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):	**Kleber Alves dos Santos**		
Vigência da Bolsa de Pesquisa:	**08/06/2026**	**Até:**	**07/09/2026**
Encerramento em:	**07/09/2026**		
Ref. Fundação:	**32725**	**Ref. Ibict:**	**008/2026-81**

Descrição das Atividades da Bolsa de Pesquisa:

Estudar, prototipar experimentalmente e avaliar arquiteturas, módulos e serviços back-end aplicáveis ao SAVITS, utilizando Java e Spring Boot como tecnologias de referência, considerando critérios de escalabilidade, manutenibilidade, desempenho, rastreabilidade e aderência aos padrões arquiteturais definidos para o projeto; Analisar estratégias de modelagem, especificação e validação experimental de APIs RESTful em Java para integração entre os módulos do SAVITS, considerando requisitos de segurança, versionamento, documentação metodológica, interoperabilidade e consistência dos fluxos de dados; Apoiar workflows de interoperabilidade com infraestruturas externas (BrCris, Oasisbr, APIs abertas) e monitorar desempenho operacional dos serviços Java da plataforma; Elaborar relatório técnico final de bolsa, consolidando resultados, atividades desenvolvidas e contribuições metodológicas da pesquisa.

1 INTRODUÇÃO

O Evidencia é um repositório digital orientado à reunião, organização e disseminação de documentos e resultados de avaliações de políticas públicas. Sua função é tornar evidências localizáveis, contextualizadas e reutilizáveis por pesquisadores, gestores e demais interessados na análise de problemas públicos, intervenções e efeitos observados. A qualidade da descoberta é, portanto, parte central da utilidade do repositório: quanto melhor o sistema relaciona uma necessidade de informação ao conteúdo disponível, mais acessível se torna o conhecimento reunido no acervo (EVIDENCIA, 2026).

Nesse contexto, a busca semântica constitui uma capacidade de acesso que complementa a recuperação convencional por palavras. Ela permite aproximar uma pergunta formulada em linguagem natural de documentos cujo vocabulário não reproduz exatamente os termos da consulta. Em vez de depender exclusivamente da coincidência literal, a pesquisa passa a considerar sinais de proximidade de significado, preservando a ligação entre o resultado apresentado, seus metadados e o item de origem do repositório.

1.1 Objetivo e escopo arquitetural

O objetivo da solução planejada é ampliar a capacidade de descoberta sem desorganizar os mecanismos já consolidados de acesso ao acervo. A busca por correspondência de termos permanece relevante para títulos, siglas, nomes próprios e expressões específicas. A camada semântica acrescenta uma alternativa para perguntas temáticas, formulações mais abertas e situações em que o documento usa termos correlatos aos empregados pelo usuário. A arquitetura híbrida organiza esses modos como sinais complementares, e não como alternativas mutuamente excludentes.

O escopo compreende a preparação do conteúdo indexável, a representação vetorial de documentos e consultas, a recuperação de candidatos por similaridade, a

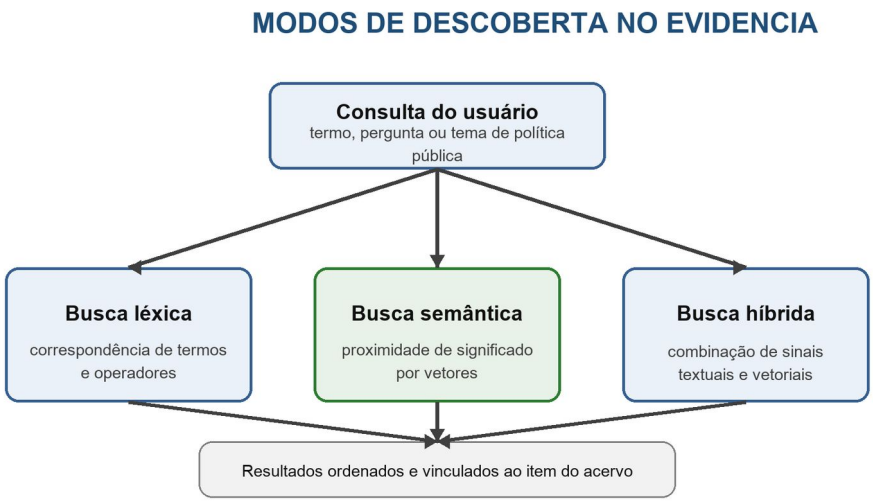
combinação com sinais léxicos e a entrega de resultados vinculados ao item original. Inclui também as condições de plataforma, *schema*, indexação, reindexação, testes e observabilidade necessárias para que o Solr 10.1-beta sustente essa capacidade com previsibilidade.

