Revista Jurídica unicuritiba

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

DIVERSIDADE LINGÚISTICA, IA E GOVERNANÇA DA INFOSFERA: POR QUE PORTAIS DE DADOS LINGUÍSTICOS DEVEM SER CONCEBIDOS COMO DATA TRUSTS E DATA COMMONS

LINGUISTIC DIVERSITY, AI, AND THE GOVERNANCE OF THE INFOSPHERE: WHY LINGUISTIC DATA PORTALS MUST BE CONCEIVED AS DATA TRUSTS AND DATA COMMONS[1]

PAOLA CANTARINI[2]

RESUMO: A rápida difusão de grandes modelos de linguagem (LLMs) transformou a linguagem em infraestrutura estratégica para produção de conhecimento, administração pública e participação econômica. Embora exista amplo consenso de que a desigualdade linguística em sistemas de IA constitui problema estrutural, há muito menos consenso sobre como enfrentá-la institucionalmente. Este artigo argumenta que iniciativas de diversidade linguística em IA devem ir além da expansão técnica e engajar-se com governança de dados e desenho institucional. Baseando-se em trabalhos contemporâneos sobre ética da informação (Floridi), governança de comuns (Ostrom) e administração fiduciária de dados (Delacroix & Lawrence), o artigo defende que portais de dados linguísticos devem ser concebidos como data trusts e/ou data commons—não como repositórios neutros. A análise demonstra que a marginalização de línguas não-hegemônicas representa não apenas falha técnica, mas sintoma de assimetrias estruturais na governança tecnológica global que ameaça justiça epistêmica e soberania informacional.

Pesquisa: Este estudo investiga a governança de dados linguísticos no desenvolvimento de sistemas de IA focalizando a marginalização estrutural de línguas não-hegemônicas em grandes modelos de linguagem. A questão central é: como

[1] As estruturas analíticas e propostas de política central a este artigo—particularmente o Índice de Impacto de Diversidade Linguística (LDII) e modelos de data trust para soberania linguística—foram inicialmente formuladas pela autora em resposta a solicitação técnico-diplomática do Ministério das Relações Exteriores do Brasil (Itamaraty). Este artigo representa elaboração acadêmica independente da autora desses conceitos. Embora informados por esse diálogo estratégico, os argumentos e conclusões apresentados são de exclusiva responsabilidade da autora e não representam necessariamente a posição oficial do Governo do Brasil ou do Itamaraty.

[2] Advogada, Professora universitária, PhD em Direito, em Filosofia, (PUC-SP) e em Filosofia do Direito (Unisalento); Pós-Doutorado em Direito, Filosofia e Sociologia (FDUSP, PUCSP-TIDD, EGS, Universidade de Coimbra/CES, IEA/USP). Pesquisadora do IEA/projeto UAI, e em pós-doutorado na USP/RP em IA. Presidente e Pesquisadora no EthikAI – ethics as a service . Membro da Comissão da Criança e do Adolescente e da Comissão de Proteção de Dados da OABSP e de IA da OAB/Santo Amaro.

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

estruturas institucionais de governança podem transformar portais de dados linguísticos de meros repositórios técnicos em instrumentos de justiça epistêmica, soberania cultural e sustentabilidade digital? A pesquisa parte de constatações empíricas: línguas indígenas e minoritárias representam menos de 1% dos dados de treinamento para LLMs; línguas africanas (30% da diversidade linguística global) correspondem a menos de 1% da produção de pesquisa em NLP; dados da América do Sul representam menos de 0,2% de datasets multimodais. Essa sub-representação não constitui problema técnico de escala, mas manifesta assimetrias estruturais com consequências para justiça epistêmica, soberania informacional e sustentabilidade cultural.

Metodologia: A pesquisa emprega análise conceitual interdisciplinar combinada com estudo comparativo de casos e desenvolvimento de framework normativo-institucional, integrando métodos de: filosofia da tecnologia (Floridi), economia institucional (Ostrom), teoria política (Pettit, Mouffe), estudos críticos de raça e tecnologia (Benjamin, Noble, Rosa & Flores), e direito/governança de dados. O framework emerge de síntese iterativa entre literatura teórica, análise de casos internacionais (CLARIN ERIC, Mozilla Common Voice, FirstVoices, DECODE, MIDATA) e princípios normativos. O processo seguiu lógica abdutiva (Peirce): identificação de padrões empíricos, formulação de hipóteses explicativas, refinamento através de confronto com teoria e casos adicionais. Limitações incluem viés de seleção (concentração no Global North), dados escassos sobre línguas indígenas brasileiras, e necessidade de validação da aplicabilidade universal do framework.

