




**GAZETTEER LITERÁRIO DE MACHADO DE ASSIS**

Literary Gazetteer of Machado de Assis

**Dilvan de Abreu Moreira**Universidade de São Paulo,  
Instituto de Ciências Matemáticas e de Computação,  
São Carlos, SP, Brasil  
dilvan@icmc.usp.br  
<https://orcid.org/0000-0002-4801-2225> **Davi Machado da Rocha**Secretaria da Educação do Estado de São Paulo,  
E.E. Prof. José Juliano Neto,  
São Carlos, SP, Brasil  
davimachado@prof.educacao.sp.gov.br  
<https://orcid.org/0009-0004-3326-6881> A lista completa com informações dos autores está no final do artigo **RESUMO**

**Objetivo:** Este estudo tem o objetivo de desenvolver uma aplicação web semântica que mapeia localidades geográficas nas obras de Machado de Assis, armazenando-as em uma triplestore. A partir da integração dos dados disponibilizados pela enciclopédia MachadodeAssis.net com as coordenadas geográficas de Geonames.org e GoogleMaps, o projeto visa oferecer uma experiência de leitura através de mapas interativos que servirão de suporte para as menções aos espaços realizadas pelo escritor ao longo do Século XIX.

**Método:** Utiliza a biblioteca Python BeautifulSoup para consultas e coleta dos dados da enciclopédia, estruturando-os de acordo com os parâmetros do schema.org. As citações coletadas são submetidas aos modelos gpt3.5-instruct e gpt4-turbo para obter os nomes atuais das localidades e a devida classificação destes espaços de acordo com a ontologia Geonames.org. Consultas SPARQL são realizadas ao portal dados.literaturabrasileira.ufsc.br para obter identificadores únicos para cada livro.

**Resultado:** A aplicação oferece uma integração entre mapas, citações e textos completos, em consonância com os padrões Linked Data.

**Conclusões:** A intersecção entre tecnologia, literatura e geolocalização pode oferecer experiências de leitura alternativas, proporcionando um terreno fértil para o desenvolvimento das chamadas humanidades digitais.

**PALAVRAS-CHAVE:** Web Semântica. Machado de Assis. Geolocalização. Linked Data. Humanidades Digitais.

**ABSTRACT**

**Objective:** This study aims to develop a semantic web application that maps geographic locations in the works of Machado de Assis, storing them in a triplestore. By integrating data from the MachadodeAssis.net encyclopedia with geographic coordinates from Geonames.org and GoogleMaps, the project aims to offer a reading experience through interactive maps that support the spatial references made by the writer throughout the 19th century.

**Methods:** The Python library BeautifulSoup is used for querying and collecting data from the encyclopedia, structuring it according to schema.org parameters. The collected citations are submitted to the gpt3.5-instruct and gpt4-turbo models to obtain the current names of the locations and the proper classification of these spaces according to the Geonames.org ontology. SPARQL queries are performed on the dados.literaturabrasileira.ufsc.br portal to obtain unique identifiers for each book.

**Results:** The application offers an integration between maps, citations, and full texts, in line with Linked Data standards.

**Conclusions:** The intersection of technology, literature, and geolocation can offer interesting reading experiences, providing fertile ground for the development of so-called digital humanities.

**KEYWORDS:** Semantic Web. Machado de Assis. Geolocation. Linked Data. Digital Humanities.

**1 INTRODUÇÃO**

Machado de Assis, um dos mais ilustres escritores do Brasil oitocentista, é reconhecido como uma figura central na literatura em língua portuguesa. Sua obra abrange

romances, poesias, peças teatrais e crônicas, marcando profundamente o cenário literário com seu estilo inconfundível e análises perspicazes da sociedade brasileira da época. Atualmente, na área das humanidades, tem se discutido a trajetória ascendente do escritor como expressão do potencial criativo das populações afrodescendentes no Brasil, cujo legado para a formação intelectual do povo brasileiro é muitas vezes negado em razão do histórico escravista que marca indelevelmente a nossa História. Cabe considerar ainda a importância do escritor para o registro dos padrões de sociabilidade e mesmo dos acontecimentos históricos ocorridos durante a sua existência, que acompanha momentos decisivos da História do Brasil, tais como a transição da Monarquia para a República, a Guerra do Paraguai e o processo de abolição do trabalho escravo.

Em razão dessa centralidade, atualmente, há diversos projetos que disponibilizam sua obra em formatos digitais variados, contribuindo para ampliar o alcance de seu legado. Na área das Tecnologias da Informação, seus escritos têm servido ainda para o treinamento de modelos de Machine Learning, chatbots e outros usos, como pode ser verificado no dataset disponível na plataforma Kaggle que reúne suas obras completas em formato digital, por se tratar de um conjunto que oferece uma escrita apurada e condizente com os mais elevados padrões da língua portuguesa, para além dos fatores já mencionados.

Assim, na esteira dos projetos de tecnologia que encontram a obra do Bruxo do Cosme Velho, o presente trabalho consiste na construção de uma triplestore de localidades presentes em sua obra, disponibilizadas em forma de verbetes pelo portal machadodeassis.net. Uma triplestore é um tipo de banco de dados projetado especificamente para armazenar e gerenciar dados organizados em triplas RDF (Resource Description Framework). Cada tripla representa uma relação entre recursos na forma de sujeito-predicado-objeto, permitindo que informações estruturadas sejam armazenadas e consultadas de maneira eficiente. Diferente de bancos de dados relacionais, uma triplestore é otimizada para consultas complexas sobre dados conectados, frequentemente utilizando SPARQL como linguagem de consulta. Essa arquitetura é amplamente usada na Web Semântica e em aplicações de Linked Data, pois facilita a integração e interoperabilidade de dados entre sistemas distintos (Bizer; Heath; Berners-Lee, 2011).

A partir da coleta dos verbetes no portal machadodeassis.net – uma iniciativa assinada pela pesquisadora Marta de Senna e que conta com fontes variadas de financiamento – que se autodefine como uma "enciclopédia" ou "dicionário" especializado na obra do escritor, procedemos a identificação de coordenadas geográficas em pares de

latitude e longitude dos locais referenciados e descritos no site a partir das bases do Geonames.org e do GoogleMaps, propondo outras experiências de leitura através de mapas interativos, que servem, em última análise, como suporte para excertos da obra machadiana.

Para a coleta dos dados, utilizamos a biblioteca BeautifulSoup em Python, apontando as tags HTML da página com os nomes, os verbetes e as citações sobre um dado local em livros diversos. Além disso, armazenaremos o resultado da coleta respeitando os parâmetros do schema.org, pensando em facilitar a localização desses registros pelo núcleo de sistemas de busca do Google, o que pode proporcionar, por exemplo, o melhor ranqueamento dos registros geográficos nesse buscador. Por fim, por se tratar de um tipo de dado específico (entidade geográfica) também usaremos a ontologia do Geonames para a descrição dos objetos e classes, oferecendo uma organização mais completa em termos semânticos e maior perenidade dos registros, considerando eventuais mudanças de domínio, por exemplo.

Do ponto vista técnico, cumpre considerar ainda que a aplicação desenvolvida neste estudo está disponível para acesso público na plataforma Hugging Face (<https://huggingface.co/spaces/histlearn/MachadodeAssis>) e, para facilitar a colaboração e desenvolvimento contínuo, o código-fonte completo do projeto está disponível em repositório público no GitHub (<https://github.com/rocdav/MachadodeAssis>), permitindo que pesquisadores e desenvolvedores possam contribuir com melhorias e adaptações. A implementação seguiu as diretrizes e melhores práticas estabelecidas na documentação oficial da OpenAI (Openai, 2023), especialmente no que diz respeito à configuração e uso dos modelos de linguagem.

No contexto dos estudos literários e das tecnologias da informação, este trabalho propõe uma aplicação que explora a geografia literária na obra de Machado de Assis, oferecendo uma experiência de leitura imersiva que combina literatura e geolocalização. Construída sobre uma triplestore de locais referenciados nas obras do autor, a aplicação visa mapear digitalmente o universo machadiano, destacando como o escritor retrata o espaço urbano e os ambientes que circundam seus personagens. Ao integrar dados do portal *machadodeassis.net* com coordenadas de *Geonames.org* e *GoogleMaps*, esta plataforma permite que leitores e pesquisadores interajam com os locais citados nos textos, enriquecendo a compreensão dos espaços culturais e sociais descritos por Machado. Assim, ao transformar as referências geográficas da obra em mapas interativos, este

projeto explora a narrativa literária sob um viés espacial, ampliando o acesso ao legado de Machado de Assis e promovendo a interdisciplinaridade entre humanidades digitais e tecnologias de georreferenciamento.

No que diz respeito aos usos possíveis da solução proposta, é relevante destacar que as relações entre espaço geográfico e análise literária ocupam um papel importante na produção acadêmica, tanto nacional quanto internacional. Entre as obras que exploram essa interface, destaca-se o trabalho do pesquisador Franco Moretti, em *Atlas do Romance Europeu*, no qual ele observa que:

os mapas literários nos permitem ver duas coisas basicamente: em primeiro lugar, realçam o *ortsgebunden*, a natureza espacial das formas literárias: cada uma delas com sua geometria peculiar, suas fronteiras, seus tabus espaciais e rotas favoritas. Em seguida, os mapas trazem à luz a lógica interna da narrativa: o domínio semiótico em torno do qual um enredo se aglutina e se organiza. (Moretti, 2003, p.15).

