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):	**Rogério Nogueira de Sousa**		
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 métodos de processamento de linguagem natural (NLP/PLN) aplicáveis às funcionalidades analíticas do SAVITS, incluindo classificação textual, análise de sentimentos e extração de informações, considerando critérios de desempenho, consistência semântica, qualidade dos dados e aderência aos objetivos do Caso Piloto; Contribuir para pesquisa, experimentação e fine-tuning de modelos LLM aplicados ao SAVITS, investigando adaptações ao domínio do projeto e alinhamento com os objetivos analíticos; Desenvolver pipelines de pré-processamento textual e integração dos componentes de NLP com as APIs do SAVITS, garantindo escalabilidade, manutenção evolutiva e documentação técnica; Elaborar relatório técnico final de bolsa, consolidando resultados, atividades desenvolvidas e contribuições metodológicas da pesquisa.

1 INTRODUÇÃO

O projeto SAVITS: Sistema de Avaliação e Validação de Impactos em Tecnologias Sociais tem como objetivo desenvolver e integrar ao ecossistema de informação científica do Ibict uma plataforma de inteligência analítica e preditiva para avaliação de impactos em tecnologias sociais. Sua arquitetura prevê módulos de ingestão, processamento e análise de dados, aprendizado de máquina e interpretação de linguagem natural, incluindo o uso de Grandes Modelos de Linguagem (LLMs), além da interoperabilidade com infraestruturas nacionais de Ciência Aberta mantidas pelo Instituto (IBICT, 2026).

No âmbito da bolsa de pesquisa, as atividades previstas concentram-se no estudo, prototipação e avaliação de métodos de Processamento de Linguagem Natural aplicáveis às funcionalidades analíticas do SAVITS, incluindo classificação textual, extração de informações, análise de sentimentos, experimentação com LLMs e desenvolvimento de pipelines de pré-processamento textual integrados às APIs da plataforma.

A partir desse escopo, a pesquisa foi direcionada à investigação de uma camada de integração capaz de acessar dados provenientes das infraestruturas de informação científica do Ibict, realizar seu tratamento e disponibilizá-los de forma adequada aos componentes analíticos de NLP e LLM do SAVITS. Essa abordagem está relacionada diretamente à concepção do projeto, que prevê a integração do SAVITS com infraestruturas como Oasisbr, BDTD e BrCris (IBICT, 2026).

Para esse estudo, foi utilizado o Harvester do Oasisbr como ambiente real de pesquisa e experimentação. O Harvester realiza a coleta de registros provenientes dos repositórios que alimentam o Oasisbr e disponibiliza informações relacionadas às fontes, às coletas, aos registros e aos resultados das regras de validação (IBICT, 2026). Dessa forma, oferece um conjunto de dados reais e heterogêneos adequado para investigar problemas relacionados à integração, preparação, normalização e qualidade de metadados antes de sua utilização por mecanismos de análise automatizada.

O trabalho buscou compreender como esses dados poderiam ser acessados, tratados e organizados por uma camada intermediária antes de serem consumidos pelos componentes analíticos do SAVITS. A arquitetura estudada preserva o Harvester como fonte dos dados de coleta, evitando sua replicação desnecessária, enquanto a camada de

integração realiza o acesso, tratamento e disponibilização das informações requeridas pelas aplicações consumidoras. Essa estratégia foi aplicada sobre dados reais de coletas, diagnósticos e registros do Oasisbr.

No contexto da preparação dos dados para integração com componentes de NLP e LLMs do SAVITS, uma etapa importante da investigação envolveu o pré-processamento dos metadados textuais, com procedimentos de normalização de títulos e nomes de instituições e agrupamento de valores associados às regras de validação. Esse tratamento permitiu identificar padrões recorrentes e investigar requisitos de preparação e organização dos dados que deverão ser considerados na integração de componentes baseados em linguagem natural ao SAVITS.

Buscando materializar e validar parte da camada de integração proposta, foi desenvolvida uma Interface do Gestor, utilizada como artefato experimental da pesquisa. A interface consome os dados disponibilizados pelo Harvester e permite acompanhar coletas, consultar diagnósticos de qualidade, identificar registros com problemas e analisar valores associados às regras de validação. Sua implementação possibilitou avaliar, em um ambiente real, mecanismos de acesso aos dados, controle de permissões, tratamento de respostas, cache e navegação entre diagnósticos e registros.

