Relatório final de atividades
Projeto: Repositório Digital de Avaliação de Políticas Públicas (RDAPP)

Coordenador do Projeto:	**Marcel Garcia de Souza**		
Nome do pesquisador:	**Rogério Nogueira de Sousa**		
Vigência da Bolsa de Pesquisa:	**01/04/2026**	**Até:**	**30/09/2026**
Encerramento em:	**30/09/2026**		
Ref. Fundação:	**32643**	**Ref. Ibict:**	**0887/2025-61**

Descrição das Atividades da Bolsa de Pesquisa:

1-Desenvolver pipeline genérico de Open Evaluation para indexação de documentos. Meta 1 2-Implementar módulo especializado de pré-processamento de PDFs e extração estruturada de texto. Meta 2 3-Converter documentos para Markdown, preservando hierarquia. Meta 2 4 Implementar estratégias de chunking inteligente (semântico e hierárquico) otimizadas para recuperação RAG. Meta 2 5-Configurar inferência local rápida com Ollama para deploy da LLM e sumarização. Meta 5 6-Documentar arquitetura, modelos e parâmetros utilizados. meta 4 7-Elaborar relatório técnico consolidado com resultados e entregas realizadas.

1 INTRODUÇÃO

Este relatório apresenta as atividades técnicas realizadas no âmbito do Projeto Repositório Digital de Avaliação de Políticas Públicas (RDAPP), com foco no desenvolvimento de uma arquitetura de ingestão, extração, indexação, recuperação e apresentação de documentos avaliativos depositados em repositório DSpace. A solução implementada parte de um item identificado por UUID no DSpace e executa, sem intervenção manual, a cadeia necessária para transformar documentos PDF em vetores pesquisáveis no Qdrant, preservando rastreabilidade, qualidade dos fragmentos e integração com a interface institucional do repositório.

1.1 Escopo técnico da pesquisa

O escopo contemplou quatro frentes principais: ingestão automatizada de PDFs a partir do DSpace; extração estruturada de texto, seções e tabelas; indexação vetorial híbrida com embeddings densos e esparsos; e disponibilização da busca semântica no frontend DSpace Angular. A arquitetura foi concebida para manter a etapa de enriquecimento por modelo de linguagem como processo opcional e desacoplado, garantindo que a pesquisabilidade dos documentos nunca dependa de uma API externa.

2 MÉTODO

O método adotado organizou o desenvolvimento em uma cadeia obrigatória de três estágios: download, extração e indexação. Uma chamada contendo o UUID de um item do DSpace aciona a recuperação dos PDFs do bundle original, submete os arquivos ao MinerU para extração de texto e estrutura, fragmenta o conteúdo em chunks estruturais com cabeçalho de seção e sobreposição, gera embeddings com o modelo bge-m3 e grava os pontos resultantes no Qdrant. Cada etapa foi implementada como task Celery em fila própria, com acompanhamento do estado do trabalho no Redis.

A separação operacional entre filas gpu e light foi definida para evitar disputa desnecessária de recursos e permitir que apenas uma cópia do modelo utilize a placa gráfica. O serviço mineru-pipeline foi adotado como padrão por entregar a mesma

cobertura funcional com uso substancialmente menor de memória de vídeo, estimado em 6,7 vezes menos VRAM em comparação com a alternativa anterior.

No pré-processamento, foram aplicados filtros de qualidade para descartar ruído de OCR, fragmentos de tabela inadequados e trechos curtos demais antes da geração dos chunks. As tabelas foram convertidas para Markdown e preservadas inteiras, evitando que informações tabulares fossem quebradas em partes sem contexto. O chunking foi orientado por tokens e por estrutura documental, mantendo o vínculo com seções e páginas para futura rastreabilidade.

