revista humanidades digitais

H2D

digital humanities journal

Estudo metodológico – Representação de conhecimento para a análise de corpus históricos

Autores:

Rafael Prezado, Universidade de Évora, Portugal

https://orcid.org/0009-0004-9795-0148

Renata Vieira, Universidade de Évora, Portugal

https://orcid.org/0000-0003-2449-5477

Como citar:

Prezado, R. & Vieira, R. (2025). Estudo metodológico — Representação de conhecimento para a análise de corpus históricos. *H2D / Revista de Humanidades Digitais, 7*.

DOI: 10.21814/h2d.6580

Article history: Submetido a 30 de maio de 2025; Aceite a 29 de novembro 2025; Publicado a 22 de dezembro de 2025

Estudo metodológico – Representação de conhecimento para a análise de corpus históricos

A Methodological Approach to Knowledge Representation in Historical Corpus Analysis

Rafael Prezado[1], Universidade de Évora, Portugal
Renata Vieira[2], Universidade de Évora, Portugal

Resumo

Este artigo apresenta um modelo teórico para a análise semântica de corpus históricos com base em ferramentas digitais interoperáveis. Recorrendo ao PaddleOCR para extração textual, a uma ontologia construída sobre o modelo CIDOC CRM e a consultas SPARQL definidas manualmente, propõe-se uma abordagem escalável para o tratamento e exploração de fontes jornalísticas digitalizadas. O estudo de caso centra-se na cobertura mediática da missão Apolo 11 em periódicos portugueses do século XX, no contexto do Estado Novo. A metodologia permitiu identificar padrões discursivos, estruturar relações semânticas entre eventos, atores e publicações, e demonstrar o potencial da modelação ontológica na análise crítica de discursos históricos. O modelo, ainda em fase exploratória, mostra-se promissor para futuras aplicações em Humanidades Digitais.

Palavras-chave: Humanidades Digitais, Análise de Corpus, Ontologias, Reconhecimento Óptico de Caracteres, SPARQL, Imprensa Histórica.

Abstract

This article presents a theoretical model for the semantic analysis of historical corpora using interoperable digital tools. By applying PaddleOCR for text extraction, an ontology based on the CIDOC CRM model, and manually defined SPARQL queries, a scalable approach is proposed for processing and exploring digitized journalistic sources. The case study focuses on media coverage of the Apollo 11 mission in Portuguese newspapers from the 20th century, within the context of the Estado Novo regime. The methodology enabled the identification of discursive patterns, the structuring of semantic relationships between events, actors, and

[1] **Rafael Prezado** frequenta atualmente dois cursos de mestrado na Universidade de Évora: um em História, com especialização em História Política, e outro em Engenharia Informática. Os seus interesses académicos incluem a História da Ciência, programação, as Humanidades Digitais e as Ontologias. Tem participado em projetos de investigação, conferências académicas e iniciativas de voluntariado que exploram as interseções entre o conhecimento histórico e as tecnologias digitais. Email: rafaeltprezado@gmail.com

[2] **Renata Vieira** obteve o grau de doutoramento em Ciência Cognitiva em 1998 pela Universidade de Edimburgo, com uma tese na área da Linguística Computacional. Desde então, tem exercido atividade académica nos domínios da Linguística Computacional, da Inteligência Artificial, da Representação do Conhecimento e da Web Semântica. É atualmente Investigadora Principal na Universidade de Évora, onde coordena o Laboratório de Humanidades Digitais. Possui experiência na coordenação de projetos interinstitucionais e internacionais, financiados por empresas e por agências de financiamento à investigação científica. Email: renatav@uevora.pt

publications, and demonstrated the potential of ontological modelling for critical discourse analysis. Although still in an exploratory phase, the model shows promise for future applications in the field of Digital Humanities.

Keywords: Digital Humanities, Corpus Analysis, Ontologies, Optical Character Recognition, SPARQL, Historical Newspapers.

---

1. Introdução

Este estudo apresenta um modelo metodológico para a análise semântica de corpora textuais históricos, com recurso a tecnologias de reconhecimento ótico de caracteres (OCR), ontologias e consultas SPARQL. O principal objetivo consiste em demonstrar de que forma estes recursos podem ser aplicados de modo integrado para estruturar, analisar e interpretar volumes de texto não estruturado, como aqueles que se encontram em arquivos relacionados com a imprensa. A partir da digitalização e do tratamento computacional dos documentos, a proposta visa oferecer uma abordagem aplicável, escalável e adaptável a diferentes projetos de investigação em humanidades digitais, com ênfase na reutilização de dados e na interoperabilidade entre disciplinas.

