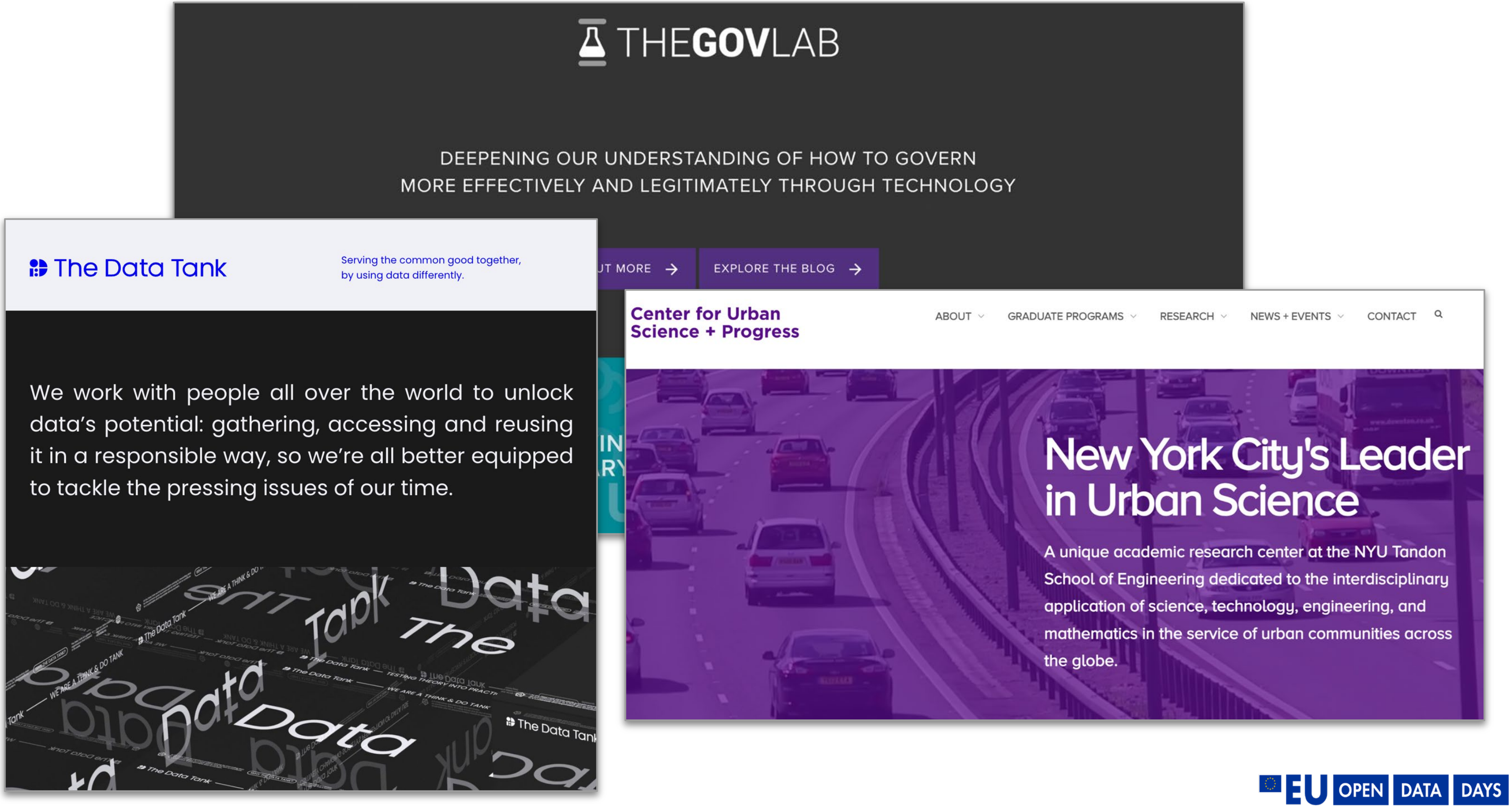
THEGOVLAB

A FOURTH WAVE OF OPEN DATA?
Combining Open Data and AI creatively for Social Impact

Stefaan G. Verhulst

EU Open Data Days, March 20, 2025



DEEPENING OUR UNDERSTANDING OF HOW TO GOVERN MORE EFFECTIVELY AND LEGITIMATELY THROUGH TECHNOLOGY

Serving the common good together, by using data differently.

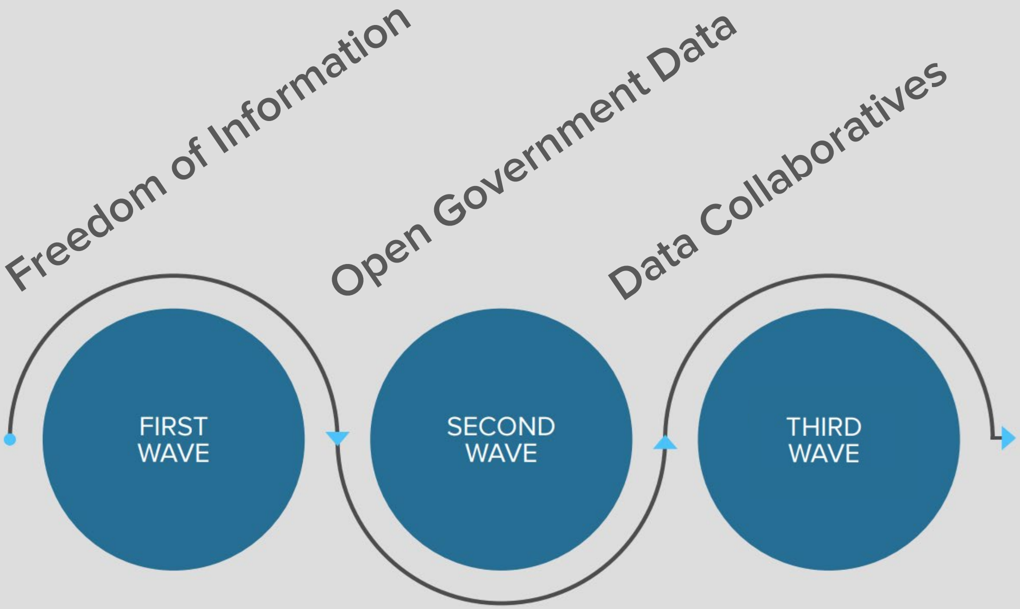
We work with people all over the world to unlock data's potential: gathering, accessing and reusing it in a responsible way, so we're all better equipped to tackle the pressing issues of our time.

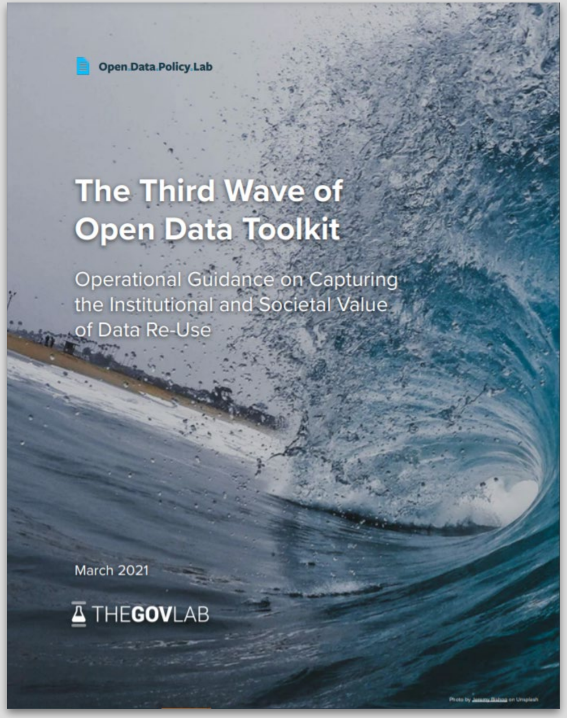
New York City's Leader in Urban Science

A unique academic research center at the NYU Tandon School of Engineering dedicated to the interdisciplinary application of science, technology, engineering, and mathematics in the service of urban communities across the globe.

1. The Past, Present and Future of Open Data

THE EVOLUTION OF OPEN DATA




Open Data Policy Lab

The Third Wave of Open Data Toolkit

Operational Guidance on Capturing the Institutional and Societal Value of Data Re-Use

March 2021

THEGOVLAB

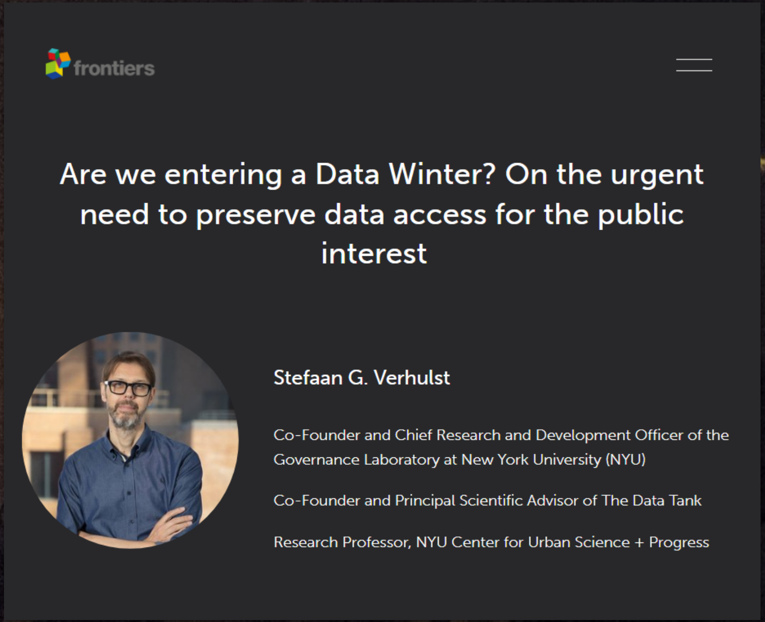
Photo by Jeremy Bishop on Unsplash

bit.ly/3iMnXNI

THE PAST, PRESENT, AND FUTURE OF OPEN DATA

First Wave	Second Wave	Third Wave
Freedom of Information	Open Government Data	Re-use of Public and Private Data
Data on Request (Right to Know)	Open by Default (Right to Share)	Publish with Purpose
Pull Focused	Push Focused	Partnerships
National	International and National	Alignment Subnational and National

02. Today's "Backwash": An Approaching Data Winter



Are we entering a Data Winter? On the urgent need to preserve data access for the public interest

Stefaan G. Verhulst

Co-Founder and Chief Research and Development Officer of the Governance Laboratory at New York University (NYU)

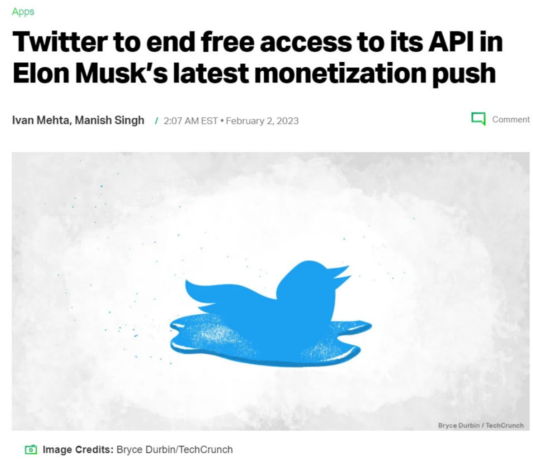
Co-Founder and Principal Scientific Advisor of The Data Tank

