Análise e Publicação de Dados de Processos Eletrônicos em Organizações Públicas

Ivan Luiz Salvadori[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Rua Cristo Rei, 19 – Vila Becker – CEP 85902-490 – Toledo – PR – Brasil

ivanlsalvadori@utfpr.edu.br

Abstract. *The Electronic Information System is used for managing documents and digital processes in Brazilian public organizations. It promotes administrative efficiency, resource savings, agility, and transparency. It facilitates the processing of digital processes, increasing public administration productivity. Despite the benefits, the system has limitations in information retrieval. This work proposes a platform for analyzing and publishing digital processes, allowing a more diverse use of public data. The platform adopts methodologies and tools to structure and publish open data, by adopting technologies such as the Semantic Web, linked data, and natural language processing.*

Resumo. *O Sistema Eletrônico de Informações é utilizado para gestão de documentos e processos eletrônicos em organizações públicas brasileiras. Promove eficiência administrativa, economia de recursos, agilidade e transparência. Facilita a tramitação de processos eletrônicos, aumentando a produtividade da administração pública. Apesar dos benefícios, o sistema possui limitações na estruturação e recuperação de informações. Este trabalho propõe uma plataforma para análise e publicação de processos eletrônicos, permitindo um uso mais amplo dos dados públicos. A plataforma utiliza metodologias e ferramentas para estruturar e publicar dados abertos por meio da adoção de tecnologias da Web semântica, dados conectados e processamento de linguagem natural.*

1. Introdução

O Sistema Eletrônico de Informações (SEI) é a solução oficial para gestão de documentos e processos eletrônicos das organizações públicas brasileiras. O SEI busca promover eficiência administrativa por meio da digitalização e automação dos procedimentos internos, resultando em economia de recursos, agilidade e transparência na disponibilização de informações. Com funcionalidades que permitem a criação, edição e tramitação de processos eletrônicos, o SEI disponibiliza diversos módulos projetados para aumentar a produtividade e modernizar a administração pública brasileira. Possui interface Web amigável, que permite o controle de prazos processuais, assinatura eletrônica de documentos e consulta processual.

Apesar dos benefícios proporcionados pelo SEI, desafios persistem na gestão de documentos eletrônicos, incluindo a heterogeneidade de dados e a dificuldade na recuperação de informações. Para enfrentar esses desafios, este trabalho propõe uma plataforma para análise e publicação de processos eletrônicos. O objetivo da plataforma é permitir um uso mais amplo e diversificado dos dados processuais gerados pela administração pública

aos cidadãos. Para isso, a plataforma reúne metodologias e ferramentas consolidadas para organização e publicação de dados abertos. Avalia e discute a adoção de tecnologias da Web semântica, dados conectados e modelos generativos de processamento de linguagem natural aplicados no cenário de processos eletrônicos. Além disso, apresenta um panorama da utilização do SEI nas organizações públicas e discute soluções para as limitações técnicas do sistema.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta os conceitos básico sobre o SEI, seus benefícios, limitações e panorama de utilização na administração pública. A plataforma proposta é apresentada na Seção 3. A Seção 4 discute e avalia a adoção da plataforma no cenário específico das universidades federais. Por fim, a Seção 5 apresenta as conclusões e sugere trabalhos futuros.

2. Sistema Eletrônico de Informações

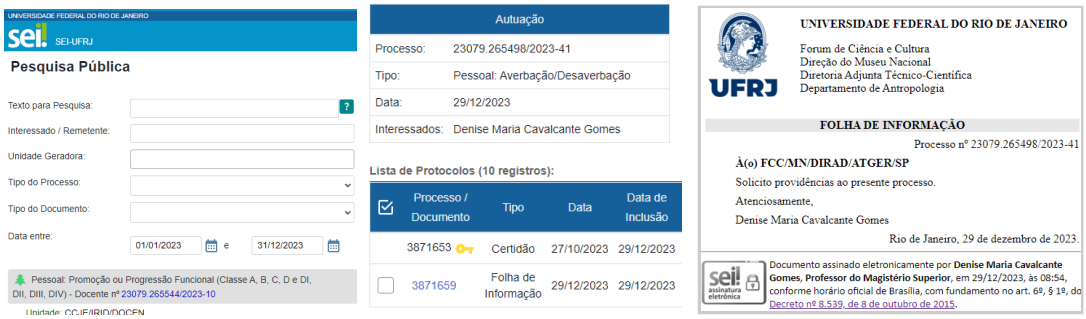
O Sistema Eletrônico de Informações (SEI), desenvolvido pelo Tribunal Regional Federal da 4ª Região (TRF4) e sob responsabilidade do Ministério da Gestão e da Inovação em Serviços Públicos, é uma ferramenta para a gestão de documentos e processos eletrônicos. Sua principal função é promover eficiência administrativa através da digitalização e automação dos procedimentos internos. O SEI é a principal solução do Processo Eletrônico Nacional (PEN)[1], um projeto colaborativo que envolve órgãos de diferentes esferas da administração pública. O objetivo do PEN é criar uma infraestrutura pública para o gerenciamento de documentos e processos administrativos que possibilitam economia de recursos e oferecer mais agilidade e transparência com a disponibilização de informações.

