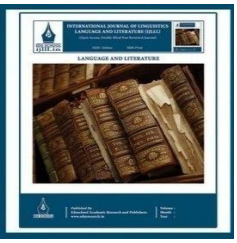
INTERNATIONAL JOURNAL OF LINGUISTICS LANGUAGE AND LITERATURE (IJLLL)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3049-3242 ISSN Print: 3139-1990



Stewardship, Not Extraction: Large Language Models and Endangered Language Revitalization

Annette Treesa Benny

Assistant Professor,Marian College, Kuttikkanam (Autonomous)Idukki District, Kerala, India

Article information

Received: 3rd July 2026
Received in revised form:7th August 2026
Accepted: 10th September 2026
Available online: 7th October 2026

Volume: 2
Issue: 4
DOI: https://doi.org/10.63090/IJLLL/3049.3242.0043

Abstract

This article examines the rapidly expanding application of large language models, automatic speech recognition, text-to-speech synthesis, and machine translation to the documentation and revitalization of endangered and Indigenous languages. It argues that the encounter between data-hungry neural architectures and the world's smallest speech communities is structurally fraught: the very statistical logic that makes contemporary language technology powerful presupposes the abundant text that endangered languages, by definition, lack, while the dominant practice of harvesting linguistic material as freely available "data" reproduces the extractive dynamics that contributed to language endangerment in the first place. Drawing on computational scholarship on low-resource languages, on the Indigenous data sovereignty literature, and on concrete cases including te reo Maori, Welsh, and Cherokee, the discussion situates technical questions of scarcity and scale within ethical questions of agency, ownership, and benefit. It contends that frameworks such as the CARE Principles, OCAP, and community-authored licensing offer a constructive path beyond a sterile opposition between technological optimism and refusal. The article concludes that artificial intelligence can meaningfully support revitalization only when subordinated to community-led documentation and, above all, to the intergenerational transmission that remains the decisive measure of a language's future.

Keywords:- Endangered Languages, Language Revitalization, Large Language Models, Low-Resource NLP, Data Sovereignty, Indigenous Languages, Machine Translation, Automatic Speech Recognition

Introduction

Linguists have long warned that the present century will witness a contraction of human linguistic diversity unmatched in recorded history. David Crystal's influential survey estimated that roughly half of the world's six to seven thousand languages stand in danger of disappearing within a hundred years, a projection echoed by the UNESCO Atlas of the World's Languages in Danger, which classifies some two and a half thousand languages across a five-step scale running from "vulnerable" to "extinct" (Crystal 19; Moseley). Against this backdrop, the maturation of artificial intelligence has been received in many quarters as a long-awaited reprieve. If neural systems can transcribe speech, translate between tongues, synthesize voices, and converse in fluent prose, the reasoning goes, then surely they can be turned to the service

of the languages most at risk. The promise is seductive precisely because the labor of documentation and teaching is so vast and the number of fluent speakers, elders especially, so finite. Technology appears to offer scale where human capacity is scarce.

Yet the encounter between large language models and endangered languages is more vexed than the optimistic narrative allows. The statistical paradigm that underwrites modern language technology is voracious in its appetite for text and audio, and an endangered language is, almost by definition, a language for which such material is scant. More troubling still, the prevailing mode of acquiring linguistic data, the wholesale harvesting of whatever can be scraped or recorded, reenacts the very logic of appropriation that drove many communities toward language loss in the first place. This article takes a critical but constructive stance toward that encounter. It surveys what the technologies can plausibly do, anatomizes the data-scarcity problem and the ethical hazards of extraction, and argues that the question is not whether artificial intelligence should be used in revitalization but under whose authority, toward whose benefit, and within what relationship of accountability. The cases of te reo Maori, Welsh, and Cherokee, considered below, suggest that the most defensible path runs through community sovereignty rather than around it.

