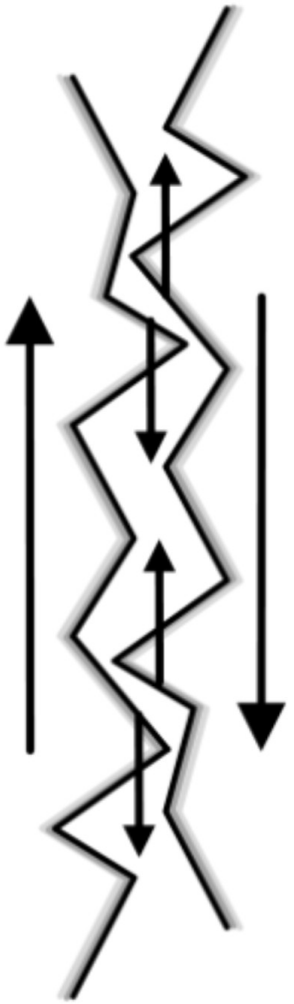
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569.

Legal frictions for the re-use of the open web for AI training

New licensing initiatives for AI training data and insights for data governance

Ramya Chandrasekhar

CommonsAI
December 10, 2025



Legal frictions for data openness

Reflections from a case-study on re-use of the open web for AI training

cnrs centre internet et societe — CUBE inno — Open Knowledge FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

https://doi.org/10.5281/zenodo.15097649

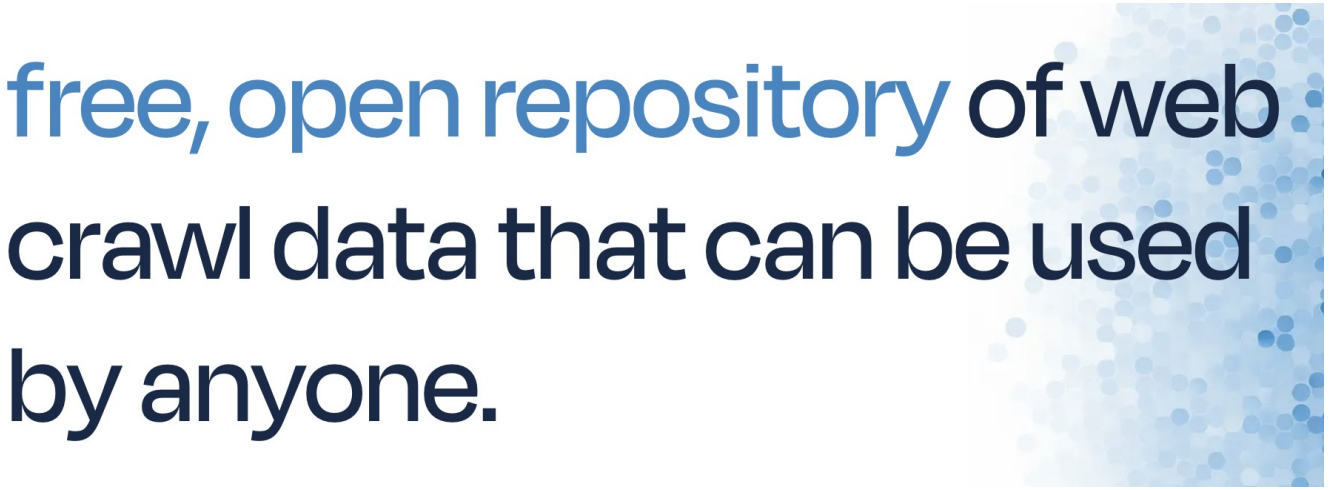
“A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks; current examples include BERT,.. GPT-3.., and CLIP...”