O SEI permite a criação, edição, tramitação e consulta de processos eletrônicos. Essas funcionalidades são implementadas em módulos projetados para aumentar a produtividade através da modernização da administração pública brasileira. Além da padronização dos processos, o SEI pode reduzir gastos com material de expediente, como verificado por [Filho e Peixe 2017]. Possui interface Web amigável que facilita a adaptação e o uso contínuo por parte dos servidores. Entre as principais vantagens gerenciais oferecidas pelo SEI, destacam-se (a) o controle de prazos processuais, realização de estatísticas e monitoramento do tempo de processamento; (b) disponibilidade de modelos de documentos nato-digitais e textos padrão com assinatura eletrônica de documentos; (c) pesquisa integral nos documentos, acompanhamento e inspeção administrativa.

2.1. Interface Web para Consulta Pública

O SEI disponibiliza módulos para a consulta pública de processos eletrônicos. Esta funcionalidade é disponibilizada por uma interface Web que permite aos cidadãos realizarem consultas de acordo com características específicas de processos e documentos eletrônicos. Conforme demonstrado pela Figura 1-(a), é possível aplicar filtros de pesquisa para encontrar um processo com um determinado número de protocolo, além de outras características como tipo de documento ou processo, unidade geradora e processos criados ou que tiveram tramitação em um intervalo de tempo. O resultado da pesquisa é representado por uma lista de processos com informações básicas de identificação e um *link* para obtenção de mais detalhes de um processo específico.

[1]https://www.gov.br/gestao/pt-br/assuntos/processo-eletronico-nacional



(a) (b) (c)

Figura 1. Interface Web SEI: (a) Pesquisa pública; (b) Detalhes de um processo; (c) Documento interno nato-digital

A interface Web de detalhes de processo apresenta dados de autuação, lista de protocolos e lista de andamentos. Os dados de autuação definem o número, tipo, data de geração e pessoas interessadas no processo. A lista de protocolos representa os documentos anexados ao processo. É importante destacar que processos e documentos eletrônicos podem ser classificados como acesso público ou restrito. A Figura 1-(b) mostra a interface Web de um processo eletrônico. Neste exemplo, a lista de anexos possui um documento de acesso público e um documento restrito.

Documentos públicos podem ser consultados em seu inteiro teor. A Figura 1-(c) mostra um documento nato-digital de acesso público. Documentos nato-digitais são informações textuais geradas a partir de modelos previamente estabelecidos pela organização pública. Estes documentos possuem um número de identificação e tratam de um determinado assunto. De forma geral, os documentos são assinados eletronicamente por um servidor da organização. A assinatura eletrônica define a autoria, o papel desempenhado pelo servidor na organização, garante a integridade dos dados, não repúdio e as demais características previstas nos mecanismos de assinatura digital de documentos.

2.2. Utilização do SEI em Organizações Públicas

A adoção do SEI em organizações públicas representa um marco significativo na modernização administrativa. A adoção envolve várias etapas que vão desde a decisão de implantação até a completa automação das rotinas de uma organização pública. A exemplo da implantação do SEI nos ministérios federais brasileiros, que teve início em 2015, trazendo vantagens e desafios. Apesar da resistência de servidores no abandono do uso do papel e a dificuldade com as ferramentas tecnológicas, a adoção do SEI e a produção de documentos nos ministérios é uma realidade estabelecida [da Silva et al. 2018], fato que pode ser comprovado pela quantidade de processos gerados pelos diversos ministérios.

A Tabela 1 resume a quantidade de processos criados ou tramitados em 2023 pelas organizações públicas observadas. Dentre os 31 ministérios federais, 22 utilizam o SEI, e disponibilizaram cerca de três milhões de processos nesse período. Além disso, 50% dos governos estaduais também adotaram o SEI, e disponibilizaram quase 15 milhões de processos em 2023. Esses números refletem a crescente adesão ao SEI tanto no âmbito federal quanto estadual, consolidando-o como uma ferramenta importante para a gestão de processos em organizações públicas.

Tabela 1. Adoção do SEI em organizações públicas

Organização	Total	Adesão	Processos	Pesq. Pública
Ministérios Federais	31	22	2.831.141	17
Governos Estaduais	26	13	14.888.964	12
Universidades Federais	68	40	1.462.195	33
Institutos Federais	37	10	165.682	7

Outro dado importante é a adoção do SEI nas instituições de ensino federal. Aproximadamente 60% das universidades federais brasileiras adotam o SEI como ferramenta de gestão para processos eletrônicos. Essas universidades disponibilizaram, no período observado, quase um milhão e meio de processos. Os institutos federais de educação iniciaram recentemente a adoção do SEI. Atualmente, dez institutos federais utilizam essa ferramenta e disponibilizaram 165.682 processos nos 12 meses de 2023. Embora a quantidade de processos disponibilizados pelos institutos federais seja muito inferior comparada às demais organizações públicas, ainda assim é um número expressivo. Em média, 46 processos são criados ou tramitados diariamente em cada instituto federal.

A Tabela 1 ainda apresenta o número de organizações que disponibilizam a pesquisa pública de seus processos eletrônicos. Conforme apurado, aproximadamente 20% das organizações não disponibilizam a pesquisa pública dos processos. Na ausência da pesquisa pública, o cidadão somente pode consultar processos pelo número do protocolo, fato que prejudica drasticamente a transparência dos dados públicos. Este é o cenário descrito como dados escuros, onde as informações não são estruturadas e disponibilizadas adequadamente, de modo que são de difícil acesso ao público e, portanto, é mais provável que permaneçam subutilizadas e eventualmente perdidas [Albuquerque e Dorneles 2022].

Pode-se afirmar que o SEI atende adequadamente às atividades administrativas das organizações públicas, pois possibilita a rápida produção e tramitação de processos eletrônicos. No entanto, a automatização promovida pelo SEI não cobre todas as fases da gestão documental. Essa deficiência pode comprometer a capacidade de decisão dos gestores devido à falta de estruturação das informações [da Silva et al. 2018].

