

Comunicação

Busca Semântica Multilíngue em Repositórios Digitais: Diversidade Linguística

Multilingual Semantic Search in Digital Repositories: Linguistic Diversity

Búsqueda Semántica Multilingüe en Repositorios Digitales: Diversidad Lingüística

Washington Luís Ribeiro de Carvalho Segundo

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasil

ORCID: <https://orcid.org/0000-0003-3635-9384>

Lattes: <http://lattes.cnpq.br/9453481318889500>

washingtonsegundo@ibict.br

Lautaro Matas

LA Referencia

ORCID: <https://orcid.org/0000-0003-0272-1592>

lautaro.matas@lareferencia.redclara.net

Luciana Mara Silva

Universidade do Estado de Santa Catarina (Udesc), Brasil

ORCID: <https://orcid.org/0000-0002-3513-2375>

Lattes: <http://lattes.cnpq.br/3005116350546030>

luciana.ms@udesc.br

Priscila Machado Borges Sena

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasil

ORCID: <https://orcid.org/0000-0002-5612-4315>

Lattes: <https://lattes.cnpq.br/0155235005204514>

priscilasena@ibict.br

Dayane Dornelles

Universidade do Estado de Santa Catarina (Udesc), Brasil

ORCID: <https://orcid.org/0000-0002-4171-6502>

Lattes: <http://lattes.cnpq.br/4971626543868135>

dayane.dornelles@udesc.br

Resumo

Esta comunicação tem por objeto discutir a implementação de uma ferramenta de busca semântica multilíngue em repositórios digitais como estratégia para ampliar a diversidade linguística na construção e divulgação do conhecimento ibero-americano. Em um cenário global é fundamental a discussão sobre infraestruturas que garantam a representação, distribuição, participação e reconhecimento na produção do saber em múltiplas linguagens. É uma pesquisa exploratória e descritiva por meio de uma revisão narrativa e adota uma abordagem teórico-argumentativa. A adoção de modelos de Processamento de Linguagem Natural (PLN), especificamente embeddings multilíngues, permite a recuperação de informação independente do idioma de busca ou do documento, rompendo silos linguísticos e ampliando a participação equitativa na ciência. O estudo demonstra que repositórios digitais não são meros depósitos técnicos, mas infraestruturas sociotécnicas de mediação que articulam tecnologia, política científica e compromisso social. Os resultados indicam que a adoção de sistemas de busca baseados em palavras-chave, associados a buscas semânticas vetoriais são tecnicamente viáveis e politicamente necessários para estabelecer a diversidade linguística representada na ciência produzida no contexto ibero-americano.

Palavras-chave:

Busca semântica multilíngue; Repositórios Digitais; Diversidade linguística; Justiça Informacional.

Abstract

This proposal aims to discuss the implementation of a multilingual semantic search tool in digital repositories as a strategy to expand linguistic diversity in the construction and dissemination of Ibero-American knowledge. In a global scenario, it is essential to discuss infrastructures that guarantee representation, distribution, participation, and recognition in the production of knowledge across multiple languages. This is exploratory and descriptive research based on a narrative review, adopting a theoretical-argumentative approach. The adoption of Natural Language Processing (NLP) models, specifically multilingual embeddings, enables information retrieval regardless of the search or document language, breaking down linguistic silos and expanding equitable participation in science. The study demonstrates that digital repositories are not merely technical deposits, but sociotechnical mediation infrastructures that articulate

technology, scientific policy, and social commitment. The results indicate that the adoption of keyword-based search systems combined with vector semantic searches is technically feasible and politically necessary to establish the linguistic diversity represented in science produced in the Ibero-American context.

Keywords:

Multilingual semantic search; Digital Repositories; Linguistic diversity; Informational Justice.