Bommasani et al., “On the Opportunities and Risks of Foundation Models, 2021, https://arxiv.org/abs/2108.07258

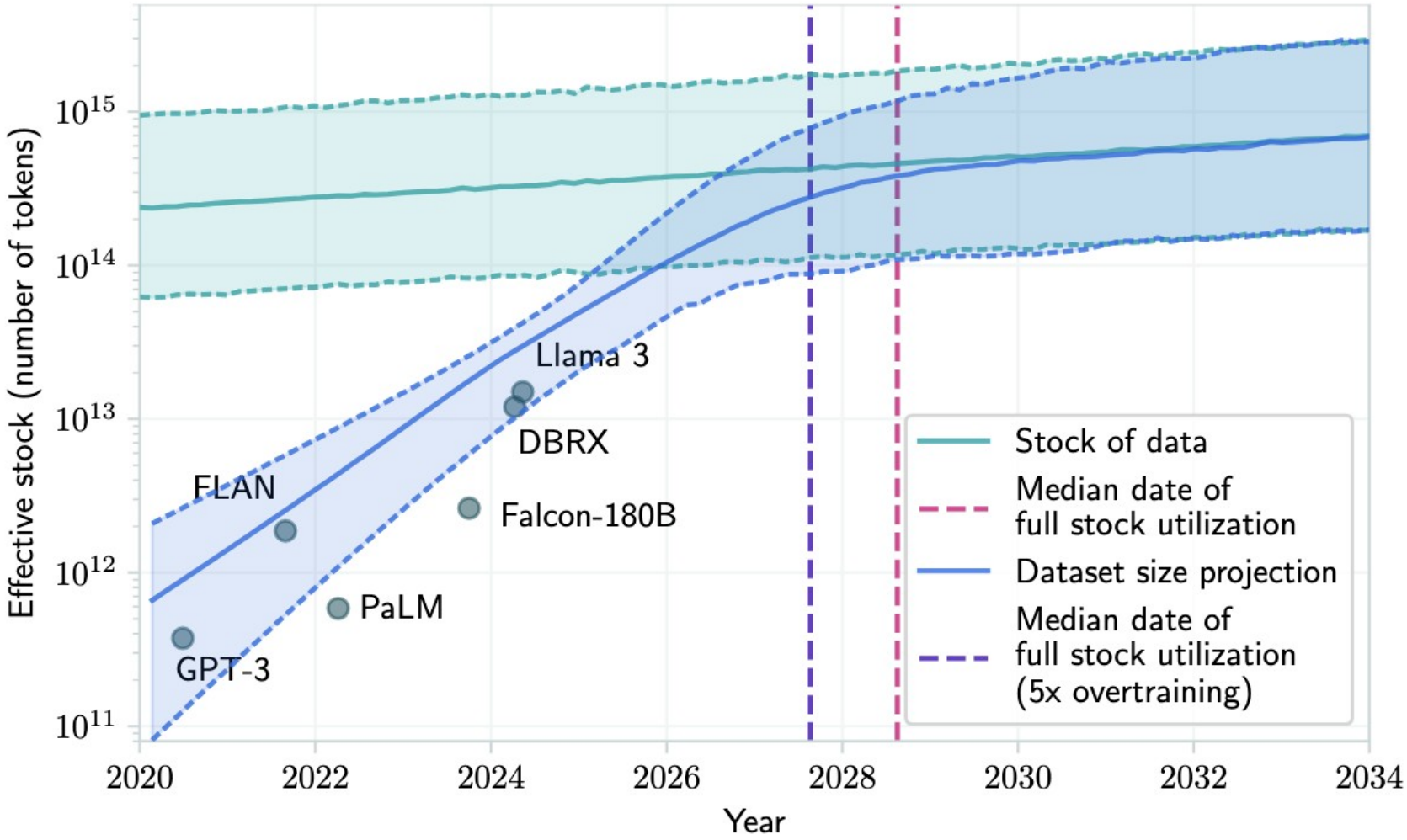
Common Pile v0.1