Os dados dos processos eletrônicos são disponibilizados por meio de interfaces Web, por elementos HTML como tabelas, campos de texto e rótulos. Esta forma de representação de dados é considerada como semi-estruturada, ou seja, possuem estrutura heterogênea. Há mais de duas décadas, pesquisadores da área de banco de dados alertavam sobre as limitações causadas pela estruturação inadequada de dados. A heterogeneidade aumenta a complexidade de pesquisas, uma vez que não existe um esquema uniforme a partir do qual uma consulta possa ser formulada. Consultas são realizadas por navegação exaustiva ou busca por palavras-chave [Mello et al. 2000].

Pesquisadores da área da ciência da informação e arquivologia estudaram o impacto da adoção do SEI em diversas organizações públicas. Estes estudos apontam a problemática da garantia de acesso a curto e longo prazo dos documentos. Destacam as dificuldades de recuperação de informações [Rodrigues e Cavalcante 2018], requerendo outros sistemas para ampararem essas atividades [Macedo e Tolfo 2017].

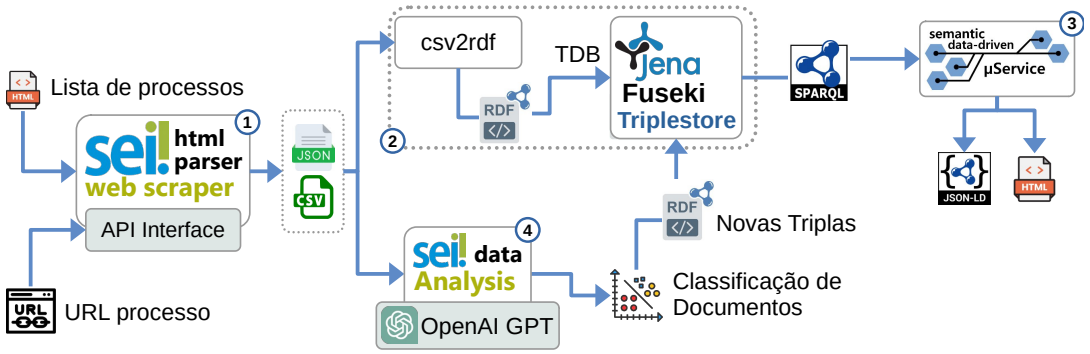


Figura 2. Componentes arquiteturais e fluxo de execução

3. Plataforma para Análise e Publicação de Processos Eletrônicos

Este trabalho propõe uma plataforma para análise e publicação de processos eletrônicos, capaz de aprimorar a estrutura e a publicação das informações processuais, possibilitando um uso mais amplo e diversificado pelos cidadãos. A plataforma combina ferramentas em um fluxo de execução de quatro etapas: obtenção e estruturação de dados semi-estruturados do SEI; enriquecimento semântico dos dados; publicação de dados conectado; e análises para classificação de documentos e identificação de pessoas mencionadas, como demonstrado na Figura 2. Inspirado na Rede-Go-Fair-Agro[2], que se propõe a desenvolver soluções para interoperabilidade de dados para as ciências agrárias, a plataforma proposta busca alcançar resultados semelhantes no campo do governo eletrônico.

3.1. Etapa 1: Obtenção e Estruturação de Dados

A publicação dos dados dos processos SEI é realizada exclusivamente através de páginas HTML. Tecnologias como Web APIs e Serviços Web poderiam ser adotadas para permitir a troca de informações entre o SEI e aplicações externas interessada em dados processuais. Devido à falta de suporte do SEI à interoperabilidade dos dados, a plataforma proposta realiza a obtenção dos dados por meio da técnica de Web *scraping*, também conhecida como raspagem de páginas Web. Para esse fim, foi desenvolvida a aplicação *sei.html.parser*, que recebe como entrada arquivos HTML resultantes da pesquisa pública, como exemplificado pela Figura 1-(a). Sendo assim, é necessário realizar manualmente uma pesquisa pública com processos de interesse e salvar em arquivo a página HTML resultante. A partir do arquivo HTML, a aplicação *sei.html.parser* identifica os *links* para os processos e obtém os detalhes de cada processo contido na listagem, além de obter os dados de todos os documentos anexados. É importante destacar que o SEI permite anexar documentos de diversos formatos, como PDFs, imagens, dentre outros. Entretanto, apenas documentos nato-digitais HTML foram considerados nesta atividade.

Como resultado, o *sei.html.parser* estrutura as informações nos formatos JSON e CSV, adequados para realização de análises e estatísticas, que são disponibilizados no portal[3] da plataforma para serem utilizados pelos cidadãos interessados. Além disso, é possível estruturar um processo eletrônico a partir da sua respectiva URL. O *sei.html.parser* disponibiliza uma Web API que pode ser utilizada para converter um determinado processo em CSV ou JSON através de requisições HTTP.

[2]https://go-fair-agro.github.io
[3]https://go-fair-gov.github.io

Figura 3. Exemplo de conversão CSV para RDF

3.2. Etapa 2: Semântica e Dados Conectados

Converter processos eletrônicos, originalmente semi-estruturados, em arquivos CSV e JSON com estrutura homogenia possibilita uma utilização mais ampla dos dados. Entretanto, a plataforma vai além e define explicitamente o significado dos dados por meio de técnicas da Web Semântica. Primeiramente, foi definida uma ontologia[4] para representar conceitualmente o domínio de processos eletrônicos. Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada [Studer et al. 1998] e, atualmente, é utilizada para organização e classificação do conhecimento em sistemas de informação. Muitos trabalhos estão disponíveis na literatura sobre construção de ontologias para administração pública e dados abertos governamentais [Pinto e Almeida 2020], [Ferneda et al. 2016] e [Mohamad et al. 2022]. A ontologia desenvolvida inspira-se na ontologia OGDPub [Pereira e Todesco 2020], que propõe um esquema de classificação de dados em linguagem compreensível ao cidadão.