Resumen

Esta propuesta tiene por objeto discutir la implementación de una herramienta de búsqueda semántica multilingüe en repositorios digitales como estrategia para ampliar la diversidad lingüística en la construcción y divulgación del conocimiento iberoamericano. En un escenario global, es fundamental discutir infraestructuras que garanticen la representación, distribución, participación y reconocimiento en la producción del saber en múltiples lenguas. Se trata de una investigación exploratoria y descriptiva mediante una revisión narrativa, adoptando un enfoque teórico-argumentativo. La adopción de modelos de Procesamiento de Lenguaje Natural (PLN), específicamente embeddings multilingües, permite la recuperación de información independientemente del idioma de búsqueda o del documento, rompiendo silos lingüísticos y ampliando la participación equitativa en la ciencia. El estudio demuestra que los repositorios digitales no son meros depósitos técnicos, sino infraestructuras sociotécnicas de mediación que articulan tecnología, política científica y compromiso social. Los resultados indican que la adopción de sistemas de búsqueda basados en palabras clave, asociados a búsquedas semánticas vectoriales, es técnicamente viable y políticamente necesaria para establecer la diversidad lingüística representada en la ciencia producida en el contexto iberoamericano.

Palabras clave:

Búsqueda semántica multilingüe; Repositorios Digitales; Diversidad lingüística; Justicia Informacional.

Introdução

A ampliação do Acesso Aberto modificou a comunicação científica, mas não resolveu a invisibilidade linguística, que atua como mecanismo de exclusão e injustiça epistêmica: conteúdos disponíveis permanecem inalcançáveis para quem pesquisa em diferentes idiomas (Márquez & Porras, 2020; Vučković & Sikimić, 2023). A comunicação científica em múltiplas línguas é condição para a efetividade social da ciência (Márquez & Porras, 2020) e a liberdade de circulação da

informação (Le Coadic, 2004) fica limitada quando a recuperação depende do domínio de idiomas hegemônicos, restringindo a produção de soluções contextualizadas.

Nos repositórios digitais, a busca lexical predominante reforça silos linguísticos, dificultando a recuperação de conteúdos em diferentes idiomas. Avanços em Processamento de Linguagem Natural, como embeddings multilíngues, ampliam a conexão entre documentos independentemente da língua (Weckmüller et al., 2025). No entanto, ainda é incipiente a articulação entre essas soluções técnicas e os princípios da Justiça Informacional, tratando a recuperação da informação como dimensão ética e política das infraestruturas científicas.

Diante disso, questiona-se: como a implementação de busca semântica multilíngue em repositórios digitais pode superar a invisibilidade linguística na comunicação científica ibero-americana e promover a Justiça Informacional no contexto da Ciência Aberta? Este artigo tem por objetivo discutir essa implementação como estratégia para ampliar a diversidade linguística na construção, circulação e apropriação do conhecimento científico.

Aspectos metodológicos

Esta comunicação caracteriza-se como uma pesquisa qualitativa, de natureza exploratória e descritiva, orientada por uma abordagem teórico-argumentativa. Fundamenta-se em uma revisão narrativa da literatura, com o objetivo de discutir a implementação de uma ferramenta de busca semântica multilíngue em repositórios digitais como estratégia para ampliar a diversidade linguística na construção, circulação e apropriação do conhecimento científico, à luz dos princípios da Justiça Informacional e da Ciência Aberta.

O corpus de análise foi constituído por documentos publicados entre 2019 e 2026, priorizando artigos de periódicos, trabalhos publicados em eventos e preprints. A seleção seguiu critérios previamente definidos: (a) pertinência direta ao problema da diversidade linguística na comunicação científica; (b) aderência ao contexto de repositórios digitais, agregadores ou infraestrutura aberta; (c) contribuição para a compreensão técnica da recuperação semântica multilíngue; e (d) disponibilidade de metadados e texto integral para análise interpretativa.

