Relatório final de atividades
Projeto: SAVITS: Sistema de Avaliação e Validação de Impactos em Tecnologias Sociais

Coordenador do Projeto:	**Washington Luís Ribeiro de Carvalho Segundo**		
Nome do(a) pesquisador(a):	**Jesiel Viana da Silva**		
Vigência da Bolsa de Pesquisa:	**01/04/2026**	**Até:**	**30/09/2026**
Encerramento em:	**30/09/2026**		
Ref. Fundação:	**32725**	**Ref. Ibict:**	**008/2026-81**

Descrição das Atividades da Bolsa de Pesquisa:

1- Investigar e desenvolver abordagens baseadas em LLMs para a análise semântica de textos qualitativos, integradas a dashboards acessíveis e inclusivos, em conformidade com as diretrizes WCAG- Meta 2-Desenvolvimento do SAVITS 2- Relatório das ativiades desenvolvidas.

1 INTRODUÇÃO

A recuperação da informação em repositórios digitais vem incorporando técnicas de representação semântica capazes de complementar os mecanismos léxicos tradicionais. Em abordagens baseadas em *embeddings*, consultas e documentos são representados por vetores densos, o que permite estimar a proximidade de significado mesmo quando os textos não compartilham exatamente os mesmos termos (SHARMA, 2025). No Apache Solr, esse paradigma é suportado por campos vetoriais densos e mecanismos de busca por vizinhos mais próximos, constituindo uma base tecnológica para aplicações de recuperação neural e semântica (APACHE SOFTWARE FOUNDATION, 2026).

No âmbito deste projeto, a atividade principal deste relatório consistiu em investigar e desenvolver abordagens baseadas em modelos de linguagem para a análise semântica de textos qualitativos. Como ambiente de experimentação e desenvolvimento, utilizou-se o DSpace, plataforma de código aberto amplamente empregada na implementação de repositórios digitais e já utilizada em diferentes iniciativas institucionais.

O trabalho concentrou-se na implementação, em caráter beta, de modalidades de busca semântica e híbrida no DSpace, utilizando *embeddings* armazenados e consultados no Apache Solr 10.1. A investigação abrangeu a geração de representações vetoriais, a recuperação multivetor, o processamento de textos extensos por *chunking*, a fusão de rankings e a integração das novas modalidades à interface de pesquisa.

Para a geração de *embeddings*, foi adotado o modelo BGE-M3. O modelo foi proposto como uma arquitetura multilíngue, multifuncional e multigranular, com suporte a mais de 100 idiomas e a textos de diferentes extensões (CHEN et al., 2024). A escolha também se apoiou em avaliação técnica realizada no contexto da LA Referencia, na qual o BGE-M3 apresentou relação adequada entre recuperação multilíngue, execução local, dimensão vetorial, tamanho do modelo e integração com Solr, tendo sido selecionado como modelo de referência para a primeira etapa de experimentação (LA REFERENCIA IOI, 2026).

2 MÉTODO

A pesquisa teve caráter aplicado e experimental, orientada à implementação e avaliação de recursos tecnológicos para recuperação semântica da informação. As atividades foram conduzidas de forma incremental, com configuração do ambiente, geração e indexação de *embeddings*, desenvolvimento das modalidades de busca, processamento de textos extensos e adaptação da interface do DSpace.

2.1 Implementação da busca semântica

Inicialmente, foi configurado no DSpace um fluxo para transformar consultas e metadados textuais em *embeddings*. Os vetores gerados são persistidos no Solr 10.1 e utilizados para recuperar itens com base na similaridade vetorial. A documentação do Solr caracteriza a representação vetorial densa como uma forma de condensar significado semântico aproximado em um número fixo de dimensões e disponibiliza o tipo DenseVectorField para indexação e busca desses vetores (APACHE SOFTWARE FOUNDATION, 2026).

A implementação permite a configuração de diferentes campos de metadados para vetorização. Foi adotada também uma estratégia multivetor, na qual um mesmo item pode possuir representações vetoriais distintas, por exemplo, para título, resumo e outros metadados selecionados. Essa separação possibilita explorar sinais semânticos específicos de cada campo e ajustar a recuperação de acordo com as características do acervo.

2.2 Processamento de textos extensos por *chunking*

Para metadados textuais extensos, foi implementado um processo de *chunking*. O conteúdo é normalizado e dividido em segmentos de tamanho configurável; cada segmento é vetorizado separadamente, mantendo o título do item como informação de contexto. Essa estratégia reduz o risco de que informações relevantes presentes apenas em uma parte de um resumo ou de uma descrição extensa sejam diluídas em uma única representação global. A abordagem é compatível com a característica multigranular do BGE-M3, que foi desenvolvido para representar desde sentenças curtas até documentos longos (CHEN et al., 2024).