Desse modo, ao introduzir conceitos espaciais como centro e periferia, entre outros, criam-se dimensões essenciais para a construção de personagens e a contextualização de certos hábitos e valores sociais que refletem as dinâmicas das cidades ao longo do tempo. Nesse contexto, a aplicação deste trabalho surge como um recurso valioso para pesquisadores que se interessam pelo espaço geográfico e social das obras literárias, assim como para os leitores do portal machadodeassis.net. Ao proporcionar uma experiência de leitura que utiliza mapas, distâncias, frequências de citações e outras métricas, promovemos uma interação alternativa com a obra de Machado de Assis através da cartografia literária.

Seguindo o que Moretti propõe sobre o papel dos mapas literários em revelar "a lógica interna da narrativa", o espaço urbano em *Dom Casmurro*, por exemplo, oferece ao leitor uma visão mais profunda da psicologia dos personagens e do cenário social do Brasil oitocentista. O Engenho Novo, onde vive o narrador, emerge como uma região periférica do Rio de Janeiro da época, um contraste em relação aos centros sociais e culturais da cidade. Ao situar Bentinho nesse espaço, Machado de Assis não apenas caracteriza o subúrbio, mas transforma-o em um reflexo da solidão e introspecção do protagonista. Esse afastamento físico do centro simboliza, de maneira mais ampla, uma alienação emocional, sugerindo um distanciamento de Bentinho em relação às convenções sociais e à vida comunitária.

Essa relação entre espaço e psicologia é acentuada na cena em que o narrador recorda o encontro com um poeta no trem, em seu trajeto de volta ao Engenho Novo. Ao

comentar: "Os vizinhos, que não gostam dos meus hábitos reclusos e calados, deram curso à alcunha, que afinal pegou" (Assis, 1994, p.2), Bentinho expõe como a geografia do Engenho Novo amplifica seu estado emocional. O subúrbio, aqui, assume a forma de uma "caverna" emocional, um espaço onde o protagonista se refugia do dinamismo social do centro, ao mesmo tempo que revive aspectos nostálgicos de sua vida.

A geografia e a subjetividade se conectam, permitindo que Machado de Assis use o espaço físico como uma metáfora do isolamento psicológico de Bentinho. Este trabalho, ao mapear essas localidades em uma triplestore, oferece uma plataforma interativa para que leitores e pesquisadores explorem como a ambientação em *Dom Casmurro* e em outras obras machadianas reforça aspectos centrais da narrativa. Ao integrar geografia e literatura, promovemos uma compreensão mais ampla da obra machadiana, possibilitando novas interpretações sobre a construção mútua entre espaço e identidade dos personagens, entre outras relações possíveis entre texto e espaço.

## 2 REFERENCIAL TEÓRICO

### 2.1 Geografia literária: fundamentos e intersecções com a tecnologia

A Geografia Literária emerge como um campo interdisciplinar que investiga as relações entre o espaço geográfico e a literatura, examinando como os textos literários não apenas representam, mas também influenciam a compreensão de lugares e identidades (Westphal, 2011). Antes da era digital, essa abordagem já permitia que a análise literária transcendesse a interpretação textual tradicional, buscando nos textos aquilo que se encontraria "fora" deles – as relações territoriais, contextuais e culturais que moldam as percepções dos espaços e, conseqüentemente, das identidades narrativas.

Em sua obra seminal *La Géocritique: Réel, Fiction, Espace*, Bertrand Westphal introduziu o conceito de "geocrítica", propondo uma abordagem multifocal para explorar as representações espaciais na literatura. Desse modo, a geocrítica tem permitido que estudiosos investiguem o espaço não como um mero cenário passivo, mas como um componente ativo que interage com o real e o imaginário, revelando como os contextos geográficos, históricos e culturais moldam a narrativa e influenciam a percepção dos leitores. Esse movimento busca, essencialmente, situar o espaço literário em uma rede de interações territoriais e sociais que ampliam e ressignificam a compreensão do mundo representado.

Com o advento das plataformas digitais, esses conceitos foram ressignificados, permitindo a criação de projetos de georreferenciamento literário, nos quais mapas digitais e tecnologias interativas revelam como os lugares descritos na literatura se relacionam com o espaço real, adicionando uma camada de visualização que reforça as conexões entre literatura e geografia. Nesse contexto, *Geocritical Explorations: Space, Place, and Mapping in Literary and Cultural Studies*, de Tally Jr. (2014), reforça a relevância da “cartografia literária” como uma ferramenta para traçar essas interseções, ao conceber mapas que refletem os percursos dos personagens e contextualizam os espaços literários em relação ao território e à história. Essas ferramentas, embora recentes, são sustentadas por uma teoria que precede a era digital, indicando a literatura como um ponto de convergência entre representações espaciais e percepções culturais complexas, enriquecendo a leitura e a análise crítica.

## 2.2 Humanidades digitais e georreferenciamento literário

A convergência entre métodos computacionais e análise literária materializa-se nas chamadas Humanidades Digitais, área interdisciplinar que oferece novos caminhos para o estudo de temas como literatura, história e cultura. Desde o final do século XX, com o aumento do uso de ferramentas digitais e a ampliação do acesso a grandes conjuntos de dados, os pesquisadores passaram a incorporar métodos como análise de redes, mapeamento geográfico e visualização de dados, enriquecendo as interpretações tradicionais. Obras como *A Companion to Digital Humanities* (Schreibman, Siemens e Unsworth, 2004) e *Digital Humanities* (Burdick et al., 2012) ajudaram a estabelecer as bases teóricas e metodológicas para o campo, definindo um espaço em que se investigam padrões e relações que ampliam a compreensão de fenômenos culturais e sociais. Essas abordagens têm sido particularmente frutíferas na análise espacial de textos e na criação de projetos de georreferenciamento literário, onde a relação entre narrativa e espaço é explorada com mais profundidade.

Projetos como *Mapping Shakespeare's London* (Johnson et al., 2019), que emprega tecnologias de SIG (Sistemas de Informação Geográfica) para mapear os lugares mencionados nas obras de Shakespeare, ilustram a aplicação prática desses conceitos. Outro exemplo, o *Literary Atlas of Europe* (Schmidt; Piatti, 2020), utiliza ontologias geográficas e mapas interativos para explorar a literatura europeia em uma perspectiva espacial. De maneira similar, o *Pelagios Project* (Simon et al., 2016) estabeleceu padrões

para anotação geográfica de textos históricos, criando uma rede interligada de referências espaciais em fontes antigas.

No contexto brasileiro, o projeto (Documenta Palmares, 2021) ilustra essa abordagem ao oferecer um mapeamento detalhado de locais históricos significativos para a resistência negra e a cultura afro-brasileira. Em sua seção mapas, o projeto facilita a visualização de áreas como mocambos, aldeias indígenas, vilas, arraiais militares, sesmarias e trajetos de expedições – referências que se destacam nas fontes históricas ligadas ao Quilombo dos Palmares. Esse mapeamento foca nos locais para os quais há registros geográficos documentados, listados na seção "Fontes".

Os dados estão projetados sobre bases cartográficas modernas, que delineiam o relevo regional e os limites dos estados e municípios contemporâneos. A localização de cada elemento é indicada por pontos, linhas e polígonos, além de ícones específicos, organizados em uma legenda que diferencia tipos de locais: aldeias, arraiais, campos ou sertões, expedições, freguesias ou vilas, mocambos ou cercas, a região dos Palmares e sesmarias, além dos caminhos das expedições. Cada local, trajeto ou área possui uma ficha descritiva que inclui o nome, uma breve explicação do elemento mapeado e critérios para a escolha de sua localização, além de links para fontes e referências usadas na verificação dos dados.

Esse recurso oferece uma visão mais ampla da história e literatura brasileiras, permitindo uma interação entre o espaço literário e real que amplia a memória territorial e as relações sociais, agregando complexidade à leitura do contexto afro-brasileiro.

### **2.3 Tecnologias semânticas e Modelos de Linguagem na análise literária**

A integração de tecnologias semânticas e modelos de linguagem na análise literária representa uma fronteira em expansão, permitindo o processamento de grandes volumes de texto com foco na classificação e identificação de entidades geográficas. (Gregory; Hardie, 2018) discutem o uso de ontologias geográficas e da web semântica para enriquecer a análise de textos literários, tornando-os acessíveis a sistemas de inteligência artificial. Paralelamente, (Zhang et al., 2022) exploram os desafios e as potencialidades do uso de modelos de linguagem como GPT para reconhecer e classificar entidades geográficas em contextos literários, uma técnica que facilita o mapeamento de lugares mencionados em obras literárias.

Esses modelos de linguagem, como os utilizados neste estudo, possibilitam identificar e atualizar os nomes de locais referenciados em obras literárias, tornando-se essenciais para a precisão dos mapas interativos de projetos como o presente. No Brasil, autores como (Santos, 2017) e (Marandola Jr., 2021) contribuem com reflexões sobre o espaço e a sociedade que são particularmente relevantes para a análise da obra machadiana, oferecendo um contexto teórico nacional que complementa as abordagens tecnológicas.

## **2.4 Desafios e Limitações**

A literatura especializada aponta desafios específicos no uso de tecnologias para análise literária espacial, particularmente quanto à precisão na identificação de lugares. Enquanto (Grover; Tobin, 2014) destacam a importância da validação e desambiguação na classificação de topônimos históricos, estudos recentes no contexto brasileiro têm evidenciado complexidades adicionais nesse processo.

A questão da variação toponímica emerge como um desafio central, uma vez que um mesmo local pode apresentar diferentes denominações ao longo do tempo, ou denominações idênticas podem referir-se a locais distintos. Esta complexidade é particularmente evidenciada nos estudos sobre toponímia no período colonial brasileiro, onde se observa tanto a co-ocorrência quanto a concorrência de diferentes nomes para um mesmo local (Santos, 2022).

