15º Congresso Nacional BAD

SHORT PAPER

SciLinkDB: uma abordagem aberta para a estruturação de metadados científicos em bibliotecas académicas

Daniel Moura Gonçalves[a], Vanessa Silva[b]

[a]*Nova School of Business and Economics, Portugal, daniel.goncalves@novasbe.pt*
[b]*Nova School of Business and Economics, Portugal, vanessa.silva@novasbe.pt*

Resumo

Os *Current Research Information Systems* (CRIS) constituem ferramentas centrais para a gestão da informação científica nas instituições de ensino superior, embora a sua adoção esteja frequentemente associada a elevados custos e a dependências tecnológicas. Este trabalho tem como objetivo analisar a viabilidade do desenvolvimento de um CRIS baseado em *Knowledge Graph*, recorrendo à Wikibase como tecnologia de suporte.

A metodologia adotada consistiu no desenvolvimento de vários testes de interoperabilidade que culminaram no desenvolvimento de um protótipo orientado à agregação e visualização de metadados científicos a partir de fontes abertas.

Os resultados evidenciam que esta abordagem permite uma elevada flexibilidade na modelação dos dados e produz visualizações eficazes para apoio à gestão da ciência.

Conclui-se que os CRIS assentes em *Knowledge Graphs* constituem uma alternativa possível aos CRIS proprietários, apresentando elevado potencial de escalabilidade e impacto institucional, embora dependa da capacidade técnica das instituições, de automação de processos e do reforço da interoperabilidade com fontes abertas e fidedignas.

Palavras-chave: CRIS, Knowledge Graph, Wikibase, Interoperabilidade, Metadados científicos, Ciência Aberta

Introdução

As bibliotecas académicas desempenham um papel fundamental no apoio às atividades de gestão da informação científica. Desde os anos 90 do século passado começaram a utilizar os designados *Current Research Information Systems* (CRIS) para efeitos de gestão da ciência produzida pelas suas instituições(Bryant et al., 2018), estabelecendo a "ponte entre a comunidade científica e a infraestrutura contribuindo assim para a abertura da ciência e da inovação" (Dias et al., 2018). Os CRIS são *softwares* robustos que permitem "a acumulação de informação sobre as atividades de investigação da organização num único local, aumenta o nível de partilha dos resultados da investigação, permite avaliar a eficácia e a eficiência das atividades de investigação e resolve o problema de prever as futuras orientações desta atividade, permitindo ajustar a estratégia atual para o seu desenvolvimento". (Udartseva 2024). Discute-se atualmente a utilização de funcionalidades, como as 'Palavras-chave da Biblioteca' no Pure CRIS, para apoiar os novos fluxos de trabalho de Acesso Aberto e permitir às instituições monitorizar e analisar a implementação das políticas de Ciência Aberta (de Castro 2024). Esta funcionalidade permite aos Bibliotecários usarem este campo para atribuir diferentes categorias às publicações. Embora estas ferramentas tragam benefícios claros para as universidades, a sua implementação exige um investimento

https://doi.org/10.48798/congressobad.3240

elevado, mesmo em soluções *open source*, devido à falta de recursos humanos e técnicos internos. No caso do Pure da Elsevier, atualmente adotado pela Universidade Nova de Lisboa, trata-se de uma solução premium que reforça de forma significativa a visibilidade, a organização e a análise da produção científica, mas que, no entanto, envolve desafios relevantes em matéria de custos, limitações de personalização e até questões de privacidade (Sab, Ahamed KK, Bagalkoti 2024). A literatura recente tem vindo a explorar a integração de Knowledge Graphs (KGs) em sistemas CRIS, com o objetivo de promover o enriquecimento semântico dos metadados através do uso de ontologias e vocabulários padronizados. Esta integração reforçaria necessariamente a precisão e a interoperabilidade na pesquisa e recuperação de informação, contribuindo para sistemas mais consistentes e semanticamente robustos (Fabre & Azeroual, 2024). Neste contexto, destaca-se o OpenAIRE Graph enquanto uma das principais infraestruturas europeias baseadas em Knowledge Graph, concebida para agregar e interligar metadados provenientes de múltiplas fontes, incluindo repositórios institucionais, sistemas CRIS e bases de dados científicas (OpenAIRE, 2026). A Wikibase[1] poderá apresentar-se como uma solução adequada e viável para a implementação de um *Knowledge Graph-based* CRIS, pois, apesar das suas limitações como a estrutura RDF plana, a ausência de regras de validação e a falta de modelos visuais nativos, permite criar ontologias próprias e assegura controlo total sobre a curadoria dos dados sem depender do modelo superior da Wikidata[2] (Rossenova 2022). Como afirmam (Ruttenberg et al., 2019)"Através do Wikidata, as bibliotecas podem aplicar os seus conhecimentos especializados na criação de dados estruturados e na descrição de recursos num ambiente aberto, reutilizável e interoperável a nível global", estas ferramentas colaborativas permitem partilhar metadados de forma aberta e legível por máquina, alinhando-se com a visão da *web* semântica de Tim Berners Lee (Tharani, 2021). A estruturação de metadados constitui, desde sempre, uma competência fundamental no trabalho dos bibliotecários, assumindo especial importância quando utilizada como base para a avaliação da produção científica de unidades de investigação. Contudo, este processo enfrenta limitações significativas, decorrentes da dispersão dos metadados por múltiplas plataformas, da dependência de bases de dados proprietárias, como a Scopus (Elsevier) e a Web of Science (Clarivate Analytics), e da incapacidade destas ferramentas para agregarem de forma exaustiva toda a produção científica de uma unidade de investigação, sobretudo a que não se encontra indexada, o que dificulta uma gestão integrada e rigorosa dos metadados necessários à avaliação científica. É neste contexto que a Biblioteca Teresa e Alexandre Soares dos Santos apresenta o SciLinkDB, uma instância Wikibase destinada a integrar e estruturar, de forma independente e sustentável, os metadados de toda a produção científica da Nova School of Business & Economics, assegurando o cumprimento dos princípios FAIR (*Findable, Accessible, Interoperable and Reusable*) e preparando o caminho para o desenvolvimento futuro de um CRIS baseado em *Knowledge Graph* totalmente aberto.

