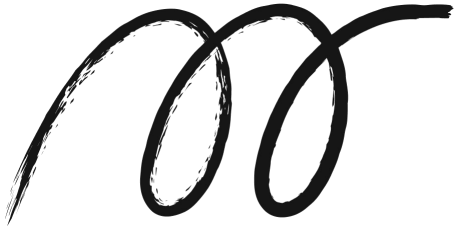
Mozilla Data Collective

Create. Curate. Control.

Your datasets, under your control



Kathy Reid | September 2025 | 424.80 PPM CO^2

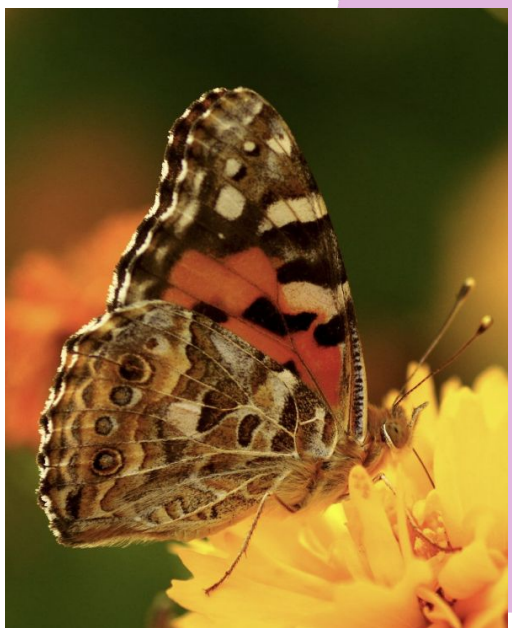
Mozilla Foundation

Mozilla Data Collective

Acknowledgement of Country

Luis Mata,
Vanessa kershawi,

Royal Botanic Gardens Melbourne,
City of Melbourne, Victoria
CC-BY-NC-SA



Via https://flic.kr/p/S3Sk7r

Kathy Reid

Speaks with machines

Engineer. Data. Machine learning.
Voice AI.
Conscientious technologist.

Knitter.
❤️s your dog.

*Contains some replacement parts
** No I haven't finished my PhD yet, don't @ me

she/her pronouns (they/them totally fine, too)