3 RESULTADOS

Os resultados obtidos demonstram a viabilidade de uma cadeia automatizada de ingestão e recuperação semântica para documentos avaliativos no RDAPP. No fluxo de indexação, um documento de 178 páginas foi processado e convertido em 695 chunks, mantendo cabeçalhos de seção, sobreposição controlada e metadados de rastreabilidade. Os embeddings foram gerados com bge-m3, combinando vetores densos de 1024 dimensões e representação esparsa, o que permitiu estruturar a base para busca híbrida no Qdrant.

Na recuperação, foram definidos cinco endpoints de busca para cobrir três frentes vetoriais independentes — texto dos PDFs, resumos e metadados estruturados — além da busca lexical full-text. As consultas vetoriais combinam resultados densos e esparsos por Reciprocal Rank Fusion (RRF) e passam por reranking com cross-encoder, o que aumenta a precisão da ordenação e permite aplicar cortes por relevância absoluta e relativa.

Os resultados retornam item_uuid, item_handle, seção e página, garantindo rastreabilidade da resposta até o documento oficial no DSpace. Também foram implementados colapso por documento nos metadados, filtros por ano, ODS, tipo, critério e território, além de fallback lexical com FTS5 para cenários em que a busca vetorial não seja suficiente ou não esteja disponível.

As facetas foram escopadas à consulta: as contagens refletem o conjunto de documentos que responde à pergunta e cada faceta ignora o próprio filtro, permitindo

exploração progressiva sem produzir becos sem saída. Essa estratégia melhora a experiência de drill-down e aproxima a busca semântica do comportamento esperado em repositórios institucionais.

3.1 Indexação automatizada

A indexação foi organizada como uma sequência reprodutível e observável. O item do DSpace é identificado por UUID, os PDFs são baixados do bundle original, o conteúdo textual e estrutural é extraído, os trechos passam por saneamento e chunking, e os pontos vetoriais são persistidos no Qdrant com metadados suficientes para auditoria e navegação.

O enriquecimento por LLM foi mantido como etapa posterior e opcional, preservando o desenho best-effort: quando não há chave de API ou quando o provedor externo está indisponível, o pipeline segue normalmente e o documento permanece pesquisável.

3.2 Recuperação híbrida com reranking

A busca combina significado semântico, correspondência lexical e recorte facetado. O uso de embeddings densos e esparsos amplia o alcance da recuperação, enquanto o reranker bge-reranker-v2-m3 recalcula a relevância dos candidatos finais com score normalizado. Essa composição fornece resultados mais consistentes para perguntas abertas, expressões exatas e filtros temáticos.

A literatura técnica e a documentação do Qdrant indicam que consultas híbridas combinam vetores densos e esparsos e podem empregar RRF para fusão de rankings, além de reranking posterior para melhorar a precisão dos resultados, estratégia compatível com a implementação adotada neste projeto.

3.3 Catalogação e metadados candidatos

A etapa de catalogação transforma o Markdown extraído em um objeto estruturado com 31 campos, incluindo título, ano, instituição, tipo de avaliação, período avaliado, programa, metodologia, achados e recomendações. Cada campo recebe uma confiança entre 0 e 1, e documentos com média inferior a 0,6 são sinalizados para

revisão humana. Em validação real, foram extraídos 7 achados e 5 recomendações em 26 segundos.

A solução diferencia explicitamente ODS citados no texto de ODS inferidos pelo tema, separando aquilo que o documento afirma daquilo que o sistema sugere como metadado candidato. Após a extração, os metadados podem ser propagados aos pontos já indexados sem reprocessar embeddings, tornando a catalogação reversível e de baixo custo operacional.

3.4 Integração com a interface DSpace

A interface foi integrada ao DSpace Angular 9 e 10, evitando a criação de um portal paralelo. Na página do item, um painel apresenta ocorrências da consulta dentro do documento, com seção, trecho e botão para abrir o PDF diretamente na página correspondente. Na página de busca, uma chave permite alternar entre a busca rápida nativa e a busca semântica, com persistência do modo selecionado entre sessões.