Para enfrentar esses desafios, torna-se fundamental a adoção de uma abordagem historiográfica na análise toponímica, que permita contextualizar historicamente os nomes de lugares. Tal metodologia auxilia não apenas na compreensão das variações, mas também na identificação mais precisa dos locais mencionados, considerando suas transformações ao longo do tempo (Zamariano, 2016). Esta perspectiva é especialmente relevante ao empregar modelos de linguagem para identificação de locais, demandando uma camada adicional de validação histórica e contextual.

## **3 SOLUÇÃO PROPOSTA**

### **3.1. Visão geral**

A solução proposta consiste em uma plataforma digital que mapeia e exhibe interativamente as localidades geográficas mencionadas nas obras de Machado de Assis. Utilizando tecnologias de georreferenciamento e linguagens de marcação da Web Semântica, a aplicação organiza as informações espaciais e literárias em um formato de



fácil navegação. A partir dessa abordagem, o leitor pode explorar as relações entre a narrativa e os locais mencionados, ampliando a compreensão das obras machadianas e oferecendo novas perspectivas de leitura através de mapas interativos.

### 3.2 Arquitetura do sistema

A arquitetura do sistema é baseada em uma integração entre dados estruturados de geolocalização, processamento de linguagem natural e visualização interativa. A aplicação é construída em uma interface intuitiva, hospedada na plataforma Hugging Face Spaces, com visualizações e mapas interativos carregados através de iframes de diferentes relatórios hospedados no Netlify. Além disso, o sistema conta com um endpoint SPARQL no Apache Jena Fuseki, permitindo consultas diretas aos dados RDF armazenados em uma triplestore.

A coleta de dados foi realizada utilizando a biblioteca Python BeautifulSoup para extrair citações de localidades mencionadas no site *MachadodeAssis.net*. Esses dados são processados por modelos de linguagem (GPT-3.5-instruct e GPT-4) para validação e atualização dos nomes das localidades, e as coordenadas geográficas são obtidas a partir das APIs do Geonames.org e Google Maps. As informações são, então, estruturadas em formato JSON-LD e carregadas na Triplestore para consultas semânticas.

### 3.3 Caso de uso

Objetivo: O principal objetivo da aplicação é identificar e explorar espaços (localizações) mencionados na obra de Machado de Assis, permitindo que usuários pesquisem e visualizem essas localizações geograficamente. A solução é destinada a pesquisadores nas áreas de literatura e história, bem como a leitores leigos interessados em literatura.

Atores:

1. **Pesquisadores nas áreas de literatura e história:** profissionais que buscam explorar a geografia literária da obra de Machado de Assis de forma analítica e acadêmica.
2. **Usuário leigo interessado em literatura:** leitores que desejam enriquecer sua experiência com a obra machadiana através da visualização dos locais mencionados nos textos.

Interações:

1. **Entrar na página Web:** o usuário acessa a plataforma online onde estão disponíveis as funcionalidades de mapeamento e consulta.



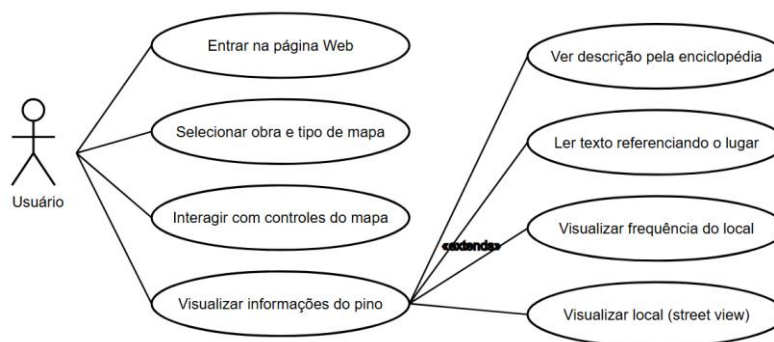
2. **Selecionar a obra e o tipo de mapa:** o usuário escolhe a obra de Machado de Assis que deseja explorar, bem como o tipo de visualização disponível (mapa de citações, mapa de calor, etc.).

3. **Interagir com os controles do mapa:** a plataforma oferece controles interativos que permitem ao usuário navegar pelo mapa, ampliar ou reduzir a área visualizada e explorar os pontos geográficos de interesse.

4. **Visualizar informações do pino:** ao selecionar um pino no mapa, o usuário obtém uma série de informações detalhadas sobre o local específico.

- **Descrição do espaço pela enciclopédia:** exibe uma breve descrição do local, extraída da enciclopédia *MachadodeAssis.net*, contextualizando a referência geográfica dentro da obra.
- **Texto referenciando o lugar:** mostra o excerto exato da obra de Machado de Assis onde o local é mencionado, oferecendo ao leitor uma conexão direta com o contexto literário.
- **Frequência que o local aparece na obra:** exibe um mapa de calor indicando a frequência com que o local é citado na obra, possibilitando uma análise da importância relativa de cada espaço dentro do enredo.
- **Visualização do local (Street View):** Embora inicialmente planejada, a implementação da visualização imersiva através da API do Google Street View não foi incluída na versão atual do sistema. Esta decisão foi tomada considerando a estrutura de custos associada ao serviço, que opera em um modelo de precificação baseado no volume de requisições. De acordo com a documentação atual do Google Maps Platform, cada carregamento de Street View é tarifado individualmente, podendo gerar custos significativos em um sistema com múltiplos usuários realizando consultas frequentes.

Figura 1: Diagrama de Caso de Uso.



Fonte: autores (2024)

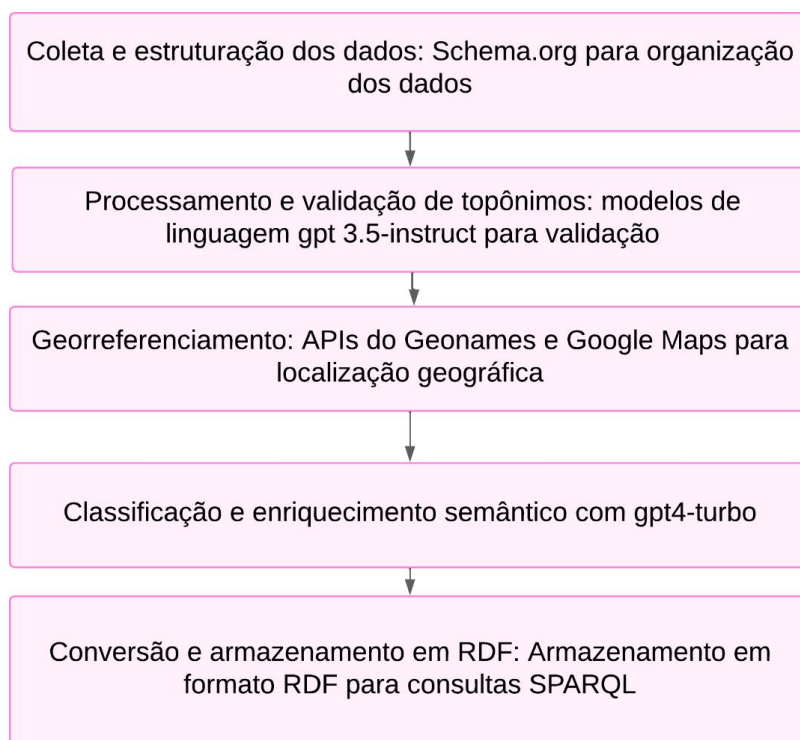
## 4 METODOLOGIA

Esta seção descreve o pipeline de desenvolvimento do gazetteer literário de Machado de Assis, detalhando desde a coleta e formatação dos dados até sua validação e enriquecimento com tecnologias de georreferenciamento e modelos de linguagem. O projeto, realizado no segundo semestre de 2023, empregou as tecnologias disponíveis à época para alcançar os objetivos de análise e visualização geoespacial dos textos machadianos.

### 4.1. Visão geral do pipeline de processamento

A metodologia foi organizada em cinco etapas principais, cada uma direcionada para um aspecto específico do processamento e análise dos dados literários, conforme ilustrado na Figura 2. Estas etapas compreendem a coleta inicial dos dados, seu processamento por modelos de linguagem, a obtenção de coordenadas geográficas, a classificação ontológica e, finalmente, a disponibilização em formato semântico para consulta.

Figura 2: Pipeline de Processamento



Fonte: autores (2024)

### 4.2 Coleta e Estruturação dos Dados

Schema.org é uma ontologia colaborativa, mantida pelo Google, Microsoft, Yahoo! e Yandex. Ela fornece um conjunto de vocabulários para descrever entidades,

relacionamentos e ações no mundo real. Esses vocabulários podem ser usados para enriquecer o conteúdo da web com metadados estruturados, que podem ser interpretados por máquinas. Os metadados estruturados podem ser usados para uma variedade de propósitos, incluindo:

a) melhorar os resultados de pesquisa: os motores de busca podem usar metadados estruturados para entender melhor o conteúdo de uma página da web. Isso pode levar a resultados de pesquisa mais relevantes e úteis para os usuários.

b) criar experiências mais ricas para os usuários: os metadados estruturados podem ser usados para entregar resultados de pesquisa enriquecidos com informações adicionais ou conteúdo interativo.

c) facilitar a automação: os metadados também podem facilitar a automação de tarefas, como a classificação de conteúdo ou a geração de relatórios.

Com isso em mente, o pipeline a seguir foi projetado para extrair e estruturar dados de lugares mencionados no site machadodeassis.net. O objetivo é criar um arquivo JSON-LD que contenha informações sobre os lugares e as obras que os mencionam. No início, são realizadas as importações das bibliotecas necessárias: requests (para fazer solicitações HTTP), JSON (para manipular dados JSON) e BeautifulSoup (uma biblioteca para análise de HTML). Também é definida uma URL específica do site, que é onde o script buscará os lugares referenciados. Assim, uma solicitação GET é feita para essa URL usando a biblioteca requests e o conteúdo da resposta é analisado pelo BeautifulSoup para facilitar a extração de informações.

