Universidade do Minho
Escola de Engenharia

Diogo António Ribeiro Gomes

RDProfile – Integração e Interoperabilidade

Outubro de 2023

RDProfile – Integração e Interoperabilidade
Diogo António Ribeiro Gomes
UMinho | 2023

Universidade do Minho
Escola de Engenharia

Diogo António Ribeiro Gomes

RDProfile – Integração e Interoperabilidade

Relatório de Trabalho de Projeto de Dissertação de Mestrado
Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação do
Professor Carlos Filipe da Silva Portela
Professor Jorge Oliveira Sá

Outubro de 2023

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

AGRADECIMENTOS

A realização desta dissertação assinala o fim de todo o percurso académico que me permitiu crescer como pessoa e, certamente, me prepara para os obstáculos que a vida profissional me reserva.

Assim sendo, quero desejar o meu sincero agradecimento a todos aqueles que me ajudaram nesta etapa nomeadamente:

Ao professor e orientador, Filipe Portela, um agradecimento por todo apoio, acompanhamento e disponibilidade desde o início até ao fim deste projeto;

Ao professor e coorientador Jorge Oliveira e Sá, um agradecimento pelo apoio ao longo deste trabalho;

Aos meus colegas e amigos que me acompanharam e apoiaram durante este percurso de académico;

À minha namorada por me ter ajudado e motivado nos momentos mais difíceis durante toda esta etapa;

À minha família, especialmente aos meus pais, pelos valores que me transmitiram e apoio incondicional durante esta fase;

Deixo também um agradecimento especial à Universidade do Minho que me acolheu e contribuiu para a pessoa que sou hoje.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio, nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

RDProfile – Integração e Interoperabilidade

Esta dissertação está enquadrada num projeto de dissertação de Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação em conjunto com o Centro de Investigação Algoritmi sobre o tema "RDProfile – Integração e Interoperabilidade".

A criação deste tema provém do crescimento exponencial de investigadores em Portugal, em conjunto com o aparecimento de várias plataformas científicas, o que gera uma grande dificuldade em gerir os perfis científicos. Surge assim, a necessidade da criação de um artefacto que seja capaz de agrupar os dados espalhados por estas plataformas científicas, cruzando as suas informações através da integração de *Application Programming Interfaces* (API) das plataformas de indexação, como SCOPUS, Google Scholar, Web of Science, entre outras, com o objetivo de os investigadores terem uma fonte precisa de informação sobre o seu impacto na comunidade científica.

Este projeto de dissertação foi dividido em duas partes durante a sua elaboração: na primeira fase foram estudadas as várias plataformas científicas com o objetivo de compreender o funcionamento das mesmas. Além destas, também foram estudadas duas plataformas similares ao RDProfile permitindo identificar as fraquezas a serem colmatadas. A segunda parte do projeto foi focada no desenvolvimento de dois algoritmos sendo o primeiro capaz de identificar referências erradas no Scopus dos investigadores e o segundo capaz de agregar e cruzar informação, nomeadamente artigos e as suas respetivas citações de diversas plataformas. Também foram detalhados o funcionamento de cada algoritmo e a estrutura de dados, tendo ainda sido realizados testes a investigadores para verificar o funcionamento de ambos os algoritmos.

Este projeto de dissertação teve como contributos o estudo de várias plataformas científicas, tais como Scopus, Google Scholar, Web of Science, ORCID, Research Gate e Semantic Scholar, um *crawler* capaz de obter informações da plataforma Google Scholar, um algoritmo capaz de identificar referências erradas no Scopus e um algoritmo capaz de aproximar as citações de um artigo ao número real, integrando 4 API externas (ORCID, Scopus, Semantic Scholar e OpenAlex).

Palavras-chave: Citação, Integração, Métricas, *Pervasive,* Plataformas Científicas

ABSTRACT

RD – Integration and Interoperability

This dissertation is part of a dissertation project for an Integrated Master's Degree in Information Systems Engineering and Management in conjunction with the Algoritmi Research Center on the subject of "RDProfile - Integration and Interoperability".

The creation of this topic stems from the exponential growth of researchers in Portugal, together with the emergence of various scientific platforms, which creates great difficulty in managing scientific profiles. Thus, the need arises to create an artifact that is capable of grouping the data spread across these scientific platforms, cross-referencing their information through the integration of Application Programming Interfaces (API) from indexing platforms such as SCOPUS, Google Scholar, Web of Science, among others, with the aim of researchers having an accurate source of information about their impact on the scientific community.

This dissertation project was divided into two parts during its development: in the first phase, the various scientific platforms were studied in order to understand how they work. In addition to these, two platforms similar to RDProfile were also studied in order to identify the weaknesses to be overcome. The second part of the project focused on developing two algorithms, the first capable of identifying incorrect references in researchers' Scopus databases and the second capable of aggregating and cross-referencing information, namely articles and their respective citations from various platforms. The functioning of each algorithm and the data structure were also detailed, and tests were carried out on researchers to verify the functioning of both algorithms.

This dissertation project contributed to the study of various scientific platforms, such as Scopus, Google Scholar, Web of Science, ORCID, Research Gate and Semantic Scholar, a crawler capable of obtaining information from the Google Scholar platform, an algorithm capable of identifying wrong references in Scopus and an algorithm capable of bringing the citations of an article closer to the real number, integrating 4 external API (ORCID, Scopus, Semantic Scholar and OpenAlex).

Keywords: Citation, Integration, Metrics, Pervasive, Scientific Platforms

ÍNDICE

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
DBLP	*Digital Bibliography & Library Project*
DOI	*Digital Object Identifier*
DSR	*Design Science Research*
EID	*Electronic Identifier*
FCT	Fundação para a Ciência e Tecnologia
HTML	*Hypertext Markup Language*
IA	Inteligência Artificial
JSON	*JavaScript Object Notation*
KPI	*Key Performance Indicators*
ORCID	*Open Researcher and Contributor Identifier*
PDF	*Portable Document Format*
URL	*Uniform Resource Locator*
VPN	*Virtual Private Network*
XML	*Extensible Markup Language*

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Este capítulo detalha o enquadramento e motivação, objetivos e resultados esperados deste projeto de dissertação, bem como a estrutura do documento.

1.1 Enquadramento e Motivação

Devido ao crescimento exponencial do número de investigadores existentes em Portugal, o surgimento e utilização de diversas plataformas para a indexação de publicações e gestão de perfis científicos aumentou, sendo exemplos: *Open Researcher and Contributor Identifier* (ORCID), Scopus, Google Scholar, entre outros (Portela, 2022).

Com esta quantidade de plataformas, é difícil a correspondência entre conteúdos, isto é, a informação não é transversal a todas as plataformas e os investigadores raramente têm os seus perfis atualizados. Além destes inconvenientes, a pesquisa de investigadores por tópicos é demorada e individualizada. Tendo em conta estes problemas, foi criado o RDProfile, uma plataforma capaz de uniformizar e agregar dados dos investigadores de várias plataformas de indexação, permitindo uma consulta em várias dimensões (tempo, grupos, equipas, laboratórios, instituição), acompanhado de um sistema de gamificação que irá permitir classificar cada investigador. Além disto, irá permitir a exportação de alguns dados fundamentais para os relatórios anuais (Portela, 2022).

Assim, este projeto de dissertação pretende dar continuidade ao desenvolvimento da plataforma RDProfile que vem preencher uma grande lacuna de informação sobre os perfis científicos dos investigadores, integrando as bibliotecas, métricas e *Key Performance Indicators* (KPI) das várias plataformas acima mencionadas numa só aplicação, facilitando a gestão por parte dos investigadores e instituições científicas.

Este trabalho foca-se na área de desenvolvimento *web*, contribuindo para a melhoria da forma como os investigadores em Portugal acedem à informação necessária aos seus projetos de investigação. Concentra-se, especificamente, na integração de *Application Programming Interfaces* (API) de diversas plataformas científicas e adição de novas funcionalidades ao nível das citações, integração e interoperabilidade de diversos perfis científicos.

1.2 Objetivos

Este projeto de dissertação tem como objetivo estudar a viabilidade da combinação de dados de diversas plataformas de forma a otimizar o protótipo existente e, acima de tudo, explorar uma alternativa mais fiável para o mecanismo de citações que tem diversas falhas, como é o caso do Scopus.

Desta forma, o presente trabalho visa responder à seguinte questão de investigação: *"De que forma é possível recolher e agrupar os dados de várias plataformas científicas?".*

Definem-se ainda como objetivos secundários deste projeto de dissertação:

- Criar um *web crawler*, um programa que sistematicamente procura e indexa artigos para que a pesquisa dos mesmos seja a mais rápida e precisa possível (Cho et al., 2001);
- Integrar diversas API das plataformas, como o ORCID, Scopus, Google Scholar, Web of Science, Research Gate, entre outras bibliotecas a explorar;
- Desenvolver algoritmos capazes de agregar e cruzar dados provenientes das várias plataformas;
- Desenvolver modelos interoperáveis, de forma que os investigadores tenham acesso às suas informações mais facilmente, como citações, número de artigos publicados, entre outros.

Estes objetivos foram repartidos em 10 *sprints* usando o guia Scrum (Schwaber & Sutherland, 2020) e foi usada a metodologia *Design Science Research* (DSR) (Peffers et al., 2007) como abordagem metodológica durante toda a dissertação. Tudo isto culmina num protótipo funcional de uma solução *web pervasive* que permita a recolha e análise de dados científicos, sendo os modelos e algoritmo obtidos o resultado científico deste trabalho.

1.3 Estrutura do documento

Esta dissertação está estruturada em cinco capítulos. No primeiro capítulo é introduzido o tema do trabalho, é detalhado o seu motivo, são apresentados os objetivos deste projeto de dissertação e a estrutura do presente documento.

No segundo capítulo é feita a revisão de literatura, iniciada pela estratégia de pesquisa em que se detalha o método de pesquisa dos artigos e quais os critérios para os selecionar. Além disto são analisadas algumas plataformas científicas integradas e plataformas similares à plataforma RDProfile.

No terceiro capítulo são descritas as abordagens metodológicas que foram usadas neste projeto de dissertação, sendo elas o guia Scrum, para o desenvolvimento do artefacto e definição do *product backlog* e *sprints*, e a metodologia DSR, como guia de criação do artefacto que cumpra os objetivos propostos. São também listadas todas as API utilizadas na solução final.

O quarto capítulo foca-se na documentação do desenvolvimento das soluções e é divido em 4 partes. Na primeira parte é descrita a arquitetura definida. Na segunda parte detalha-se o processo manual de deteção de referências erradas na plataforma Scopus e é especificado o algoritmo concebido para automatizar o processo. Na terceira parte é evidenciado o algoritmo de agregação e cruzamento de dados, bem como os processos de obtenção de dados das plataformas científicas. Por fim, na quarta parte, são listadas as funcionalidades do RDProfile.

No quinto capítulo são expostos os testes realizados aos algoritmos desenvolvidos neste projeto de dissertação.

Finalmente, o sexto e último capítulo, contém as considerações finais, os riscos e dificuldades encontradas durante o desenvolvimento do projeto e sugestões para trabalho futuro.

2. REVISÃO DE LITERATURA

Este capítulo é composto pela estratégia de investigação, onde é detalhado o processo de pesquisa da informação, são enumerados os critérios de seleção da mesma e enuncia também todas as plataformas científicas exploradas neste projeto de dissertação para auxiliar o êxito do desenvolvimento da solução final.

2.1 Estratégia de pesquisa

Para a elaboração da revisão da literatura foram usadas diversas plataformas de indexação, como Scopus, Google Scholar e Google, com a finalidade de obter o máximo de informação adequada a partir de fontes credíveis para este projeto de dissertação.

Primeiramente foram estudadas todas as plataformas integradas no RDProfile, obtendo o máximo de informação sobre a forma como funcionam. As plataformas analisadas foram:

- Scopus;
- ORCID;
- Google Scholar;
- Web of Science;
- Research Gate;
- Semantic Scholar.

A seleção de artigos para revisão de literatura atendeu aos seguintes critérios:

- Idioma: Estar escrito em português ou inglês;
- Ano de publicação: Ter menos de 10 anos, salvo à exceção de artigos revelantes na área;
- Citações: Possuir, pelo menos, mais de 5 citações;
- Conteúdo do *abstract*, introdução e conclusão serem pertinentes para o tema.

2.2 RDProfile

Segundo Portela (2022), o RDProfile visa criar um solução inclusiva, fiável e escalável que pode facilmente incluir informação diversa dos perfis dos investigadores, sendo um agregador de conhecimento dispersos, disponível ao público e de fácil acesso. Além disso, a gamificação ajudará a identificar os melhores perfis científicos e a melhorar/corrigir indicadores que permitam aumentar o número de projetos I&D financiados com base na melhoria da qualidade das métricas dos investigadores". Por outras palavras, o RDProfile é uma plataforma com o objetivo de juntar toda a informação disponível num só sítio, fazendo com que os investigadores tenham acesso a informação crucial do seu impacto na comunidade científica. Assim, os investigadores poderão ser avaliados com base nas suas métricas num sistema de classificação justo e competitivo.

A plataforma tem vindo a ser desenvolvida, logo já possui algumas funcionalidades, como (Portela, 2022):

- Carregar automaticamente publicações anexas;
- Adicionar informação sobre investigadores e o seu trabalho científico;
- Interoperabilidade com bibliotecas externas (DBLP, SCOPUS, ISI Web of Science, Google Scholar, CROSS-REF);
- Consultar indicadores bibliométricos sobre investigadores e instituições;
- Consultar métricas (por exemplo, quartis, *h-index*, citações, SCI, SJR, JIF) sobre publicações científicas e investigadores;
- Consultar métricas sobre projetos e revisões;
- Consultar lista de publicações;
- Consultar estatísticas sobre as publicações (p.e. citações, menções, recomendações);
- Cruzar, fundir e validar dados de várias plataformas distintas;
- Exportar indicadores para relatórios da FCT e projetos de I&D;
- Filtros que facilitam referências cruzadas (por exemplo, publicações indexadas por tipo de livraria);

Neste projeto de dissertação, foram adicionadas duas funcionalidades:

- Detetar citações erradas na plataforma Scopus;
- Consultar o número real de citações dos artigos científicos.

2.3 Scopus

O Scopus[1] é uma base de dados abrangente, especializada em resumos e citações com dados enriquecidos e literatura académica ligada através de uma grande variedade de disciplinas. Esta plataforma científica é propriedade da Elsevier e cobre várias disciplinas académicas, como ciências, tecnologia, medicina, ciências sociais e humanidades. O Scopus tem um grande número de editoras indexadas e é uma das maiores bases de dados de resumos e citações da literatura revista por pares (Elsevier, n.d.).

Esta plataforma científica oferece a possibilidade de pesquisar com filtros avançados por autores, instituições, publicações e, adicionalmente, verificar a influência e o impacto que um indivíduo tem na comunidade científica. Neste projeto de dissertação, o foco é o processo de citação que será detalhado posteriormente e que usualmente sofre falhas fazendo com que a informação não seja correta.

Na Figura 1 apresentado um perfil de um investigador gerado automaticamente pelo Scopus em que é possível ter conhecimento de algumas informações, tais como o nome, o ID do investigador no Scopus e o ID do ORCID. São ainda apresentadas algumas estatísticas, como o número de citações e a quantidade de documentos, o número de coautores e o *h-index* do investigador.

Figura 1: Perfil do investigador (Scopus)

Esta plataforma científica também possui uma *dashboard* em que o utilizador pode visualizar detalhadamente as citações dos documentos. Na Figura 2 é demonstrado um gráfico com as citações dos artigos do utilizador acompanhado de alguns filtros, como intervalo de tempo, exclusão de autocitações e exclusão citações de livros.

[1] https://www.scopus.com/

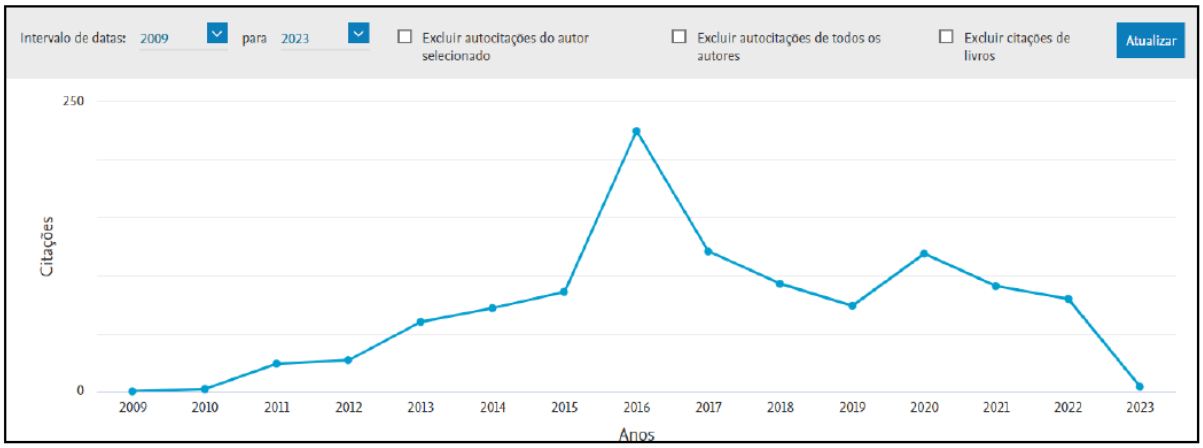


Figura 2: Gráfico de citações

Além disso, o utilizador tem ainda a possibilidade de visualizar o número de citações por documentos em cada ano. Na Figura 3 é apresentada uma tabela com o nome do documento, o ano de publicação e o número de citações feitas por ano. Adicionalmente, é indicado, por linha, o número total de citações do documento e, por coluna, o número total de citações do ano.

	Documentos	Citações	<2009	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Subtotal	>2023	Total
		Total	0	0	2	24	27	60	72	86	225	121	93	74	119	91	80	4	1078	0	1078
1	Apoio à decisão pervasivo e inteligente em medicina intensiva...	2014								9	19	4	9		3	2			46		46
2	A próxima geração de agentes de interoperabilidade na área da saúde	2014							2	4	13	1	5	4	5	4	6		44		44
3	INTCARE: Abordagem multiagente para decisões inteligentes em tempo real...	2011				4	2	10	5	4	6	3	3	1	4	1			43		43
4	Habilitando uma abordagem abrangente para suporte inteligente à decisão...	2011					3	8	5	5	6	1	1		2				31		31
5	Modelos de mineração de dados para prever a readmissão do paciente em int...	2014							1	3	9	5	1	3	4	3			29		29
6	Uma análise de benchmarking de inteligência de negócios de código aberto...	2016										5	7	3	7	4	1		27		27
7	Descoberta de conhecimento para inteligência difusa e em tempo real ...	2011					4	9	4	2	3	1	2	1	1				27		27
8	Prontuário eletrônico em serviço de dermatologia	2011					3	2	5	2	4	2		5	1	2	1		27		27
9	Prevendo o tipo de parto pela identificação do procedimento obstétrico...	2015									1	5	4	4	2	6	3	1	26		26

Figura 3: Tabela de citações (Scopus)

Cada investigador tem à sua disposição vária métricas que permitem verificar a relevância dos seus trabalhos científicos (Elsevier, 2023b):

- ***H-Index*:** É o número de artigos com o número de citações ≥ h como um índex que caracteriza o *output* de um investigador e quantifica o seu impacto cumulativo na investigação científica. Um investigador tem um índex h se os seus N artigos têm pelo menos h citações e os outros (N – h) têm ≤ h citações cada (Hirsch, 2005). Por exemplo, se um investigador possuir 7 artigos com 23, 15, 14, 10, 6, 4, 2 citações respetivamente, o seu *H-Index* será 4, de acordo com a fórmula.

- ***CiteScore*:** Mede o número médio de citações por documento. O valor desta métrica resulta da divisão do número de citações de documentos revistos por pares, como artigos, papéis de conferência, capítulos de livros, entre outros, pelo número total de documentos criados no intervalo de 4 anos.
- ***Citation count*:** É a soma de todas as citações recebidas num intervalo de 4 anos por documentos publicados também nesse mesmo intervalo (numerador da métrica *CiteScore*).
- ***Document Count*:** É a soma de todos os documentos publicados no intervalo de 4 anos (denominador da métrica *CiteScore*).

No que toca ao processo de citação, o Scopus usa um algoritmo próprio para calcular o número de citações que um determinado artigo ou autor recebeu. Primeiramente, este algoritmo procura frequentemente por novos artigos ou trabalhos académicos e adiciona-os à base de dados. Quando um artigo é encontrado, este é analisado pelo algoritmo com especial enfoque nas citações presentes no texto e na lista de referências. Todas as citações encontradas são adicionadas ao contador do artigo citado, que será usado para determinar o impacto do mesmo e o *h-index* do autor. No fim deste processo, a informação é partilhada com os utilizadores do Scopus, onde pode ser usada para averiguar o impacto do artigo, a quantidade de citações, entre outras métricas.