1.2 Busca semântica, vetorial e híbrida

Na busca semântica, a consulta e os documentos são convertidos em vetores comparáveis. A recuperação seleciona os itens mais próximos do vetor de consulta segundo uma função de similaridade definida para o índice. Esse mecanismo é especialmente útil quando a pergunta contém uma descrição, uma intenção ou uma relação conceitual que pode aparecer no acervo com formulações diferentes. A documentação de referência para busca semântica recomenda a recuperação dos vetores mais próximos e, quando disponível, o tratamento adequado para o caráter distinto de consultas curtas e documentos mais extensos (SENTENCE TRANSFORMERS, 2026).

A busca vetorial é a operação de consulta sobre essas representações numéricas. Já a busca híbrida combina a evidência vetorial com a evidência léxica, permitindo equilibrar a aproximação por significado com a precisão de termos observáveis.

Figura 1 - Modos de descoberta aplicáveis ao Evidencia



Fonte: Elaborado pelo autor (2026)

A escolha da combinação deve ser verificável por meio de conjuntos de consulta e regras de ordenação documentadas. Dessa maneira, resultados semanticamente próximos não ocultam a contribuição de uma expressão exata, e consultas específicas não deixam de aproveitar a ampliação oferecida pela representação vetorial.

2. MÉTODO

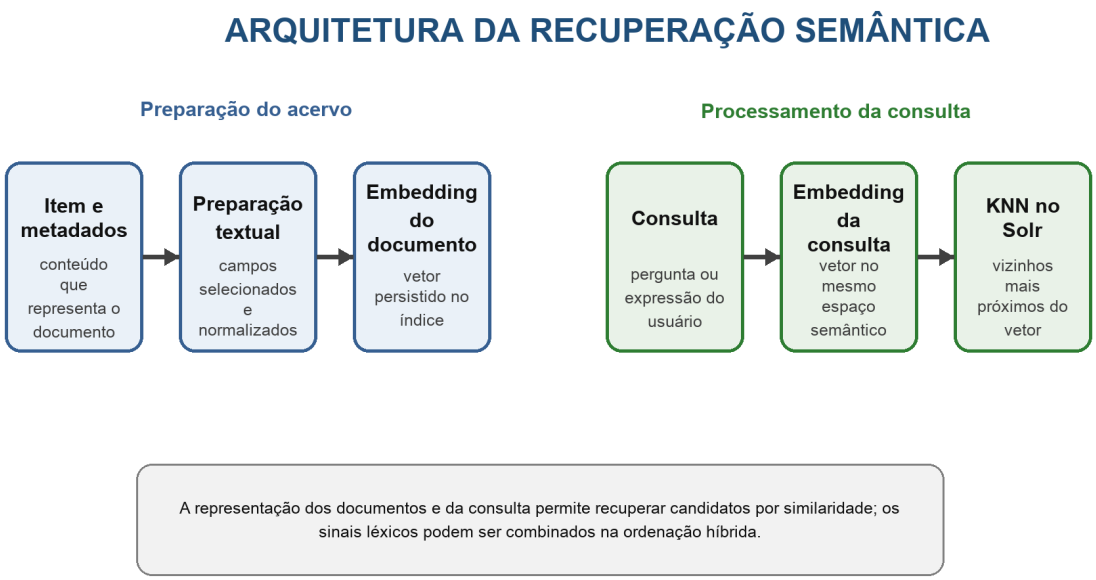
A solução adotou uma perspectiva aplicada, baseada no propósito público do Evidencia, na documentação oficial do Apache Solr e em referências de recuperação semântica. A validação arquitetural foi organizada em quatro dimensões: consistência do conteúdo representado, qualidade dos resultados recuperados, estabilidade da infraestrutura e capacidade de acompanhar o comportamento da solução. A intenção não é declarar que um único modo de busca é superior em toda situação, mas estabelecer critérios para o uso responsável de modalidades complementares.