@**KathyReid** in most of the places

Australian National University

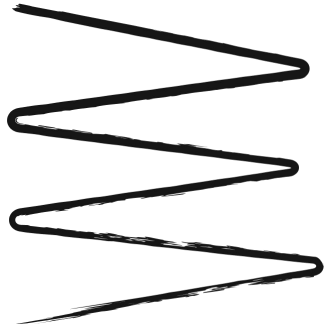
School of Cybernetics

https://cybernetics.anu.edu.au/people/kathy-reid/

Volunteers

<3

Agenda



01 **Tokenomics**:
why human-generated data is so valuable

02 **Token harvesting:**
Extractive and exclusive

03 **Is there a better way?**
YES! Mozilla Data Collective

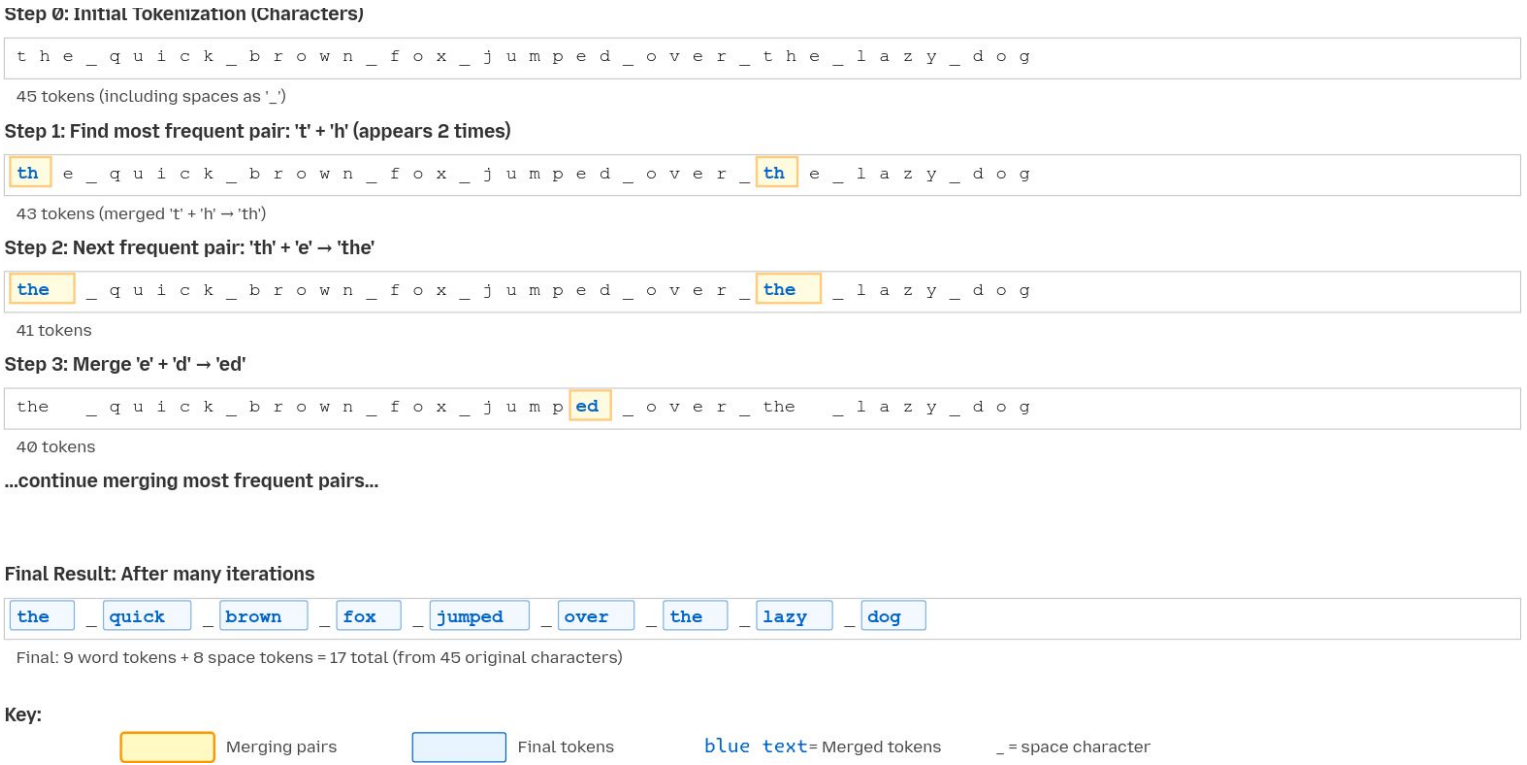
01

Tokenomics

“

Tokens are **building blocks** of data.