A definição explícita da semântica aos dados é realizada pela aplicação *csv2rdf* [Salvadori et al. 2019], que converte arquivos CSV em triplas RDF. A partir do arquivo CSV resultante da etapa anterior, da ontologia de domínio e de um arquivo de mapeamento entre as colunas CSV e os termos definidos na ontologia, a aplicação *csv2rdf* é capaz de converter os dados tabulares em triplas RDF conforme a especificação da ontologia utilizada, como mostra a Figura 3. Por fim, como demonstrado na Figura 2-(2) o arquivo RDF gerado é carregado em uma instância do Apache Jena Fuseki, um banco de dados semântico (*Triple Store*) que gerencia dados RDF e disponibiliza uma interface SPARQL para consultas.

A conversão HTML em JSON (Etapa 1) poderia ser armazenada em bancos de dados NoSQL orientados a documentos. Dessa forma, dados estruturados poderiam ser disponibilizados via Web APIs ou Serviços Web para utilização de aplicações externas. Apesar de proporcionar um nível muito mais elevado de interoperabilidade, a disponibilização de dados estruturados em JSON não suporta a descrição explícita da semântica. Como alternativa, poderiam ser adotados bancos de dados orientados à grafos. O *neo4j*[5] é um exemplo de banco de dados orientados a grafos que suporta dados semânticos e que está sendo adotado como tecnologia de persistência de dados semânticos em diferentes domínios, como extração de petróleo [Gong et al. 2018] e cultura [Drakopoulos et al. 2019].