A arquitetura de busca semântica foi organizada como uma camada de recuperação integrada ao repositório. O acervo permanece a fonte de autoridade dos

documentos e metadados; a camada de indexação prepara representações adequadas à consulta; e o componente de descoberta recebe a pergunta do usuário, identifica os candidatos mais próximos e devolve resultados que apontam para os itens originais. A separação dessas responsabilidades favorece a manutenção do repositório e impede que a representação vetorial seja tratada como substituta do conteúdo curado.

Em termos de fluxo, há dois momentos distintos. No primeiro, o conteúdo do acervo é preparado e transformado em uma representação que pode ser armazenada no índice. No segundo, uma consulta é convertida em representação compatível e utilizada para recuperar os candidatos mais próximos. Esses candidatos podem ser combinados com resultados obtidos por correspondência textual antes da apresentação final. A divisão torna explícito que a geração de vetores de documentos é uma tarefa de indexação, enquanto a geração do vetor da consulta acontece no caminho de atendimento ao usuário.

Figura 2 - Arquitetura da recuperação semântica no Evidencia



Fonte: Elaborado pelo autor (2026)

Os limites entre esses momentos são intencionais. A indexação pode ser executada de forma controlada, com verificação de cobertura e possibilidade de reprocessamento, enquanto a consulta deve responder com previsibilidade ao usuário.

A arquitetura também preserva a independência entre conteúdo, representação e ordenação. O item original não é alterado para atender à busca, e a representação vetorial é tratada como um artefato indexado que pode ser reconstruído.

2.1 Preparação do conteúdo e representação vetorial

A preparação do conteúdo é uma etapa decisiva para a qualidade da recuperação. A arquitetura considera que cada item do acervo pode conter campos com funções distintas: identificação, contextualização, descrição temática e acesso ao documento. Nem todo campo precisa ter o mesmo peso na representação para descoberta. A seleção

deve privilegiar informações que exprimam o tema, o objetivo, o objeto avaliado e os resultados descritos, mantendo os campos de identificação e proveniência disponíveis para exibição e verificação do resultado.

Após a seleção, o texto é normalizado de maneira compatível com o mecanismo de representação adotado. O objetivo não é alterar o documento de origem, mas produzir uma entrada de indexação estável e reproduzível. Essa entrada precisa permitir reindexação controlada quando houver atualização de metadados, correção de conteúdo ou mudança de modelo. A identificação do item acompanha o vetor durante todo o fluxo, garantindo que a proximidade semântica sempre possa ser reconciliada com o registro do repositório.

Os embeddings de documento são calculados antes da consulta e armazenados junto aos demais campos indexados. Essa estratégia desloca a parte mais custosa do processamento para o ciclo de indexação, favorecendo a resposta em tempo de pesquisa.

2.2 Processamento de consultas e recuperação por similaridade

No caminho de consulta, a pergunta do usuário é recebida como linguagem natural e convertida em vetor compatível com o índice. A consulta vetorial recupera um conjunto limitado de candidatos por proximidade, evitando tratar todos os itens do acervo como comparações diretas. A lista resultante é um conjunto de candidatos semanticamente relacionados, que ainda pode receber regras de ordenação, deduplicação e apresentação. A resposta final conserva a referência ao item, permitindo que o usuário examine título, descrição e demais elementos de contexto antes de aprofundar a leitura.