Um dos principais desafios associados ao trabalho com fontes históricas digitalizadas reside no seu formato desestruturado, o qual dificulta a análise sistemática através de métodos tradicionais. A classificação manual dos dados, prática comum na historiografia, limita tanto a escalabilidade como a reprodutibilidade dos resultados, e frequentemente inviabiliza uma leitura crítica dos discursos dominantes, tal como proposto por Fairclough (2009) na sua abordagem do discurso como forma de prática social.

Esta proposta aborda tal problemática através de uma metodologia semi-automatizada que combina ferramentas digitais e modelos semânticos, com o intuito de transformar *plain text* em dados estruturados e reutilizáveis, possibilitando o desenvolvimento de análises tanto quantitativas como qualitativas.

2. Metodologia

Conforme destacado por Meroño-Peñuela et al., a adoção de tecnologias da Web Semântica no âmbito da investigação histórica representa uma resposta concreta aos problemas clássicos do tratamento de dados históricos, nomeadamente, a heterogeneidade das fontes, a dificuldade de documentação estruturada e a necessidade de interoperabilidade semântica (Meroño-Peñuela et al., 2015, pp. 10–12).

Este paradigma favorece a construção de modelos de conhecimento robustos, alinhados com padrões internacionais, e potencia a reutilização dos dados em múltiplos contextos de análise histórica. Nos termos do ciclo de vida da informação histórica definido por Meroño-Peñuela et al. (2015, pp. 4–7), a nossa abordagem ocupa as seis etapas principais:

- *Creation*: Inclui o planeamento do projeto, a digitalização das fontes (PDFs, JPGs, PNGs) e a conversão para texto via OCR, com recurso ao PaddleOCR, garantindo a criação inicial dos dados em formato digital estruturado.
- *Enrichment*: Nesta fase, os dados extraídos são enriquecidos com metainformação, como a identificação de entidades (nomes, instituições, locais), e estruturados semanticamente com base numa ontologia própria, inspirada no modelo CIDOC-CRM.

- *Editing*: A etapa de edição inclui o processo de validação manual de segmentos com baixo grau de confiança no OCR, assim como a modelação semântica em RDF, que organiza as entidades e relações de forma coerente.
- *Retrieval*: A exploração dos dados é feita através de consultas SPARQL no Apache Jena Fuseki, permitindo recuperar informação específica de forma eficiente.
- *Analysis*: Os dados estruturados são analisados com base em padrões de coocorrência, ligações semânticas e recorrência temática ao longo do corpus.
- *Presentation*: Por fim, os resultados são apresentados sob a forma de visualizações interativas, esquemas ontológicos e exemplos extraídos do corpus, permitindo comunicar os resultados de forma clara e acessível.

Contextualizando o processo no âmbito do ciclo de vida da informação histórica, a etapa de *creation* foi operacionalizada através da conversão de fontes digitalizadas (PDF, PNG ou JPG) em texto legível por máquina. Para esse fim, utilizou-se o PaddleOCR, em articulação com scripts em Python, automatizando a extração textual a partir dos documentos digitalizados. Numa fase inicial, foram realizados testes com o Tesseract OCR, uma ferramenta de código aberto desenvolvida pelos HP Labs e disponibilizada como software livre em 2005 (Smith, 2007, pp. 1–2).

Apesar da sua arquitetura ser robusta, baseada em análise por componentes conectados e reconhecimento adaptativo em dois passos, verificou-se que o PaddleOCR apresentava melhores resultados em termos de precisão, especialmente em páginas de imprensa de menor qualidade, ligeiramente degradadas ou com layouts irregulares.

Por essa razão, a ferramenta foi adotada na fase principal do projeto, tirando partido da sua arquitetura leve e modular (PP-OCR), que integra capacidades otimizadas de deteção, retificação geométrica e reconhecimento de texto em contextos exigentes (Du et al., 2020, p. 10). A Figura 1 exemplifica o resultado obtido numa manchete da imprensa histórica, evidenciando a precisão da ferramenta na deteção e extração de texto.

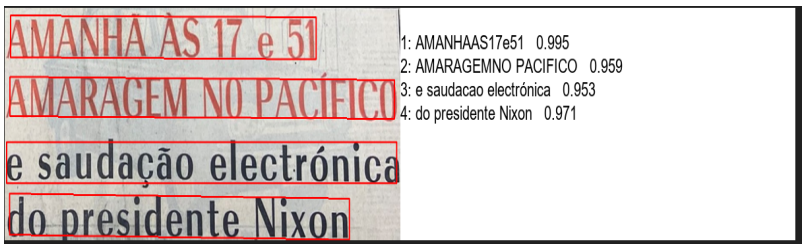


Figura 1. Resultado do OCR com PaddleOCR: áreas de deteção de texto com valores de confiança.