Resultados: A pesquisa produziu framework conceitual tripartite articulando dimensões: **(A) Normativa** – diversidade linguística como biodiversidade infosférica (Floridi), entropia informacional qualitativa, justiça epistêmica (Fricker) como critério avaliativo; **(B) Institucional** – princípios de Ostrom aplicados a recursos linguísticos, data trusts como mecanismo fiduciário, modelos híbridos commons-trust; **(C) Político-Jurídica** – princípios CARE, não-dominação (Pettit) como soberania informacional, interoperabilidade legal. Análise de casos internacionais (CLARIN ERIC, Mozilla Common Voice, FirstVoices, DECODE, MIDATA) gerou oito princípios de design: administração contextual, arquitetura federada, acesso graduado, participação significativa, transparência/accountability, fluxos de benefício, sustentabilidade longo prazo, integração técnico-legal. Inovações conceituais incluem: **Interoperabilidade Cultural** (métrica avaliando fidelidade ontológica, sensibilidade pragmática, respeito ético, accountability histórica); **Licenciamento Recíproco** (sustentabilidade via taxação comercial redistribuída); **Linguicídio Digital** ("pegada de carbono digital"); **LDII** (instrumento avaliativo).

Contribuições: Avanços Teóricos Centrais: síntese sistemática inovadora integrando ética informacional, economia institucional, teoria política, estudos críticos raça/tecnologia e soberania indígena; expansão da teoria da infosfera ao domínio linguístico-digital ("biodiversidade infosférica"); resolução de tensões commons-trust via demonstração de complementaridade funcional; teorização de "colonialismo semântico" capturando perpetuação colonial através de categorias linguísticas em IA. **Aplicações Práticas e Normativas:** Framework de Interoperabilidade Cultural operacionalizando justiça epistêmica em critérios avaliativos; Licenciamento Recíproco oferecendo modelo financeiro sustentável; LDII como ferramenta de procurement público fornecendo blueprint para agência do Sul Global; argumento

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

ético fundamentando diversidade como necessidade epistêmica (não apenas preservação cultural); reframing de questão multilíngue de desafio técnico para governança institucional, visibilizando conexões entre marginalização digital e padrões históricos coloniais/racistas.

Palavras-Chave: Diversidade Linguística**;** Governança de Dados**;** Data Trusts; Data Commons; Inteligência Artificial

ABSTRACT: *The rapid diffusion of large language models (LLMs) has transformed language into strategic infrastructure for knowledge production, public administration, and economic participation. While there is broad consensus that linguistic inequality in AI systems constitutes a structural problem, there is far less agreement on how to address it institutionally. This article argues that linguistic diversity initiatives in AI must move beyond technical expansion and engage with data governance and institutional design. Drawing on contemporary work in information ethics (Floridi), commons governance (Ostrom), and data stewardship (Delacroix & Lawrence), the article contends that linguistic data portals should be conceived as data trusts and/or data commons—not as neutral repositories. The analysis demonstrates that the marginalization of non-hegemonic languages represents not merely a technical failure, but a symptom of structural asymmetries in global technological governance that threatens both epistemic justice and informational sovereignty.*

Research: *This study investigates the governance of linguistic data in AI system development, focusing on the structural marginalization of non-hegemonic languages in large language models. The central question is: how can institutional governance structures transform linguistic data portals from mere technical repositories into instruments of epistemic justice, cultural sovereignty, and digital sustainability? The research begins from empirical findings: Indigenous and minority languages represent less than 1% of training data for LLMs; African languages (30% of global linguistic diversity) account for less than 1% of NLP research output; South American data represents less than 0.2% of multimodal datasets. This underrepresentation does not constitute a technical scaling problem, but manifests structural asymmetries with consequences for epistemic justice, informational sovereignty, and cultural sustainability.*

Methodology: *The research employs interdisciplinary conceptual analysis combined with comparative case study and normative-institutional framework development, integrating methods from: philosophy of technology (Floridi), institutional economics (Ostrom), political theory (Pettit, Mouffe), critical race and technology studies (Benjamin, Noble, Rosa & Flores), and law/data governance. The framework emerges from iterative synthesis among theoretical literature, analysis of international cases (CLARIN ERIC, Mozilla Common Voice, FirstVoices, DECODE, MIDATA), and normative principles. The process followed abductive logic (Peirce): identification of empirical patterns, formulation of explanatory hypotheses, refinement through confrontation with theory and additional cases. Limitations include selection bias (concentration in the Global North), scarce data on Brazilian Indigenous languages, and the need for validation of the framework's universal applicability.*

Results: *The research produced a tripartite conceptual framework articulating three dimensions: (A) Normative—linguistic diversity as infospheric biodiversity (Floridi),*

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

qualitative informational entropy, epistemic justice (Fricker) as evaluative criterion; (B) Institutional—Ostrom's principles applied to linguistic resources, data trusts as fiduciary mechanism, hybrid commons-trust models; (C) Political-Legal—CARE principles, non-domination (Pettit) as informational sovereignty, legal interoperability. Analysis of international cases (CLARIN ERIC, Mozilla Common Voice, FirstVoices, DECODE, MIDATA) generated eight design principles: contextual stewardship, federated architecture, graduated access, meaningful participation, transparency/accountability, benefit flows, long-term sustainability, technical-legal integration. Conceptual innovations include: Cultural Interoperability (metric assessing ontological fidelity, pragmatic sensitivity, ethical respect, historical accountability); Reciprocal Licensing (sustainability via redistributed commercial taxation); Digital Linguicide ("digital carbon footprint"); LDII (evaluative instrument).

