A Framework for Ethical Data Removal from Language Resources: An Example from Low-Resource Language Communities

Sarah Luger[1,3], **Rafael Mosquera-Gómez**[2,3], **Pedro Ortiz-Suárez**[4], **Thom Vaughan**[4]
[1]iMerit
[2]Factored AI
[3]MLCommons
[4]Common Crawl
sarah@mlcommons.org, rafael.mosquera@mlcommons.org

Abstract

As data resources such as Common Crawl, Mozilla Common Voice, The Pile, and LAION increasingly serve as the raw material for foundational models, the ethical implications of data collection practices become more complex. This position paper addresses some of the growing concerns regarding data removal from said resources. Further, this paper presents a framework for data resource hosts to follow when deciding when to remove data. It also presents a process for individuals and/or communities to follow when seeking to have their data removed. Finally, numerous technical challenges and societal trade-offs are addressed.

Introduction

The scale of data collection for training foundational models has grown exponentially over the past decade. Datasets like Common Crawl, Mozilla Common Voice, The Pile, and LAION have become the backbone of modern language and multi-modal models, containing billions of text samples, audio recordings, and images sourced from across the internet. While this unprecedented scale has enabled remarkable advances in Artificial Intelligence (AI), it has also introduced complex ethical challenges that the Machine Learning (ML) community still needs to address.

At the heart of these challenges lies a fundamental tension: the data that powers our most capable models often includes content from individuals and communities who never consented to its use, may be actively harmed by its inclusion, or have since requested its removal. Unlike traditional datasets curated with explicit consent protocols, web-scale resources aggregate content under broad assumptions of public availability that may not align with the ethical standards we expect from responsible AI development. Here, we explore ethical imperatives, technical barriers, and proposed solutions for removing data from widely used datasets.

Related Work

Ethical Considerations on AI Social Impact

One of the main concerns surrounding ML systems such as Large Language Models and Text-to-Image generators is their tendency to reproduce and deepen societal biases. Hovy and Spruit (2016) identify three key risks associated with those biases: exclusion, where populations are inadequately represented in a dataset, overgeneralization, where patterns learned for a specific task are applied for other tasks inappropriately, and exposure, where populations might face discrimination as topic overexposure creates new biases.

Likewise, Bender et al. (2021) argue that even though datasets have been increasing consistently in size, it is not clear whether they have become more diverse, as internet access itself (from which most of the crawled data comes from) is not evenly distributed. Similarly, Birhane et al. (2022) concludes that ML tends to favor the needs of research communities and large firms over broader social needs.

Indigenous Data Sovereignty

When engaging with data related to Indigenous communities, distinct considerations emerge, particularly concerning their rights to self-determination and sovereignty. Indigenous Data Sovereignty (IDS) emphasizes that Indigenous Peoples have the authority to govern the collection, ownership, and application of data pertaining to their communities, lands, and resources.

To address these concerns, various Indigenous communities have developed specific frameworks that articulate principles for ethical data governance. For instance, the CARE Principles—Collective Benefit, Authority to Control, Responsibility, and Ethics—were established to complement existing FAIR data principles, ensuring that data initiatives respect Indigenous rights and worldviews (Carroll et al. 2020). These principles advocate for data practices that result in collective benefits for Indigenous communities, uphold their authority over data, ensure responsible data stewardship, and adhere to ethical standards rooted in Indigenous values (Barrowcliffe et al. 2025).

Likewise, in Latin America and the Caribbean, a report from Zepeda and Pinto (2023) highlights the importance of integrating Indigenous perspectives into AI development. It underscores the need for AI systems to be designed and implemented in ways that respect Indigenous knowledge systems, cultural values, and data governance preferences.

Technical Mechanisms for Data Governance

The Robots Exclusion Protocol (REP) serves as the main mechanism through which website administrators can

state their preferences regarding permissions to automated scripts, such as web crawlers. This protocol was extended and can be found as RFC 9309 (Koster et al. 2022). Nevertheless, Chang and He (2025) mention the problems that arise when talking about enforceability, detailing the complications that could arise, such as stifled innovation and reduced competitiveness, if stricter regulations were implemented. Ongoing refinement of REP includes enhanced parameters for describing AI data (Vaughan 2025). Beyond the vocabulary extensions for expressing content preferences around AI training are increased transparency around how data is used to train models (Hazael-Massieux 2024) and considerations for implementing data rightholder opt-outs (Keller 2024).