Método

Inspirado num projeto-piloto[3] anterior desenvolvido pela Nova School of Business & Economics (NSBE) em cooperação com o Social Sciences Datalab, este projeto consiste num estudo bibliométrico observacional, baseado na análise descritiva dos metadados científicos produzidos pela NSBE. O trabalho é suportado por uma instância Wikibase própria (SciLinkDB) criada na *wikibase.cloud*[4], dedicada ao armazenamento e estruturação desses metadados, permitindo controlar o processo de importação sem descurar o respeito pelos princípios FAIR.

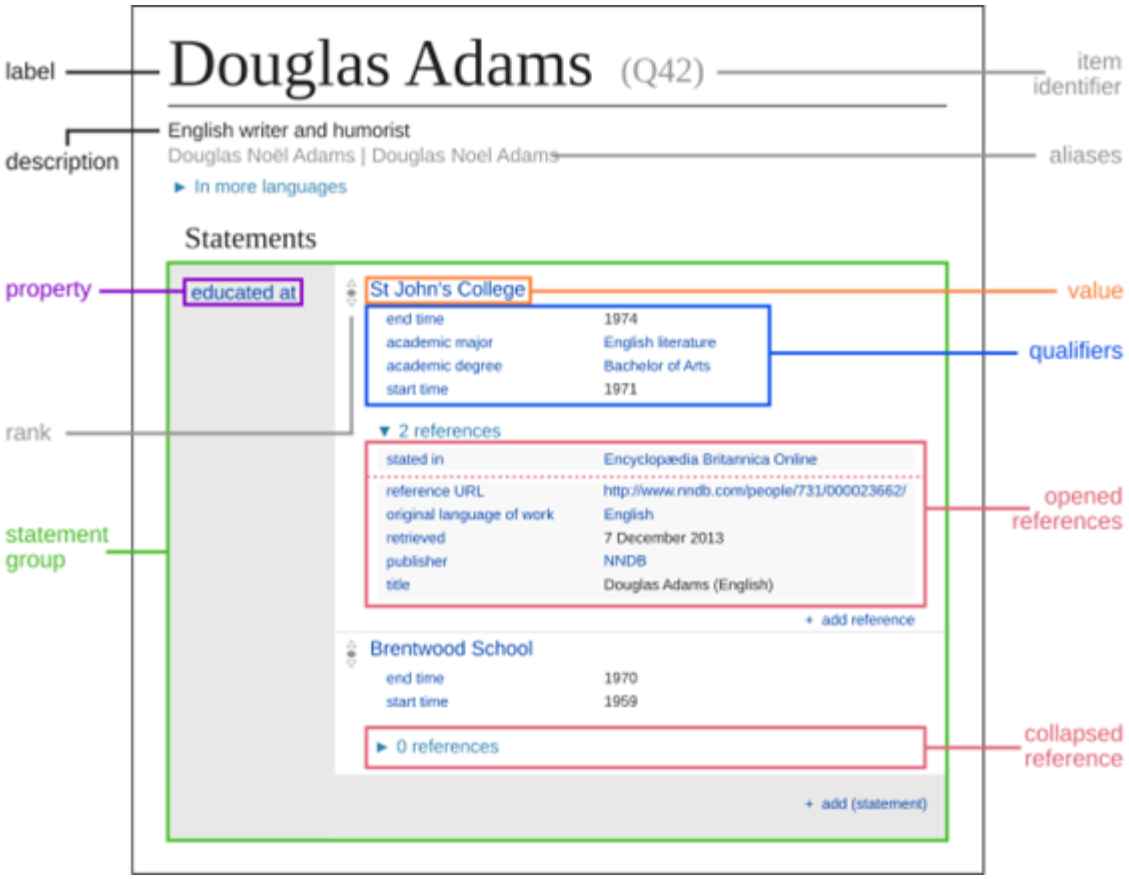


Figura 1 - Esquema de representação do modelo de dados no Wikidata, com declarações agrupadas e referências abertas, criado por Kritschmar (2022)ao abrigo da licença Creative Commons CC0 1.0 Universal Public Domain Dedication.

As ontologias[5] foram construídas com base nas propriedades (PIDs) já mapeadas no projeto-piloto, às quais foram acrescentadas outras mediante a necessidade de descrição de cada item. Uma vez definidas as ontologias, tornou-se necessário proceder à seleção e extração de metadados a partir de fontes fidedignas que assegurassem o nível de qualidade exigido para o projeto. Nesse contexto, optou-se pelos sistemas Renates[6] e Pure[7] uma vez que os metadados neles disponíveis foram objeto de processos de curadoria previamente realizados pelos profissionais responsáveis pela sua manutenção e atualização. No caso do RENATES, procedeu-se-à extração de dados relativos aos programas de doutoramento da Nova SBE, enquanto que no sistema PURE foram recolhidos os metadados associados à produção científica dos investigadores da Nova SBE. Os metadados foram extraídos em formato csv, analisados e refinados inicialmente no Excel, importados para o *OpenRefine*[8] onde foram limpos, reconciliados[9] e estruturados e posteriormente importados para a SciLinkDB.