[4]https://go-fair-gov.github.io/ontology/e-process/
[5]https://neo4j.com/

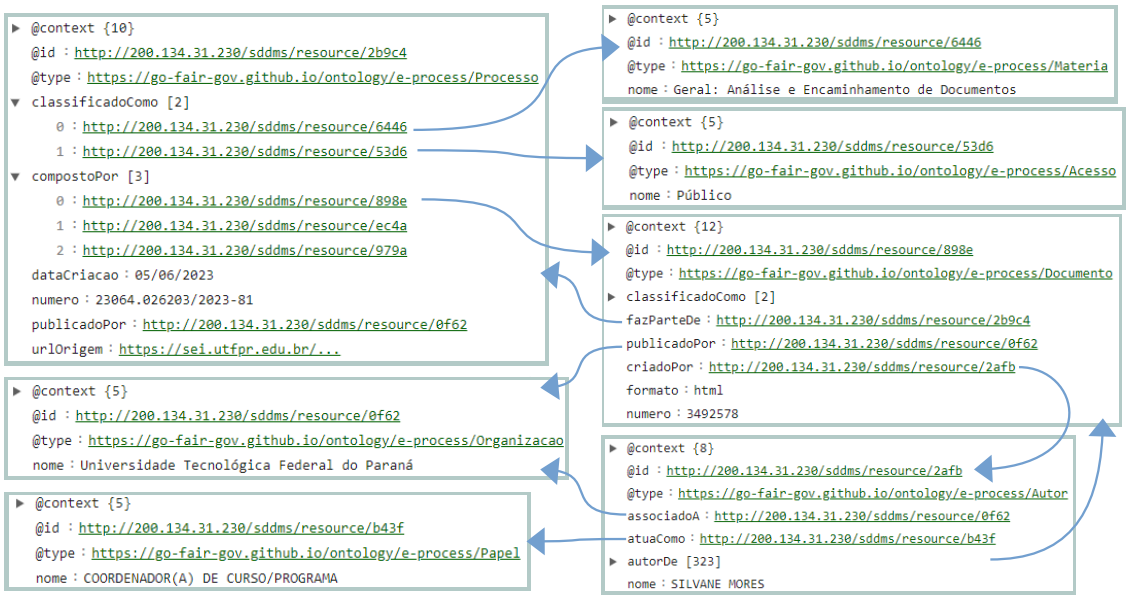


Figura 4. Representações JSON-LD interconectadas

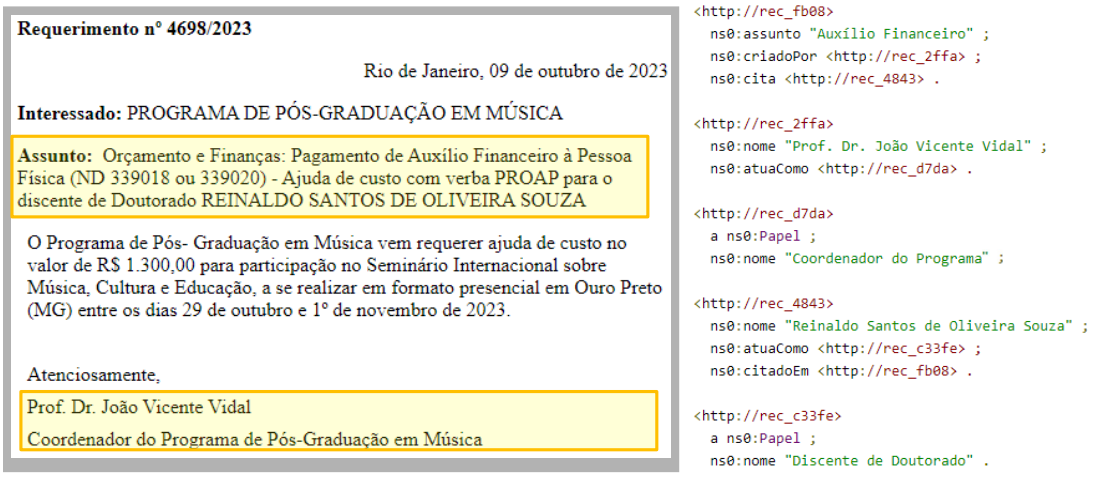
3.3. Etapa 3: Acesso aos Dados

Interfaces SPARQL permitem consultas elaboradas e precisas sobre os dados RDF. Entretanto, aplicações Web modernas utilizam outras formas para acessar fontes de dados, sendo mais comum o acesso por meio de Web APIs. Como discutido anteriormente, a disponibilização de arquivos CSV ou JSON apenas confere aos dados uma representação estrutural homogênea, mas não suporta descrição explícita do seu significado. A disponibilização dos dados na forma de grafos suporta descrição semântica, mas não realiza a interconexão de dados relacionados. Dessa forma, é necessário adotar tecnologias capazes de interligar dados relacionados por *links*.

A plataforma, além de disponibilizar arquivos CSV, JSON, RDF e consultas SPARQL, também disponibiliza uma interface de acesso REST. Para isso, utiliza-se a aplicação *sddms* [Salvadori et al. 2019], capaz de criar links entre recursos relacionados e disponibilizar representações semânticas JSON-LD através de chamadas HTTP. Dessa forma, o *sddms* simplifica o acesso aos dados, convertendo chamadas HTTP REST em consultas SPARQL. A Figura 4 ilustra uma representação JSON-LD de um processo eletrônico enriquecido semanticamente e interconectado por *links*. Por exemplo, um processo pode referenciar sua classificação e documentos anexados, que por sua vez, referenciam autores e organizações associadas. Essa cadeia de referências cria uma rede interconectada de recursos que permite a exploração de informações através da navegação entre recursos relacionados.

3.4. Etapa 4: Análise de Dados

As informações mais relevantes sobre os processos eletrônicos estão contidas nos documentos anexados. Por se tratar de dados semi-estruturados, é possível obter, via raspagem, poucas informações sobre o documento. Entretanto, os assuntos associados aos documentos são descritos de forma livre e sem uma categorização adequada, fato que prejudica a



Requerimento nº 4698/2023

Rio de Janeiro, 09 de outubro de 2023

Interessado: PROGRAMA DE PÓS-GRADUAÇÃO EM MÚSICA

Assunto: Orçamento e Finanças: Pagamento de Auxílio Financeiro à Pessoa Física (ND 339018 ou 339020) - Ajuda de custo com verba PROAP para o discente de Doutorado REINALDO SANTOS DE OLIVEIRA SOUZA

O Programa de Pós- Graduação em Música vem requerer ajuda de custo no valor de R$ 1.300,00 para participação no Seminário Internacional sobre Música, Cultura e Educação, a se realizar em formato presencial em Ouro Preto (MG) entre os dias 29 de outubro e 1º de novembro de 2023.

Atenciosamente,

Prof. Dr. João Vicente Vidal
Coordenador do Programa de Pós-Graduação em Música

(a)

```
<http://rec_fb08>
  ns0:assunto "Auxílio Financeiro" ;
  ns0:criadoPor <http://rec_2ffa> ;
  ns0:cita <http://rec_4843> .

<http://rec_2ffa>
  ns0:nome "Prof. Dr. João Vicente Vidal" ;
  ns0:atuaComo <http://rec_d7da> .

<http://rec_d7da>
  a ns0:Papel ;
  ns0:nome "Coordenador do Programa" ;

<http://rec_4843>
  ns0:nome "Reinaldo Santos de Oliveira Souza" ;
  ns0:atuaComo <http://rec_c33fe> ;
  ns0:citadoEm <http://rec_fb08> .

<http://rec_c33fe>
  a ns0:Papel ;
  ns0:nome "Discente de Doutorado" .
```

(b)

Figura 5. Sumarização e identificação de pessoas mencionadas em texto livre: (a) Documento nato-digital SEI; (b) Novas triplas RDF geradas

consulta. A Figura 5 mostra um documento cujo assunto é descrito em texto livre e representado por dezenas de palavras. Apesar de a pesquisa pública permitir a consulta de documentos filtrados pelo assunto, na prática, essa pesquisa traz resultados insatisfatórios. Observou-se a ausência de assinatura digital em alguns documentos, sendo impossível identificar, via raspagem, o autor. Nesses casos, a autoria é descrita em texto livre, pela presença do nome e do papel desempenhado pelo autor no final do texto. Além disso, outras pessoas são frequentemente mencionadas nos documentos, cuja referência se dispersa ao longo do texto.

Técnicas baseadas em BERT [Devlin et al. 2019] são comumente utilizadas para processamento de linguagem natural [Bizer 2023]. Muitos trabalhos disponíveis na literatura adotam o BERT para segmentação, classificação, análise de polaridade (análise de sentimento) e identificação de tópicos. Neste contexto, foram propostas melhorias e modificações em modelos BERT para solução de uma variedade problemas, como, por exemplo, processamento de linguagem natural em textos curtos [Amorim et al. 2022], modelos adaptados para informações publicadas em diário oficial [Constantino et al. 2022], e de forma mais abrangente, modelos adaptados para idiomas específicos. A exemplo do BERTimbau [Souza et al. 2020], que disponibiliza modelos BERT pré-treinados na língua portuguesa. Apesar de sua grande adoção, modelos baseados em BERT exigem um significativo esforço em etapas de treinamento. Sendo assim, autores afirmam que modelos que empregam técnicas *Generative Pre-training Transformer* (GPT) podem ser mais eficientes e robustos que os baseados em BERT [Bizer 2023].