2.3 Busca híbrida e fusão de rankings

Além da modalidade exclusivamente semântica, foi implementada a busca híbrida, combinando a recuperação textual tradicional do DSpace/Solr com a recuperação vetorial. Os dois rankings são fundidos por Reciprocal Rank Fusion (RRF), técnica que combina listas ordenadas a partir das posições ocupadas pelos documentos, reduzindo a dependência da escala absoluta dos escores de cada mecanismo. O Solr documenta a aplicação de RRF para a composição de conjuntos de resultados, permitindo integrar sinais de recuperação distintos (APACHE SOFTWARE FOUNDATION, 2026).

2.4 Integração com a interface do DSpace

Na interface de pesquisa foram incorporadas três modalidades selecionáveis: léxica, semântica e híbrida. Também foram adicionados elementos para identificar explicitamente o modo de busca em uso e apresentar a pontuação de relevância associada aos resultados. As modalidades semântica e híbrida podem ser habilitadas de maneira independente por configuração, permanecendo identificadas como funcionalidades beta durante o período de avaliação.

2.5 Recuperação semântica multilíngue

A adoção do BGE-M3 possibilita a recuperação semântica multilíngue, pois consultas e documentos em idiomas diferentes são projetados em um espaço vetorial compartilhado. Assim, uma consulta formulada, por exemplo, em inglês pode recuperar documentos semanticamente equivalentes em português, mesmo sem tradução explícita da consulta ou coincidência direta dos termos. Essa característica é particularmente relevante para repositórios que agregam produção científica e técnica multilíngue (CHEN et al., 2024; LA REFERENCIA IOI, 2026).

3 RESULTADOS

Como principal resultado tecnológico, foi obtida uma versão beta do DSpace com suporte integrado a três estratégias de recuperação: busca léxica, busca semântica e busca híbrida. A implementação demonstra a viabilidade de incorporar técnicas de recuperação

baseada em *embeddings* ao fluxo de pesquisa de um repositório digital sem eliminar os mecanismos tradicionais de correspondência textual.

3.1 Busca semântica baseada em *embeddings*

A modalidade semântica passou a recuperar documentos a partir da proximidade entre o vetor da consulta e os vetores associados aos itens. Na prática, essa abordagem amplia a possibilidade de localizar conteúdos conceitualmente relacionados mesmo quando o vocabulário utilizado na consulta difere daquele presente nos metadados. O Solr 10.1 atua como mecanismo de armazenamento e recuperação vetorial, enquanto o BGE-M3 é empregado na geração das representações semânticas.

A escolha do BGE-M3 mostrou-se adequada ao caráter multilíngue e institucional da pesquisa. O estudo de seleção utilizado como referência registra o modelo com 1024 dimensões e 568 milhões de parâmetros, destacando seu equilíbrio entre a qualidade de recuperação multilíngue, o custo de execução local e a integração com a infraestrutura de busca (LA REFERENCIA IOI, 2026). Esses fatores são relevantes para soluções institucionais que precisam manter controle sobre a infraestrutura e processar acervos heterogêneos.

3.2 Recuperação multivetor e *chunking*

A indexação multivetor possibilitou representar separadamente partes semanticamente relevantes de um mesmo registro. Em conjunto com o *chunking*, essa abordagem permite que segmentos específicos de textos extensos contribuam para a recuperação do item. O título preservado como contexto em cada segmento auxilia a manter a relação entre o trecho vetorizado e o objeto informacional ao qual pertence.

Esse resultado é especialmente relevante para a análise semântica de textos qualitativos, pois descrições, resumos e outros campos narrativos podem concentrar conceitos importantes em passagens localizadas. A segmentação configurável cria uma base técnica para futuras avaliações comparativas de tamanho de chunk, sobreposição entre segmentos, campos utilizados e estratégias de agregação dos escores.