Background

Data sovereignty in the AI era is a crucial topic for Indigenous communities (Pickens 2025). Voice data not only records words, but also documents immutable, biometric information about the speaker. Historically, there are aspects of collecting language training data that overlap with colonial mindsets and behaviors that reinforce stereotypes, control access to self-determinism, and force dependence of members of the Global South to colonial power structures (King 2020). Even language collection in the name of "language preservation" has come with archaic, geopolitical artifacts or baggage. There are legitimate linguistic projects focused on recording languages threatened with extinction, but some early linguistic-anthropological documentation is linked with furthering cultural assimilation (Baker 2021).

This work was motivated by recent conversations between low-resource machine translation experts after the presentation of novel, low-resource data sources to this community. The low-resource machine translation community balances improving communication technology with the right of some communities to opt out of digital recordings and artifacts that they feel violate their cultural autonomy (Caswell et al. 2020). New data sources are incredibly valuable for building robust language processing tools and products so newly surfaced data is shared widely (Kreutzer et al. 2022), (Kargaran et al. 2023). One such source contained significant amount of Maori language data. The Maori people have opted out of digitizing their language and actively deplatform data that violates what they consider good practice (Kukutai 2024); this approach has also been taken by other Indigenous peoples (Hutchinson et al. 2025).

It was unclear how this Maori language data could be removed by someone who knew that the data should not be recorded online, had supporting evidence that the data should not be crawled, but was not a member of the Maori people. The subsequent discussion is the result of these conversations around best practice, ethical data removal.

Key Technical Challenges

The implementation of ethical data removal frameworks faces technical obstacles that extend far beyond simple deletion operations. Understanding the constraints can help both data owners, policy makers, and other parties develop realistic expectations around data removal from web archives and training datasets.

What Can Be Accomplished

Redaction and Index Removal: the most practical (and currently the most common) approach to data removal involves preventing data from appearing in search results, or surfacing through other discovery mechanisms. In this way, dataset hosts can achieve the functional equivalent of removing the data while maintaining the integrity of a web archive.

Limitations

Physical Deletion: Given the scale of modern web archives, as well as the most common archival formats used for that purpose, removing chunks of data imply recalculating and rewriting entire archive files as offsets of all subsequent entries in the archive get invalidated. This is computationally expensive, and impractical.

Data Integrity: Given that web archives use check-sums and integrity verification systems to ensure data is not corrupted having to regenerate check-sums across massive datasets could lead to potential errors across these systems.

Operational Challenges

Versioning and Temporal Consistency: Large datasets often exist in multiple versions and get distributed across different platforms. Even after removing data from the source, that does not guarantee removal from downstream applications, cached copies or derivative datasets created before the actual removal occurred.

The Language Identification Paradox: For low-resource languages, a request for complete removal will often require accurate language identification to locate relevant content. However, the scarcity of training data to improve language identification models creates a circular problem where the communities most in need of data sovereignty tools are least served by current technical capabilities, as insufficient training data hampers the development of robust language identification systems necessary for targeted removal.

Proposed Framework

We propose a basic framework as seen in Figure 1 to help individuals (the requester) navigate removing data from a dataset. The goal of Stage 1 in the flowchart is to distinguish data ownership. There are three personas that this flowchart captures. The first is that of the data owner, the second is that of a member and/or leader of a recognized group that this data represents, and the third is a person outside of the recognized community who is aware of the data management wishes of this recognized group. In all cases, whether someone is a member of a tribe or recognized group, is a concerned member of the public, or feels that the data may be their property, the next step, seen at Stage 2, is to produce evidence to support one of the three personas.