Research Professor, NYU Center for Urban Science + Progress

Changing Dynamics: A Data Winter?

https://unsplash.com/photos/person-walking-at-night-with-snow-UaxkX3rJh68

EXHIBIT 1: ACCESS TO PLATFORM DATA



Apps

Twitter to end free access to its API in Elon Musk's latest monetization push

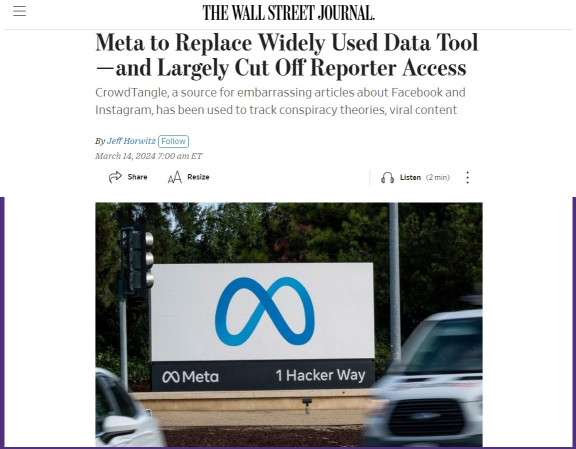
Ivan Mehta, Manish Singh / 2:07 AM EST • February 2, 2023

Comment

Bryce Durbin / TechCrunch

Image Credits: Bryce Durbin/TechCrunch

https://bit.ly/3JQ6rCj



THE WALL STREET JOURNAL.

Meta to Replace Widely Used Data Tool —and Largely Cut Off Reporter Access

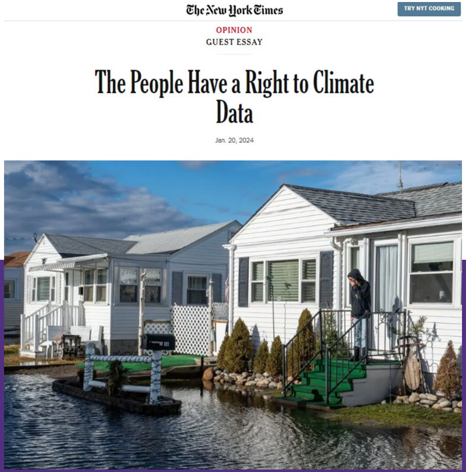
CrowdTangle, a source for embarrassing articles about Facebook and Instagram, has been used to track conspiracy theories, viral content

By *Jeff Horwitz* Follow

March 14, 2024 7:00 am ET

bit.ly/4b6bZoc

EXHIBIT 2: ACCESS TO CLIMATE DATA



The New York Times

TRY NYT COOKING

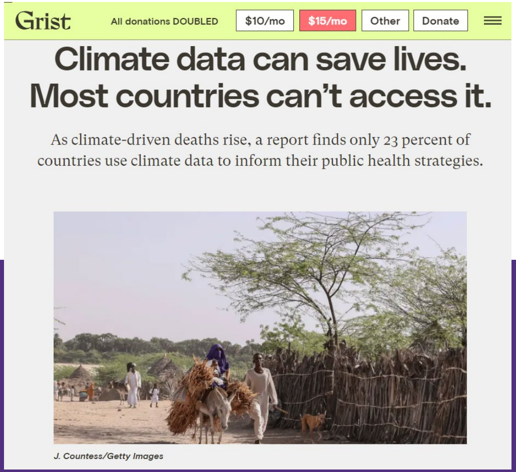
OPINION

GUEST ESSAY

The People Have a Right to Climate Data

Jan. 20, 2024

bit.ly/4cuzNTK



Grist

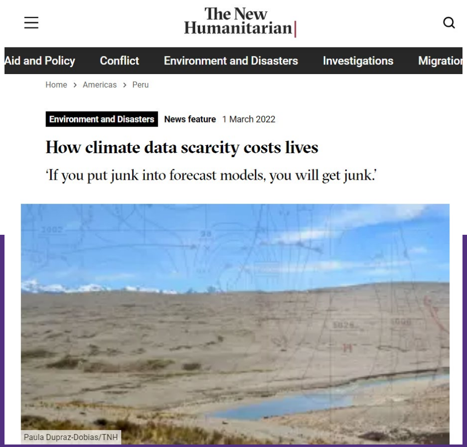
All donations DOUBLED $10/mo $15/mo Other Donate

Climate data can save lives. Most countries can't access it.

As climate-driven deaths rise, a report finds only 23 percent of countries use climate data to inform their public health strategies.

J. Countess/Getty Images

bit.ly/3KcW3EV



The New Humanitarian

Aid and Policy | Conflict | Environment and Disasters | Investigations | Migratio

Home › Americas › Peru

Environment and Disasters News feature 1 March 2022

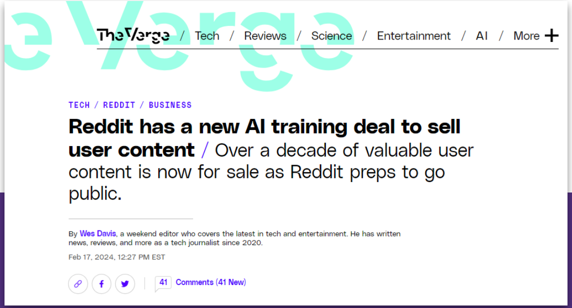
How climate data scarcity costs lives

'If you put junk into forecast models, you will get junk.'

Paula Dupraz-Dobias/TNH

bit.ly/4b85st7

EXHIBIT 3: Generative AI-nxiety



The Verge / Tech / Reviews / Science / Entertainment / AI / More +

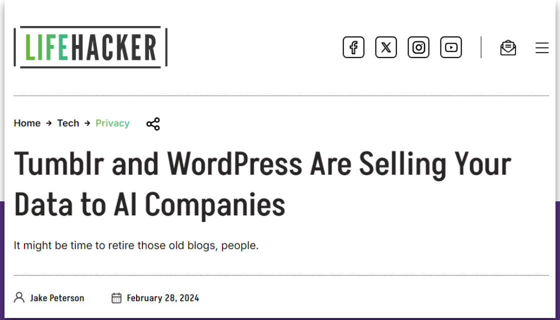
TECH / REDDIT / BUSINESS

Reddit has a new AI training deal to sell user content / Over a decade of valuable user content is now for sale as Reddit preps to go public.

By Wes Davis, a weekend editor who covers the latest in tech and entertainment. He has written news, reviews, and more as a tech journalist since 2020.

Feb 17, 2024, 12:27 PM EST

41 Comments (41 New)



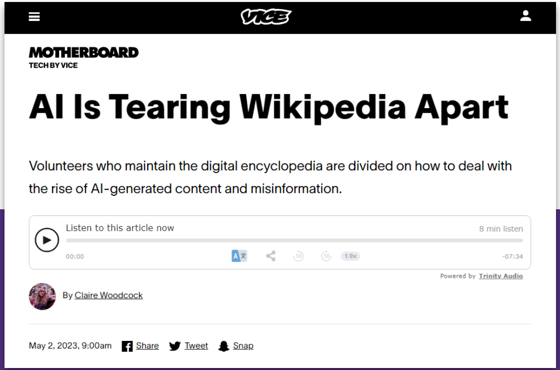
LIFEHACKER

Home → Tech → Privacy

Tumblr and WordPress Are Selling Your Data to AI Companies

It might be time to retire those old blogs, people.

Jake Peterson February 28, 2024



VICE

MOTHERBOARD
TECH BY VICE

AI Is Tearing Wikipedia Apart

Volunteers who maintain the digital encyclopedia are divided on how to deal with the rise of AI-generated content and misinformation.