Nesse contexto, o objetivo da pesquisa desenvolvida no período foi investigar uma camada de integração para acesso, tratamento e preparação de dados provenientes das infraestruturas de informação científica do Ibict, utilizando o Harvester do Oasisbr como ambiente experimental e considerando sua posterior utilização pelos componentes analíticos previstos no SAVITS.

1.1 Escopo técnico da pesquisa

O escopo da pesquisa concentrou-se na investigação de mecanismos de integração entre dados provenientes das infraestruturas de informação científica do Ibict e os componentes analíticos previstos no SAVITS. Para essa finalidade, o Harvester do Oasisbr foi utilizado como ambiente experimental, permitindo trabalhar com dados reais de repositórios, coletas, diagnósticos e registros.

A pesquisa foi organizada em cinco frentes principais:

- compreensão dos fluxos e mecanismos de acesso aos dados disponibilizados pelo Harvester do Oasisbr;
- definição de uma camada de integração capaz de acessar e tratar os dados sem replicação desnecessária das informações de coleta;
- tratamento e preparação dos metadados textuais, com aplicação de procedimentos de normalização e organização dos valores;
- investigação de requisitos para posterior utilização desses dados por componentes de NLP e LLMs previstos no SAVITS;
- validação da abordagem por meio de um protótipo funcional e de experimentos com dados reais.

No tratamento dos metadados, foram investigados procedimentos de normalização textual e agrupamento de valores associados às regras de validação, com o objetivo de reduzir variações de representação, identificar padrões recorrentes e preparar os dados para etapas posteriores de processamento baseado em linguagem natural.

A arquitetura estudada preservou o Harvester como fonte dos dados de coleta, evitando sua replicação em uma nova base. A camada de integração passou a realizar o acesso e o tratamento das informações necessárias às aplicações consumidoras, enquanto dados próprios da aplicação, como contas, perfis, vínculos, avisos e pedidos de coleta, permaneceram em banco específico.

Buscando materializar e validar parte da camada de integração proposta, foi desenvolvida uma Interface do Gestor, utilizada como artefato experimental da pesquisa. A interface permitiu testar mecanismos de acesso aos dados, controle de permissões, tratamento de respostas, navegação entre coletas, diagnósticos e registros e apresentação dos resultados do processamento.

A definição e o acompanhamento desse escopo ocorreram em conjunto com profissionais do Ibict envolvidos na administração do Harvester e no desenvolvimento de sistemas de informação científica. A Tabela 1 apresenta os participantes diretamente envolvidos na definição, revisão e validação das atividades desenvolvidas.

Tabela 1 – Participantes envolvidos na definição e validação da pesquisa

Pessoa	Vínculo	Papel no projeto

Gabriel Silveira Marques	Administrador do Harvester (Ibict)	Apresentou a demanda; validou o sistema
Blena Estevam dos Santos	Administradora do Harvester (Ibict)	Apresentou a demanda; validou o sistema
Millena Cordeiro Matos de Lima	Administradora do Harvester (Ibict)	Apresentou a demanda; validou o sistema
Cássio Teixeira de Morais	Administrador do Harvester (Ibict)	Validou o sistema
Priscila Lima	Servidora do Ibict, lotada na CODIC	Apresentou a demanda; revisou as histórias de usuário
Rogério Nogueira de Sousa	Bolsista	Levantamento, especificação, desenvolvimento, testes e implantação

Fonte: Elaborado pelo autor (2026).

2 MÉTODO

A pesquisa foi conduzida de forma aplicada e incremental, organizada em cinco etapas: planejamento, análise de requisitos, definição da arquitetura, desenvolvimento e validação. O Harvester do Oasisbr foi utilizado como ambiente experimental, permitindo trabalhar com dados reais de repositórios, coletas, diagnósticos e registros para investigar mecanismos de integração, tratamento e preparação de dados relacionados às necessidades do SAVITS.

Como referência para o enquadramento da pesquisa no SAVITS, considerou-se a metodologia adotada pelo projeto, estruturada a partir do CRISP-DM (CHAPMAN et al., 2000; IBICT, 2026). Dentro desse ciclo, o recorte realizado durante o período da bolsa relaciona-se principalmente às atividades de entendimento dos dados, preparação dos dados, desenvolvimento experimental e avaliação. O projeto SAVITS prevê que os dados sejam preparados para alimentar de forma confiável seus modelos analíticos e estabelece a interoperabilidade como princípio para integração com Oasisbr, BDTD e BrCris.

