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):	**Fábio Lorensi do Canto**		
Vigência da Bolsa de Pesquisa:	**20/03/2026**	**Até:**	**20/09/2026**
Encerramento em:	**20/09/2026**		
Ref. Fundação:	**32725**	**Ref. Ibict:**	**008/2026-81**

Descrição das Atividades da Bolsa de Pesquisa:

1- Realizar estudo conceitual e benchmarking sobre repositórios digitais, acessibilidade e IA aplicada; 2- Realizar estudos de usuários com participantes do piloto para adequação dos instrumentos às necessidades diversas. 3- Aplicar testes de acessibilidade em todos os materiais de coleta. 4- Desenvolver, documentar e validar instrumentos padronizados de coleta de dados, com linguagem acessível e sensibilidade às diversidades, a serem aplicados no âmbito do programa piloto por equipes capacitadas; 5- Definir e validar procedimentos padronizados de consentimento livre e esclarecido, contínuo e adaptativo, a serem adotados pelos parceiros executores, em conformidade com os preceitos éticos e normativos aplicáveis; 6- Estabelecer e documentar protocolos padronizados de custódia, anonimização, segurança e governança dos dados, aplicáveis desde o momento da coleta e ao longo de todo o ciclo de processamento e análise. 7- Relatório das atividades desenvolvidas. - Associadas a Meta 1.

1 INTRODUÇÃO

O presente relatório apresenta uma síntese das atividades desenvolvidas no âmbito do projeto SAVITS (Sistema de Avaliação e Validação de Impactos em Tecnologias Sociais), durante o período de vigência da bolsa. As atividades são associadas à Meta 1 do projeto, dedicada aos estudos preliminares, à coleta de dados e à aplicação de protocolos, abrangendo estudos conceituais e de prototipagem envolvendo o uso de IA em repositórios e infraestruturas abertas de informação científica e tecnológica.

Além disso, considerando a alta demanda de trabalho decorrente do planejamento e execução do processo seletivo destinado à composição da equipe multidisciplinar do projeto, parte significativa das atividades desenvolvidas nessa etapa ficaram a cargo desse amplo processo de seleção da equipe de execução do projeto.

A partir disso, foi dado enfoque às atividades de pesquisa teórica e prática sobre o uso da inteligência artificial na recuperação da informação, com base em dados disponíveis em repositórios mantidos pelo Ibict.

Essas atividades resultaram no desenvolvimento de dois protótipos, ambos com amplo potencial de aplicação e aproveitamento no âmbito do projeto SAVITS.

2 MÉTODO

As atividades foram realizadas em articulação com a coordenação e com os demais integrantes do projeto, de acordo com as demandas identificadas no período e com as atividades previstas.

Foram utilizados métodos ágeis de gestão de projetos e de prototipagem de ferramentas de IA, com desenvolvimento incremental e ciclos sucessivos de implementação, teste e ajuste. A prototipagem partiu da identificação de casos de uso relacionados à recuperação da informação e ao aproveitamento de dados e infraestruturas já disponíveis no Ibict, seguida da seleção de modelos e tecnologias adequados a cada aplicação.

Os protótipos foram desenvolvidos em ambiente experimental, priorizando inicialmente a validação das funcionalidades essenciais e da viabilidade técnica das soluções. Foram utilizados modelos abertos de IA e conjuntos de dados institucionais para testar mecanismos de geração de embeddings, recuperação por similaridade semântica e interação com modelos de linguagem. Os resultados obtidos em cada etapa foram analisados qualitativamente e utilizados para orientar ajustes na arquitetura, na interface e nos parâmetros empregados.

Essa abordagem permitiu testar rapidamente diferentes alternativas tecnológicas antes de uma eventual implantação em maior escala, reduzindo custos de desenvolvimento e possibilitando avaliar a adequação das soluções às necessidades do projeto.