A plataforma proposta inclui uma etapa de análise de dados, capaz de reclassificar documentos de forma que o assunto represente de fato os seus conteúdos. Para isso, foi desenvolvido o sistema *sei.data.analysis*, que submete o inteiro teor do documento para a OpenAI API[6] a fim de reclassificar o documento. Adotou-se o modelo *gpt-3.5-turbo-0125*[7], adequado para ser utilizado uma ampla gama de aplicações de processamento

[6]https://platform.openai.com/docs/api-reference/chat
[7]https://platform.openai.com/docs/models/gpt-3-5-turbo

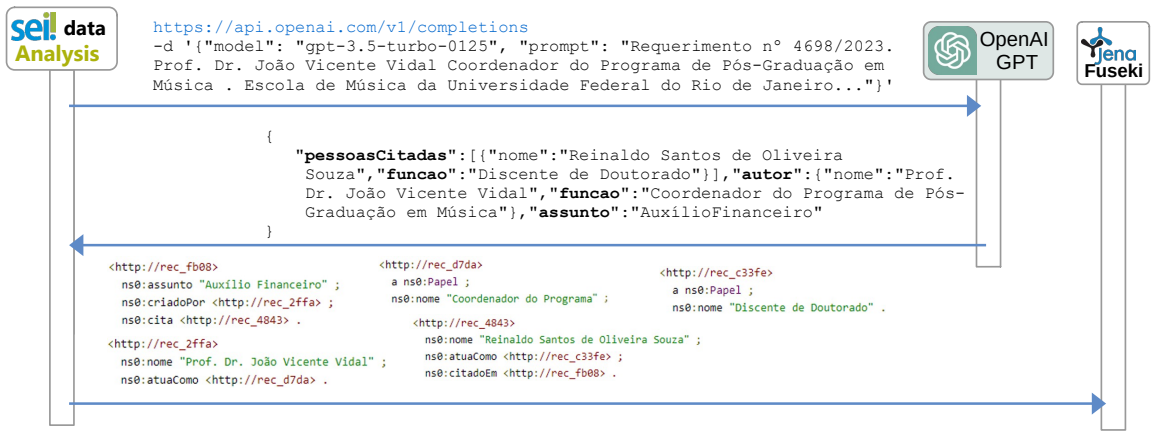


Figura 6. Fluxo de operações da etapa de análise de dados

de linguagem natural. Além disso, utiliza-se a OpenAI API para identificar os nomes das pessoas mencionadas no texto, o autor e seus respectivos papéis. A Figura 6 ilustra o fluxo de operações desta etapa. Primeiramente, o sistema *sei.data.analysis* submete uma requisição HTTP para OpenAI API com o inteiro teor do documento. A resposta, representada no formato JSON, é convertida em novas triplas RDF que são adicionadas ao *Triple Store* Apache Fuseki, complementando as informações processuais obtidas nas etapas anteriores.

4. Cenário de Aplicação

Para avaliar as técnicas e ferramentas de análise e publicação de dados de processos eletrônicos, foi escolhido o cenário das universidades federais brasileiras. O SEI foi instituído por meio da Portaria nº 1.042, do dia 4 de novembro de 2015 do Ministério da Educação, como o sistema oficial de informações, documentos e processos eletrônicos das entidades ligadas ao Ministério. Neste estudo, foram consideradas as seguintes universidades: Universidade de Brasília (UnB), Universidade Federal de Goiás (UFG), Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal de Minas Gerais (UFMG) e Universidade Tecnológica Federal do Paraná (UTFPR). Os processos eletrônicos dessas universidades foram obtidos através da pesquisa pública e submetidos para a plataforma de acordo com o fluxo descrito na Figura 2.

Neste cenário de aplicação[8], foram considerados os processos eletrônicos de 2023. A Tabela 2 apresenta o volume de dados correspondente a cada instituição. Destaca-se a UnB como a instituição com o maior número de processos, mas com a menor quantidade de documentos. Outro aspecto importante é o volume necessário para representar os dados processuais em diversos formatos. Arquivos RDF[9] são cerca de 422% maiores que arquivos no formato CSV. Conforme o fluxo de execução da plataforma, os arquivos RDF são carregados e indexados em um *Trile Store* para fins de desempenho em consultas (TDB), duplicando praticamente seu tamanho original.

A Tabela 3 mostra o número de indivíduos de cada classe definida na ontologia. Nota-se o número extremamente elevado de assuntos. A UFRJ, por exemplo, classifica

[8]Instruções para reprodução do experimento em https://figshare.com/s/9e35eb8352fb009e5ddc
[9]Materialização RDF-*Turtle*.