É importante salientar que o algoritmo só é capaz de contar citações de artigos corretamente indexados pelo Scopus, logo artigos citados, mas não indexados não serão contados para o número de citações. Na Figura 4 é representado um diagrama do funcionamento do algoritmo, de forma sintetizada.

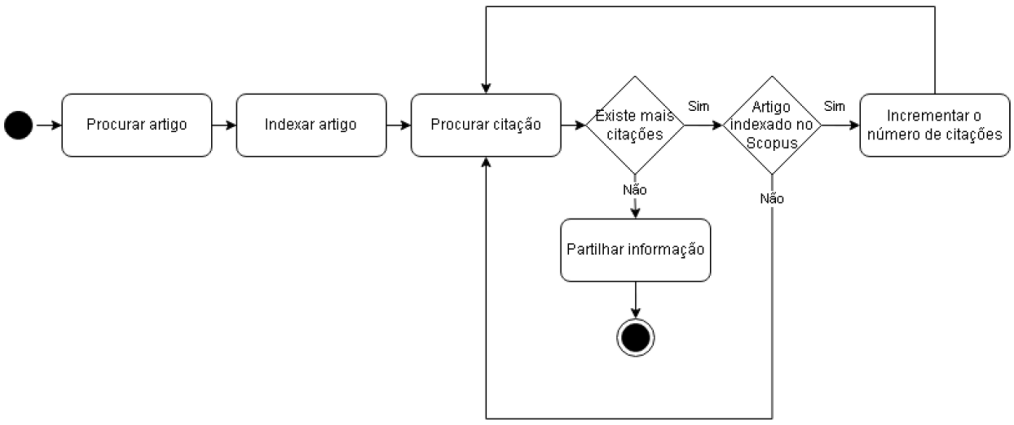


Figura 4: Fluxograma do processo de citações

Apesar do Scopus ser uma plataforma científica bastante completa com várias métricas e *dashboards* pelas quais os investigadores se podem guiar e verificar o seu impacto na comunidade científica, esta apresenta algumas falhas, como a não correta citação de artigos, o que prejudica o trabalho dos investigadores, visto que não têm uma medição precisa do seu impacto. Assim, o RDProfile surge na tentativa de colmatar esta falha, agregando os dados das várias plataformas para os investigadores terem maior confiança nos valores das métricas dos seus perfis.

2.4 Google Scholar

O Google Scholar[2] é um motor de busca que indexa vários artigos, editoras e bases de dados científicas. A partir deste motor de busca é possível pesquisar artigos, teses, livros, resumos e opiniões judiciais, de editoras académicas, repositórios, universidades e outros sítios da *web* (Google, n.d.-a).

O mecanismo de indexação do Google Scholar funciona na base de *web crawlers* que são capazes de navegar pela rede *web* para identificar e indexar ficheiros. Este *software* necessita que o *website* esteja estruturado de uma certa forma, de maneira que os *crawler*s obtenham os *Uniform Resource Locators* (URL) de todos os artigos e, também, para ser possível periodicamente atualizar o conteúdo do *website*. Os requisitos necessários para que o *web crawler* do Google Scholar indexe os artigos são (Google, 2023):

- **Formatos dos ficheiros:** Os ficheiros necessitam de estar no formato *Hypertext Markup Language* (HTML) ou *Portable Document Format* (PDF). Os ficheiros PDF necessitam de conter texto que seja possível de ser pesquisado. Cada ficheiro não pode exceder os 5MB de tamanho.
- **Interface de navegação:** É necessária uma interface de navegação para que os *crawlers* consigam descobrir os URLs dos artigos. A melhor prática é todos os artigos serem descobertos a partir da página inicial do *website.* Algumas recomendações do Google Scholar são:
    - Caso o investigador tenha um número reduzido de publicações hospedadas, como artigos escritos por um autor ou um pequeno grupo de autores, é

[2] https://scholar.google.com/

recomendada uma lista numa página HTML, por exemplo www.example.edu/~professor/publications.html e incluir as hiperligações para o texto do artigo em formato PDF.

- Caso o *website* tenha milhares de publicações, é necessário listar os artigos por ordem da data de publicação ou data da entrada do artigo.
- Caso o *website* tenha dezenas de milhares de artigos, é necessário criar uma interface de navegação que liste apenas artigos adicionados nas duas últimas semanas. Com esta dimensão reduzida da lista, os *crawlers* conseguem atualizar mais frequentemente os artigos.

Este *web crawler*, chamado Googlebot, atua em três fases:

- **Descoberta de URL:** Nesta primeira fase, o *web crawler* procura por páginas que existam na *web*. Por não existir um registo central de todas as páginas *web*, o *crawler* tem de pesquisar constantemente por páginas novas e atualizadas, adicionando-as posteriormente à sua lista de páginas conhecidas.
- ***Crawl*:** Depois de encontrar a página *web*, este *software* analisa o seu conteúdo para a categorizar. Neste processo vários computadores com este *software* são usados para categorizar milhões de páginas *web*, usando um algoritmo que determina quais são os *websites* que devem ser indexados, com que frequência e quantas páginas devem buscar em cada um deles.
- **Indexar:** Nesta fase final, o *software,* a partir do conteúdo, tenta descobrir qual é o tema. Este processo inclui processar e analisar o conteúdo textual, *tags* e atributos HTML da página, como o elemento *"<title>"* e atributos *alt*, como imagens, vídeos, entre outros.

O Google Scholar fornece algumas métricas adicionais, tais como (Google, n.d.-b):

- ***H-core*:** Conjunto de h artigos mais citados na publicação e é baseado no *h-index*, isto é, se um investigador tem 5 artigos com 17, 9, 6, 3 e 2 de *h-index* possui um *h-core* de 3;
- ***H-median*:** Mediana das contagens de citações do *h-core*, assim, a *h-median* do exemplo anterior é 9. Este indicador é uma forma de medir a distribuição de citações do artigo no *h-core*.

Apesar do Google Scholar ser uma ferramenta bastante útil para encontrar literatura académica, este possui algumas limitações em comparação com outras plataformas científicas (Christopher Newport University, 2018):

- Os *crawlers* do Google Scholar só indexam revistas públicas ou documentos que as editoras permitam indexar;
- Resultados de pesquisa podem incluir *powerpoints*, notícias ou documentos não publicados, como livros ou artigos;
- Cobre bastantes disciplinas, mas o seu ponto forte são as ciências técnicas e ponto fraco as humanidades;
- Reencaminha para publicadoras que cobram para ter acesso a documentos;
- Só é possível pesquisar por palavras-chave, isto é, não é possível pesquisar por área ou documento;
- Possui poucas opções de filtragem dos resultados;
- O conteúdo apresentado não é curado ou organizado por profissionais;
- Dificuldade em identificar fontes revistas por pares.

Concluindo, o Google Scholar possui um perfil básico de investigador por ser um motor de busca de documentos científicos. No entanto, é uma fonte válida de documentos, pois indexa várias revistas e pode ser usado para integrar a plataforma RDProfile.

2.5 Web of Science

O Web of Science[3], propriedade da Clarivate, é uma base de dados global de citação independente de editoras. Esta plataforma científica oferece aos investigadores informação e tecnologia para a comunidade global de investigação científica. Também fornece dados e análises sobre o perfil do investigador, bem como ferramentas e serviços profissionais personalizados aos investigadores e a toda a comunidade de investigação, como universidades, instituições de investigação, governos nacionais e locais e organizações privadas e públicas de financiamento da investigação em todo o mundo. Suporta ainda mais de 95% das principais instituições de investigação do mundo, múltiplos governos e agências

[3] https://www.webofknowledge.com/

nacionais de investigação com cerca de 20 milhões de investigadores em mais de 7000 organizações de investigação (Clarivate, 2023a).

As plataformas científicas Web of Science e Scopus são bastante similares em termos de apresentação do perfil do investigador. Na Figura 5 apresenta-se uma visão geral do perfil do investigador que contém informações, como o nome, a sua afiliação, o seu ID na plataforma, os nomes pelo qual é apresentado nas suas publicações, organizações às quais o investigador esteve afiliado, palavras-chave dos assuntos que o investigador se foca e outros indentificadores.

Figura 5: Perfil do investigador (Web of Science)

Na Figura 6 é possível verificar os artigos publicados do autor indexados pelo Web of Science com informações gerais, como o título, nome dos autores, ano de publicação e a revista na qual foi publicada. É possível ordenar os artigos por data ou por citação usando os filtros.

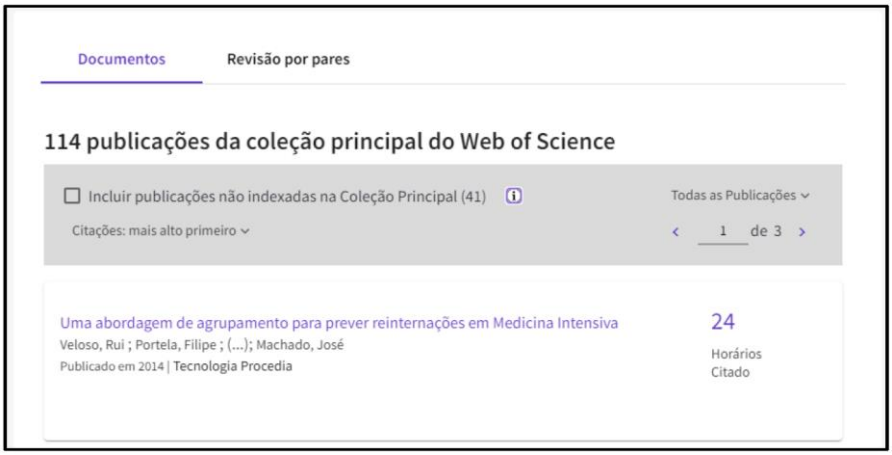


Figura 6: Lista de publicações (Web of Science)

Esta plataforma científica possui uma *dashboard* com várias estatísticas importantes para os investigadores. Na Figura 7 estão representadas estatísticas, como o total de publicações na plataforma, a soma de citações de todas as publicações até ao momento e o *h-index*. Também apresenta um gráfico, similar ao Scopus, que relaciona o total de citações por ano com o número de publicações por ano.

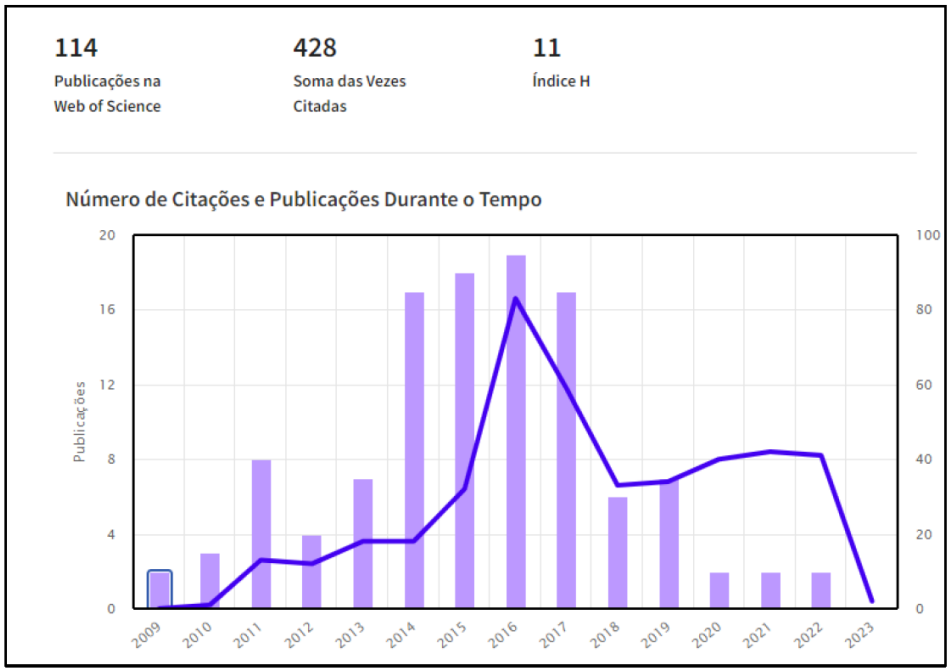


Figura 7: Gráfico de citações (Web of Science)

O Web of Science possui ainda algumas métricas adicionais, tais como (Clarivate, 2023c):

- **Revisão por pares:** Número total de revisões por pares verificadas adicionadas ao perfil;
- **Revisão por pares verificadas (últimos 12 meses):** Número total de revisões por pares verificadas realizadas no último ano adicionadas ao perfil;
- **Rácio de revisão por pares por publicação:** Rácio que compara o número de revisão por pares por cada documento publicado.

Na Figura 8 é demonstrado um exemplo destas estatísticas com informações adicionais, como o percentil e a mediana, acompanhadas por um gráfico de barras de revisão por pares na Figura 9.

Figura 8: Métricas para revisão de pares (Web of Science)

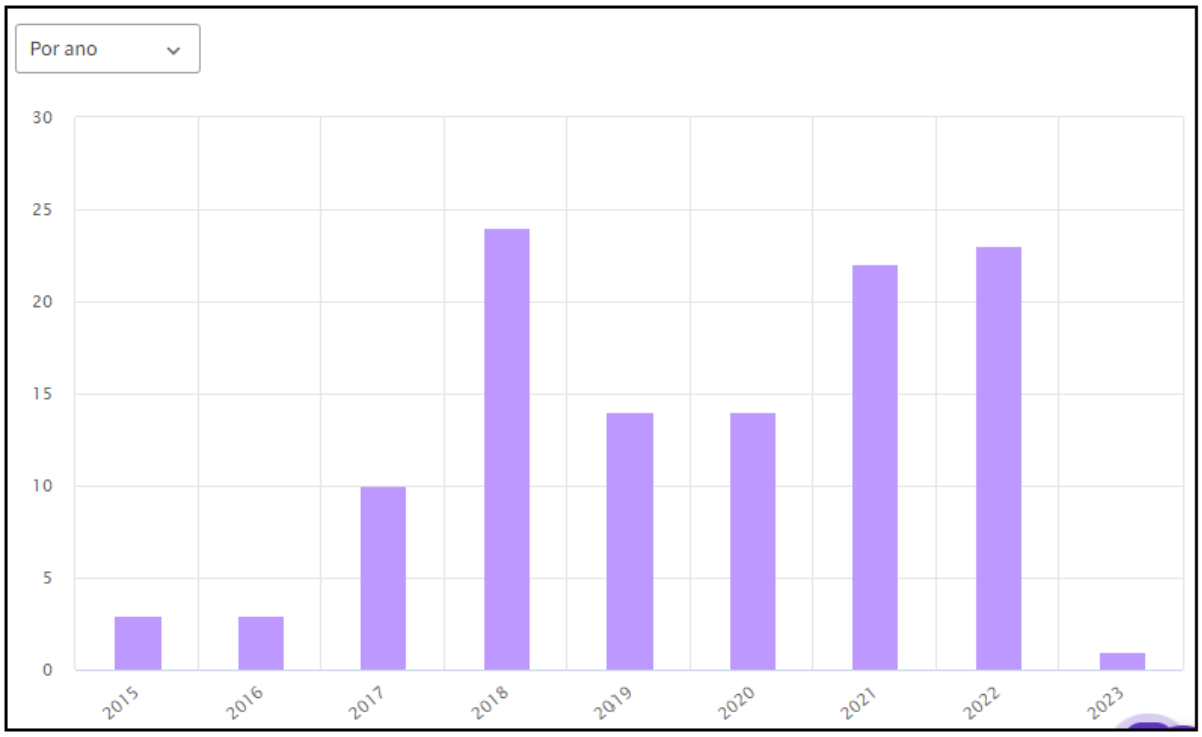


Figura 9: Gráfico de revisão por pares (Web of Science)

Esta plataforma cientifica possui mais quatro métricas, no entanto estas necessitam de uma subscrição no Web of Science para poderem ser visualizadas (Clarivate, 2023c):

- Gráfico do impacto do autor;
- Mapa de citação gráfica;
- Lista de coautores;
- Análise da posição do autor.

Posto isto, é possível concluir que o Web of Science é uma plataforma bastante completa, mas sofre de alguns problemas com a contagem de citações devido ao artigo citado não constar na base de dados ou a referência ao artigo estar erradamente formatada. No entanto, é uma plataforma válida para ser usada como fonte de documentos para a plataforma RDProfile no sentido de agregar dados.

2.6 ORCID

O *Open Researcher and Contributor ID*[4] (ORCID) é uma plataforma global, sem fins lucrativos, sustentada pelos próprios membros com o objetivo de identificar todos os que participam em investigações, bolsas de estudo e inovação de forma única e transversal a todas as disciplinas, fronteiras e tempo. Esta plataforma permite que investigadores registem todas as suas informações pertinentes de uma forma centralizada para que seja mais fácil a partilha de artigos e publicações. O ORCID fornece um ID único e persistente ao utilizador que pode ser partilhado com outros investigadores ou integrado em artigos para que outros investigadores possam ter acesso ao perfil do utilizador mais facilmente (ORCID, n.d.).

Na Figura 10 está representado um perfil de um utilizador da plataforma que contém várias informações. Por exemplo, na aba à esquerda está incluído o ID único do utilizador, *websites* ou redes sociais que o utilizador deseja partilhar, *keywords*, ou seja, palavras-chave que auxiliam os motores de busca a exibir este utilizador a outros investigadores, e o país onde o utilizador conduz as suas pesquisas. No centro, representado por abas cinzentas, são apresentados o nome e o apelido do utilizador e uma breve biografia escrita pelo mesmo.

[4] https://orcid.org/

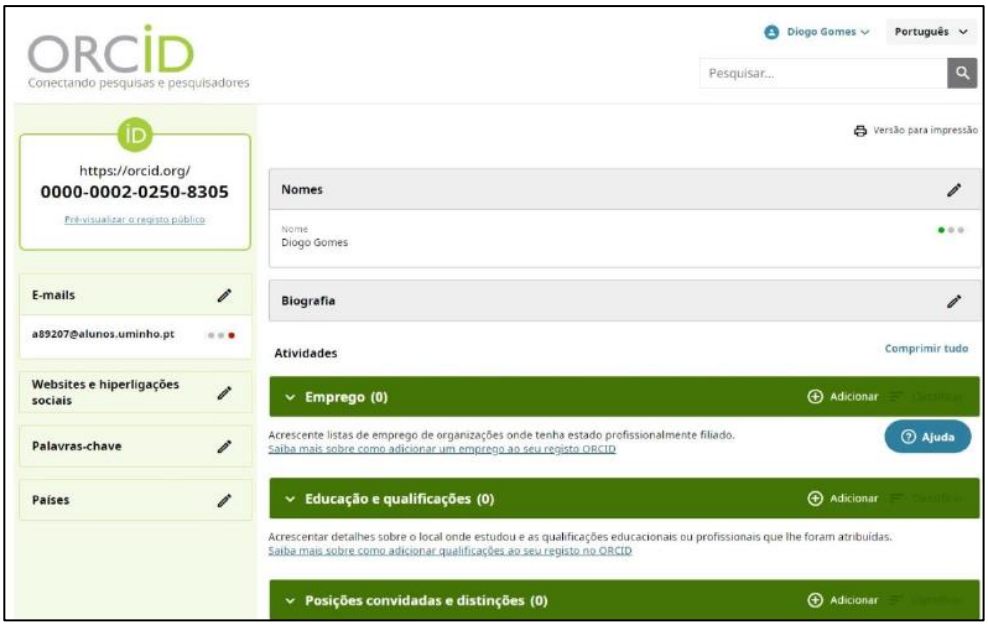


Figura 10: Perfil de Investigador (ORCID)

São ainda apresentadas seis atividades, representadas por abas verdes, que exibem informação detalhada sobre o investigador acerca de diversos temas, como:

- **Emprego:** Lista de organizações em que o utilizador esteve ou esteja ativo;
- **Educação e Qualificações:** Lista de todas as instituições académicas que o utilizador frequentou e o nível de educação;
- **Posições convidadas e distinções:** Lista de todas as posições que o utilizador ocupou e recompensas ou prémios que recebeu por reconhecimento das suas conquistas;
- **Adesão e serviço:** Lista de todas as afiliações ou associações e doações de tempo ou outros recursos em serviço de uma organização;
- **Projetos e trabalhos financiados:** Lista de todos os prémios e mecanismos de financiamento que o utilizador recebeu do seu trabalho;
- **Trabalhos:** Lista de todos os resultados de investigações, como publicações, *data sets*, apresentações de conferências, entre outros.

Na Figura 11 é apresentada uma das funcionalidades do ORCID, “adicionar trabalhos”.

