Enhancing Visibility Across Languages: Semantic Multilingual Search for Scholarly Content

by Lautaro Matas on behalf of COAR

Version 1.0 - November 6, 2025

Acknowledgements

First and foremost, I would like to thank **Kathleen Shearer** for her vision and confidence in exploring this path. I am also deeply grateful to the colleagues who generously contributed their time and insights to review this document and improve it with their valuable comments and suggestions: **Washington Segundo, Luciana Mara Silva, Carlos Nusch, Martin Klein,** and **Paul Walk**.

Abstract

Scholarly knowledge is created and shared through a wide range of sources — repositories, journals, data platforms, and other scholarly information systems — and in hundreds of languages. Yet, most discovery tools continue to privilege a few dominant languages, leaving large portions of research effectively invisible to global audiences. This paper explores the concept of *semantic multilingual search*: an emerging approach that retrieves information by meaning rather than by exact wording, enabling users to search in any supported language and discover relevant work across all languages.

Instead of proposing a fixed technical design, the document invites the community to consider how semantic multilingual search could evolve within the broader ecosystem of scholarly communication. It reflects on early experiences, shared principles, and collective responsibilities to ensure that this new generation of discovery tools advances openness, equity, and linguistic diversity.

Developed within COAR's broader vision of openness, multilingualism, and bibliodiversity, this paper aims to stimulate dialogue and advance collaboration toward discovery systems that reflect the full linguistic and cultural diversity of global scholarship. It contains a "Call to the Community" for feedback about this approach so that, **when you ask a question in your language, the entire world should have a chance to understand and respond.**

Confederation of Open Access Repositories

Table of Contents

Introduction

Scholarly knowledge is produced in hundreds of languages across every region of the world. Yet, the tools we use to discover that knowledge often act as if only a few of those languages truly matter. Researchers, librarians, policymakers, and the public face an invisible barrier: a vast amount of relevant work remains hidden simply because it is written in another language.

This document explores a new approach — **semantic multilingual search** — that makes it possible to search in any supported language and retrieve results from all languages by meaning, not by exact wording. It outlines the limitations of current search methods, presents a conceptual model for a meaning-based, language-inclusive discovery layer, and describes practical implementation paths that can work with existing scholarly information sources and aggregator infrastructures.

This approach is not about translating content, but about using new machine learning technologies to enable users to search seamlessly across multiple languages. The goal is simple yet ambitious: **to make language no longer a barrier to discovering and using scholarly knowledge.**

This work builds on the vision promoted by the Confederation of Open Access Repositories (COAR) and other community-led initiatives: a scholarly communication system that is open, inclusive, multilingual, and interoperable. By aligning with COAR's strategic vision, semantic multilingual search can be implemented as part of a broader, coordinated effort across repository networks worldwide.

1. A world of knowledge divided by language

The global research landscape is vast, diverse, and inherently multilingual. Although the language of international publishing has traditionally been English, many many scholars across continents publish work in their own languages — Spanish in Argentina, Portuguese in Brazil, Arabic in Egypt, Japanese in Japan, Swahili in Kenya, and hundreds of others. This linguistic diversity is not accidental; it reflects the deep connection between knowledge creation and the communities that sustain it. Research in local or regional languages ensures that findings are accessible to those most directly affected — policymakers, practitioners, educators, and citizens in the same linguistic and cultural context.

But while the creation of knowledge is multilingual, the systems we use to find that knowledge often are not. The global infrastructure of scholarly discovery — journals, repositories, academic search engines, bibliographic indexes, and portals — operates under the promise of inclusivity, but in practice, it privileges a narrow set of dominant languages. Chief among them is English, whose dominance is reinforced not only by its large volume of publications but also by the way search algorithms, ranking systems, and interface designs are built.

The result is an imbalance with real consequences. An ecologist in Mexico searching for studies on *cambio climático may* never encounter critical datasets published in Mandarin or French. A public health worker in Senegal might miss Portuguese-language studies from Brazil that address the very health challenges they are confronting. A policymaker in Vietnam could remain entirely unaware of English-language research that directly answers their pressing policy question. This is not a problem of research quality — it is a problem of visibility and accessibility.

At the heart of this problem lies a reliance on keyword-based search, a method that looks for exact word matches in metadata or text. This works well when the query and the indexed content are in the same language, but it fails in a multilingual environment. Without translations, multilingual thesauri, or complex cross-language mapping, a search for *renewable energy* will never find *energía renovable*, *énergies renouvelables*, or 再生可能エネルギー — even though they mean exactly the same thing.