Na etapa de planejamento e análise de requisitos, foram realizadas reuniões com a equipe responsável pela administração do Harvester e com a CODIC para compreender os fluxos de informação, as formas de acesso disponíveis e os problemas relacionados ao

acompanhamento da qualidade dos metadados. Os requisitos identificados foram organizados em sete histórias de usuário, descritas no formato “Como..., quero..., para...”, conforme a abordagem de histórias de usuário apresentada por Cohn (2004), e acompanhadas de critérios de aceite, delimitação do escopo e comportamento esperado na ausência de dados. Uma planilha compartilhada foi utilizada para acompanhar as alterações solicitadas e registrar a validação das funcionalidades desenvolvidas. Uma planilha compartilhada foi utilizada para acompanhar as alterações solicitadas e registrar a validação das funcionalidades desenvolvidas.

Tabela 2 - Histórias de usuário utilizadas na experimentação e validação

História	Tela	Para quem
HU0000	Meus Repositórios: situação da última coleta de cada repositório	Gestor
HU0001	Análise da coleta: diagnóstico da qualidade e registros com problema	Gestor
HU0002	Notificações: avisos do Ibict aos gestores, com confirmação de leitura	Gestor e administrador
HU0003	Perfis e permissões: contas, perfis e vínculo com repositórios	Administrador
HU0004	Solicitação de coleta: pedido do gestor e resposta do Ibict	Gestor e administrador
HU0005	Administração de repositórios: todos os repositórios numa tabela	Administrador
HU0006	Página do repositório: cadastro e histórico de coletas	Gestor e administrador

Fonte: Elaborado pelo autor (2026).

Na etapa de arquitetura, foi adotada uma estratégia que preserva o Harvester como fonte dos dados das coletas. Informações sobre coletas, diagnósticos e registros são consultadas na infraestrutura de origem, enquanto apenas informações próprias da aplicação, como contas, perfis, vínculos entre usuários e repositórios, avisos e pedidos de coleta, são mantidas em banco específico. Essa decisão permitiu experimentar uma camada de integração sem replicação desnecessária dos dados mantidos pelo Harvester. A estratégia

também se alinha ao princípio de interoperabilidade previsto no SAVITS e aos princípios FAIR, especialmente quanto à capacidade de acesso, interoperabilidade e reutilização das informações (WILKINSON et al., 2016).

Como a consulta depende de uma infraestrutura externa e de um volume elevado de informações, foram implementados mecanismos de cache com tempos de retenção distintos. Dados de coletas concluídas, que não sofrem novas alterações, podem permanecer armazenados por períodos maiores, enquanto informações ainda mutáveis são renovadas com maior frequência. Em caso de falha de comunicação, a requisição é repetida e o erro não é armazenado como resposta válida. O controle de acesso também foi aplicado no servidor, de forma que um gestor possa acessar somente os repositórios associados à sua conta.

Para a obtenção dos metadados foram considerados os padrões empregados pelo ecossistema do Oasisbr. O protocolo OAI-PMH foi utilizado como base para a coleta de metadados em repositórios (OPEN ARCHIVES INITIATIVE, 2002), enquanto o Dublin Core forneceu os elementos descritivos presentes nos registros analisados (DUBLIN CORE METADATA INITIATIVE, 2020).

No contexto da preparação dos dados para componentes de NLP e LLMs previstos no SAVITS, foram analisados os metadados textuais disponibilizados pelos repositórios ao Oasisbr, especialmente campos Dublin Core coletados por OAI-PMH, padrões amplamente utilizados para descrição e interoperabilidade de metadados em repositórios digitais (OPEN ARCHIVES INITIATIVE, 2002; DUBLIN CORE METADATA INITIATIVE, 2020). Como parte desse conteúdo é formada por texto livre, foram observadas variações de grafia, acentuação, caixa e pontuação. Foram aplicados procedimentos de normalização a títulos e nomes de instituições, incluindo uniformização de caixa, remoção de diacríticos, eliminação de pontuação e segmentação em palavras, operações associadas à preparação textual para tarefas de processamento de linguagem natural (JURAFSKY; MARTIN, 2026). Também foram agrupados os valores associados às regras de validação por frequência de ocorrência, permitindo identificar padrões recorrentes.

Esses procedimentos foram utilizados como etapa de preparação e organização dos dados para análises posteriores baseadas em linguagem natural, uma vez que o pré-processamento busca reduzir variações de representação antes da aplicação de métodos

analíticos sobre o texto (JURAFSKY; MARTIN, 2026). O trabalho realizado no período concentrou-se nessa camada preparatória e de integração, não sendo atribuídos aos procedimentos de normalização resultados de tarefas posteriores, como classificação textual, análise de sentimentos ou fine-tuning de LLMs, que dependem da continuidade da investigação.