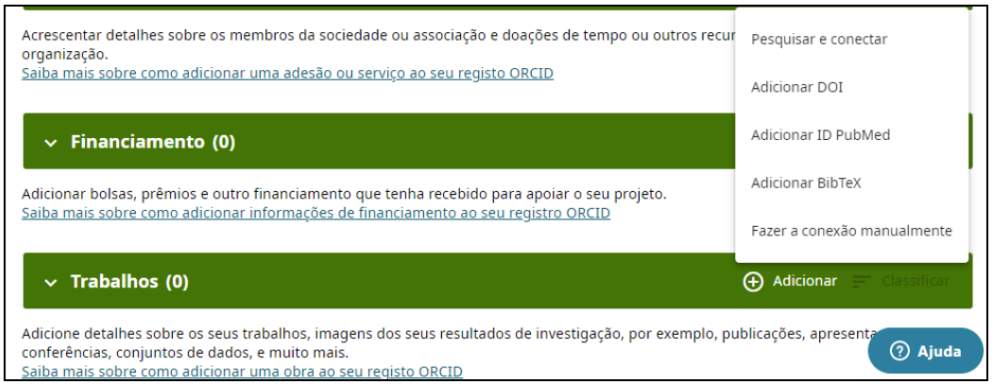


Figura 11: Funcionalidade “adicionar trabalhos” (ORCID)

A plataforma ORCID oferece a possibilidade de adicionar trabalhos das seguintes formas:

- **Pesquisar e conectar:** Permite ao investigador conectar o perfil ORCID a outras plataformas científicas, como o Scopus, MLA International Bibliography, CrossRef Metadata Search, entre outras;
- **Adicionar o DOI:** Permite adicionar o Digital Object Identifier (DOI) dos trabalhos elaborados;
- **Adicionar o ID PubMed:** Permite adicionar o identificador do motor de busca PubMed, base de dados que contém citações de literatura biomédica;
- **Adicionar o BibTex:** Permite adicionar citações de ficheiros BibTex, *software* que gere referências;
- **Fazer a conexão manualmente:** Permite adicionar trabalhos manualmente, preenchendo alguns campos representados na Figura 12.

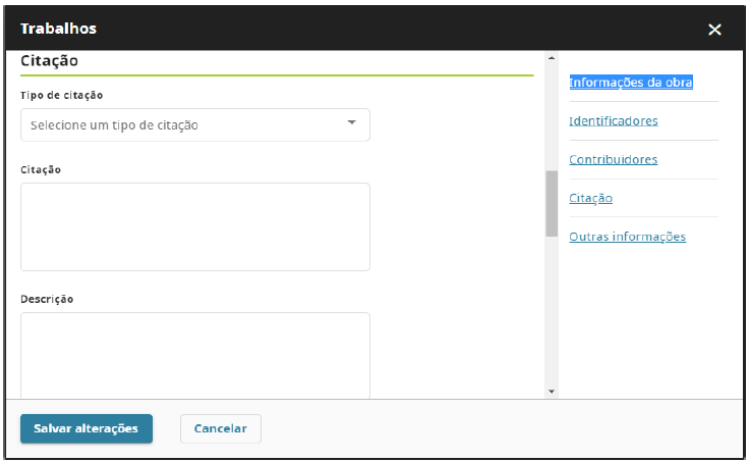


Figura 12: Funcionalidade “conexão manual” (ORCID)

Uma das limitações desta plataforma é ser puramente informacional, isto é, não exibe métricas aos seus utilizadores, como artigos citados ou publicados. No entanto, é uma fonte válida para obter artigos de um determinado investigador por permitir ao utilizador adicionar ao perfil todas as suas publicações, revistas, artigos, entre outros trabalhos.

2.7 Research Gate

O Research Gate[5] é uma rede social, criada em 2008, para cientistas e investigadores e possibilita que estes se conectem, facilitando a partilha e o acesso a conhecimento. É composta por 20 milhões de investigadores de várias áreas de conhecimento em mais de 190 países (Research Gate, 2023).

O Research Gate oferece uma página de perfil mais simples quando comparado com as plataformas científicas anteriormente mencionadas, como o Scopus ou o Web of Science. Na Figura 13 estão ilustradas as informações apresentadas no perfil de um investigador, contendo uma fotografia, o nome, o grau académico, a afiliação, o país e algumas métricas, como o *Research Interest Score,* métrica própria da plataforma, o número de citações e o *h-index*.

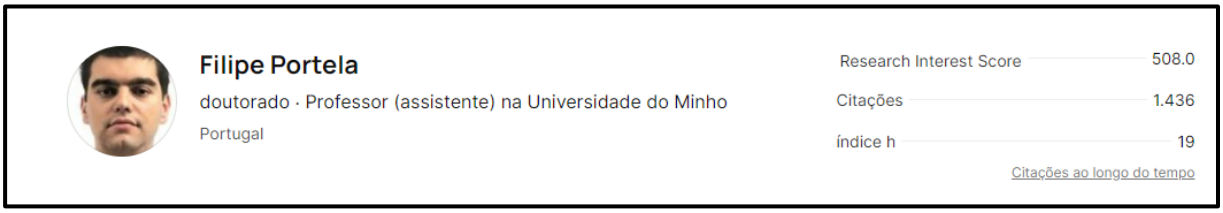


Figura 13: Perfil de Investigador (Research Gate)

Esta plataforma científica também possui um resumo geral sobre o investigador. Na Figura 14 está um exemplo de informações que podem ser apresentadas no perfil do mesmo, tais como disciplinas, habilidades e conhecimentos do investigador e a sua atividade na plataforma.

[5] https://www.researchgate.net/

Figura 14: Informações do investigador (Research Gate)

Na Figura 15 são apresentadas as pesquisas em destaque com algumas informações, tais como o tipo de publicação, data da publicação, a editora em que foi publicada, número de leituras e citações.

Figura 15: Pesquisa em destaque (Research Gate)

Em comparação com outras plataformas científicas, o Research Gate apresenta estatísticas básicas sem filtros avançados para uma melhor análise da informação do investigador devido ao seu foco ser uma rede social em vez de uma base de dados científica. Na Figura 16 são demonstradas as métricas que o Research Gate apresenta: o *Research Interest Score* (métrica própria do Research Gate), o número de leituras (número de vezes que alguém vê o resumo de uma aplicação, como o título, resumo e lista de autores ou descarrega a publicação), o número de citações e recomendações, o *h-index* com autocitações e sem autocitações e ainda a investigação mais citada.

Figura 16: Métricas do Research Gate

Concluindo, o Research Gate possui estatísticas básicas quando comparado com outras plataformas científicas, uma vez que o seu principal propósito é partilhar conhecimento entre investigadores, não se focando tanto na indexação e citação de documentos. No entanto, é uma fonte válida de documentos para a plataforma RDProfile fazendo com que os investigadores possuam um perfil mais alinhado com a realidade do seu impacto na comunidade científica.

2.8 Semantic Scholar

O Semantic Scholar[6] é um motor de busca alimentado por Inteligência Artificial (IA) que indexa mais de 200 milhões de artigos académicos provenientes de parcerias com editoras, fornecedores de dados e pesquisas na *web* desenvolvida pelo *Allen Institute for Artificial Intelligence*. Algumas características inovadoras com o auxílio da IA são:

- Selecionar as melhores palavras-chave sem depender do autor ou editor para as introduzir, utilizando técnicas semelhantes à "leitura automática" (Jones, 2015);
- Capacidade de discernir que referências citadas num artigo tiveram impacto genuíno no mesmo, ao invés de terem sido mencionadas acidentalmente (Jones, 2015);
- Capaz de recomendar ao investigador quais são os melhores artigos para ler tendo em conta a sua pesquisa, mantendo-o sempre atualizado com a investigação mais recente (Semantic Scholar, n.d.-b);

[6] https://www.semanticscholar.org

- Apresentar resumos das citações dos artigos que sumariza todo o conteúdo da citação para auxiliar o investigador a captar apenas os pontos essenciais e, assim, aumentar a sua produtividade (Semantic Scholar, n.d.-a).

Em suma, esta plataforma é bastante similar ao Google Scholar no que toca à indexação de artigos, mas não possui um perfil do investigador, nem estatísticas, como *h-index*, citações, entre outras. No entanto, é uma plataforma válida para ser usada como fonte de documentos para a plataforma RDProfile com vista a agregação de dados.

2.9 Comparação das plataformas

Na Tabela 1 são comparadas as funcionalidades das plataformas anteriormente estudadas, separando-as por:

- **Plataformas indexadoras:** Bases de dados científicas que permitem indexar e catalogar os documentos;
- **Plataformas agregadoras:** Plataformas que agregam os dados das plataformas indexadoras.

Tabela 1: Comparação das plataformas

	Plataformas indexadoras				**Plataformas agregadoras**	
Plataformas / Funcionalidade	**Scopus**	**Google Scholar**	**Semantic Scholar**	**Web of Science**	**ORCID**	**Research Gate**
Filtros avançados	X		X	X		
Métricas (ex*: h-index*, contagem de citações…)	X	X		X		X
Indexação de artigos	X	X	X	X		
Dashboards	X	X		X		X

	Plataformas indexadoras				Plataformas agregadoras	
Plataformas / Funcionalidades	**Scopus**	**Google Scholar**	**Semantic Scholar**	**Web of Science**	**ORCID**	**Research Gate**
Exportação de dados (ex: para .csv, *mendeley*…	X				X	
Cruzamento de dados entre plataformas					X	X

Então, pode concluir-se que o Scopus é a plataforma mais completa com várias funcionalidades, no entanto, não permite o cruzamento de dados com outras plataformas. Apesar do ORCID possuir menos funcionalidades por ser uma plataforma que se foca apenas no perfil do investigador, esta cruza os dados das plataformas indexadoras. A plataforma Web of Science aproxima-se bastante do Scopus, mas contém funcionalidades que necessitam de subscrição especial e não facilita a exportação de dados do investigador. Assim, o RDProfile irá agrupar todos os dados das plataformas analisadas para colmatar as limitações das mesmas e permitir que todos os investigadores portugueses tenham acesso a uma plataforma onde podem consultar informação mais precisa.

2.10 Plataformas similares ao RDProfile

Nesta secção são apresentadas plataformas similares ao RDProfile, nomeadamente o Authenticus e o Ciência Vitae, onde são descritas algumas funcionalidades e limitações que irão ser melhoradas pelo mesmo.

2.10.1 Authenticus

O Authenticus é uma plataforma que associa de forma automática autores de publicações a investigadores e a instituições portuguesas, importando de múltiplas bases de dados de indexação, como o ISI Web of Science, Scopus, CrossRef e Digital Bibliography &

Library Project (DBLP), com o objetivo de construir um repositório nacional digital da produção científica.

A plataforma possui várias funcionalidades, como (CRACS & Inesc TEC, 2023):

- Um algoritmo de identificação multicritério para associar de forma automática autores de publicações a investigadores e instituições portuguesas;
- Interfaces especializadas para investigadores e instituições com acesso à sua informação e através das quais podem, por exemplo, validar ou descartar as associações propostas pelo algoritmo de identificação do Authenticus, bem como decidir sobre publicações redundantes;
- Importação semanal de publicações a partir de múltiplas bases de dados de indexação, nomeadamente da ISI Web Of Science, Scopus, CrossRef e DBLP;
- Atualização dinâmica de informação estatística, em particular de citações, das múltiplas fontes;
- Múltiplos formatos de exportação para que os investigadores possam fazer uso da informação.

O Authenticus possui um algoritmo multicritério de classificação que associa autores, investigadores e instituições. Primeiramente, para cada autor de uma publicação o algoritmo seleciona investigadores candidatos com base em regras de construção de nomes e, para cada candidato, calcula uma medida de semelhança a partir de um conjunto de parâmetros, nomeadamente, nome de autores, emails, endereços institucionais, nomes de revistas e conferências, subáreas científicas, palavras-chave e coautores. O candidato que obtiver o maior valor de semelhança será selecionado como autor da publicação. Posteriormente, estas identificações serão validadas pelos próprios investigadores.

Esta plataforma permite aos utilizadores pesquisar por publicação, investigadores ou instituições, como ilustrado na Figura 17, que demonstra ainda a forma de pesquisa por publicações, permitindo ao utilizador pesquisar pelo título, autor, ano, tema ou identificador das plataformas científicas suportadas pelo Authenticus.

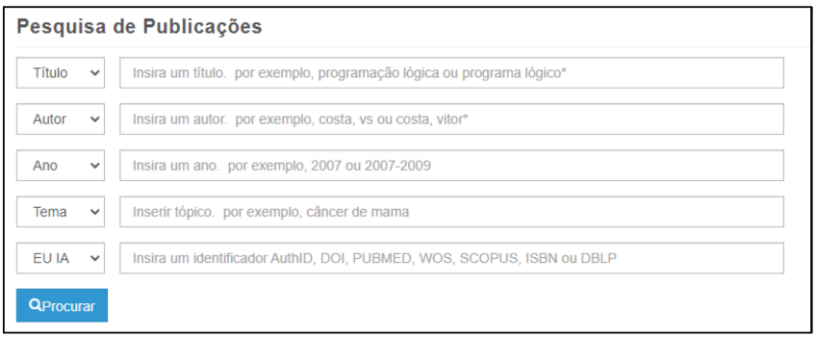


Figura 17: Caixa de pesquisa de publicações (Authenticus)

A Figura 18 apresenta a forma de pesquisa por investigadores, em que é possível procurar pelo nome, ID do ORCID, Researcher AuthID (ID do Authenticus) ou email. Adicionalmente, à esquerda do nome do autor, representada pela cor azul, é exibido o número de publicações validadas pelos próprios investigadores.

Figura 18: Caixa de pesquisa de investigadores (Authenticus)

Na Figura 19 encontra-se a forma de pesquisa por categorias, como apresentado na aba esquerda, listando todas as instituições registadas na plataforma. Também é possível pesquisar por instituições onde é possível procurar pelo nome, acrónimo ou AuthID da instituição, como é visível na Figura 20.

Figura 19: Caixa de pesquisa por categorias

Figura 20: Caixa de pesquisa de instituições (Authenticus)

No centro da Figura 21 é exibida a lista das publicações validadas pelo investigador no Authenticus, contendo o título da publicação, os autores, o ano de publicação, a fonte da qual a publicação foi retirada e a plataforma científica em que está indexada. No topo da Figura 21 são evidenciados os diversos filtros que permitem filtrar as publicações pela plataforma que advêm, isto é, pelas plataformas em que estão indexadas, o tipo de documento, o intervalo de tempo em que foi publicada e a possibilidade de ordenar a lista por ano, citações e título.

Figura 21: Lista de publicações (Authenticus)

Além da lista de publicações, o Authenticus também possui estatísticas gerais sobre as publicações do investigador. Na Figura 22 é apresentado um gráfico de barras, similar ao da plataforma Scopus, com o número de publicações realizadas em cada ano, porém, o Authenticus apresenta a possibilidade de ver o número de publicações totais ou por plataforma científica.

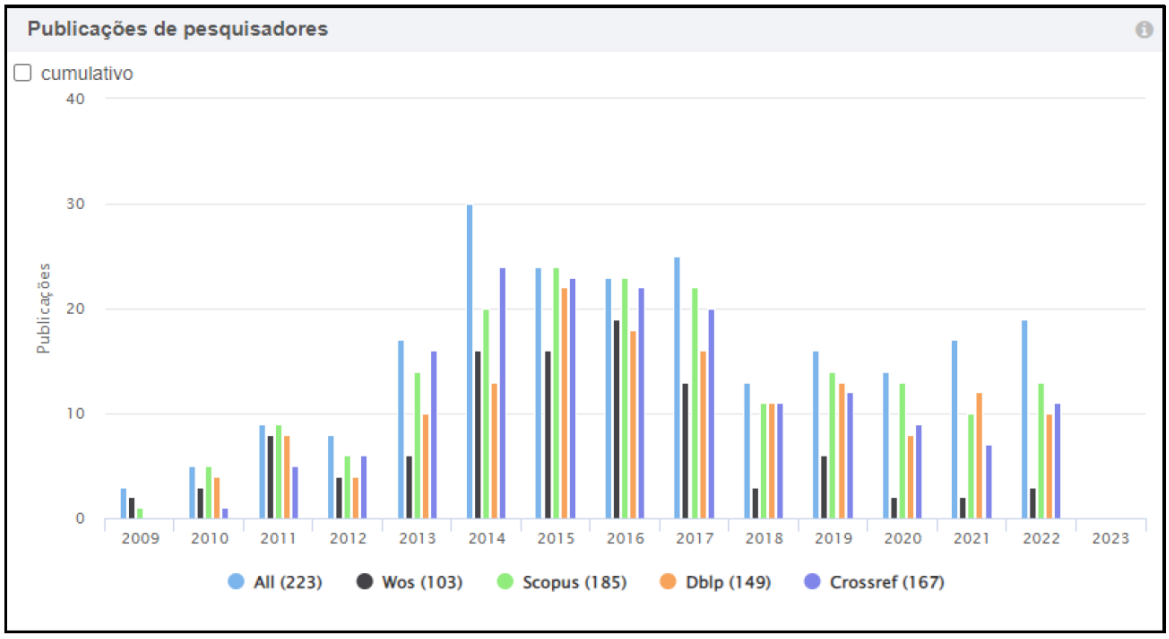


Figura 22: Gráfico de barras do número de publicações (Authenticus)

Na Figura 23 é exibido o número total de citações nas plataformas científicas Web of Science, Scopus e Crossref, enquanto na Figura 24 é apresentado o *h-index* e o *i10-index* (número de publicações com pelo menos 10 citações) das plataformas.

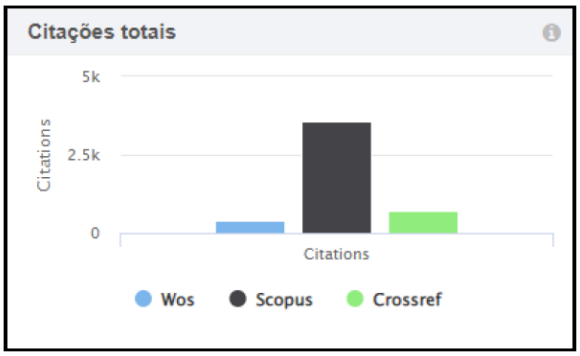


Figura 23: Gráfico de barras de citações totais (Authenticus)

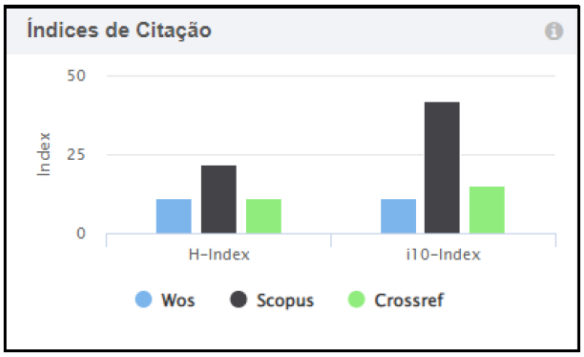


Figura 24: Gráfico de barras do h-index e i10-index (Authenticus)

Além destas estatísticas, o Authenticus possui métricas adicionais, tais como:

- Principais publicações publicadas em periódicos por fator de impacto;
- Topo das publicações por citações;
- Distribuição das publicações por quartis;
- Distribuição de publicações por tipos de documentos;
- Distribuição de coautorias por países.

Assim sendo, o Authenticus é uma plataforma bastante similar ao RDProfile mas que possui algumas fraquezas, tais como:

- Interface com alguns erros;
- Indicador *h-index* nem sempre está de acordo com as plataformas à qual se refere;
- Número de citações nem sempre estão corretos;
- Não possui dados do Google Scholar e ORCID;
- Informação dos quartis é dúbia ou, em algumas situações, errada.

2.10.2 Ciência Vitae

O Ciência Vitae é uma plataforma científica criada pela Fundação para a Ciência e Tecnologia (FCT) que permite aos investigadores em Portugal criarem o seu *Curriculum Vitae* científico. Esta plataforma agrega, num único sítio, informação dispersa de múltiplas plataformas de uma forma simples, harmonizada e estruturada. Integra também vários sistemas nacionais e internacionais, tais como ORCID, FCT|SIG, com o objetivo de reutilizar informação e facilitar a disseminação do currículo do investigador. Permite ainda que investigadores, estudantes ou outros profissionais criem e mantenham um registo de informações, tais como educação e qualificações, histórico de emprego, publicações, colaborações, entre outros (Fundação para a Ciência e Tecnologia, n.d.).

As duas plataformas Ciência Vitae e ORCID apresentam bastantes similaridades no que toca à apresentação da informação. Na Figura 25 é apresentado um exemplo de perfil de um investigador, na plataforma Ciência Vitae, onde se pode concluir que adotam a mesma estrutura de apresentação.