updated Jun 6

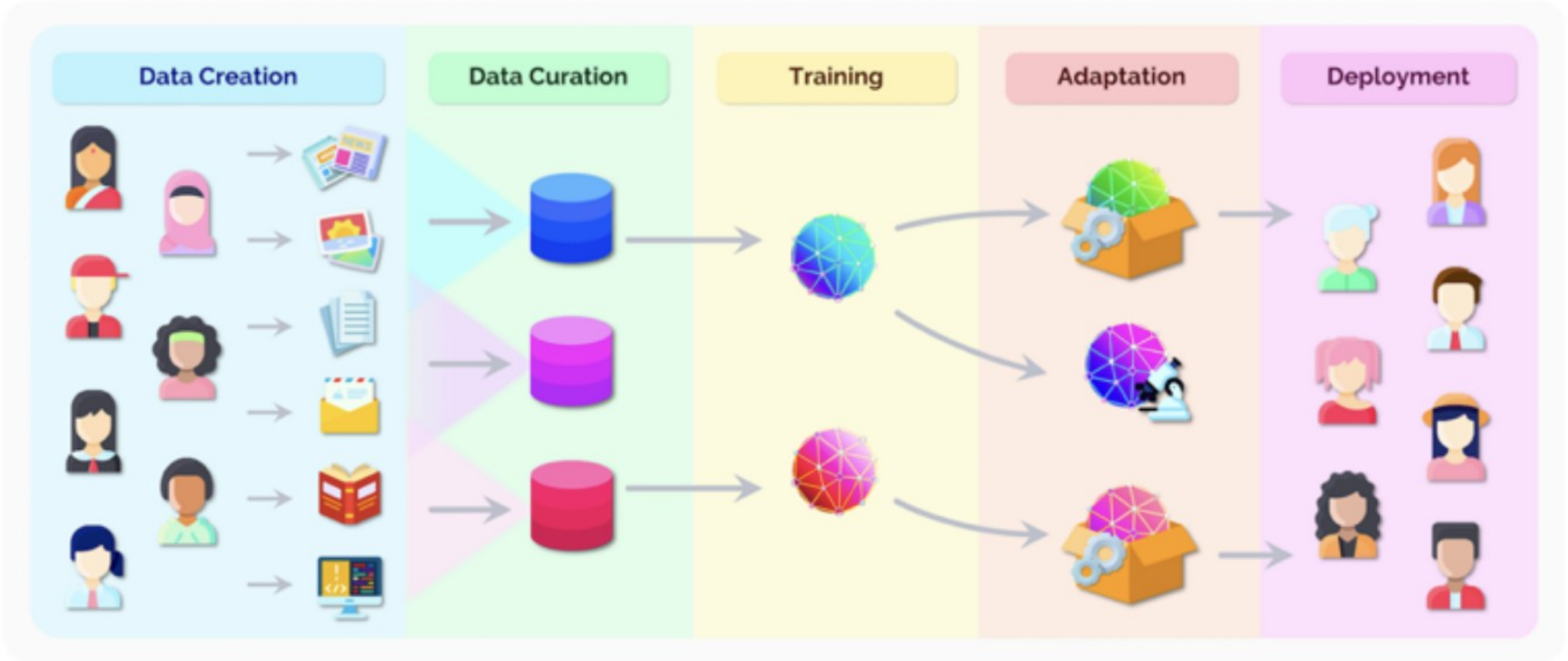
All resources related to Common Pile v0.1, an 8TB dataset of public domain and openly licensed text