Os resultados podem ser devolvidos com uma quantidade de candidatos adequada à interface e à latência esperada. A definição desse limite é uma decisão de serviço: poucos candidatos podem reduzir a cobertura, enquanto uma lista excessivamente ampla aumenta custo e dificulta a ordenação posterior. O acompanhamento deve relacionar o tamanho do conjunto recuperado ao tempo de

resposta, à qualidade observada e ao comportamento de uso, de forma que ajustes sejam guiados por evidências e não somente por parâmetros abstratos.

2.3 Estratégia híbrida e preservação da busca léxica

A estratégia híbrida é o ponto de encontro entre as modalidades de descoberta. A busca léxica reconhece palavras e expressões presentes nos campos indexados; a busca vetorial reconhece proximidade entre representações de significado. Em vez de eliminar uma dessas fontes, a arquitetura pode calcular candidatos por ambos os caminhos e aplicar uma regra de combinação. A escolha dos pesos, do conjunto de candidatos e das regras de desempate deve ser documentada para que mudanças de comportamento possam ser avaliadas ao longo do tempo.

Essa preservação é especialmente importante em um repositório de avaliações de políticas públicas. Nomes de programas, acrônimos, marcos normativos, localidades e identificadores institucionais exigem a precisão que a busca textual oferece.

Quadro 1 - Uso complementar dos sinais de busca

Situação de consulta	Sinal prioritário	Tratamento arquitetural
Sigla, nome próprio ou expressão normativa	Léxico	Preservar correspondência exata e destacar registros diretamente relacionados.
Pergunta temática em linguagem natural	Semântico	Recuperar candidatos por proximidade de significado e validar a aderência temática.
Consulta com termo específico e contexto amplo	Híbrido	Conciliar correspondência textual com candidatos semânticos, segundo regra documentada.

Situação de consulta	Sinal prioritário	Tratamento arquitetural
Resultado de alta ambiguidade	Revisão de sinais	Ajustar combinação, conjunto de candidatos ou representação a partir de consultas de referência.

Fonte: Elaborado pelo autor (2026)

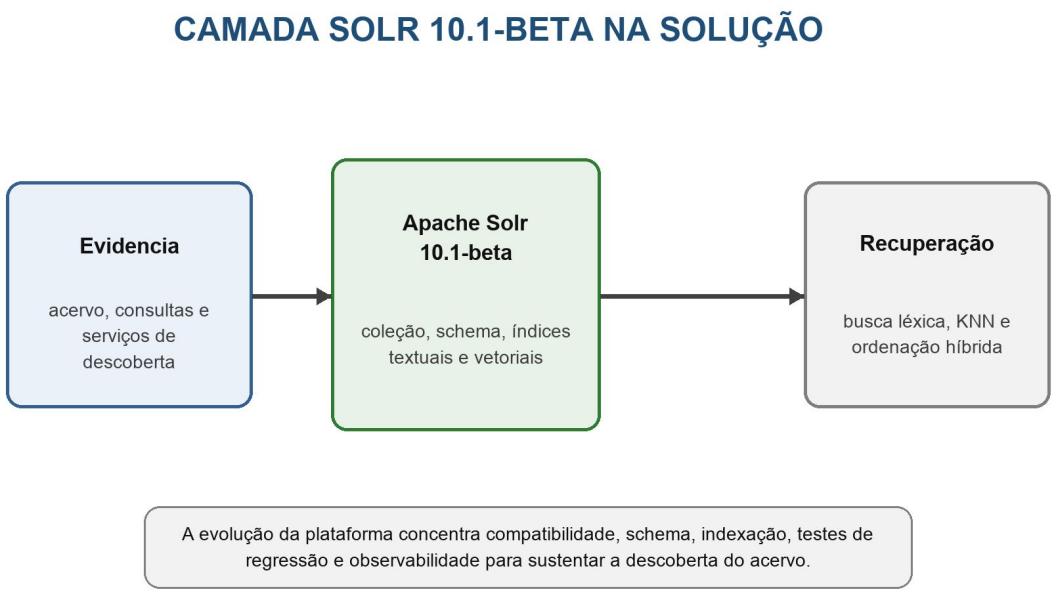
Ao mesmo tempo, perguntas como “efeitos de uma intervenção sobre determinado público” podem se beneficiar de documentos que empregam terminologia correlata. A arquitetura híbrida acomoda esses dois movimentos: especificidade quando o usuário a fornece e ampliação semântica quando ela é necessária para explorar o acervo.