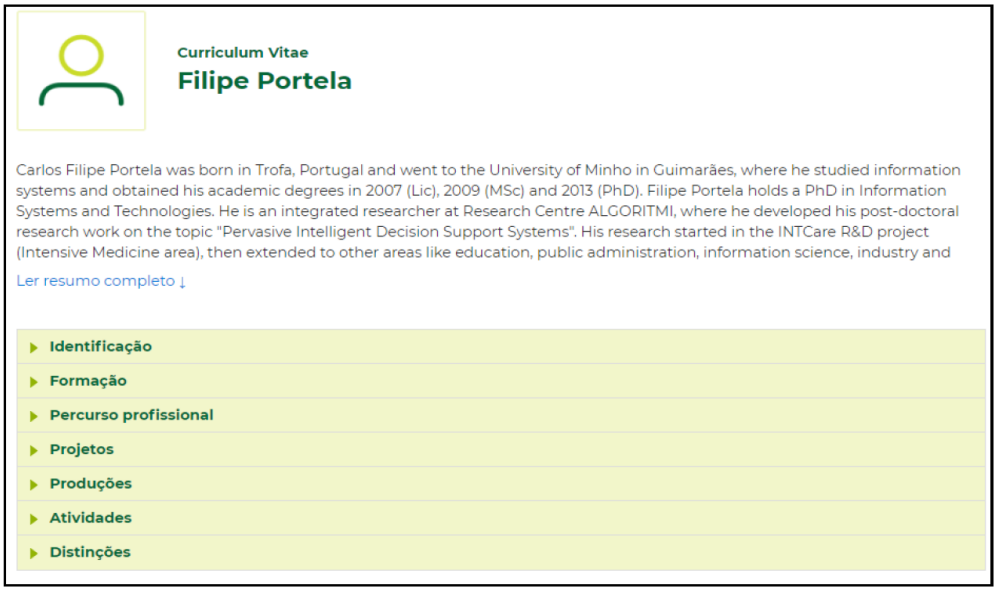


Figura 25: Perfil de Investigador (Ciência Vitae)

São apresentadas 6 categorias no perfil do investigador sendo estas:

- **Identificação:** Informações gerais sobre o investigador, como o nome, endereços de contacto, áreas de atuação, entre outras;
- **Formação:** Informações sobre cursos e especializações do investigador;
- **Percurso profissional:** Informações sobre o histórico de emprego durante os anos do investigador, indicando os cargos e funções desempenhadas;
- **Projetos:** Lista de projetos que o investigador fez parte, mencionando o intervalo de tempo em que atuou e financiadores;
- **Produções:** Lista de todas as publicações, como artigos em revista, artigos em conferência, capítulos de livros para os quais o investigador contribuiu;
- **Atividades:** Lista de todas as atividades que o investigador realizou ou está a realizar, a sua função nas mesmas e onde está a ser realizada;
- **Distinções:** Lista de prémios recebidos pelo investigador.

O Ciência Vitae sofre da mesma limitação que o ORCID devido a ser uma plataforma puramente informacional.

3. ABORDAGEM METODOLÓGICA

No desenvolvimento do projeto de dissertação foram utilizadas as seguintes metodologias:

- Design Science Research;
- Scrum.

A metodologia DSR foi usada como guia para criar um artefacto que, no caso deste projeto de dissertação, é a melhoria do processo de citação e deteção de referências erradas e o Scrum como *framework* para o desenvolvimento do mesmo. Também são exploradas e detalhadas as API das plataformas estudadas na secção 2 e as ferramentas utilizadas durante o desenvolvimento do artefacto.

3.1 Design Science Research

Segundo Peffers et al. (2007), a metodologia DSR envolve um processo rigoroso para desenhar um artefacto capaz de resolver problemas, fazer contribuições para investigação, avaliar o *design* e comunicar os resultados para as audiências apropriadas. Divide-se em seis atividades, ilustradas na Figura 26 (Peffers et al., 2007):

- **Identificação do problema:** Nesta fase do processo é necessário definir o problema de investigação e justificar o valor da solução, pois esta definição irá ser usada para desenvolver um artefacto que, efetivamente, soluciona o problema em questão. A identificação do problema encontra-se detalhada na secção 1.1 deste documento;
- **Definição de objetivos de uma solução:** A partir da identificação do problema são inferidos os objetivos da solução e o conhecimento do que é possivel e viável. Os objetivos podem ser quantitativos, a solução final é melhor que a anterior, ou qualitativos, descrição de como o novo artefacto irá resolver os problemas, até agora, não resolvidos. Os objetivos desta solução encontram-se detalhados na secção 1.2;
- ***Design* e Desenvolvimento:** Criação do artefacto que pode corresponder a construções, modelos, métodos ou instanciação. Esta atividade inclui determinar qual será a funcionalidade do artefacto, a sua arquitetura e a criação do próprio. Neste

projeto de dissertação, o artefacto é o desenvolvimento de novas funcionalidades na plataforma RDProfile, descrito na secção 4;

- **Demonstração:** Demonstração do uso do artefacto para resolver em uma ou várias instâncias do problema. Esta atividade pode envolver o uso do artefacto num experimento, simulação, caso de estudo, prova de conceito, entre outros. Neste trabalho é realizado um experimento do artefacto desenvolvido, demonstrado no capítulo 0;
- **Avaliação:** Observar e verificar se o artefacto desenvolvido é capaz de resolver o problema identificado na primeira atividade. Também envolve comparar os objetivos da solução com os resultados observados do artefacto durante a demonstração e é necessário possuir conhecimentos sobre métricas relevantes e técnicas de análise para avaliar objetivamente os resultados. No fim da avaliação é possível voltar à segunda atividade (Definição de objetivos de uma solução) e redefinir os objetivos da solução ou à terceira atividade (*Design* e Desenvolvimento) para melhorar e refinar alguns requisitos do artefacto. Neste projeto de dissertação o artefacto RDProfile foi avaliado conforme os objetivos definidos na secção 1.2;
- **Comunicação:** Comunicar o problema e a sua importância e, quanto ao artefacto, a sua utilidade e a sua eficácia para os investigadores e outras audiências relevantes. Neste projeto de dissertação a comunicação é feita através da elaboração deste documento e da apresentação do mesmo.

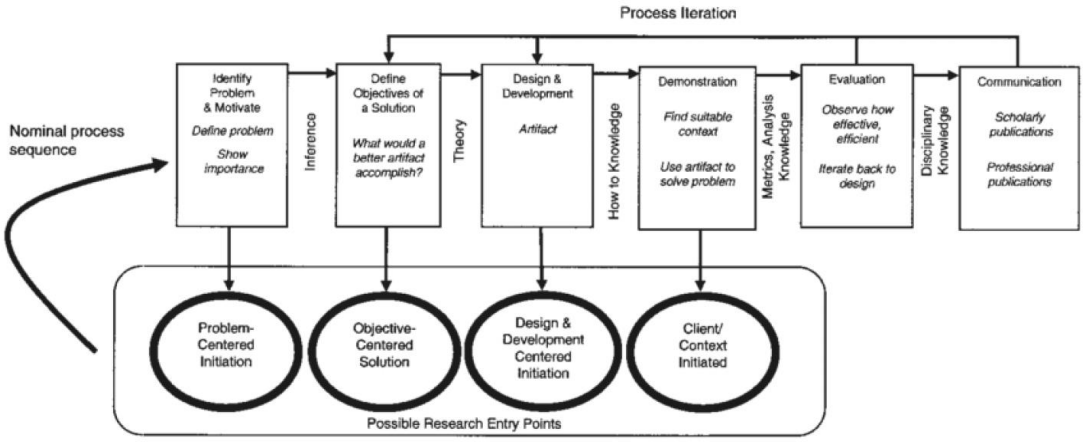


Figura 26: Metodologia Desgin Science Research

3.2 Scrum

Segundo Schwaber & Sutherland (2020), o Scrum é uma *framework* que ajuda desenvolvedores, equipas e organizações a gerar valor através de soluções adaptativas para problemas complexos, baseando-se no empirismo, ou seja, que todo o conhecimento provém da experiência e das decisões feitas com base no que é observado e no pensamento magro, isto é, o foco é sempre o essencial do projeto.

Resumindo, Scrum implementa uma abordagem interativa e incremental para otimizar a previsibilidade e controlar o risco, e incentiva um conjunto de pessoas que, coletivamente, possuem todo o conhecimento e habilidades para realizar o projeto e partilhar tais habilidades conforme necessário.

Para usar esta metodologia é necessário criar uma *Scrum Team,* uma equipa pequena que não ultrapasse os 10 elementos que será responsável por todas atividades relativas ao desenvolvimento do produto desde a colaboração com os *stakeholders*, verificação, manutenção, operação, experimentação, investigação e desenvolvimento, entre outras. Esta equipa é composta pelos seguintes elementos:

- ***Scrum Master*:** Elemento responsável por ajudar o resto da equipa a implementar a metodologia, reforçando as suas regras, boas práticas e encarregado de manter a eficácia da equipa. Neste projeto de dissertação, o *Scrum Master*, é o Professor Carlos Filipe da Silva Portela;
- ***Product Owner*:** Elemento responsável por maximizar o valor do produto que advém do trabalho da *Scrum Team.* É também responsável por gerir o *Product Backlog,* comunicando-o à equipa, ordenar os itens que o compõe por ordem de prioridade e assegurar que este é transparente visível e entendido por toda a equipa. Neste projeto de dissertação, o *Product Owner* é o professor Carlos Filipe da Silva Portela;
- ***Developers*:** Elementos responsáveis pelo desenvolvimento do produto, criando incrementos usáveis em cada *Sprint*. Neste caso, o *Developer* é o aluno Diogo Gomes.

Na Figura 27 está representado um diagrama da metodologia Scrum.

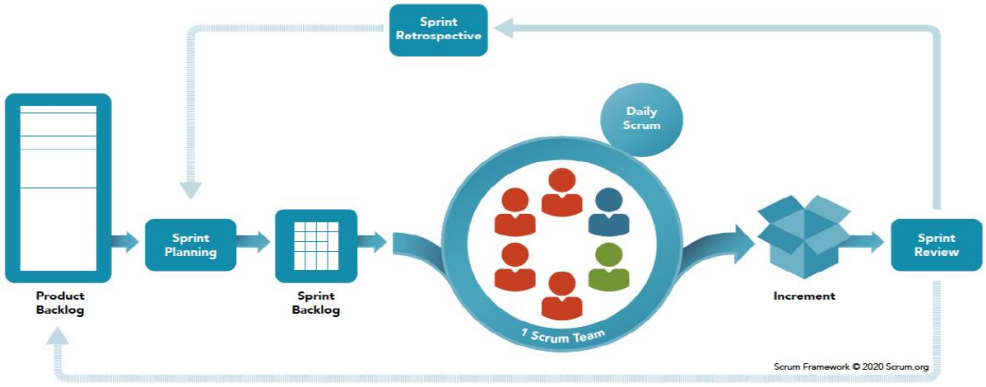


Figura 27: Metodologia Scrum

Este processo é iniciado pelo *product backlog*, uma lista ordenada de tudo o que é necessário melhorar no produto. Esta lista é refinada durante o desenvolvimento conforme o conhecimento adquirido sobre o produto, tornando-se cada vez mais precisa (Schwaber & Sutherland, 2020).

Após a lista estar concluída, o passo seguinte será o *sprint planning* em que a equipa e, se necessário, alguns *stakeholders,* discutem quais dos itens do *product backlog* serão desenvolvidos na próxima *sprint*. Depois de os elementos estarem de acordo, é criada a *sprint backlog*, composta pelo objetivo da *Sprint*, itens selecionados do *product backlog* e um plano acionável para entregar o incremento do produto (Schwaber & Sutherland, 2020).

Seguidamente, começa a *sprint*, altura em que a equipa de desenvolvimento trabalha nos itens decididos anteriormente na *sprint backlog*. As *sprints* costumam durar menos de um mês para manter a consistência e diminuir o risco. Estas *sprints* são acompanhadas pelo *daily scrum*, evento diário onde a equipa numa reunião de 15 minutos adapta o *sprint backlog* conforme que o aconteceu no dia anterior para que o desenvolvimento do projeto seja mais previsível (Schwaber & Sutherland, 2020).

No final da *sprint*, o resultado será um incremento, uma adição concreta e usável ao produto que é utilizável com todos os incrementos anteriores. Posteriormente, dá-se a *sprint review*, evento em que a equipa inspeciona o resultado da *sprint* determinando futuras adaptações e apresenta aos *stakeholders* os resultados e progresso na realização do produto final. No final é realizada a *sprint retrospective*, evento onde se planeiam formas de aumentar a qualidade e a eficácia, ou seja, é um evento onde a equipa avalia como a *sprint* anterior correu, que problemas foram encontrados e como estes problemas foram ou não resolvidos.

Com esta aprendizagem, o *product backlog* também poderá ser alterado, finalizando assim o ciclo do Scrum (Schwaber & Sutherland, 2020).

3.2.1 *Product Backlog*

Na Tabela 2 está representado o *Product Backlog* da fase de desenvolvimento da solução final. Os requisitos do produto estão ordenados por ordem decrescente de prioridade que vai de 1 a 5, sendo o 1 o mínimo e 5 o máximo. Além dessa classificação de prioridade, os requisitos são avaliados com base no esforço necessário, também classificado de 1 a 5. Existe ainda a coluna "Concluído" que indica o *status* de cada requisito, com um valor de 2 para totalmente concluído, 1 para parcialmente concluído e 0 para não concluído.

Tabela 2: *Product Backlog*

ID	Requisito	Prioridade	Esforço	Concluído
1	Elaborar o algoritmo de deteção de referências erradas no Scopus	5	5	2
2	Elaboração do algoritmo de agregação e cruzamento de dados	5	5	2
3	Criar um *crawler* do Google Scholar	5	4	1
4	Implementar a API do Scopus no algoritmo de deteção de referências erradas no Scopus	3	2	2
5	Criar *crawler* do Research Gate	3	4	0
6	Implementar a API do Scopus no algoritmo de agregação e cruzamento de dados	3	2	2
7	Implementar a API do Web of Science	3	2	0
8	Implementar API do ORCID	3	2	2
9	Implementar a API do Semantic Scholar	2	2	2
10	Implementar a API da OpenALex	2	2	2
11	Realizar testes a utilizadores	2	3	2

Analisando a Tabela 2, o terceiro requisito foi parcialmente concluído por não ser possível extrair as citações dos artigos, uma vez que o Google Scholar não permite a extração de informação em massa por *crawlers*, ou seja, só foi possível extrair os títulos de artigos, mas

não as suas citações. Relativamente ao quinto requisito, criar *crawler* da plataforma Research Gate, houve bastantes dificuldades em interagir com o *website* de forma dinâmica, fazendo com que o algoritmo fosse bastante instável na extração de informações. No que toca ao sétimo requisito, não houve resposta da plataforma para fornecer a chave necessária para usar a API que permitia extrair citações dos artigos.

3.2.2 *Sprints*

Na Tabela 3 estão representados todos os *sprints* definidos e realizados durante o desenvolvimento da solução com as respetivas datas de início e fim. Foram realizados 10 *sprints* com duração de 2 semanas cada um.

Tabela 3: *Sprint Backlog*

ID	***Sprint***	**Data de Inicio**	**Data do Fim**	**Descrição da Tarefa**
1	*Sprint* 1	14/02/2023	03/03/2023	Compreender o processo manual da deteção de referências erradas no Scopus
2	*Sprint* 2	06/02/2023	23/03/2023	Criar o *crawler* do Google Scholar
3	*Sprint 3*	24/03/2023	12/04/2023	Implementar API do Scopus
4	*Sprint 4*	13/04/2023	02/05/2023	Desenvolvimento do algoritmo de deteção de referências erradas no Scopus
5	*Sprint 5*	03/05/2023	22/05/2023	Implementar API do Scopus no algoritmo de algoritmo de agregação e cruzamento de dados
				Implementar API do ORCID
				Implementar API do Web of Science
6	*Sprint 6*	23/05/2023	09/06/2023	Criar *crawler* do Research Gate
7	*Sprint 7*	12/06/2023	29/06/2023	Implementar API do Semantic Scholar
8	*Sprint 8*	30/06/2023	19/07/2023	Implementar API do OpenAlex

ID	*Sprint*	Data de Inicio	Data do Fim	Descrição da Tarefa
9	*Sprint 9*	20/07/2023	08/08/2023	Realizar testes em utilizadores
10	*Sprint 10*	09/08/2023	28/08/2023	Elaborar o documento de dissertação

3.3 API

Nesta seção são apresentadas algumas API usadas para obter dados das plataformas científicas a fim de integrá-las na plataforma RDProfile. Também são mencionadas algumas API e a razão pela qual não foram implementadas. Na Tabela 4 é apresentado um resumo dos dados retirados de cada API utilizada.

Tabela 4: Dados retirados das API

Plataforma	Dados retirados da API
Scopus	• Pesquisa de artigos; • Extração de título, DOI e EID; • Extração de citações; • Extração de referências.
ORCID	• Extração de artigos;
Semantic Scholar	• Extração de título e DOI; • Extração de citações.
OpenAlex	• Extração de título e DOI.

3.3.1 Scopus API

Neste projeto de dissertação é explorada a API do Scopus com o objetivo de melhorar o processo de obtenção das referências erradas dos artigos dos investigadores e para obter os artigos e respetivas citações para posteriormente apresentar essa informação no RDProfile. As chamadas à API utilizadas no desenvolvimento da solução são apresentadas na Tabela 5 (Elsevier, 2023a).

Tabela 5: API do Scopus

Nome da API	API	Descrição
Scopus Search API	https://api.elsevier.com /content/search/scopus?query=TITLE()	Esta chamada API foi utilizada para verificar quais os artigos existentes na plataforma Scopus.
	https://api.elsevier.com /content/search/scopus?query=AU-ID()&field=dc:title,eid ,prism:doi&count=200&start=0	Esta chamada API foi utilizada para extrair o título, Scopus *Electronic Identifier* (EID) e DOI dos artigos de um determinado autor utilizando o seu AuthorID.
	https://api.elsevier.com /content/search/scopus?field=dc:title& query=eid&count=200&start=0	Esta chamada API foi utilizada para extrair todos os títulos das citações de um determinado artigo utilizando o seu EID.
Abstract Retrieval API	https://api.elsevier.com /content/abstract/eid/eid?view=REF	Esta chamada API foi utilizada para extrair todas as referências do artigo utilizando o seu EID.

3.3.2 Google Scholar API

Por não existir uma API oficial do Google Scholar, foi criado um *crawler* que fosse capaz de obter os trabalhos dos investigadores a partir do seu ID do Google Scholar. A partir deste *crawler* foi possível extrair todos os títulos dos artigos. Posteriormente, houve uma tentativa de melhoria do *crawler* para que este também conseguisse extrair citações, no entanto, não foi possível uma vez que a plataforma não permite a extração de informação a partir do uso de *scrapers*. Esta tentativa é explicada ao detalhe na secção 4.3.2.

3.3.3 ORCID API

O ORCID oferece uma API pública que foi usada neste projeto de dissertação para ser integrada na plataforma RDProfile. Esta API permite aos utilizadores conectarem os seus IDs de ORCID com outras plataformas científicas com o objetivo de facilitar a manutenção do seu perfil científico. No caso do ORCID, este oferece dois tipos de API (Blackburn, 2022):

- **API Pública:** Esta API é pública para qualquer indivíduo usar e é conectada com ID do ORCID do utilizador que a pretende utilizar. Esta API permite extrair IDs de ORCID autenticados e informações publicamente disponíveis de registos no ORCID;
- **API para Membros:** Esta API só está disponível para membros. Além de ter as mesmas funcionalidades que a API Pública, pode ser extraída informação sobre terceiros de confiança e conectar informação com registos no ORCID, como autores e publicações.

Neste projeto de dissertação foi usada a API pública do ORCID e na Tabela 6 são apresentadas as chamadas a API utilizadas:

Tabela 6: API do ORCID

Nome da API	API	Descrição
Read ORCID works	https://pub.orcid.org /v3.0/ORCIDID/works	Esta chamada API foi utilizada para extrair o título e o DOI de cada artigo de um determinado autor utilizando o seu ORCID ID.

3.3.4 Web of Science API

O Web of Science oferece uma API RESTful com formato de dados *JavaScript Object Notation* (JSON) e *Extensible Markup Language* (XML) e capacidades avançadas de pesquisa e filtragem (Clarivate, 2023b). Contudo as API disponíveis, Web of Science API Lite e Web of Science Starter API mostram-se suficientes para atingir o objetivo de obter artigos juntamente com as citações. Foi feito um pedido de acesso à API à empresa Clarivate, mas até a data desta dissertação, não houve resposta, levando à decisão de descartar o uso desta API.

3.3.5 Semantic Scholar API

O Semantic Scholar API oferece uma API RESTful permitindo o acesso a uma base de dados académica que contém uma coleção de trabalhos, artigos e literatura académica (Redocly, 2023). Neste projeto de dissertação as chamadas à API utilizadas estão representadas na Tabela 7.

Tabela 7: API do Semantic Scholar

Nome da API	**API**	**Descrição**
Artigos de um autor	https://api.semanticscholar.org /graph/v1/author/authorID/ papers?offset=${offset}&fields=externalIds,title	Esta chamada API foi utilizada para extrair o título e o DOI de cada artigo de um determinado autor utilizando o seu AuthorID.
Citações de um artigo	https://api.semanticscholar.org /graph/v1/paper/idPaper /citations?offset=${offset}	Esta chamada API foi utilizada para extrair todos os títulos das citações de um determinado artigo utilizando o seu paperID.