GPT-3 powers the next generation of apps

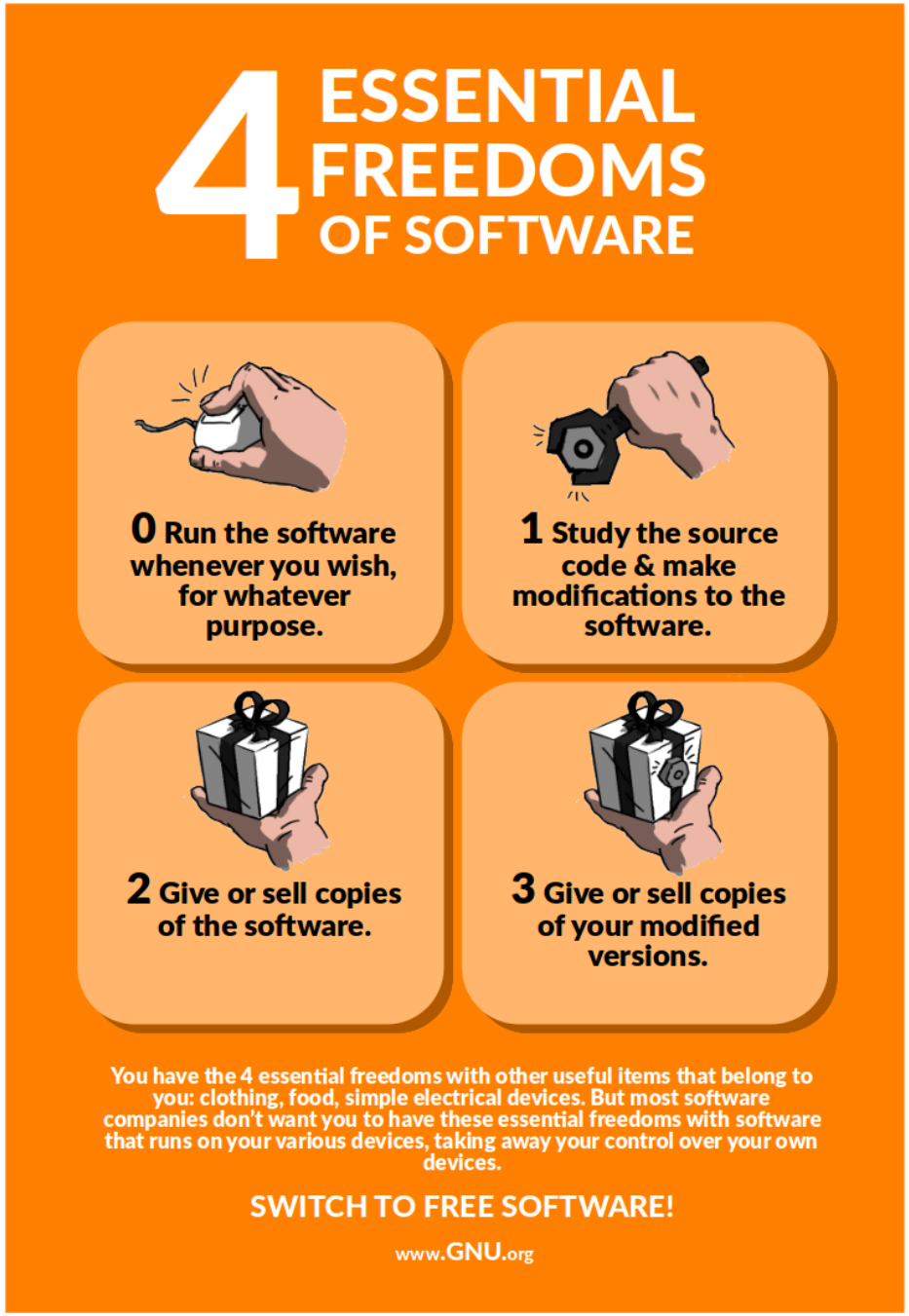


Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Marius Hobbhahn
Proceedings of the 41st International Conference on Machine Learning, PMLR 235:49523-49544, 2024.



Bommasani et al., “On the Opportunities and Risks of Foundation Models, 2021, https://arxiv.org/abs/2108.07258