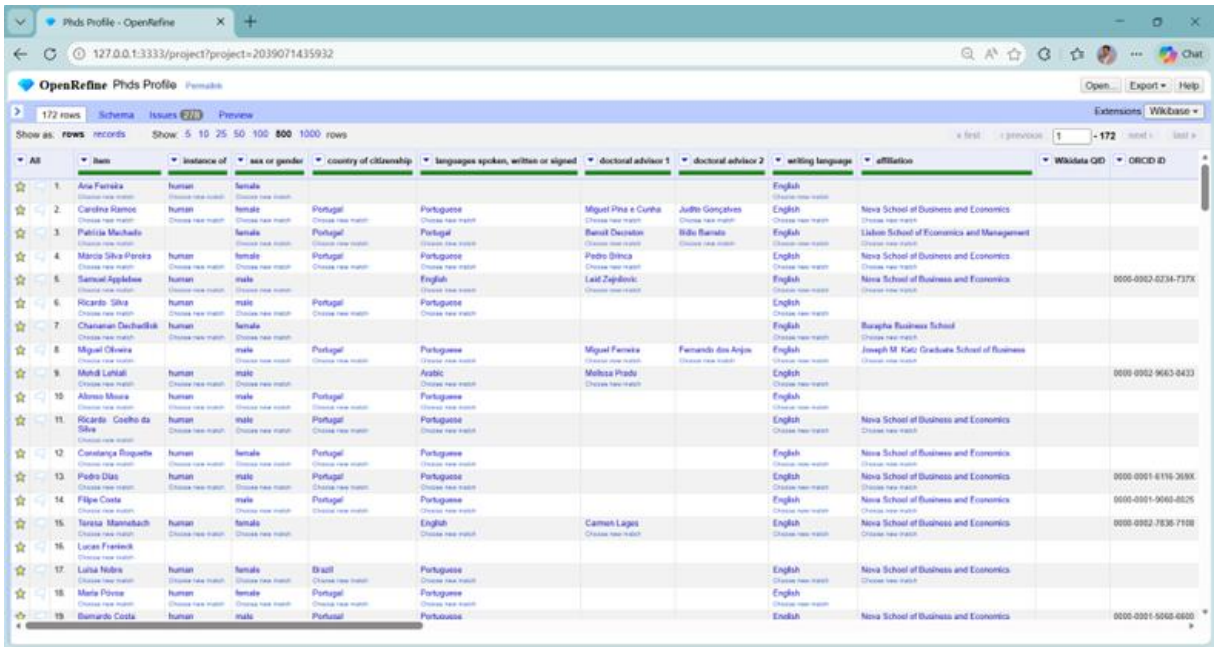
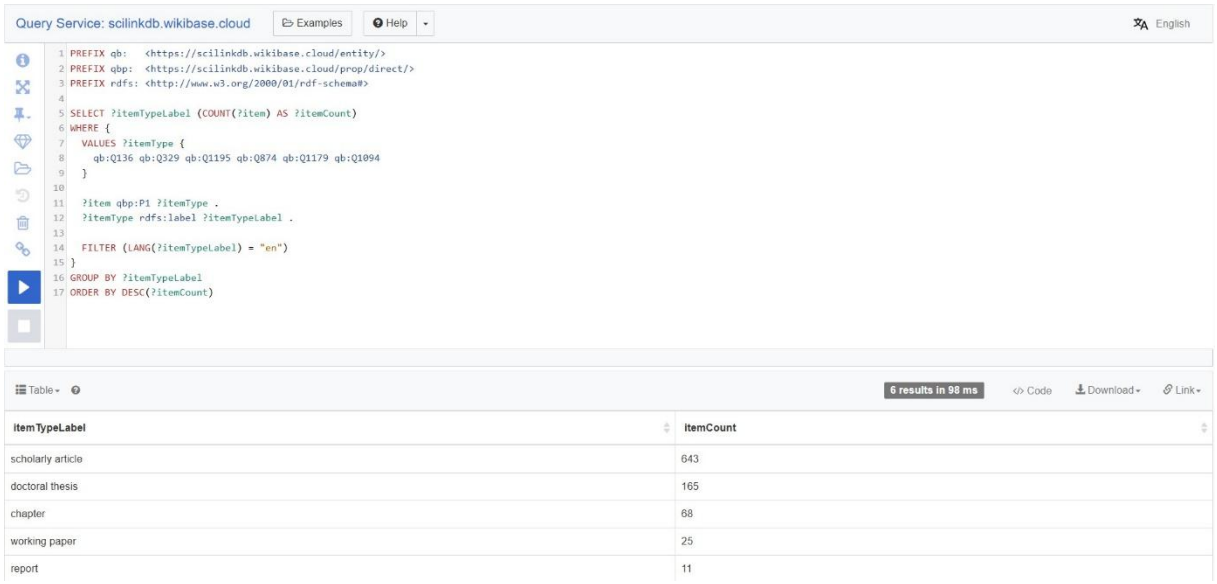


Figura 2 - Consulta SPARQL realizada no Query Service nativo da SciLinkDB, esta consulta permite saber o número total de items por tipo de documento.

O *endpoint*[10] para efeitos de reconciliação foi configurado num *container Docker*[11]. Para a análise dos dados recorreu-se num primeiro momento ao *Query Service*[12] nativo da Wikibase que permite efectuar consultas através da linguagem SPARQL[13].



Query Service: scilinkdb.wikibase.cloud

```
PREFIX qb:   <https://scilinkdb.wikibase.cloud/entity/>
PREFIX qbp:  <https://scilinkdb.wikibase.cloud/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?itemTypeLabel (COUNT(?item) AS ?itemCount)
WHERE {
  VALUES ?itemType {
    qb:Q136 qb:Q329 qb:Q1195 qb:Q874 qb:Q1179 qb:Q1094
  }

  ?item qbp:P1 ?itemType .
  ?itemType rdfs:label ?itemTypeLabel .

  FILTER (LANG(?itemTypeLabel) = "en")
}
GROUP BY ?itemTypeLabel
ORDER BY DESC(?itemCount)
```

6 results in 98 ms

itemTypeLabel	itemCount
scholarly article	643
doctoral thesis	165
chapter	68
working paper	25
report	11

Figura 3 - Consulta SPARQL realizada no Query Service nativo da Scilink DB, esta consulta permite saber o número total de items por tipo de documento.