3.3.6 OpenAlex API

O OpenAlex é um projeto *open-source* que indexa vários trabalhos de investigadores a nível global (Priem et al., 2022). Esta API, no momento da escrita desta dissertação, ainda se encontra numa fase inicial de desenvolvimento, no entanto já contém um número significativo de trabalhos indexados. A API utilizada está apresentada na Tabela 8.

Tabela 8: API do OpenAlex

Nome da API	API	Descrição
Obter lista de trabalhos	https://api.openalex.org /works?filter=author.id:authorID	Esta chamada API foi utilizada para extrair o título e o DOI de cada artigo de um determinado autor utilizando o seu AuthorID.

3.3.7 Outras API exploradas

Durante a elaboração desta dissertação foram exploradas outras possíveis API, tais como:

- **Crossref API:** Esta API é bastante lenta e instável, fazendo com que, por vezes, não fosse possível obter todas as publicações dos investigadores ou demorasse bastante tempo para obter publicações de investigadores com uma grande coletânea de trabalhos;
- **SCITE_:** Uma plataforma indexadora que oferece uma API que necessita de um pedido por e-mail para ser possível obter a documentação e acesso à API. Até à data da realização desta dissertação não foi possível obter resposta.
- **Dimensions API:** Uma plataforma indexadora que também necessita de pedido por e-mail para ser possível obter acesso à API. No entanto, esta plataforma não aconselha o uso de API para obtenção informação em massa, algo que é necessário neste projeto de dissertação.

3.4 Ferramentas

Nesta seção são apresentadas todas as ferramentas usadas durante o desenvolvimento da solução. Assim, é apresentado na Tabela 9 o nome e a funcionalidade de cada ferramenta.

Tabela 9: Ferramentas utilizadas

Ferramenta	Funcionalidade
JavaScript	JavaScript é uma linguagem de programação maioritariamente usada no desenvolvimento *web* (MDN web docs, 2023). No entanto, esta ferramenta foi utilizada para criar o *middleware* da solução que tem a função de comunicar com as API das várias plataformas científicas, bases de dados e, também, criar os algoritmos necessários para a solução final.
Node.js	Node.js é um ambiente *open-source* que permite o desenvolvimento de aplicações e a criação de ferramentas do lado do servidor na linguagem JavaScript. Através do node.js é possível executar código fora de um *browser*, facilitando assim o desenvolvimento (OpenJS Foundation, n.d.) . Esta ferramenta foi utilizada para executar código JavaScript, juntamente com as seguintes bibliotecas: • **Axios:** Biblioteca que permite realizar pedidos às API das plataformas implementadas; • **Puppeteer:** Biblioteca que permite controlar *browser* por código com o objetivo de extrair informação de plataformas que não fornecem API.
Visual Studio Code	Visual Studio Code é um editor de texto gratuito desenvolvido pela Microsoft que permite a criação de *software* e oferece ferramentas de *debugging* para auxiliar a correção erros que possam surgir no código (Microsoft, n.d.). Este editor de texto suporta várias linguagens, no entanto foi usado para escrever código JavaScript, gerir bibliotecas do Node.js e como ferramenta de correção de erros durante a execução do código.

Ferramenta	Funcionalidade
Postman	Postman é um *software* que permite testar pedidos HTTP a API com os diferentes métodos GET, POST, PUT, PATCH, DELETE (Postman, n.d.). Esta ferramenta foi utilizada para testar as API das plataformas, pois apresenta uma interface interativa onde rapidamente se pode testar o output de qualquer API, facilitando o desenvolvimento da solução.
MongoDB	MongoDB é uma base de dados não relacional que permite o armazenamento de um grande volume de dados (MongoDB, n.d.). Neste projeto de dissertação, o MongoDB é usado para armazenar todas as publicações e citações dos investigadores.
Mongoose	O Mongoose é uma biblioteca de modelação de dados para o MongoDB (Mongoose, n.d.). Neste projeto de dissertação auxilia na estruturação dos dados para armazenar na base dados MongoDB.

4. DESIGN E DESENVOLVIMENTO

Esta etapa da dissertação tem como objetivo a resolução de dois de problemas que afetam os investigadores do centro Algoritmi, sendo estes a deteção de citações erradas na plataforma Scopus e o número real de citações dos artigos científicos.

Contextualizando o primeiro problema, deteção de citações erradas na plataforma Scopus, durante a publicação dos artigos científicos na mesma, as referências por vezes são erradamente formatadas, resultando em erros na secção de referências do *website*. Esses erros são identificados pela falta de hiperligação nas referências. Assim, os investigadores, periodicamente, necessitam de realizar o processo tedioso de verificar manualmente se estas estão corretamente formatadas. De uma forma resumida, cada investigador possui uma lista de artigos que podem estar ou não publicados no Scopus. Dos artigos publicados é necessário manualmente verificar se as referências estão corretamente formatadas, anotar o Scopus EID do artigo em questão, o número e o título da referência bibliográfica para posteriormente corrigir o erro de referência. Este processo é repetido para todos os artigos, o que o pode tornar massudo e moroso caso o autor possua uma quantidade considerável de artigos científicos.

Relativamente ao segundo problema, número real de citações dos artigos científicos no Scopus, os artigos científicos dos investigadores são frequentemente citados por outros e, dependendo, de onde o investigador que cita publique o artigo, a contagem da citação irá contar apenas dentro dessa plataforma. Por exemplo, caso o investigador citado tenha o artigo publicado na plataforma Scopus e Web of Science e investigador que o cita publique um artigo que apenas esteja corretamente identificado na segunda plataforma, apesar de o artigo estar indexado nas duas, o número de citações irá aumentar apenas na plataforma Web of Science e não no Scopus. Este problema descrito dificulta, aos investigadores, a perceção do verdadeiro impacto do seu artigo científico uma vez que a informação se encontra dispersa pelas várias plataformas.

4.1 Arquitetura

Nesta secção está representada a arquitetura do sistema que foi desenvolvido para a plataforma RDProfile. Conforme ilustrado na Figura 28, tudo é controlado por um *middleware* que tem a principal função de comunicar via API com as plataformas ORCID, Scopus, Semantic Scholar e OpenAlex para obter os trabalhos dos investigadores. No caso do Google Scholar, por não ter uma API é usado um *web scraper* para obter as informações necessárias. Este *middleware* também é responsável por encaminhar os dados para ambos os algoritmos, o algoritmo de agregação e cruzamento de dados e o de deteção de referências erradas na plataforma Scopus, e comunicar com a base de dados MongoDB para armazenar informação. Esses processos são agendados para ocorrer mensalmente, garantindo a atualização contínua das informações. Em suma, a arquitetura é composta por 5 plataformas, 1 base de dados e 2 algoritmos.

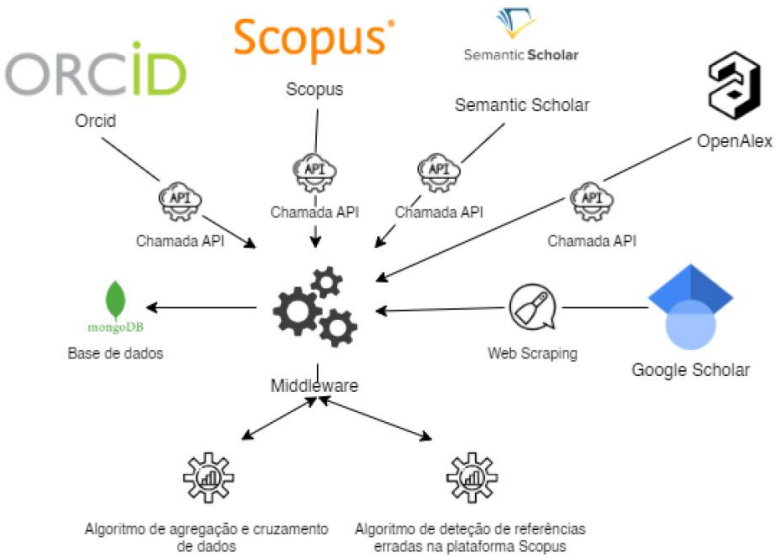


Figura 28: Arquitetura do projeto

4.2 Deteção de referências erradas na plataforma Scopus

Nesta secção é demonstrado o processo da elaboração do algoritmo de automatização da deteção de referências erradas na plataforma Scopus.

Primeiramente foi necessário obter uma fonte viável que possuísse uma lista com todos ou a maior parte dos artigos científicos dos investigadores em Portugal. Por aconselhamento do Professor Orientador Filipe Portela, a plataforma Google Scholar era a melhor opção para obter a lista visto que indexa, frequentemente, artigos científicos de várias plataformas.

Tendo em conta esta informação, no sentido de iniciar a elaboração do algoritmo, foi averiguada a existência de uma API oficial do Google Scholar com o objetivo de obter os títulos dos artigos científicos. Após explorar o *website* e verificar que este não possui uma API, foi necessário explorar outra API alternativa que conseguisse alcançar o mesmo objetivo.

Depois de alguma pesquisa, a única API viável tinha custo monetário ao fim de algum tempo de uso, o que fez com que fosse descartada como opção para este projeto de dissertação. Assim, foi criado um *scraper* que fosse capaz de aceder ao perfil do investigador e extrair todos os títulos dos artigos científicos. Depois de concluída esta parte do algoritmo, foi usada a API do Scopus para obter informações sobre os artigos e as suas referências com objetivo de identificar quais destas estariam erradas.

4.2.1 Processo manual da deteção de referências erradas na plataforma Scopus

Nesta secção é descrito o processo manual que os investigadores, periodicamente, necessitam de fazer para verificar quais as referências que estão erradas. No caso do Professor Filipe Portela, este mantém a sua lista de trabalhos da sua autoria num ficheiro Excel, como se pode observar na Figura 29.

1		Scopus	PaperScopus	IDPaperScopus	CitationNumber	CitationText	CitationRight?	
128	90	Developing a Multi-Agent Platform Support	YES	eid=2-s2.0-85038816822	13	Healthcare interope	NO	TBC
129	292	Early prediction of ICU readmissions using c	YES	eid=2-s2.0-85078992522	14	A clustering approa	NO	TBC
130	297	Early prediction of ICU readmissions using c	YES	eid=2-s2.0-85078992522	14	A clustering approa	NO	TBC
131	232	Effect of conditioning temperature on pelle	YES	eid=2-s2.0-85069671387	29	Proteolytic Inhibitors	NO	TBC
132	86	Development of a research-oriented systen	YES	eid=2-s2.0-85043340745	7	Pervasive universal	NO	TBC

Figura 29: Lista de ficheiro Excel dos artigos

Além da lista de artigos, o ficheiro Excel possui alguns campos adicionais, como se pode observar na Figura 29, sendo estes o seu significado:

- ***PaperScopus*:** Indica se o artigo existe no Scopus ou não, colocando "*Yes*" caso exista ou "*No*" caso não;
- ***IDPaperScopus*:** Contém o EID do artigo original na plataforma Scopus, preenchido quando há pelo menos uma referência mal formatada;
- ***CitationNumber*:** Contém o número da posição da referência mal formatada na plataforma Scopus;
- ***CitationText*:** Contém o texto da referência mal formatada na plataforma Scopus;

- ***CitationRight*:** Indica se o artigo contém referências corretas ou não, colocando "*Yes*" caso estejam ou "*No*" caso não.

A partir da lista, é necessário percorrer cada artigo, copiar o título correspondente e aceder à plataforma Scopus para realizar a pesquisa do mesmo. Durante esta etapa, podem ocorrer três situações: (i) o artigo pode não existir, (ii) pode existir ou (iii) podem aparecer vários artigos com palavras-chave idênticas ao artigo original. Isto requer uma verificação adicional para determinar se o artigo está entre os demais. Se ocorrer a primeira situação, ou seja, o artigo não existir, é somente necessário preencher o campo *PaperScopus* com "*No*". Caso o artigo exista, preenche-se o campo com "*Yes*" e acede-se às referências do mesmo, verificando quais as que estão erradas.

Na Figura 30 está retratado um exemplo de uma referência errada e uma correta, ou seja, a primeira está representada como texto simples, enquanto a segunda possui uma hiperligação para o artigo original. Depois, por cada referência errada, é necessário preencher os campos "*IDPaperScopus*" com o EID do artigo, "*CitationNumber*" que se situa ao lado esquerdo da referência, "*CitationText*" que se situa entre os autores e o nome da revista e "*CitationRight*" com "*No*", por não ter as referências corretas.

Figura 30: Exemplos de Referência errada e correta, respetivamente

Com o objetivo de entender o processo manual e compreender melhor o problema, foi realizado o processo manual primeiro antes da elaboração da solução. Foi fornecida uma lista de trabalhos com cerca de 318 artigos. O tempo de análise de cada artigo varia pois depende de dois fatores: o número de artigos que são apresentados durante a pesquisa e a quantidade de referências que o artigo possui. No entanto, neste caso em específico, o tempo médio de análise de cada artigo foi 45 segundos, ou seja, o processo demorou cerca de 3 horas e 58 minutos para analisar toda a lista de trabalhos. No final existiam 31 artigos com referências erradas.

Assim, o processo manual de deteção de referências erradas na plataforma Scopus é um processo moroso e, por tal razão, foi criada uma solução capaz de resolver este problema. No capítulo 0 foram realizados testes a três utilizadores para provar a sua eficácia.

4.2.2 Algoritmo de deteção de referências erradas na plataforma Scopus

O algoritmo de deteção de referências erradas na plataforma Scopus pode ser dividido em duas partes, sendo a primeira a extração dos títulos da página de perfil do investigador e a segunda a deteção de referências erradas.

Esta parte do algoritmo funciona na base de um *browser* controlado por *software*, por outras palavras, todas as ações executadas no *browser* são previamente escritas por código. Na Figura 31 é apresentado um diagrama que sintetiza o funcionamento do algoritmo como um todo. Primeiramente é necessário obter o ID do Google Scholar do investigador, após obtê-lo, o *browser* controlado por *software* acede à página de artigos do mesmo. Posteriormente, se todos os artigos estiverem visíveis na página, isto é, não estiverem ocultos devido ao número de artigos ser maior do que aquele que a página carrega numa primeira visita ao *website*, são extraídos os títulos dos artigos. Caso contrário, são carregados todos os artigos antes de extrair os títulos dos mesmos.

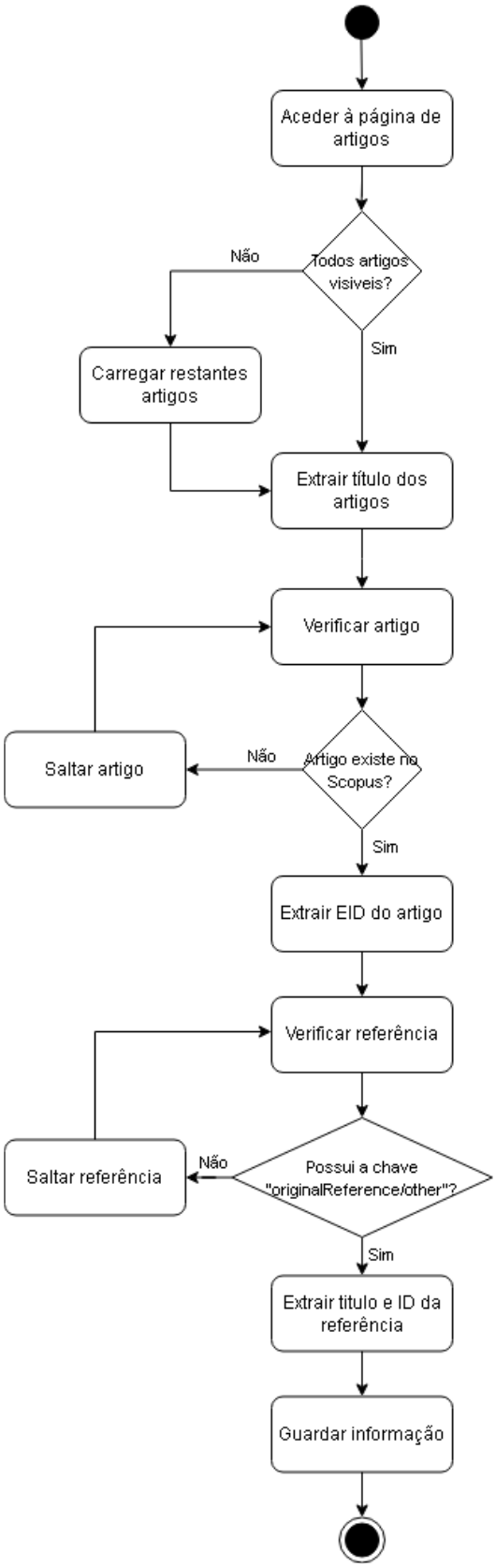


Figura 31: Diagrama do processo de deteção de referências erradas

Na segunda parte do algoritmo, após a obtenção de todos os títulos dos artigos, é verificado um por um se existem na plataforma Scopus. Caso o artigo não exista, o algoritmo salta para o próximo. Caso exista, é extraído o Scopus EID do artigo para poder ter acesso às citações. No caso da plataforma Scopus, as citações erradas têm a chave "*originalReference/other*", representado na Figura 32 que significa que a citação não possui hiperligação, sendo sinalizada como errada. De seguida extrai-se o título e o ID da referência errada e guarda a informação juntamente com o EID do respetivo artigo.

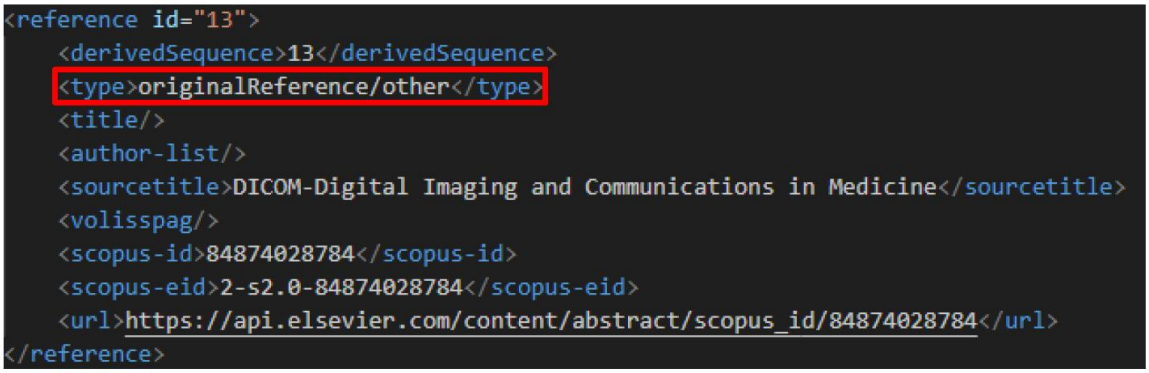


Figura 32: Exemplo de referência errada na resposta API

No final da execução é obtida a seguinte resposta detalhada na Tabela 10.

Tabela 10: Modelo de dados do algoritmo de deteção de referências erradas

Nome do campo	Tipo de dados	Valores possíveis	Descrição
title	*string*	*"The next generation of interoperability agents in healthcare"*	Título do artigo
eid	*string*	"2-s2.0-84900870917"	Identificador único da plataforma Scopus
status	*string*	"Existe no Scopus"	Indica que o artigo existe no Scopus
	string	"Necessita de revisão manual"	Indica que, por algum motivo, a API encontrou mais que um artigo
	string	"Não existe no Scopus"	Indica que o artigo não existe no scopus

Nome do campo	Tipo de dados	Valores possíveis	Descrição
references	*array*	[{"title": "titulo","ID": "10"}]	Possui um *array* com título e a posição da referência na lista de referências

4.3 Número real de citações dos artigos científicos

Depois da criação do algoritmo de deteção de citações erradas na plataforma Scopus, o próximo passo é criar o algoritmo capaz de obter todas as citações de um só artigo de várias plataformas científicas. Nesta secção é detalhada a forma como os subalgoritmos do algoritmo principal funcionam em todas as plataformas, bem como os dados que são extraídos e a forma como estes são organizados e é clarificada a razão de não ter sido possível extrair dados de algumas plataformas. No final é explicado o subalgoritmo que agrega e cruza os artigos e as suas respetivas citações.

4.3.1 ORCID

Para obter todos os artigos da plataforma ORCID, foi utilizada a API correspondente. O funcionamento do algoritmo é ilustrado na Figura 33. Inicialmente, é necessário fornecer o ORCID ID do autor. Em seguida, são extraídas as informações dos artigos, incluindo o título, o EID e o DOI, a partir desse ID. Esses dados são armazenados para posterior processamento. No final, tudo é consolidado numa lista para ser processada pelo algoritmo que agrupa todas as citações únicas. No caso da plataforma ORCID, as publicações não possuem citações, pois trata-se de uma plataforma agregadora, no entanto, esta plataforma adiciona mais informação à coletânea de artigos.