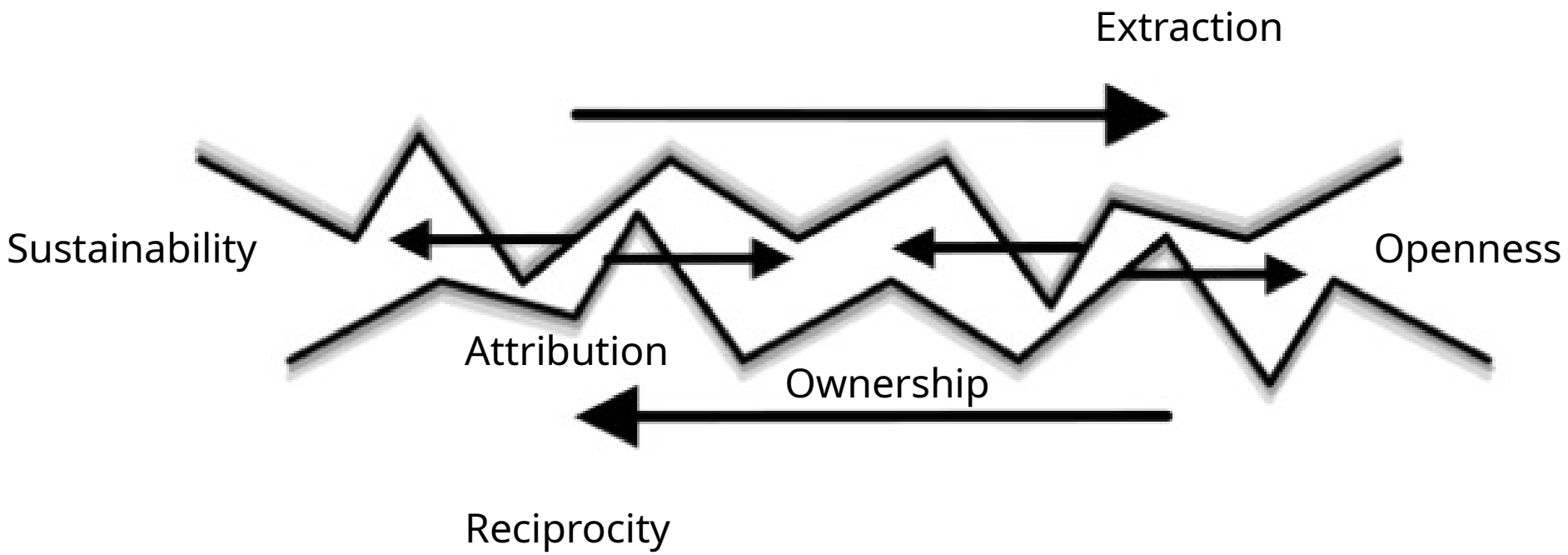
Training data and the four freedoms



Legal controversies on the use of public data for AI training

- Content from shadow libraries – Kadrey v. Meta, Bartz v. Anthropic
- Openly licensed code – Doe v. Github Inc
- Publicly accessible (copyrighted) works – Getty Images v. Stability AI (UK), Robert Kneschke v. LAION
- News articles – New York Times v. Microsoft
- Publicly available personal data – Orders against Meta Platforms Inc (Brazil DPA), Verbraucherzentrale NRW v. Meta

https://lite.framacalc.org/w3wokslmgn-adf3



Foundational licenses

Traditional permissive licenses (MIT, Apache, BSD, CC-BY, CC Zero)

- Traditional free+copyleft licenses (GPL, AGPL, CC-BY-SA, CC-BY-SA-ND)
- Community Data License Agreement (CDLA) Permissive 2.0

Responsible Use Licenses

RAIL Licenses

- OpenRAIL Licenses
- Montreal License
- AI2Impact Licenes

Emerging alternative licenses

- Copyfarleft
- Kaitiakitanga License
- Nwulite Obodo Licenseand Esethu License
- UsageRight License
- ~~Open~~ Data Commons License
- co-pilot resistant licenses
- Microsoft data use agreements for AI training

RAIL and OpenRAIL Licenses

- Behavioural use restrictions – licensee + downstream
- Applicable to data, application/software, model, source code

“The Open Source movement has been critical to the growth of transparency and reproducibility in science allowing people to license code easily and with understandable clauses. We introduce the idea of Open RAIL licenses as an attempt to provide practitioners with more control over how what they create is used, while also creating a simple and understandable mechanism for them to license material broadly and *permissively*.”

https://www.licenses.ai/blog/2022/8/18/naming-convention-of-responsible-ai-licenses

Montreal License

New standardised taxonomy for data use in the context of AI/ML

Misha Benjamin et al, “Towards Standardization of Data Licenses: The Montreal Data License”, 2019, https://arxiv.org/pdf/1903.12262

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						

~~Open~~ Data Commons License

- “Extend open licenses to all data types, including personal and technical data, while enabling granular control over data with modular clauses which allow licensors to restrict access outside of certain boundaries (e.g. authorised users and uses) – thereby ultimately fostering a commons-based approach.”
- Compulsory elements – share alike, privacy pledge, right to erasure
- Optional elements – attribution, derivatives, confidentiality, scope of use and users.