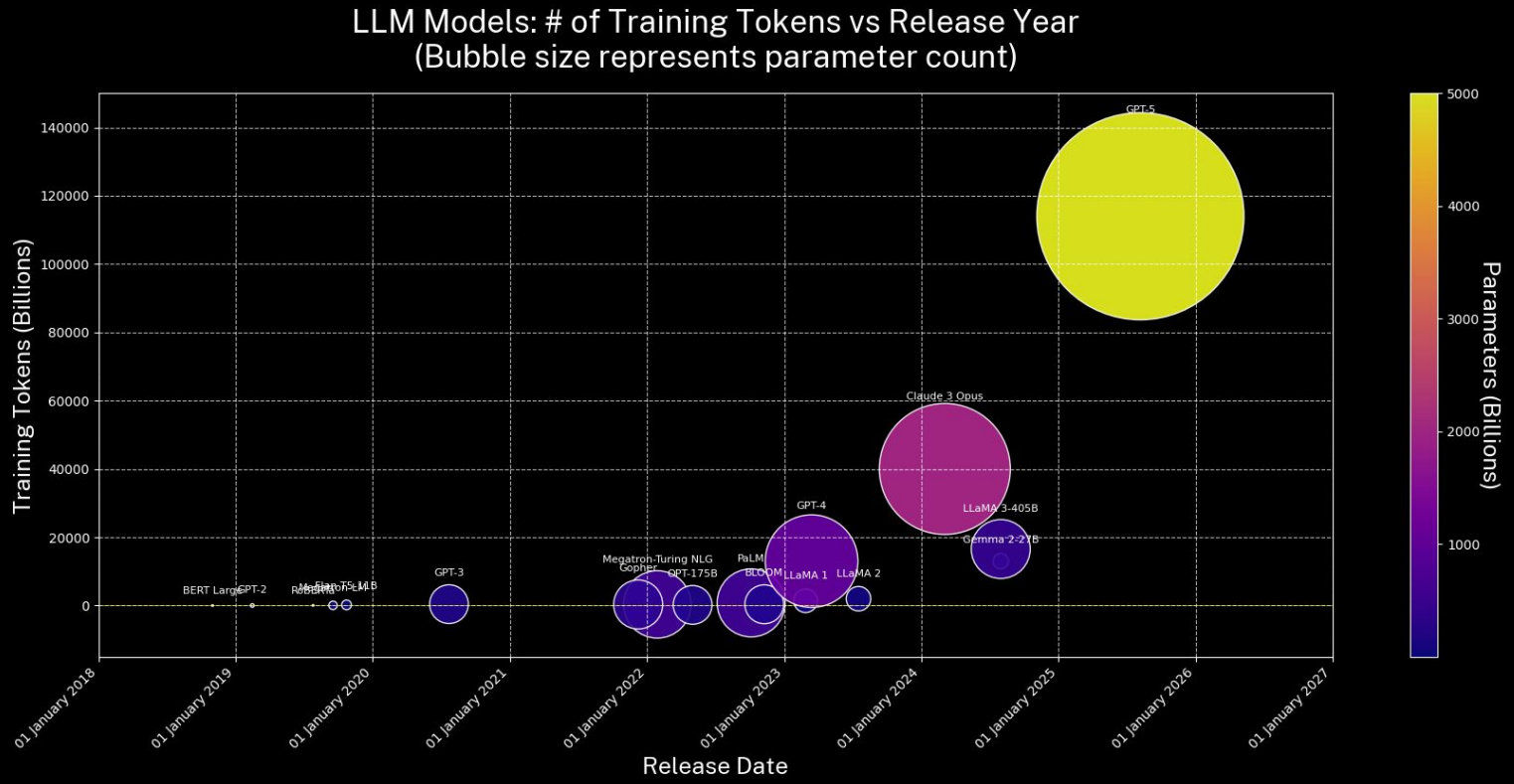
Tokenisation example: Byte Pair



Relationships between tokens

the quick brown fox jumped over the lazy dog

	the	quick	brown	fox	jumped	over	the	lazy	dog
the	0.92	0.03	0.02	0.45	0.01	0.01	0.15	0.02	0.02
quick	0.05	0.85	0.15	0.42	0.02	0.01	0.02	0.03	0.02
brown	0.04	0.12	0.88	0.48	0.02	0.01	0.02	0.03	0.02
fox	0.05	0.10	0.12	0.90	0.45	0.15	0.02	0.03	0.02
jumped	0.02	0.02	0.02	0.40	0.92	0.35	0.02	0.03	0.02
over	0.02	0.02	0.02	0.10	0.38	0.88	0.02	0.03	0.02
the	0.02	0.02	0.02	0.05	0.02	0.02	0.90	0.40	0.45
lazy	0.02	0.02	0.02	0.05	0.02	0.02	0.05	0.88	0.42
dog	0.02	0.02	0.02	0.05	0.02	0.02	0.05	0.15	0.92



https://github.com/KathyReid/token-wars-dataviz

How many tokens are there in the world?

60-160 trillion tokens

Cummins M.

How much LLM training data is there, in the limit?

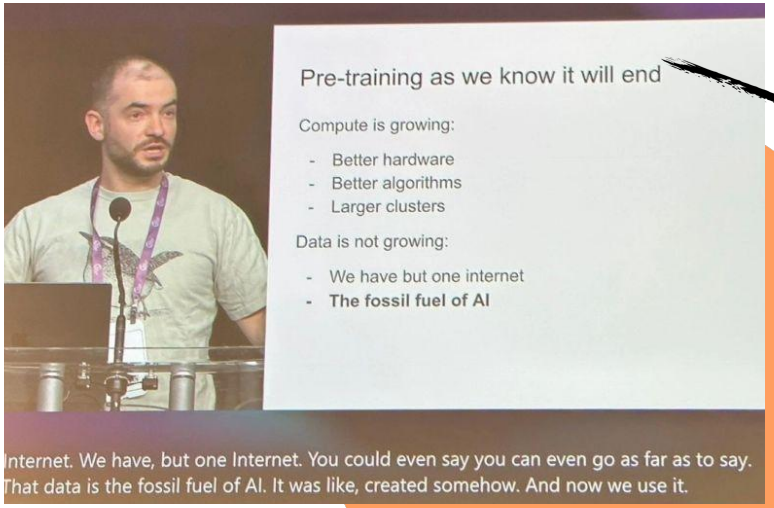
Educating Silicon. May 9, 2024.