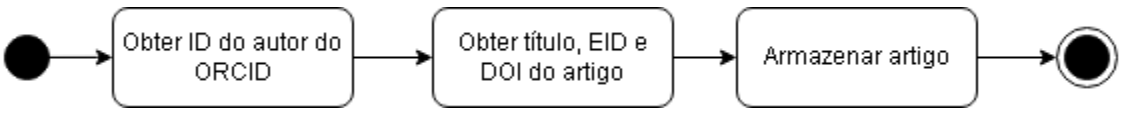


Figura 33: Diagrama do processo (ORCID)

Na Tabela 11 são apresentados os campos do ficheiro JSON que serão posteriormente tratados pelo algoritmo que agrupa as citações e na Figura 34 um exemplo de resposta JSON. Nota importante, é necessário que modelo de dados possua um *array* de citações para o algoritmo de agregação e cruzamento de dados funcionar corretamente.

Tabela 11: Modelo de dados (ORCID)

Nome do campo	Tipo de dados	Valores possíveis	Descrição
title	*string*	*"The next generation of interoperability agents in healthcare"*	Título do artigo
eid	*string*	"2-s2.0-84900870917"	Identificador único do Scopus
id	*string*	"10.4230/OASIcs.ICPEC.2020.23"	Identificador único de objetos
fromPlatform	*string*	"ORCID"	Identificador da plataforma proveniente
citations	*array*	[{"title": "IoEduc-Bring Your Own Device to the Classroom","doi": "10.4230/OASIcs.ICPEC.2020.23"}]	Conjunto de citações dos artigos

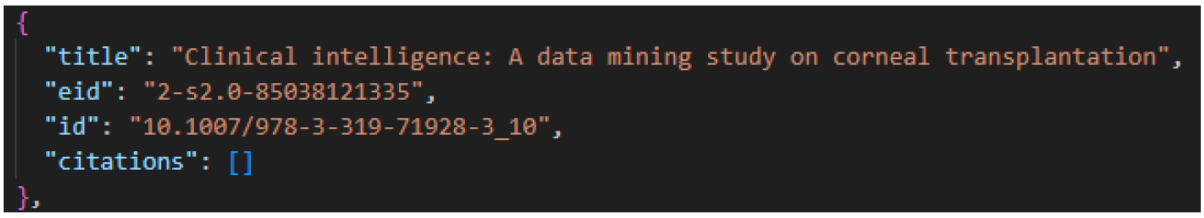


Figura 34: Exemplo de resposta JSON (ORCID)

De notar que o ORCID é uma plataforma agregadora, ou seja, esta plataforma apenas fornece publicações de investigadores sem as suas citações. No entanto, para o investigador obter um perfil mais preciso da quantidade de trabalhos, serão adicionados todos os trabalhos que esta plataforma fornece para que futuramente sejam adicionadas as citações dos trabalhos provenientes de outras plataformas.

4.3.2 Google Scholar

O Google Scholar não disponibiliza uma API oficial que possa ser utilizada e as alternativas existentes não apresentam as funcionalidades necessárias para a obtenção da informação requerida, além de envolverem custos monetários. Com base no algoritmo previamente desenvolvido para deteção de referências erradas na plataforma Scopus, foi adicionada a funcionalidade de extrair todos os títulos das citações de cada artigo científico. Uma vez que a prática de *scraping* não é permitida no Google Scholar, ao fim de alguns testes ao algoritmo este era interrompido por um CAPTCHA, um desafio cognitivo para diferenciar máquinas de humanos.

No entanto foram testadas várias estratégias, sendo uma das primeiras a troca de IP através de uma *Virtual Private Network* (VPN) que permite a alteração virtual da localização do computador com o propósito de evitar a perceção de que os pedidos ao *website* originavam da mesma fonte. Apesar da realização de vários testes, o algoritmo continuava a ser bloqueado pelo CAPTCHA. Depois de alguma pesquisa, a próxima solução explorada foi o uso de *proxies*, servidores que atuam como intermediário entre o computador e o servidor do Google Scholar, ou seja, um sistema idêntico ao anterior no que toca a evitar a identificação dos pedidos como provenientes de uma única máquina. Entretanto, também esta solução não se revelou eficaz, visto que, após uma série de testes, o CAPTCHA continuava a interromper o funcionamento do algoritmo. Juntamente com estas soluções descritas também foi testada a diminuição e aleatorização do tempo dos pedidos com o objetivo do algoritmo assemelhar-se a um humano e evitar que o CAPTCHA o interrompesse.

Em suma, mesmo com todas as modificações no algoritmo, este não era capaz de obter todos os títulos das citações sem ser detetado, descartando esta plataforma como fonte de citações. No entanto, a partir do algoritmo de deteção de referências erradas na plataforma Scopus foi usada a primeira parte do mesmo, como se pode observar na Figura 35 , ou seja, a parte do algoritmo responsável por extrair os títulos dos artigos.

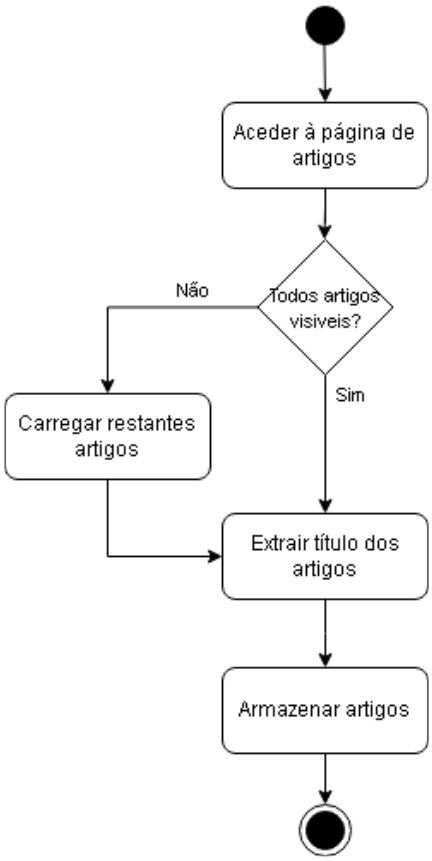


Figura 35: Primeira parte do algoritmo de deteção de referências erradas

Na Tabela 12 estão apresentados os campos do ficheiro JSON que serão posteriormente tratados pelo algoritmo de agregação e cruzamento de dados e na Figura 36 um exemplo de uma resposta JSON. De notar que apesar de o subalgoritmo não extrair citações do Google Scholar, é necessário que o modelo de dados possua um *array* de citações para que o algoritmo de agregação e cruzamento de dados funcione corretamente.

Tabela 12: Modelo de dados (Google Scholar)

Nome do campo	**Tipo de dados**	**Valores possíveis**	**Descrição**
title	*string*	*"The next generation of interoperability agents in healthcare"*	Título do artigo
fromPlatform	*string*	"Google_Scholar"	Identificador da plataforma proveniente

Nome do campo	Tipo de dados	Valores possíveis	Descrição
citations	*array*	[{“title": "IoEduc-Bring Your Own Device to the Classroom", "doi": "10.4230/OASIcs.ICPEC.2020.23"}]	Conjunto de todas as citações do artigo

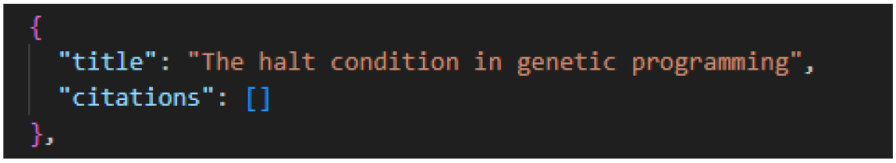


Figura 36: Exemplo de resposta JSON (Google Scholar)

4.3.3 Research Gate

Similarmente ao Google Scholar, a plataforma Research Gate também não possui uma API oficial. Então, foi necessário criar um *scraper* que fosse capaz de obter os artigos e as suas respetivas citações. Apesar de não haver problemas com o CAPTCHA, a página *web* apresentada no *browser* controlado por *software* era uma versão diferente e instável da versão que o *browser* normal apresenta. A instabilidade dessa versão ocasionalmente causava o desaparecimento de elementos HTML quando interagidos. Durante o processo de desenvolvimento do algoritmo, essa instabilidade impactou negativamente a sua capacidade de extrair informações sem intervenção humana durante a execução.

Foi testada outra alternativa, que consistia em fazer pedidos GET à plataforma, ou seja, ao invés de interagir com os elementos HTML da página, o conteúdo era solicitado a partir do URL da página com o propósito de evitar interações com esses elementos. No entanto, esta abordagem não é permitida pelo *website* por este método alertar a possibilidade de ser um robô a interagir com o conteúdo.

Resumindo, a obtenção de informações dessa plataforma também se mostrou inviável, já que a prática de *scraping* não é permitida.

4.3.4 Scopus

Com o objetivo de obter as logo artigos citados, mas não indexados não serão contados científicos, foi utilizada a API do Scopus. Na Figura 37 está representado o diagrama do algoritmo. Primeiramente, é necessário obter o ID do autor, de seguida, são extraídas todas as informações dos artigos, ou seja, o título, o EID do artigo e o DOI, para posteriormente serem armazenadas. A partir do ID de cada artigo são extraídas as suas respetivas citações caso estas existam. No final é tudo armazenado numa lista para posteriormente ser processado pelo algoritmo de agregação e cruzamento de dados.

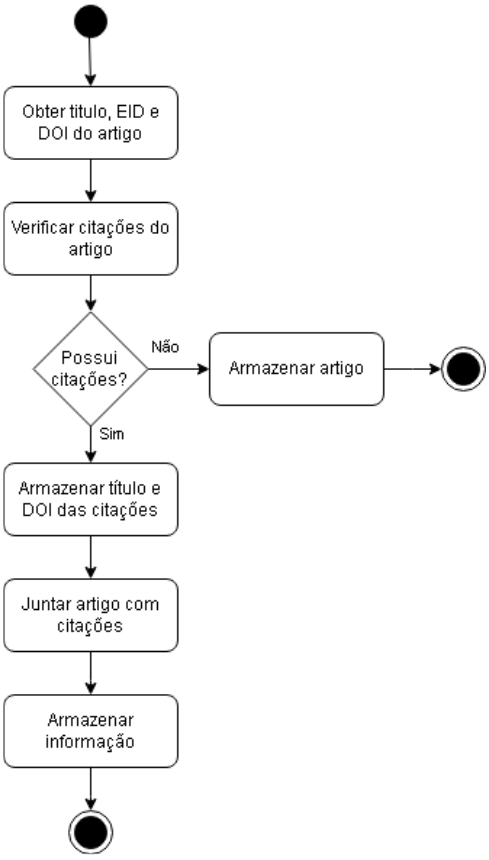


Figura 37: Diagrama do processo (Scopus)

Na Tabela 13 estão apresentados os campos do ficheiro JSON que serão posteriormente tratados pelo algoritmo que agrupa as citações e naFigura 38 um exemplo de uma resposta JSON.

Tabela 13: Modelo de dados (Scopus)

Nome do campo	Tipo de dados	Valores possíveis	Descrição
title	*string*	*"The next generation of interoperability agents in healthcare"*	Título do artigo
eid	*string*	"2-s2.0-84900870917"	Identificador único da plataforma Scopus
doi	*string*	"10.4230/OASIcs.ICPEC.2020.23"	Identificador único de objetos
fromPlatform	*string*	"Scopus"	Identificador da plataforma proveniente
citations	*array*	[{"title": "IoEduc-Bring Your Own Device to the Classroom", "doi": "10.4230/OASIcs.ICPEC.2020.23" FromPlatform: "Scopus"}]	Conjunto de todas as citações do artigo juntamente com o DOI e plataforma proveniente

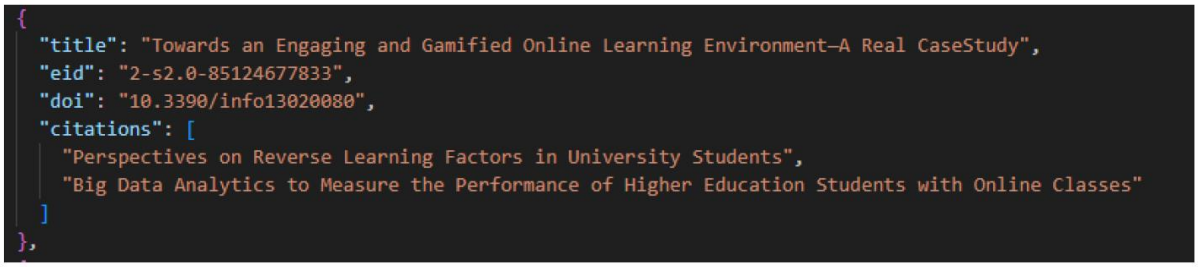


Figura 38: Exemplo de resposta JSON (Scopus)

4.3.5 Semantic Scholar

Com o objetivo de obter todos os artigos da plataforma Semantic Scholar foi usada a API da mesma. Na Figura 39 está representado o funcionamento do algoritmo. Primeiramente é necessário fornecer o ID do autor desta plataforma, de seguida, a partir desse mesmo ID são extraídas as informações dos artigos, ou seja, o título, o ID do artigo e o DOI, que são posteriormente armazenadas. A partir do ID de cada artigo são extraídas as suas respetivas citações caso estas existam. No final é tudo armazenado numa lista para posteriormente ser processado pelo algoritmo de agregação e cruzamento de dados.

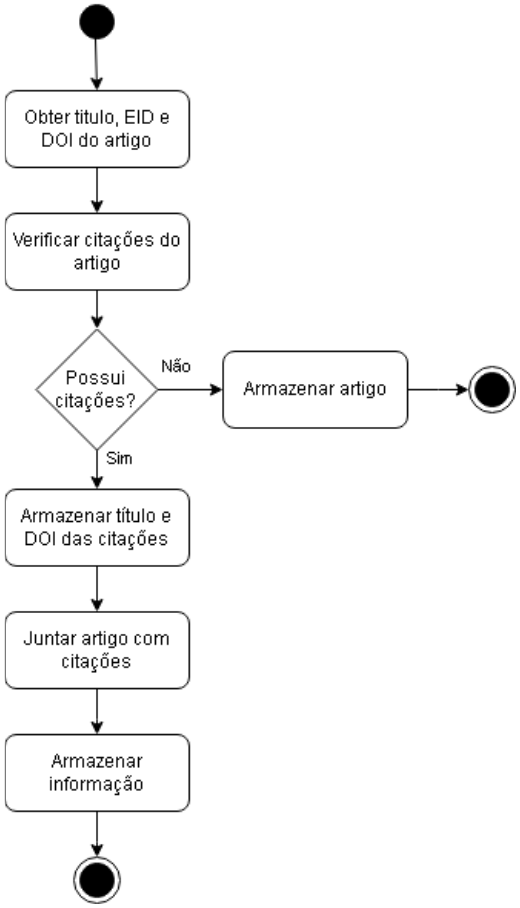


Figura 39: Diagrama do processo (Semantic Scholar)

Na Tabela 14 estão apresentados os campos do ficheiro JSON que serão posteriormente tratados pelo algoritmo de agregação e cruzamento de dados e na Figura 40 um exemplo de uma resposta JSON.

Tabela 14: Modelo de dados (Semantic Scholar)

Nome do campo	Tipo de dados	Valores possíveis	Descrição
title	*string*	*"The next generation of interoperability agents in healthcare"*	Título do artigo
id	*string*	"2a65451b17977f3a2a32aa74d730bedea4b15693"	Identificador único do Semantic Scholar
fromPlatform	*string*	"Semantic_Scholar"	Identificador da plataforma proveniente
doi	*string*	"10.4230/OASIcs.ICPEC.2020.23"	Identificador único de objetos
citations	*array*	[{*"title": "IoEduc-Bring Your Own Device to the Classroom",* "doi": "10.4230/OASIcs.ICPEC.2020.23"}, *"fromPlatform": "Semantic_Scholar"]*	Conjunto de todas as citações do artigo juntamente com o DOI e plataforma proveniente

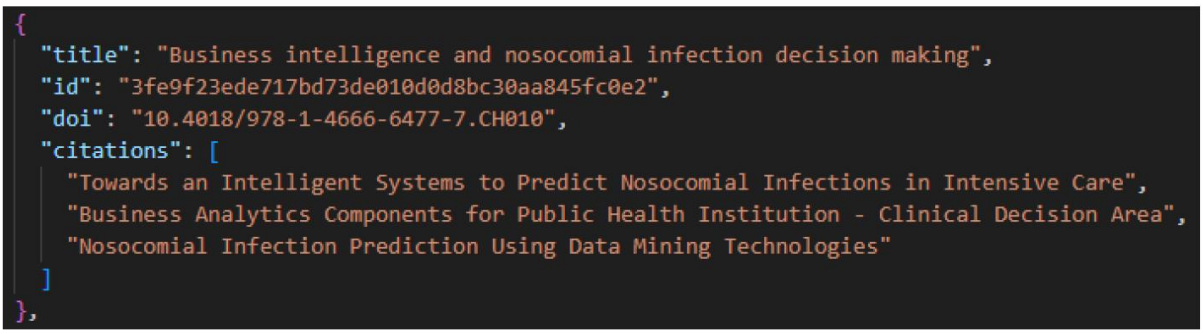


Figura 40: Exemplo de resposta JSON (Semantic Scholar)

4.3.6 OpenAlex

Para obter todos os artigos da plataforma OpenALex, foi utilizada a API da mesma. O funcionamento do algoritmo é ilustrado na Figura 41. Inicialmente, é necessário fornecer o ID do autor na plataforma OpenAlex. Em seguida, são extraídas as informações dos artigos, incluindo o título e o DOI, a partir desse ID.

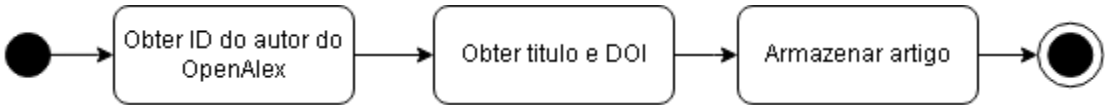


Figura 41: Diagrama do processo (OpenAlex)

Na Tabela 15 estão apresentados os campos do ficheiro JSON que serão posteriormente tratados pelo algoritmo de agregação e cruzamento de dados e na Figura 42 um exemplo de resposta JSON. Nota importante, é necessário que o modelo de dados possua um *array* de citações para que o algoritmo de agregação e cruzamento de dados funcione corretamente.

Tabela 15: Modelo de dados (OpenAlex)

Nome do campo	Tipo de dados	Valores possíveis	Descrição
title	*string*	*"The next generation of interoperability agents in healthcare"*	Título do artigo
doi	*string*	"10.4230/OASIcs.ICPEC.2020.23"	Identificador único de objetos
fromPlataform	*string*	"OpenAlex"	Identificador da plataforma proveniente
citations	*array*	[{"title": "IoEduc-Bring Your Own Device to the Classroom", "doi": "10.4230/OASIcs.ICPEC.2020.23"}]	Conjunto de citações dos artigos


```
{
  "title": "Electronic Health Record in Dermatology Service",
  "doi": "10.1007/978-3-642-24352-3_17",
  "citations": []
},
```


Figura 42: Exemplo de resposta JSON (OpenAlex)

4.3.7 Algoritmo de agregação e cruzamento de dados

Após reunir todos os artigos e as suas respetivas citações das plataformas mencionadas, foi desenvolvido um algoritmo capaz de integrar essas informações de forma a produzir uma lista final contendo apenas artigos e citações únicas, ou seja, nenhum artigo ou citação aparece mais de duas vezes na lista.

Na Figura 43 encontra-se o diagrama que ilustra o funcionamento desse algoritmo. Primeiramente são obtidas as informações sobre os artigos provenientes das plataformas científicas, ou seja, o título do artigo, o DOI, EID e as suas citações, como seu respetivo DOI. Depois, por cada artigo é verificado se possui um DOI, caso tenha, é verificado se este já existe na lista de artigos únicos. Caso já conste na lista, não é adicionado, mas é verificado se o artigo que iria ser adicionado possui citações que o já existente não possui. Se houver citações em falta, estas são adicionadas ao artigo já existente e é descartada o resto da informação. Este processo é igual para os artigos e citações que não possuem DOI, mas no caso destes é usado o título como identificador na procura. No fim é criada uma lista com artigos e citações únicas.