Para garantir a fiabilidade dos dados extraídos, foi aplicado um controlo de qualidade baseado nos valores de confiança gerados pelo PaddleOCR. Sempre que esses valores ficavam abaixo de um limiar definido, os segmentos eram revistos manualmente, sobretudo quando envolvidos títulos, nomes próprios ou elementos relevantes para a modelação semântica. Este procedimento assegurou que apenas informação validada integrasse a fase seguinte da análise.

Uma vez extraído o conteúdo textual, procedeu-se a uma fase de análise semi-automatizada destinada à identificação de entidades-chave, tais como nomes próprios, locais, eventos, instituições e conceitos temáticos. Para esse fim, foi desenvolvida uma ontologia própria que permite estruturar

logicamente as relações entre entidades, em conformidade com normas internacionais aplicadas à gestão do património cultural e documental.

O modelo ontológico, referente ao caso de estudo apresentado a seguir, foi concebido para representar de forma explícita os vínculos entre:

- periódicos e acontecimentos históricos,
- atores políticos e científicos,
- elementos visuais, como caricaturas, fotografias, infografias, manchetes, localizações físicas e contextos ideológicos.

Estas relações foram codificadas sob a forma de RDF triples, permitindo a exploração do corpus através de consultas semânticas em SPARQL. O ambiente de análise adotado foi o Apache Jena Fuseki, onde as bases de dados RDF foram armazenadas e consultadas.

Importa destacar que a ontologia desenvolvida não foi aplicada diretamente como sistema de anotação textual sobre os documentos analisados. Em vez de marcar semanticamente os textos originais, optou-se por uma estratégia conceptual: as entidades e as relações foram sistematizadas a partir da informação extraída e organizadas num repositório independente. Esta abordagem permitiu estruturar logicamente os dados sem interferir nos documentos de origem, assegurando maior interoperabilidade e otimizando a eficiência das consultas semânticas.

Contudo, o processo metodológico apresentou limitações importantes. A qualidade do OCR variou consoante o estado físico dos periódicos digitalizados, sendo afetada por fatores como o desgaste do papel ou a tipografia. Adicionalmente, a extração semântica exigiu validações manuais periódicas, sobretudo em casos de baixo grau de confiança, para garantir a fiabilidade dos dados integrados. Finalmente, a representação de conteúdos visuais, como caricaturas, diagramas ou fotografias, implicou desafios técnicos adicionais, dado que requerem processos de anotação distintos ainda em fase de experimentação.

3. Caso de Estudo

O fluxo de trabalho foi aplicado a um conjunto de periódicos portugueses do século XX, com foco num evento-chave da corrida espacial: a chegada da missão Apolo 11 à Lua, em 1969. Este corpus foi selecionado pela sua relevância simbólica nos discursos sobre modernidade, progresso científico e propaganda política, no contexto do regime autoritário do Estado Novo em Portugal.

A informação extraída permitiu analisar tanto a cobertura mediática como os discursos editoriais. Numa abordagem quantitativa, foram geradas tabelas que registam a frequência de ocorrência de palavras-chave, a distribuição temática pelas diferentes secções dos periódicos e a proporção de referências em primeira página. Cada periódico analisado, nomeadamente *O Século*, *Diário de Lisboa*, *A Flama* e *Avante!*, foi sistematizado em tabelas que especificam a localização dos documentos, o tipo de conteúdo e o enquadramento temporal.

Do ponto de vista qualitativo, identificaram-se padrões discursivos como a exaltação da ciência ocidental, o silenciamento de atores soviéticos e a apropriação ideológica do acontecimento por parte do regime. A articulação destes padrões com o modelo CIDOC CRM permitiu não apenas descrever, mas também estratificar as relações semânticas entre atores, eventos e meios de comunicação.

Tabela 1. Material documental integrado na ontologia.