Accessed July 14, 2025.

The Token Crisis

Attribution:
@JohnRushX via X/Twitter
taken at NeurIPS 2024,
Ilya Sutskever speaking

Video link:
https://www.youtube.com/watch?v=1yvBqasHLZs



“

Synthetic data is data that has been generated using a purpose- built mathematical model or algorithm, with the aim of solving a (set of) data science task(s).

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., ... & Weller, A. (2022). Synthetic Data--what, why and how?. arXiv preprint arXiv:2205.03257.

Model Collapse

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755-759.

“

Peak token is the point in time when the most human-generated, authentic data is available on the web.

Kathy Reid, 2025

Authentic, human-generated data is endangered

1 Scarce

models need trillions of tokens to train on, none left

2 Rare

human-created tokens are polluted with AI slop

3 Contested

because tokens are endangered, they're highly sought after

02

Token harvesting

Token harvesting is extractive

Behind 'miracle' AI is an army of 'ghost workers' — and they're speaking out about Appen

By technology reporter **Ariel Bogle**

ABC Science | AI

Fri 14 Oct 2022

Companies building artificial intelligence rely on an army of global workers. *(Getty Images: The Good Brigade)*

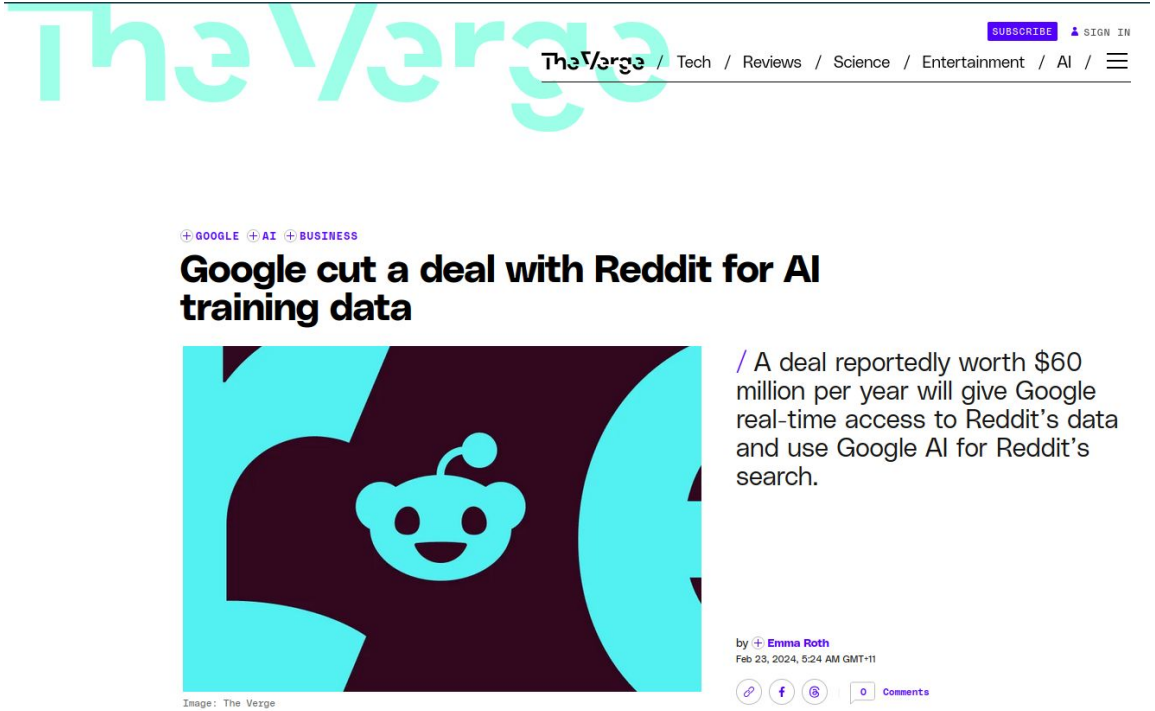
Australian authors challenge Productivity Commission's proposed copyright law exemption for AI