Num segundo momento efetuaram-se alguns testes de integração com o Power BI[14] através da opção de carregamento de dados *Blank Query* e um *script* em linguagem M[15].

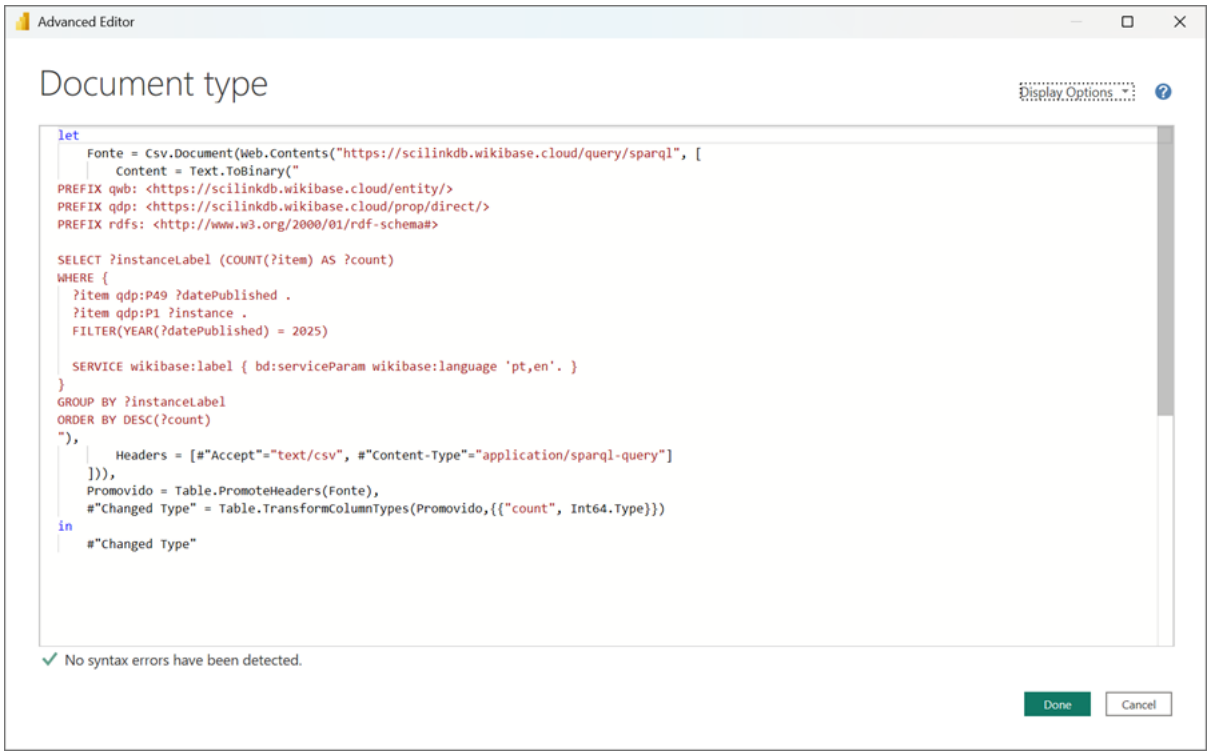


Figura 4 - Exemplo de um script construído em linguagem M para integração com o Power BI.

Recentemente foi desenvolvido o protótipo de visualização de metadados designado *Nova Research Graph* que consiste numa aplicação web construída em *Flask*[16] que funciona como uma plataforma de exploração académica baseada em entidades (autores, instituições e publicações). Ela integra dados provenientes da SciLinkDB (via consultas SPARQL), do OpenAlex[17] e Crossref[18] (citações e *abstracts*), criando perfis completos e navegáveis para diferentes tipos de entidades científicas. A produção de inferências e de todos os desenvolvimentos posteriores só foi possível graças à técnica de *vibe coding*[19], utilizando o LLM Copilot da Microsoft.

Resultados e Discussão

A SciLinkDB integra 237 propriedades e um número superior a 6 000 items[20], que descrevem recursos bibliográficos, investigadores, instituições, assuntos, entre outras entidades. Cada item apresenta, em média, 16 declarações[21], o que evidencia um nível de estruturação dos dados na plataforma considerado razoável. Tendo por inspiração o trabalho de Roszkowski (2023) sobre os padrões de representação das teses de Doutoramento na Wikidata, foram introduzidas na SciLinkDB os metadados relacionados com as teses[22] da Nova SBE disponíveis no repositório RUN[23]. A partir destes metadados foi possível gerar inferências no *Query Service* da SciLinkDB e gerar um *dashboard* em Microsoft Power BI.