Contributions: *Core Theoretical Advances: First systematic synthesis integrating information ethics, institutional economics, political theory, critical race/technology studies, and Indigenous sovereignty; expansion of infosphere theory to the linguistic-digital domain ("infospheric biodiversity"); resolution of commons-trust tensions via demonstration of functional complementarity; theorization of "semantic colonialism" capturing colonial perpetuation through linguistic categories in AI. Practical and Normative Applications: Cultural Interoperability framework operationalizing epistemic justice into evaluative criteria; Reciprocal Licensing offering sustainable financial model; LDII as public procurement tool providing blueprint for Global South agency; ethical argument grounding diversity as epistemic necessity (not merely cultural preservation); reframing of multilingual question from technical challenge to institutional governance, making visible connections between digital marginalization and historical colonial/racist patterns.*

Keywords: *Linguistic Diversity; Data Governance; Data Trusts; Data Commons; Artificial Intelligence*

1. INTRODUÇÃO: A NATUREZA ESTRUTURAL DA EXCLUSÃO LINGUÍSTICA

A língua tornou-se substrato central da inteligência artificial contemporânea. Em sistemas de IA em larga escala, dados linguísticos não mais funcionam meramente como meio de comunicação ou expressão cultural, mas como recurso central que molda autoridade epistêmica, capacidade institucional e inclusão social.

A sub-representação linguística em sistemas de IA gera efeitos que se estendem muito além do desempenho técnico. Análises empíricas revelam disparidades sistemáticas: a maioria dos datasets utilizados em LLMs consiste predominantemente de conteúdo em inglês (Longpre et al., 2024). Línguas indígenas e minoritárias representam menos de 1% dos dados de treinamento disponíveis (Joshi et al., 2020). Línguas africanas, representando 30% da diversidade linguística global,

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

correspondem a menos de 1% da produção de pesquisa em processamento de linguagem natural (NLP). Dados da América do Sul representam menos de 0,2% de datasets multimodais (Longpre et al., 2024), enquanto línguas indígenas brasileiras enfrentam marginalização ainda mais severa.

Línguas com presença digital limitada tendem a exibir menor precisão em tarefas de processamento de linguagem natural e disponibilidade reduzida em serviços digitais públicos. Com o tempo, essas disparidades consolidam-se em assimetrias epistêmicas, moldando quais categorias, significados e visões de mundo são codificadas em sistemas automatizados (Rosa & Flores, 2017).

1.1 RACISMO LINGUÍSTICO ALGORÍTMICO E ASSIMETRIAS DE PODER

O fenômeno do "racismo linguístico algorítmico" (Dovchin, 2020; Rosa & Flores, 2021) descreve como sistemas de IA perpetuam hierarquias linguísticas ao privilegiar variedades dominantes enquanto marginalizam práticas comunicativas de comunidades racializadas e povos indígenas. Isso não é um acidente técnico, mas manifestação de estruturas político-econômicas mais amplas.

O conceito de "algoritmos de opressão" de Noble (2018) demonstra como mecanismos de busca e sistemas automatizados reforçam hierarquias raciais através de seu design e dados. Aplicada à IA linguística, essa estrutura revela como a sub-representação não é meramente quantitativa, mas qualitativa, afetando qual conhecimento é considerado válido, quais expressões são marcadas como "padrão" e quais práticas comunicativas são algoritmicamente policiadas como desviantes (Benjamin, 2019).

Os relatórios da UNESCO (2023, 2024) sobre IA generativa e diversidade cultural reconhecem explicitamente a sub-representação e assimetria linguística como desafios estruturais com forte viés favorecendo o inglês e um pequeno número de línguas em modelos fundacionais e sistemas de IA generativa.

A concepção de infosfera de Luciano Floridi fornece estrutura normativa crucial para enfrentar esses desafios (Floridi, 2013). Floridi caracteriza sociedades contemporâneas como embebidas em um ambiente informacional onde interações entre agentes humanos e artificiais geram resultados moralmente significativos. A ética da informação, como Floridi a articula, requer atenção à saúde e florescimento

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

de todo o ecossistema informacional, não meramente proteção contra danos individuais, mas também coletivos e sociais.

O enquadramento ecológico de Floridi é particularmente relevante porque desloca a atenção de danos isolados para degradação cumulativa do ambiente informacional. Sua ênfase em qualidade da informação, variedade, pluralismo e acesso posiciona a diversidade linguística como condição constitutiva de uma infosfera saudável, e não como preocupação cultural auxiliar. Assim como biodiversidade é essencial para resiliência ecológica, a diversidade linguística funciona como "biodiversidade infosférica" (Floridi, 2014).