2.4 Evolução da infraestrutura SOLR para 10.1-BETA

O Apache Solr é a camada de recuperação que organiza os campos indexados e atende às consultas necessárias à descoberta no Evidencia. A evolução para a versão 10.1-beta foi tratada como qualificação da plataforma para suportar, de maneira coerente, a busca textual já estabelecida e a camada vetorial incorporada à arquitetura. A atualização não é apenas uma troca de versão: ela envolve compatibilidade de ambiente, revisão de configurações, comportamento de indexação, consultas de verificação e acompanhamento de riscos próprios de uma versão beta.

A plataforma deve, portanto, ser observada como uma fundação compartilhada pelas modalidades de descoberta. A evolução precisa assegurar que a busca por termos, a recuperação por similaridade e a combinação híbrida usem coleções e campos coerentes, sem comprometer a disponibilidade do serviço. Essa visão evita tratar a capacidade vetorial como um componente isolado: seu comportamento depende de decisões de *schema*, do estado da indexação, da capacidade do ambiente e das regras que convertem a consulta em candidatos recuperáveis.

Figura 3 - Camada Solr 10.1-beta na arquitetura da solução



Fonte: Elaborado pelo autor (2026)

O uso de uma versão beta reforça a necessidade de um ciclo de aceitação explícito. A documentação técnica da atualização, a validação em cenário controlado e o acompanhamento inicial de operação formam evidências para decidir sobre a continuidade da adoção. Essa abordagem não antecipa um resultado automático; ela cria condições para evoluir a infraestrutura com transparência, medição e possibilidade de resposta rápida diante de qualquer variação relevante.

2.5 Compatibilidade de ambiente e estratégia de atualização

A compatibilidade de ambiente deve ser examinada antes de qualquer conclusão sobre a qualidade da busca. O servidor Solr, a versão de Java, os serviços de apoio, o armazenamento do índice e os recursos de rede formam uma mesma camada operacional. Uma mudança isolada pode iniciar corretamente e ainda assim produzir efeitos indesejados em indexação, disponibilidade ou tempo de resposta. A estratégia de atualização, portanto, estabelece uma sequência controlada: preparar ambiente

compatível, validar inicialização, aplicar configurações de coleção, executar indexação de teste, comparar consultas e então avançar para a operação acompanhada.

A compatibilidade não é avaliada apenas pela ausência de falha durante a inicialização. Também precisam ser observados o carregamento da coleção, a leitura dos campos, a capacidade de receber documentos, a persistência do índice, a execução das consultas habituais e a resposta sob carga compatível com o uso previsto. Esses aspectos formam a base para que a camada vetorial possa ser adicionada sem degradar os serviços de descoberta já oferecidos pelo repositório.

3 RESULTADOS

A aplicação da busca semântica no Evidencia amplia a descoberta de avaliações de políticas públicas ao aproximar perguntas em linguagem natural dos documentos organizados no repositório. A arquitetura mantém o acervo como fonte de referência, representa documentos e consultas de modo comparável, recupera candidatos por similaridade e preserva os sinais textuais necessários a pesquisas exatas. O resultado é uma camada de acesso complementar, capaz de lidar com variações de vocabulário sem remover a rastreabilidade nem a necessidade de leitura crítica dos itens recuperados.

A evolução do Apache Solr para a versão 10.1-beta constitui o suporte de infraestrutura para essa capacidade. A atualização envolve mais que compatibilidade de versão: requer ambiente adequado, *schema* coerente com campos textuais e vetoriais, indexação controlada, testes de regressão, observabilidade e possibilidade de retorno seguro. A busca vetorial com *DenseVectorField* e a recuperação KNN apoiada por *HNSW* oferecem os mecanismos necessários para aproximar desempenho e qualidade, desde que sejam avaliados por dados e consultas representativos (APACHE SOLR, 2026a; 2026b).