The Statistical Paradox of Low-Resource Languages

The defining property of contemporary natural language processing is its dependence on quantity. Transformer-based systems learn the distributional regularities of a language by ingesting enormous corpora, and their fluency scales with the volume and diversity of what they are fed. This dependence produces a stark stratification of the world's languages. In their widely cited taxonomy, Joshi and colleagues sort languages into tiers according to the labeled and unlabeled resources available for them, demonstrating that a tiny handful of "winners" command the overwhelming share of digital text and tooling while the vast majority languish in the lowest tiers, effectively invisible to the field (Joshi et al. 6284). Their analysis punctures the comfortable assumption that modern models are "language agnostic": the architectures may be general, but their competence is parasitic on data that most languages simply do not possess. Kornai's earlier and bleaker estimate, that fewer than five percent of the world's languages are positioned to achieve any digital presence at all, frames the same disparity as a coming "digital language death," a die-off driven not by the silencing of speakers but by exclusion from the computational infrastructures through which contemporary life is increasingly conducted (Kornai 1).

For endangered languages the paradox is acute. The technologies most celebrated for their linguistic prowess are the ones least applicable to the languages most in need, because those languages lack the textual mass on which such prowess depends. The difficulty is not merely one of volume. Many endangered and Indigenous languages exhibit morphological and typological features poorly served by tools developed on a narrow base of European languages. Mager and colleagues, surveying the Indigenous languages of the Americas, observe that polysynthesis, rich agglutinative morphology, and extensive dialectal variation compound the scarcity of parallel and annotated data, so that even modest tasks become research problems in their own right (Mager et al. 64). It is for this reason that some scholars resist the now-ubiquitous label "low-resource" itself, since it frames these languages negatively, as deficient instances of a high-resource norm, rather than as fully realized linguistic systems whose documentation simply has not been digitized or, in many cases, ought not to be released without restriction. The term smuggles in an industrial standard of "enough data" that has nothing to do with a language's vitality among its speakers.

From Documentation to Reclamation: What the Technologies Can and Cannot Do

It is useful to disaggregate the loose phrase "AI for endangered languages" into the

distinct technical capacities it conflates, for each bears a different relationship to the work of revitalization. Automatic speech recognition (ASR) promises to ease what Bird calls the "transcription bottleneck," the chronic mismatch between the hours of recorded speech a documentation project can gather and the far smaller quantity it can afford to transcribe, since careful transcription of a single hour of audio can occupy a skilled linguist for many hours (Bird, "Sparse Transcription" 713). Text-to-speech synthesis can lend a voice to written materials and support learners who have no fluent speakers nearby. Machine translation can, in principle, render community texts into a language of wider communication and back again. And large language models hold out the further prospect of conversational tutors, predictive keyboards, and generative aids for writers working in a language that the wider digital world ignores.

Sober assessment of these capacities, however, reveals that the bottleneck is rarely where the hype locates it. Bird's notion of "sparse transcription" is instructive: rather than assuming that an endangered language must first be transcribed fully and phonetically before it can be processed, he proposes interpretive, iterative, word-level workflows that invite wider community participation and treat the spoken word, not the written transcript, as primary (Bird, "Sparse Transcription" 713). The reframing matters because it shifts the locus of value from the machine's output to the human relationships and interpretive acts that documentation entails.

The generative capacities of large language models invite a particular caution. A conversational tutor or a predictive writing aid is attractive because it appears to supply the immersive interaction that learners of an endangered language so rarely find, and because it can operate at any hour without taxing the small number of fluent speakers. But a model that has learned a near-extinct language from a few thousand sentences will, when prompted beyond that thin evidence, generate text that is grammatically plausible and culturally hollow, or simply wrong. The stochastic-parrot critique applies with redoubled force in this setting: where a high-resource language can absorb a model's occasional errors without consequence, a community reconstructing a written norm from fragmentary documentation has little capacity to detect or correct fluent fabrications, which may quietly propagate into teaching materials and learner usage (Bender et al. 617). Synthesis poses an analogous hazard. A text-to-speech voice trained on limited data may flatten the prosody, vowel length, or tonal distinctions that carry meaning, and a synthetic voice presented as authoritative can displace the embodied speech of elders rather than supplement it. The technologies are most trustworthy, in short, precisely where they are most modest, retrieving, organizing, and transcribing material that humans have authenticated, and least trustworthy where they are asked to originate language on a community's behalf. Littell and colleagues, in their careful assessment of Indigenous language technologies across the roughly sixty languages of Canada, similarly emphasize that the technologies most feasible and most useful, spell-checkers, text prediction, electronic dictionaries, and pedagogical tools, are often the unglamorous ones, and that the appropriate measure of a system is whether it serves a community's own goals rather than whether it advances the state of the art (Littell et al. 2620). Machine translation between an endangered language and a dominant one remains exceptionally hard at the scale of available data; the Cherokee-English system assembled by Zhang, Frey, and Bansal, trained on a painstakingly compiled corpus of only some fourteen thousand sentence pairs, achieved BLEU scores that, while a genuine contribution, signal serviceable assistance rather than fluent automatic translation (Zhang, Frey, and Bansal 577). The lesson is not that the technologies are useless but that their value is auxiliary: they can amplify documentation and learning, but they cannot manufacture the speakers, contexts, and intergenerational ties on which a living language depends.