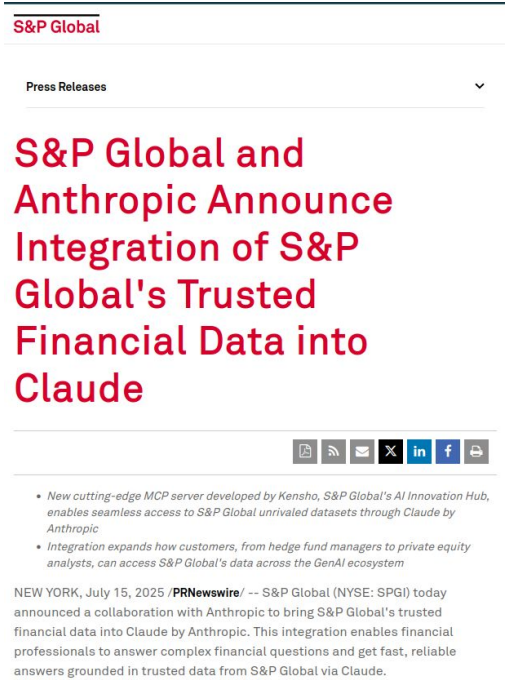
By Nicola Heath | ABC Arts | AI

Wed 13 Aug

Authors Danielle Clode, Rhett Davis and Kate Kruimink are concerned about the effects of proposed changes to copyright law. *(ABC Arts: Christian Harimanow)*

Small group of players shape AI innovation





Billions of people are not represented

WIRED SECURITY POLITICS THE BIG STORY BUSINESS SCIENCE CULTURE REVIEWS

PARESH DAVE BUSINESS MAY 31, 2023 7:00 AM

ChatGPT Is Cutting Non-English Languages Out of the AI Revolution

AI chatbots are less fluent in languages other than English, threatening to amplify existing bias in global commerce and innovation.

Search | Free Newsletters QUARTZ Editions

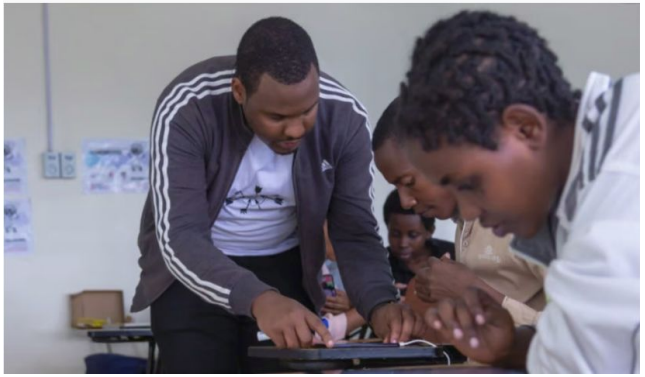
HOME LATEST BUSINESS NEWS MONEY & MARKETS TECH & INNOVATION A.I. LIFESTYLE LEADERSHIP EMAILS PODCASTS

TECH & INNOVATION

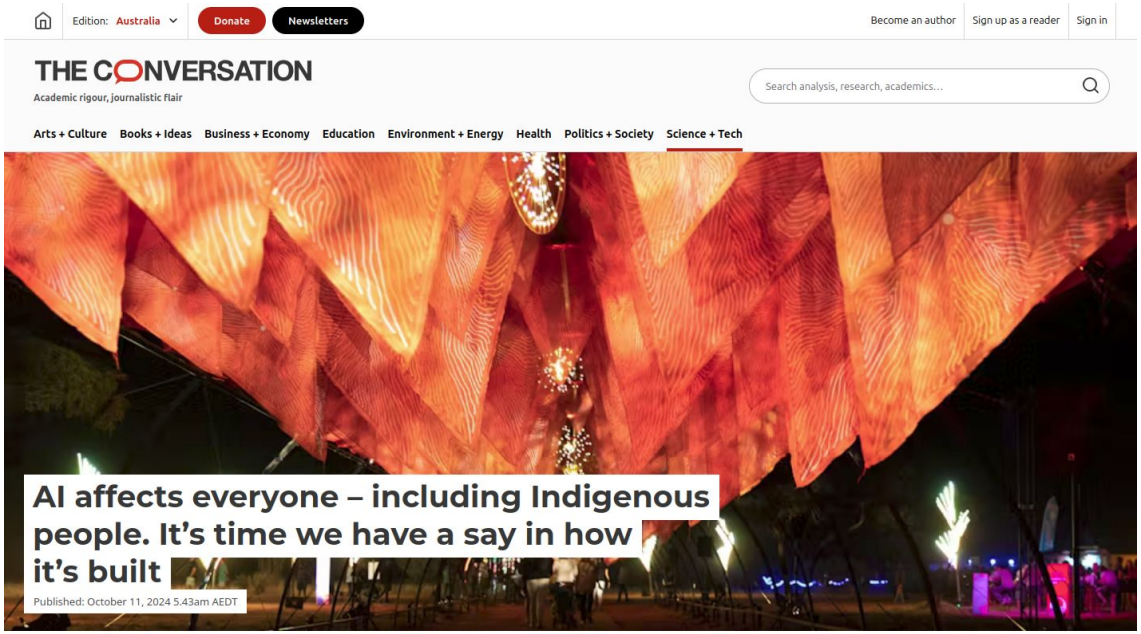
Siri and Alexa still don't support African languages