A análise orientou-se pela articulação entre dimensões sociotécnicas, operacionais e ético-políticas da recuperação da informação, buscando evidenciar como determinadas escolhas tecnológicas podem tanto reproduzir quanto mitigar desigualdades informacionais. Dessa forma, o estudo não se limita à descrição de soluções técnicas, mas examina criticamente suas implicações para o acesso, a visibilidade e a participação de diferentes comunidades linguísticas no ecossistema científico. Como resultado desse processo analítico, desenvolve-se uma proposição conceitual de modelo de busca semântica multilíngue estruturado em camadas, concebido como artefato conceitual-operacional orientado à promoção da diversidade linguística e da Justiça Informacional em repositórios digitais. Ressalta-se que não se trata de um estudo

experimental de desempenho, mas de uma construção teórica fundamentada na literatura, com potencial de aplicação prática e validação em pesquisas futuras.

Fundamentação teórica

O direito à informação pode ser reconhecido como um direito humano fundamental, intrinsecamente ligado à liberdade de expressão, ao desenvolvimento cultural e à participação social. Se a ciência é um empreendimento global, suas infraestruturas de acesso, mediação e circulação também precisam refletir a pluralidade linguística da humanidade. Nessa direção, a Justiça Informacional, conceituada por Mathiesen (2015), oferece um arcabouço ético relevante para analisar as desigualdades nos ecossistemas de informacionais, ao problematizar quem pode acessar, produzir, controlar, compartilhar e se beneficiar da informação em condições efetivamente justas.

A exclusão informacional, nesse sentido, compromete para além da equidade, a própria capacidade de sociedades periféricas produzirem inovação e responderem a seus próprios contextos sociais, científicos e culturais. Em termos normativos, essa desigualdade pode ser compreendida como problema de simultaneamente informacional, epistêmico e político. Ford e Alemneh (2024) demonstram que a exclusão de pesquisadores de países de baixa e média renda resulta de uma combinação de barreiras econômicas, infraestruturais e linguísticas, produzindo um ecossistema informacional menos saudável e menos democrático.

Vučković e Sikimić (2023), evidenciam que a centralidade de uma língua franca na ciência produz perdas epistêmicas, porque conhecimentos formulados em outros repertórios linguísticos não são simplesmente “traduzidos”: muitas vezes, eles são reduzidos, descontextualizados ou não chegam a ser incorporados ao debate internacional. Assim, a questão linguística não pode ser lida como obstáculo meramente comunicacional, mas como dimensão estruturante das condições de visibilidade, participação e reconhecimento na comunicação científica.

Nessa perspectiva, a Justiça Informacional exige considerar não apenas quem acessa a informação, mas sob quais condições esse acesso ocorre, quem participa de sua produção e circulação e quais sujeitos e saberes são reconhecidos nas infraestruturas que organizam o conhecimento (Mathiesen, 2015). Garcês-da-Silva et al. (2021) reforçam essa compreensão ao delinearem princípios orientadores da Justiça Informacional no campo biblioteconômico-informacional, entre os quais se destacam o combate às injustiças sociais e representacionais, a crítica às assimetrias de poder na informação, a defesa do acesso como direito humano, a valorização de licenças e tecnologias abertas, a centralidade da formação crítica para o uso, a produção e a apropriação da informação. Essa formulação é especialmente pertinente ao debate aqui proposto, pois evidencia que tecnologias e sistemas informacionais não são neutros: ao serem projetados, implementados e utilizados, podem reproduzir exclusões históricas ou contribuir para sua mitigação.