ID	Periódico e Data	Descrição

ID1	O Século – 1969-07-24	Representação da reentrada atmosférica
ID2	Pravda – 1969-07-22	Análise de Alexander Vinogradov
ID3	Diário de Lisboa – 1957-10-06	Manchete relacionada com o Sputnik-1
ID4	Diário de Lisboa – 1957-10-07	Figura com a trajetória do Sputnik-1
ID5	Diário Popular – 1969-07-16	Diagrama e cronologia da missão Apolo 11
ID6	Diário Popular – 1969-07-21	Fotografia
ID7	O Século – 1969-07-21	Representação de Armstrong na Lua
ID8	Diário Popular – 1969-07-20	Representação científica da Lua
ID9	Diário de Lisboa – 1969-07-19	Representação da descolagem do módulo lunar Eagle
ID10	O Século – 1969-07-25	Manchete: Benefícios da exploração espacial
ID11	Diário Popular – 1969-07-18	Artigo de Isaac Asimov
ID12	Diário de Lisboa – 1969-07-21	Artigo de Joshua Lederberg
ID13	O Século – 1969-07-25	Manchete: Os benefícios do espaço
ID14	Diário Popular – 1969-07-17	Fotografias de Virgílio, Pires e Amaral
ID15	A Capital – 1969-07-23	Fotografias de Ribeiro Alves e Nunes Machado
ID16	Diário de Lisboa – 1969-07-16	Fotografias de Grácio e Paulino
ID17	Diário Popular – 1969-07-17	Fotografia de Alfredo Campos
ID18	Diário Popular – 1969-07-17	Manchete: Esperança no espaço
ID19	Diário do Sul – 1969-07-22	Imagem do Sagres-Houston
ID20	Diário de Lisboa – 1969-07-18	Caricatura: Adolf Hitler e Eva Braun na Lua
ID21	Diário de Lisboa – 1969-07-21	Caricatura: Eusébio a receber a Lua para jogar no Benfica
ID22	Diário Popular – 1969-07-16	Caricatura: “Riso Amarelo” – Foguetes
ID23	Diário Popular – 1969-07-19	Caricatura: “Riso Amarelo” – Construções

Tal como verificável na Tabela 1, a cada material documental considerado relevante foi atribuído um identificador único (ID), que permitiu a sua posterior integração sistemática na estrutura ontológica desenvolvida. Esta codificação revelou-se essencial para garantir a rastreabilidade dos dados no corpus e assegurar a correspondência semântica entre os objetos descritos e as instâncias ontológicas modeladas no Protégé.

A articulação entre o conteúdo documental e a camada ontológica facilitou a organização lógica das fontes, respeitando simultaneamente o princípio de interoperabilidade definido pelo CIDOC CRM. A diversidade de documentos, entidades e relações representada na tabela ilustra o modo como o modelo foi aplicado ao caso de estudo, permitindo uma exploração semântica rigorosa e contextualizada dos materiais analisados.

3.1. Modelação ontológica no CIDOC CRM

A ontologia desenvolvida no contexto deste estudo encontra-se em construção contínua, evoluindo de forma incremental à medida que os periódicos são analisados e novas entidades, eventos ou relações relevantes são identificados. Esta abordagem iterativa permite adaptar a estrutura ontológica às especificidades do corpus em tempo real, assegurando coerência interna e pertinência semântica. A modelação foi realizada no software Protégé, com base no esquema do CIDOC Conceptual Reference Model (CIDOC CRM), a

ontologia de referência para documentação de património cultural, reconhecida como norma internacional (ISO 21127:2014).

O CIDOC CRM fornece uma estrutura formal para descrever eventos, entidades e relações em termos espaço-temporais, permitindo representar não apenas objetos históricos, mas também os contextos em que foram criados, utilizados ou transformados. A sua natureza *event-centric* é particularmente adequada para o estudo histórico-mediático, pois permite articular narrativas, atores e suportes materiais em redes relacionais consistentes.

A escolha do CIDOC CRM foi também fundamentada em precedentes relevantes no campo da história e das humanidades digitais. Trabalhos recentes, como o desenvolvido no âmbito do projeto SeaLiT (2022), demonstram como o modelo CIDOC CRM é eficaz na integração de dados dispersos e heterogéneos relacionados com fontes históricas, como registos navais, trajetórias de indivíduos ou atividades económicas (Fafalios et al., 2023). Neste contexto, a SeaLiT Ontology utiliza a versão 7.1.1 do CIDOC CRM, explorando a sua flexibilidade para representar informação complexa sobre eventos, agentes, localizações e objetos, lógica que inspirou diretamente a estratégia adotada neste estudo (Fafalios et al., 2023).

O uso deste modelo favorece a interoperabilidade com outras ontologias e sistemas de documentação histórica e cultural, promovendo a sustentabilidade a longo prazo dos dados estruturados. Ao articular conceitos como E5 Event, P14 carried out by, P4 has time-span ou P7 took place at, a modelação semântica passa a permitir inferências sobre o corpus mediático, facilitando a integração futura com outros datasets históricos compatíveis.