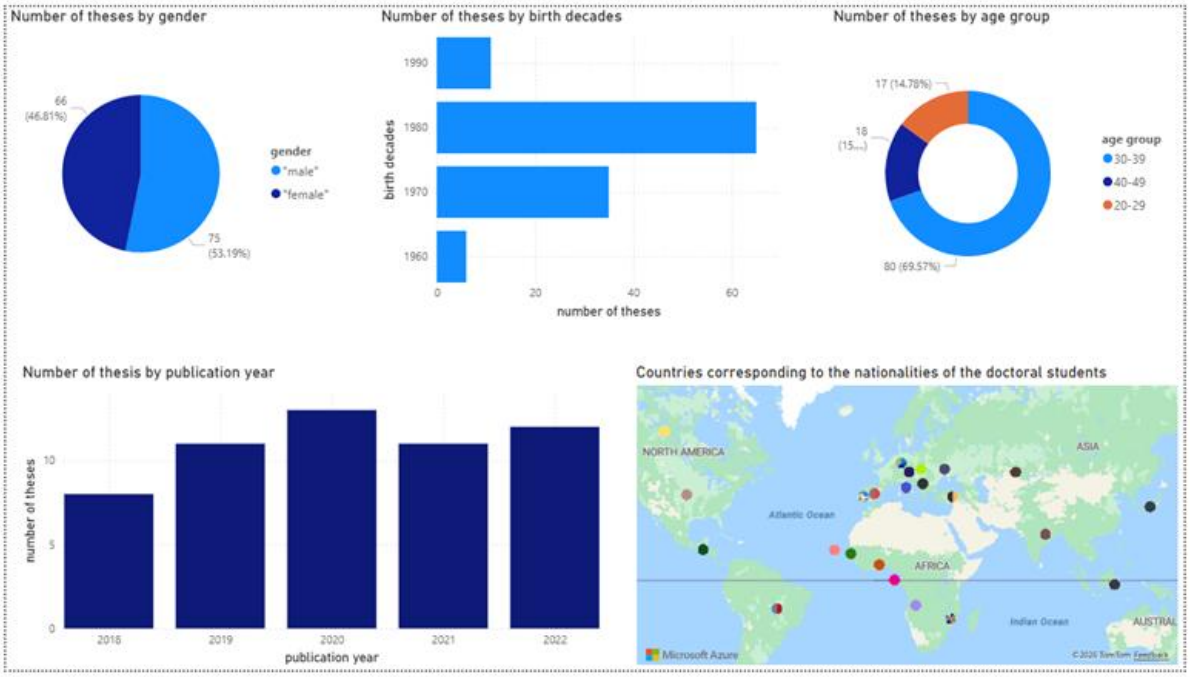


Figura 5 - Dashboard em Power BI com os dados relacionados com as teses de doutoramento da Nova SBE, criado no dia 11 setembro de 2025.

Para além das teses de Doutoramento foram também introduzidos na SciLinkDB metadados relacionados com os artigos científicos produzidos pela Nova SBE em 2025. Estes dados permitiram também gerar inferências no Power BI.

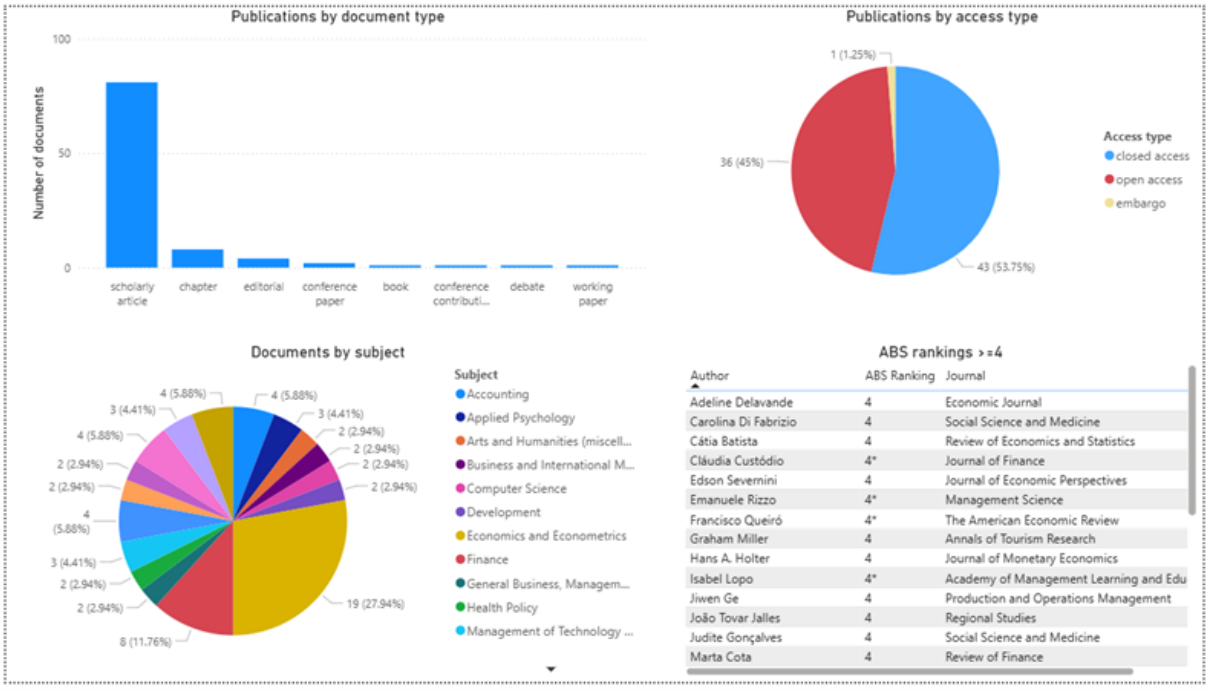


Figura 6 - Dashboard em Power BI com os dados relacionados com os resultados da investigação científica da Nova SBE, criado no dia 20 outubro de 2025.

A consciência da necessidade de desenvolvimento de uma ferramenta de visualização de dados, surge após a insatisfação com os resultados da experiência de emulação da ferramenta Scholia, para que esta corresse localmente, quer nas visualizações que a ferramenta possibilita, quer no trabalho hercúleo que

exigiria a reconfiguração do endpoint e das consultas existentes. O NRG visa a integração, exploração e análise sistemática de dados académicos provenientes de infraestruturas abertas, com foco na representação estruturada da atividade científica. Para o seu desenvolvimento recorreu-se à linguagem Python, utilizando a *framework* Flask para a construção da aplicação web.