Ao se considerar a comunicação científica contemporânea, a promoção da Justiça Informacional pode ser analisada a partir de três dimensões complementares e interdependentes: distributiva, participativa e de reconhecimento. A dimensão distributiva refere-se à superação das barreiras estruturais e financeiras que restringem o acesso equitativo aos recursos e resultados de pesquisa, combatendo injustiças que limitam a capacidade de consumir e disseminar conhecimento de forma paritária nos canais tradicionais (Ford & Alemneh, 2024). A dimensão participativa envolve a garantia de que grupos e comunidades de diversas origens geopolíticas e linguísticas possam atuar ativamente na produção e circulação do conhecimento, tanto como acadêmicos quanto como cidadãos (Márquez & Porras, 2020), mitigando vieses linguísticos e a “marginalização hermenêutica” que negam atuação plena nas experiências sociais (Ford & Alemneh, 2024). Já a dimensão de reconhecimento requer a valorização ativa de sujeitos, línguas, saberes tácitos e expressões historicamente invisibilizadas (Ford & Alemneh, 2024; Garcês-da-Silva et al., 2021; Mathiesen, 2015; Sena, 2023), defendendo um multilinguismo equilibrado no qual línguas nativas e não-inglesas sejam documentadas e valorizadas (Dony et al., 2024). Esses três eixos não são acessórios, mas estruturantes de uma ciência mais plural, aberta e socialmente relevante: articulados, garantem a diversidade epistemológica e ajudam a construir um ecossistema de pesquisa global mais justo, saudável e plural, no qual a diversidade linguística deixa de ser apenas atributo da comunicação científica para constituir condição de justiça na própria organização do conhecimento. É justamente nesse horizonte que os repositórios digitais precisam ser compreendidos. Mais do que depósitos técnicos de objetos informacionais, eles operam como infraestruturas sociotécnicas de mediação, capazes de articular tecnologia, política, ética e compromisso público. Essas infraestruturas conectam produção, preservação, circulação e reuso do conhecimento, sustentadas por princípios como transparência, interoperabilidade, acessibilidade e Acesso Aberto. Nessa condição, os repositórios não apenas armazenam documentos: eles moldam possibilidades de descoberta, visibilidade e apropriação social do conhecimento. Quando orientados por princípios de Justiça Informacional, podem se tornar infraestruturas de convergência entre Ciência Aberta, democracia informacional e cidadania (Sena et al., 2026).

Sob esse prisma, repositórios que permanecem dependentes exclusivamente de mecanismos de busca lexical correm o risco de se converter em arquivos formalmente abertos, porém materialmente limitados. Ainda que a diversidade linguística esteja presente no acervo, ela permanece pouco acionável se o sistema não permitir que diferentes comunidades formulem consultas em seus próprios idiomas e encontrem documentos semanticamente relevantes. A desigualdade linguística, portanto, não se reproduz apenas na publicação científica, mas também nos mecanismos de organização, recuperação e ranqueamento da informação. Isso significa que a busca semântica multilíngue não deve ser entendida apenas como aprimoramento técnico, mas como estratégia de mediação compatível com os princípios da Justiça Informacional, ao ampliar condições de acesso, participação e reconhecimento em repositórios digitais.

Por outro lado, a busca semântica vetorial não representa apenas uma evolução técnica, mas uma transformação nos sistemas de recuperação da informação. Diferentemente da busca tradicional baseada em palavras-chave, que depende da correspondência exata entre os termos digitados pelo usuário e os termos presentes nos metadados ou no texto indexado, a busca semântica procura recuperar documentos a partir de sua proximidade de significado. Em outras palavras, o foco deixa de estar restrito à coincidência lexical e passa a estar na relação semântica entre consulta e documento (Weckmüller et al., 2025). Essa mudança é especialmente relevante porque a busca lexical tradicional, ao privilegiar a exatidão formal dos termos, tende a reproduzir e reforçar silos linguísticos e assimetrias epistêmicas já existentes.

Essa abordagem torna-se viável com o avanço dos modelos de linguagem e, em particular, dos embeddings textuais. Embeddings são representações numéricas densas de fragmentos de texto, como títulos, resumos, palavras-chave ou trechos de documentos, geradas por modelos treinados para capturar regularidades semânticas. Ao converter textos em vetores posicionados em um espaço multidimensional, esses modelos permitem calcular similaridades entre unidades textuais mesmo quando elas não compartilham os mesmos termos (Weckmüller et al., 2025).

Assim, duas expressões distintas podem ser reconhecidas como semanticamente próximas, ainda que utilizem vocabulários diferentes, sinônimos, paráfrases ou formulações disciplinares variadas. No contexto multilíngue, esse avanço é ainda mais significativo. Modelos de linguagem multilíngues são capazes de projetar enunciados de diferentes idiomas em um mesmo espaço semântico compartilhado. Isso significa que uma consulta formulada em português pode ser aproximada, em termos de significado, de documentos escritos em espanhol, inglês ou francês, sem depender de tradução prévia manual, tesouros bilíngues exaustivos ou mapeamentos terminológicos rigidamente definidos. Busca-se a correspondência conceitual entre significados expressos em línguas distintas (Litschko et al., 2022).