Despite speech increasingly becoming one of the main ways people interact with devices, voice technology remains largely closed off to Africa's languages, accents, and speech patterns. Case in point: The world's most popular voice assistants, Siri, Alexa and Google Assistant, still don't support any African languages. The continent has more than 1,000 languages.

By Carlos Mureithi and Carlos Mureithi Updated July 21, 2022



Current approaches bulldoze rights

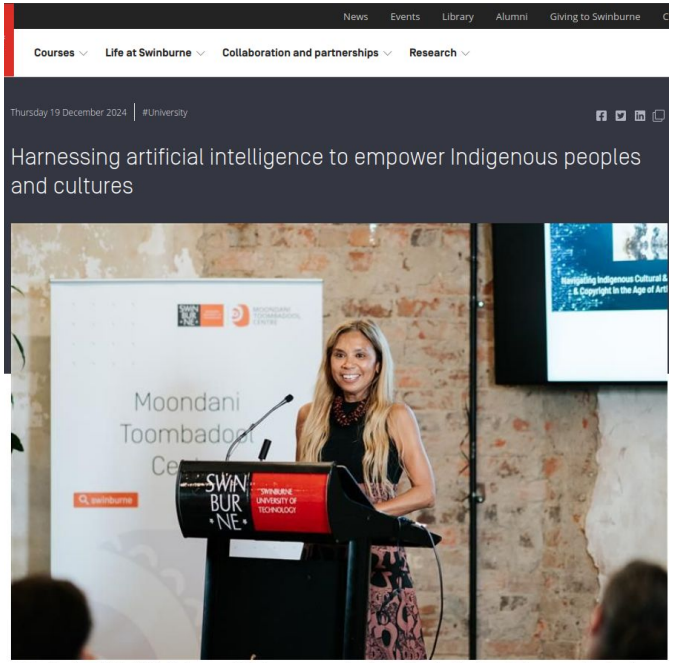


The 2022 Parrtjima Festival, held on Mparntwe (Alice Springs). Tamati Smith/Getty Images

Since artificial intelligence (AI) became mainstream over the past two years, many of the risks it poses have been widely documented. As well as fuelling deep fake porn, threatening personal privacy and accelerating the climate crisis, some people believe the emerging technology could even lead to human extinction.

Author

Tamika Worrell
Senior Lecturer in the Department of Critical Indigenous Studies, Macquarie University



Dr Terri Janke, a Wuthathi, Yadhaigana and Meriam woman, and an international authority on Indigenous cultural and intellectual property gave the 2024 Barak Wonga Oration

Home > News > 2024 > December > harnessing-artificial-intelligence-to-empower-indigenous-peoples-and-cultures

03

A better way: Mozilla Data Collective

“

In an industry that relies on extractive practices, Mozilla Data Collective is **rebuilding the AI data ecosystem - with communities at the centre.**

Data contributors can share their data on their own terms; data users can access high quality data in a transparent and ethical way.