This linguistic fragmentation creates silos of knowledge. It traps research in the language in which it was published, accessible only to those who can search using the right terms in that same language. The result is a persistent visibility gap that disadvantages researchers publishing in less dominant languages and limits the perspectives and evidence available to the global community.

This is not just a technical flaw; it is a structural inequity. By reinforcing language barriers, our current discovery systems inadvertently concentrate visibility, influence, and impact in a

small subset of the scholarly world. The diversity of global scholarship — the richness of perspectives, methodologies, and cultural contexts — is flattened into a narrower, less representative record. If we are serious about making knowledge a truly global public good, we must first confront this hidden, yet deeply consequential, divide.

2. Limitations of current search approaches

Most discovery systems in journals, repositories, scholarly indexes, and search portals still rely on a foundation that has changed little in decades: keyword-based retrieval. In its simplest form, this approach works by matching the exact words in a user's query with those stored in the system's index, sometimes aided by light linguistic processing such as lemmatization or stemming. While these techniques can improve matching within a single language, they remain limited when facing the complexity of multilingual scholarship.

The problem is not that keyword search is inherently flawed — it is that it was designed for a world where the query and the content are assumed to share the same language. In a global research environment, this assumption does not hold true. Today, the scholarly record is fragmented across hundreds of languages and writing systems, each with its own terminology, morphology, and syntax. Keyword search is simply not equipped to bridge those divides.

Several factors contribute to this issue

Exact word bias – Keyword matching is literal. It cannot infer that "sustainable power generation" is related to "renewable energy" unless both phrases are explicitly present in the metadata or fulltext (depending on the system). This rigidity means that relevant research articulated differently — whether through synonyms, regional expressions, or discipline-specific jargon — is invisible.

Language silos – In most systems, queries only retrieve results in the same language as the search terms. A Spanish-language search for *cambio climático* will not surface documents indexed only in English as *climate change*, despite the identical meaning. This creates parallel, disconnected pools of knowledge for each language.

Dominance of certain languages – Even when multilingual metadata exists, ranking algorithms often favor content in dominant languages — most frequently English — because of how term frequencies, stemming, and scoring models interact. The effect is that less widely used languages, even when highly relevant, are pushed lower in the results and risk being overlooked entirely.

Challenges with non-Latin scripts – Languages such as Chinese, Japanese, Arabic, or Russian introduce additional complexity. They may require specialised tokenisation,

stemming, or normalisation to be searchable at all. Without proper processing, titles and abstracts can be indexed incorrectly, rendering them effectively invisible unless a user knows the precise full string to search for.

Mixed-language records – Many scholarly records contain more than one language — for example, an English abstract paired with a Spanish title and body text. Standard keyword search may handle only one component effectively, leaving other equally important parts invisible to queries.

These limitations are not hypothetical. They manifest in measurable ways: search results that consistently overrepresent one language, query terms that fail to retrieve semantically relevant content in other languages, and discovery experiences that are incomplete for the user.

The consequence is a **visibility gap** in the global research landscape. This gap does not reflect the relevance or quality of the work itself — only the inability of current search mechanisms to connect meaning across languages. Without a change in approach, these structural shortcomings will continue to constrain how knowledge flows between communities, disciplines, and regions.

3. A different approach — searching by meaning, not words

If the limitations of current approaches come from their dependence on exact words, the solution lies in shifting the focus from words to meaning. This principle is already well established in **semantic search within a single language**, where machine learning techniques such as embeddings allow systems to retrieve documents that express similar ideas even when the exact words differ.

Semantic multilingual search builds on this foundation. Instead of only asking, "Do these words appear in this document in the same language?", the system extends the question to: "Does this document express the same idea as the query, regardless of language or phrasing?" In this context, when we say that the system "asks" a question, it is only a metaphor: there is no conversational interaction taking place. What it really means is that, instead of matching keywords, the system computes whether a document represents the same underlying idea as the query — regardless of language or phrasing.

The core enabler of this shift is a family of machine learning techniques known as **multilingual embeddings**. In simple terms, **embeddings are numerical representations of text that capture its meaning in a way that can be compared mathematically**. When trained on multiple languages, these models learn to map semantically equivalent phrases — whether written in English, Spanish, Arabic, or Japanese — to nearly the same location in a shared "meaning space."