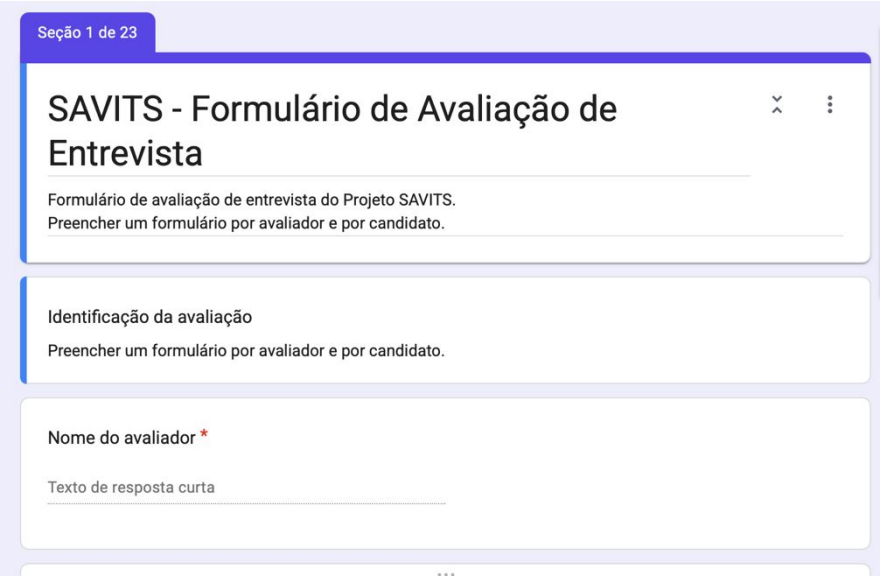
Os protótipos resultantes serviram, assim, como provas de conceito para identificar possibilidades de aplicação futura dessas tecnologias no âmbito do SAVITS.

3 RESULTADOS

3.1 Planejamento e execução do processo seletivo para compor equipe do projeto

A partir do da colaboração no planejamento e na execução do processo seletivo de pesquisadores foi possível compor a equipe multidisciplinar do projeto SAVITS. A atividade foi relacionada às necessidades de estruturação e execução do projeto, uma vez que a formação da equipe técnica e científica constitui condição necessária para o desenvolvimento das diferentes frentes previstas, incluindo ciência de dados, aprendizado de máquina, processamento de linguagem natural, arquitetura e governança de dados, infraestrutura digital e desenvolvimento de sistemas.

Figura 1 - Formulário de avaliação de entrevista



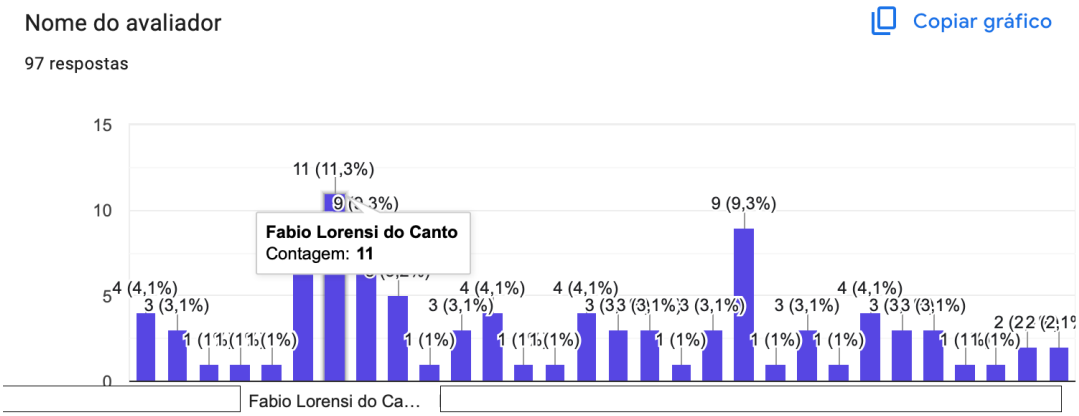
Fonte: Google formulários do projeto (2026).

O processo seletivo recebeu aproximadamente mil inscrições, o que demandou esforço significativo para organização, distribuição e avaliação das candidaturas. Na primeira etapa, foi realizada uma seleção preliminar baseada na análise dos currículos e das respostas fornecidas pelos candidatos no questionário de inscrição. Nessa fase, foram considerados aspectos como formação acadêmica, experiência profissional e aderência às competências requeridas para cada perfil.

No período da bolsa houve participação direta dessa atividade, com avaliação de aproximadamente 89 candidaturas relacionadas aos perfis de especialista em aprendizado de máquina, especialista em processamento de linguagem natural, arquiteto de dados/informação, especialista em infraestruturas digitais, especialista em interfaces construídas com JavaScript e especialista em aplicações PHP.

Houve colaboração no planejamento, na organização e na realização das entrevistas com candidatos pré-selecionados, participando de bancas de avaliação e do registro dos respectivos pareceres. As entrevistas utilizaram formulário estruturado, contemplando experiência, motivação e alinhamento com os objetivos do SAVITS, disponibilidade de carga horária, organização e gestão do tempo e questões específicas relacionadas ao perfil profissional. Os registros produzidos foram posteriormente utilizados na consolidação das avaliações e na definição dos candidatos selecionados.

Figura 2 – Quantidade de entrevistas realizadas



Fonte: Google formulários do projeto (2026).