O script procura todos os elementos Div que têm a classe "content-card-wrap" reference. Cada uma dessas divs representa um lugar referenciado no site. Para cada div encontrada, o título (nome do lugar) e o link associado são extraídos. Em seguida, uma nova solicitação GET é feita usando o link extraído para obter detalhes adicionais sobre o lugar. Do novo conteúdo carregado, o script extrai o Type-id e o Info-text, que fornecem informações adicionais sobre o lugar. Além disso, o código verifica se o título (lugar) já existe em um dicionário chamado structured\_data. Se não existir, uma nova entrada é criada para o título com detalhes como nome, URL e descrição. Ao final desta etapa, realiza-se a busca por informações sobre as obras (romances, contos, crônicas...) associadas a esse lugar. Para cada obra encontrada, são extraídos o título da obra, o tipo, o ano de publicação e o texto referenciado.

- Lugar
  - Nome (título)
  - Descrição
  - Lista de títulos das obras que referenciam esse lugar
- Obra
  - Nome (título)
  - Ano
  - Gênero
  - Citações do lugar no livro

Estas informações da obra são então estruturadas como CreativeWork (conforme o schema.org) e adicionadas à lista subjectOf da entrada correspondente do lugar em structured\_data. A propriedade "subjectOf" é uma inversão da propriedade "about", e refere-se ao item que é o assunto da coisa (por exemplo, um Place que é o assunto de um CreativeWork). Finalmente, o script salva todos os dados estruturados no formato JSON-LD em um arquivo chamado 'structured\_data.jsonl'.

Para a definição dos identificadores dos livros em que os locais são citados, faremos a coleta das URL's das obras no portal <http://dados.literaturabrasileira.ufsc.br/> a partir de uma query SPARQL, pensando também em facilitar o acesso aos textos integrais pelos leitores. Em uma situação ideal, o @id seria o ISBN do livro, no entanto, por razões que envolvem a complexidade desse tipo de coleta, considerando o volume de obras e a quantidade de edições e reedições ao longo do tempo, a própria definição do ISBN adequado demandaria uma pesquisa aprofundada. Por isso, entendemos que a definição do ID a partir da URI de uma versão digital das obras hospedadas por uma instituição referenciada como a Universidade Federal de Santa Catarina é a melhor alternativa para essa atividade, por dois motivos principais: facilidade de acesso aos textos integrais em versão digital e sua estruturação consonante com as melhores práticas em Web Semântica pelo referido portal.

Assim, para coleta e atribuição dos IDs, utilizamos uma query SPARQL que obtém os campos URI, títulos e títulos alternativos, pensando em situações de obras que componham um mesmo livro, como é o caso das crônicas e contos. As obras que possuem um mesmo URI no portal da UFSC, terão a propriedade OWL "partOf" adicionadas a triplstore e a inserção do título na URI em um formato do tipo #nome+do+conto, garantindo URIs únicos para cada item na triplstore.

Em resumo, cada obra listada no site machadodeassis.net é mapeada para a respectiva URI na base de dados da UFSC, utilizando um mapeamento manual para alguns títulos que têm grafias ligeiramente diferentes. Quando uma URI é encontrada com base em um título alternativo, o título principal correspondente é adicionado como book e o título da obra é considerado partOf deste book. Entradas com 'N/A' são removidas antes de adicionar as informações em structured\_data.

### 4.3. Processamento com Modelos de Linguagem

Para lidar com as referências geográficas literárias que mencionam locais com nomes históricos ou alternativos, adotamos o modelo GPT-3.5-instruct para validar e padronizar os nomes dos locais extraídos do portal MachadodeAssis.net. Essa escolha, entre os modelos de linguagem disponíveis em 2023, foi fundamentada nas características específicas do InstructGPT, desenvolvido pela OpenAI, que se destacava por sua capacidade de seguir instruções humanas com maior precisão, graças ao uso da técnica de "instruction fine-tuning" (Ouyang et al., 2022). Esse método de treinamento com aprendizado por reforço e feedback humano (RLHF) possibilitou uma adaptação do modelo para gerar respostas mais alinhadas com instruções específicas.

Diferentemente dos modelos convencionais da época, o InstructGPT oferecia:

- **Saídas Estruturadas:** o modelo demonstrava maior consistência na geração de formatos estruturados, como JSON e CSV, essenciais para a integração com APIs de georreferenciamento e para garantir uniformidade nos dados.
- **Controle de Formato:** apresentava previsibilidade na formatação das respostas, crucial para a automatização do pipeline e para minimizar o retrabalho manual.
- **Redução de Alucinações:** devido ao treinamento orientado a instruções, reduzia-se a tendência de gerar informações inconsistentes ou irrelevantes, um problema comum nos modelos não-instruct disponíveis à época.

Essas características foram particularmente relevantes para nossa tarefa de identificar e padronizar topônimos históricos e locais mencionados nas obras de Machado de Assis. Um exemplo prático é ilustrado no **Quadro 1**, que mostra o prompt utilizado e a resposta obtida para um verbete específico. Os parâmetros do modelo foram cuidadosamente ajustados para priorizar respostas mais focadas e estruturadas, com destaque para duas configurações principais:

A temperatura (temperature: 0.5) foi definida buscando um equilíbrio entre criatividade e precisão. Valores mais baixos (próximos a 0) produzem respostas mais determinísticas e consistentes, enquanto valores mais altos (próximos a 1) geram mais variabilidade. Para nossa tarefa de validação geográfica, era essencial manter consistência nas respostas, mas também permitir flexibilidade suficiente para interpretar descrições históricas complexas.

O limite de tokens (max\_tokens: 50) foi estabelecido considerando que respostas válidas deveriam ser sucintas e estruturadas. Este limite ajuda a evitar que o modelo gere informações extras ou divague, mantendo o foco na identificação precisa do local, cidade e país. Testes com valores maiores resultaram em respostas mais verbosas sem ganho significativo de precisão.

Quadro 1 – Estrutura do prompt submetido ao modelo gpt3.5-instruct

```
INPUT: {"role": "user", "content": = "Dada a seguinte descrição, responda: nome atual do lugar, cidade e país quando aplicável? Nome do lugar: Aterrado. Descrição: "No vasto alagadiço que era o mangue da Cidade Nova, desde o antigo Rossio Pequeno (atual praça Onze de Junho) foi construído um longo e estreito aterro. No tempo de D. João VI era o caminho usado pela família real para chegar a São Cristóvão. Assim foram adotados os nomes de "caminho das Lanternas" e "rua do Aterrado". A rua do Aterrado desapareceu com a construção da avenida Presidente Vargas, na década de 1940."
```

```
OUTPUT: {"role": "assistant", "content": { "Nome atual do lugar: Praça Onze de Junho, Cidade: Rio de Janeiro, País: Brasil"
```

Fonte: OpenAI API Reference

Como se vê no quadro acima, a saída do modelo manteve uma estrutura JSON consistente para todos os locais processados, seguindo sempre o padrão: {"role": "assistant", "content": { "Nome atual do lugar: [local], Cidade: [cidade], País: [país]"}. Esta padronização, como exemplificado no output {"role": "assistant", "content": { "Nome atual do lugar: Praça Onze de Junho, Cidade: Rio de Janeiro, País: Brasil"}}, foi mantida em todas as respostas do modelo, independente da complexidade do verbete de entrada. Esta uniformidade na estrutura de dados é particularmente valiosa em processamentos automatizados, pois permite:

1. Parsing consistente dos dados usando bibliotecas JSON padrão
2. Mapeamento direto para estruturas de dados em memória
3. Integração simplificada com APIs de georreferenciamento
4. Redução de necessidade de tratamento de casos especiais no código

O formato JSON, sendo uma notação leve e amplamente suportada, facilitou significativamente o pipeline de processamento subsequente. A consistência na estrutura das

respostas - mantendo sempre os três campos (nome atual do lugar, cidade e país) no mesmo formato - permitiu uma automação mais robusta e reduziu a necessidade de tratamentos de exceção no código, um benefício significativo considerando o volume de dados processados.

Essa padronização foi especialmente relevante para a etapa seguinte de georreferenciamento, onde os campos cidade e país serviram como contexto adicional para resolver ambiguidades nas APIs do Geonames e Google Maps, melhorando a precisão na identificação das coordenadas geográficas.

Para ilustrar melhor o funcionamento do modelo, trazemos outro exemplo que demonstra como o processamento lida com estabelecimentos históricos que não mais existem, mas cuja localização pode ser precisamente determinada através dos endereços originais preservados no traçado urbano do Rio de Janeiro.

Quadro 2 – Estrutura do prompt submetido ao modelo gpt3.5-instruct

INPUT: {"role": "user", "content": = f"Dada a seguinte descrição, responda: nome atual do lugar, cidade e país quando aplicável? Nome do lugar: Loja do Valais. Descrição: "A joalheria de Carlos Valais ficava na rua do Ouvidor, n. 81. Muito popular em meados do século XIX, anunciava-se no Almanak Laemmert de 1864 como 'joalheiros de Suas Majestades Imperiais'. Em 1873, a joalheria deixa de pertencer a Valais, passando às mãos de Alfredo Dreux."}
OUTPUT: {"role": "assistant", "content": { "Nome atual do lugar: Rua do Ouvidor, n. 81, Cidade: Rio de Janeiro, País: Brasil"}}

Fonte: OpenAI API Reference

Este exemplo é particularmente ilustrativo das possibilidades e limitações do processamento automatizado de referências histórico-geográficas. Embora a resposta do modelo não capture a natureza histórica do estabelecimento, que não mais existe, ela preserva a informação mais crucial para nossos propósitos de georreferenciamento: o endereço exato. Esta característica é especialmente valiosa considerando que grande parte do centro histórico do Rio de Janeiro manteve seu traçado e numeração originais, permitindo que locais mencionados por Machado de Assis possam ser posicionados com precisão em mapas contemporâneos. Assim, mesmo que o estabelecimento não exista mais, sua localização pode ser determinada com exatidão, contribuindo para uma representação precisa da geografia literária machadiana. Esta particularidade é especialmente relevante para ruas historicamente importantes como a do Ouvidor, que preserva até hoje muito de sua configuração oitocentista, permitindo uma sobreposição confiável entre o Rio de Janeiro de Machado de Assis e a cidade contemporânea.