Scarcity, Scale, and the Lure of the Large Model

The recent turn toward massively multilingual models has reframed the scarcity problem as one of transfer rather than abundance. The premise is that a single model trained jointly across hundreds of languages can draw on what it learns from data-rich languages to bootstrap competence in data-poor ones. The No Language Left Behind effort, which scaled neural machine translation to two hundred languages and published a substantial portion of its findings in the scientific literature, exemplifies the ambition and the genuine technical achievement of this approach, reporting marked gains over prior systems and releasing its models and evaluation benchmark openly (NLLB Team 841). Such projects rightly insist that historically de-prioritized communities deserve to be served, and their open-source posture is preferable to the proprietary enclosure of linguistic capacity.

And yet scale carries its own distortions. A model that "covers" two hundred languages does not thereby grant each an equal voice; coverage is uneven, evaluation for the smallest languages is thin, and the inclusion of a language in a benchmark can create a misleading impression of solved competence that masks brittle performance on the registers, domains, and dialects that matter to actual speakers. Bender and colleagues' critique of ever-larger language models, the "stochastic parrots" argument, bears directly here: systems trained on vast uncurated corpora tend to encode the perspectives of the dominant and well-resourced, carry significant environmental and financial costs, and produce fluent text without understanding, such that an apparent gain in coverage for a marginalized language may amount to little more than statistically plausible mimicry untethered from cultural ground truth (Bender et al. 610). For an endangered language, fluent-seeming but unreliable output is not a neutral byproduct; it can introduce errors into a fragile written norm, dilute orthographic and grammatical standards that communities have labored to establish, and arrogate to a machine the authority that properly belongs to elders and fluent speakers. The seductive proposition that scale will eventually reach every language thus obscures a deeper asymmetry: the languages are made to fit the technology, rather than the technology being shaped to fit the languages and the people who speak them.

Data Sovereignty and the Ethics of Care

If the technical literature exposes the limits of scale, the Indigenous data sovereignty literature exposes the politics of data itself. Much of the machinery of contemporary NLP rests, often tacitly, on an assumption that text and recordings are raw material freely available for collection. Bird names this assumption directly, observing that Indigenous people frequently respond with hostility to the proposition that their languages are simply "data ready for the taking," and warning that treating Indigenous knowledge as a commodity risks disenfranchising the local authorities who are its rightful custodians and reenacting the very dynamics of dispossession that endangered the languages to begin with (Bird, "Decolonising Speech and Language Technology" 3504). The point reframes the data-scarcity problem in moral rather than merely engineering terms: for many communities, the relevant material is not scarce so much as deliberately withheld, restricted to particular persons, seasons, or ceremonial contexts, and not available for indiscriminate ingestion at all.

The asymmetry of benefit sharpens the ethical stakes. When the linguistic material of a marginalized community is folded into a model whose principal beneficiaries are a technology firm, an academic career, or a benchmark leaderboard, the relationship reproduces a familiar colonial geometry in which value is extracted from the periphery and accumulates at the center, with the source community receiving, at best, a citation. Bird's analysis insists that this is not an incidental flaw to be patched with better consent forms but a structural feature of treating language as a commodity, one that can disenfranchise the very knowledge-holders, elders,

ceremonial specialists, and community linguists, whose authority sustains the language (Bird, "Decolonising Speech and Language Technology" 3504). Nor is consent a simple remedy, since the individual who reads a phrase into a microphone may not be the appropriate authority to release collectively held knowledge, and a permission granted today cannot anticipate the uses to which a general-purpose model might later put the data. These difficulties explain why the most thoughtful responses have moved from the language of individual consent to the language of collective governance and ongoing stewardship.