Portais de dados linguísticos, portanto, operam como sítios institucionais onde inclusão, autoridade e legitimidade são ativamente negociadas. A questão não é simplesmente se certas línguas são "incluídas" em sistemas de IA, mas sob quais arranjos de governança tal inclusão ocorre, quem controla os termos de uso e como benefícios e riscos são distribuídos.

2. *DATA COMMONS, DATA TRUSTS* E RECURSOS LINGUÍSTICOS COMPARTILHADOS

A governança de dados linguísticos demanda estruturas institucionais capazes de equilibrar múltiplos imperativos: fomentar inovação, proteger direitos culturais, assegurar acesso equitativo e manter *accountability.*

A teoria dos comuns, desenvolvida por Elinor Ostrom (1990), oferece fundação robusta para governar recursos compartilhados sem recorrer à privatização ou controle centralizado. O trabalho empírico de Ostrom por sua vez demonstra que comuns podem ser sustentavelmente geridos através de regras claramente definidas, mecanismos de monitoramento, arranjos de escolha coletiva e sanções graduadas.

Os oito princípios de design de Ostrom para instituições de recursos comuns duradouros merecem enumeração explícita no contexto da governança de dados linguísticos:

1. Limites claramente definidos (quem tem direitos de acesso a dados linguísticos)

2. Equivalência proporcional entre benefícios e custos

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

3. Arranjos de escolha coletiva (partes afetadas participam na elaboração de regras)

4. Monitoramento por atores responsabilizáveis

5. Sanções graduadas para violações de regras

6. Mecanismos de resolução de conflitos

7. Reconhecimento mínimo de direitos por autoridades externas

8. Empreendimentos aninhados para sistemas complexos

Aplicada a dados linguísticos, a governança baseada em comuns enfatiza administração compartilhada e benefício coletivo. Corpora linguísticos, quando concebidos como comuns, são compreendidos como recursos cujo valor deriva do uso coletivo enquanto requerem proteção contra depleção ou cercamento.

Data trusts fornecem modelo de governança complementar baseado em responsabilidade fiduciária. Como articulado pelo Open Data Institute (Delacroix & Lawrence, 2019), data trusts envolvem administradores independentes gerindo dados em nome de outros sob propósitos e deveres claramente definidos. A estrutura de trust importa princípios de séculos de direito fiduciário anglo-americano, estabelecendo separação entre controle legal (detido por trustees) e interesse beneficiário (detido por beneficiários).

Para dados linguísticos—particularmente onde direitos culturais coletivos estão implicados—data trusts possibilitam separação entre autoridade e administração, reduzindo riscos de uso indevido enquanto apoiam pesquisa e inovação legítimas. Incorporam compromissos éticos diretamente na prática institucional, em vez de tratá-los como restrições externas.

O modelo de data trust aborda desafio fundamental na governança de dados linguísticos: como possibilitar usos benéficos enquanto se previne danos. Diferentemente de regimes simples de licenciamento, trusts estabelecem relações continuadas de responsabilidade. Trustees devem ativamente monitorar usos, avaliar impactos e intervir quando termos são violados.

Os desenhos institucionais mais promissores podem combinar elementos de ambos os modelos de commons e trust. Um portal de dados linguísticos poderia

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

operar como commons na definição de recursos compartilhados e regras de acesso, enquanto incorpora mecanismos de trust para supervisão fiduciária.

Pluralismo de governança no sentido de coexistência de múltiplos arranjos de governança sobreposto pode ser necessário para abordar a diversidade de contextos de dados linguísticos. Línguas de alto recurso com tradições literárias estabelecidas enfrentam desafios diferentes de línguas indígenas ameaçadas mantidas por pequenas comunidades de falantes.

3. LÍNGUAS INDÍGENAS E GOVERNANÇA BASEADA EM CARE

A governança de dados linguísticos indígenas levanta considerações normativas distintas enraizadas em histórias de colonialismo, práticas extrativistas de pesquisa e lutas continuadas por autodeterminação.

Os Princípios CARE para Governança de Dados Indígenas—Benefício Coletivo, Autoridade para Controlar, Responsabilidade e Ética—fornecem estrutura explicitamente desenhada para contrabalançar limitações nos amplamente adotados princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) (Carroll et al., 2020). Enquanto FAIR enfatiza abertura técnica e reutilização, CARE coloca em primeiro plano benefício coletivo, autoridade para controlar, responsabilidade e ética a partir de perspectivas indígenas.

Benefício Coletivo requer que ecossistemas de dados sejam desenhados para possibilitar que povos indígenas derivem valor de seus dados de maneiras que se alinhem com suas visões de mundo e objetivos coletivos.