#### 4.4. Processo de Validação e Desambiguação

Para garantir que o gazetteer literário fosse composto apenas por locais reais e historicamente válidos, implementamos uma lógica de filtragem no código que busca identificar e excluir entradas potencialmente fictícias, ambíguas ou inadequadas para georreferenciamento. Esse processo foi orientado pelo output gerado pelo modelo GPT, que nos auxiliou a identificar descrições ambíguas e nos ajudou a implementar um sistema de exclusão automática para essas entradas. A partir disso, utilizamos as funções `is_not_applicable` e `is_complex_or_ambiguous`, que verificam se a descrição do local contém termos como "não especificada", "não mencionada" ou expressões ambíguas ("vários países", "também conhecida como"). Além disso, foi incluído um processo de padronização e limpeza das entradas por meio da função `process_location_text`, que simplifica descrições complexas. Essa abordagem permitiu reduzir significativamente a presença de locais fictícios no dataset, embora ainda tenhamos realizado verificações manuais para garantir a precisão e qualidade dos dados coletados.

O modelo GPT-3.5-instruct foi empregado para fornecer uma primeira camada de validação dos topônimos, mas a validação final foi realizada manualmente pelos autores para garantir consistência histórica e literária, particularmente em locais do Rio de Janeiro oitocentista. Embora já existissem métodos automatizados de desambiguação toponímica em 2023, como *Entity Linking* com bases de conhecimento geográfico e *clustering* geográfico, optamos por um processo de revisão humana. Essa escolha foi motivada pela necessidade de um conhecimento contextual específico, além das limitações dos métodos automatizados frente à complexidade histórica dos dados literários.

Para projetos futuros, sugerimos o uso de técnicas mais automatizadas para desambiguação, incluindo:

- **Redes neurais para Entity Linking:** baseadas em bases de conhecimento geográfico.
- **Métricas de acurácia:** como a Distância de Haversine entre pontos preditos e reais para locais históricos conhecidos.
- **Taxa de concordância:** entre múltiplas fontes de georreferenciamento.

Essas técnicas poderiam fornecer métricas quantitativas de precisão, complementando o trabalho manual e oferecendo maior confiabilidade na validação dos dados.

#### 4.5. Obtenção das coordenadas geográficas com Geonames e Google Maps

Durante a abordagem, identificou-se que a coleta em GeoNames apresenta maior dificuldade em identificar locais específicos das cidades, como bairros, praças, ruas e outros logradouros se comparado ao GoogleMaps. Apesar de serem soluções parecidas, GeoNames e GoogleMaps descrevem de formas diferentes seus objetos, sendo GeoNames uma ontologia específica para entidades geográficas e GoogleMaps um serviço de geolocalização que não usa quaisquer ontologias de web semântica para descrever seus objetos. Assim, todos os dados coletados via Google Maps deverão ser submetidos ao GeoNames para a definição do código de característica, isto é, a definição desta ontologia para cada tipo de entidade geográfica. Ao final, o que não for localizado será identificado pela URL do Google Maps.

O pipeline representado a seguir chama a API GeoNames para enriquecer entradas JSON-LD com informações geográficas. Primeiro, são importados os módulos necessários: JSON para manipulação de arquivos JSON e geocoder para a interação com a API GeoNames. Em seguida, uma função chamada `get_geonames_data` é definida para aceitar um nome de lugar (`place_name`) como argumento. Se a consulta for bem-sucedida, a função constrói um dicionário que contém informações como tipo, nome, latitude, longitude, código de característica (se a entidade é uma Cidade, Estado, País e etc. segundo a nomenclatura da base), além de uma URI específica do GeoNames para aquele local. A URI é criada usando o `'geonameld'` da resposta da API. Em suma, a função retorna o dicionário de dados geográficos.

As informações geográficas são adicionadas ao dicionário sob a chave "geo" e, assim, a função processa recursivamente cada item da lista de modo a adequar os dados à estrutura do schema.org, que possui a propriedade `GeoCoordinates` como uma subclasse do tipo "geo", que armazena detalhes relativos à localização de entidades geográficas nesta ontologia.

A integração entre GeoNames e Google Maps, embora eficaz para a obtenção de coordenadas, apresentou desafios relacionados à classificação ontológica dos locais, especialmente aqueles obtidos via Google Maps. Para endereçar esta limitação e garantir uma taxonomia uniforme em nossa base de dados, implementamos uma etapa adicional de classificação utilizando o modelo GPT-4, que se tornou disponível durante o desenvolvimento do projeto.

#### 4.6. GPT 4 como classificador de dados geográficos

Como etapa complementar do processo de validação iniciado com o GPT-3.5-instruct, implementamos o GPT-4 para classificação mais granular dos dados geográficos. Esta escolha foi motivada pela necessidade de categorizar precisamente as entidades geográficas segundo a ontologia GeoNames, especialmente para locais coletados via Google Maps que não possuíam classificação ontológica nativa.

O modelo GPT-4, disponibilizado pela OpenAI em 2023, representou um avanço significativo em relação aos modelos anteriores, especialmente na compreensão de contextos específicos e na consistência de respostas estruturadas. Para nossa aplicação, desenvolvemos uma classe Gpt4Turbo que interage com a API através do endpoint gpt-4-1106-preview, configurada para otimizar a classificação de entidades geográficas.

A implementação foi estruturada para maximizar a precisão das classificações através de parâmetros cuidadosamente ajustados. O método gptCall\_json foi desenvolvido com controles específicos de temperatura e limites de tokens, visando manter a consistência nas classificações enquanto permite flexibilidade suficiente para interpretar descrições complexas de locais históricos. O processamento segue um padrão estruturado onde cada entidade geográfica é analisada individualmente, resultando em classificações padronizadas segundo a ontologia GeoNames. Para ilustrar o processo, apresentamos a estrutura do prompt utilizado:

Quadro 3 – Estrutura do prompt submetido ao modelo gpt4-turbo

SYSTEM: {"role": "system", "content": "Você é um assistente de classificação de dados geográficos e responde no formato JSON."}
INPUT: {"role": "user", "content": "Qual é o gn:featureCode e o gn:featureCodeName mais adequado do GeoNames para o local chamado '{place_name}' descrito como: {description}?"}
OUTPUT: {"role": "assistant", "content": {"gn:featureCode": "PCLI", "gn:featureCodeName": "independent political entity"}}

Fonte: OpenAI API Reference

A saída padronizada do modelo, mantendo sempre a estrutura com gn:featureCode e gn:featureCodeName, permitiu a integração direta com nossa base de dados, enriquecendo as entidades geográficas com classificações consistentes segundo a ontologia GeoNames. Este processo foi particularmente valioso para normalizar a classificação de locais provenientes de diferentes fontes (GeoNames e Google Maps), estabelecendo uma taxonomia uniforme em nossa base de dados.

É importante notar que, assim como na etapa anterior com GPT-3.5-instruct, mantivemos um processo de validação manual das classificações geradas, especialmente para casos ambíguos ou locais históricos com características particulares. Esta abordagem híbrida - combinando a eficiência do processamento automatizado com a precisão da validação humana - mostrou-se eficaz para garantir a qualidade dos dados geográficos em nosso gazetteer literário.

#### **4.7. Conversão de JSON-LD para TURTLE com Apache Jena**

Apache Jena é um framework para trabalhar com dados ligados e semânticos que suporta vários formatos, incluindo JSON-LD e Turtle. JSON-LD é um formato baseado em JSON, muito útil para representar informações de maneira organizada e legível tanto para máquinas quanto para humanos. Já Turtle é um formato de serialização para dados RDF, focado em ser conciso e facilmente legível, comumente usado em aplicações de dados ligados por ser mais compacto e mais fácil de escrever e entender do que outros formatos RDF, como RDF/XML.

Para converter JSON-LD para Turtle, usamos o riot, uma ferramenta do Apache Jena que transforma a estrutura de dados de um formato para o outro, mantendo a semântica, e alterando a sintaxe. Há algumas razões para converter JSON-LD para Turtle, dentre as quais, destaca-se a facilidade de leitura e escrita: Turtle, com sua sintaxe mais concisa facilita a escrita e leitura de consultas SPARQL. Isso torna mais simples trabalhar com dados ligados, especialmente para consultas complexas. Cabe ressaltar ainda a compatibilidade e eficiência, pois alguns sistemas e ferramentas que trabalham com SPARQL podem ser otimizados para trabalhar com Turtle. Por fim, destaca-se a padronização, já que Turtle é um formato amplamente adotado para representar dados RDF, o que favorece a interoperabilidade, possibilitando a integração e o compartilhamento de dados.

#### **4.8. Consulta ao arquivo de dados**

Para realização de consultas SPARQL, inicialmente, é preciso definir os prefixos e namespaces. No contexto desta aplicação, temos o “PREFIX schema: <http://schema.org/>”, isto é, um prefixo chamado schema para o namespace <http://schema.org/>. O uso de prefixos em SPARQL é uma maneira conveniente de abreviar URIs longas. No restante da consulta, qualquer vez que “schema:” for usado, ele se refere a <http://schema.org/>.