Em aplicação recente de busca semântica multilíngue baseada em embeddings, Weckmüller et al. (2025) demonstram que fluxos de representação vetorial podem operar com boa responsividade e escalabilidade, inclusive com agregação de embeddings e tempos interativos compatíveis com uso real. Ainda que o estudo esteja situado em dados geo-textuais urbanos, sua contribuição é conceitualmente transferível para repositórios digitais: ele evidencia que a recuperação semântica multilíngue já é tecnicamente viável em ambientes que exigem grande volume de dados, múltiplos idiomas e interfaces exploratórias.

No entanto, essa projeção em um espaço semântico “compartilhado” não elimina as assimetrias estruturais. Modelos treinados predominantemente dominados pelo inglês e por línguas de alta disponibilidade de recursos tendem a apresentar desempenho desigual, muitas vezes comprometendo ou distorcendo as nuances de línguas periféricas e conhecimentos formulados em repertórios não hegemônicos. Dessa forma, o progresso técnico alcançado pode reproduzir ou até aprofundar perdas epistêmicas, caso não seja acompanhada de uma reflexão crítica acerca

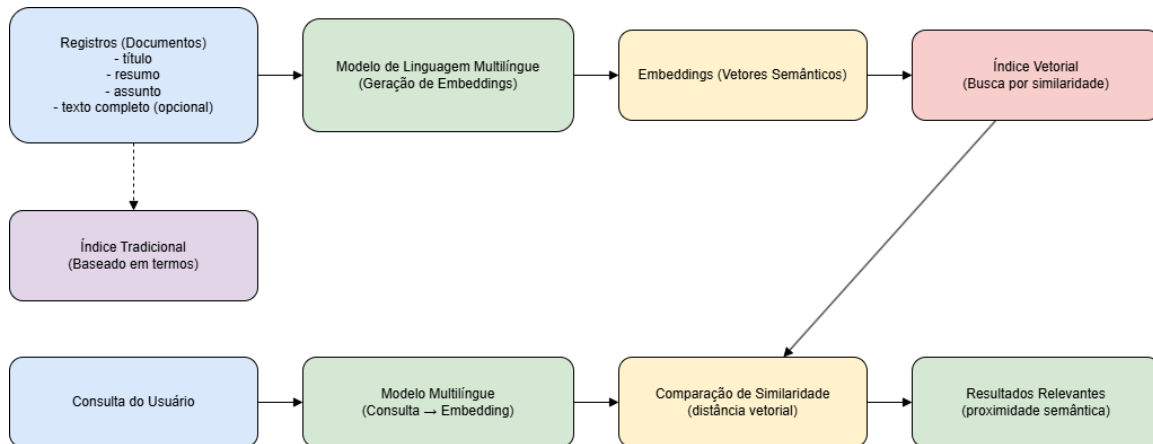
de quem define o espaço semântico comum e sob quais condições diferentes línguas e saberes são representados e reconhecidos.

Essa capacidade altera de forma substantiva as possibilidades de descoberta em repositórios digitais. Em sistemas convencionais, uma busca por correspondência lexical tende a reproduzir silos linguísticos: documentos só são encontrados com facilidade quando compartilham o idioma e, muitas vezes, a mesma formulação terminológica da consulta (Ford & Alemneh, 2024). Essa limitação não é meramente técnica, ela reflete e reforça desigualdades informacionais e epistêmicas mais profundas, tornando grande parte do conhecimento produzido em línguas periféricas ou com repertórios conceituais distintos praticamente inacessível ou invisível para a maioria dos usuários.