A constructive response has emerged in the form of explicit governance principles. The widely adopted FAIR principles, which hold that research data should be Findable, Accessible, Interoperable, and Reusable, were devised chiefly to maximize the technical reusability of data and say little about the rights and interests of the people from whom data derive (Wilkinson et al. 1). It is precisely this lacuna that the CARE Principles for Indigenous Data Governance address. Formulated by Carroll and colleagues, CARE stands for Collective Benefit, Authority to Control, Responsibility, and Ethics, and is offered not as a replacement for FAIR but as a necessary complement, insisting that any use of Indigenous data advance the self-determined wellbeing of the communities concerned and proceed under their authority (Carroll et al. 1). In the same spirit, the First Nations principles of OCAP, Ownership, Control, Access, and Possession, articulate a community's collective right to govern the creation, custody, and application of information about it, with possession functioning as the concrete mechanism by which ownership is asserted and protected (First Nations Information Governance Centre). These frameworks reorient the conversation from what a model can be trained to do toward who decides whether it may be trained at all, and on what terms the resulting benefit must flow back to the source community.

Te Reo Maori and the Practice of Guardianship

No initiative has demonstrated the practical force of data sovereignty more clearly than the work of Te Hiku Media, a charitable broadcaster of New Zealand's far north that has built automatic speech recognition for te reo Maori. Confronted with the prospect of handing hard-won language recordings to external technology firms, Te Hiku Media instead developed its own corpus through a community crowdsourcing campaign and released it under a bespoke instrument, the Kaitiakitanga License, grounded in the Maori concept of guardianship. The license rests on a deliberate refusal of the ownership model: data, it holds, is not owned but cared for under the principle of kaitiakitanga, with the organization acting as guardian rather than proprietor, and any benefit derived from the data flowing back to the people from whom it descends (Te Hiku Media). When approached by outside parties seeking to license or purchase the resulting models and data, the organization declined arrangements that would have alienated the community's control, on the ground that decisions about te reo must remain with its source community.

The Te Hiku case is instructive precisely because it does not reject the technology. It embraces ASR enthusiastically while subordinating the technical project to a relationship of accountability, demonstrating that sovereignty and innovation are not opposed. The model that emerges is one in which the community sets the purpose, retains the data, governs downstream use, and harnesses the speech recognizer to a thoroughly local end, the transcription of archival broadcasts and the support of learners, rather than to the abstract goal of expanding a global model's language inventory. It is the inversion of extraction: rather than a language being mined to enrich a distant system, a system is built and held within the community to enrich the language. This is data sovereignty not as theoretical principle but as engineering practice, and it offers a replicable template precisely because it began from the question of authority rather than appending ethics as an afterthought.

Scale Versus Specificity: Welsh, Cherokee, and the Limits of Generality

Beyond the Pacific, comparable efforts illuminate the persistent tension between technological scale and linguistic-cultural specificity. The Welsh experience demonstrates what sustained, well-resourced, community-aligned investment can accomplish for a language that, though far more vital than many, remains minoritized within an Anglophone state. Through the crowdsourced Paldaruo Speech Corpus, gathered via a purpose-built mobile application that allowed speakers across Wales to contribute recordings, researchers assembled the speech data necessary to develop Welsh speech technologies and, in turn, applications such as a Welsh-language voice assistant (Cooper, Jones, and Prys 4). The Welsh case is notable for its institutional scaffolding: governmental commitment, a dedicated technology unit, and a national strategy that treats language technology as public infrastructure rather than as a research curiosity. Notably, the corpus was designed with metadata, dialect coverage, and licensing in view from the outset, reflecting an understanding that data gathered carelessly serves no one well.