SELECT \* WHERE é a parte principal da consulta SPARQL. SELECT \* significa que você quer selecionar todas as variáveis disponíveis (?sujeito, ?predicado, ?objeto) nos padrões de triplas correspondentes. WHERE é a cláusula onde se define um padrão de tripla que se quer procurar no grafo RDF. Dentro da cláusula WHERE, temos dois padrões de triplas: o primeiro, ?s schema:name "África" procura por todas as triplas no grafo RDF onde o predicado é schema:name e o objeto é o literal "África". A variável ?s será qualquer sujeito que tenha "África" como um schema:name.

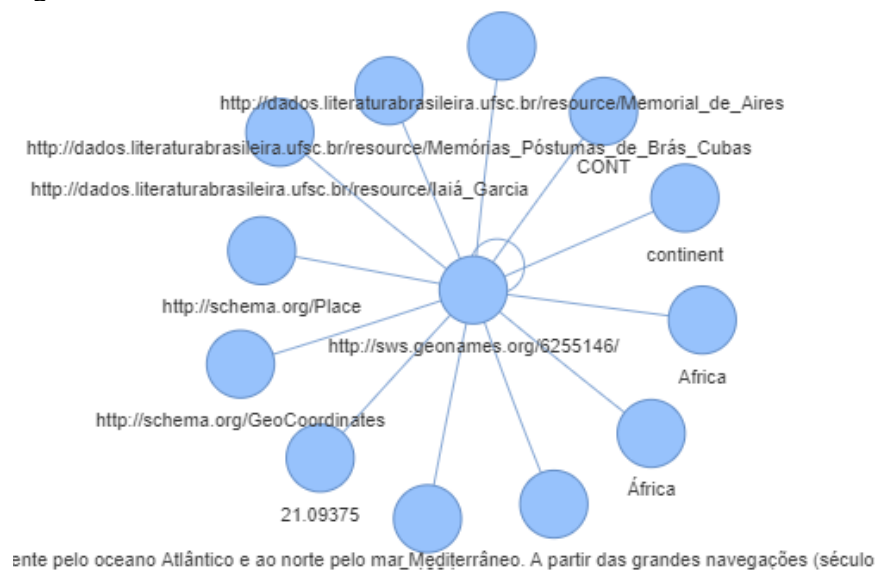
Já a notação ?p ?o aponta um padrão mais geral e procura por todas as triplas no grafo RDF onde o sujeito (?s) é o mesmo encontrado na primeira tripla (ou seja, qualquer sujeito que tenha "África" como schema:name). ?p e ?o são variáveis que representam respectivamente qualquer predicado e objeto associados a esse sujeito. Portanto, a consulta SPARQL busca todas as triplas no grafo RDF onde o sujeito tem um nome "África" segundo o schema.org, e retorna todas as informações (predicado e objeto) associadas a esses sujeitos como localização, descrições, e as citações ao local na obra de Machado de Assis.

Quadro 4 – Sintaxe da consulta ao *schema:name* África

```
PREFIX schema: <http://schema.org/>
SELECT * WHERE {
?s schema:name "África" .
?s ?p ?o .
}.
```

Fonte: autores (2024)

Figura 3: Grafo com os dados da consulta ao *schema:name* África.



Fonte: autores (2024)

## 5 RESULTADOS E DISCUSSÃO

### 5.1. Visualizações Implementadas

Uma vez coletados os dados de Latitude e Longitude e estruturadas as informações dos textos e localidades, tentaremos demonstrar as visualizações possíveis a partir da estruturação dos dados. Inicialmente, serão criados pinos em um mapa, onde cada pino carrega e exibe o verbete sobre o local no portal MachadodeAssis.net. Cumpre considerar que, apesar do nosso esforço em validar as coordenadas geográficas, alguns pinos estão em locais errados, como no "Morro do Castelo", acidente geográfico que ocupava a área central do Rio de Janeiro e que foi removido em 1922 em um amplo movimento de reestruturação da então capital do Brasil. No processo de coleta, GeoNames identificou um local homônimo na Ilha de Paquetá. Há outros erros do tipo e inconsistências que foram tratadas manualmente quando identificadas, uma vez que os modelos da OpenAI usados aqui são sensíveis a alterações de temperatura e podem oscilar na apresentação dos resultados.

#### 5.1.1. Mapa de locais coletados com descritores das localidades

A imagem a seguir apresenta um mapa mundi onde marcadores em azul indicam as localidades mencionadas nas obras de Machado de Assis, conforme catalogadas pela enciclopédia MachadodeAssis.net. É possível observar uma significativa concentração de marcadores no Brasil, especialmente na região do Rio de Janeiro, o que era esperado dado o contexto da obra do autor. Há também uma presença marcante de referências na Europa, principalmente em países como França, Inglaterra, Itália e Portugal, além de menções dispersas em outros continentes como África e Ásia. Cada marcador, ao ser selecionado, exibe o verbete correspondente da enciclopédia, oferecendo contexto sobre como aquele local específico aparece na obra machadiana.

Esta visualização demonstra o alcance geográfico das referências presentes nos textos de Machado de Assis, revelando como o autor, mesmo escrevendo majoritariamente sobre o Rio de Janeiro, construiu uma obra com dimensão verdadeiramente cosmopolita, incluindo referências a locais em praticamente todos os continentes.

Figura 4: Mapa de locais citados no conjunto da obra



Fonte: autores (2024)

### 5.1.2. Mapa de citações por local

Já na visualização que segue, demonstra-se uma funcionalidade do gazetteer que permite uma experiência de leitura mais contextualizada. O usuário pode navegar (através das setas de slideshow) pelas diferentes citações onde o local - neste caso, Áden, no Iêmen - aparece nas obras de Machado de Assis, enquanto mantém a referência geográfica visível no mapa. No exemplo mostrado, vemos um trecho do conto "O Imortal" (1882), com o marcador posicionado precisamente em Áden, acompanhado da descrição da enciclopédia que classifica o local como "seat of a first-order administrative division".

Esta interface integrada permite que o leitor compreenda simultaneamente:

1. A localização exata do lugar mencionado
2. O contexto narrativo em que aparece (através do trecho da obra)
3. A informação enciclopédica sobre o local
4. As múltiplas ocorrências do mesmo local em diferentes obras (através do slideshow)

Essa abordagem multifacetada enriquece a experiência de leitura, permitindo que o leitor estabeleça conexões entre a geografia real e o uso literário que Machado faz destes espaços em suas narrativas.

Figura 5: Mapa de citações por local

Fonte: autores (2024)

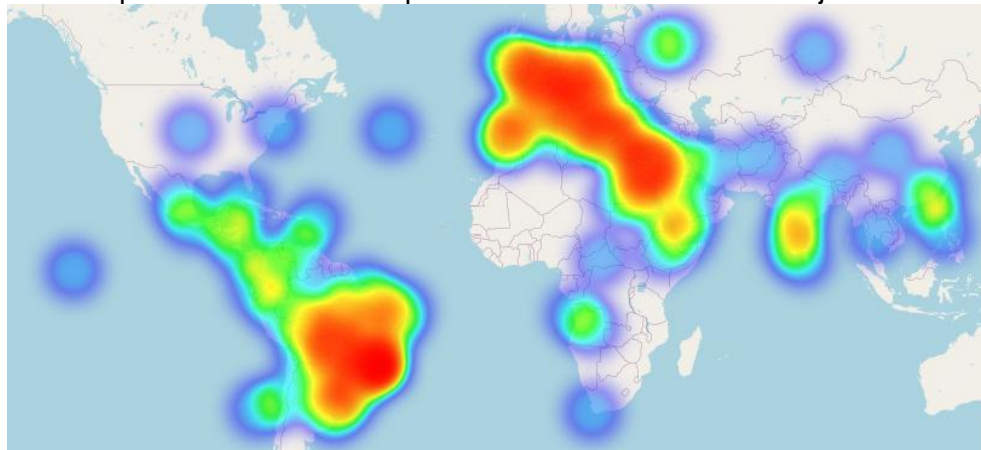
### 5.1.3. Mapa de calor com o conjunto das citações

O mapa de calor em questão revela padrões na distribuição geográfica das citações na obra de Machado de Assis. As áreas em vermelho indicam maior concentração de menções, enquanto as gradações em amarelo, verde e azul mostram frequências decrescentes.

Destacam-se três principais zonas de intensidade:

1. A mais intensa (vermelho) sobre o Brasil, especialmente na região do Rio de Janeiro, evidenciando o forte vínculo da narrativa machadiana com seu espaço imediato
2. Uma concentração significativa (vermelho-amarelo) na Europa, particularmente na região do Mediterrâneo e Europa Central, refletindo as frequentes referências à cultura e história europeias
3. Uma presença moderada (verde-azul) em pontos dispersos da Ásia, África e América do Norte, demonstrando o alcance global das referências do autor

Figura 6: Mapa de calor com a frequência de locais citados no conjunto da obra



Fonte: autores (2024)

Esta visualização térmica permite compreender não apenas onde Machado situa suas narrativas, mas também a intensidade com que diferentes regiões do globo aparecem em sua obra, revelando um escritor que, embora profundamente ancorado no Rio de Janeiro, mantinha um diálogo constante com diferentes partes do mundo em seus textos.

### 5.1.4. Citações por Obra

Esta visualização demonstra como o gazetteer pode revelar camadas históricas na obra de Machado de Assis. No trecho do "Memorial de Aires", vemos uma referência à Batalha de Tuiuti (Guerra do Paraguai), narrada através das memórias do Conselheiro Aires,



um diplomata que representa o olhar do escritor, membro do funcionalismo público imperial, sobre eventos históricos importantes.