Buscando materializar e validar os mecanismos investigados, a Interface do Gestor foi utilizada como artefato experimental. Sua implementação ocorreu de forma incremental e sobre dados reais do Harvester, permitindo testar acesso às informações, controle de permissões, tratamento das respostas, cache e navegação entre repositórios, coletas, diagnósticos e registros. Para que a equipe do Ibict pudesse avaliar a solução, foi disponibilizada uma infraestrutura específica de validação.

Na construção da interface foi adotado o Padrão Digital de Governo (BRASIL, [s.d.]), considerando requisitos de acessibilidade e internacionalização. Os critérios de acessibilidade, especialmente contraste visual e organização dos elementos da interface, foram orientados pelo Modelo de Acessibilidade em Governo Eletrônico – eMAG (BRASIL, 2014) e pelas Web Content Accessibility Guidelines – WCAG 2.1 (W3C, 2018).

A validação foi conduzida em três frentes: testes automatizados, experimentação com dados reais e avaliação pelos administradores do Harvester. Os testes automatizados utilizaram respostas simuladas do coletor, evitando dependência da disponibilidade da infraestrutura externa e permitindo a repetição das verificações. Entre os aspectos avaliados estava a regra de controle de acesso, segundo a qual gestores não podem consultar repositórios não associados à sua conta.

Como parte da experimentação com dados reais, foi realizado um teste para avaliar o mecanismo de identificação do endereço correspondente ao documento no repositório de origem. Foram utilizados 1.033 registros pertencentes a 305 repositórios e comparadas a estratégia anteriormente utilizada e uma nova forma de descoberta do endereço. Para cada registro, verificou-se se o endereço recuperado conduzia efetivamente ao documento. Também foi verificado o acesso ao metadado bruto disponibilizado pela fonte de origem.

3 RESULTADOS ALCANÇADOS

Os resultados obtidos demonstraram a viabilidade da utilização de uma camada intermediária para acesso e tratamento dos dados disponibilizados pelo Harvester, preservando a infraestrutura de origem como responsável pelos dados de coletas, diagnósticos e registros. A Interface do Gestor encontra-se funcionando com dados reais no ambiente de validação e foi avaliada pelos administradores do Harvester, com o acompanhamento das validações realizado em planilha compartilhada.

3.1 Validação da camada de integração

A implementação incremental permitiu avaliar diferentes funções relacionadas ao acesso, organização e disponibilização dos dados. As entregas realizadas foram reorganizadas de acordo com sua contribuição para a experimentação da camada de integração.

Tabela 3 – Principais resultados da implementação experimental

Eixo	Resultados alcançados
Acesso e controle	Autenticação; associação de gestores a repositórios; controle de permissões aplicado no servidor; painel administrativo
Consulta aos dados	Acesso a repositórios, coletas, diagnósticos e registros provenientes do Harvester
Qualidade dos metadados	Identificação das regras de validação; apresentação dos valores responsáveis pelas falhas; acesso aos detalhes de cada registro
Navegação e análise	Filtros de registros; paginação; manutenção de filtros no endereço da página; navegação entre diagnóstico e registros
Comunicação	Avisos aos gestores, confirmação de leitura e solicitação de nova coleta
Exportação e rastreabilidade	Exportação dos registros e do histórico de coletas; acesso à página do item e ao metadado disponibilizado pela fonte

Validação	Histórias HU0000 a HU0006 confrontadas com o sistema e validação em ambiente com dados reais