3.3 Busca híbrida

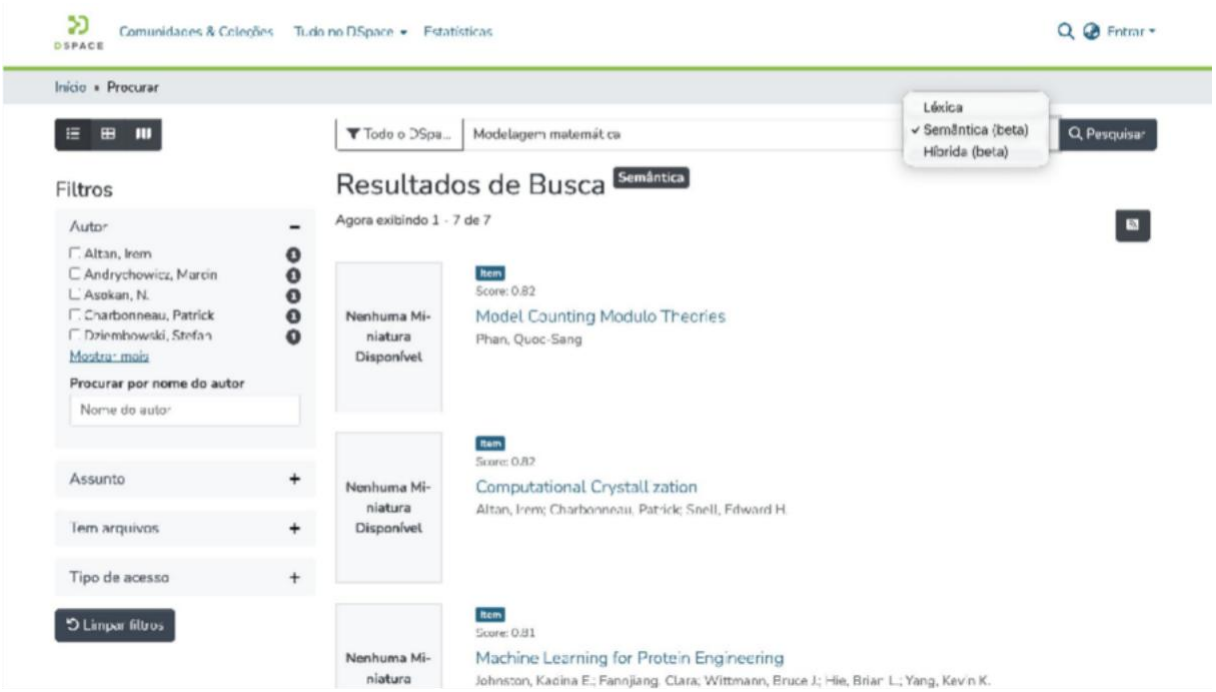
A modalidade híbrida integrou o comportamento consolidado da busca léxica com a capacidade de generalização semântica da busca vetorial. A utilização de RRF permitiu

combinar os rankings sem exigir que os escores produzidos pelos dois métodos possuíssem a mesma escala. Com isso, itens com correspondência direta de termos podem permanecer bem posicionados, ao mesmo tempo em que documentos semanticamente relacionados passam a competir nas primeiras posições.

3.4 Interface e transparência do modo de busca

A interface foi adaptada para tornar explícita a estratégia de recuperação selecionada pelo usuário. As opções léxica, semântica e híbrida foram incorporadas ao portal, acompanhadas da identificação do modo utilizado e da exibição da pontuação de relevância. A ativação independente das modalidades semântica e híbrida por configuração permite a implantação gradual, a experimentação controlada e a desativação sem impacto sobre a busca léxica convencional.

Figura 1 - Interface de busca do DSpace com as opções Léxica, Semântica e Híbrida.



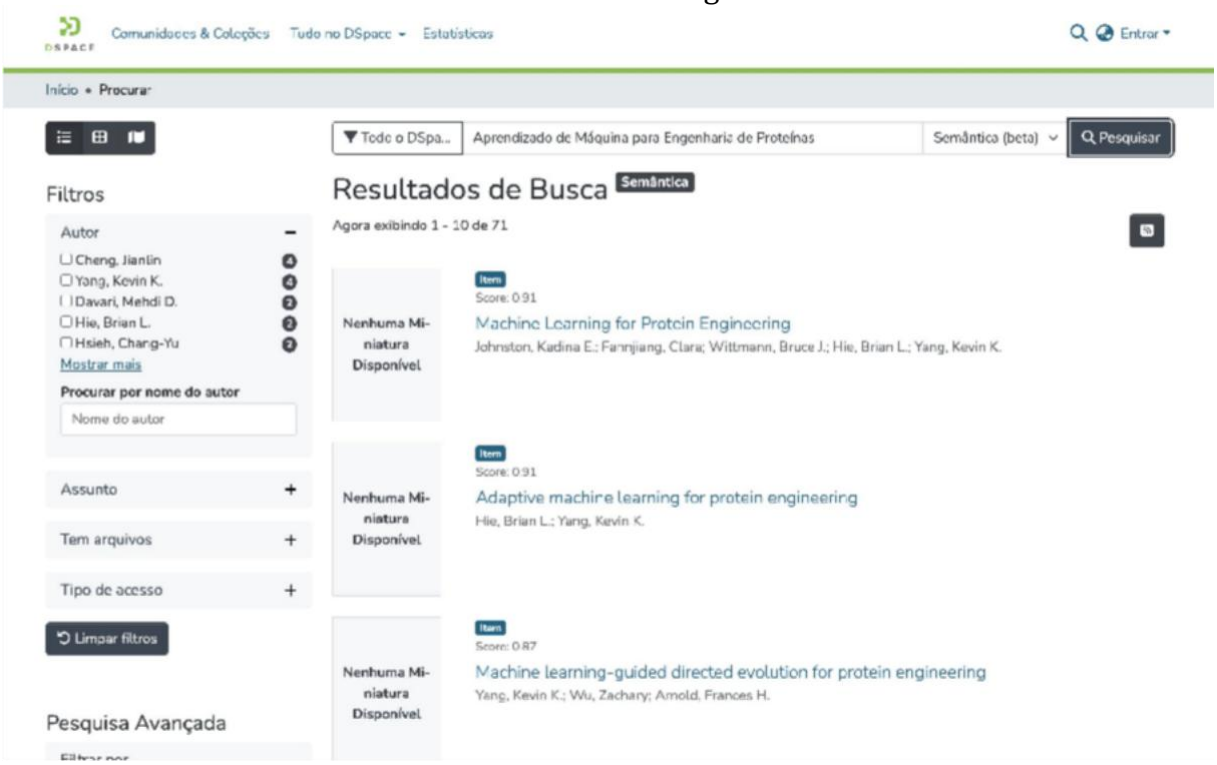
Fonte: Captura de tela do autor (2026).