Cherokee occupies the opposite end of the vitality spectrum, and its case reveals how thin the margin can be. As Zhang, Frey, and Bansal note, fluent first-language speakers of Cherokee number only around two thousand and are declining, which renders every act of documentation an emergency and every corpus an irreplaceable artifact (Zhang, Frey, and Bansal 577). Their machine-translation work was undertaken explicitly in the service of revitalization and in collaboration with the community, yet the modest results testify to a hard truth: at such scales, statistical methods can assist but cannot substitute for the human labor of teaching and speaking. The contrast between Welsh and Cherokee maps onto the sociolinguistic stages that revitalization theory has long emphasized. Fishman's framework of reversing language shift locates the decisive variable not in technology but in intergenerational transmission, the passing of a language from parents to children in the home, without which institutional and technological supports float free of any self-sustaining base (Fishman). The classic practitioners' literature on revitalization concurs, foregrounding immersion schooling, master-apprentice mentorship between elders and learners, and the deliberate creation of domains in which the language is spoken (Hinton and Hale). Technology can enrich each of these, supplying dictionaries, transcribed archives, and learning aids, but it operates downstream of the human relationships that alone produce new speakers.

The matter of fidelity deserves particular emphasis, for it is where the abstract tension between scale and specificity becomes concrete. Endangered and Indigenous languages frequently present features that general systems handle poorly: unsettled or competing orthographies, substantial dialectal variation across small populations, and morphological complexity that strains tools calibrated on analytic European languages, as Mager and colleagues document for the Americas (Mager et al. 64). A model optimized for average performance across many languages has every incentive to smooth such variation away, collapsing dialects into a single standardized register and normalizing orthographic choices that a community may regard as identity-bearing. Yet for a language with few speakers, dialectal and orthographic particularity is not noise to be filtered but precisely the substance worth preserving. The Welsh corpus designers' attention to dialect coverage and to carefully structured metadata reflects an awareness that data gathered without such care can entrench a single variety at the expense of others, a soft homogenization that a revitalization effort can ill afford (Cooper, Jones, and Prys 4). Specificity, in other words, is not a luxury layered atop a working system; it is the very thing the community is trying to keep alive, and a technology that erodes it has defeated its own purpose even as its metrics improve.

This is the crux of the tension between scale and specificity. A general-purpose model optimizes for breadth and average performance; a revitalization program needs depth, fidelity,

and cultural appropriateness for a single community. The features that make a model commercially impressive, vast coverage, fluent generation, indifference to provenance, are frequently orthogonal or even hostile to what a small community requires. Pine and Turin, surveying the field of language revitalization, observe that the global concentration of communication into a handful of technologized world languages is itself a principal engine of language loss, which should give pause to any assumption that more powerful general technology is automatically beneficial (Pine and Turin). The risk is that endangered languages are enrolled as tokens in a project whose center of gravity lies elsewhere, gaining nominal representation in a benchmark while losing the specificity, and the control, that revitalization demands.

Toward a Constructive Settlement

A purely cautionary conclusion would be as inadequate as an uncritically optimistic one. The hazards canvassed here, extraction, the dilution of fragile norms, the illusion of competence, the displacement of community authority, are real, but so are the genuine contributions that well-governed technology can make to documentation and learning. The constructive settlement lies in inverting the default order of operations. In the dominant paradigm, data is gathered first and questions of ethics, benefit, and consent are entertained afterward, if at all; the alternative, exemplified by Te Hiku Media and increasingly codified in the CARE and OCAP frameworks, begins from the community's authority and purpose and treats the technical work as instrumental to ends the community has set (Te Hiku Media; Carroll et al. 1; First Nations Information Governance Centre). Under this inversion, the right questions are not "how much data can we obtain" and "how large a model can we train," but "what does this community want for its language," "who governs the resulting resources," and "how does the benefit return to the source."