A busca semântica multilíngue permite superar parcialmente essa fragmentação, ao conectar conteúdos por afinidade conceitual em vez de mera coincidência lexical. Com isso, o idioma deixa de operar como filtro implícito estruturante de relevância, permitindo que a recuperação da informação se torne potencialmente mais inclusiva, mais sensível à diversidade de expressões do conhecimento. No entanto, essa superação não é automática nem neutra: ela depende de como os espaços semânticos são construídos e de até que ponto os modelos conseguem representar, sem distorcer, saberes formulados fora dos padrões hegemônicos da comunicação científica global. Do ponto de vista técnico, esse tipo de sistema costuma envolver algumas etapas articuladas. Inicialmente, determinados campos dos registros, como título, resumo, assunto ou, em alguns casos, trechos do texto completo, são processados por um modelo de linguagem multilíngue capaz de gerar embeddings semânticos. Esses vetores passam a compor um índice vetorial, distinto, mas complementar ao índice tradicional baseado em termos. Ao realizar uma busca, a consulta do usuário é convertida em embedding e comparada com os vetores armazenados, identificando registros semanticamente próximos independentemente do idioma ou da formulação exata (ver Figura 1) (Jiang, et al., 2020; Litschko et al. 2022; Weckmüller et al., 2025).

Figura 1.

Diagrama Explicativo Sobre o Funcionamento da Busca Semântica Versus Busca Tradicional



Nota: Elaboração dos autores (2026).

Entretanto, isso não significa que a busca semântica deva substituir integralmente a busca lexical. Em repositórios digitais, a solução mais promissora é a combinação de ambas, uma busca híbrida. A busca por palavras-chave continua sendo relevante para consultas precisas, nomes próprios, expressões controladas, siglas, identificadores e filtros específicos. Já a busca semântica amplia a capacidade de descoberta ao recuperar documentos conceitualmente relacionados, inclusive em outros idiomas. Essa articulação permite equilibrar precisão e abrangência, mas não resolve por si só as desigualdades estruturais presentes nos sistemas de recuperação.

A adoção de modelos multilíngues traz desafios significativos. O desempenho semântico desigual entre idiomas, especialmente no caso de línguas com menor presença em grandes corpora de treinamento. Além disso, a geração, armazenamento e atualização de embeddings demandam infraestrutura computacional elevada, critérios claros de curadoria e processos contínuos de avaliação. Modelos de linguagem podem reproduzir assimetrias, regionais e epistêmicas presentes em seus dados de treinamento, o que reforça, em vez de mitigar, injustiças informacionais. Por essa razão, a implementação da busca semântica multilíngue não deve ser tratada como solução neutra ou automática. Trata-se de uma escolha infraestrutural que deve ser orientada por princípios de Justiça Informacional: transparência, avaliação contínua e responsabilidade pública diante das assimetrias de poder na comunicação científica.

Em síntese, infere-se que a busca semântica multilíngue em repositórios digitais representa a convergência entre inovação tecnológica e compromisso ético-político. Esta tecnologia permite operacionalizar a diversidade linguística de maneiras tangíveis: ao reduzir barreiras de acesso, aumentar a visibilidade da ciência periférica e fortalecer a mediação informacional entre demandas sociais e produção científica global. Como reforçam Garcês-da-Silva et al. (2021), a

Justiça Informacional exige acesso equitativo, possibilita a participação justa nos processos de produção do conhecimento e assegura o reconhecimento de diferentes sujeitos e saberes. Assim, implementar sistemas de busca semântica multilíngue constitui uma ação política concreta que materializa esses princípios ou reproduz exclusões históricas, caso não seja acompanhada de crítica e governança adequada.

Nesse sentido, a relevância dessa tecnologia ultrapassa a esfera puramente operacional. Ao permitir que um usuário pesquise em sua própria língua e encontrem resultados relevantes produzidos em outros contextos linguísticos, ela amplia a circulação do conhecimento, reduz assimetrias de visibilidade e fortalece condições mais equitativas de participação na ciência. Contudo, seu verdadeiro potencial reside em contribuir para ecologia informacional mais plural e equitativa, desde que explicitamente orientada pelos eixos distributivo, participativo e de reconhecimento da Justiça Informacional.