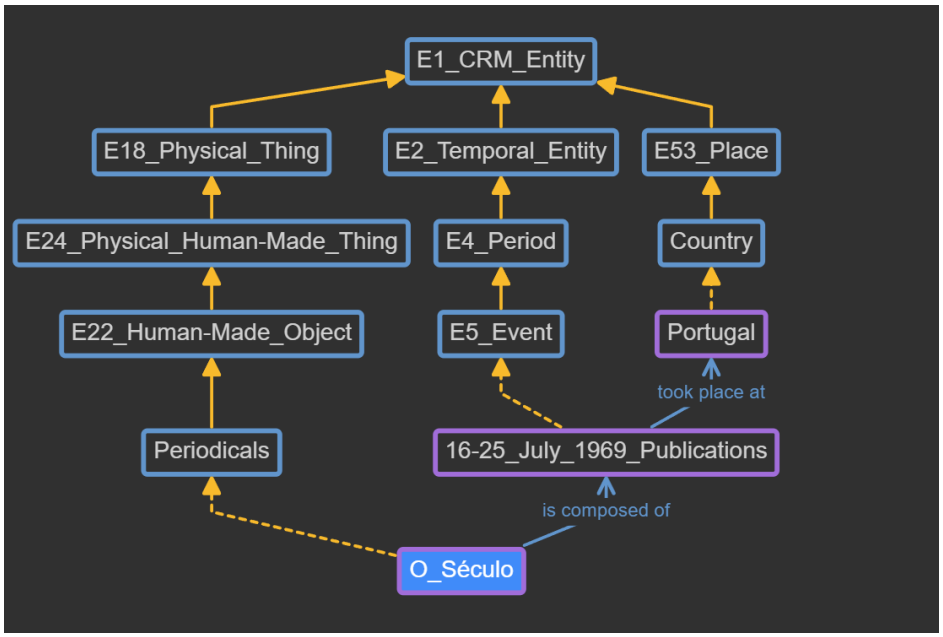
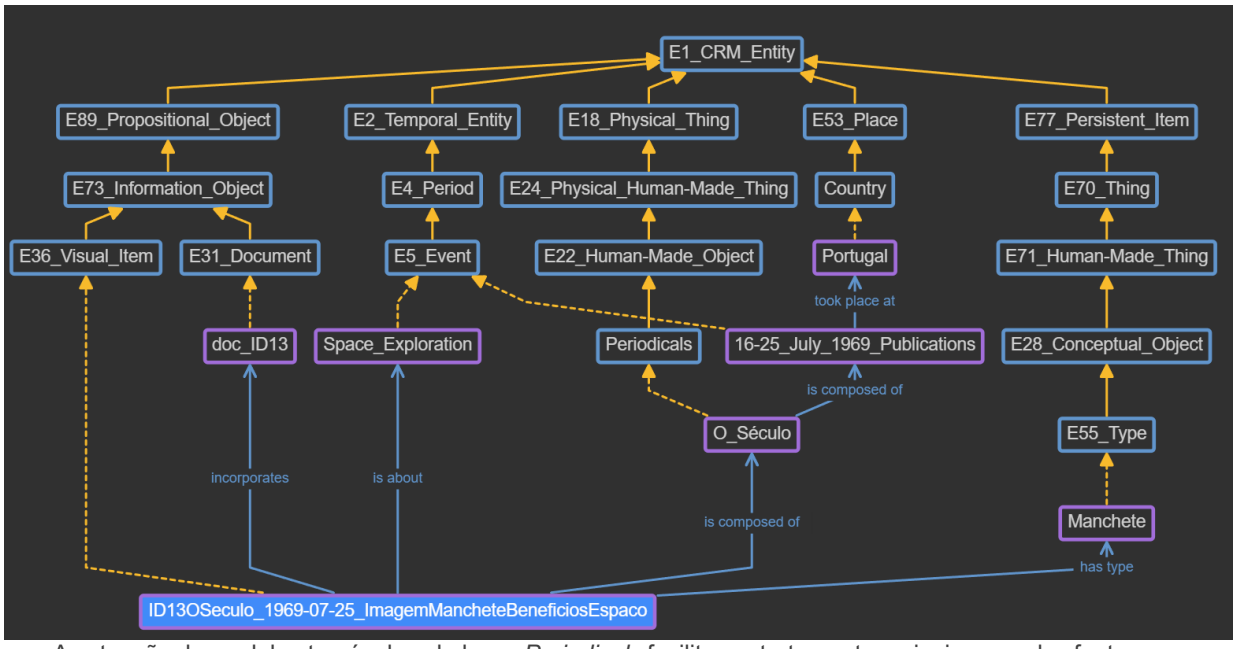


Figura 2. Representação ontológica do periódico *O Século* no Protégé

A Figura 2 apresenta a instância do periódico *O Século*, modelada no Protégé como parte de uma subclasse específica, *Periodicals*, derivada de *E22_Human-Made_Object*, no âmbito do modelo CIDOC CRM (Bekiari et al., 2024, p. 74).

Esta instância está ligada às publicações do período compreendido entre 16 e 25 de julho de 1969 através da propriedade P46_isComposedOf, o que permite representar de forma estruturada a relação entre o objeto físico (o periódico) e as suas edições específicas.