Benhamou, Y., Dulong de Rosnay, M. Open Licensing and Data Trust for Personal and Non-Personal Data: A Blueprint to Support the Commons and Privacy. *IIC* (2025). https://doi.org/10.1007/s40319-025-01636-y

AI2Impact License

- Risk-based use restrictions – high, medium, low
- Public release of Derivative Impact Reports
- Public disclosure of license violators

https://medium.com/ai2-blog/the-ai2-impact-license-a-new-way-to-think-about-ai-licensing-bc90ff26a9ee

Nwulite Obodo License

“Equitable Open Data Licensing”, with tiered share-alike

https://licensingafricandatasets.com/resources

3. Key freedoms for recipients in Africa & developing countries

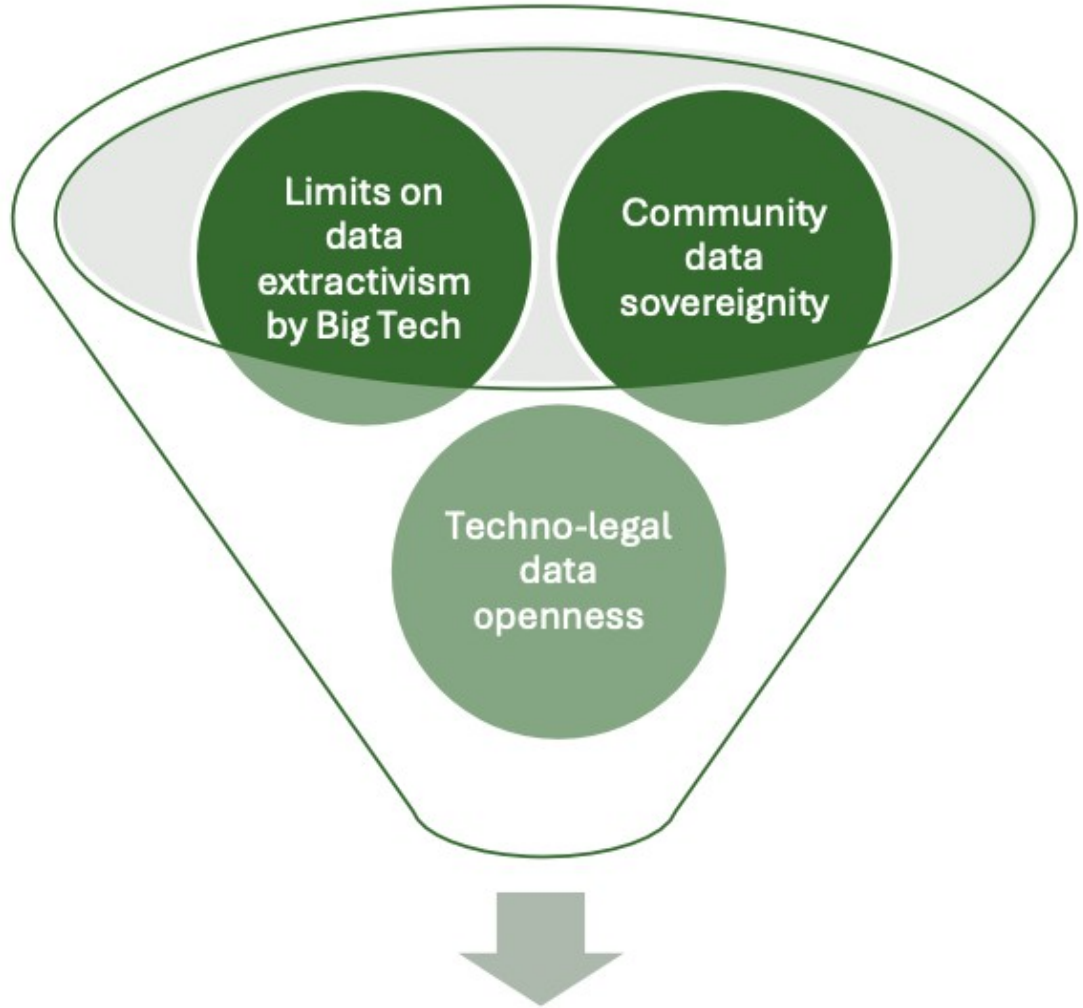
If you’re a recipient from Africa or another developing country, you are granted a **worldwide, royalty-free, non-exclusive, irrevocable right** to:

- Use the dataset.
- Adapt or remix it.
- Share it **within** developing countries—under **the same licence**.
- Engage in **commercial use**, but only **outside** Africa/developing countries.