This allows a user's query to be understood independently of the specific language it was entered in. For example, the English phrase *renewable energy*, the French *énergie renouvelable*, and the Spanish *energía renovable* would all end up as vectors positioned very close to each other in this semantic space. A search for any one of them would retrieve documents containing the others, even if no explicit translation or keyword mapping exists.

The implications for discovery are profound

- **True cross-language retrieval** – A researcher can write a query in their own language and find relevant results in many others without ever translating a word themselves.
- **Resilience to phrasing differences** – Semantic retrieval captures meaning even if the query and the document use different terminology or syntax.
- **Independence from pre-built thesauri** – The model learns associations directly from multilingual data, reducing the need for manually maintained cross-language vocabularies.

Importantly, semantic search does not need to replace existing keyword-based systems. Instead, it can operate as a complementary layer. In practice, this means that semantic retrieval can be used alongside traditional keyword search, allowing both exact-term precision and meaning-based discovery to coexist. This ensures continuity with current systems while expanding their reach across languages. The process, in practice, follows a clear sequence:

1. **Representation** – Selected text fields (such as titles, abstracts, subject terms, or portions of full text) *could be processed* through a multilingual transformer model capable of capturing meaning across languages. Each processed segment *might then be transformed* into a fixed-length vector, or *embedding*, representing its semantic content rather than its literal wording.

2. **Storage** – These embeddings *could be stored* in a dedicated vector index designed for semantic retrieval, while the original metadata *would remain* in the traditional keyword index used for filtering and exact-term matching.

3. **Querying** – When a user submits a search query, it *could be normalized* and converted into its own embedding using the same multilingual model, ensuring that both queries and documents are represented within a shared semantic space.

4. **Semantic retrieval** – The system *could perform* a nearest-neighbour search within the vector index to identify documents whose embeddings are most similar in meaning to the query, regardless of language or phrasing.

5. **Hybrid ranking** – Semantic similarity scores *could be combined* with keyword-based relevance signals and other contextual factors to produce a balanced ranking that integrates precision and semantic breadth.

6. **Presentation** – Results *could be displayed* in a familiar interface, optionally enriched with language indicators and on-demand translations of snippets or abstracts to help users quickly assess relevance across languages.

Because OAI-PMH already offers a standard mechanism for data exchange, it *could potentially serve* as the backbone for sharing and reusing semantic representations across scholarly information sources and aggregators. In such a scenario, embeddings generated at one level of the network *might be exposed and harvested* by others, creating opportunities for federated exchange, reduced redundancy, and greater alignment — allowing the semantic layer to *evolve naturally* within the existing OAI-PMH ecosystem.

This approach is inherently open, modular, and interoperable. It can be deployed alongside existing scholarly information discovery systems software like Solr or Elasticsearch, integrated into OAI-PMH harvesting workflows, and scaled from a single institutional source (e.g. repository) to an entire regional or global discovery network.

Most importantly, it tackles the **visibility gap** created by language boundaries. By enabling search systems to recognise and connect meaning across languages, semantic multilingual search transforms the way scholarship is discovered — making it possible for research in any supported language to reach audiences it could never reach before.

4. Proof in practice — what proofs of concept have suggested

Semantic multilingual search is still an emerging approach in scholarly discovery, but initial proofs of concept have started to show how it could work in practice. Over the past several months, we have run small-scale tests and engaged in expert consultations to better understand both the potential for, and the limitations of, meaning-based retrieval in multilingual contexts.

These activities were designed with a **lightweight, exploratory mindset**: not to produce final answers, but to gather experience, learn what works, and identify what needs more refinement before large-scale deployment.

How the proofs of concept were structured

Each experiment combined three main elements:

- **Multilingual sentence embeddings** – Generated for titles, abstracts, using small models trained across multiple languages. (ie:paraphrase-multilingual-mpnet-base-v2)
- **Vector search** – To retrieve documents based on semantic similarity, enabling cross-language matches without relying on exact keyword overlap.
- **Hybrid ranking** – Combining semantic similarity with traditional keyword matching to balance broader recall with precise filtering.