A extensão do modelo através da subclasse *Periodicals* facilita um tratamento mais rigoroso das fontes em contextos históricos e semânticos, permitindo compatibilizar a granularidade documental com os princípios de interoperabilidade e reutilização de dados do CIDOC CRM.

Figura 3. Instância D13.

A Figura 3 apresentada corresponde à estruturação semântica de um dos materiais analisados no presente trabalho, identificado por ID13OSeculo_1969-07-25_ImagemMancheteBeneficiosEspaco. Este representa uma manchete publicada n'O Século a 25 de julho de 1969, e está modelada segundo o padrão CIDOC CRM.

Na figura observa-se que o documento (doc_ID13), categorizado como E31_Document, incorpora um item visual (E36_Visual_Item) e um conteúdo proposicional (E89_Propositional_Object), ambos subordinados à categoria mais ampla E73_Information_Object. O conteúdo refere-se ao tema da exploração espacial (Space_Exploration), modelado como E5_Event.

O periódico O Século aparece como integrante do conjunto de publicações entre 16 e 25 de julho de 1969, associadas à cobertura mediática da alunagem. Estas publicações estão contextualizadas geograficamente em Portugal (E53_Place), reforçando a localização sociopolítica da recepção mediática. Finalmente, a instância é do tipo Manchete (E55_Type), destacando a sua posição editorial de relevo. Este alinhamento semântico procura seguir os padrões definidos pelo CIDOC CRM (Bekiari et al., 2024, pp. 57–58).

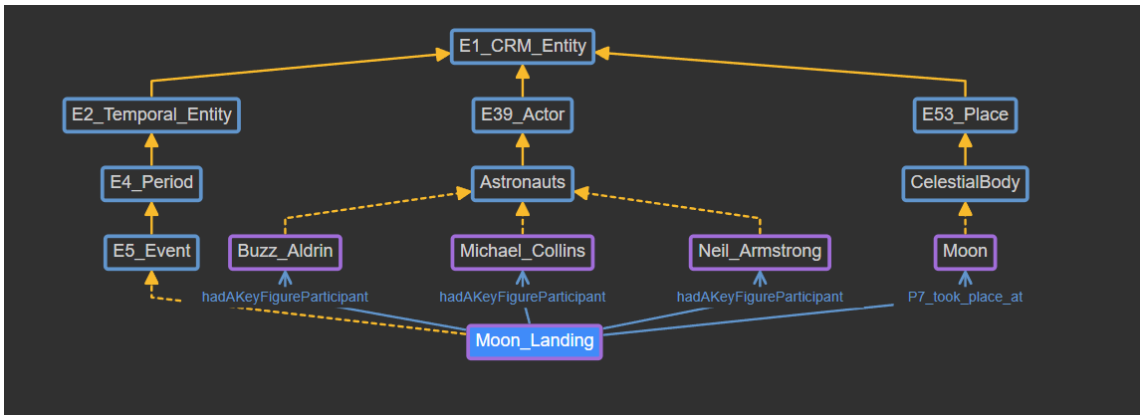


Figura 4. Instância da *Moon Landing* representada no Protégé

A Figura 4 representa o evento *Moon_Landing* como uma instância de E5_Event no Protégé, de acordo com o modelo CIDOC CRM. O evento está ligado ao local *Moon*, modelado como uma instância de E53_Place e pertencente a uma subclasse designada *CelestialBody*, através da propriedade P7_took_place_at.

Os astronautas Buzz Aldrin, Michael Collins e Neil Armstrong são representados como instâncias de *E39_Actor*, agrupadas sob a entidade *Astronauts*, e relacionados com o evento através da propriedade *hadAKeyFigureParticipant*, uma extensão derivada de *P11_hadParticipant*, que assinala o seu papel central na missão (Bekiari et al., 2024, pp. 121–122).

Estes nós semânticos permitem a realização de consultas detalhadas em SPARQL. Foram executados exemplos reais, entre os quais se destaca a seguinte interrogação:

• Como foram enquadrados os discursos sobre ciência e tecnologia num regime autoritário?

Estas perguntas não apenas demonstram a capacidade analítica do modelo, como também evidenciam o seu potencial para gerar novas linhas de investigação a partir de dados já existentes.

Apesar de a ontologia desenvolvida estar ainda em fase exploratória e restrita ao âmbito do caso de estudo, o modelo apresenta várias vantagens concretas. A estrutura modular adotada permitiu a reutilização de dados estruturados com base em normas semânticas consolidadas, promovendo a interoperabilidade com outros projetos de Humanidades Digitais.

Além disso, a delimitação temática e temporal do corpus não compromete a escalabilidade, já que a ontologia foi concebida para permitir expansão progressiva à medida que novas fontes forem integradas. A formalização dos vínculos semânticos também fornece uma base metodológica clara e reproduzível, aplicável a contextos linguísticos e culturais diversos, incluindo estudos em português, espanhol ou inglês.