Autoridade para controlar reconhece os direitos de povos indígenas de governar coleta, propriedade e aplicação de seus dados. Isso estende o princípio de consentimento livre, prévio e informado (FPIC) da ética de pesquisa para governança continuada de dados.

Responsabilidade enfatiza que aqueles trabalhando com dados indígenas devem assegurar que seu uso apoia autodeterminação e benefício coletivo de povos indígenas.

Ética requer alinhamento com estruturas éticas indígenas, não meramente princípios bioéticos ocidentais. Éticas indígenas frequentemente enfatizam relacionalidade, reciprocidade e respeito por interesses coletivos e intergeracionais.

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

Os princípios CARE alinham-se com instrumentos internacionais de direitos humanos, incluindo a Declaração das Nações Unidas sobre os Direitos dos Povos Indígenas (UNDRIP), que afirma os direitos de povos indígenas de manter, controlar, proteger e desenvolver seu patrimônio cultural e conhecimento tradicional (ONU, 2007).

A Convenção da UNESCO de 2005 sobre a Proteção e Promoção da Diversidade das Expressões Culturais similarmente reconhece diversidade cultural como direito fundamental e afirma a importância de proteger expressões culturais de povos indígenas e minorias.

Incorporar esses princípios em portais de dados linguísticos é essencial para evitar práticas extrativistas e assegurar alinhamento com padrões internacionais de soberania de dados indígenas.

Sem intervenção, a trajetória do desenvolvimento de IA ameaça o que pode ser denominado "linguicídio digital"—o apagamento sistemático da diversidade linguística em sistemas computacionais, contribuindo para acelerado perigo de línguas (Bender et al., 2021). Das aproximadamente 7.000 línguas do mundo, muitas são faladas por pequenas comunidades já enfrentando pressões de mudança linguística.

O Relatório UNESCO de 2025 do Grupo Independente de Especialistas em Inteligência Artificial e Cultura (CULTAI) identifica explicitamente povos indígenas, comunidades locais e grupos historicamente marginalizados como enfrentando riscos particulares, incluindo apropriação cultural, descontextualização de conhecimento tradicional, exploração comercial sem retorno e violação de consentimento livre, prévio e informado.

4. PRIVACIDADE, VIGILÂNCIA, A NATUREZA DUAL DE DADOS LINGUÍSTICOS E EXTRAÇÃO LINGUÍSTICA

Dados linguísticos ocupam posição peculiar: simultaneamente patrimônio cultural coletivo e informação pessoal potencialmente identificadora. Essa dualidade demanda estruturas de governança abordando ambas dimensões.

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

Estilometria moderna e técnicas de atribuição de autoria demonstram que padrões linguísticos funcionam como biometria comportamental. Estilo de escrita, seleção de vocabulário, preferências sintáticas e padrões discursivos podem identificar indivíduos com alta precisão (Brennan et al., 2012).

O Regulamento Geral de Proteção de Dados (GDPR) da União Europeia reconhece essa realidade. Sob o GDPR, dados pessoais incluem qualquer informação relativa à pessoa identificável. A Lei Geral de Proteção de Dados do Brasil (LGPD) similarmente define dados pessoais amplamente.

A análise de Shoshana Zuboff sobre "capitalismo de vigilância" (2019) revela como dados comportamentais são extraídos, analisados e comodificados para prever e influenciar comportamento. Dados linguísticos constituem forma particularmente rica de dados comportamentais.

O conceito de "colonialismo de dados" (Couldry & Mejias, 2019) descreve como extração de dados contemporânea espelha padrões coloniais históricos: recursos fluem de periferias para centros, valor é apropriado por atores dominantes e comunidades locais arcam com custos enquanto recebem benefícios mínimos.

Tecnologias de aprimoramento de privacidade (PETs) oferecem potenciais complementos técnicos para governança institucional. Privacidade diferencial fornece garantias matemáticas de que registros individuais em dataset não podem ser identificados (Dwork & Roth, 2014).

Contudo, a percepção crítica é que soluções técnicas sozinhas são insuficientes. Tecnologias de aprimoramento de privacidade podem complementar governança institucional, mas não podem substituí-la. Decisões sobre trade-offs aceitáveis de privacidade-utilidade, usos apropriados de dados linguísticos e distribuição justa de benefícios são inerentemente normativas e políticas, requerendo deliberação democrática.

5. ARQUITETURAS DE GOVERNANÇA: MODELOS INTERNACIONAIS

Experiências internacionais demonstram que modelos de governança baseados em *commons* e *trust* para dados linguísticos são viáveis.

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

A *Common Language Resources and Technology Infrastructure* (CLARIN ERIC) reúne dezenas de centros de pesquisa através de mais de 20 países europeus, fornecendo acesso federado a recursos linguísticos. Opera através de modelo de governança distribuído que combina administração nacional de dados com padrões técnicos compartilhados, procedimentos de revisão ética e regimes de acesso graduados.