A practical demonstration of these concepts is available in the notebook: "*Semantic Multilingual Search for Scholarly Content"*. It offers a simple, step-by-step exploration of how multilingual embeddings can represent meaning across languages and illustrate, in an accessible way, how semantic similarity works in practice. The notebook can be reviewed as a companion piece to this paper, showing the basic mechanics behind the ideas discussed here. https://tinyurl.com/semanticmultiligualsearch	

Although exploratory in scope, the proofs of concept offered several encouraging signs. Searching across languages felt intuitive, with users able to formulate queries in their own language and still discover relevant materials written in others, after which they could read using their preferred translation method. Content in less dominant languages also gained greater visibility, appearing more prominently than it typically would under traditional keyword-based search. Just as importantly, the semantic layer integrated smoothly with existing metadata harvesting and indexing workflows, showing that meaning-based retrieval can complement current infrastructures without major disruption.

At the same time, the experiments revealed important areas for further investigation and refinement. Performance varied across languages — stronger when training data were abundant, weaker when linguistic resources were scarce. Generating and storing embeddings introduced new computational and storage demands, underscoring the need for scalable approaches. Finally, assessing the quality of cross-language results proved complex, requiring tailored evaluation methods beyond conventional relevance metrics. Together, these insights provided a clearer picture of both the promise and the practical challenges of bringing semantic multilingual search into real-world scholarly environments.

What we learned

Even at this early stage, these proofs of concept confirmed that **meaning-based retrieval can uncover connections that keyword search alone would miss**. They also showed that it is

possible to integrate semantic methods into federated, multilingual scholarly environments without dismantling or significantly changing existing infrastructure.

Still, these are **first steps** and not yet close to being a minimal viable product. The insights gathered now serve as a foundation for refining the models, improving efficiency, and designing robust governance for a system that can scale while remaining fair, transparent, and sustainable.

5. Pathways to implementation — exploring practical routes

The question now is not only whether semantic multilingual search could work, but how it might take shape across the diverse landscape of scholarly information sources and aggregators. Because the proofs of concept were intentionally small and exploratory, the next steps should emphasise progressive, low-risk adoption — allowing each organisation to experiment at its own pace and scale up as the benefits become clear.

Rather than converging on a single deployment model, two complementary routes could evolve, each reflecting different balances between decentralisation, coordination, and technical capacity.

1. **The "Embedding Commons" or decentralised model**
   In this approach, participating information sources and networks agree on a shared multilingual embedding model, schema, and exchange format. Each participant could generate its own embeddings locally or rely on those produced by peers within the network. This collective framework would enable the reuse of semantic representations across different layers, supporting both local autonomy and global interoperability.

   Such a model maximises openness and reusability, respects institutional and national sovereignty, and can reduce redundant computation — though it requires strong coordination, shared governance, and version control to ensure consistency across participants.

2. **The "Harvest, Process, and Serve" or more centralised model**
   An alternative path would involve a consortium or service provider that handles harvesting, embedding generation, and the provision of semantic search capabilities through APIs or downloadable vector datasets. Institutions could then integrate these resources into their own discovery portals without running the embedding infrastructure themselves.

This lowers the barrier to adoption and is particularly useful where local machine-learning capacity is limited. However, it introduces a dependency on the central operator and demands sustainable resources for compute, storage, and maintenance.

Integration through plugins and add-ons

In either case, semantic multilingual search will only become truly accessible if it can integrate smoothly with the platforms researchers and institutions are already using. Developing plugins and add-ons for widely adopted systems — such as DSpace, Dataverse, or VuFind — will therefore be essential. These extensions could operate in "consume" mode, connecting to hosted embeddings, or in "generate" mode, creating them locally when resources permit. This layer of integration ensures flexibility, sustainability, and user familiarity across a heterogeneous ecosystem.

A phased and inclusive path forward

These routes are not mutually exclusive. A realistic roadmap could begin with more centralised pilot services, gradually evolving toward shared, decentralised frameworks as experience and capacity grow. Supporting both centralised and distributed options will allow institutions of all sizes and capabilities to participate meaningfully.

The overarching goal is to minimise disruption to current systems while maximising opportunities to test, refine, and align this technology within real-world scholarly environments — so that semantic multilingual search can mature from a promising idea into a trusted, sustainable part of the global discovery ecosystem.

6. Guardrails for success — fairness, transparency, and sustainability

Deploying semantic multilingual search is not only a technical challenge — it is also a matter of fairness, transparency, and long-term sustainability. A system designed to connect knowledge across languages must do so responsibly, ensuring that results truly reflect relevance and accuracy rather than reproducing hidden biases.