Listen to this article now 8 min listen

00:00 1.0x -07:34

Powered by Trinity Audio

By Claire Woodcock

May 2, 2023, 9:00am Share Tweet Snap

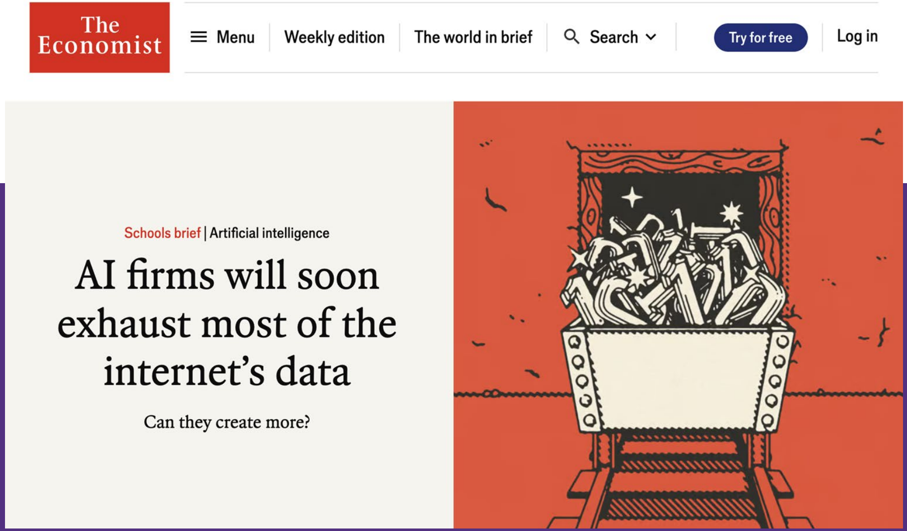
Reddit selling access to data

Tumblr and Wordpress being used as training data

Fears at Wikipedia

EU OPEN DATA DAYS

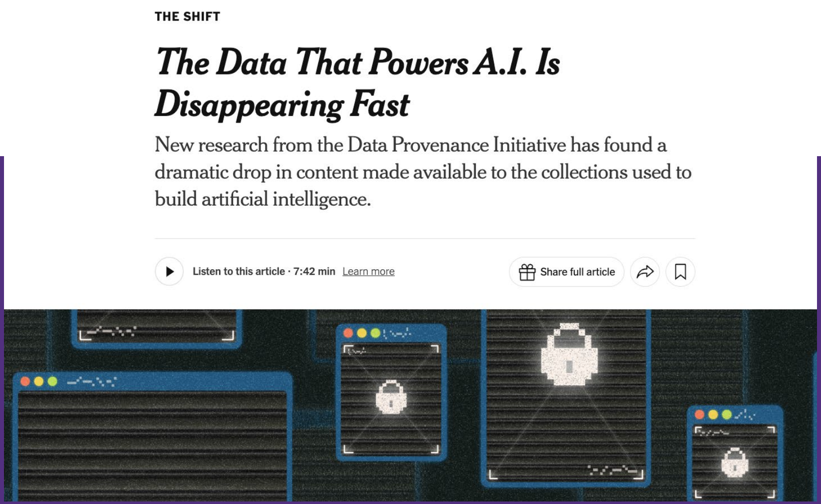
EXHIBIT 5: ACCESS TO TRAINING DATA



Schools brief | Artificial intelligence

AI firms will soon exhaust most of the internet's data

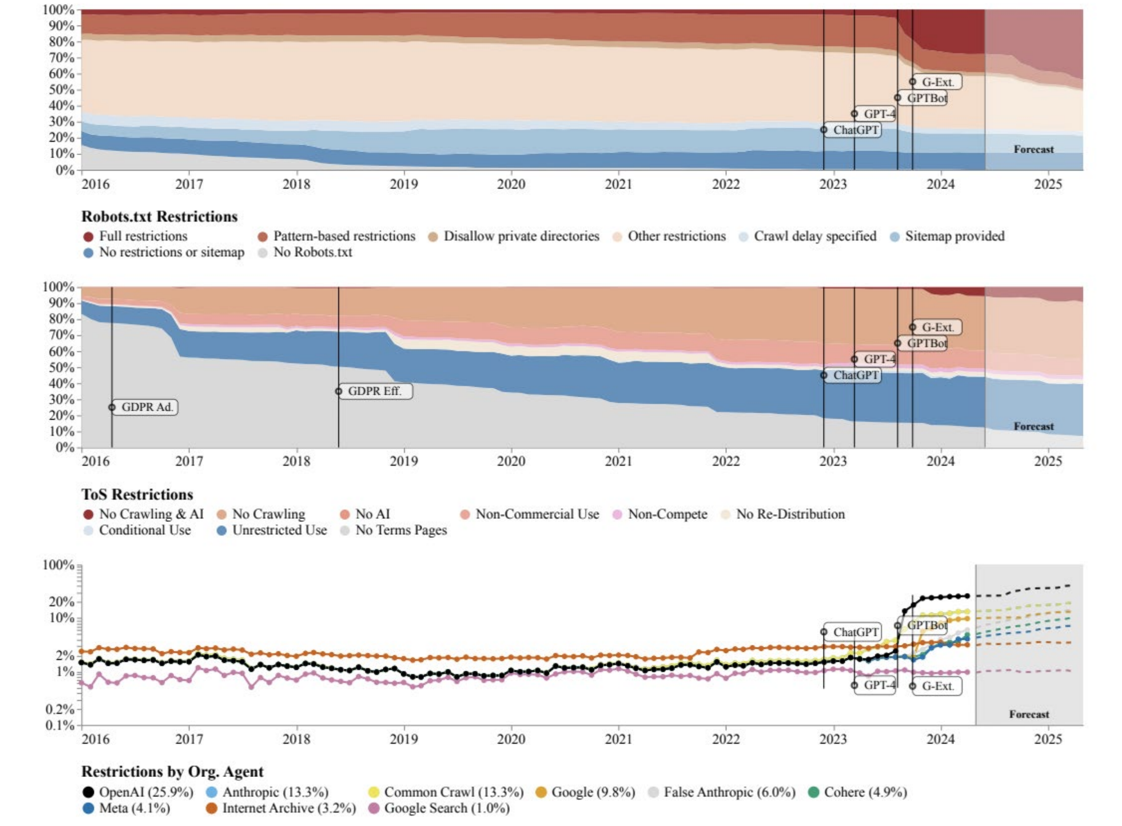
Can they create more?


THE SHIFT

The Data That Powers A.I. Is Disappearing Fast

New research from the Data Provenance Initiative has found a dramatic drop in content made available to the collections used to build artificial intelligence.

EXHIBIT 6: CLOSURE OF THE PUBLIC WEB



OpenData Barometer

Data for everyone

Data people need

Data people can use

Global Report
Fourth Edition

opendatabarometer.org

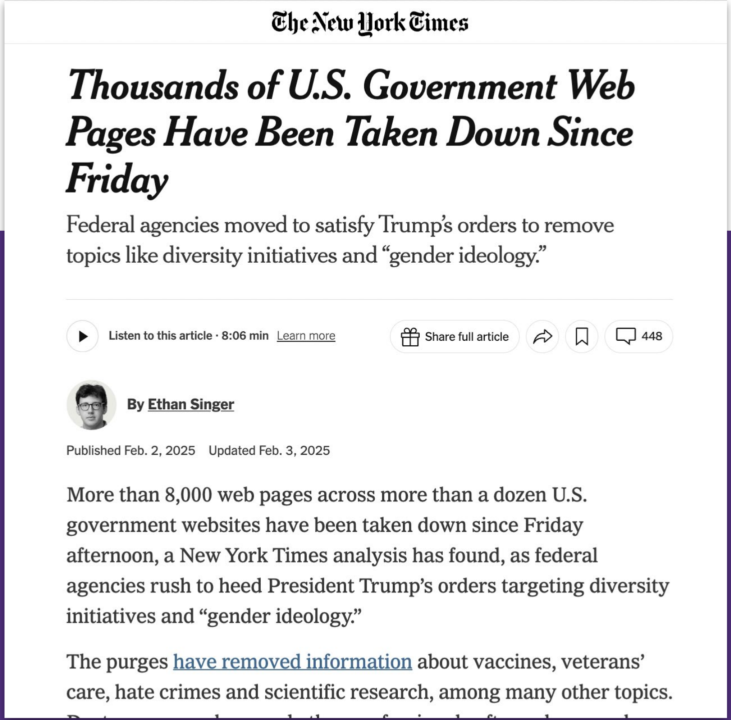
WORLD WIDE WEB
FOUNDATION

EXHIBIT 7: Open Data Barometer

"…this year's Barometer shows that governments are slowing and stalling in their commitment to open data. In some cases, progress has even been undone."

– Open Data Barometer, bit.ly/3y6Vx8u

EXHIBIT 8: GOV SITES BEING TAKEN DOWN



The New York Times

Thousands of U.S. Government Web Pages Have Been Taken Down Since Friday

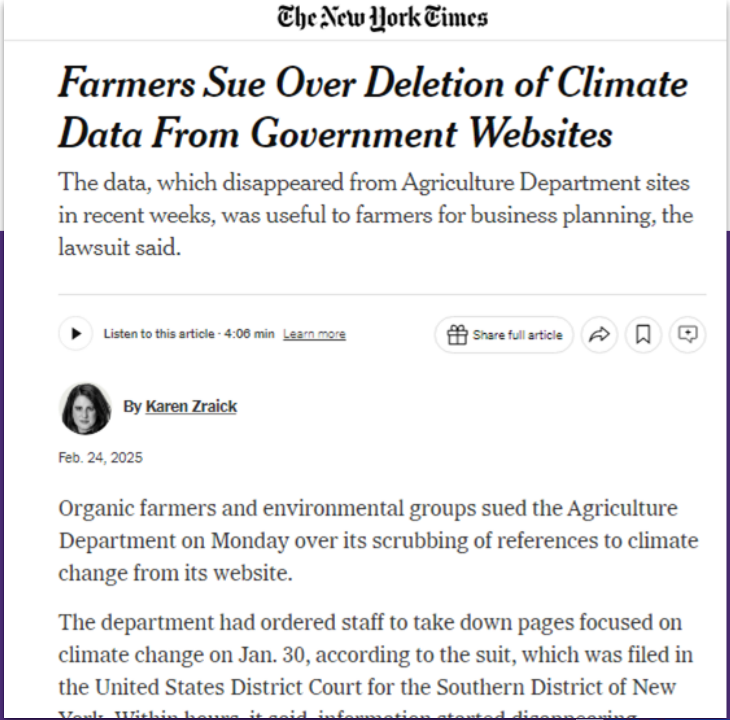
Federal agencies moved to satisfy Trump's orders to remove topics like diversity initiatives and "gender ideology."

By Ethan Singer

Published Feb. 2, 2025 Updated Feb. 3, 2025

More than 8,000 web pages across more than a dozen U.S. government websites have been taken down since Friday afternoon, a New York Times analysis has found, as federal agencies rush to heed President Trump's orders targeting diversity initiatives and "gender ideology."

The purges have removed information about vaccines, veterans' care, hate crimes and scientific research, among many other topics.



The New York Times

Farmers Sue Over Deletion of Climate Data From Government Websites

The data, which disappeared from Agriculture Department sites in recent weeks, was useful to farmers for business planning, the lawsuit said.

By Karen Zraick

Feb. 24, 2025

Organic farmers and environmental groups sued the Agriculture Department on Monday over its scrubbing of references to climate change from its website.

The department had ordered staff to take down pages focused on climate change on Jan. 30, according to the suit, which was filed in the United States District Court for the Southern District of New York. Within hours, it said, information started disappearing.

03. A Fourth Wave?

OPEN DATA AND GENERATIVE AI

Improving AI

- Quality of AI Output
- Augmentation of AI (RAG)
- Public Interest Use Cases

A Fourth Wave of Open Data?
Exploring the Spectrum of Scenarios for
Open Data and Generative AI
Hannah Chafetz, Sampriti Saxena, and Stefaan G. Verhulst
May 2024

THEGOVLAB Open.Data.Policy.Lab

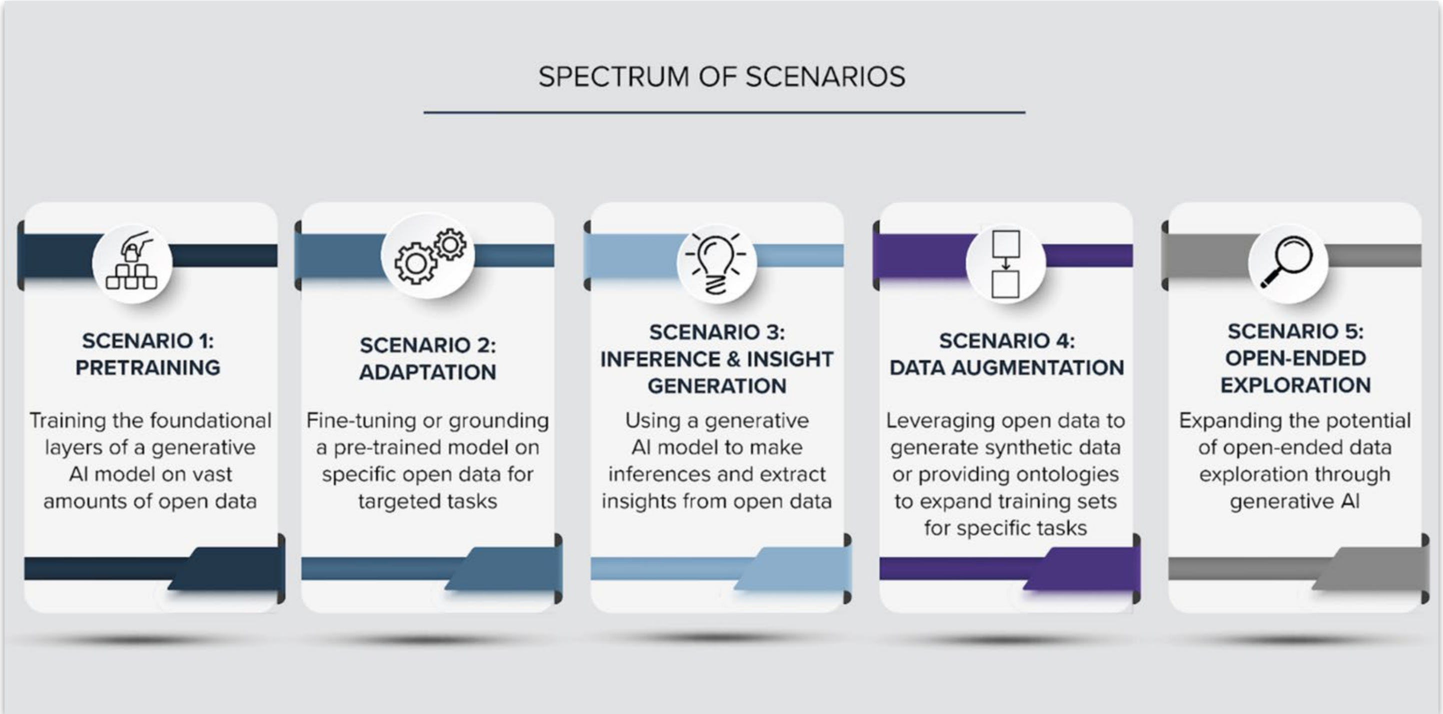
Democratizing Data

- Conversational Interfaces
- Synthetic Data and Data Quality
- Code Book Generation