Como continuidade, recomenda-se manter um conjunto vivo de consultas de referência, registrar as decisões que afetam a representação e a ordenação, acompanhar os indicadores de serviço e revisar periodicamente a cobertura do acervo. A combinação

entre governança de conteúdo, arquitetura de busca e disciplina operacional permitirá que o Evidencia evolua sua capacidade de descoberta de forma verificável, responsiva e alinhada ao propósito público do repositório.

3.1 Resultados da aplicação da busca semântica

O principal resultado da atividade foi a incorporação de uma camada de descoberta que trata documentos e consultas como representações comparáveis, sem deslocar o acervo de sua condição de fonte de autoridade. Cada resultado continua vinculado ao item que o originou e aos seus metadados, de modo que a aproximação por significado não substitui a leitura contextualizada do documento. Essa separação preserva a curadoria do repositório e oferece ao usuário uma nova via para encontrar avaliações descritas por vocabulários distintos. A estratégia é compatível com a recuperação semântica baseada em embeddings, na qual corpus e consulta ocupam um mesmo espaço vetorial e os itens mais próximos são recuperados por similaridade (SENTENCE TRANSFORMERS, 2026; REIMERS; GUREVYCH, 2019).

A atividade também estabeleceu uma distinção operacional entre a preparação do acervo e o atendimento da consulta. Na preparação, os conteúdos selecionados para descoberta são normalizados e representados de forma estável, permitindo reindexação quando houver atualização de metadados ou mudança na estratégia de representação. No atendimento, a pergunta do usuário é convertida para uma representação compatível e utilizada para recuperar candidatos semanticamente relacionados. O resultado dessa divisão é a previsibilidade do fluxo: o processamento mais custoso permanece associado ao ciclo de indexação, enquanto a pesquisa concentra-se na recuperação e na apresentação de candidatos vinculados aos itens originais.

A aplicação não retirou a importância da busca léxica. Nomes de programas, siglas, expressões normativas, localidades e identificadores institucionais continuam exigindo correspondência textual precisa. A camada semântica passou a complementar esse comportamento em consultas temáticas ou formuladas como perguntas, nas quais

uma avaliação relevante pode empregar termos correlatos aos usados pelo pesquisador. A combinação dos dois sinais evita que uma consulta específica perca precisão e, ao mesmo tempo, amplia a possibilidade de descoberta quando a formulação do usuário não coincide literalmente com a redação presente no acervo.

Outro resultado é a definição de critérios verificáveis para a evolução da qualidade da recuperação. O conjunto de consultas de referência pode reunir perguntas temáticas, termos institucionais e casos de ambiguidade para observar se os candidatos recuperados permanecem pertinentes e rastreáveis. Essa prática é necessária porque a recuperação aproximada envolve escolhas de representação, quantidade de candidatos e ordenação. A documentação especializada destaca que algoritmos de busca aproximada equilibram cobertura e tempo de resposta; por isso, a avaliação deve confrontar consultas representativas com os resultados efetivamente apresentados, em vez de assumir que a proximidade vetorial, isoladamente, determina a relevância (SENTENCE TRANSFORMERS, 2026).

3.2 Resultados da evolução do Solr para 10.1 beta

A evolução do Apache Solr para a versão 10.1-beta resultou em uma base de infraestrutura preparada para atender, de forma coerente, à busca textual existente e à recuperação vetorial incorporada ao Evidencia. O resultado não se limita à atualização do servidor: inclui a revisão das condições de execução, da configuração da coleção, do schema e da indexação. Essa abordagem torna explícitas as dependências entre ambiente, conteúdo indexado e consulta, reduzindo o risco de que a capacidade vetorial seja tratada como um elemento isolado ou seja introduzida sem critérios de continuidade operacional.

No nível da recuperação, a evolução disponibiliza uma configuração compatível com campos vetoriais densos e consultas de vizinhos mais próximos. A documentação do Solr define o DenseVectorField para indexar e pesquisar vetores numéricos densos, com dimensões e função de similaridade estabelecidas no schema. A consulta KNN

recupera os itens mais próximos do vetor de entrada e permite delimitar o conjunto de candidatos a ser considerado. Dessa forma, a arquitetura passa a ter um mecanismo de recuperação apropriado para o uso de embeddings, sem alterar a referência dos documentos nem a função dos campos textuais utilizados pela pesquisa convencional (APACHE SOLR, 2026a).