A participação nessa atividade permitiu ao bolsista contribuir para a estruturação inicial da equipe responsável pelas etapas subsequentes do SAVITS, ampliando sua atuação para além das atividades originalmente estabelecidas na Meta 1.

A composição da equipe multidisciplinar fortaleceu a capacidade operacional do projeto e possibilitou a distribuição das atividades entre pesquisadores com competências especializadas necessárias ao desenvolvimento de seus componentes metodológicos e tecnológicos.

3.1 Estudo sobre aplicação de IA em repositórios digitais

No que diz respeito ao estudo sobre aplicação de funcionalidades baseadas em modelos de IA em repositórios digitais, observa-se uma tendência de incorporação de ferramentas em repositórios, bibliotecas digitais, catálogos e sistemas de recuperação da informação. Inicialmente, essas aplicações estiveram associadas principalmente à classificação automática, extração de entidades e conceitos, enriquecimento de metadados, recomendação de documentos e processamento de linguagem natural.

Os avanços em modelos abertos de embeddings, modelos de linguagem e IA generativa vem permitindo incorporar recursos de busca semântica, nos quais documentos podem ser recuperados pela proximidade de significado com a pergunta do usuário, mesmo quando não existe correspondência exata entre as palavras utilizadas na consulta e aquelas presentes nos documentos (Kasenchak, 2019).

A recuperação semântica representa uma evolução importante em relação aos mecanismos tradicionais baseados predominantemente na correspondência lexical.

Coneglian e Santarem Segundo (2022), por exemplo, aproximam IA, processamento de linguagem natural e tecnologias da Web Semântica para propor modelos nos quais o significado e o contexto dos dados possam ser utilizados no processo de recuperação da informação.

Outra linha de aplicação de crescente importância envolve *Large Language Models* (LLMs) e arquiteturas de *Retrieval-Augmented Generation* (RAG) (Lewis, 2020). Nesses sistemas, os conteúdos armazenados em repositórios ou bases institucionais podem ser segmentados e indexados semanticamente. Assim, diante de uma pergunta do usuário, o sistema primeiro recupera documentos ou trechos considerados relevantes e, posteriormente, fornece essas evidências a um modelo generativo para produzir uma resposta em linguagem natural. Essa arquitetura possibilita desenvolver interfaces conversacionais sobre acervos específicos e, quando implementada com mecanismos adequados de rastreabilidade, permite vincular as respostas aos documentos utilizados como evidência (Lewis, 2020).

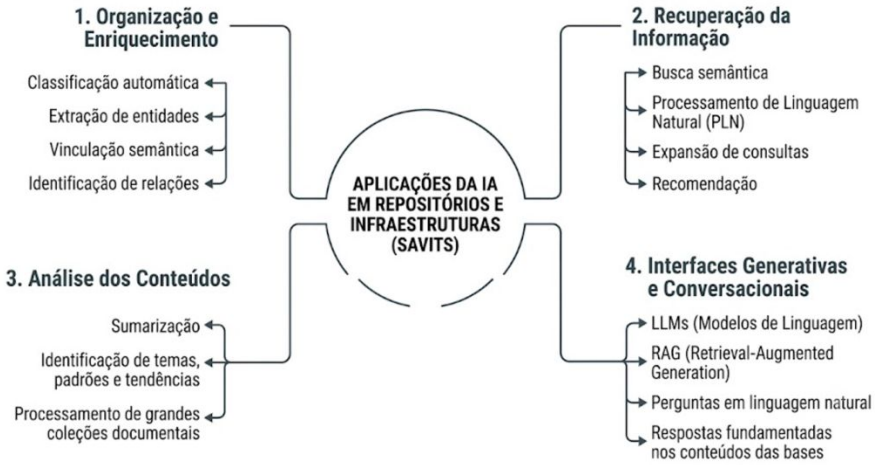
Entre as possibilidades de aplicação de tecnologias de IA no Ibict estão a geração ou sugestão de estratégias de busca, classificação automática de áreas do conhecimento, tratamento de nomes de pessoas e instituições, vinculação de entidades a identificadores persistentes, geração de resumos, tradução, identificação de documentos semelhantes, recomendação de conteúdos e apoio à detecção de duplicidades.

Da perspectiva do projeto SAVITS, podem ser identificados quatro grandes grupos de aplicação da IA em repositórios e infraestruturas de informação:

1. organização e enriquecimento dos dados e metadados, por meio de classificação, extração de entidades, vinculação semântica e identificação de relações;
2. recuperação da informação, incorporando busca semântica, processamento de linguagem natural, expansão de consultas e recomendação;
3. análise dos conteúdos, com sumarização, identificação de temas, padrões e tendências e processamento de grandes coleções documentais, e;
4. interfaces generativas e conversacionais, especialmente por meio de LLMs e RAG, que possibilitam formular perguntas em linguagem natural e obter respostas fundamentadas nos conteúdos das próprias bases.