Fair representation, not artificial balance

The goal of semantic multilingual search is not to make all languages appear equally in every results list, but to ensure that language is *not* a barrier to relevance. A well-designed system should be able to find the most meaningful content for a user's query — regardless of the language in which the query or the documents were written.
To achieve this, governance processes should focus on:

- Monitoring how different languages are represented in search results to detect systematic biases.
- Adjusting models and parameters when technical imbalances appear, without forcing artificial equality.
- Including multilingual and cross-regional perspectives in evaluation and decision-making.

Transparency and accountability

Semantic retrieval should not operate as a "black box." The models used, their training data, and their known limitations should be openly documented and shared in accessible terms. When possible, open-source models and openly licensed datasets should be preferred, enabling community review, reproducibility, and collective improvement.

Ethical and sustainable use of AI

Language models inevitably reflect the biases of their training data. Continuous evaluation is therefore essential to identify gaps across languages or disciplines and to correct them when needed. Embedding and processing of texts must also comply with privacy laws and institutional policies.

Sustainability requires regular model updates, re-indexing of collections, and stable resources — human, computational, and financial — so that institutions can maintain and evolve the service over time.

Interoperability and integration

For maximum impact, the semantic layer must remain interoperable with existing infrastructures and standards such as OAI-PMH, OpenSearch APIs, and shared metadata profiles. This ensures that improvements in discovery flow throughout the broader scholarly communication ecosystem, instead of remaining confined to isolated platforms.

Evaluating what truly matters

Evaluation should go beyond technical precision and recall. It should capture how well the system helps users discover relevant content across languages and how it expands visibility for diverse scholarly communities. Useful indicators include:

- **Cross-language coverage:** how often a query in one language retrieves meaningful documents in others.
- **Equity of visibility:** whether non-dominant languages appear proportionally when they are relevant, without being artificially boosted.

- **User experience:** whether researchers, students, or policymakers feel they are finding more diverse and useful results.
- **Institutional value:** evidence that multilingual discovery increases the reach and visibility of local research outputs.
- **Operational sustainability:** stability, response times, and efficiency of the system under realistic conditions.

Why this matters

By combining fairness, transparency, and robust evaluation, semantic multilingual search can evolve into a trusted infrastructure that represents knowledge as it exists — diverse, multilingual, and interconnected. The objective is not to engineer equality by language, but to enable *equity in access*: a system where the best results emerge from meaning, not from the accident of shared vocabulary.

7. The vision ahead — toward a truly multilingual scholarly commons

The promise of semantic multilingual search is not just faster results or more sophisticated technology. It is the possibility of a discovery layer that reflects the true diversity of global scholarship — a system in which the language of a query is no longer a gatekeeper to the world's knowledge.

Imagine an agronomist in Peru searching in Spanish and discovering studies from Japan, France, and Ethiopia that share the same findings. Or a policymaker in Ghana retrieving Portuguese research from Brazil and English reports from Canada with equal ease. In this vision, relevance is defined by meaning, not by the accident of shared vocabulary.

From concept to shared infrastructure

Achieving this vision requires more than a single piece of software. It demands a shared approach across institutions, networks, and regions — one that respects local autonomy while enabling interoperability at a global scale. The models and indexes must be openly documented, the governance transparent, and the deployment flexible enough to serve both a single scholarly information source and an entire federation.

A phased journey

This does not necessarily require a "big bang" change. It can start small — with a proof of concept in a single network, a limited set of languages, or just one field like abstracts. From

there, the system can evolve, expanding its coverage and improving its models as more institutions join.

8. A call to the community

The technical foundation is ready. The concepts have been tested. What remains is a collective choice:

- To invest in infrastructure that values every language.
- To adopt standards and processes that make semantic discovery an open, shared resource.
- To embed governance, ethics, and sustainability into the system from the outset.

If we take this path, the result will be more than better search — it will be a transformation of how scholarship is found, shared, and valued. It will mean that the scholarly record we hand to the next generation is broader, fairer, and closer to the full richness of human knowledge.

COAR, an organisation with a large and extensive international representation and a strong interest in promoting multilingualism and bibliodiversity, is uniquely positioned to advance such an approach at the international level.

To that end, we are actively seeking input from the COAR community and beyond about this conceptual model for improving multilingual search and the feasibility of the approach outlined here. Based on this feedback, COAR will consider the possibility of launching a more comprehensive project to implement and evaluate these technologies in the context of our current repository ecosystem.

In the end, the goal is simple: **When you ask a question in your language, the entire world should have a chance to understand and respond.**