Tabela 2. Volume de dados do cenário de aplicação

Instituição	Processos	Documentos	Triplas	CSV	RDF	TDB
UnB	141.287	432.891	5.890.895	328 MB	1.27 GB	2.42 GB
UFG	59.166	591.694	6.350.872	312 MB	1.26 GB	2.51 GB
UFRJ	65.043	853.862	9.713.128	453 MB	2.02 GB	3.80 GB
UFMG	75.656	703.804	7.617.452	328 MB	1.51 GB	2.97 GB
UTFPR	51.711	470.934	5.400.024	273 MB	1.12 GB	2.18 GB

Tabela 3. Indivíduos por classe definidas na ontologia

Classes	UnB	UFG	UFRJ	UFMG	UTFPR
Processo	141.287	59.166	65.043	75.656	51.711
Documento	432.891	591.694	853.862	703.804	470.934
Assunto	32.599	12.544	81.766	16.031	36.812
Autor	10.448	0	8.079	470	5.227
Papel	2.146	0	248	81	149
Matéria	562	497	397	439	366
Acesso	8	16	14	15	10
Total	619.941	663.917	1.009.409	796.496	565.209

documentos em mais de 80 mil assuntos, inviabilizando a consulta por este parâmetro. Observa-se que a UFG não possui autores e papéis, pois todos os documentos publicados são de acesso restrito, permitindo assim a obtenção apenas dos metadados. De forma semelhante, a UFMG restringiu[10] o acesso a maior parte dos documentos. Dentre os mais de 700 mil documentos, apenas 4.386 são de acesso público. Nas demais instituições, observa-se um elevado número de autores. Como exemplo, a UnB registra mais de 10 mil autores, fato que indica um uso generalizado do SEI entre os servidores.

A interligação entre as informações processuais é feita por meio das relações estabelecidas na ontologia. A Tabela 4 mostra o número de relações entre as informações processuais. Cada relação é expressa por uma tripla RDF, logo, as relações representam cerca de 50% dos dados. Observou-se que a quantidade média de documentos produzidos por cada autor e a média de documentos anexados em cada processo foram de 20,95 e 8,92 documentos, respectivamente. O grau da relação *criadoPor* mostra que 69% dos documentos não possuem autoria definida, causado pela ausência de assinatura digital ou por se tratar de documentos PDF ou imagens.

O estudo avaliou a eficiência do modelo *gpt-3.5-turbo-0125* em documentos eletrônicos das universidades federais. Foi examinada a capacidade do modelo identificar a presença de pessoas mencionadas nos documentos, bem como de determinar seus autores. Realizou-se análise manual de 100 documentos, selecionados aleatoriamente e igualmente distribuídos entre as instituições[11] de ensino. Os documentos foram classificados de acordo com o conteúdo, incluindo atas de reuniões, requisições e declarações, assuntos financeiros e comunicações por e-mail.

[10]Acesso restrito provisoriamente em razão da necessidade de reclassificação de nível de acesso.

[11]UFG foi descartada pois não disponibiliza nenhum documento público.

Tabela 4. Relacionamento entre os dados

Relação	UnB	UFG	UFRJ	UFMG	UTFPR
associadoA	10.448	0	8.079	470	5.227
atuaComo	10.448	0	8.079	470	5.227
autorDe	229.898	0	266.108	1.684	114.054
criadoPor	229.898	0	266.108	1.684	114.054
compostoPor	432.891	591.694	853.862	703.804	470.934
fazParteDe	432.891	591.694	853.862	703.804	470.934
publicadoPor	574.178	650.860	918.905	779.460	522.645
classificadoComo	1.146.393	1.291.007	1.819.213	1.558.915	1.044.937
Total	2.633.154	3.125.255	4.994.216	3.750.291	2.748.012

Tabela 5. Efetividade do modelo *gpt-3.5-turbo-0125*

Instituição	Pessoas citadas			Autoria		
	Precisão	Acurácia	Revocação	Precisão	Acurácia	Revocação
Atas	0,81	0,83	0,96	0,65	0,50	0,85
Req.&Decl.	1,00	0,94	0,93	1,00	0,70	0,83
Financeiro	0,60	0,60	1,00	0,60	0,70	0,86
Email	0,62	0,60	0,71	0,71	0,50	0,62

A Tabela 5 mostra o desempenho do modelo nos diferentes tipos de documentos. Embora apresente bom desempenho na detecção de pessoas mencionadas, a capacidade de identificar os autores é limitada em atas de reuniões. Estes documentos possuem autoria coletiva e alta ocorrência de citação de pessoas. Em contrapartida, teve bom desempenho em requisições e declarações, mas teve dificuldade ao lidar com documentos financeiros, associando frequentemente nomes de empresas e suas respectivas razões sociais a indivíduos. Troca de mensagens via e-mail são frequentemente anexadas aos processos. Neste tipo de documento, o modelo tende a interpretar incorretamente o nome de pessoas com base em endereços de e-mail.

Por fim, avaliou-se a capacidade de sumarização de texto. Originalmente, identificou-se que os documentos foram classificados de acordo com três padrões. De forma geral, adotam a natureza do documento, como *parecer*, *ofício* ou *despacho*, com ocorrências de acréscimo de um identificador interno, ou sumarização extremamente longa e específicas, sem possibilidade de reúso. Sendo assim, a classificação é prejudicada tanto pelo generalismo quanto pela especialização dos assuntos. Observou-se que o modelo *gpt-3.5-turbo-0125* foi capaz de reclassificar adequadamente os documentos com base em seus conteúdos.

5. Conclusões e Trabalhos Futuros

Este trabalho abordou a adoção de metodologias e ferramentas existentes para viabilizar a interoperabilidade de dados entre sistemas, e assim ampliar o consumo e reuso de dados em um cenário real. Por meio de uma plataforma, que reúne uma coleção de ferramentas que executam tarefas em um determinado fluxo de execução, foi possível converter dados semi-estruturados de processos eletrônicos em uma rede interconectada de informações

semanticamente descritas. Como resultado, conclui-se que o uso combinado dessas ferramentas aprimora a qualidade das informações e, consequentemente, permite que sejam reutilizadas por mais pessoas e sistemas computacionais para finalidades diversas.