Data contributors

Researchers and non-profits

want to increase the lifetime discoverability, value, and impact of their datasets

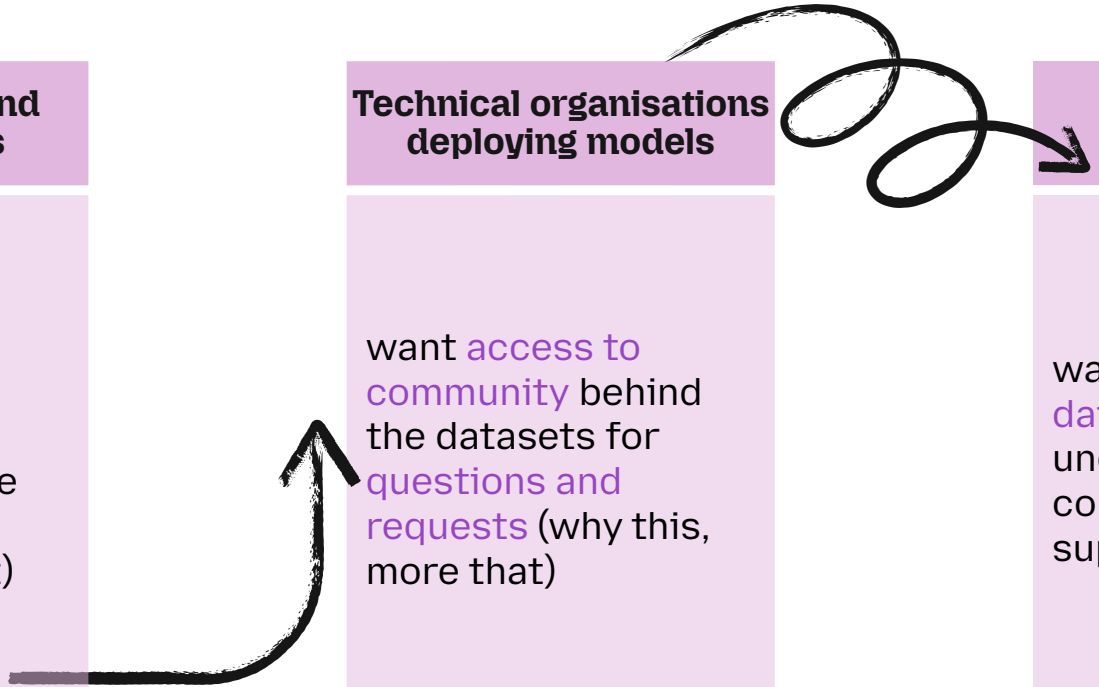
Creatives, media organisations & SMEs

want to generate revenue and benefits for the people who created it

Governments, funders & public knowledge orgs

want to open up more datasets in a transparent manner to unlock data for AI innovation that benefits all

Data consumers



Model trainers and data engineers	Technical organisations deploying models	Compliance professionals
want affordable, right-sized, high quality datasets (stop wasting time and money with trash in, trash out)	want access to community behind the datasets for questions and requests (why this, more that)	want to de-risk using data by understanding its context, intent, and supply chain

“

We are launching a **platform** that unlocks **data abundance** by giving **people** and **communities** control over their datasets.

Features for data contributors

1 Promote datasets

Store, **host** and **list** datasets with clear stewardship & robust datasheet

2 Governance dashboards

Share data in line with your values and needs - under particular licenses, or only for particular purposes

3 New value from existing data

By opting in to cross-walking, blending or augmenting their data with other datasets

Datasheets

common-voice/**cv-datasheets**

13 Contributors | 1 Issue | 4 Stars | 13 Forks

Features for data consumers

1 Discover datasets

With **transparent documentation** for licensing, intended task and source

2 Connect with community

to contextualise data and enable more collaboration and creation (data flywheels)