https://arxiv.org/abs/2405.04333

Improving AI

IMPROVING AI AND PUBLIC INTEREST USE CASES



SPECTRUM OF SCENARIOS

SCENARIO 1: PRETRAINING

Training the foundational layers of a generative AI model on vast amounts of open data

SCENARIO 2: ADAPTATION

Fine-tuning or grounding a pre-trained model on specific open data for targeted tasks

SCENARIO 3: INFERENCE & INSIGHT GENERATION

Using a generative AI model to make inferences and extract insights from open data

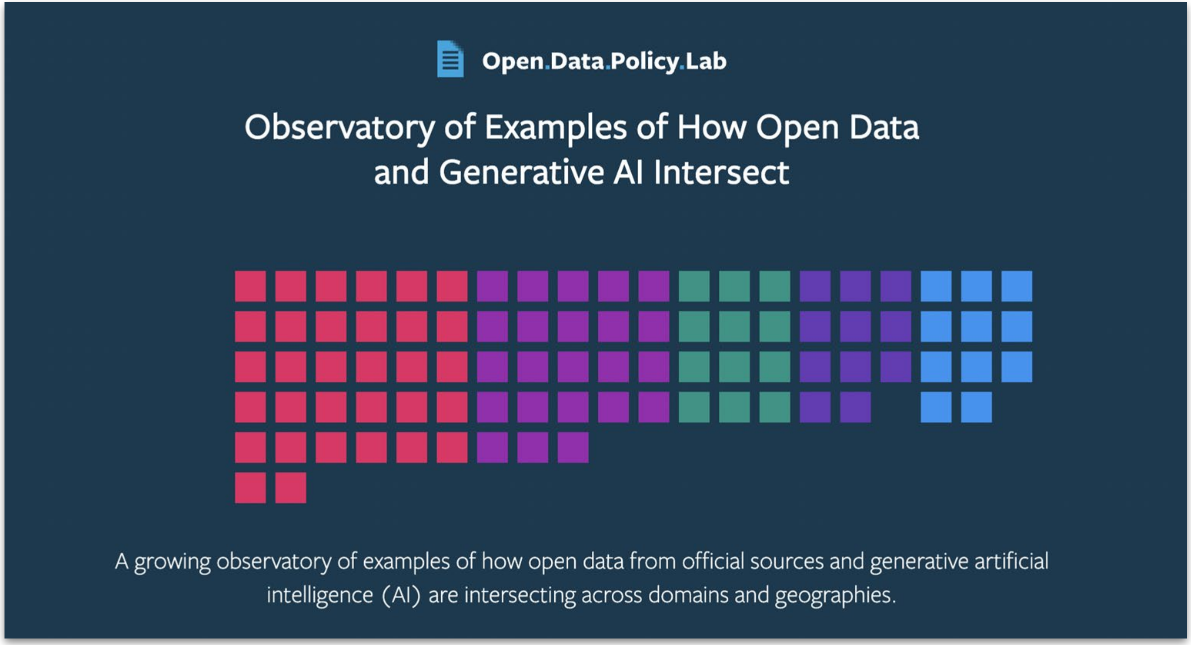
SCENARIO 4: DATA AUGMENTATION

Leveraging open data to generate synthetic data or providing ontologies to expand training sets for specific tasks

SCENARIO 5: OPEN-ENDED EXPLORATION

Expanding the potential of open-ended data exploration through generative AI

OPEN DATA AND GENERATIVE AI CAN INTERSECT IN SEVERAL WAYS



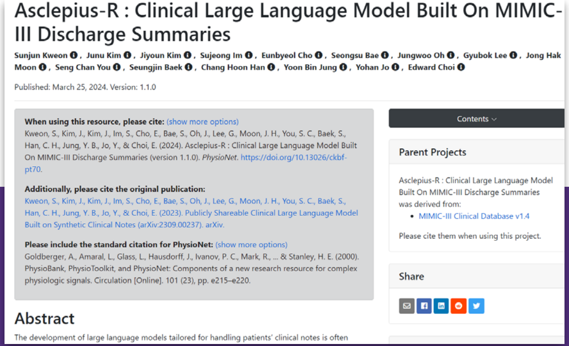
Open.Data.Policy.Lab

Observatory of Examples of How Open Data and Generative AI Intersect

A growing observatory of examples of how open data from official sources and generative artificial intelligence (AI) are intersecting across domains and geographies.

https://repository.opendatapolicylab.org/genai/

AUGMENTATION OF AI (RAG) FOR PUBLIC INTEREST USE CASES



Asclepius-R : Clinical Large Language Model Built On MIMIC-III Discharge Summaries

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, Edward Choi

Published: March 25, 2024. Version: 1.1.0

When using this resource, please cite: (show more options)
Kweon, S., Kim, J., Kim, J., Im, S., Cho, E., Bae, S., Oh, J., Lee, G., Moon, J. H., You, S. C., Baek, S., Han, C. H., Jung, Y. B., Jo, Y., & Choi, E. (2024). Asclepius-R : Clinical Large Language Model Built On MIMIC-III Discharge Summaries (version 1.1.0). *PhysioNet*. https://doi.org/10.13026/ckbf-pt70.

Additionally, please cite the original publication:
Kweon, S., Kim, J., Kim, J., Im, S., Cho, E., Bae, S., Oh, J., Lee, G., Moon, J. H., You, S. C., Baek, S., Han, C. H., Jung, Y. B., Jo, Y., & Choi, E. (2023). Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes (arXiv:2309.00237). arXiv.

Please include the standard citation for PhysioNet: (show more options)
Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Abstract

The development of large language models tailored for handling patients' clinical notes is often

Contents

Parent Projects

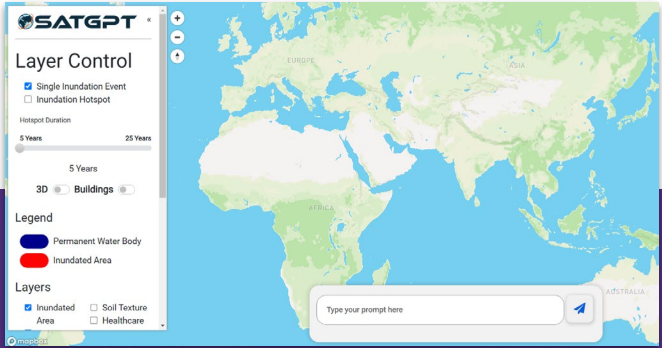
Asclepius-R : Clinical Large Language Model Built On MIMIC-III Discharge Summaries was derived from:

- MIMIC-III Clinical Database v1.4

Please cite them when using this project.

Share

bit.ly/4iml6V0



SATGPT

Layer Control

☑ Single Inundation Event
☐ Inundation Hotspot

Hotspot Duration

5 Years — 25 Years

5 Years

3D ○ Buildings ○

Legend

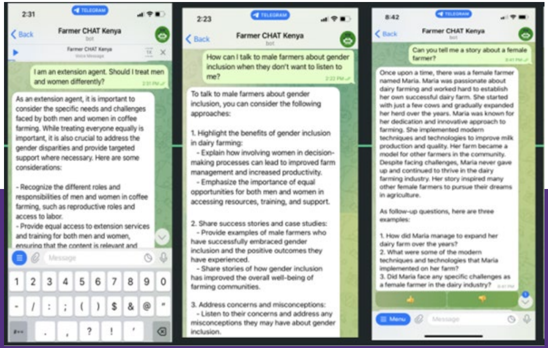
🟦 Permanent Water Body
🟥 Inundated Area

Layers

☑ Inundated Area
☐ Soil Texture
☐ Healthcare

Type your prompt here

satgpt.net



2:31 — Farmer CHAT Kenya (bot)

> I am an extension agent. Should I treat men and women differently?

As an extension agent, it is important to consider the specific needs and challenges faced by both men and women in coffee farming. While treating everyone equally is important, it is also crucial to address the gender disparities and provide targeted support where necessary. Here are some considerations:

- Recognize the different roles and responsibilities of men and women in coffee farming, such as reproductive roles and access to labor.
- Provide equal access to extension services and training for both men and women, ensuring that the content is relevant and

2:23 — Farmer CHAT Kenya (bot)

> How can I talk to male farmers about gender inclusion when they don't want to listen to me?

To talk to male farmers about gender inclusion, you can consider the following approaches:

1. Highlight the benefits of gender inclusion in dairy farming:
   - Explain how involving women in decision-making processes can lead to improved farm management and increased productivity.
   - Emphasize the importance of equal opportunities for both men and women in accessing resources, training, and support.

2. Share success stories and case studies:
   - Provide examples of male farmers who have successfully embraced gender inclusion and the positive outcomes they have experienced.
   - Share stories of how gender inclusion has improved the overall well-being of farming communities.

3. Address concerns and misconceptions:
   - Listen to their concerns and address any misconceptions they may have about gender inclusion.

8:42 — Farmer CHAT Kenya (bot)

> Can you tell me a story about a female farmer?