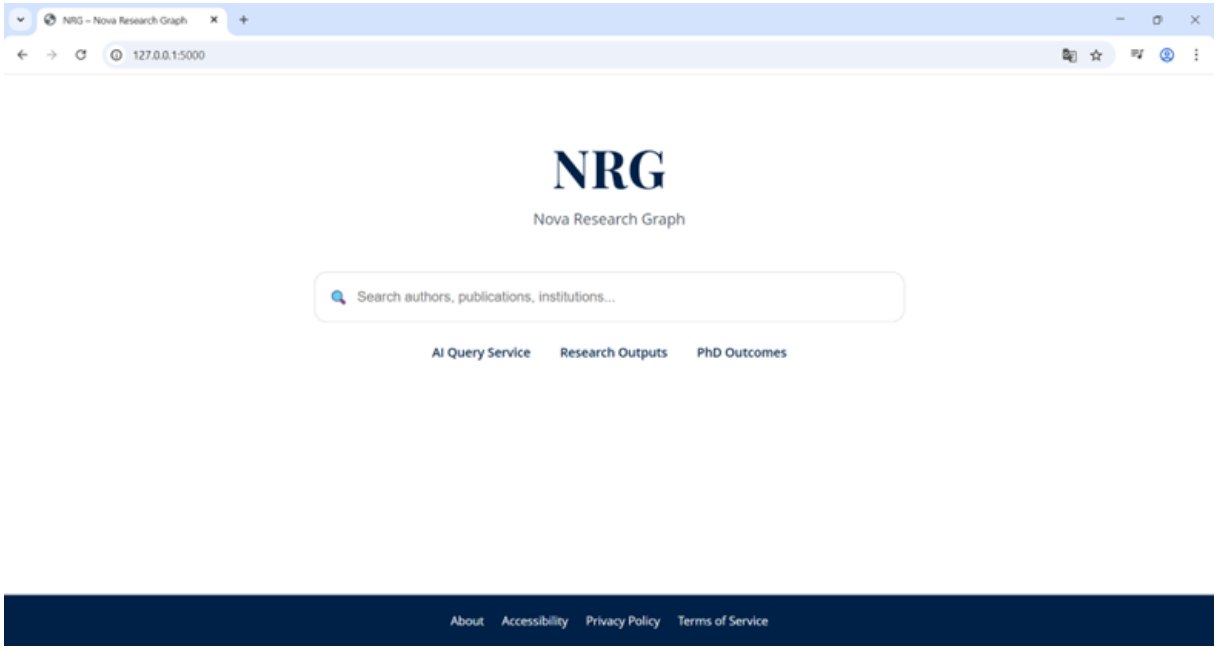


Figura 7 - Página principal do protótipo Nova Research Graph, contém uma caixa de pesquisa e três submenus que são o AI Query Service, Research Outputs e PhD Outcomes.

A integração de dados científicos é assegurada através da SciLinkDB, explorada por meio de consultas SPARQL, da API do Open Alex para obter o número de citações e do CrossRef para obter os resumos de cada artigo científico. Para além dos submenus da página inicial, o NRG permite neste momento a consulta por instituição, publicação, investigador, considerando que a qualidade dos resultados depende dos dados que estes perfis tenham, ou não, associados na SciLinkDB.

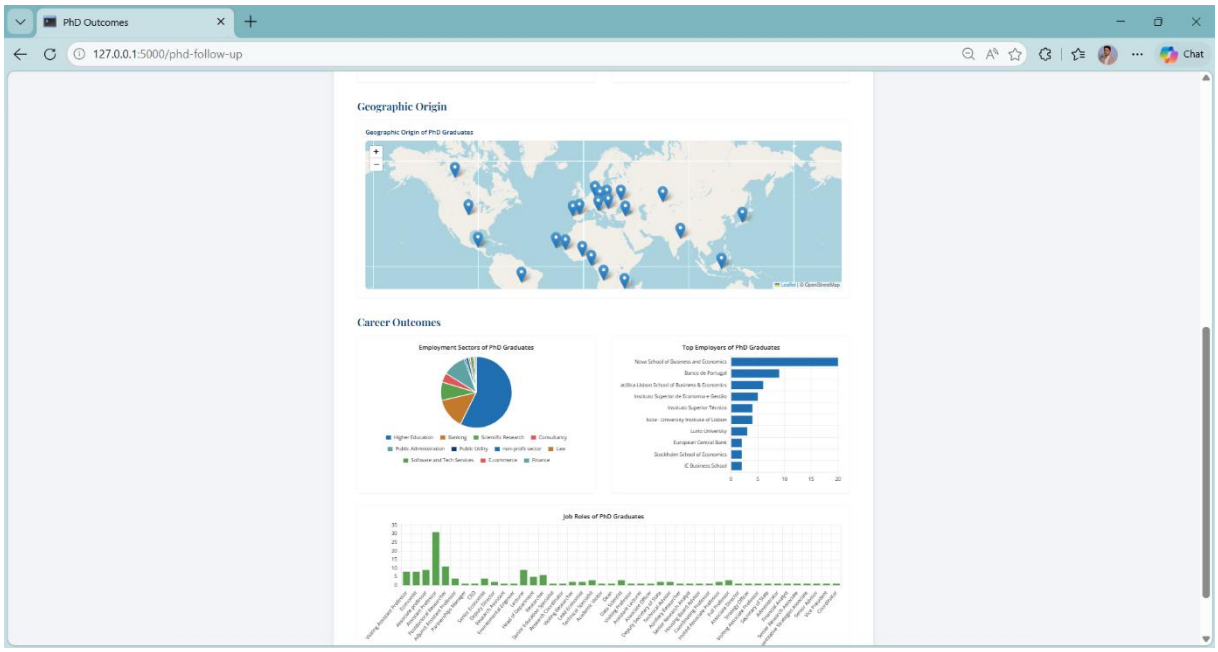


Figura 8 - Detalhe do submenu "PhD Outcomes" do protótipo Nova Research Graph, permite visualizações de caracter demográfico e geográfico, assim como de carreira profissional dos Doutorados.

Adicionalmente, a plataforma incorpora um *AI Query Service*, baseado no modelo de inteligência artificial local qwen2.5:3b-instruct[24], que permite a interpretação de questões formuladas em linguagem natural e a sua conversão automática em consultas semânticas, possibilitando a consulta a utilizadores que não dominem o SPARQL. A devolução do SPARQL permite validar a resposta e replicar a consulta na SciLinkdb.