4. Resultados e Visualização

Os resultados foram visualizados em múltiplos níveis. Através do Apache Jena Fuseki, organizaram-se as respostas a consultas específicas, gerando visualizações relativas a:

• O volume de referências por periódico e por data.

journal	name	totalReferences
1<http://example.org/space_race/O_Século>	O Século	"177"^^<http://www.w3.org/2001/XMLSchema#integer>
2<http://example.org/space_race/Diário_Popular>	Diário Popular	"129"^^<http://www.w3.org/2001/XMLSchema#integer>
3<http://example.org/space_race/Diário_de_Lisboa>	Diário de Lisboa	"108"^^<http://www.w3.org/2001/XMLSchema#integer>
4<http://example.org/space_race/A_Capital>	A Capital	"91"^^<http://www.w3.org/2001/XMLSchema#integer>
5<http://example.org/space_race/Diário_do_Sul>	Diário do Sul	"13"^^<http://www.w3.org/2001/XMLSchema#integer>
6<http://example.org/space_race/Flama>	Flama	"10"^^<http://www.w3.org/2001/XMLSchema#integer>
7<http://example.org/space_race/Avante>	Avante	"0"^^<http://www.w3.org/2001/XMLSchema#integer>

Figura 5. Organização por total de referências – Apache Jena Fuseki

Na figura 5 consta uma visualização gerada no Apache Jena Fuseki, com base em consultas SPARQL, que indica o número total de referências encontradas em cada periódico do corpus. Esta organização permite comparar quantitativamente a cobertura mediática dos eventos analisados, facilitando a identificação de padrões discursivos em função do meio de comunicação e do contexto temporal.

name	firstPageReferences
Diário de Lisboa	"17"^^<http://www.w3.org/2001/XMLSchema#integer>

Figura 6. Consulta SPARQL no Apache Jenna Fuseki sobre o *Diário de Lisboa*.

A Figura 6 ilustra uma consulta realizada no Apache Jena Fuseki, onde foi aplicada uma query SPARQL sobre a base ontológica construída. Esta consulta visava responder à pergunta: *"Quantas referências relacionadas com a missão Apolo 11 foram destacadas na primeira página do Diário de Lisboa entre 16 e 25 de julho de 1969?"*

Para além da pesquisa estruturada via SPARQL, considera-se como horizonte de evolução metodológica a adoção de sistemas de busca textual sobre grafos RDF, de modo a tornar a exploração dos dados mais acessível e dinâmica. Nesse sentido, o trabalho de Elas4RDF: Multi-perspective Triple-Centered Keyword Search over RDF (Kadilierakis et al., 2020) demonstra uma abordagem que adapta um sistema de recuperação de informação textual tradicional, Elasticsearch, para indexar e recuperar triplas RDF como unidades de busca, oferecendo resultados refinados e informativos com base em palavras-chave.

Essa alternativa é interessante porque permitiria, no futuro, oferecer ao utilizador final uma interface tipo "search engine": basta escrever termos-chave (por exemplo: "Apolo 11", "primeira página", "Diário de Lisboa") para obter um conjunto de triplas ou entidades relevantes, beneficiando da expressividade da modelação semântica, mas sem exigir conhecimentos em SPARQL.

Além disso, esse método de busca sobre RDF pode ser combinado com a abordagem atual: a base ontológica continua a servir como respaldo formal e estrutural; o Elasticsearch serve como camada de acesso mais intuitiva e amigável. Assim, amplia-se o potencial de reutilização dos dados, facilitando análises exploratórias, cruzamentos rápidos e extração de padrões inesperados, sem comprometer a coerência semântica.

Por fim, a adoção desta solução também responde a uma das críticas metodológicas frequentemente associadas ao uso de SPARQL: a exigência de competências técnicas e a rigidez das consultas. A busca por palavras-chave torna os dados mais acessíveis a investigadores com diferentes perfis, democratizando o uso da base construída e abrindo caminho a colaborações e reutilizações externas.

5. Discussão

A abordagem aqui desenvolvida diferencia-se de iniciativas como o sistema Synthesis, usado no projeto RICONTRANS (Fafalios et al., 2021), ao priorizar a modelação semântica de padrões discursivos e a representação contextualizada de eventos mediáticos. Enquanto o Synthesis se foca na documentação colaborativa de artefactos históricos e suas transferências, o presente modelo orienta-se para a análise semântica de fontes textuais e visuais.

Com efeito, as vantagens do modelo assentam em fundamentos operacionais claros. A reutilização de dados estruturados é viabilizada pela atribuição de identificadores únicos a cada item relevante, o que permite a sua integração e reaproveitamento em diferentes fases ou projetos.