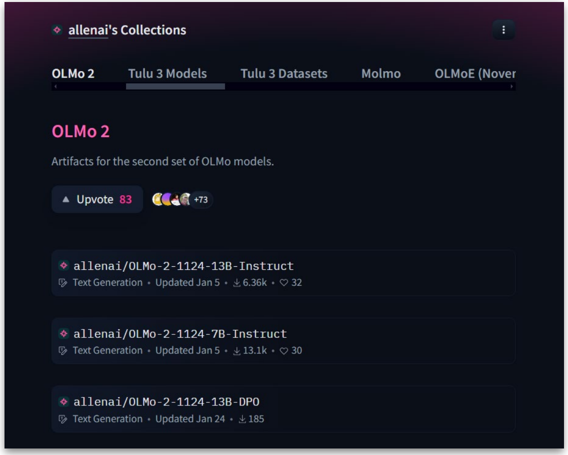
Once upon a time, there was a female farmer named Maria. Maria was passionate about dairy farming and worked hard to establish her own successful dairy farm. She started with just a few cows and gradually expanded her herd over the years. Maria was known for her dedication and innovative approach to farming. She implemented modern techniques and technologies to improve milk production and quality. Her farm became a model for other farmers in the community. Despite facing challenges, Maria never gave up and continued to thrive in the dairy farming industry. Her story inspired many other female farmers to pursue their dreams in agriculture.

As follow-up questions, here are three examples:

1. How did Maria manage to expand her dairy farm over the years?
2. What were some of the modern techniques and technologies that Maria implemented on her farm?
3. Did Maria face any specific challenges as a female farmer in the dairy industry?

https://arxiv.org/pdf/2409.08916

OPEN DATA ENABLES OPEN SOURCE AI


allenai's Collections

OLMo 2 | Tulu 3 Models | Tulu 3 Datasets | Molmo | OLMoE (Nover

OLMo 2

Artifacts for the second set of OLMo models.

▲ Upvote 83 +73

allenai/OLMo-2-1124-13B-Instruct
Text Generation • Updated Jan 5 • 6.36k • ♡ 32

allenai/OLMo-2-1124-7B-Instruct
Text Generation • Updated Jan 5 • 13.1k • ♡ 30

allenai/OLMo-2-1124-13B-DPO
Text Generation • Updated Jan 24 • 185

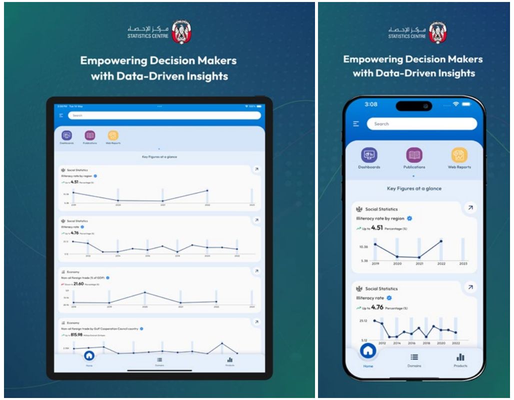
https://allenai.org/olmo

The Allen Institute for AI have been creating **OLMo**, an open language model to promote the study of large-scale NLP systems.

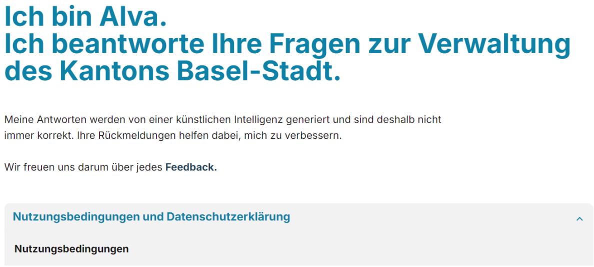
The model relies on an open dataset known as **Dolma**, which is a diverse mix of web content, academic publications, code, books, and encyclopedic materials. Openly available for download on the HuggingFace Hub, Dolma is the largest open dataset to date.

Democratizing Data

DEMOCRATIZING DATA : CONVERSATIONAL INTERFACE



Bayaan Platform



Ich bin Alva.
Ich beantworte Ihre Fragen zur Verwaltung des Kantons Basel-Stadt.

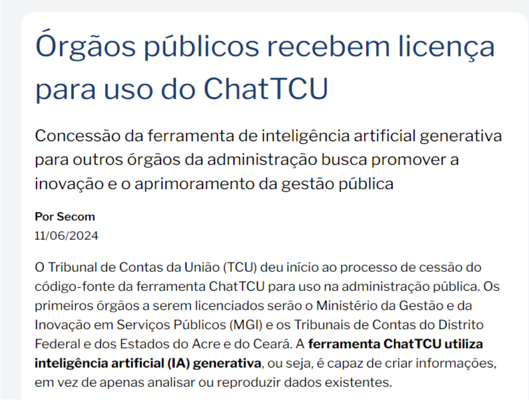
Meine Antworten werden von einer künstlichen Intelligenz generiert und sind deshalb nicht immer korrekt. Ihre Rückmeldungen helfen dabei, mich zu verbessern.

Wir freuen uns darum über jedes **Feedback.**

Nutzungsbedingungen und Datenschutzerklärung

Nutzungsbedingungen

Alva



Órgãos públicos recebem licença para uso do ChatTCU

Concessão da ferramenta de inteligência artificial generativa para outros órgãos da administração busca promover a inovação e o aprimoramento da gestão pública

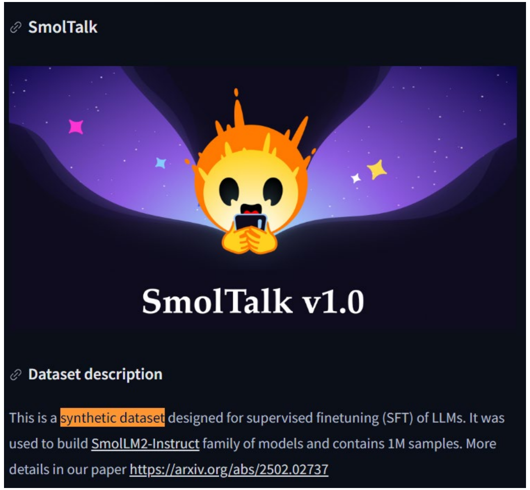
Por Secom
11/06/2024

O Tribunal de Contas da União (TCU) deu início ao processo de cessão do código-fonte da ferramenta ChatTCU para uso na administração pública. Os primeiros órgãos a serem licenciados serão o Ministério da Gestão e da Inovação em Serviços Públicos (MGI) e os Tribunais de Contas do Distrito Federal e dos Estados do Acre e do Ceará. A **ferramenta ChatTCU utiliza inteligência artificial (IA) generativa**, ou seja, é capaz de criar informações, em vez de apenas analisar ou reproduzir dados existentes.

ChatTCU

DEMOCRATIZING AI: SYNTHETIC DATA AND FILLING IN DATA GAPS

Synthetic data generation can fill data gaps

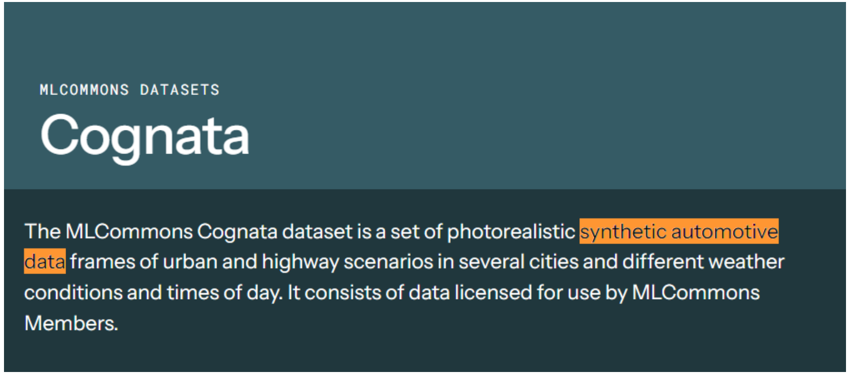


SmolTalk

Dataset description

This is a synthetic dataset designed for supervised finetuning (SFT) of LLMs. It was used to build SmolLM2-Instruct family of models and contains 1M samples. More details in our paper https://arxiv.org/abs/2502.02737

Smoltalk
https://bit.ly/4i1OyQj



MLCOMMONS DATASETS

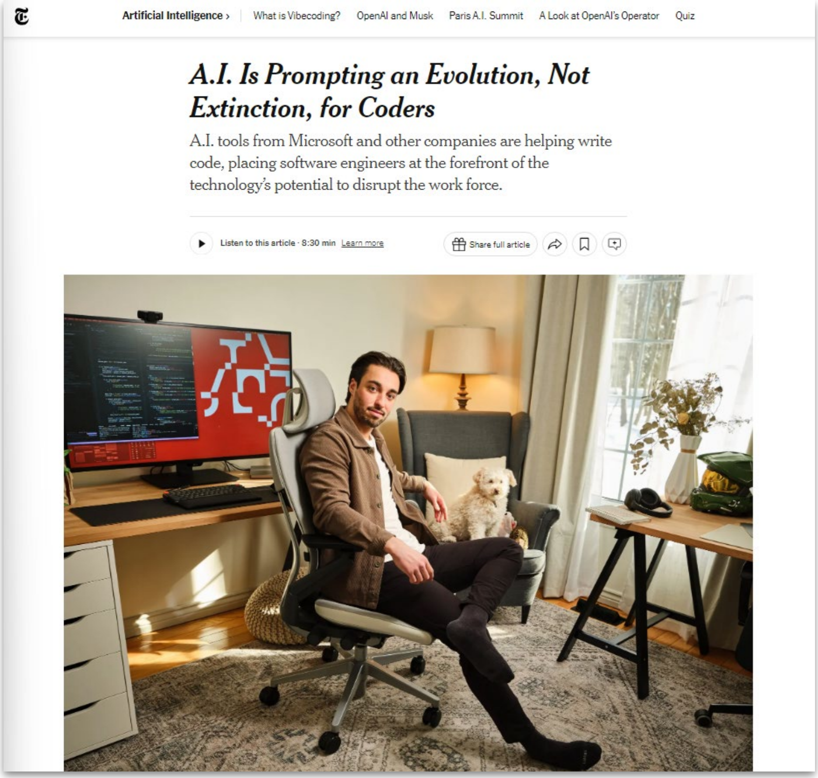
Cognata

The MLCommons Cognata dataset is a set of photorealistic synthetic automotive data frames of urban and highway scenarios in several cities and different weather conditions and times of day. It consists of data licensed for use by MLCommons Members.

ML Commons Cognata
https://mlcommons.org/datasets/cognata/

DEMOCRATIZING AI: CODE BOOK DEVELOPMENT

- Generative AI can expedite the creation of code books or metadata documents describing datasets.
- This automation facilitates users to comprehend the data better, significantly reducing the time and effort needed to exploit open data.



Artificial Intelligence ›
What is Vibecoding?
OpenAI and Musk
Paris A.I. Summit
A Look at OpenAI's Operator
Quiz

A.I. Is Prompting an Evolution, Not Extinction, for Coders

A.I. tools from Microsoft and other companies are helping write code, placing software engineers at the forefront of the technology's potential to disrupt the work force.