A interface permite visualizar simultaneamente:

1. A localização geográfica de Tuiuti no Paraguai (marcador azul no mapa)
2. O trecho onde o narrador relembra o episódio histórico, conectando sua experiência diplomática com a cobertura jornalística do evento
3. A classificação do local como "populated locality" conforme a ontologia utilizada

Esta funcionalidade pode ser útil para pesquisadores interessados na intersecção entre literatura e história, pois permite mapear como Machado de Assis, através de seus personagens (neste caso, o Conselheiro Aires), incorporava eventos históricos significativos em sua narrativa, oferecendo um panorama único do Brasil oitocentista.

Figura 7: Mapa de citações por obra



Fonte: autores (2024)

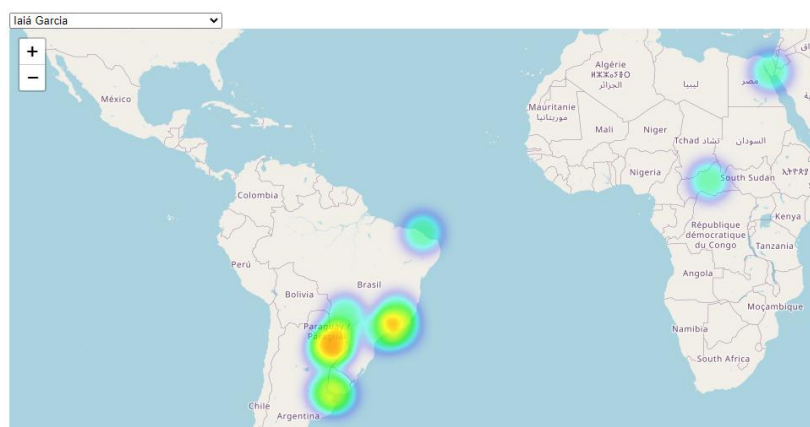
### 5.1.5. Citações por Obra

A próxima imagem apresenta um mapa de calor que mostra a frequência dos locais mencionados na obra *laiá Garcia* de Machado de Assis. As áreas com maior intensidade de cor indicam uma frequência maior de referências a esses locais na narrativa. Esse romance, conhecido por abordar temas como a escravidão e a Guerra do Paraguai, inclui diversas localizações na América do Sul, especialmente no Brasil e no Paraguai, refletindo o contexto histórico e social da época.

Esse tipo de visualização é útil para identificar a geografia literária e o impacto das temáticas históricas em *laiá Garcia* ou em qualquer outra obra de preferência do usuário,

permitindo aos leitores e pesquisadores explorarem como esses espaços geográficos se relacionam com a trama e os eventos abordados por Machado de Assis.

Figura 8: Mapa de calor de locais citados por obra  
Escolha uma Obra:



Fonte: autores (2024)

### 5.1.6. StreetView Panorama

Esta visualização demonstra uma das possibilidades mais interessantes do projeto: a integração com o Google Street View, que permite uma experiência imersiva ao conectar os locais mencionados por Machado de Assis com suas configurações atuais. No exemplo, vemos Copacabana como aparece em "Esaú e Jacó", onde o Conselheiro Aires busca refúgio em suas caminhadas, junto com uma vista panorâmica atual da praia.

Embora a funcionalidade completa de visualização 360° não tenha sido implementada devido aos custos elevados da API do Google Street View, este exemplo ilustra o potencial da ferramenta para estabelecer pontes entre o Rio de Janeiro histórico de Machado e a cidade contemporânea. A justaposição do trecho literário com a imagem atual permite ao leitor compreender como os espaços descritos pelo autor se transformaram ao longo do tempo, enriquecendo a experiência de leitura com uma dimensão visual e temporal.

Esta funcionalidade, mesmo em versão limitada, demonstra como a tecnologia pode criar novas camadas de significado na leitura da obra machadiana, permitindo um diálogo entre passado e presente através dos espaços urbanos.

Figura 9: Visão de local e citação sobre Copacabana com Google StreetView.



Fonte: autores (2024)

## 5.2 Validação e contribuições da comunidade

Até o presente momento, o projeto foi reconhecido como “relevante” por representantes acadêmicos da área dos estudos literários, que recomendaram esta publicação para periódicos voltados à tecnologia da informação. No entanto, a validação e as contribuições da comunidade são parte do compromisso deste projeto em manter-se em constante evolução, visando a melhoria contínua por meio da participação ativa da comunidade acadêmica. Embora o gazetteer literário seja fundamentado nos dados previamente curados e validados pela plataforma *machadodeassis.net*, a publicação deste artigo busca ir além de uma documentação técnica: atua como um convite à colaboração, incentivando a participação de pesquisadores e leitores de Machado de Assis. Esse engajamento da comunidade é visto como essencial para o aprimoramento da ferramenta.

Para facilitar essa colaboração, criamos mecanismos específicos que permitem que os interessados participem de maneira prática. Um dos principais canais é um repositório público no GitHub, onde o código-fonte do projeto estará disponível. Por meio desse repositório, pesquisadores e desenvolvedores poderão examinar a implementação técnica, propor melhorias por meio de *pull requests*, reportar problemas na seção de *issues* e até mesmo sugerir novas funcionalidades. Esse ambiente colaborativo visa atrair contribuições que ajudem a expandir e refinar o projeto.

Além do repositório no GitHub, também planejamos oferecer um formulário de feedback acessível via interface web, voltado para o público em geral. Esse formulário permitirá que os usuários reportem inconsistências nas coordenadas geográficas, sugiram correções para classificações de locais históricos e compartilhem suas percepções sobre a

experiência de uso. Assim, usuários de diferentes perfis poderão contribuir não só com ajustes técnicos, mas também com ideias para novas visualizações e análises que possam enriquecer a exploração dos dados.

Dentro dessas iniciativas, algumas áreas são priorizadas para a validação com o apoio da comunidade. A precisão geográfica é um aspecto central: buscamos validar as coordenadas atribuídas a locais históricos, identificar e resolver ambiguidades toponímicas e corrigir eventuais erros de georreferenciamento. Também nos concentramos na classificação ontológica, revisando as categorias atribuídas automaticamente pelos modelos de inteligência artificial, refinando a taxonomia de lugares e verificando a consistência histórica das classificações.

A experiência de usuário é outra área de foco. Avaliamos a utilidade das diferentes visualizações disponíveis, exploramos necessidades específicas dos pesquisadores e coletamos sugestões que possam inspirar novas formas de interação com os dados. A integração com os estudos literários também é de extrema importância, permitindo validar se as análises espaciais trazem insights relevantes para a crítica literária. Espera-se, assim, identificar padrões geográficos significativos e desenvolver novas abordagens analíticas que dialoguem com a obra de Machado de Assis.

Este projeto é, portanto, uma iniciativa em constante transformação, na qual o feedback da comunidade acadêmica e dos usuários será fundamental para o aprimoramento técnico e o enriquecimento do escopo acadêmico. Através das contribuições recebidas, planejamos refinar os algoritmos de processamento, melhorar a precisão das classificações e implementar funcionalidades adicionais que respondam às necessidades emergentes.

## **6 CONSIDERAÇÕES FINAIS**

Como tentamos demonstrar, a intersecção entre tecnologia, literatura e geolocalização pode oferecer experiências de leitura interessantes, muitas delas inexploradas nesse estudo, o que proporciona um terreno fértil para o desenvolvimento das chamadas humanidades digitais. Ao unir essas diferentes áreas do conhecimento, pode-se enriquecer as narrativas literárias com uma dimensão espacial palpável, bem como explorar novas maneiras de engajar e interpretar textos clássicos e contemporâneos. Nesta atividade, por exemplo, pudemos mergulhar em algumas obras de Machado de Assis, não apenas seguindo a trama de forma imaginária, mas também acompanhando os passos dos

personagens em mapas digitais que falam do universo de possibilidades conhecidas e imaginadas por um dos grandes expoentes do Século XIX no Brasil.

À medida que a história se desenrola, o leitor pode explorar os locais reais ou fictícios mencionados, compreendendo melhor o contexto cultural e histórico que molda a narrativa. Essa imersão geográfica concede uma camada de compreensão que permite que a literatura respire fora das páginas dos livros e ganhe vida em outros suportes. Nesse contexto, a tecnologia é apenas uma lente através da qual podemos reexaminar obras clássicas sob uma luz diferente. Com isso, não se pretende substituir o contato com o livro, mas complementar a experiência de leitura a partir de outros métodos e ferramentas para questionar, compreender e apreciar a literatura.

Do ponto de vista dos estudos acadêmicos em tecnologia da informação o presente artigo busca oferecer conceitos, abordagens e mesmo incentivar projetos multidisciplinares que envolvam a estruturação de dados segundo os parâmetros da chamada Web Semântica a fim de enriquecer o corpo de conhecimento qualificado disponível. Ao empregar padrões como RDF (Resource Description Framework), SPARQL (uma linguagem de consulta) e ontologias OWL (Web Ontology Language), é possível estruturar dados literários de maneira que máquinas possam "entender" e processar relações complexas, facilitando a pesquisa interdisciplinar e permitindo a criação de redes de conhecimento que podem revelar padrões e tendências previamente ocultos dentro de grandes volumes de texto.

Em um aspecto mais técnico, o desafio está em como representar e interligar dados de forma que eles sejam ao mesmo tempo acessíveis para análise computacional e visualização humana. O uso de Linked Data e padrões de interoperabilidade assegura que diferentes conjuntos de dados possam ser combinados, fornecendo um quadro mais completo e multidimensional de informações.

Por fim, a Web Semântica não é apenas um instrumento para melhorar a pesquisa e educação nas humanidades digitais, mas também um convite à colaboração transdisciplinar. Ela encoraja o diálogo entre cientistas da computação, bibliotecários digitais, historiadores, geógrafos e literatos, todos contribuindo com suas perspectivas únicas para a construção de uma infraestrutura de conhecimento mais rica e conectada.