Essas quatro perspectivas são ilustradas na Figura 3:

Figura 3 – Perspectivas do uso de IA em repositórios e infraestruturas



Fonte: Elaborado pelo autor (2026).

Com base nessas quatro perspectivas, foram realizadas atividades práticas de prototipagem a partir de modelos abertos de IA e de dados disponíveis em repositórios mantidos pelo Ibict.

3.2 Protótipo de semântica de teses e dissertações da BDTD

Como atividade prática da bolsa foi desenvolvido um protótipo de busca semântica aplicado a teses e dissertações, utilizando inicialmente um conjunto de 2.531 documentos da Universidade do Estado de Santa Catarina (UDESC) indexados na Biblioteca Brasileira de Teses e Dissertações (BDTD), mantida pelo Ibict.

A escolha desse corpus permitiu realizar experimentos em uma coleção controlada e com metadados estruturados, avaliando o potencial de técnicas de recuperação baseadas em representações vetoriais para complementar os mecanismos tradicionais de busca por palavras-chave.

O protótipo contemplou a preparação dos metadados textuais das teses e dissertações e a geração de embeddings a partir de campos como título e resumo. Foram testados diferentes modelos de embeddings, entre eles BGE-M3, EmbeddingGemma 300M e Qwen3-Embedding-0.6B, permitindo comparar características como dimensão dos vetores, capacidade de representação semântica, desempenho computacional e qualidade dos resultados recuperados.

A partir dos vetores gerados, foi implementado um mecanismo de recuperação por similaridade semântica, no qual consultas em linguagem natural eram convertidas em embeddings e comparadas com as representações dos documentos.

Além disso, os modelos de embeddings utilizados possuem capacidade multilíngue, permitindo representar semanticamente consultas e documentos em diferentes idiomas em um mesmo espaço vetorial. Dessa forma, uma consulta formulada em português, inglês, espanhol e até chinês pode recuperar teses e dissertações semanticamente relacionadas, mesmo quando o idioma da consulta difere daquele utilizado no documento.

Figura 4 – Interface do protótipo de busca semântica de teses e dissertações



Fonte: Elaborado pelo autor (2026).

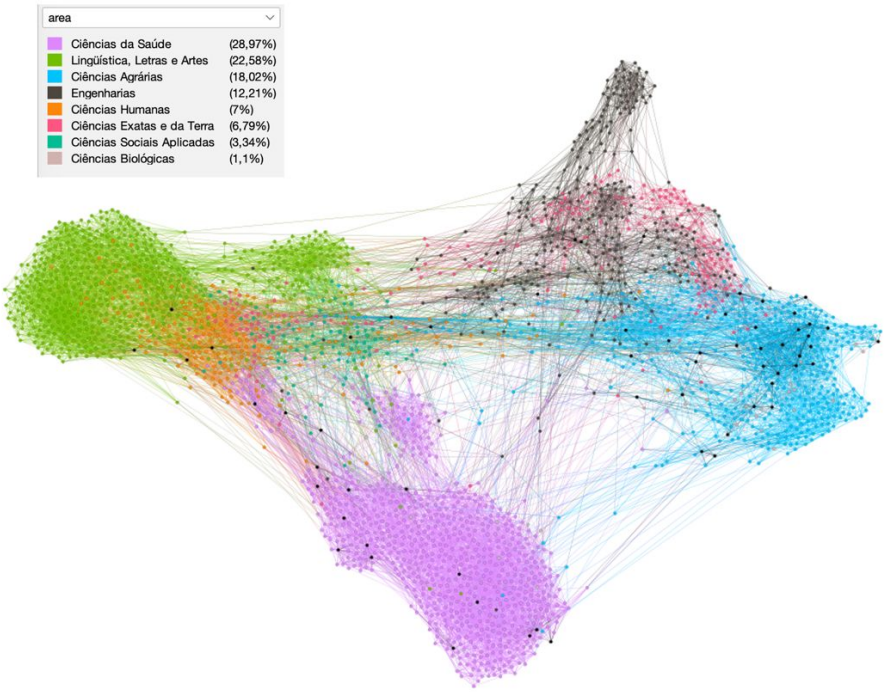
Essa característica amplia as possibilidades de acesso à produção científica indexada na BDTD e favorece sua descoberta por usuários de diferentes comunidades linguísticas. Consequentemente, a recuperação multilíngue pode contribuir para aumentar a visibilidade e a circulação internacional das teses e dissertações brasileiras.