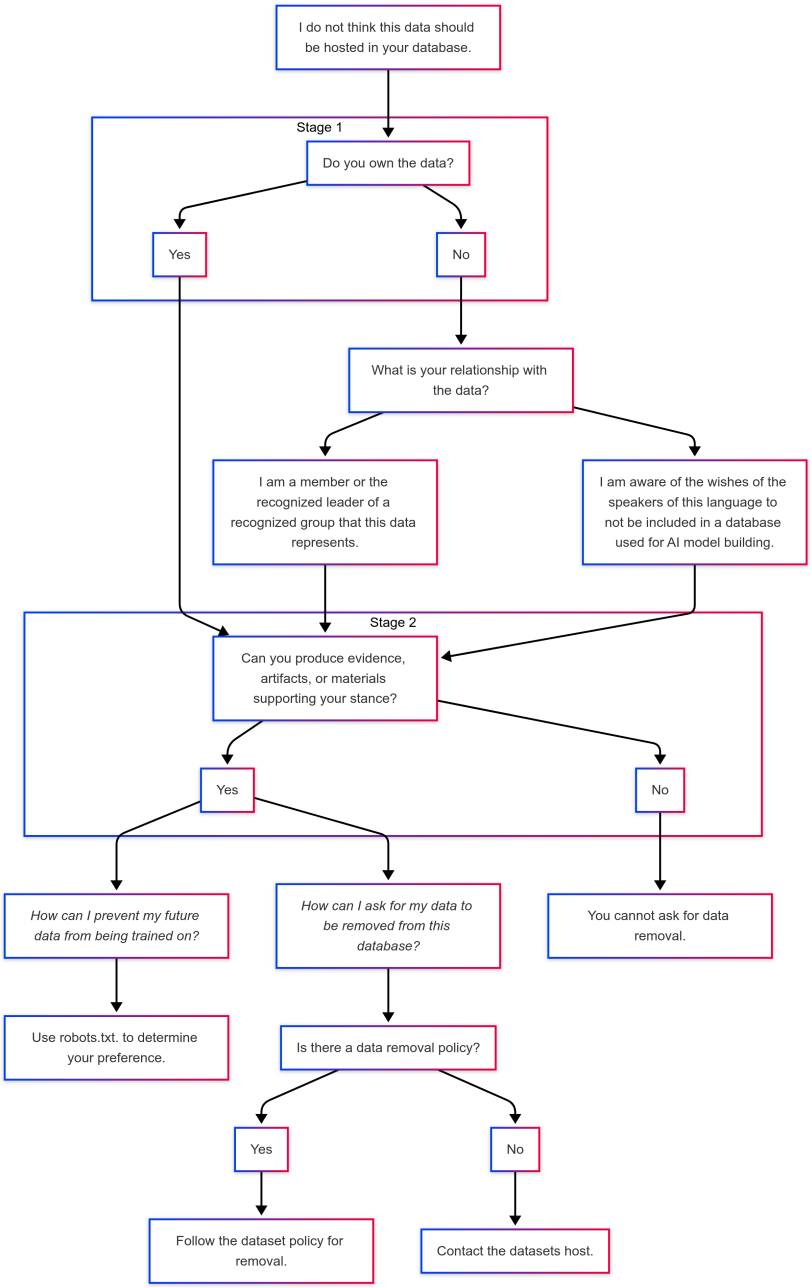

Figure 1: A flowchart showing how to navigate removing data from a data source.

At this point in Stage 2, if there is no supporting evidence or artifacts that confer data ownership or data sovereignty, data removal is not possible. This is the terminal state of a scenario where the requester has no supporting evidence. In contrast, if the requester can show evidence of data ownership or data sovereignty there are two subsequent scenarios where the requester can seek data removal. In the first case, the requester is seeking to prevent their future data from being trained on by AI systems. The Robots Exclusion Protocol, or ”robots.txt” can be added to websites and allows the site owner to manage their preferences for both if their data is crawled and what the crawled data can be used for. Adding ”robots.txt” is the terminal state of a data owner wanting to control access to their future data (as their content changes.)

In the second case, the requester is seeking to remove past data from datasets. The requester should look for data removal policies from the requisite crawlers. If there is a dataset removal policy listed, follow it. If there is no policy listed, best practice is to contact the dataset host and inquire how to remove your data. Following the posted dataset removal policy and contacting the dataset host are the terminal states of how to remove past data from a dataset.

Conclusion

The massive-scale data collection required to build robust Large Language Models often has serious ethical consequences. Unauthorized data gathered from web crawling introduces opportunities to de-platform information that does not respect Indigenous communities' right to data sovereignty or the rights of other data owners. Mainstream, Western data ownership rights are in flux. In this paper, we present a framework for identifying and de-platforming data of questionable provenance.