Several practical commitments follow. Documentation methods that broaden participation and respect the primacy of the spoken word, such as Bird's sparse transcription, align technical workflows with community involvement rather than supplanting it (Bird, "Sparse Transcription" 713). Tooling priorities should follow the unglamorous but genuinely useful needs that Littell and colleagues identify, rather than the prestige tasks that dominate the research literature (Littell et al. 2620). Licensing and governance should be built in from the beginning, on the model of the Paldaruo corpus and the Kaitiakitanga License, rather than retrofitted (Cooper, Jones, and Prys 4; Te Hiku Media). And the field would do well to abandon the framing of endangered languages as deficient "low-resource" cases, a framing that, as Joshi and colleagues' taxonomy inadvertently makes vivid, measures languages against an industrial standard of data abundance that has nothing to do with their worth or their futures (Joshi et al. 6284). Researchers, for their part, must accept that the appropriate role of an outside technologist is often a supporting one, lending technical capacity to a community-led effort rather than authoring it, and measuring success by the strength of the relationship and the usefulness of the result to the community rather than by publication or benchmark performance. Such a reorientation asks the field to relinquish some of its accustomed authority, but it is the price of doing genuine rather than merely nominal good.

Conclusion

The arrival of large language models, speech recognizers, synthesizers, and translation systems has genuinely altered the landscape of language documentation, and it would be a failure of imagination to refuse these tools outright. They can compress the transcription of archives, support isolated learners, and lend new reach to communities determined to reclaim their tongues. But the central argument of this article is that the value of the technology is conditional and instrumental, never primary. The statistical paradigm that makes these systems

powerful is structurally mismatched to the smallest languages, and its default mode of acquiring data reproduces the extractive relations that helped endanger those languages in the first place. The decisive questions are therefore ethical and political rather than computational: questions of authority, benefit, consent, and control, for which the Indigenous data sovereignty literature and community-authored instruments such as the Kaitiakitanga License supply a workable answer. Above all, the measure of a language's future remains what it has always been, the willingness and ability of one generation to speak it to the next. Artificial intelligence can serve that transmission, enriching the contexts in which a language lives, but it cannot perform it. A language is not, finally, a dataset, and the most sophisticated model is no substitute for a grandmother speaking to a grandchild in the words her own grandmother gave her. The promise of these technologies for endangered languages will be realized only insofar as they are placed, deliberately and from the outset, in the service of that irreplaceable human relationship.

Works Cited

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2021, pp. 610–23.

Bird, Steven. "Decolonising Speech and Language Technology." *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, 2020, pp. 3504–19.

---. "Sparse Transcription." *Computational Linguistics*, vol. 46, no. 4, 2020, pp. 713–44.

Carroll, Stephanie Russo, et al. "The CARE Principles for Indigenous Data Governance." *Data Science Journal*, vol. 19, no. 1, 2020, pp. 1–12.

Cooper, Sarah, Dewi Bryn Jones, and Delyth Prys. "Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology." *Information*, vol. 10, no. 8, 2019, article 247.

Crystal, David. *Language Death*. Cambridge UP, 2000.

First Nations Information Governance Centre. "The First Nations Principles of OCAP." *FNIGC*, fnigc.ca/ocap-training/. Accessed 28 June 2026.

Fishman, Joshua A. *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Multilingual Matters, 1991.

Hinton, Leanne, and Ken Hale, editors. *The Green Book of Language Revitalization in Practice*. Academic Press, 2001.

Joshi, Pratik, et al. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 6282–93.

Kornai, András. "Digital Language Death." *PLoS ONE*, vol. 8, no. 10, 2013, article e77056.

Littell, Patrick, et al. "Indigenous Language Technologies in Canada: Assessment, Challenges, and Successes." *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, 2018, pp. 2620–32.

Mager, Manuel, et al. "Challenges of Language Technologies for the Indigenous Languages of the Americas." *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, 2018, pp. 55–69.

Moseley, Christopher, editor. *Atlas of the World's Languages in Danger*. 3rd ed., UNESCO Publishing, 2010.

NLLB Team. "Scaling Neural Machine Translation to 200 Languages." *Nature*, vol. 630, 2024, pp. 841–46.

Pine, Aidan, and Mark Turin. "Language Revitalization." *Oxford Research Encyclopedia of Linguistics*, Oxford UP, 2017.

Te Hiku Media. "Data Sovereignty and the Kaitiakitanga License." *Te Hiku Media*, 6 May 2022, tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license. Accessed 28 June 2026.

Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data*, vol. 3, 2016, article 160018.

Zhang, Shiyue, Benjamin Frey, and Mohit Bansal. “ChrEn: Cherokee-English Machine Translation for Endangered Language Revitalization.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2020, pp. 577–95.