Além disso, este trabalho apresentou um panorama da utilização do SEI em algumas organizações públicas. Demonstrou o custo, em termos do volume de dados, para converter informações semi-estruturadas em representações semânticas em diversos formatos RDF, bem como o custo de indexação e armazenamento em um *Triple Store*. Sendo assim, demonstrou-se que a estruturação, enriquecimento semântico e indexação podem resultar em banco de dados dez vezes maiores. Por fim, o modelo generativo para processamento de linguagem natural *gpt-3.5-turbo-0125* mostrou-se adequado para reclassificação de assuntos e identificação de autores e pessoas mencionadas em documentos no cenário de processos eletrônicos, embora haja limitações em alguns tipos de documentos específicos.

Este trabalho reafirma a importância do Sistema Eletrônico de Informações como uma ferramenta fundamental na modernização administrativa das organizações públicas brasileiras. Os resultados obtidos pela adoção das metodologias e ferramentas para manipulação e tratamento de dados existentes, organizadas de acordo com o fluxo de execução proposto pela plataforma indicam um caminho promissor para aprimorar ainda mais a gestão documental e processual das organizações públicas. Trabalhos futuros podem adotar a plataforma proposta em cenários distintos, identificando as limitações e pontos de adaptações. Este trabalho avaliou apenas um modelo GPT, entretanto, outros modelos podem ser avaliados e ter seus resultados comparados para, assim, permitir a escolha mais adequada para os mais diversos objetivos e cenários de aplicação. Se faz necessária uma avaliação quantitativa com um corpus textual maior de documentos para identificar o potencial da sumarização em agrupar assuntos em categorias ou tópicos.

Referências

Albuquerque, A. and Dorneles, C. (2022). Dados escuros à luz do controle público. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 78–89, Porto Alegre, RS, Brasil. SBC.

Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D., and Paes, A. (2022). Modelagem de tópicos em textos curtos: uma avaliação experimental. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*.

Bizer, C. (2023). Gpt-4 Versus Bert: Which Foundation Model Is More Suitable for Integrating Data from the Web? In *Proceedings of the 19th International Conference on Web Information Systems and Technologies, WEBIST 2023, Rome, Italy*.

Constantino, K., Cruz, V. A. L., Zucheratto, O. M. M., França, C., Carvalho, M., Silva, T. H. P., Laender, A. H. F., and Gonçalves, M. A. (2022). Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 304–316.

da Silva, M. C. G., Donato, J. A., and Cardoso, L. G. (2018). O cenário arquivístico na implantação do Sistema Eletrônico de Informações (SEI) nos ministérios federais brasileiros. *RACIn - Revista Analisando em Ciência da Informação*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *NAACL-HLT (1)*. Association for Computational Linguistics.

Drakopoulos, G., Spyrou, E., Voutos, Y., and Mylonas, P. (2019). A semantically annotated json metadata structure for open linked cultural data in neo4j. In *Proceedings of the 23rd Pan-Hellenic Conference on Informatics*, PCI '19, page 81–88, New York, NY, USA. Association for Computing Machinery.

Ferneda, E., Cruz, F. W., do Prado, H. A., Guadagnin, R. d. V., dos Santos, L. C., dos Santos, D. L. N., and da Costa, O. L. (2016). Potential of ontology for interoperability in e-government: discussing international initiatives and the brazilian case. *Brazilian Journal of Information Science: research trends*, 10(2).

Filho, S. S. L. and Peixe, B. C. S. (2017). Estudo da eficiência na execução da despesa pública com materialde expediente face a adoção ao sistema eletrônico de informações em órgãos públicos federais. In *Congresso Brasileiro De Custos*.

Gong, F., Ma, Y., Gong, W., Li, X., Li, C., and Yuan, X. (2018). Neo4j graph database realizes efficient storage performance of oilfield ontology. *PLOS ONE*, 13(11):1–16.

Macedo, E. and Tolfo, R. (2017). Do processo eletrônico ao documento público: uma análise da conservação dos autos como arquivos permanentes. *Revista Eletrônica do Curso de Direito da UFSM*, 12:709.

Mello, R. d. S., eles, C. F., Kade, A., Braganholo, V., and HEUSER, C. A. (2000). Dados semi-estruturados. *XV Simpósio Brasileiro de Banco de Dados*.

Mohamad, A., Sylvester, A., and Campbell-Meier, J. (2022). Towards a taxonomy of emerging topics in open government data: A bibliometric mapping approach. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Honolulu, HI 96822. Hamilton Library.

Pereira, L. and Todesco, J. (2020). Ogdpub - uma ontologia para publicação de dados abertos governamentais. In *Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico*, pages 72–83, Porto Alegre, RS, Brasil. SBC.

Pinto, J. A. and Almeida, M. B. (2020). Ontologias públicas sobre governo eletrônico: Uma revisão sistemática da literatura. *Brazilian Journal of Information Science: research trends*, 14.

Rodrigues, D. R. and Cavalcante, S. M. d. A. (2018). Acesso à longo prazo de documentos arquivísticos: os impactos da adesão ao sistema eletrônico de informações (sei) na universidade federal do ceará. *RACIn - Revista Analisando em Ciência da Informação*.

Salvadori, I. L., Huf, A., and Siqueira, F. (2019). Semantic data-driven microservices. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 402–410.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data and Knowledge Engineering*, 25(1):161–197.