A Figura 1 apresenta a interface de pesquisa desenvolvida, destacando as modalidades de busca disponíveis ao usuário. As opções semântica e híbrida permanecem identificadas como funcionalidades beta e podem ser habilitadas de forma independente por configuração.

3.5 Recuperação multilíngue

Os testes funcionais demonstraram a possibilidade de realizar recuperação entre idiomas nas modalidades semântica e híbrida. Como o BGE-M3 produz representações multilíngues em um espaço vetorial comum, uma consulta em inglês pode recuperar documentos equivalentes ou semanticamente relacionados em português. Na busca híbrida, esse sinal semântico é combinado ao ranking léxico por RRF, preservando simultaneamente a contribuição da correspondência textual e da proximidade semântica.

Figura 2 - Resultados de busca multilíngue: consulta em português com recuperação de documentos em inglês.



Fonte: Captura de tela do autor (2026).

A Figura 2 deverá registrar um exemplo de consulta formulada em inglês e a recuperação de resultados relevantes em português, evidenciando experimentalmente a capacidade multilíngue da representação semântica utilizada.

3.6 Considerações finais

As atividades realizadas produziram uma base experimental para análise semântica de textos qualitativos e recuperação de informação baseada em embeddings,

integrada ao DSpace e ao Solr 10.1. O resultado reúne busca semântica, recuperação multivetor, chunking, busca híbrida com RRF e recuperação multilíngue, permitindo ampliar a descoberta de conteúdos para além da correspondência lexical tradicional.

A classificação das funcionalidades semântica e híbrida como beta é coerente com o estágio atual de pesquisa e desenvolvimento, permitindo que métricas de recuperação, desempenho e comportamento em diferentes idiomas sejam acompanhadas antes de uma adoção definitiva. Como continuidade, recomenda-se ampliar as avaliações com conjuntos de consultas representativos e métricas consolidadas de recuperação da informação.

REFERÊNCIAS

APACHE SOFTWARE FOUNDATION. Dense Vector Search. Apache Solr Reference Guide. 2026. Disponível em: https://solr.apache.org/guide/solr/latest/query-guide/dense-vector-search.html. Acesso em: 9 set. 2026.

APACHE SOFTWARE FOUNDATION. Stream Decorator Reference: Reciprocal Rank Fusion (RRF). Apache Solr Reference Guide. 2026. Disponível em: https://solr.apache.org/guide/solr/latest/query-guide/stream-decorator-reference.html. Acesso em: 9 set. 2026.

CHEN, Jianlv et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. 2024. arXiv:2402.03216. Disponível em: https://arxiv.org/abs/2402.03216. Acesso em: 9 set. 2026.

LA REFERENCIA IOI. Seleção de um modelo para busca semântica multilíngue. Infografia A4, Componente A, jan.-jun. 2026. Disponível em: https://la-referencia-ioi.github.io/reports/productos/a4-seleccion-modelos/pt/index.html. Acesso em: 9 set. 2026.

SHARMA, Puneet et al. A comprehensive survey of text-based semantic similarity with potential applications. International Journal of Intelligent Information and Database Systems, v. 17, n. 2, p. 143-185, 2025.

gov.br
Documento assinado digitalmente
JESIEL VIANA DA SILVA
Data: 06/10/2026 15:11:28-0300
Verifique em https://validar.iti.gov.br

Jesiel Viana da Silva
Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 08/10/2026 06:03:23-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo
Coordenador do projeto