Mozilla Common Voice exemplifica data *commons* comunitário para reconhecimento de fala, convidando falantes mundialmente a contribuir gravações de voz em suas línguas (Ardila et al., 2020). O projeto coletou mais de 20.000 horas de fala em 100+ línguas.

Várias comunidades indígenas desenvolveram abordagens inovadoras de governança para arquivos linguísticos. O First Peoples' Cultural Council na Colúmbia Britânica, Canadá, opera FirstVoices—plataforma online para documentação e revitalização de línguas indígenas controlada por comunidades indígenas. Cada comunidade linguística determina suas próprias políticas de acesso.

Esses exemplos internacionais sugerem princípios convergentes de design:

1. **Administração Contextual**: Arranjos de governança devem ser adaptados a contextos linguísticos, culturais e institucionais específicos

2. **Arquitetura Federada**: Hospedagem distribuída de dados com coordenação centralizada

3. **Acesso Graduado**: Sistemas de acesso em camadas acomodam diversos níveis de sensibilidade

4. **Participação de Stakeholders**: Envolvimento significativo de comunidades linguísticas

5. **Transparência e Accountability**: Documentação clara de fontes, usos e impactos de dados

6. **Fluxos de Benefício**: Arranjos explícitos assegurando que comunidades recebam benefícios tangíveis

7. **Sustentabilidade de Longo Prazo**: Estruturas institucionais e financeiras capazes de manter funções de governança através de décadas

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

8. **Integração Técnico-Legal**: Combinando instrumentos legais de governança com arquiteturas técnicas

6. POR QUE GOVERNANÇA IMPORTA - DIVERSIDADE LINGUÍSTICA COMO NECESSIDADE EPISTÊMICA

O caso para conceber portais de dados linguísticos como data trusts e commons repousa não meramente em vantagens funcionais, mas em compromissos normativos fundamentais.

Diversidade linguística não é simplesmente patrimônio cultural a ser preservado, mas recurso epistêmico essencial para florescimento humano coletivo. Línguas diferentes codificam ontologias, epistemologias e sistemas de conhecimento prático distintos acumulados ao longo de milênios (Maffi, 2001).

Quando sistemas de IA são treinados predominantemente em dados em inglês refletindo categorias culturais anglo-americanas, eles codificam visões de mundo particulares como universais, marginalizando estruturas alternativas para compreender realidade. Isso constitui o que Miranda Fricker (2007) denomina "injustiça epistêmica".

A concepção republicana de liberdade como não-dominação de Philip Pettit fornece estrutura para compreender soberania informacional (Pettit, 1997). Dominação existe não apenas quando interferência ocorre, mas quando interferência arbitrária permanece possível devido a assimetrias de poder.

Data trusts e estruturas de commons podem reduzir tal dominação estabelecendo estruturas institucionais limitando interferência arbitrária. Deveres fiduciários restringem discrição de trustees. Governança de commons dá voz a stakeholders em decisões coletivas.

O conceito de "colonialismo semântico" captura como sistemas de IA podem perpetuar padrões coloniais de extração e imposição de conhecimento. O Índice de Impacto de Diversidade Linguística (LDII) proposto deve evoluir além de medidas de desempenho gramatical para avaliação de "Interoperabilidade Cultural"—a capacidade de modelos de preservar ontologias nativas e respeitar contextos históricos de comunidades minoritárias e povos indígenas.

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

Interoperabilidade Cultural requer que sistemas de IA mantenham:

- **Fidelidade ontológica**: Preservando categorias conceituais
- **Sensibilidade pragmática**: Reconhecendo significados dependentes de contexto
- **Respeito ético**: Honrando restrições sobre linguagem sagrada ou cerimonial
- **Accountability histórica**: Reconhecendo histórias coloniais

O relatório UNESCO CULTAI (2025) introduz o conceito de "dados culturais" como valor cognitivo coletivo, distinguindo entre expressões culturais explícitas (protegidas por direitos autorais) e expressões culturais implícitas/latentes (traços digitais, linguagem cotidiana) que formam fundação não-reconhecida e não-compensada para treinamento de IA.

Se dados linguísticos representam trabalho cognitivo coletivo—práticas comunicativas acumuladas de comunidades ao longo de gerações—então sua apropriação para desenvolvimento comercial de IA sem compensação constitui enriquecimento injusto.

7. DA PRINCÍPIO À PRÁTICA: IMPLEMENTANDO GOVERNANÇA

7.1 ESTRUTURAS LEGAIS E REGULATÓRIAS

Estabelecer data trusts e commons linguísticos requer fundações legais reconhecendo direitos coletivos sobre dados e possibilitando arranjos fiduciários:

- **Direitos Sui Generis sobre Bases de Dados**: Diretiva de base de dados da UE oferece base potencial
- **Estruturas de Direitos Bioculturais**: Protocolo de Nagoya estabelece requisitos para consentimento prévio informado e compartilhamento de benefícios
- **Regimes de Proteção de Dados**: GDPR e LGPD estabelecem princípios aplicáveis
- **Proteção de Patrimônio Cultural**: Convenção UNESCO 2003 reconhece língua como veículo de patrimônio cultural intangível

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

O Plano Brasileiro de IA (PBIA, 2024) propõe explicitamente Ação 9: curar datasets nacionais e apoiar desenvolvimento de modelos fundacionais, particularmente LLMs especializados em português.