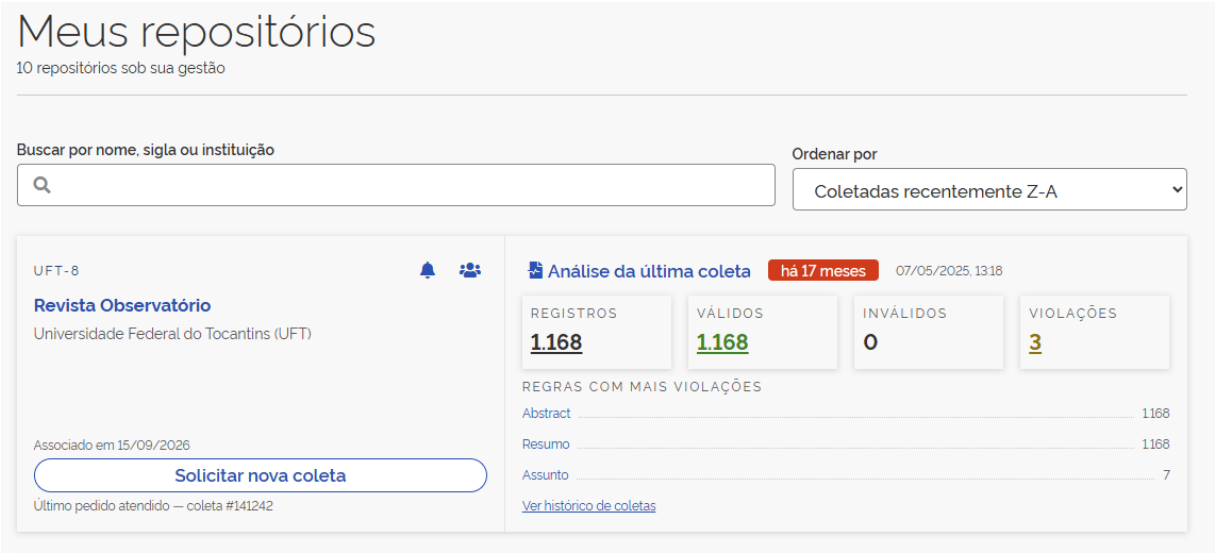
Fonte: Elaborado pelo autor (2026).

Os resultados mostram que a camada experimental conseguiu consumir os dados do Harvester sem necessidade de manter uma cópia das coletas em sua própria base. Essa separação permitiu manter os dados operacionais sob responsabilidade da infraestrutura de origem e concentrar na camada intermediária as funções de acesso, tratamento, apresentação e controle necessárias às aplicações consumidoras.

3.2 Interface, acessibilidade e navegação

A Interface do Gestor utiliza o Padrão Digital de Governo e apresenta a identidade visual adotada pelos projetos do Ibict. Nos casos em que componentes existentes não atendiam às necessidades de acessibilidade ou internacionalização, foram produzidas adaptações mantendo a mesma identidade visual. As cores utilizadas em alertas também foram verificadas quanto ao contraste, e as medições realizadas foram registradas durante o desenvolvimento. A interface foi disponibilizada em português, espanhol e inglês.