Os resultados semânticos são renderizados pelo componente nativo de resultado do DSpace, preservando a aparência da busca tradicional. A navegação por página foi resolvida em duas camadas — visualizador e componente angular — com abertura direta do arquivo e fallback para a rota do item. Também foram aplicados elementos de identidade institucional gov.br, estados explícitos de carregamento, vazio e erro, filtros avançados recolhíveis e fork próprio versionado no GitHub.

3.5 Caso de referência: OpenEval.fi

Como referência externa para o RDAPP, foi analisada a plataforma OpenEval.fi, iniciativa pública do Ministry for Foreign Affairs of Finland voltada a tornar pesquisáveis e reutilizáveis evidências contidas em relatórios de avaliação da cooperação e da política de desenvolvimento financiadas pela Finlândia. O caso é relevante porque demonstra, em ambiente governamental real, como relatórios avaliativos extensos podem ser transformados em uma camada de busca, filtros, visualizações e resumos assistidos por IA, mantendo vínculo com os documentos originais e responsabilidade humana sobre o uso dos resultados.

A plataforma finlandesa combina busca por palavras-chave, filtros temáticos, geográficos, temporais e por critérios de avaliação, exportação de resultados, geração de resumos por IA e visualizações de cobertura de evidências, incluindo recortes por ODS/SDG. Seus resultados apresentam título do relatório, tipo de avaliação, ano, trecho textual, página e link para o PDF original, reforçando a rastreabilidade entre a resposta produzida pela interface e a fonte primária consultável.

Do ponto de vista tecnológico, as fontes analisadas descrevem uma arquitetura baseada em serviços Microsoft Azure, incluindo Azure Document Intelligence, Azure AI Search, modelos GPT-4o, processamento de linguagem natural, busca semântica e aprendizado de máquina. Embora a pilha tecnológica seja diferente da adotada no RDAPP, há convergência conceitual: ingestão de PDFs, extração e estruturação de conteúdo, indexação semântica, recuperação com filtros e apresentação de evidências com referência ao documento de origem.

O principal aprendizado para o RDAPP é que a camada de IA deve ser tratada como infraestrutura de governança de evidências, e não apenas como chatbot ou buscador genérico. O diferencial do OpenEval.fi está na combinação entre corpus público, filtros estruturados, visualizações analíticas, exportação, avisos de limitação, vínculos com páginas e PDFs originais e explicitação de que conteúdos gerados por IA devem ser verificados em fontes primárias. Esse desenho reforça decisões já adotadas no RDAPP, como preservar item_uuid, item_handle, seção e página nos resultados e manter o enriquecimento por LLM como etapa candidata e revisável.

Para adaptação brasileira, o caso sugere priorizar um corpus público de baixa sensibilidade, normalização consistente de metadados, taxonomia temática em português, mapeamento de ODS, revisão humana de metadados candidatos e registro das consultas e filtros usados nas análises. A experiência também indica que a utilidade pública da ferramenta depende menos da sofisticação isolada do modelo de linguagem e mais da integração entre documentação oficial, busca rastreável, interface institucional e governança clara do uso das evidências.

REFERÊNCIAS

QDRANT. Hybrid Queries. Disponível em: https://qdrant.tech/documentation/search/hybrid-queries/. Acesso em: 17 ago. 2026.

QDRANT. Hybrid Search with Reranking. Disponível em: https://qdrant.tech/documentation/tutorials-basics/reranking-hybrid-search/. Acesso em: 17 ago. 2026.

YUNIKO SOFTWARE. BGE-M3 Qdrant sample: hybrid search and reranking. Disponível em: https://github.com/yuniko-software/bge-m3-qdrant-sample. Acesso em: 17 ago. 2026.

gov.br
Documento assinado digitalmente
ROGERIO NOGUEIRA DE SOUSA
Data: 16/10/2026 15:34:07-0300
Verifique em https://validar.iti.gov.br

Rogério Nogueira De Sousa
Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
MARCEL GARCIA DE SOUZA
Data: 29/10/2026 09:07:25-0300
Verifique em https://validar.iti.gov.br

Marcel Garcia de Souza
Coordenador do projeto