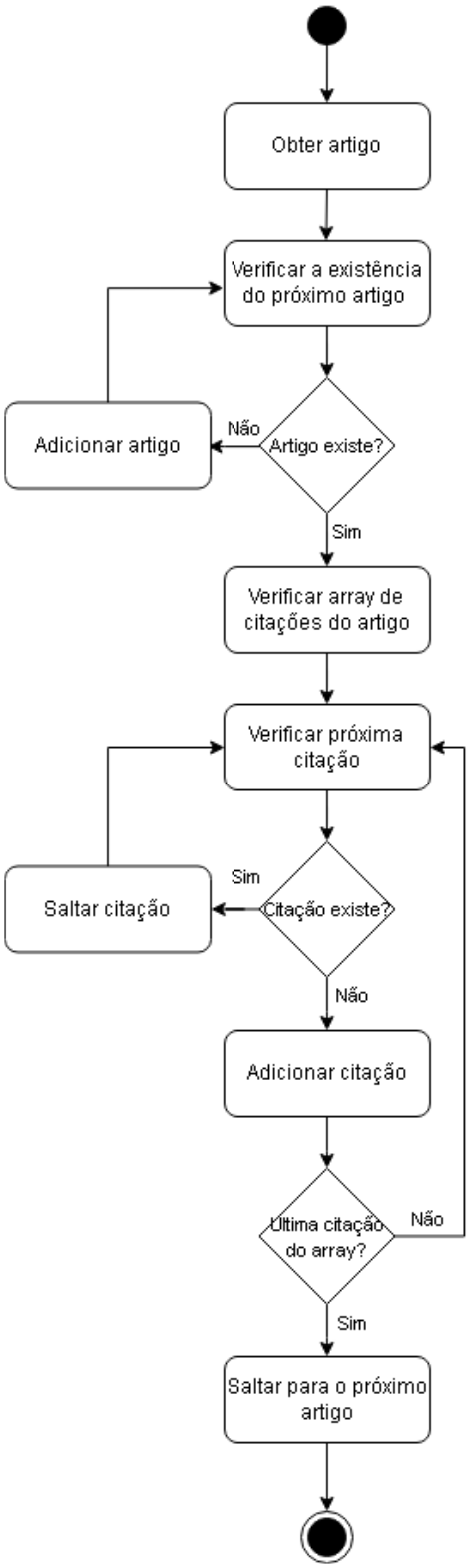


Figura 43: Diagrama do processo de agregação e cruzamento de dados

Na Tabela 16 é apresentado o modelo de dados que possui o título do artigo, o EID, o DOI e as citações únicas das várias plataformas.

Tabela 16: Modelo de dados do algoritmo de agregação e cruzamento de dados

Nome do campo	**Tipo de dados**	**Valores possíveis**	**Descrição**
title	*string*	“*The next generation of interoperability agents in healthcare*”	Título do artigo
eid	*string*	"2-s2.0-84900870917"	Identificador único do Scopus
doi	*string*	"10.4230/OASIcs.ICPEC.2020.23"	Identificador único de objetos
fromPlatform	*string*	“Scopus”	Identificador da plataforma proveniente
citations	*array*	[{“title": "IoEduc-Bring Your Own Device to the Classroom", "doi": "10.4230/OASIcs.ICPEC.2020.23"}, “fromPlataform”: “Scopus”]	Conjunto de todas as citações do artigo juntamente com o DOI e plataforma proveniente

5. TESTES DE UTILIZADORES

Nesta secção foram realizados testes a ambos os algoritmos com o objetivo de testar a robustez e a eficácia dos mesmos. Os testes foram realizados em três investigadores do centro Algoritmi, o professor Filipe Portela, o professor Jorge Sá e o professor José Machado. Foi testado, primeiramente, o algoritmo de deteção de referências erradas na plataforma e de seguida o número real de citações dos artigos científicos.

5.1 Algoritmo de deteção de referências erradas na plataforma Scopus

Para relembrar a forma como as referências erradas são detetadas, na secção 4.2.1, na Figura 30, as referências erradas não possuem hiperligação e na secção 4.2.2, na Figura 32, a chave detetada pelo algoritmo para sinalizar como referência errada.

O primeiro utilizador a ser testado é o professor Filipe Portela que possui 237 artigos no Google Scholar. Na Figura 44 está apresentado um diagrama dos resultados do algoritmo e, como se pode observar, dos 237 artigos, 54 artigos não existem no Scopus e dos 183 artigos que existem no Scopus, estes possuem 1641 referências erradas.

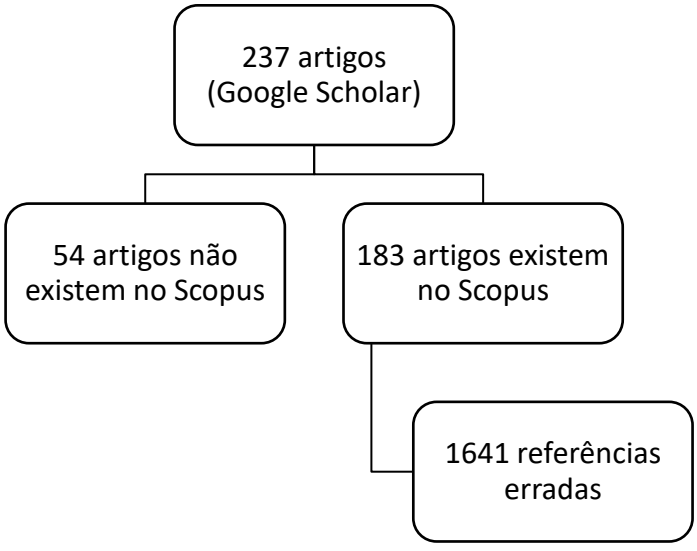


Figura 44: Diagrama de resultados (Filipe Portela)

Então, tendo em conta o tempo médio de 45 segundos de análise por cada artigo, este processo realizado manualmente demoraria 2 horas e 57 minutos aproximadamente, enquanto o *script* demorou aproximadamente 3 minutos e 42 segundos, ou seja, houve uma redução de 97,91% de tempo, provando a eficácia do algoritmo a detetar as referências erradas no Scopus.

O segundo utilizador a ser testado é o professor Jorge Sá que possui 60 artigos no Google Scholar. Na Figura 45 está apresentado um diagrama dos resultados do algoritmo e como se pode observar, dos 60 artigos, 15 artigos não existem no Scopus e dos 45 artigos que existem no Scopus, estes possuem 468 referências erradas.

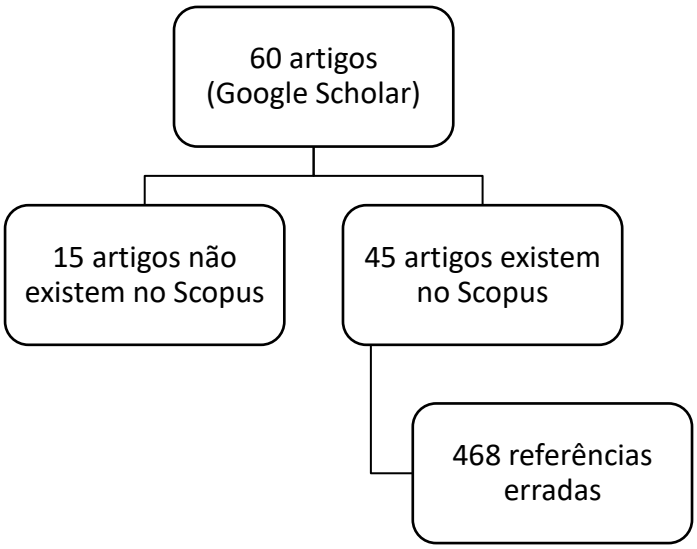


Figura 45: Diagrama de resultados (Jorge Sá)

Então, tendo em conta o tempo médio de análise por cada artigo, este processo realizado manualmente demoraria 45 minutos aproximadamente, enquanto o *script* demorou aproximadamente 1 minuto e 1 segundo, ou seja, houve uma redução de 97,74% de tempo, provando a eficácia do algoritmo a detetar as referências erradas no Scopus.

O terceiro utilizador a ser testado é o professor José Machado que possui 428 artigos no Google Scholar. Na Figura 46 está apresentado um diagrama dos resultados do algoritmo e como se pode observar, dos 428 artigos, 107 artigos não existem no Scopus e dos 321 artigos que existem no Scopus, estes possuem 2651 referências erradas.

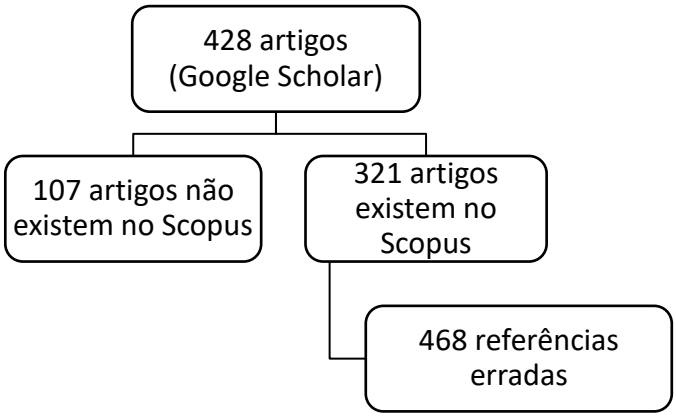


Figura 46: Diagrama de Resultados (José Machado)

Então, tendo em conta o tempo médio de análise por cada artigo, este processo realizado manualmente demoraria 5 horas e 21 minutos aproximadamente, enquanto o *script* demorou aproximadamente 6 minutos e 19 segundos, ou seja, houve uma redução de 98,03% de tempo, provando a eficácia do algoritmo a detetar as referências erradas no Scopus.

5.2 Algoritmo de agregação e cruzamento de dados

O primeiro investigador a ser testado foi o professor e orientador Filipe Portela. Dos vários artigos, foi destacado o artigo *"Towards the development of a data science modular solution"* da plataforma Scopus na Figura 47 como exemplo do sucesso do algoritmo. Este artigo possui 2 citações na plataforma Scopus e 2 citações na plataforma Semantic Scholar. Após a execução do algoritmo, o artigo passou a ter 4 citações, sendo este o número mais próximo da realidade.


```
"title": "Towards the development of a data science modular solution",
"doi": "10.1109/FiCloudW.2019.00030",
"fromPlatform": "Scopus",
"citations": [
  {
    "title": "Intelligent Dashboards for Car Parking Flow Monitoring",
    "doi": "10.1007/978-3-031-30514-6_3",
    "fromPlatform": "Scopus"
  },
  {
    "title": "Augmented Virtual Reality in Data Visualization",
    "doi": "10.1007/978-3-031-20319-0_22",
    "fromPlatform": "Scopus"
  },
  {
    "title": "Implementasi OLAP pada Data Kerja Praktik dan Tugas Akhir Menggunakan Framework Modular Cube JS",
    "doi": "10.19184/isj.v6i3.27614",
    "fromPlatform": "Semantic_Scholar"
  },
  {
    "title": "A Practical Solution to Synchronise Structured and Non-structured Repositories",
    "doi": "10.1007/978-3-030-72654-6_35",
    "fromPlatform": "Semantic_Scholar"
  }
]
},
```


Figura 47: Artigo destacado (Filipe Portela)

Para provar o funcionamento do algoritmo está representado na Figura 48 o artigo destacado com as citações provenientes da plataforma Scopus.

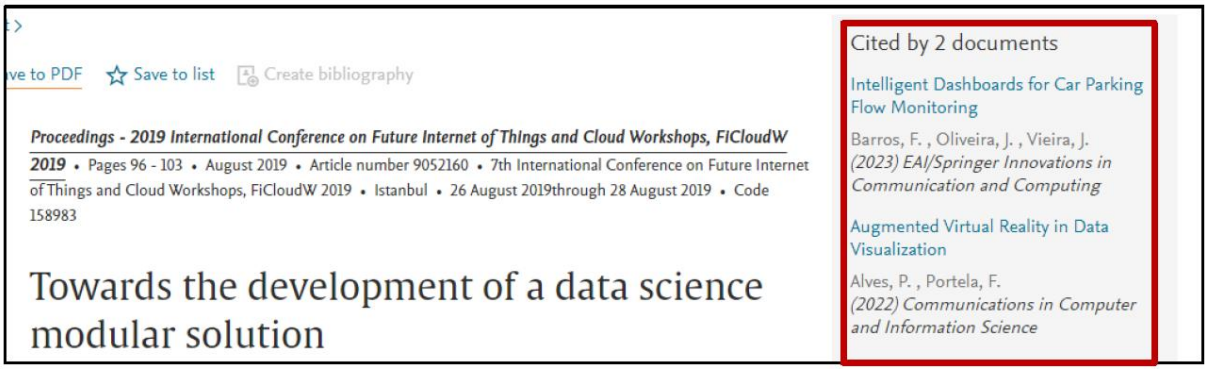


Figura 48: Citações do artigo destacado na plataforma Scopus (Filipe Portela)

Quanto às citações provenientes da plataforma Semantic Scholar, estas estão representadas na Figura 49.

Figura 49: Citações do artigo destacado na plataforma Semantic Scholar (Filipe Portela)

O segundo utilizador a ser testado foi o professor Jorge Sá. Dos vários artigos, foi destacado o artigo *"Performance Evaluation in IST Projects: A Case Study"* da plataforma Scopus na Figura 50 como exemplo do sucesso do algoritmo. Este artigo possui 1 citação na plataforma Scopus e 1 citação na plataforma Semantic Scholar. Após a execução do algoritmo, o artigo passou a ter 2 citações, sendo este o número mais próximo da realidade.

```
{
  "title": "Performance Evaluation in IST Projects: A Case Study",
  "doi": "10.1007/978-3-030-14850-8_2",
  "fromPlatform": "Scopus",
  "citations": [
    {
      "title": "The Interplay of Management Information Systems in Industry 4.0: A Bibliometric Review",
      "doi": "10.1007/978-3-030-96150-3_21",
      "fromPlatform": "Scopus"
    },
    {
      "title": "Approaches and Methods for Individual Performance Assessment in Information Systems Projects",
      "doi": "10.4018/978-1-7998-1279-1.ch024",
      "fromPlatform": "Semantic_Scholar"
    }
  ]
},
```


Figura 50: Artigo destacado (Jorge Sá)

Para provar o funcionamento do algoritmo, está representado na Figura 51 o artigo destacado com a citação proveniente da plataforma Scopus.

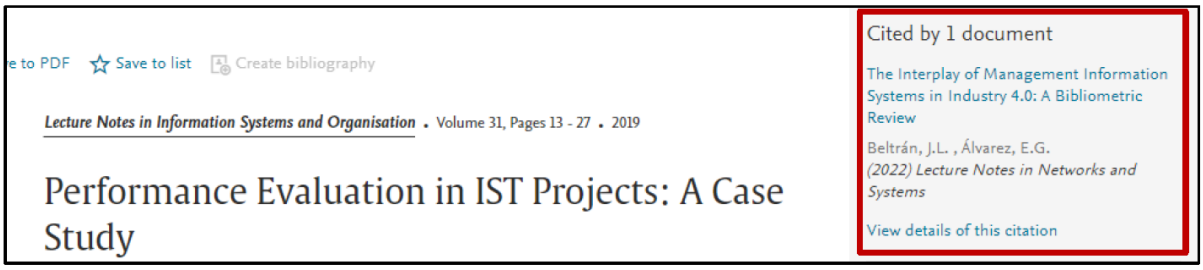


Figura 51: Citações do artigo destacado na plataforma Scopus (Jorge Sá)

Quanto à citação proveniente da plataforma Semantic Scholar, esta está representada na Figura 52.

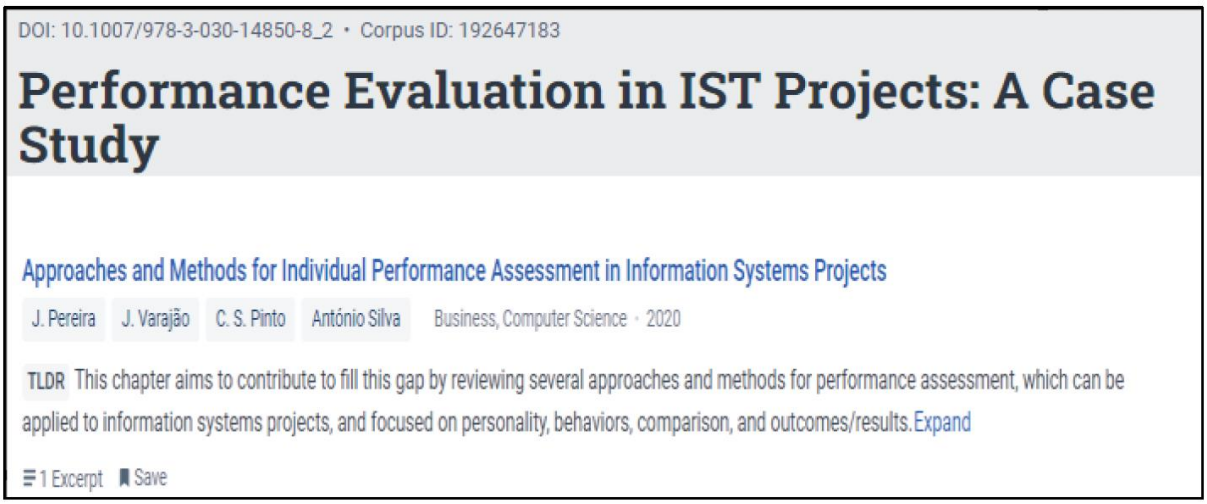


Figura 52: Citações do artigo destacado na plataforma Semantic Scholar (Jorge Sá)

O terceiro e último investigador a ser testado foi o professor José Machado. Dos vários artigos, foi destacado o artigo *"Diagnosis of Diabetic Retinopathy Using Data Mining Classification Techniques"* da plataforma Scopus na Figura 53 como exemplo do sucesso do algoritmo. Este artigo possui 1 citação na plataforma Scopus e 2 citações na plataforma Semantic Scholar. Após a execução do algoritmo, o artigo passou a ter 3 citações, sendo este o número mais próximo da realidade.


```
{
  "title": "Diagnosis of Diabetic Retinopathy Using Data Mining Classification Techniques",
  "doi": "10.1007/978-3-030-71782-7_18",
  "fromPlatform": "Scopus",
  "citations": [
    {
      "title": "Prediction of diabetic Retinopathy using Health Records with Machine Learning Classifiers and data Science
      "doi": "10.4018/IJRQEH.299959",
      "fromPlatform": "Scopus"
    },
    {
      "title": "Prediction of Diabetic Retinopathy Using Health Records with Machine Learning Classifiers and Data Science
      "doi": "10.4018/ijrqeh.299959",
      "fromPlatform": "Semantic_Scholar"
    },
    {
      "title": "Diabetic Retinopathy Classification using Support Vector Machine with Hyperparameter Optimization",
      "fromPlatform": "Semantic_Scholar"
    }
  ]
},
```


Figura 53: Artigo destacado (José Machado)

Para provar o funcionamento do algoritmo, está representado na Figura 54 o artigo destacado com a citação proveniente da plataforma Scopus.

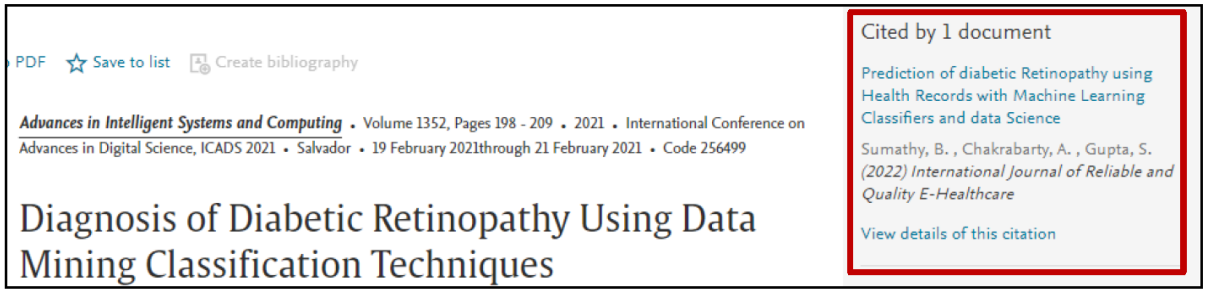


Figura 54: Citações do artigo destacado na plataforma Scopus (José Machado)

Quanto às citações provenientes da plataforma Semantic Scholar, estas estão representadas na Figura 55.

Figura 55: Citações do artigo destacado na plataforma Semantic Scholar (José Machado)

Deve ter-se em consideração que durante o desenvolvimento do algoritmo foram feitos testes com conjunto de dados mais pequenos para ser possível visualizar o correto funcionamento do algoritmo antes de ser testado no conjunto de dados maiores dos investigadores. Nestes testes realizados, não é possível provar que o algoritmo produza um ficheiro com artigos e citações únicas devido às respostas das API das plataformas serem inconsistentes entre si, isto é, caso o algoritmo tenha de usar o título como método de comparação, esse mesmo título numa plataforma pode conter um caracter especial que outra plataforma não contém, levando o algoritmo a considerar, incorretamente, que os artigos são diferentes. Para combater este problema são usadas técnicas como: colocar todas as palavras em minúsculo e remover caracteres especiais com o objetivo de manter a consistência durante a comparação.

Apesar disto, o algoritmo foi capaz de cruzar os dados com alguma taxa de acerto devido à maior parte dos artigos e citações possuírem um DOI, aproximando os artigos do seu número de citações real. No entanto, podem existir casos fora do alcance do algoritmo que podem ser corrigidos conforme este é usado pelos investigadores.