Figura 1- Tela Meus Repositórios

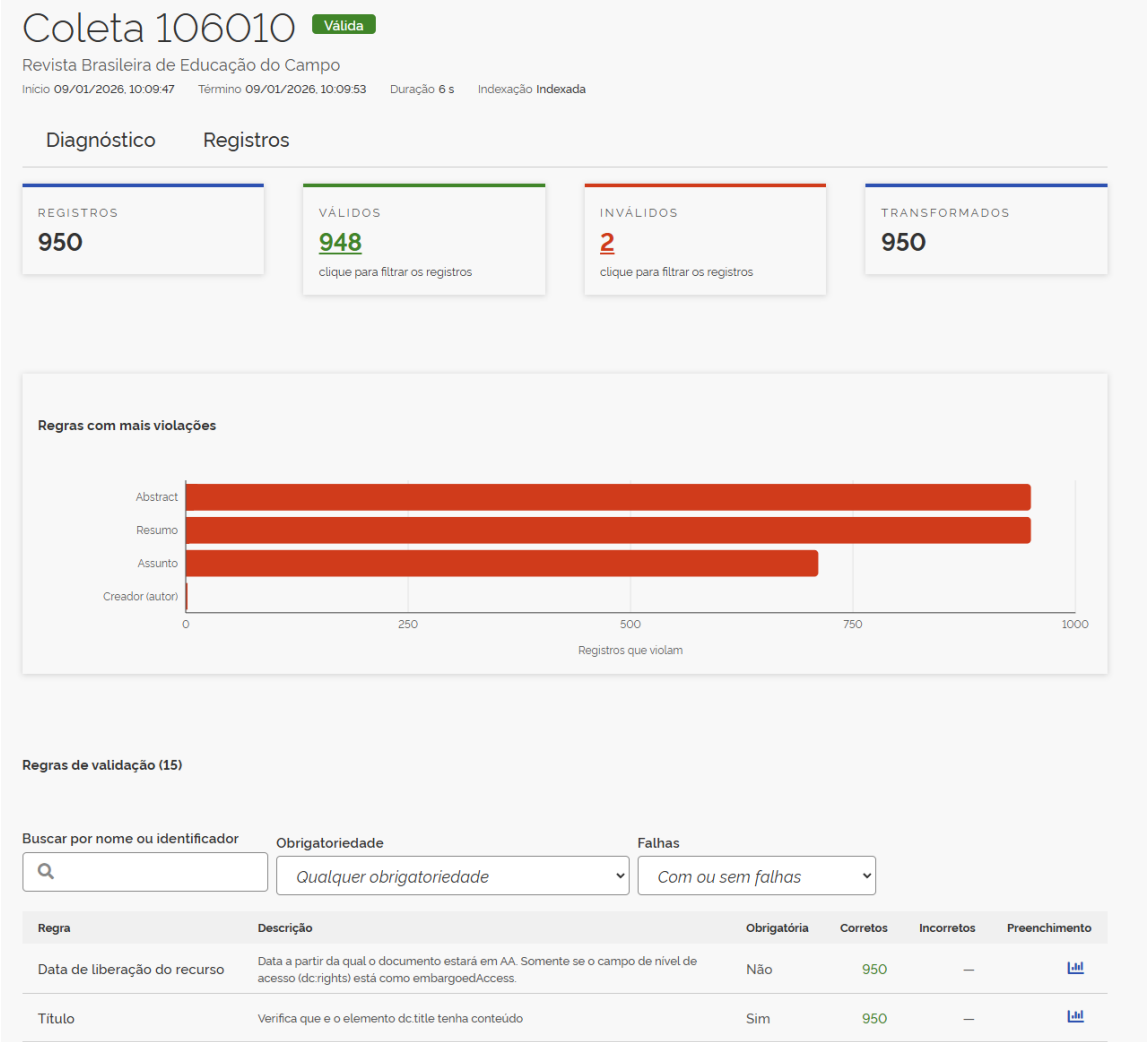


Fonte: Elaborado pelo autor (2026).

Os filtros relacionados à validade e à transformação dos registros foram mantidos no próprio endereço da página. Dessa forma, a seleção realizada no diagnóstico conduz aos

registros correspondentes já filtrados, o estado da consulta pode ser preservado durante a navegação e o mesmo recorte pode ser compartilhado por meio do endereço da página.

Figura 2- Tela Análise da coleta



Fonte: Elaborado pelo autor (2026).

3.3 Preparação dos metadados para NLP e LLMs

Os registros utilizados são provenientes dos repositórios integrados ao Oasisbr e apresentam campos descritivos em Dublin Core coletados por OAI-PMH. Por se tratarem, em grande parte, de informações preenchidas como texto livre, foram observadas diferenças de grafia, acentuação, uso de maiúsculas e minúsculas e pontuação.

O processamento concentrou-se principalmente em dois grupos de informação: títulos e nomes de instituições. Para os títulos, as métricas foram calculadas sobre uma amostra de

318.953 registros pertencentes a 1.573 fontes, correspondente a aproximadamente 200 registros e 12 páginas de resultados por fonte. Desse conjunto, 295.767 registros apresentavam título e foram submetidos à normalização textual. Para os nomes de instituições, a análise considerou o universo das 2.183 fontes disponíveis, vinculadas inicialmente a 755 nomes de instituições distintos.

A normalização foi realizada de forma determinística, aplicando decomposição Unicode NFKD com remoção de diacríticos, conversão para caixa baixa e eliminação de pontuação. O efeito do processamento foi medido pela cardinalidade, comparando-se a quantidade de valores distintos antes e depois da normalização. Nos títulos, o número de formas distintas foi reduzido de 275.039 para 274.096 (−0,3%). Nos nomes de instituições, a cardinalidade passou de 755 para 745 (−1,3%). A redução relativamente pequena indica que apenas uma parcela da heterogeneidade observada decorre de diferenças superficiais de grafia, caixa, acentuação e pontuação, sugerindo que a identificação de equivalências mais amplas exige etapas adicionais de processamento semântico.

Também foi analisada a concentração dos valores associados às regras de validação. As violações foram agrupadas por forma textual idêntica, ordenadas por frequência e avaliadas pela participação acumulada, tanto no conjunto total quanto individualmente por regra. Das 1.304.739 ocorrências de valores que violam alguma regra, as 10 formas mais frequentes concentram 52,8% das violações, enquanto apenas 7 formas respondem por metade do total. Em regras específicas, a concentração mostrou-se ainda mais elevada: na validação de idioma segundo o padrão ISO 639-3, as 10 formas mais recorrentes concentram 98,6% das violações. Esse resultado indica que parcela expressiva das inconsistências pode ser priorizada a partir de um conjunto reduzido de formas recorrentes.