Dentro da infraestrutura técnica para acesso governado os componentes técnicos incluem:

- **Gestão de Identidade e Acesso (IAM)**: Sistemas de autenticação federados
- **Rastreamento e Auditoria de Uso**: Logs imutáveis de acesso e uso de dados
- **Tecnologias de Aprimoramento de Privacidade**: Privacidade diferencial, computação multipartidária segura
- **Aplicação Automatizada de Políticas**: Arquiteturas técnicas traduzindo políticas de governança em código executável

7.2 MODELOS DE FINANCIAMENTO E SUSTENTABILIDADE

- **Investimento Público como Infraestrutura**: Tratando governança de dados linguísticos como infraestrutura digital pública
- **Mecanismos de Compartilhamento de Benefícios**: Requerendo desenvolvedores comerciais de IA a contribuir financeiramente
- **Dotações Filantrópicas**: Estabelecendo endowments gerando retornos para operações de governança
- **Consórcios Multipartes**: Combinando contribuições de governos, universidades e parceiros industriais

Um Framework de Licenciamento Recíproco crítico: o Trust funciona como administrador fiduciário gerindo acesso aos commons linguísticos. Enquanto acesso permanece gratuito ou baixo custo para pesquisa acadêmica e desenvolvimento comunitário, entidades comerciais—particularmente desenvolvedores de IA em larga escala—seriam requeridos a pagar uma "Taxa de Infraestrutura Linguística". Esses fundos são redistribuídos às comunidades linguísticas para revitalização de línguas e programas de letramento digital.

Revista Jurídica unicuritiba
Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: 2316-753X

8 CONSIDERAÇÕES FINAIS: RUMO A UMA ECOLOGIA LINGUÍSTICA E SUSTENTABILIDADE DIGITAL

A governança da diversidade linguística deve ser compreendida como pilar fundamental da Ecologia Digital. Assim como monoculturas biológicas ameaçam resiliência ambiental, "monoculturas linguísticas" em sistemas de IA comprometem sustentabilidade do ecossistema global de conhecimento. A erosão de línguas não-hegemônicas representa forma de "Pegada de Carbono Digital"—custo oculto do avanço tecnológico que coloca em risco a diversidade cognitiva e cultural do mundo.

Promover infosfera plurilíngue não é meramente ato de preservação cultural; é requisito essencial para sustentabilidade de longo prazo da própria IA. Adotando Data Trusts como salvaguardas institucionais, asseguramos que desenvolvimento de inteligência artificial esteja alinhado com princípios de Justiça Epistêmica e preservação do patrimônio simbólico humano.

O que está em jogo na governança de dados linguísticos não é meramente inclusão técnica, mas capacidade de sociedades reterem agência significativa sobre condições linguísticas através das quais sistemas automatizados interpretam, classificam e agem sobre o mundo. Na ausência de instituições robustas de governança, dados linguísticos continuarão fluindo de comunidades para corporações, de periferias para centros, de prática viva para extração computacional, perpetuando desigualdades estruturais.

A visão de Floridi da infosfera como ambiente ético requerendo administração ativa fornece horizonte normativo. Infosfera saudável requer não apenas quantidade de informação, mas qualidade, diversidade e acesso equitativo. Diversidade linguística funciona como biodiversidade infosférica—essencial para resiliência, criatividade e florescimento de comunidades humanas em ambientes digitais.

A demonstração de Ostrom de que comuns podem ser governados com sucesso (1990, 2015), e trabalho contemporâneo sobre data trusts mostrando que responsabilidade fiduciária pode ser institucionalizada para benefício coletivo (Delacroix & Lawrence, 2019), fornecem fundações práticas.

As estruturas de governança aqui delineadas—combinando administração compartilhada baseada em commons com supervisão fiduciária baseada em trust, incorporando princípios CARE para dados indígenas, abordando preocupações de

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

privacidade e vigilância através de mecanismos técnicos e institucionais—representam não solução final, mas repertório institucional adaptável a contextos diversos.

A experiência brasileira sugere modelo onde diversidade linguística torna-se não obstáculo à padronização tecnológica, mas vantagem comparativa estratégica. Posicionando-se como custodiante de "biodiversidade linguística" e conhecimento tradicional, o Brasil poderia liderar desenvolvimento de "IA Multimodal Resiliente"—sistemas treinados em dados diversos exibindo menos alucinação cultural e maior capacidade de generalização ética, transformando diversidade brasileira em ativo indispensável para criar sistemas de IA globalmente competitivos e menos enviesados.