Listen to this article · 8:30 min Learn more
Share full article

DEMOCRATIZING DATA: IMPROVING DATA QUALITY

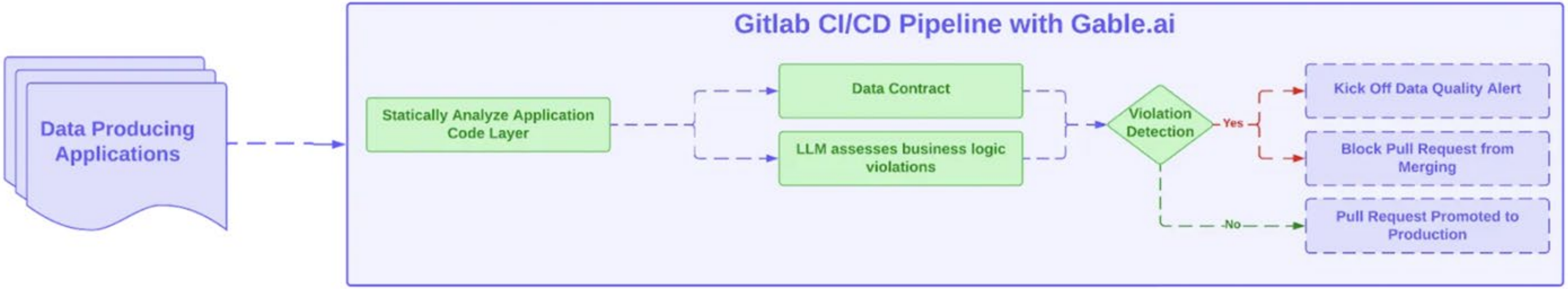
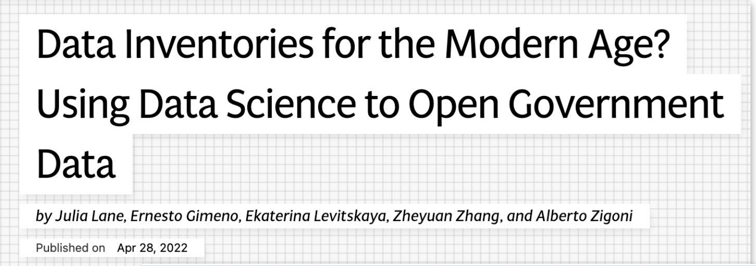

Gitlab CI/CD Pipeline with Gable.ai
Data Producing Applications
Statically Analyze Application Code Layer
Data Contract
LLM assesses business logic violations
Violation Detection
Yes
No
Kick Off Data Quality Alert
Block Pull Request from Merging
Pull Request Promoted to Production

Figure 1. Enhancing quality guards at the application code layer.

https://medium.com/glassdoor -engineering/data -quality -at-petabyte -scale-building -trust -in-the-data-lifecycle -7052361307a4

DEMOCRATIZING DATA: DATA DISCOVERY



Data Inventories for the Modern Age? Using Data Science to Open Government Data

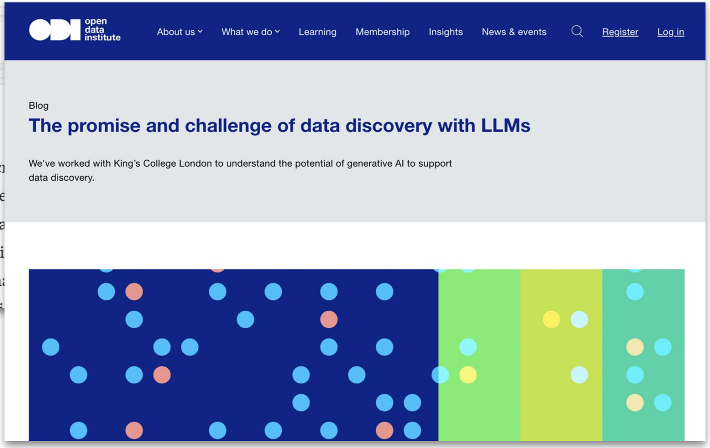
by Julia Lane, Ernesto Gimeno, Ekaterina Levitskaya, Zheyuan Zhang, and Alberto Zigoni

Published on Apr 28, 2022

last released
3 years ago

ABSTRACT

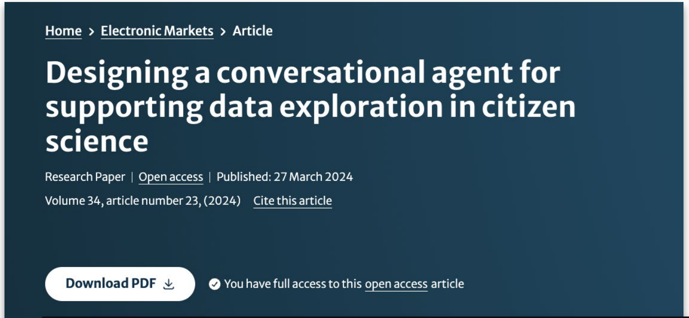
This article desc
natural language
government data
established—usi
new and sustain
States and on ci



Blog

The promise and challenge of data discovery with LLMs

We've worked with King's College London to understand the potential of generative AI to support data discovery.

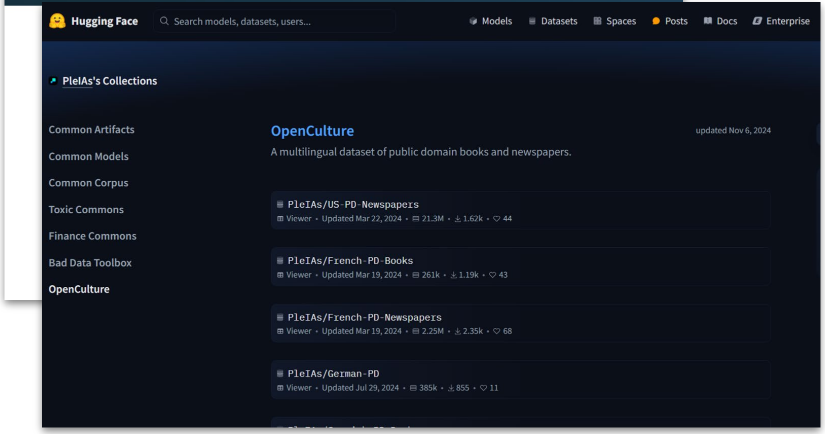


Home › Electronic Markets › Article

Designing a conversational agent for supporting data exploration in citizen science

Research Paper | Open access | Published: 27 March 2024

Volume 34, article number 23, (2024) Cite this article



PleIAs's Collections

Common Artifacts
Common Models
Common Corpus
Toxic Commons
Finance Commons
Bad Data Toolbox
OpenCulture

OpenCulture updated Nov 6, 2024

A multilingual dataset of public domain books and newspapers.

PleIAs/US-PD-Newspapers
Viewer • Updated Mar 22, 2024 • 21.3M • 1.62k • ♡ 44

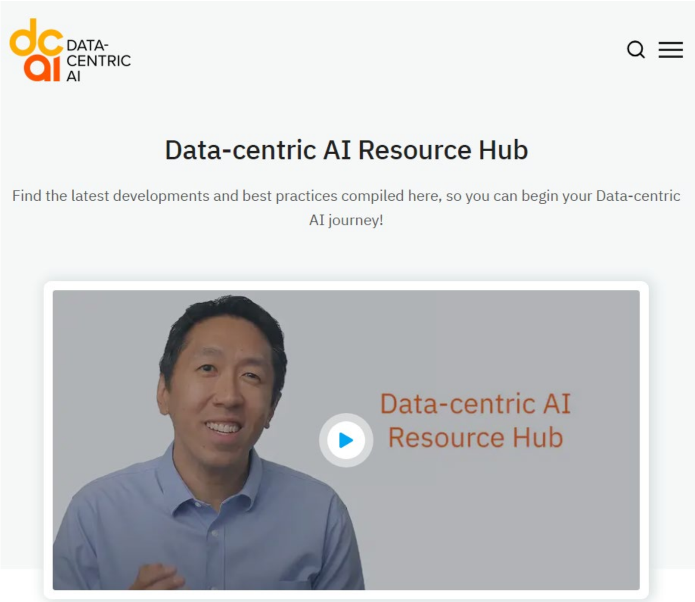
PleIAs/French-PD-Books
Viewer • Updated Mar 19, 2024 • 261k • 1.19k • ♡ 43

PleIAs/French-PD-Newspapers
Viewer • Updated Mar 19, 2024 • 2.25M • 2.35k • ♡ 68

PleIAs/German-PD
Viewer • Updated Jul 29, 2024 • 385k • 855 • ♡ 11

04. Riding the Wave?

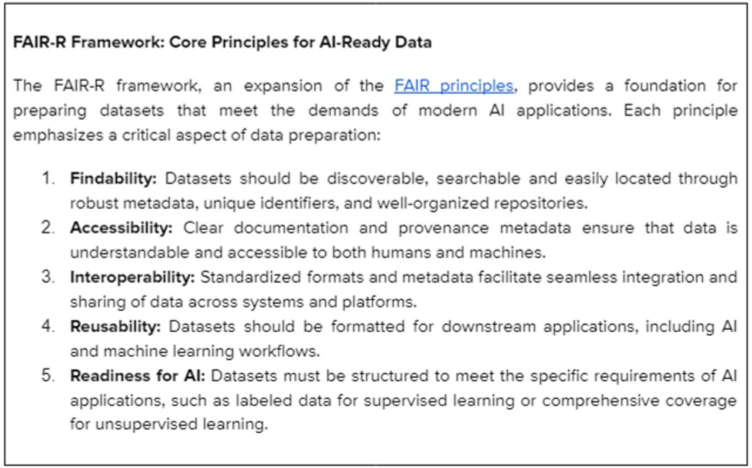
(1) MAKING DATA AI-READY:



Data-centric AI Resource Hub

Find the latest developments and best practices compiled here, so you can begin your Data-centric AI journey!

Data-centric AI Resource Hub



FAIR-R Framework: Core Principles for AI-Ready Data

The FAIR-R framework, an expansion of the FAIR principles, provides a foundation for preparing datasets that meet the demands of modern AI applications. Each principle emphasizes a critical aspect of data preparation:

1. **Findability:** Datasets should be discoverable, searchable and easily located through robust metadata, unique identifiers, and well-organized repositories.
2. **Accessibility:** Clear documentation and provenance metadata ensure that data is understandable and accessible to both humans and machines.
3. **Interoperability:** Standardized formats and metadata facilitate seamless integration and sharing of data across systems and platforms.
4. **Reusability:** Datasets should be formatted for downstream applications, including AI and machine learning workflows.
5. **Readiness for AI:** Datasets must be structured to meet the specific requirements of AI applications, such as labeled data for supervised learning or comprehensive coverage for unsupervised learning.

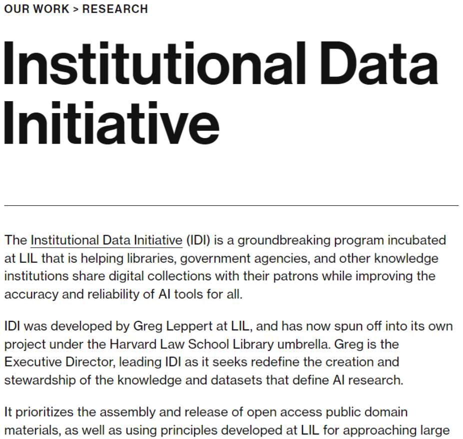
FAIR-R Principles

(2) IMPROVE PROVENANCE:

DEVELOPING A PUBLIC-INTEREST TRAINING COMMONS OF BOOKS

Posted December 5, 2024

Public-Interest Book Training Commons



OUR WORK > RESEARCH

Institutional Data Initiative

The Institutional Data Initiative (IDI) is a groundbreaking program incubated at LIL that is helping libraries, government agencies, and other knowledge institutions share digital collections with their patrons while improving the accuracy and reliability of AI tools for all.