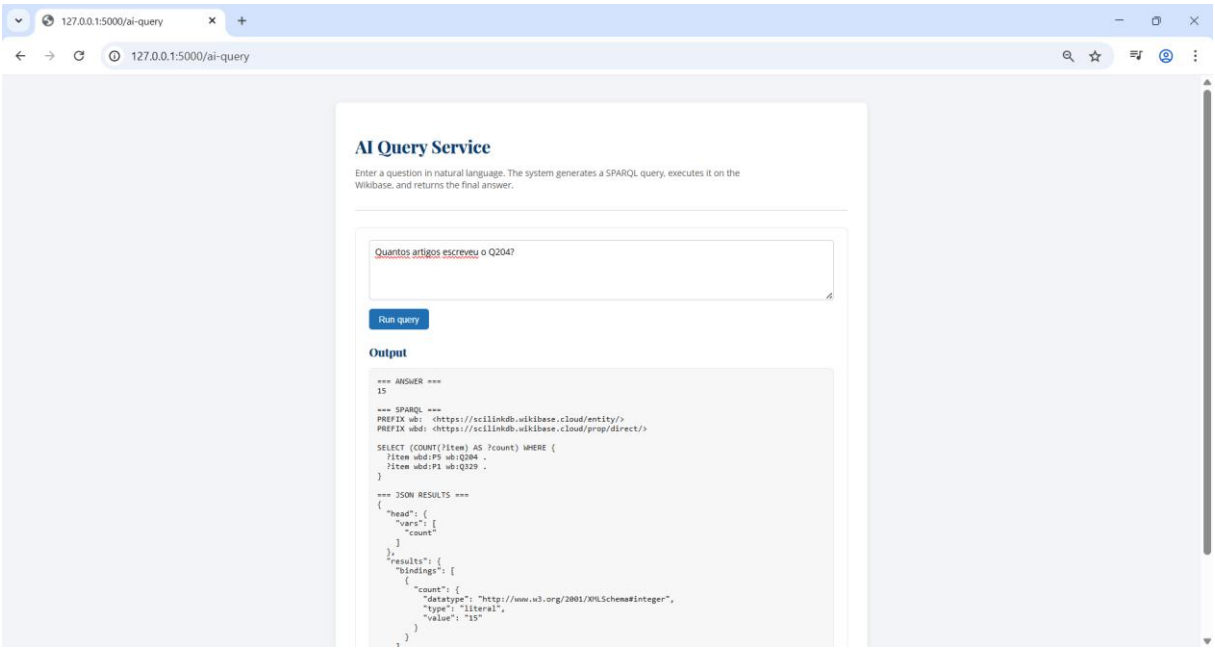


Figura 9 - Submenu "AI Query Service"do protótipo Nova Research Graph, permite consultas em linguagem natural e retorna não só a resposta à pergunta, como também o SPARQL gerado pelo modelo AI

Os CRIS constituem ferramentas indiscutivelmente robustas para a gestão da informação científica, permitindo um controlo rigoroso e sistemático da atividade de investigação. Contudo, a sua adoção acarreta custos de aquisição elevados para as instituições de ensino superior, bem como uma dependência significativa ao nível da usabilidade, uma vez que se tratam, na sua maioria, de soluções proprietárias. Por contraste, a Wikibase oferece um elevado grau de liberdade no acesso às fontes de informação e na modelação dos metadados, embora transfira para a instituição a totalidade do esforço associado à sua implementação e manutenção, dado não ter sido concebida especificamente para este fim. Ainda assim, a partir das visualizações produzidas pelo protótipo NRG, é possível sustentar que o desenvolvimento de um CRIS assente num *Knowledge Graph* é viável e capaz de gerar resultados satisfatórios no que respeita à visualização de metadados científicos. Perspetiva-se que, num futuro próximo, este projeto adquira a escala necessária para produzir um impacto significativo na gestão da ciência na Nova SBE, o que exigirá a automatização de processos e o desenvolvimento de novas formas de interoperabilidade com fontes abertas e fidedignas.

Dimensão	CRIS proprietários	Wikibase	CRIS baseado em Knowledge Graph (NRG)
Adequação à gestão da ciência	Elevada, por serem sistemas concebidos especificamente para esse fim.	Limitada, por não ter sido desenvolvida como sistema CRIS.	Elevada, resultante da adaptação do modelo de *Knowledge Graph* ao contexto CRIS.
Custos e dependência	Custos elevados e forte dependência do fornecedor.	Custos diretos reduzidos, com elevada exigência técnica institucional.	Custos de licenciamento reduzidos, exigindo investimento em desenvolvimento e automação.

Modelação e flexibilidade dos dados	Limitada aos modelos definidos pelo sistema.	Elevada liberdade na modelação de metadados.	Elevada flexibilidade semântica, adequada à informação científica.
Visualização de metadados científicos	Funcional, mas pouco flexível.	Não nativa, dependente de desenvolvimento adicional.	Visualizações eficazes, conforme demonstrado pelo protótipo NRG.
Escalabilidade e impacto institucional	Condicionada por custos e *vendor lock-in.*	Limitada enquanto solução isolada.	Elevado potencial de impacto, dependente da interoperabilidade e automatização futura.

Figura 10 - Tabela síntese comparativa das abordagens analisadas, construída com recurso ao M365 Copilot.

Conclusões

Este trabalho analisou as potencialidades e limitações da Wikibase enquanto suporte à gestão da informação científica, propondo uma abordagem alternativa dos CRIS baseados em *Knowledge Graph*. Os resultados obtidos a partir do protótipo NRG demonstram que é possível desenvolver um CRIS flexível e funcional, capaz de produzir visualizações eficazes de metadados científicos, constituindo uma alternativa viável a soluções proprietárias. Conclui-se que esta abordagem apresenta elevado potencial para apoiar a gestão da ciência na Nova SBE, desde que acompanhada por estratégias de automação e por um reforço da interoperabilidade com fontes abertas e fidedignas, condição essencial para a sua escalabilidade e sustentabilidade futura.