Uma interface demonstrativa permitiu visualizar os documentos recuperados e seus respectivos escores de similaridade, possibilitando a análise qualitativa dos

resultados. A Figura 4 mostra a interface experimental para execução das buscas semânticas e análise dos resultados a partir de uma consulta realizada em idioma chinês, que recupera documentos em português relacionados ao tema pesquisado.

Também foram exploradas formas de representação das relações de similaridade entre documentos por meio de redes e ferramentas de visualização, como o Gephi. A Figura 5 mostra a rede de similaridade vetorial das teses e dissertações da UDESC com agrupamento de cores por grande área do conhecimento.

Figura 5 – Rede de similaridade vetorial das teses e dissertações



Fonte: Elaborado pelo autor (2026).

Observa-se que a similaridade vetorial resultante do modelo agrupa as teses e dissertações na rede de acordo com as respectivas grandes áreas, o que demostra a precisão dos modelos atual de LLM/embeddings para recuperação da informação científica e tecnológica disponíveis em repositórios.

A atividade constituiu uma prova de conceito para a aplicação de IA na recuperação da informação da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). Os experimentos forneceram subsídios para uma arquitetura mais ampla de busca semântica, combinando recuperação lexical, vetorial e híbrida, além de

fundamentarem estudos posteriores envolvendo textos completos, recuperação de trechos e técnicas de Retrieval-Augmented Generation (RAG) aplicadas ao conteúdo científico da BDTD

A partir dos resultados promissores desse protótipo, essa atividade teve dois desdobramentos relevantes:

a) apresentação da palestra “Busca semântica de teses e dissertações”, no encontro da Rede Brasileira de Repositórios Digitais, realizado em Florianópolis, entre 10 e 11 de junho de 2026;

b) em conjunto com outros colaborares do Ibict, submissão do projeto intitulado “BDTD@IA: recuperação semântica e RAG a partir de teses e dissertações brasileiras” para a Chamada Universal CNPq/FNDCT nº 06/2026 do CNPq, solicitando recursos financeiros para aplicação desse modelo em toda coleção da BDTD, e posteriormente, para outras infraestruturas do Ibict, incluindo as decorrentes do projeto SAVITS.

3.2 Protótipo de chatboot com base em LLM institucional do Ibict

Outro resultado de destaque entre as atividades do período da bolsa foi o desenvolvimento de um protótipo de interface para uso do modelo aberto Qwen 3 hospedado no novo cluster como agente de LLM conversacional de propósito geral na rede interna do Ibict (via VPN).

A partir de um primeiro script de chamada remota via linha de comando desenvolvido pelo bolsista Renan Carneiro, foi desenvolvido um protótipo de interface para facilitar o uso por usuários sem conhecimento avançado e para permitir configurações do modelo na tela (temperatura, número de tokens gerados, função raciocínio, etc).

A Figura 6 abaixo mostra e tela inicial do protótipo de chatbot para RAG institucional do Ibict, que foi manteve o nome Contextus, já utilizado em outros modelos de IA.

Figura 6 – Interface do protótipo de chatbot institucional do Ibict



Fonte: Elaborado pelo autor (2026).

Esse protótipo está em fase de homologação e, dependendo dos resultados de testes de viabilidade, poderá ser colocado em produção para uso interno via VPN de todos os colaboradores do Ibict. Também está em análise o desenvolvimento de uma API para uso desse e de outros modelos de abertos de IA, que possibilitará aplicações em tarefas de processamento de dados em larga escala, como as que estão previstas no SAVITS.

REFERÊNCIAS

CONEGLIAN, Caio Saraiva; SANTAREM SEGUNDO, José Eduardo. **Inteligência artificial e ferramentas da web semântica aplicadas à recuperação da informação**: um modelo conceitual com foco na linguagem natural. Informação & Informação, v. 27, n. 1, 2022. DOI: 10.5433/1981-8920.2022v27n1p625.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: **Advances in Neural Information Processing Systems – NeurIPS** 2020. 2020.

KASENCHAK, Robert T. What is Semantic Search? And why is it important? **Information Services & Use**, v. 39, n. 3, 2019. DOI: 10.3233/ISU-190045.

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** (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. p. 3982–3992. DOI: 10.18653/v1/D19-1410

gov.br
Documento assinado digitalmente
FABIO LORENSI DO CANTO
Data: 25/09/2026 05:58:13-0300
Verifique em https://validar.iti.gov.br

Fábio Lorensi do Canto
Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 28/09/2026 17:16:51-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo
Coordenador do projeto