IDI was developed by Greg Leppert at LIL, and has now spun off into its own project under the Harvard Law School Library umbrella. Greg is the Executive Director, leading IDI as it seeks redefine the creation and stewardship of the knowledge and datasets that define AI research.

It prioritizes the assembly and release of open access public domain materials, as well as using principles developed at LIL for approaching large

Institutional Data Initiative

Data
&Trust
Alliance

Data Provenance
Standards
Executive Overview

07.2024

The first cross-industry metadata standards to bring transparency to the origin and use of datasets for AI and traditional data applications.

Data Provenance Standards — Executive Overview

©2024 Data & Trust Alliance. All rights reserved.

Data Provenance Standard

(3) PRIORITIZE PUBLIC INTEREST USE CASES

Instead of treating all data equally, the focus should be on **high-value datasets** with clear public interest benefits.

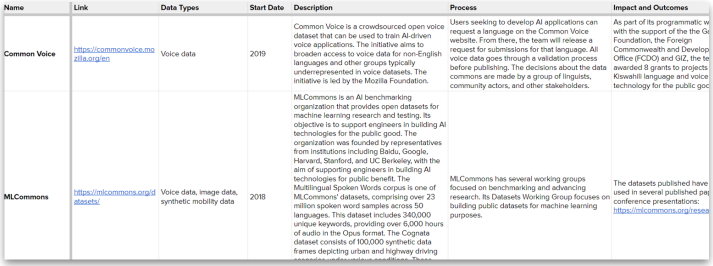
Priority areas include:

- **Climate data** for sustainability efforts.
- **Health data** (e.g., synthetic medical records) for research.
- **Disaster response and pandemic-related data** .

THE 100 QUESTIONS

UNLESS WE DEFINE THE QUESTIONS WELL...
TO UNLOCK THE POTENTIAL OF DATA AND DATA SCIENCE
HOW CAN WE PROVIDE ANSWERS THAT MATTER?

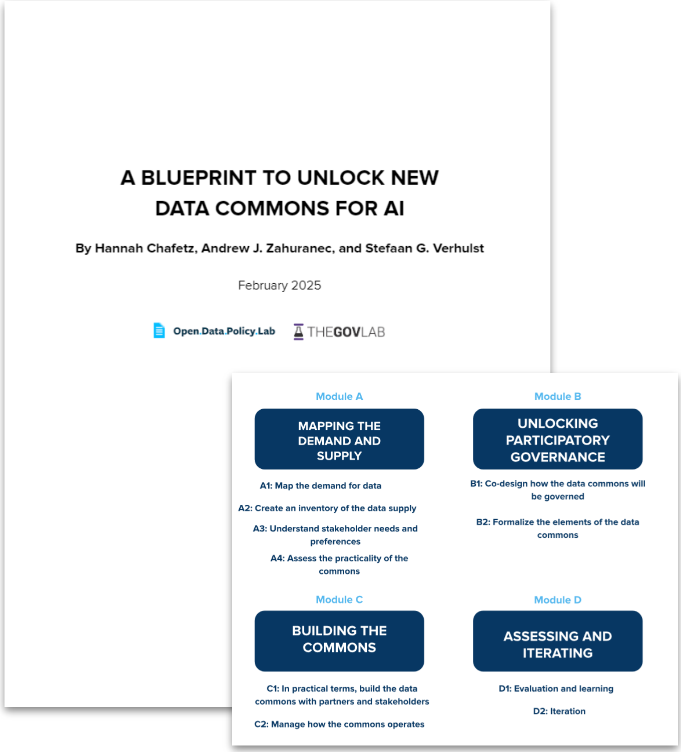
The 100 Questions



Name	Link	Data Types	Start Date	Description	Process	Impact and Outcomes
Common Voice	https://commonvoice.mozilla.org/en	Voice data	2019	Common Voice is a crowdsourced open voice dataset that can be used to train AI-driven voice applications. The initiative aims to broaden access to voice data for non-English languages and other groups typically underrepresented in voice datasets. The initiative is led by the Mozilla Foundation.	Users seeking to develop AI applications can request a language on the Common Voice website. From there, the team will release a request for submissions for that language. All voice data goes through a validation process before publishing. The decisions about the data commons are made by a group of linguists, community actors, and other stakeholders.	As part of its programmatic w with the support of the the Ga Foundation, the Foreign Commonwealth and Develop Office (FCDO) and GIZ, the te awarded 8 grants to projects Kiswahili language and voice technology for the public goc
MLCommons	https://mlcommons.org/datasets/	Voice data, image data, synthetic mobility data	2018	MLCommons is an AI benchmarking organization that provides open datasets for machine learning research and testing. Its objective is to support engineers in building AI technologies for the public good. The organization was founded by representatives from institutions including Baidu, Google, Harvard, Stanford, and UC Berkeley, with the aim of supporting engineers in building AI technologies for public benefit. The Multilingual Spoken Words corpus is one of MLCommons' datasets, comprising over 23 million spoken word samples across 50 languages. This dataset includes 340,000 unique keywords, providing over 6,000 hours of audio in the Opus format. The Cognata dataset consists of 100,000 synthetic data frames depicting urban and highway driving scenarios under various conditions. These	MLCommons has several working groups focused on benchmarking and advancing research. Its Datasets Working Group focuses on building public datasets for machine learning purposes.	The datasets published have used in several published pa conference presentations: https://mlcommons.org/resea