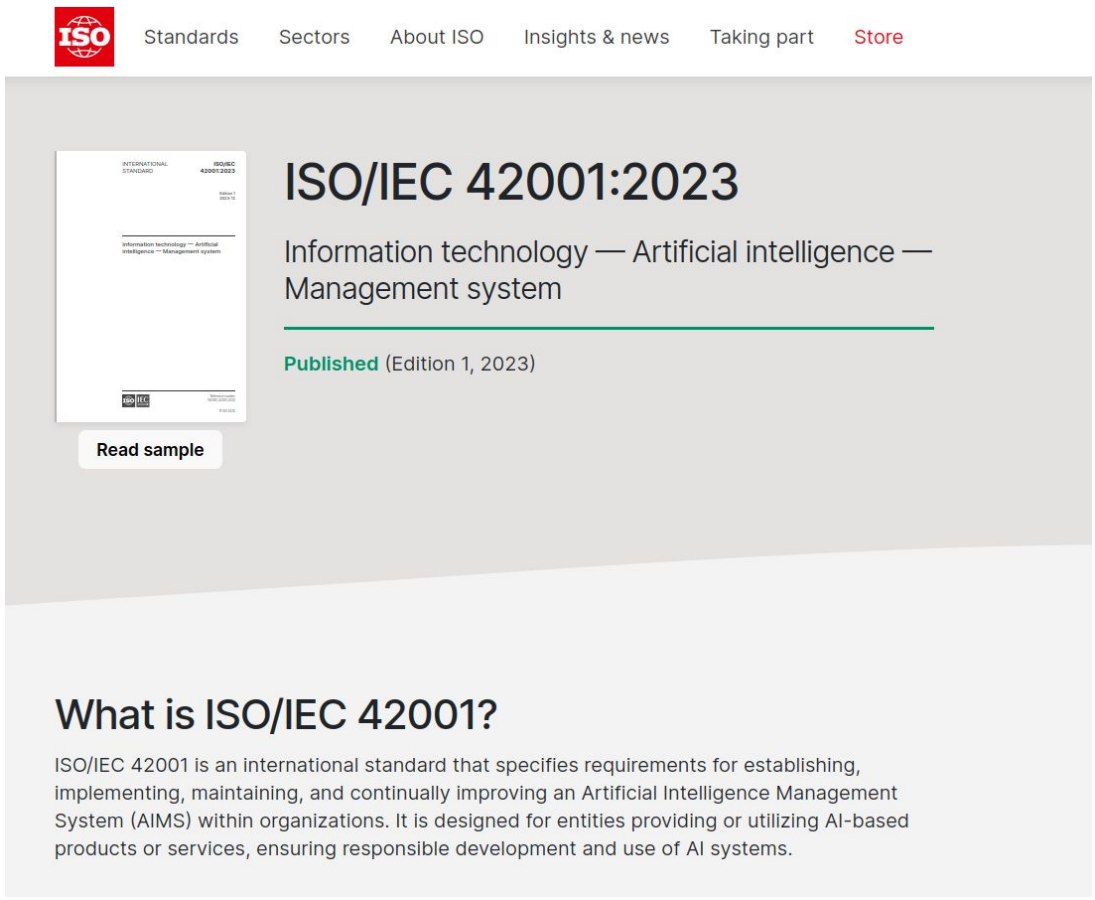
3 Supply chain transparency

enables compliance in shifting legal landscape

ISO 42001:

International standard for Artificial Intelligence Management System

(also Australian standard)



Takeaways

1. Authentic tokens are scarce, getting more so, and are highly sought after

2. Token harvesting from the web is exploitative in many ways

3. There's a better way to solve the needs of data producers & consumers

Next steps

https://discord.gg/4TjgEdq25Y

mozdatacollective.bsky.social

mozilladatacollective@
fosstodon.org

Come chat with us!
We run regular office hours

Send us feedback!

Tell us about a dataset!

mozilladatacollective@
mozillafoundation.org

https://datacollective.mozillafoundation.org/

Thank you

Thank you to **EM Lewis-Jong**, **Jessica Rose** and **Justin Grant** for feedback, iteration and assistance with this deck

04

Additional information

Further reading

- Associated Press. AI startup Anthropic agrees to pay $1.5bn to settle book piracy lawsuit. The Guardian. September 5, 2025.
- Cummins M. How much LLM training data is there, in the limit? Educating Silicon. May 9, 2024.
- Global Indigenous Data Alliance. CARE Principles for Indigenous Data Governance.
- Hao K. Artificial intelligence is creating a new colonial world order. MIT Technology Review. April 19, 2022.
- Heath N. Authors warn AI copyright exception a “free pass” for Big Tech to steal work. August 13, 2025.
- Jones PL, Mahelona K, Duncan S, Leoni G. Kaitiaki: closing the door on open Indigenous data. *International Journal on Digital Libraries*. 2025;26(1):1. doi:10.1007/s00799-025-00410-2
- Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., ... & Weller, A. (2022). Synthetic Data--what, why and how?. arXiv preprint arXiv:2205.03257.
- OpenAI. OpenAI and Reddit Partnership. May 16, 2024.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755-759.