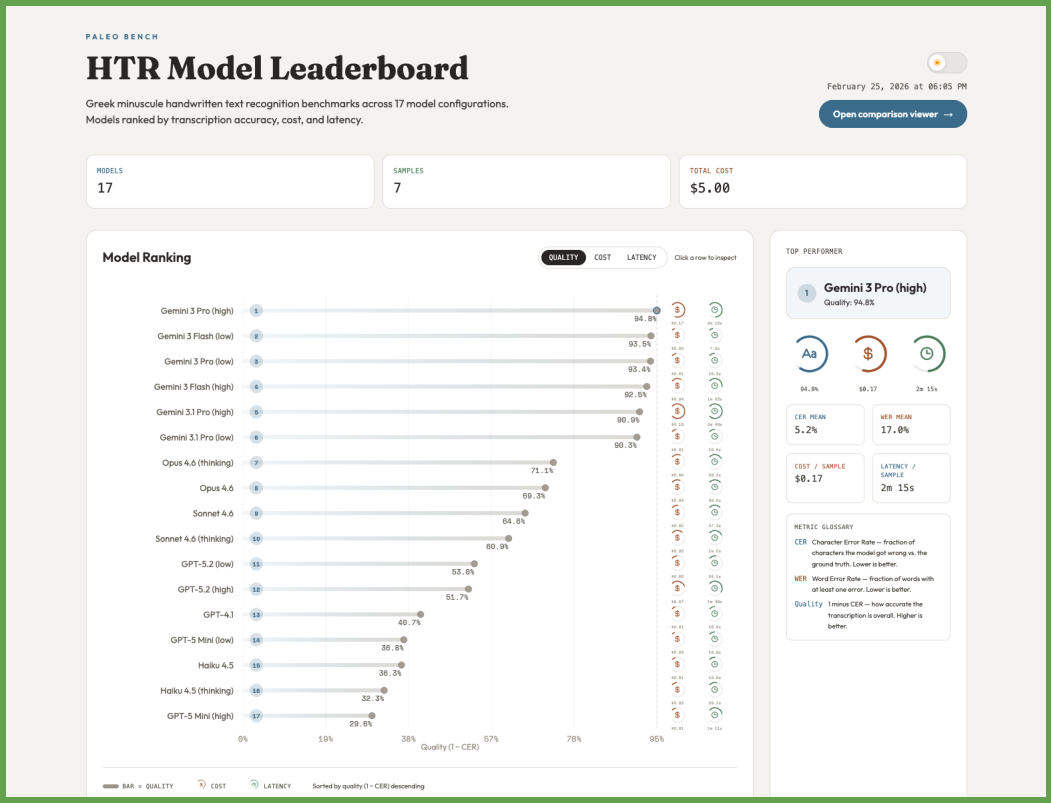
* Percentages based on directories in books_txt_full. Some books cross-listed.

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





https://menus.nypl.org/



[Paleo Bench Leaderboard], ranking Greek minuscule HTR performance

Fonts:

Klima

Katwijk Mono

both by [Matthew Hinders-Anderson]