6. CONCLUSÃO

Nesta secção são feitas considerações finais sobre este projeto de dissertação onde é respondida à questão *"De que forma é possível recolher e agrupar os dados de várias plataformas científicas?"* e é também feito um resumo sobre o trabalho desenvolvido. São ainda descritos os riscos inerentes ao projeto e quais os que, evidentemente, aconteceram durante o desenvolvimento da solução. Finalmente são dadas sugestões para o trabalho futuro no sentido de desenvolver continuamente a plataforma RDProfile.

6.1 Considerações finais

Depois da conclusão deste projeto de dissertação, é necessário averiguar se o mesmo foi de encontro aos objetivos inicialmente propostos. O objetivo principal deste projeto era estudar a viabilidade da combinação de dados de diversas plataformas de forma a otimizar o protótipo existente e, acima de tudo, explorar uma alternativa mais fiável para o mecanismo de citações que tem diversas falhas, como é o caso do Scopus.

O objetivo proposto foi alcançado com sucesso e a questão de investigação "*"De que forma é possível recolher e agrupar os dados de várias plataformas científicas?"* foi respondida através da concretização dos objetivos secundários e da elaboração da revisão da literatura e do desenvolvimento da solução. Na Tabela 17 estão representados os objetivos definidos e resultados obtidos na conclusão do projeto de dissertação.

Tabela 17: Objetivos e Resultados

Objetivos	Resultados
Criar um *web crawler*	*Web crawler* capaz de extrair dados do Google Scholar
Integrar diversas API	Integração das API das plataformas: • ORCID; • Google Scholar; • Scopus; • Semantic Scholar; • Open Alex.

Objetivos	Resultados
Desenvolver algoritmos capazes de agregar e cruzar dados	• Algoritmo de deteção de referências erradas no Scopus; • Algoritmo de agregação e cruzamento de dados.
Desenvolver modelos interoperáveis	• Modelo de dados ORCID; • Modelo de dados Google Scholar; • Modelo de dados Scopus; • Modelo de dados Semantic Scholar; • Modelo de dados OpenAlex; • Modelo de dados de referências erradas; • Modelo de dados de agregação e cruzamento de dados.

Apesar de inicialmente ter sido proposta a integração das plataformas Research Gate e Web Of Science, devido à primeira plataforma não permitir *scraping* e não possuir uma API oficial e, a segunda plataforma não ter respondido ao pedido de acesso à API necessária para obter os dados necessários para o algoritmo de agregação e cruzamento de dados. No entanto, o algoritmo está estruturado de forma que a adição de novas plataformas seja um processo simples.

Durante o desenvolvimento deste projeto de dissertação, foi elaborado um artigo intitulado "*Towards a Universal and Interoperable Scientific Data Model*" publicado na conferência *Slate* que teve o objetivo de criar uma plataforma robusta e interoperável baseada num algoritmo inovador de fusão de bibliotecas (Apêndice II – Publicação Científica).

Com a conclusão deste projeto de dissertação foram integradas 5 plataformas, nomeadamente o ORCID, Scopus, Google Scholar (*web crawler*), Semantic Scholar e OpenAlex, criados 2 algoritmos, algoritmo de deteção de referências erradas na plataforma Scopus e algoritmo de agregação e cruzamento de dados e 7 modelos de dados (modelo de dados ORCID, modelo de dados Google Scholar, modelo de dados Scopus, modelo de dados Semantic Scholar, modelo de dados OpenAlex, modelo de dados de referências erradas e modelo de dados de agregação e cruzamento de dados).

6.2 Riscos

Todos os projetos têm riscos inerentes durante a sua execução, desta forma foi criada a Tabela 18 que lista alguns riscos que são possíveis de acontecer.

A Tabela 18 contém uma breve descrição do risco, a probabilidade de o mesmo acontecer, o impacto que irá causar no projeto, a severidade, resultado da multiplicação destas duas últimas variáveis e, no fim, a ação mitigadora caso o risco aconteça. As variáveis Probabilidade (P) e Impacto (I) partilham a mesma escala que vai de 1 (pouco provável/impacto) até 5 (muito provável/impacto), enquanto a Severidade (P*I) tem a escala de 1 (pouco severo) a 25 (muito severo). Os riscos estão ordenados por ordem descendente, conforme o grau de severidade.

Tabela 18: Tabela de riscos

ID	Risco	(P) (1-5)	(I) (1-5)	(P*I) (1-25)	Ocorreu?	Ação realizada
1	Falta de experiência	3	4	12	Sim	Ler documentação, ver vídeos sobre a tecnologia, pedir ajuda ao orientador
2	Elevada complexidade do projeto	3	4	12	Sim	Pedir auxílio dos orientadores, de modo que todas as dúvidas sejam esclarecidas em tempo útil
3	Falta de reposta para acesso às API	2	5	10	Sim	Contactar continuamente a entidade ou encontrar uma alternativa à API
4	Incumprimento de prazos	2	5	10	Não	-

ID	Risco	(P) (1-5)	(I) (1-5)	(P*I) (1-25)	Ocorreu?	Ação realizada
5	Incumprimento do plano de trabalho	3	3	6	Não	-
6	Falta de experiência com as metodologias do projeto	3	3	6	Não	-
7	Perda de arquivos	1	5	5	Não	-

Durante o desenvolvimento deste projeto de dissertação existiram alguns problemas, principalmente associados à falta de experiência com esta tecnologia. O uso do Puppeteer, uma biblioteca que permite controlar o *browser* por código, em conjunto com a falta de conhecimento técnico do carregamento das páginas *web*, dificultou a integração de plataformas científicas que não possuem uma API. Para mitigar este risco foi lida a documentação da biblioteca e estudado o funcionamento das páginas *web.* Além deste, o risco "elevada complexidade do projeto" aconteceu devido a plataformas, como o Google Scholar, necessitarem do uso de outras tecnologias adicionais para extrair dados. Para mitigar este risco foi discutido com o orientador a impossibilidade de extrair a informação necessária nesta plataforma e a necessidade de explorar outras alternativas para contornar esta dificuldade. Finalmente o risco "falta de resposta para acesso às API" aconteceu, por exemplo, com a plataforma Web of Science em que foi feito um pedido de acesso à API, no entanto, até à data da finalização desta dissertação, não foi fornecido o acesso à mesma, apesar das várias tentativas de contacto. Para mitigar este risco foram adicionadas outras plataformas como alternativa.

6.3 Trabalho futuro

Os objetivos propostos para este projeto de dissertação foram atingidos, no entanto ainda existem bastantes aspetos a melhorar na plataforma desenvolvida.

O trabalho futuro passará por integrar mais plataformas científicas, conforme estas disponibilizem API ou permitam *scraping*, para tornar a plataforma RDProfile mais robusta e com indicadores cada vez mais precisos.

Como foi mencionado na secção 5.2, quanto mais esta versão do algoritmo de agregação e cruzamento de dados for testada e usada, mais casos não contemplados neste projeto de dissertação serão adicionados tornando o algoritmo cada vez mais sólido e capaz de determinar o número real de citações.

REFERÊNCIAS

Blackburn, R. (2022, November 14). *What are the differences between the public and member APIs?* https://info.orcid.org/ufaqs/what-are-the-differences-between-the-public-and-member-apis/

Cho, J., Hebj, G., & Principaladvisei, I. (2001). *Crawling the web: Discovery and maintenance os large scale web data*. https://www.researchgate.net/publication/2575416_Crawling_The_Web_Discovery_And_Maintenance_Of_Large-Scale_Web_Data

Christopher Newport University. (2018, October 9). *Google scholar: An introduction*. https://cnu.libguides.com/GoogleScholarGuide/Home

Clarivate. (2023a). *About web of science*. http://webofscience.help.clarivate.com/en-us/Content/home.htm

Clarivate. (2023b). *Web of science API expanded*. https://developer.clarivate.com/apis/wos

Clarivate. (2023c). *Web of science researcher profile metrics*. http://webofscience.help.clarivate.com/en-us/Content/researcher-profile-metrics.html?Highlight=metrics

CRACS & Inesc TEC. (2023). Sobre o authenticus. https://www.authenticus.pt/pt/home/view_article/1

Elsevier. (n.d.). *Expertly curated abstract & citation database*. Abstract and Citation Database. *Retrieved February 7, 2023, from* https://www.elsevier.com/solutions/scopus

Elsevier. (2023a). *Elsevier developer portal*. https://dev.elsevier.com/api_docs.html

Elsevier. (2023b). *Metrics to see who engages with your institution's research*. https://www.elsevier.com/solutions/scopus/how-scopus-works/metrics

Fundação para a Ciência e Tecnologia. (n.d.). Mais informação. *Retrieved February 23, 2023, from* https://www.cienciavitae.pt/mais-informacao/

Google. (n.d.-a). *About google scholar*. *Retrieved February 7, 2023, from* https://scholar.google.com/intl/pt-PT/scholar/about.html

Google. (n.d.-b). *Google scholar metrics*. *Retrieved February 13, 2023, from* https://scholar.google.com/intl/en/scholar/metrics.html#metrics

Google. (2023, May 23). *In-depth guide to how google search works*. https://developers.google.com/search/docs/fundamentals/how-search-works

Hirsch, J. E. (2005). *An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Jones, N. (2015). *Artificial-intelligence institute launches free science search engine. Nature*. https://doi.org/10.1038/nature.2015.18703

MDN web docs. (2023, September 25). *JavaScript*. https://developer.mozilla.org/en-US/docs/Web/JavaScript

Microsoft. (n.d.). *Visual Studio Code*. *Retrieved May 25, 2023, from* https://code.visualstudio.com/docs

MongoDB. (n.d.). O que é o MongoDB? *Retrieved May 30, 2023, from* https://www.mongodb.com/pt-br/what-is-mongodb

Mongoose. (n.d.). *Mongoose. Retrieved May 25, 2023, from* https://mongoosejs.com/

OpenJS Foundation. (n.d.). *About Node.js*. *Retrieved May 25, 2023, from* https://nodejs.org/en/about

ORCID. (n.d.). *About ORCID*. *Retrieved February 7, 2023, from* https://info.orcid.org/what-is-orcid/

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). *A design science research methodology for information systems research. Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Portela, F. (2022). Ficha técnica.

Postman. (n.d.). *What is Postman? Retrieved May 27, 2023, from* https://www.postman.com/product/what-is-postman/

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. *Retrieved September 8, 2023, from* *https://openalex.org/about*

Redocly. (2023). *Academic Graph API*. https://api.semanticscholar.org/api-docs/

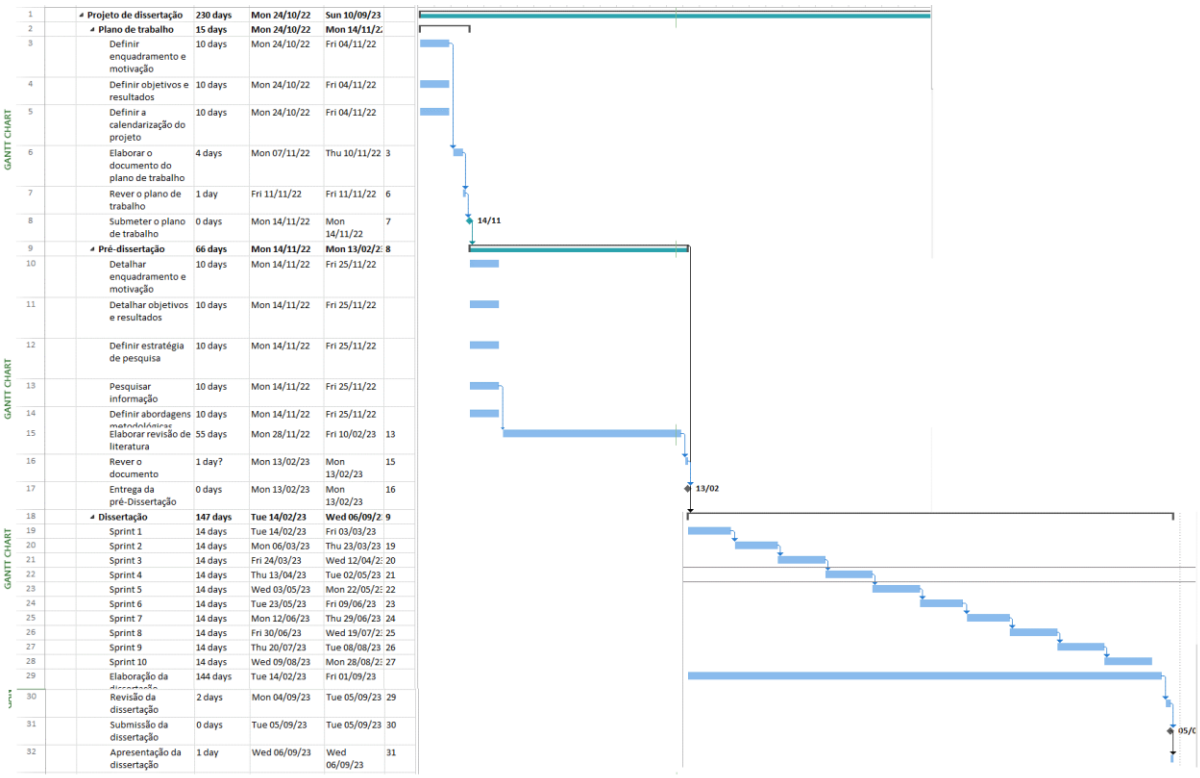
Research Gate. (2023). *About us*. https://www.researchgate.net/about

Schwaber, K., & Sutherland, J. (2020). *The scrum guide: The definitive guide to scrum*. https://scrumguides.org/

Semantic Scholar. (n.d.-a). *Introducing semantic reader*. *Retrieved August 19, 2023, from* https://www.semanticscholar.org/product/semantic-reader

Semantic Scholar. (n.d.-b). *Research feeds: What are research feeds? Retrieved August 19, 2023, from* https://www.semanticscholar.org/faq/what-are-research-feeds

APÊNDICE I – DIAGRAMA DE GANTT



1		**Projeto de dissertação**	**230 days**	**Mon 24/10/22**	**Sun 10/09/23**	
2		**Plano de trabalho**	**15 days**	**Mon 24/10/22**	**Mon 14/11/2**	
3		Definir enquadramento e motivação	10 days	Mon 24/10/22	Fri 04/11/22	
4		Definir objetivos e resultados	10 days	Mon 24/10/22	Fri 04/11/22	
5		Definir a calendarização do projeto	10 days	Mon 24/10/22	Fri 04/11/22	
6		Elaborar o documento do plano de trabalho	4 days	Mon 07/11/22	Thu 10/11/22	3
7		Rever o plano de trabalho	1 day	Fri 11/11/22	Fri 11/11/22	6
8		Submeter o plano de trabalho	0 days	Mon 14/11/22	Mon 14/11/22	7
9		**Pré-dissertação**	**66 days**	**Mon 14/11/22**	**Mon 13/02/2**	**8**
10		Detalhar enquadramento e motivação	10 days	Mon 14/11/22	Fri 25/11/22	
11		Detalhar objetivos e resultados	10 days	Mon 14/11/22	Fri 25/11/22	
12		Definir estratégia de pesquisa	10 days	Mon 14/11/22	Fri 25/11/22	
13		Pesquisar informação	10 days	Mon 14/11/22	Fri 25/11/22	
14		Definir abordagens metodológicas	10 days	Mon 14/11/22	Fri 25/11/22	
15		Elaborar revisão de literatura	55 days	Mon 28/11/22	Fri 10/02/23	13
16		Rever o documento	1 day?	Mon 13/02/23	Mon 13/02/23	15
17		Entrega da pré-Dissertação	0 days	Mon 13/02/23	Mon 13/02/23	16
18		**Dissertação**	**147 days**	**Tue 14/02/23**	**Wed 06/09/2**	**9**
19		Sprint 1	14 days	Tue 14/02/23	Fri 03/03/23	
20		Sprint 2	14 days	Mon 06/03/23	Thu 23/03/23	19
21		Sprint 3	14 days	Fri 24/03/23	Wed 12/04/2	20
22		Sprint 4	14 days	Thu 13/04/23	Tue 02/05/23	21
23		Sprint 5	14 days	Wed 03/05/23	Mon 22/05/2	22
24		Sprint 6	14 days	Tue 23/05/23	Fri 09/06/23	23
25		Sprint 7	14 days	Mon 12/06/23	Thu 29/06/23	24
26		Sprint 8	14 days	Fri 30/06/23	Wed 19/07/2	25
27		Sprint 9	14 days	Thu 20/07/23	Tue 08/08/23	26
28		Sprint 10	14 days	Wed 09/08/23	Mon 28/08/2	27
29		Elaboração da dissertação	144 days	Tue 14/02/23	Fri 01/09/23	
30		Revisão da dissertação	2 days	Mon 04/09/23	Tue 05/09/23	29
31		Submissão da dissertação	0 days	Tue 05/09/23	Tue 05/09/23	30
32		Apresentação da dissertação	1 day	Wed 06/09/23	Wed 06/09/23	31

APÊNDICE II – PUBLICAÇÃO CIENTÍFICA

Título: *Towards a Universal and Interoperable Scientific Data Model*

Autores: João Oliveira, Diogo Gomes, Francisca Santana, Jorge Oliveira e Sá, Filipe Portela

Conferência: Slate 2023

Estado: Publicado (doi.org/10.4230/OASIcs.SLATE.2023.14)

Resumo: *The growing number of researchers in Portugal has intensified the appearance of several scientific platforms that allow the indexation of publications and the management of scientific profiles. The diversity and high number of platforms brings problems at the level of crossover and integrity of the information, i.e., the researchers' profiles are rarely updated, and their data are not properly grouped and cross-referenced. Hence, the need arises for a more global platform that enables the synchronization of information, free from constraints imposed by existing data.The study and work carried out aims to solve this problem by creating a robust and interoperable platform based on an innovative library merge algorithm. Thus, this platform includes information regarding publications, researchers, and scientific indicators, by crossing and grouping data from several platforms.*

Palavras-Chave: RDProfile, Researchers, Scientific Platforms, Scientific Data

ANEXO I

Lista de Atividades

Na Tabela 19 é apresentado o planeamento de projeto com as tarefas principais a serem executadas e as suas respetivas durações, respeitando os prazos definidos.

Tabela 19: Planeamento do projeto

ID	Tarefa	Data de início	Data do fim	Dependências
1	**Projeto de dissertação**	**24/10/2022**	**28/07/2023**	
1.1	***Plano de trabalho***	***24/10/2022***	***14/11/2022***	
1.1.1	Definir enquadramento e motivação	24/10/2022	04/11/2022	
1.1.2	Definir objetivos e resultados	24/10/2022	04/11/2022	
1.1.3	Definir a calendarização do projeto	24/10/2022	04/11/2022	
1.1.4	Elaborar o documento do plano de trabalho	07/11/2022	10/11/2022	1.1.1
1.1.5	Rever o plano de trabalho	11/11/2022	11/11/2022	1.1.4
1.1.6	(M) Submeter o plano de trabalho	14/11/2022	14/11/2022	1.1.5
1.2	***Pré-Dissertação***	***15/11/2022***	***13/02/2023***	1.1.6
1.2.1	Detalhar o enquadramento e motivação	14/11/2022	25/11/2022	
1.2.2	Detalhar objetivos e resultados	14/11/2022	25/11/2022	
1.2.3	Definir estratégia de pesquisa	14/11/2022	25/11/2022	
1.2.4	Pesquisar informação	14/11/2022	25/11/2022	
1.2.5	Definir abordagens metodológicas	14/11/2022	25/11/2022	
1.2.6	Elaborar revisão de literatura	28/11/2022	10/02/2023	1.2.4

ID	Tarefa	Data de início	Data do fim	Dependências
1.2.7	Rever documento	13/02/2023	13/02/2023	1.2.6
1.2.8	(M) Entregar a pré-dissertação	13/02/2023	13/02/2023	1.2.7
1.3	***Dissertação***	***14/02/2023***	***06/09/2023***	*1.2*
1.3.1	Sprint 1	14/02/2023	03/03/2023	
1.3.2	Sprint 2	06/02/2023	23/03/2023	1.3.1
1.3.3	Sprint 3	24/03/2023	12/04/2023	1.3.2
1.3.4	Sprint 4	13/04/2023	02/05/2023	1.3.3
1.3.5	Sprint 5	03/05/2023	22/05/2023	1.3.4
1.3.6	Sprint 6	23/05/2023	09/06/2023	1.3.5
1.3.7	Sprint 7	12/06/2023	29/06/2023	1.3.6
1.3.8	Sprint 8	30/06/2023	19/07/2023	1.3.7
1.3.9	Sprint 9	20/07/2023	08/08/2023	1.3.8
1.3.10	Sprint 10	09/08/2023	28/08/2023	1.3.9
1.3.11	Elaborar a dissertação	14/02/2023	01/09/2023	
1.3.12	Rever dissertação	04/09/2023	05/09/2023	1.3.11
1.3.13	(M) Submeter de dissertação	05/09/2023	05/09/2023	1.3.12
1.3.14	Apresentar a dissertação	06/09/2023	06/09/2023	1.3.13

Legenda:

M – *Milestone*