## REFERÊNCIAS

- ADIBOZZI, D.; et al. Towards a Human-like Open-Domain Chatbot. [S.l.]: **Google Research**, 2020. Disponível em: <https://research.google/pubs/towards-a-human-like-open-domain-chatbot/>. Acesso em: 22 set. 2023.
- ASSIS, J. M. M. **Dom Casmurro**. v.I. Rio de Janeiro: Nova Aguilar, 1994b. Disponível em: [http://www.dominiopublico.gov.br/pesquisa/DetalheObraDownload.do?select\\_action=&co\\_obra=16931&co\\_midia=2](http://www.dominiopublico.gov.br/pesquisa/DetalheObraDownload.do?select_action=&co_obra=16931&co_midia=2) Acesso em: 7 out. 2023.
- BERNERS-LEE, T; HENDLER, J; LASSILA, O. The Semantic Web. **Scientific American**, v. 284, n. 5, p. 34-43, 2001.
- BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. In: \_\_\_\_\_. **Semantic Web – Interoperability, Usability, Applicability**. 2011. Disponível em: <https://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf>. Acesso em: 3 out. 2023.
- BURDICK, A.; DRUCKER, J.; LUNENFELD, P.; PRESNER, T.; SCHNAPP, J. *Digital Humanities*. Cambridge: MIT Press, 2012.
- CHALHOUB, S. **Machado de Assis**: historiador. São Paulo: Companhia das Letras, 2003.
- DIEGO, M. Entrevista com Marta de Senna. **Machado de Assis em Linha**, São Paulo, v. 13, n. 29, p. 181-189, abr. 2020. DOI: 10.1590/1983-68212020132913. Disponível em: <https://doi.org/10.1590/1983-68212020132913>. Acesso em: 20 ago. 2023.
- DO NASCIMENTO, J. G. O branco imposto e o negro conquistado: Machado de Assis na propaganda da Caixa Econômica Federal. **Revista da Associação Brasileira de Pesquisadores/as Negros/as** (ABPN), v. 8, n. 20, p. 74-85, 2016.
- DOCUMENTA** Palmares. Campinas, SP: UNICAMP/IFCH/CECULT, 2021. (Base de Dados). Disponível em: <https://www.palmares.ifch.unicamp.br/>. Acesso em: 28 out. 2023.
- GREGORY, I. N.; HARDIE, Andrew. **The Geography of Words: Mapping Language Using GIS**. Cambridge: Cambridge University Press, 2018.
- GROVER, C.; T. Richard. A Gazetteer and Georeferencing for Historical English Documents. In: \_\_\_\_\_. **Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014**. Gothenburg, Sweden: Association for Computational Linguistics, 26 Abril 2014.
- HETLAND, M. Python and the Web. In: \_\_\_\_\_. **Beginning Python From Novice to Professional**. New York: Apress, 2005. pp. 313–339. Disponível em: [https://doi.org/10.1007/978-1-4302-0072-7\\_15](https://doi.org/10.1007/978-1-4302-0072-7_15). Acesso em: 19 nov. 2023.
- ILIAIDIS, A.; ACKER, A.; STEVENS, W. One schema to rule them all: How Schema.org models the world of search. **Journal of the Association for Information Science and Technology**, 2022. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24744>. Acesso em: 12 set. 2023.
- JOHNSON, P; HARDIE, A; GROVER, C; TOBIN, R. **Mapping Shakespeare's London: A Gazetteer and Georeferencing for Historical English Documents**. In: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Gothenburg, Sweden: Association for Computational Linguistics, 2019.

MARANDOLA JR., E.; MARANDOLA, E.. **Geografia e Literatura: Perspectivas e Desafios**. São Paulo: Editora da Universidade de São Paulo, 2021.

MORETTI, F. **Atlas do Romance Europeu 1800-1900**. São Paulo: Boitempo, 2003. Disponível em: <https://www.boitempoeditorial.com.br/produto/atlas-do-romance-europeu-1800-1900-73>. Acesso em: 17 set. 2023.

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; et al. Training language models to follow instructions with human feedback. In: \_\_\_\_\_ **Advances in Neural Information Processing Systems** (NeurIPS). 2022. Disponível em: [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf). Acesso em: 25 nov. 2023.

PENG, B.; LI, C.; HE, P.; GALLEY, M.; GAO, J. **Instruction tuning with GPT-4**. ArXiv preprint arXiv:2304.03277. 2023. Disponível em: <https://arxiv.org/abs/2304.03277> Acesso em: 14 dez. 2023.

PÉREZ, J.; ARENAS, M.; GUTIERREZ, C. Semantics and complexity of SPARQL. **ACM Transactions on Database Systems**, 2009. Disponível em: <https://dl.acm.org/doi/abs/10.1145/1567274.1567278> Acesso em: 20 set. 2023.

RICHARDSON, L. **Beautiful Soup Documentation**. Disponível em: <https://www.crummy.com/software/BeautifulSoup/>. Acesso em: 12 nov. 2024.

SANTOS, C. A. N. (2022). Notas sobre variação toponímica: co-ocorrência e concorrência de nomes de lugares no Brasil Colônia. **Letras**, (60), 255–282. <https://doi.org/10.5902/2176148542326> Acesso em: 13 de novembro de 2024.

SANTOS, D. **Futuro risonho: prolegómenos para uma colaboração entre a Linguateca e o NuPILL**. 2022. Disponível em: <https://www.duo.uio.no/bitstream/handle/10852/98444/1/SantosNuPILL.pdf> Acesso em: 8 dez. 2023.

SANTOS, M. **Chatterbot baseado em obras de Machado de Assis: uma plataforma para o estímulo a leitura de literatura clássica**. Bauru: UNISAGRADO, 2021. Disponível em: <https://secure.usc.br/handle/handle/102>. Acesso em: 22 set. 2023.

SCHMIDT, B; PIATTI, B. **The Literary Atlas of Europe: Ontologies and Interactive Maps for the Exploration of Literary Spaces**. In: *Proceedings of Mapping Literary Modernism*, Zurich, 2020. SIEMER, S. Exploring the Apache Jena Framework. George August University, Göttingen, 2019. Disponível em: <http://www.dbis.informatik.uni-goettingen.de/Teaching/Theses/PDF/FPrakt-Siemer-MSc-jun-2019.pdf>. Acesso em: 22 ago de 2023.

SCHREIBMAN, S.; SIEMENS, R.; UNSWORTH, John (Ed.). *A Companion to Digital Humanities*. Oxford: Blackwell Publishing, 2004.

SCHWARZ, R. **Um mestre na periferia do capitalismo: Machado de Assis**. São Paulo: Duas Cidades, 1990.

SEGARAN, T; EVANS, Colin; TAYLOR, Jamie. **Programming the Semantic Web**. Sebastopol: O'Reilly, 2009. pp. 23-26.

SIMON, R. et al. Pelagios Commons: Establishing a Community for Linked Ancient Geodata. In: **Digital Humanities 2016: Conference Abstracts**. Kraków: Jagiellonian University & Pedagogical University, 2016.

TALLY JR., R. T. **Geocritical Explorations: Space, Place, and Mapping in Literary and Cultural Studies**. New York: Palgrave Macmillan, 2014.

WESTPHAL, B. **La Géocritique: Réel, Fiction, Espace**. Paris: Editions de Minuit, 2011.

ZAMARIANO, M. Cartografiação de dados toponímicos no Brasil: perspectiva historiográfica. **Revista do GELNE**, [S. l.], v. 14, n. 1 Ed. Esp, p. 77–98, 2016. Disponível em: <https://periodicos.ufrn.br/gelne/article/view/9384>. Acesso em: 13 nov. 2024.

ZHANG, W.; et al. **Spatial Humanities: Past, Present, and Future**. New York: Routledge, 2022.

## NOTAS

### AGRADECIMENTOS

Ricardo Marcondes Marcacini e equipe do portal MachadodeAssis.net

### CONTRIBUIÇÃO DE AUTORIA

**Concepção e elaboração do manuscrito:** D. A. Moreira, D. M. Rocha

**Coleta de dados:** D. M. Rocha

**Análise de dados:** D. A. Moreira, D. M. Rocha

**Discussão dos resultados:** D. A. Moreira, D. M. Rocha

**Revisão e aprovação:** D. A. Moreira

### CONJUNTO DE DADOS DE PESQUISA

<https://huggingface.co/spaces/histlearn/MachadodeAssis>

### FINANCIAMENTO

Não se aplica.

### CONSENTIMENTO DE USO DE IMAGEM

Não se aplica

### APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA

Não se aplica.

### CONFLITO DE INTERESSES

Não se aplica.

### LICENÇA DE USO

Os autores cedem à **Encontros Bibli** os direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob a [Licença Creative Commons Attribution](#) (CC BY) 4.0 International. Esta licença permite que **terceiros** remixem, adaptem e criem a partir do trabalho publicado, atribuindo o devido crédito de autoria e publicação inicial neste periódico. Os **autores** têm autorização para assumir contratos adicionais separadamente, para distribuição não exclusiva da versão do trabalho publicada neste periódico (ex.: publicar em repositório institucional, em site pessoal, publicar uma tradução, ou como capítulo de livro), com reconhecimento de autoria e publicação inicial neste periódico.

### PUBLISHER

Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no [Portal de Periódicos UFSC](#). As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

### EDITORES

Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Jônatas Edison da Silva, Mayara Madeira Trevisol, Edna Karina da Silva Lira e Luan Soares Silva.

### HISTÓRICO

Recebido em: 26-07-2024 – Aprovado em: 12-12-2024 – Publicado em: 14-03-2025