Todas as métricas foram calculadas de forma determinística sobre um conjunto de dados congelado, de modo que a mesma entrada e os mesmos procedimentos de normalização e agrupamento produzam os mesmos resultados. Esses resultados caracterizam uma etapa de preparação dos dados necessária antes da aplicação de técnicas mais avançadas de NLP e LLMs, como classificação, extração de informações, análise semântica e sumarização. Embora não tenha sido avaliado, neste período, o impacto desse pré-processamento sobre o desempenho de modelos de NLP ou LLMs, foi possível mensurar seu efeito sobre a cardinalidade dos valores e a concentração das violações, fornecendo

evidências sobre a qualidade e a estrutura dos dados de entrada para as etapas analíticas posteriores.

3.4 Experimento com repositórios reais

O principal experimento quantitativo realizado no período avaliou a capacidade da camada de integração de identificar corretamente a página do documento no repositório de origem. Foram verificados 1.033 registros pertencentes a 305 repositórios, comparando-se a estratégia de descoberta utilizada inicialmente com uma nova abordagem desenvolvida durante a pesquisa.

Tabela 4 – Resultados do experimento de identificação da página do item

Indicador	Resultado
Registros avaliados	1.033
Repositórios avaliados	305
Acerto da estratégia anterior	74,9%
Acerto da nova estratégia	86,5%
Variação absoluta	+11,6 pontos percentuais
Links que levavam a destino incorreto após a nova abordagem	1
Acesso bem-sucedido ao metadado bruto da origem	98,6%

Fonte: Elaborado pelo autor (2026).

A nova estratégia elevou a proporção de registros em que o endereço conduzia corretamente ao documento de 74,9% para 86,5%, uma diferença de 11,6 pontos percentuais. Além disso, os casos em que o endereço conduzia a destinos distintos do documento esperado, como páginas de autenticação ou identificadores DOI não registrados, foram reduzidos de 139 para 1. O acesso ao metadado bruto disponibilizado pela origem foi possível em 98,6% dos registros avaliados.

O experimento evidencia que a integração com fontes heterogêneas não depende apenas da disponibilidade dos metadados, mas também da capacidade de interpretar corretamente as informações disponibilizadas por diferentes repositórios. Esse aspecto é relevante para a integração com o SAVITS porque componentes posteriores de NLP e LLMs dependem da recuperação consistente dos dados e de seus respectivos contextos de origem.

3.5 Situação atual e próximos passos

Ao final do período da bolsa, a Interface do Gestor encontra-se em funcionamento com os dados reais do Harvester em ambiente de validação. A solução já permite experimentar os principais mecanismos de acesso e tratamento definidos para a camada de integração.

Permanecem como próximos passos a definição da hospedagem definitiva, em substituição à infraestrutura utilizada apenas para validação; a conclusão dos critérios de aceite ainda marcados como “a implementar” nas histórias de usuário, entre eles a manutenção dos filtros da tabela administrativa no endereço da página, a busca independente de acentuação, o aviso ao administrador sobre novos pedidos de coleta e os ajustes no gráfico do histórico; e a automatização do atendimento aos pedidos de coleta, atualmente realizado de forma manual.

A continuidade da pesquisa irá prover ao SAVITS uma camada estruturada de integração e preparação de dados, capaz de apoiar a experimentação dos componentes analíticos previstos na plataforma, incluindo técnicas de classificação, extração de informações e aplicações baseadas em LLMs, a partir de dados previamente tratados, organizados e qualificados.

4 CONSIDERAÇÕES FINAIS

A pesquisa desenvolvida no período permitiu investigar, em uma infraestrutura real do ecossistema de informação científica do Ibict, mecanismos necessários para acesso, tratamento e preparação de dados que podem subsidiar a integração com os componentes analíticos do SAVITS. O Harvester do Oasisbr foi utilizado como ambiente experimental por permitir acesso a um conjunto amplo e heterogêneo de repositórios, registros e resultados de validação.

A principal contribuição foi a experimentação de uma camada intermediária que mantém o Harvester como fonte dos dados de coleta e concentra funções de acesso, controle, tratamento e disponibilização das informações. Essa abordagem foi materializada por meio da Interface do Gestor, que atuou como artefato experimental para testar a arquitetura proposta sobre dados reais.

No âmbito de NLP e LLMs, o trabalho concentrou-se na etapa que antecede a aplicação dos modelos: a qualidade, normalização e organização dos dados de entrada. A análise dos metadados permitiu identificar variações de representação, aplicar procedimentos de normalização e estruturar valores associados às regras de qualidade. Esses resultados estabelecem uma base para a continuidade da investigação com técnicas mais avançadas de processamento de linguagem natural.