A interoperabilidade decorre da adoção do CIDOC CRM, cuja base semântica garante compatibilidade com outros sistemas em Humanidades Digitais. A escalabilidade é assegurada pela natureza modular da ontologia: novos dados são incorporados progressivamente sem comprometer a coerência do modelo.

Por fim, o pipeline definido, da análise e digitalização das fontes, à validação, identificação e modelação semântica, estabelece um percurso metodológico reproduzível e adaptável a diferentes contextos linguísticos e culturais.

Neste sentido, a solução baseou-se num projeto recente (Liang et al., 2025) que aplicava o CIDOC CRM ao estudo de artefactos culturais, organizando dados multimodais e relacionando-os com dimensões simbólicas e contextuais. Com base nessa lógica, estruturou-se um repositório digital, no github, onde se integraram imagens e documentos provenientes dos periódicos analisados, através da criação de uma propriedade específica de dados (data property type “Link”) que associa diretamente os recursos visuais às entidades e relações extraídas.

6. Conclusões

Este trabalho propôs uma metodologia robusta para a análise semântica de corpora históricos, apoiada em ferramentas computacionais interoperáveis. A conjugação de tecnologias de reconhecimento ótico de caracteres (Tesseract e PaddleOCR), modelação ontológica baseada no CIDOC CRM e consultas SPARQL no ambiente Apache Jena permitiu estruturar e explorar com rigor fontes históricas desestruturadas, nomeadamente da imprensa periódica.

Apesar de se encontrar numa fase exploratória, a aplicação ao estudo da cobertura da missão Apolo 11 nos jornais portugueses do Estado Novo demonstrou o potencial da abordagem para identificar padrões discursivos e representar relações semânticas entre entidades, eventos e conceitos. Esta validação empírica confirma a utilidade do modelo como instrumento de análise histórico-cultural.

A proposta contribui para o desenvolvimento metodológico nas Humanidades Digitais, ao oferecer um fluxo de trabalho replicável, escalável e compatível com normas internacionais de representação do conhecimento. O seu carácter modular favorece a adaptação a outros contextos linguísticos, históricos ou temáticos, assim como a integração em infraestruturas digitais já existentes.

Espera-se que este modelo possa apoiar futuras investigações que articulem a análise crítica das fontes com tecnologias semânticas, promovendo práticas de investigação mais abertas, reprodutíveis e sustentáveis no campo da história digital.

Referências

Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Michon, P., Ore, C.-E., Stead, S., & Velios, A. (2024). *Definition of the CIDOC Conceptual Reference Model*. CIDOC CRM SIG.

Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., & Wang, H. (2020). *PP-OCR: A practical ultra lightweight OCR system* (arXiv:2009.09941). arXiv. https://doi.org/10.48550/arXiv.2009.09941

Fafalios, P., Marketakis, Y., Samaritakis, G., Patramanis, D., & Tzitzikas, Y. (2021). Towards Semantic Interoperability in Historical Research: Documenting Research Data and Knowledge with Synthesis. In *Hotho, A., et al.* (Eds.), *The Semantic Web – ISWC 2021*. Lecture Notes in Computer Science, Vol. 12922. Springer, Cham. https://doi.org/10.1007/978-3-030-88361-4_40

Fafalios, P., Kritsotaki, A., & Doerr, M. (2023). *The SeaLiT Ontology – An Extension of CIDOC-CRM for the Modeling and Integration of Maritime History Information*. ACM Journal on Computing and Cultural Heritage, 16(3), Article 60, 21 pages. https://doi.org/10.1145/3586080

Fairclough, N. (2009). *Discourse and social change*. Polity Press.

Kadilierakis, G., Fafalios, P., Marketakis, Y., Tzitzikas, Y., & Doerr, M. (2020). Keyword Search over RDF using Document-Centric Information Retrieval Systems. In *A. Harth et al. (Eds.), The Semantic Web – ESWC 2020*. Lecture Notes in Computer Science, Vol. 12123. Springer. https://doi.org/10.1007/978-3-030-49461-2_8

Liang, Y., Xie, B., Tan, W., & Zhang, Q. (2025). Ontology-based construction of embroidery intangible cultural heritage knowledge graph: A case study of Qingyang sachets. *PLOS ONE, 20*(1), e0317447. https://doi.org/10.1371/journal.pone.0317447

Smith, R. W. (2007). An overview of the Tesseract OCR engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Meroño-Peñuela, A., Ashkpour, A., van Erp, M. G. J., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, K. S., & van Harmelen, F. A. H. (2015). *Semantic Technologies for Historical Research: A Survey.* Semantic Web, 6(6), 539–564. https://doi.org/10.3233/SW-140158