Use Case Mapping

(4) ESTABLISH DATA COMMONS



A BLUEPRINT TO UNLOCK NEW DATA COMMONS FOR AI

By Hannah Chafetz, Andrew J. Zahuranec, and Stefaan G. Verhulst

February 2025

Open Data Policy Lab THEGOVLAB

Module A

MAPPING THE DEMAND AND SUPPLY

A1: Map the demand for data

A2: Create an inventory of the data supply

A3: Understand stakeholder needs and preferences

A4: Assess the practicality of the commons

Module B

UNLOCKING PARTICIPATORY GOVERNANCE

B1: Co-design how the data commons will be governed

B2: Formalize the elements of the data commons

Module C

BUILDING THE COMMONS

C1: In practical terms, build the data commons with partners and stakeholders

C2: Manage how the commons operates

Module D

ASSESSING AND ITERATING

D1: Evaluation and learning

D2: Iteration

Data Governance in Open Source AI

Enabling Responsible and Systemic Access

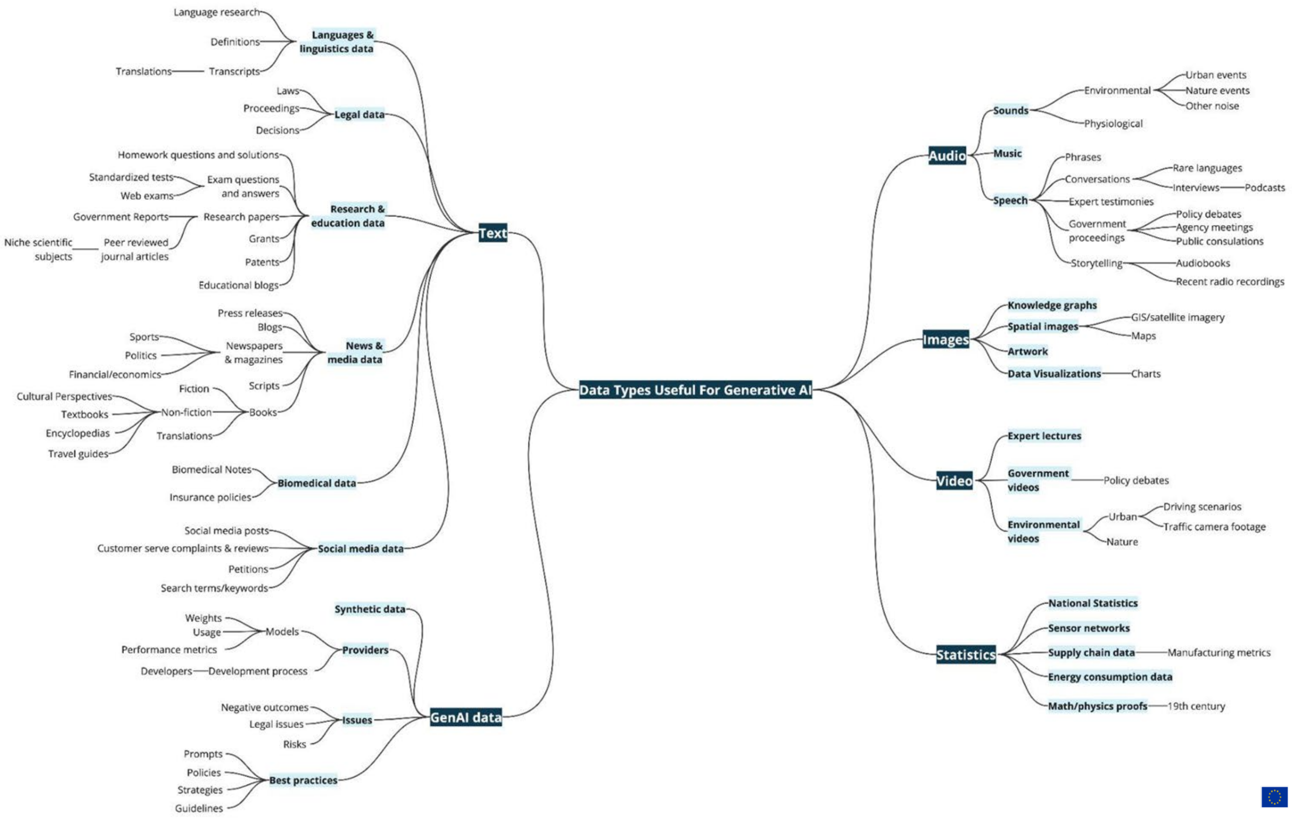
Alek Tarkowski, Open Future

in partnership with the Open Source Initiative

open source initiative OPEN _FUTURE

https://bit.ly/3Dkp2GY

(5) BROADEN OUR CONCEPT OF OPEN DATA



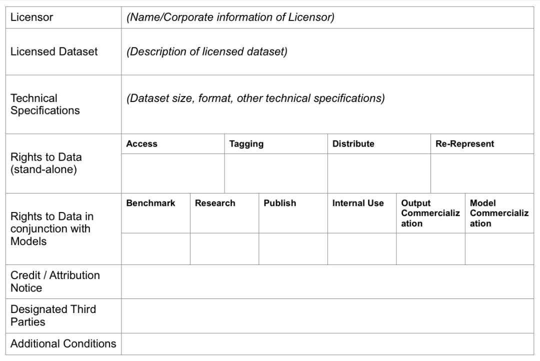
- **Data Types Useful For Generative AI**
  - **Text**
    - **Languages & linguistics data**
      - Language research
      - Definitions
      - Transcripts
        - Translations
    - **Legal data**
      - Laws
      - Proceedings
      - Decisions
    - **Research & education data**
      - Homework questions and solutions
      - Exam questions and answers
        - Standardized tests
        - Web exams
      - Research papers
        - Government Reports
        - Peer reviewed journal articles
          - Niche scientific subjects
      - Grants
      - Patents
      - Educational blogs
    - **News & media data**
      - Press releases
      - Blogs
      - Newspapers & magazines
        - Sports
        - Politics
        - Financial/economics
      - Scripts
      - Books
        - Fiction
        - Non-fiction
          - Cultural Perspectives
          - Textbooks
          - Encyclopedias
          - Travel guides
        - Translations
    - **Biomedical data**
      - Biomedical Notes
      - Insurance policies
    - **Social media data**
      - Social media posts
      - Customer serve complaints & reviews
      - Petitions
      - Search terms/keywords
  - **GenAI data**
    - **Synthetic data**
    - **Providers**
      - Models
        - Weights
        - Usage
        - Performance metrics
      - Development process
        - Developers
    - **Issues**
      - Negative outcomes
      - Legal issues
      - Risks
    - **Best practices**
      - Prompts
      - Policies
      - Strategies
      - Guidelines
  - **Audio**
    - **Sounds**
      - Environmental
        - Urban events
        - Nature events
        - Other noise
      - Physiological
    - **Music**
    - **Speech**
      - Phrases
      - Conversations
        - Rare languages
        - Interviews
          - Podcasts
      - Expert testimonies
      - Government proceedings
        - Policy debates
        - Agency meetings
        - Public consulations
      - Storytelling
        - Audiobooks
        - Recent radio recordings
  - **Images**
    - **Knowledge graphs**
    - **Spatial images**
      - GIS/satellite imagery
      - Maps
    - **Artwork**
    - **Data Visualizations**
      - Charts
  - **Video**
    - **Expert lectures**
    - **Government videos**
      - Policy debates
    - **Environmental videos**
      - Urban
        - Driving scenarios
        - Traffic camera footage
      - Nature
  - **Statistics**
    - **National Statistics**
    - **Sensor networks**
    - **Supply chain data**
      - Manufacturing metrics
    - **Energy consumption data**
    - **Math/physics proofs**
      - 19th century

(6) RETHINKING OPEN DATA LICENSING

Open Data Licensing must start to address:

- **Attribution & provenance** for data used in AI models.
- **Compensation and benefit - sharing** in cases where communities contribute valuable data.

Updating Freedom of Information (FOI) laws to align with AI and digital-era challenges.



Licensor	*(Name/Corporate information of Licensor)*					
Licensed Dataset	*(Description of licensed dataset)*					
Technical Specifications	*(Dataset size, format, other technical specifications)*					
Rights to Data (stand-alone)	**Access**	**Tagging**	**Distribute**	**Re-Represent**		
Rights to Data in conjunction with Models	**Benchmark**	**Research**	**Publish**	**Internal Use**	**Output Commercialization**	**Model Commercialization**
Credit / Attribution Notice						
Designated Third Parties						
Additional Conditions						

Montreal Data License

RESPONSIBLE AI
LICENSES

RAIL

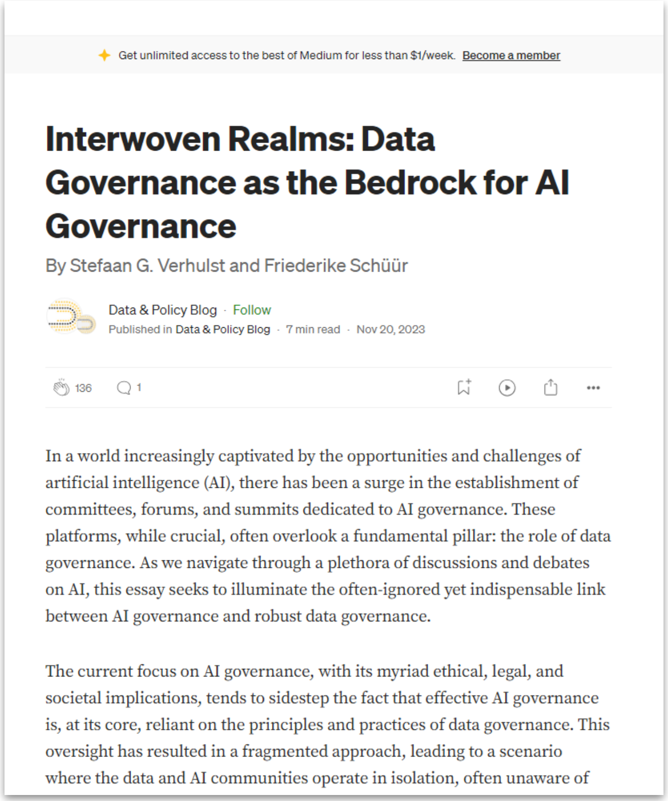
AI2 ImpACT Licenses

Licensing African datasets

Nwulite Obodo Open Data License

(7) UPGRADE DATA GOVERNANCE

- Governments and institutions must **update policies** to reflect the new AI-driven open data landscape.
- Key policy considerations:
  - **AI readiness assessments** for government datasets.
  - **Better legal frameworks** for balancing **privacy, competition, and openness** .
  - **Funding mechanisms** for AI-powered open data initiatives.



Get unlimited access to the best of Medium for less than $1/week. Become a member

Interwoven Realms: Data Governance as the Bedrock for AI Governance

By Stefaan G. Verhulst and Friederike Schüür

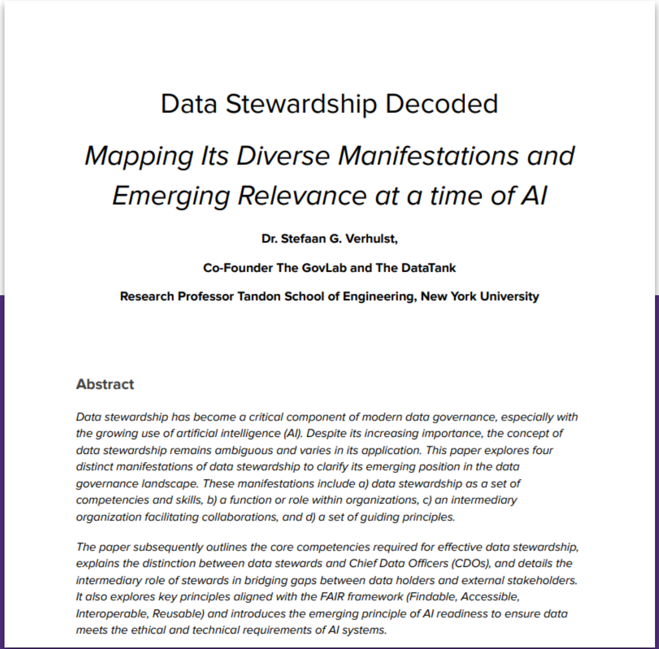
Data & Policy Blog · Follow
Published in Data & Policy Blog · 7 min read · Nov 20, 2023

136 1

In a world increasingly captivated by the opportunities and challenges of artificial intelligence (AI), there has been a surge in the establishment of committees, forums, and summits dedicated to AI governance. These platforms, while crucial, often overlook a fundamental pillar: the role of data governance. As we navigate through a plethora of discussions and debates on AI, this essay seeks to illuminate the often-ignored yet indispensable link between AI governance and robust data governance.

The current focus on AI governance, with its myriad ethical, legal, and societal implications, tends to sidestep the fact that effective AI governance is, at its core, reliant on the principles and practices of data governance. This oversight has resulted in a fragmented approach, leading to a scenario where the data and AI communities operate in isolation, often unaware of

(8) EMBRACE DATA STEWARDSHIP



Data Stewardship Decoded

Mapping Its Diverse Manifestations and Emerging Relevance at a time of AI

Dr. Stefaan G. Verhulst,

Co-Founder The GovLab and The DataTank

Research Professor Tandon School of Engineering, New York University

Abstract

Data stewardship has become a critical component of modern data governance, especially with the growing use of artificial intelligence (AI). Despite its increasing importance, the concept of data stewardship remains ambiguous and varies in its application. This paper explores four distinct manifestations of data stewardship to clarify its emerging position in the data governance landscape. These manifestations include a) data stewardship as a set of competencies and skills, b) a function or role within organizations, c) an intermediary organization facilitating collaborations, and d) a set of guiding principles.

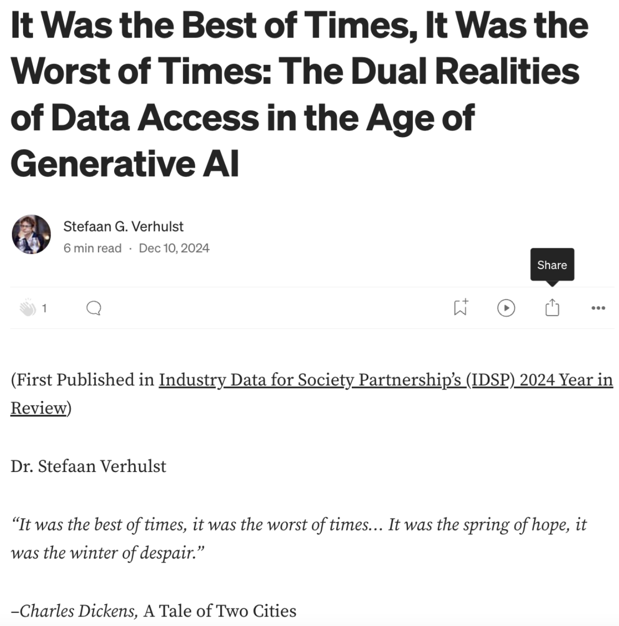
The paper subsequently outlines the core competencies required for effective data stewardship, explains the distinction between data stewards and Chief Data Officers (CDOs), and details the intermediary role of stewards in bridging gaps between data holders and external stakeholders. It also explores key principles aligned with the FAIR framework (Findable, Accessible, Interoperable, Reusable) and introduces the emerging principle of AI readiness to ensure data meets the ethical and technical requirements of AI systems.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5124555

04. Final Considerations

BEST OF TIMES, WORST OF TIMES?

> *..it was the spring of hope, it was the **winter of despair**...*
>
> *— Charles Dickens*



It Was the Best of Times, It Was the Worst of Times: The Dual Realities of Data Access in the Age of Generative AI

Stefaan G. Verhulst
6 min read · Dec 10, 2024

(First Published in Industry Data for Society Partnership's (IDSP) 2024 Year in Review)

Dr. Stefaan Verhulst

"It was the best of times, it was the worst of times… It was the spring of hope, it was the winter of despair."

–*Charles Dickens,* A Tale of Two Cities

THEGOVLAB

www.thegovlab.org