O caminho adiante demanda ação coordenada: estruturas legais reconhecendo direitos coletivos sobre dados linguísticos; infraestruturas técnicas possibilitando acesso e uso governados; mecanismos de financiamento sustentáveis; construção de capacidade possibilitando participação comunitária significativa; proteções de privacidade prevenindo vigilância e discriminação; e cooperação internacional estabelecendo normas e reconhecimento mútuo.

As instituições que construímos agora—ou falhamos em construir—determinarão qual futuro se materializa: um onde poder tecnológico se concentra nas mãos de poucas corporações e nações-estado, impondo homogeneização linguística e cultural; ou um onde comunidades diversas retêm soberania sobre seu patrimônio linguístico, participam significativamente na modelagem de sistemas de IA afetando suas vidas e beneficiam-se equitativamente do valor que suas práticas linguísticas criam.

REFERÊNCIAS

ARDILA, R., et al. (2020). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of LREC 2020*.

BENDER, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *ACM FAccT 2021*, 610-623.

BENJAMIN, R. (2019). *Race After Technology*. Polity Press.

BRASIL. Lei nº 13.709/2018. Lei Geral de Proteção de Dados Pessoais (LGPD).

BRASIL. MCTI (2024). *Plano Brasileiro de Inteligência Artificial (PBIA)*.

BRENNAN, M., Afroz, S., & Greenstadt, R. (2012). Adversarial Stylometry. *ACM TISSEC*, 15(3), 1-22.

CARROLL, S. R., et al. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), 43.

COULDRY, N., & Mejias, U. A. (2019). *The Costs of Connection*. Stanford University Press.

DELACROIX, S., & Lawrence, N. D. (2019). Bottom-up Data Trusts. *International Data Privacy Law*, 9(4), 236-252.

DOVCHIN, S. (2020). Linguistic Racism and International Students. *IJBEB*, 23(7), 804-818.

DWORK, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in TCS*, 9(3-4), 211-407.

FLORIDI, L. (2013). *The Ethics of Information*. Oxford University Press.

FLORIDI, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

FLORIDI, L. (2018). Soft Ethics and the Governance of the Digital. *Philosophy & Technology*, 31(1), 1-8.

FRICKER, M. (2007). *Epistemic Injustice*. Oxford University Press.

FRISCHMANN, B. M., Madison, M. J., & Strandburg, K. J. (Eds.). (2014). *Governing Knowledge Commons*. Oxford University Press.

JOSHI, P., et al. (2020). The State and Fate of Linguistic Diversity in NLP. *Proceedings of ACL 2020*, 6282-6293.

KREUTZER, J., et al. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *TACL*, 10, 50-72.
Kukutai, T., & Taylor, J. (Eds.). (2016). *Indigenous Data Sovereignty: Toward an Agenda*. ANU Press.

LONGPRE, S., et al. (2024). The Data Provenance Initiative. *arXiv preprint arXiv:2408.04110*.

MAFFI, L. (Ed.). (2001). *On Biocultural Diversity*. Smithsonian Institution Press.

MARIANI, J. J. (2024). Language Technology for All: A Challenge. *UNESCO Report on Languages*.

Submetido em: 18/09/2025
Aprovado em: 10/01/2026
Avaliação: Double Blind Reviewe
ISSN: **2316-753X**

MAZZUCATO, M. (2018). *The Value of Everything*. PublicAffairs.

MOUFFE, C. (2000). *The Democratic Paradox*. Verso Books.

NISSENBAUM, H. (2009). *Privacy in Context*. Stanford University Press.

NOBLE, S. U. (2018). *Algorithms of Oppression*. NYU Press.

OSTROM, E. (1990). *Governing the Commons*. Cambridge University Press.

OSTROM, E. (2015). *Governing the Commons* (Canto Classics edition). Cambridge University Press.

PETTIT, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford University Press.

ROSA, J., & FLORES, N. (2017). Unsettling Race and Language. *Language in Society*, 46(5), 621-647.

ROSA, J., & FLORES, N. (2021). Hearing Language Gaps and Reproducing Social Inequalities. In *The Handbook of Language and Race* (pp. 345-362). Routledge.

SKUTNABB-KANGAS, T. (2000). *Linguistic Genocide in Education*. Lawrence Erlbaum Associates.

STIEGLER, B. (1998). *Technics and Time, 1*. Stanford University Press.

STIEGLER, B. (2014). *Symbolic Misery, Volume 1*. Polity Press.

UNESCO (2023). *Generative AI and the Diversity of Cultural Expressions*. Paris: UNESCO.

UNESCO (2024). *Guidance for Governing AI for the Diversity of Cultural Expressions*. Paris: UNESCO.

UNESCO (2025). *Report of the Independent Group of Experts on AI and Culture (CULTAI)*. Paris: UNESCO.

UNITED NATIONS (2007). *Declaration on the Rights of Indigenous Peoples*. UN Doc. A/RES/61/295.

ZUBOFF, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.