Acknowledgments

The authors acknowledge the support from our colleagues at Common Crawl, MLCommons, Factored AI, and iMerit without which this work would not be possible. We also appreciate the thoughtful conversations on indigenous data rights with Isaac Caswell of Google Research's Low-Resource Languages group. Finally, we thank the members of the MLCommons Datasets Working Group who helped us survey disparate perspectives on this topic and formalize best practice.

References

Baker, L. D. 2021. The Racist Anti-Racism of American Anthropology. *Transforming Anthropology*, 29(2): 127–142.

Barrowcliffe, R.; Hutchinson, B.; Abdilla, A.; Acres, L.; Beetson, B.; Bell, A.; Benton, P.; Bligh, B.; Bowen, R.; Burton, N.; et al. 2025. Envisioning Aboriginal and Torres Strait Islander AI Futures Communique: March 2025. *Journal of Global Indigeneity*, 9(1): 1–8.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. arXiv:2106.15590.

Carroll, S. R.; Garba, I.; Figueroa-Rodríguez, O. L.; Holbrook, J.; Lovett, R.; Materechera, S.; Parsons, M.; Raseroka, K.; Rodriguez-Lonebear, D.; Rowe, R.; Sara, R.; Walker, J. D.; Anderson, J.; and Hudson, M. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*.

Caswell, I.; Breiner, T.; van Esch, D.; and Bapna, A. 2020. Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 6588–6608. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Chang, C.; and He, X. 2025. The Liabilities of Robots.txt. arXiv:2503.06035.

Hazael-Massieux, D. 2024. Managing exposure of Web content to AI systems. Internet-Draft. Work in Progress.

Hovy, D.; and Spruit, S. L. 2016. The Social Impact of Natural Language Processing. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598. Berlin, Germany: Association for Computational Linguistics.

Hutchinson, B.; Louro, C. R.; Collard, G.; and Cooper, N. 2025. Designing Speech Technologies for Australian Aboriginal English: Opportunities, Risks and Participation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 108–124. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

Kargaran, A. H.; Imani, A.; Yvon, F.; and Schuetze, H. 2023. GlotLID: Language Identification for Low-Resource Languages. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6155–6218. Singapore: Association for Computational Linguistics.

Keller, P. 2024. Considerations for Implementing Rightholder Opt-Outs by AI Model Developers. Internet-Draft. Work in Progress.

King, C. 2020. *Gods of the Upper Air: How a Circle of Renegade Anthropologists Reinvented Race, Sex, and Gender in the Twentieth Century*. Knopf Doubleday Publishing Group. ISBN 9780525432326.

Koster, M.; Illyes, G.; Zeller, H.; and Sassman, L. 2022. Robots Exclusion Protocol. RFC 9309.

Kreutzer, J.; Caswell, I.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A.; Subramani, N.; Sokolov, A.; Sikasote, C.; Setyawan, M.; Sarin, S.; Samb, S.; Sagot, B.; Rivera, C.; Rios, A.; Papadimitriou, I.; Osei, S.; Suarez, P. O.; Orife, I.; Ogueji, K.; Rubungo, A. N.; Nguyen, T. Q.; Müller, M.; Müller, A.; Muhammad, S. H.; Muhammad, N.; Mnyakeni, A.; Mirzakhalov, J.; Matangira, T.; Leong, C.; Lawson, N.; Kudugunta, S.; Jernite, Y.; Jenny, M.; Firat, O.; Dossou, B. F. P.; Dlamini, S.; de Silva, N.; Çabuk Ballı, S.; Biderman, S.; Battisti, A.; Baruwa, A.; Bapna, A.; Baljekar, P.; Azime, I. A.; Awokoya, A.; Ataman, D.; Ahia, O.; Ahia, O.; Agrawal, S.; and Adeyemi, M. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10: 50–72.

Kukutai, T. 2024. Science. *Science*, 384(6691): eado9298.

Pickens, R. D. 2025. Prioritizing Data and Tribal Sovereignty in Global AI Policy. Internet-Draft. Work in Progress.

Vaughan, T. 2025. Vocabulary for Expressing Content Preferences for AI Training (aipref). Internet-Draft.

Zepeda, L. E. G.; and Pinto, C. E. M. 2023. *Inteligencia Artificial centrada en los Pueblos Indígenas: Perspectivas desde América Latina y el Caribe*. París: UNESCO.