4. Key requirements for recipients outside Africa/developing countries

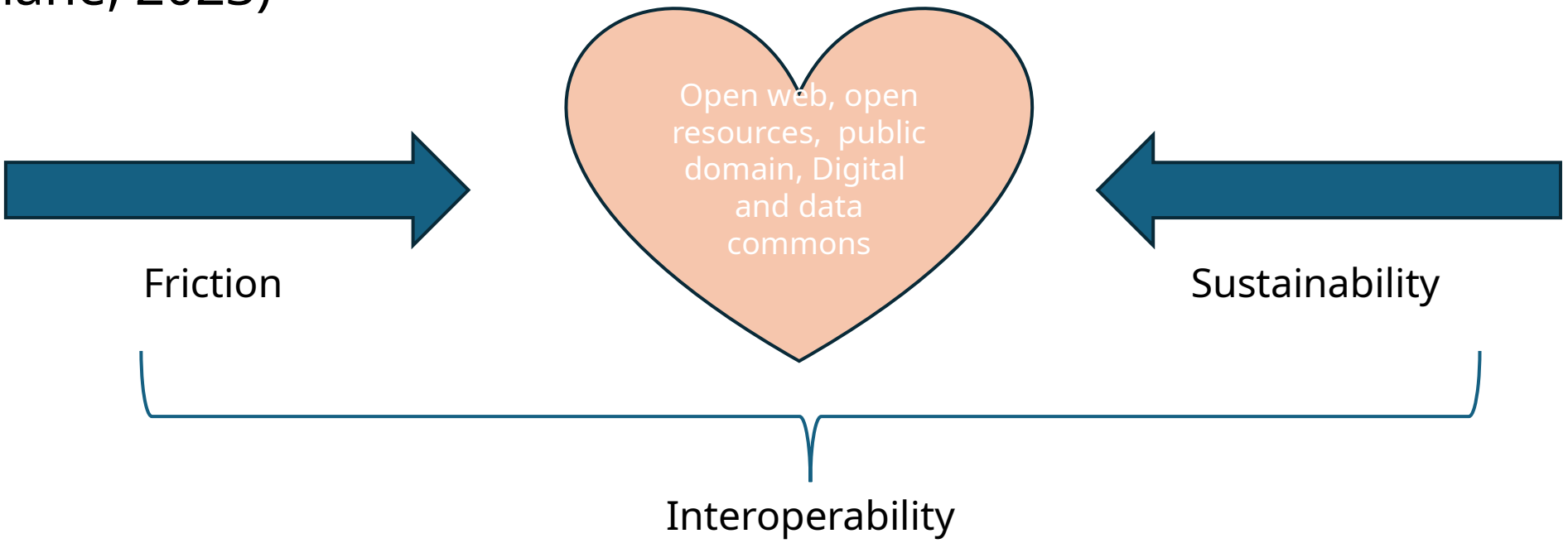
If you’re based **outside** Africa or other developing countries:

- You may use or adapt the dataset **only** under the same licence terms.
- You must provide a **benefit (that could include royalties)** to the original Dataset Provider (e.g., monetary payment, partnership, other agreed benefits).
- You may then share adapted data back to Africa or other recipients under the same licence ***provided that*** you respect the licensing terms.


Limits on data extractivism by Big Tech
Community data sovereignity
Techno-legal data openness
Data openness of training data

Provocation – Friction is necessary; friction can be productive

"…friction is the human condition, part of the human condition, we need friction in our daily interaction, to make sense of the world, it's through friction, we realise our, you know, common ground, we realise our differences, and we realise, you know, our own values." (Interview by Abeba Birhane, 2023)



Thank you!

Feel free to reach out to discuss more!

ramya.chandrasekhar@cnrs.fr