Resultados

Tendo como base nos desafios da invisibilidade linguística e nos fundamentos da Justiça Informacional discutidos anteriormente, propõe-se um modelo de busca semântica multilíngue estruturado em camadas, concebido para operacionalizar esses princípios no contexto de repositórios digitais. A implementação dessa abordagem pode ser concebida como uma camada adicional de mediação informacional, integrada às infraestruturas já existentes de coleta, indexação e disseminação de metadados científicos. A estruturação do modelo em camadas evidencia que a busca semântica multilíngue opera como uma infraestrutura sociotécnica orientada à promoção da Justiça Informacional. Cada camada incide sobre dimensões específicas de desigualdade: a ingestão e a representação atuam no reconhecimento da diversidade linguística; a recuperação híbrida amplia as condições de acesso e uso da informação; a mediação favorece a participação de diferentes comunidades linguísticas; e a governança assegura a incorporação de princípios éticos e de equidade no funcionamento do sistema.

Em termos conceituais, o modelo é organizado em cinco camadas interdependentes que transformam a diversidade linguística em um recurso de equidade informacional. O fluxo inicia-se com a ingestão e normalização, que prepara os registros por meio da coleta de metadados e limpeza textual para garantir uma base consistente, seguida pela indexação semântica, onde modelos de linguagem geram embeddings para permitir a proximidade conceitual entre idiomas distintos. A eficácia do sistema é alcançada na recuperação híbrida, que combina a precisão da busca lexical com a abrangência da descoberta vetorial, enquanto a camada de mediação e apresentação torna os resultados inteligíveis ao usuário através de filtros e justificativas de correspondência. Por fim, a governança e avaliação assegura que a infraestrutura permaneça transparente e alinhada aos princípios da Justiça Informacional, monitorando vieses e

promovendo a participação da comunidade (Baglioni et al., 2019; Dony et al., 2024; Garcês-da-Silva et al., 2021; Litschko et al., 2022; Mathiesen, 2015; Sena, 2023; Weckmüller et al., 2025).

Dessa forma, o modelo oferece uma via tecnicamente viável e politicamente significativa para fortalecer a justiça informacional em repositórios digitais, ao transformar a diversidade linguística de obstáculo em condição de ampliação do diálogo científico.

Considerações Finais

A revisão narrativa mostrou que a diversidade linguística está no centro do debate contemporâneo sobre Justiça Informacional e comunicação científica. Também demonstra que já existem fundamentos técnicos consistentes para conectar documentos por significado entre línguas distintas, sobretudo com o uso de embeddings multilíngues e recuperação translinguística. A busca semântica multilíngue não deve ser entendida apenas como um aperfeiçoamento funcional dos sistemas de busca, mas como uma condição estratégica para enfrentar desigualdades de visibilidade e acesso no ecossistema da comunicação científica. Se a justiça informacional pressupõe que diferentes sujeitos e comunidades possam acessar e se beneficiar do conhecimento em condições mais equitativas, então a capacidade de buscar para além das fronteiras linguísticas torna-se parte constitutiva desse horizonte. Nesse sentido, incorporar modelos de recuperação semântica multilíngue em repositórios digitais significa avançar na construção de infraestruturas abertas não apenas no acesso, mas também na possibilidade efetiva de encontro entre saberes. Em última instância, repositórios digitais só cumprem plenamente sua função pública quando o conhecimento que abrigam pode também ser encontrado, reconhecido e apropriado por comunidades diversas. A busca semântica multilíngue constitui, nesse sentido, uma via promissora para aproximar a infraestrutura da ciência aberta dos princípios da justiça informacional, ao transformar a diversidade linguística de obstáculo em condição de ampliação do diálogo científico.

Conflito de Interesses

Os autores declaram não haver conflitos de interesses.