A adoção do mecanismo de vizinhos mais próximos também formaliza o equilíbrio entre tempo de resposta e cobertura dos resultados. O índice HNSW organiza os vetores em uma estrutura de grafo que favorece a busca aproximada em espaços de alta dimensionalidade. Essa característica é adequada ao contexto de descoberta, desde que os parâmetros de consulta e a quantidade de candidatos sejam verificados com dados representativos. A própria natureza aproximada do método exige que a avaliação considere a possibilidade de recuperar resultados pertinentes com menor custo de busca, sem transformar uma escolha de desempenho em afirmação automática de qualidade (Malkov; Yashunin, 2020; APACHE SOLR, 2026a).

A atividade de atualização também produziu um referencial operacional para as próximas validações do ambiente. A documentação das mudanças da série Solr 10 indica requisito mínimo de Java 21 para o servidor e registra alterações que afetam a telemetria, a configuração e a compatibilidade com componentes do ecossistema. As notas de atualização orientam a revisar customizações, alterações de versão e o comportamento de consultas antes da adoção continuada. Para o Evidencia, isso se traduz em uma rotina de validação que observa inicialização, carregamento da coleção, recebimento de documentos, persistência do índice e resposta das pesquisas usuais antes de ampliar o uso da nova infraestrutura (APACHE SOLR, 2026b; 2026c).

Como resultado conjunto, as duas atividades deixam uma arquitetura de descoberta mais explícita e passível de acompanhamento. A busca semântica amplia o acesso ao acervo quando há variação de vocabulário, enquanto a evolução do Solr oferece os recursos de indexação e recuperação necessários para sustentar essa capacidade. A continuidade depende de reindexações controladas, consultas de

referência, observação da disponibilidade e revisão periódica dos parâmetros adotados. Esse conjunto de práticas permite aperfeiçoar a solução com base no comportamento observado, mantendo o repositório, seus metadados e a busca textual como elementos permanentes da experiência de pesquisa.

REFERÊNCIAS

APACHE SOLR. Dense Vector Search. [S. l.], 2026a. Disponível em: https://solr.apache.org/guide/solr/latest/query-guide/dense-vector-search.html. Acesso em: 11 set. 2026.

APACHE SOLR. Major Changes in Solr 10. [S. l.], 2026b. Disponível em: https://solr.apache.org/guide/solr/latest/upgrade-notes/major-changes-in-solr-10.html. Acesso em: 11 set. 2026.

APACHE SOLR. Solr Upgrade Notes. [S. l.], 2026c. Disponível em: https://solr.apache.org/guide/solr/latest/upgrade-notes/solr-upgrade-notes.html. Acesso em: 11 set. 2026.

EVIDENCIA. Repositório Digital de Avaliação de Políticas Públicas. [S. l.], 2026. Disponível em: https://devrdapp.ibict.br/home. Acesso em: 11 set. 2026.

MALKOV, Yu. A.; YASHUNIN, D. A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 42, n. 4, p. 824-836, 2020. Disponível em: https://doi.org/10.1109/TPAMI.2018.2889473. Acesso em: 11 set. 2026.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. p. 3982-3992. Disponível em: https://aclanthology.org/D19-1410/. Acesso em: 11 set. 2026.

SENTENCE TRANSFORMERS. Semantic Search. [S. l.], 2026. Disponível em: https://www.sbert.net/examples/sentence_transformer/applications/semantic-search/README.html. Acesso em: 11 set. 2026.

gov.br
Documento assinado digitalmente
KLEBER ALVES DOS SANTOS
Data: 21/10/2026 15:03:22-0300
Verifique em https://validar.iti.gov.br

Kleber Alves dos Santos
Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 23/10/2026 10:05:08-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo
Coordenador do projeto