O experimento realizado com 1.033 registros de 305 repositórios também forneceu evidência quantitativa da possibilidade de melhorar mecanismos de integração com fontes heterogêneas. A nova estratégia de descoberta da página do item elevou o percentual de acertos de 74,9% para 86,5%, ao mesmo tempo em que reduziu de 139 para 1 os casos em que o endereço recuperado conduzia a um destino inadequado.

Os resultados obtidos indicam, portanto, que o trabalho realizado durante o período da bolsa pode ser entendido como uma etapa de investigação e preparação para a integração do SAVITS com o ecossistema informacional do Ibict. A continuidade da pesquisa poderá avançar sobre essa base para incorporar mecanismos de análise semântica, classificação, extração de informações e LLMs, mantendo a qualidade e a rastreabilidade dos dados como requisitos do processo.

4.1 Possíveis desdobramentos da pesquisa

Além da continuidade direta da integração com os componentes analíticos do SAVITS, a exploração dos dados históricos do Harvester abriu novas possibilidades de pesquisa relacionadas à qualidade e à atualização das fontes de informação científica.

Entre as questões identificadas para investigações futuras estão:

- A incorporação de novidade semântica e novidade quantitativa melhora a priorização de fontes para recoleta em comparação com uma estratégia baseada apenas em defasagem, volume histórico e completude?

- Quais características operacionais das fontes estão associadas à persistência de falhas de coleta ao longo do tempo?
- Em que medida a qualidade dos metadados varia entre fontes e quais dimensões são mais úteis para distinguir problemas estruturais de problemas de conteúdo?
- A estimativa de novidade potencial de uma fonte melhora a capacidade de identificar recoletas com maior benefício informacional em relação ao uso apenas do volume histórico e do tempo desde a última coleta?

Essas questões surgiram a partir dos dados e problemas observados durante a investigação e representam possibilidades de continuidade científica relacionadas à interoperabilidade, qualidade de metadados e uso de métodos analíticos e semânticos sobre infraestruturas de Ciência Aberta.

REFERÊNCIAS

BRASIL. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Logística e Tecnologia da Informação. *Modelo de Acessibilidade em Governo Eletrônico: eMAG*. Versão 3.1. Brasília, DF: MP/SLTI, 2014.

BRASIL. Governo Digital. *Padrão Digital de Governo: Design System*. Brasília, DF: Governo Federal, [s.d.]. Disponível em: gov.br/design-system. Acesso em: 17 ago. 2026.

CHAPMAN, Pete et al. *CRISP-DM 1.0: step-by-step data mining guide*. [S.l.]: SPSS, 2000.

COHN, Mike. *User stories applied: for agile software development*. Boston: Addison-Wesley Professional, 2004.

DUBLIN CORE METADATA INITIATIVE. *DCMI Metadata Terms*. DCMI Recommendation, 2020. Disponível em: dublincore.org/specifications/dublin-core/dcmi-terms/. Acesso em: 17 ago. 2026.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). *SAVITS: Sistema de Avaliação e Validação de Impactos em Tecnologias Sociais*. Brasília, DF: Ibict, 2026. Projeto de pesquisa.

JURAFSKY, Daniel; MARTIN, James H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with*

language models. 3. ed. draft. 2026. Manuscrito online. Disponível em: web.stanford.edu/~jurafsky/slp3/. Acesso em: 17 ago. 2026.

OPEN ARCHIVES INITIATIVE. *The Open Archives Initiative Protocol for Metadata Harvesting: OAI-PMH*. Version 2.0. 2002. Disponível em: openarchives.org/OAI/openarchivesprotocol.html. Acesso em: 17 ago. 2026.

WILKINSON, Mark D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, v. 3, art. 160018, 2016. DOI: 10.1038/sdata.2016.18.

WORLD WIDE WEB CONSORTIUM (W3C). *Web Content Accessibility Guidelines (WCAG) 2.1*. W3C Recommendation, 5 June 2018. Disponível em: w3.org/TR/WCAG21/. Acesso em: 17 ago. 2026.

gov.br
Documento assinado digitalmente
ROGERIO NOGUEIRA DE SOUSA
Data: 23/10/2026 16:33:03-0300
Verifique em https://validar.iti.gov.br

Rogério Nogueira de Sousa
Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 29/10/2026 07:55:48-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo
Coordenador do projeto