CRedit – Contribuições dos Autores

Washington Luís Ribeiro de Carvalho Segundo | Conceitualização, Escrita – redação original, Supervisão, Recolha de dados

Lautaro Matas | Conceitualização, Escrita – redação original, Supervisão, Investigação

Luciana Mara Silva | Conceitualização, Escrita – redação original, Supervisão, Recolha de dados
Priscila Machado Borges Sena | Conceitualização, Escrita – redação original, Supervisão, Recolha de dados

Dayane Dornelles | Conceitualização, Escrita – redação original, Supervisão, Recolha de dados

Referências

- Baglioni, M., Bardi, A., Kokogiannaki, A., Manghi, P., Iatropoulou, K., Principe, P., Vieira, A., Nielsen, L. H., Dimitropoulos, H., Foufoulas, I., Manola, N., Atzori, C., La Bruzzo, S., Lazzeri, E., Artini, M., De Bonis, M., & Dell'Amico, A. (2019). The OpenAIRE research community dashboard: On blending scientific workflows and scientific publishing. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital Libraries for Open Knowledge* (pp. 56–69). Springer. https://doi.org/10.1007/978-3-030-30760-8_5
- Dony, C., Kuchma, I., & Ševkušić, M. (2024). *Dealing with multilingualism and non-English content in open repositories: Challenges and perspectives*. Zenodo. <https://doi.org/10.5281/zenodo.11060284>
- Ford, A. Y., & Alemneh, D. G. (2024). Inclusive global scholarly communication: Toward a just and healthier information ecosystem. *Journal of the Association for Information Science and Technology*, 75(10), 1058–1069. <https://doi.org/10.1002/asi.24899>
- Garcês-da-Silva, F. C., Mello, F. R. de, Lima, J. C. de, & Vechi, F. R. (2021). Justiça para quem? Justiça social, informacional, racial e de gênero em bibliotecas. In *Anais do 21º Encontro Nacional de Pesquisa em Ciência da Informação* (pp. 1–16). ANCIB; IBICT. <https://www.researchgate.net/publication/358128586>
- Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., & Zhao, L. (2020). Cross-lingual information retrieval with BERT. In *Proceedings of the Workshop on Cross-Language Search and Summarization of Text and Speech* (pp. 26–31). European Language Resources Association. <https://aclanthology.org/2020.clssts-1.5/>
- Le Coadic, Y.-F. (2004). *A ciência da informação* (2nd ed.). Briquet de Lemos.
- Litschko, R., Vulić, I., Ponzetto, S. P., & Glavaš, G. (2022). On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2), 149–183. <https://doi.org/10.1007/s10791-022-09406-x>
- Márquez, M. C., & Porras, A. M. (2020). Science communication in multiple languages is critical to its effectiveness. *Frontiers in Communication*, 5. <https://doi.org/10.3389/fcomm.2020.00031>
- Mathiesen, K. (2015). Informational justice: A conceptual framework for social justice in library and information services. *Library Trends*, 64(2), 198–225. <https://doi.org/10.1353/lib.2015.0044>
- Sena, P. M. B. (2023). Justiça informacional em ciência, tecnologia e inovação no Brasil: Reflexões e ações necessárias em ciência da informação. *Encontros Bibli*, 28(Dossier Especial), e93046. <https://doi.org/10.5007/1518-2924.2023.e93046>

Sena, P. M. B., Bonetti, L. G., & Carvalho Segundo, W. L. R. de. (2026). Repositórios institucionais como tecnologia convergente de saberes institucionais. In L. R. Silveira, S. O. Mendes, & C. Pecegueiro (Eds.), *Convergências para o acesso aberto: Políticas, tecnologias e práticas em transformação* (pp. 72–91). EDUFMA.
<https://doi.org/10.5281/zenodo.18944390>

Vučković, A., & Sikimić, V. (2023). How to fight linguistic injustice in science: Equity measures and mitigating agents. *Social Epistemology*, 37(1), 80–96.
<https://doi.org/10.1080/02691728.2022.2109531>

Weckmüller, D., Dunkel, A., & Burghardt, D. (2025). Embedding-based multilingual semantic search for geo-textual data in urban studies. *Journal of Geovisualization and Spatial Analysis*, 9, Article 31. <https://doi.org/10.1007/s41651-025-00232-5>