Referências bibliográficas

Bryant, R., Clements, A., De Castro, P., Cantrell, J., Dortmund, A., Fransen, J., Gallagher, P., & Mennielli, M. (2018). *Practices and Patterns in Research Information Management Findings from a Global Survey*.

de Castro, P. (2024). Improving CRIS features to support new Open Access implementation workflows at institutions. *Procedia Computer Science*, *249*(C), 179–185. https://doi.org/10.1016/j.procs.2024.11.062

Dias, J., Boavida, C. P., & Amante, M. J. (2018). As bibliotecas de Ensino Superior e a gestão de ciência. *Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas*, (13). https://publicacoes.bad.pt/revistas/index.php/congressosbad/article/view/1747

Fabre, R., & Azeroual, O. (2024). Knowledge Graphs – The Future of Integration in CRIS Systems for Uses of Assistance to Scientific Reasoning. *Procedia Computer Science*, *249*(C), 264–279. https://doi.org/10.1016/j.procs.2024.11.072

Kritschmar, C. (2022). *Datamodel in Wikidata*. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Datamodel_in_Wikidata_PT-EU.svg

Mimoso Correia, M. (2019). *Visualização de Q392754_ Criação de perfis académicos com Wikidata e Scholia – projecto-piloto.*

OpenAIRE. (2026). *OpenAIRE Graph - What is the OpenAIRE Graph*. https://graph.openaire.eu/what-is-the-openaire-graph#Id-Model

Rossenova, L. (2022). *Examining Wikidata And Wikibase In The Context Of Research Data Management Applications A Look At Wikidata's Past And Present*. https://swib.org/swib21/slides/05-03-gayo.pdf

Roszkowski, M. (2023). Modelling doctoral dissertations in Wikidata knowledge graph: Selected issues. *Journal of Academic Librarianship*, *49*(1). https://doi.org/10.1016/j.acalib.2022.102658

Ruttenberg, J., Task Force on Wikimedia, A., & Open Data, L. (2019). *ARL White Paper on Wikidata: Opportunities and Recommendations*.

Sab, C. M., Ahamed KK, M., & Bagalkoti, V. (2024). Unlocking Research Potential with Pure Portal: A Review of Elsevier's RIMS Solution. *Journal of Data Science, Informetrics, and Citation Studies*, *3*(3), 374–379. https://doi.org/10.5530/jcitation.3.3.41

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship*, *47*(2), 102326. https://doi.org/10.1016/J.ACALIB.2021.102326

Udartseva, O. M. (2024). Webometric Assessment of Foreign Information Systems of Current Research. *Organizatsiya i Metodika Informatsionnoi Raboty*, *51*(2), 6–14. https://doi.org/10.3103/S0147688224700023

[1] Software livre e de código aberto que permite aos utilizadores criar e gerir dados estruturados e interligados, bem como construir as suas próprias bases de conhecimento. Disponível em: https://diff.wikimedia.org/2025/07/20/the-basics-of-wikibase/
[2] Base de conhecimento livre e aberta que pode ser lida e editada tanto por humanos como por máquinas. Disponível em: https://www.wikidata.org/wiki/Wikidata:Main_Page
[3] Neste projeto-piloto foi utilizada a Wikidata como plataforma para organizar e enriquecer dados científicos da NSBE (Mimoso Correia 2019)
[4] Espaço colaborativo para indivíduos e grupos contribuírem, editarem e organizarem informação de forma estruturada, ajudando-o a transformar os seus dados em conhecimento significativo. Disponível em: https://www.wikibase.cloud/
[5] Representações formais, explícitas e compartilhadas de um domínio de conhecimento.
[6] Plataforma que recolhe informação oficial sobre teses de doutoramento e dissertações de mestrado realizadas em Portugal. Disponível em: http://renates.dgeec.mec.pt
[7] Sistema de gestão de informação científica. Disponível em: https://www.elsevier.com/products/pure
[8] Aplicação de ambiente de trabalho de código aberto para limpeza e transformação de dados noutros formatos. Disponível em: https://en.wikipedia.org/wiki/OpenRefine
[9] Processo de comparar dados de múltiplas fontes para garantir consistência e precisão, identificando e corrigindo discrepâncias.
[10] "Fim" de um canal de comunicação. Dependendo do contexto, refere-se geralmente a um dispositivo físico ou a um endereço de software específico.
[11] Unidade padrão de software que engloba o código e todas as suas dependências, permitindo que a aplicação seja executada de forma rápida e fiável em diferentes ambientes computacionais. Disponível em: https://www.docker.com/resources/what-container/
[12] Disponível em: https://scilinkdb.wikibase.cloud/query/
[13] Linguagem padrão da web semântica para a consulta de grafos RDF.
[14] Software interativo de visualização de dados desenvolvido pela Microsoft.
[15] Linguagem de programação usada no Power Query para importar, transformar e preparar dados antes da análise.
[16] Microframework de desenvolvimento web criado para quem quer construir aplicações web. Definição retirada de: https://hub.asimov.academy/blog/flask-o-que-e-para-que-serve/
[17] Base de dados bibliográfica em acesso aberto e gratuito.
[18] Organização sem fins lucrativos que atua como infraestrutura digital aberta para a comunidade científica global.
[19] Técnica de programação orientada por Inteligência Artificial.
[20] Consulta disponível em: https://tinyurl.com/29m6n65c
[21] Consulta disponível em: https://tinyurl.com/2ac8zx4v
[22] Consulta disponível em: https://tinyurl.com/23np7jdp
[23] Repositório institucional da Universidade Nova de Lisboa.
[24] Modelo de linguagem grande (LLM) da família Qwen 2.5, desenvolvida pela Alibaba Cloud.