RESIGeT On-Line
Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica
Vol. 14 Nro. 1 2024
Uni-FACEF Centro Universitário Municipal de Franca

WEB SCRAPING E A BUSCA PELA GARANTIA DA QUALIDADE DOS DADOS EM UMA COLETA AUTOMATIZADA A PARTIR DE UMA REVISÃO BIBLIOGRÁFICA

Anna Laura Magalhães Porto
Discente do curso Gestão da Informação - Universidade Federal de Goiás.
annaporto@discente.ufg.br

Douglas Farias Cordeiro
Docente da Universidade Federal de Goiás.
cordeiro@ufg.br

Resumo: O artigo busca analisar o uso de web scraping como uma ferramenta automatizada para a coleta de dados em ambientes digitais, destacando sua relação com a garantia da qualidade dos dados coletados. Utilizando uma metodologia mista, integrando elementos quantitativos e qualitativos, a pesquisa utiliza uma revisão bibliográfica de 40 artigos publicados entre 2015 e 2024, e aborda os desafios técnicos, éticos e legais da técnica, como mudanças frequentes na estrutura dos sites e questões relacionadas à extração de dados incompletos ou desatualizados. Os resultados reforçam a necessidade de práticas consistentes para assegurar a confiabilidade e a relevância dos dados coletados.

Palavras-chave: Web scraping; revisão bibliográfica; qualidade de dados.

Abstract: The article seeks to analyze the use of web scraping as an automated tool for collecting data in digital environments, highlighting its relationship with ensuring the quality of the data collected. Using a mixed methodology, integrating quantitative and qualitative elements, the research uses a bibliographic review of 40 articles published between 2015 and 2024, and addresses the technical, ethical and legal challenges of the technique, such as frequent changes in the structure of websites and issues related to data extraction. incomplete or outdated data. The results reinforce the need for consistent practices to ensure the reliability and relevance of the data collected.

Word Keys: Web scraping; bibliographic review; data quality.

1 Introdução

Nas últimas décadas, os avanços das Tecnologias de Informação e Comunicação (TICs) transformaram a forma como dados são gerados, consumidos e utilizados em diversos setores. Segundo Oliveira e Moura (2015), as TICs podem ser definidas como “um conjunto de recursos tecnológicos integrados entre si, que proporcionam, por meio das funções de software e telecomunicações, a automação e comunicação dos processos de negócios, da pesquisa científica e de ensino e aprendizagem”. Essa integração tecnológica levou ao crescimento na geração de dados, criando o fenômeno conhecido como Big Data. Como aponta McAfee (2012), a análise desses grandes

volumes de dados desempenha um papel fundamental no apoio à tomada de decisões e na geração de informações estratégicas.

Apesar da grande quantidade de dados disponíveis, as necessidades específicas de organizações e pesquisadores nem sempre são atendidas por informações acessíveis de forma explícita e direta. Dados relevantes frequentemente encontram-se dispersos, armazenados em formatos não padronizados ou restritos a interfaces de difícil acesso. Nessas situações, torna-se necessário recorrer a abordagens alternativas para obter essas informações. Uma das técnicas mais utilizadas nesse contexto é o *web scraping*, que permite a extração automatizada de dados diretamente de páginas da web. Segundo Chapagain (2023, p. 4), "a raspagem é um processo de extração, cópia, triagem ou coleta de dados. A raspagem ou extração de dados da web (sites, páginas da web e recursos relacionados à internet) para determinadas necessidades é normalmente chamada de *web scraping*". Essa prática se destaca por oferecer uma solução eficiente para capturar informações que, de outra forma, não estariam disponíveis, o que se torna essencial em cenários onde o acesso direto a dados é limitado (VIEIRA; CORDEIRO, 2023; Cordeiro et al., 2024).

Embora o *web scraping* ofereça uma solução eficaz para a coleta de dados, seu uso apresenta desafios significativos. Problemas como a extração de dados incompletos, inconsistentes ou desatualizados podem comprometer a confiabilidade das análises. Além disso, mudanças frequentes na estrutura das páginas web podem tornar inviáveis as rotinas automatizadas, exigindo atualizações constantes nos métodos de coleta. Nesse contexto, a avaliação da qualidade dos dados obtidos torna-se indispensável para garantir que as informações extraídas sejam consistentes, precisas e relevantes para os objetivos pretendidos. Diante desses desafios, torna-se evidente a necessidade de explorar como o *web scraping* pode ser melhorado para assegurar a qualidade dos dados coletados.

A partir disso, o objetivo geral é analisar a relação entre web scraping e a qualidade dos dados, destacando as práticas que asseguram a consistência e a integridade das informações extraídas da coleta automatizada, com base em um estudo bibliográfico. A justificativa para pesquisar sobre web scraping e a garantia da qualidade dos dados é motivada pela complexidade e pelas inconsistências frequentemente encontradas na coleta automatizada. Embora o *web scraping* seja uma alternativa valiosa para a coleta de dados, ele é muitas vezes a última opção e pode ser uma das menos confiáveis. Isso se deve ao fato de que um código desenvolvido para raspar dados de uma página pode funcionar hoje, mas não necessariamente funcionará no futuro devido às mudanças na estrutura das páginas web e na dinâmica de manutenção desses sites.

Essa dependência da estrutura das páginas e dos sistemas utilizados para o scraping pode resultar em dados ausentes ou problemáticos, exigindo uma verificação constante por parte dos desenvolvedores. Além disso, os códigos de web scraping podem introduzir "sujeira" ou "lixo" na base de dados, comprometendo a qualidade dos dados coletados. Com o uso contínuo dessa técnica e a crescente demanda por informações úteis para a tomada de decisões, surgem novos desafios e a principal questão que se coloca é: como assegurar a qualidade desses dados obtidos? E até que ponto o acesso a dados em páginas da internet por meio de mecanismos automatizados está em conformidade com a legislação? Por essa razão, essa pesquisa é crucial, pois dados imprecisos podem levar a conclusões erradas, afetando estratégias de negócios e pesquisas acadêmicas.

2 Referencial Teórico

Em 2010, Tim Berners-Lee, inventor da Web conhecida atualmente, introduziu o sistema de estrelas para classificar o nível de abertura dos dados, incentivando órgãos governamentais a disponibilizarem suas informações de maneira mais acessível e estruturada. Segundo Berners-Lee (2010, *online*) "o sistema ajuda a diagnosticar o nível de abertura de dados dos órgãos públicos e fornece degraus alcançáveis para se chegar a níveis cada vez mais refinados de dados abertos". Esse sistema é particularmente relevante para a parte de acesso aos dados, pois estabelece critérios claros que ajudam a avaliar e melhorar a qualidade da informação disponibilizada. Para efetivar essa abertura, é essencial realizar um levantamento das estratégias de obtenção de dados, que podem incluir Sistemas de Gerenciamento de Banco de Dados (SGBD), APIs e a disponibilização direta de arquivos para download. Esses arquivos podem variar em formato, abrangendo dados estruturados, semi-estruturados e não estruturados, além de formatos que não são processáveis por máquinas. Como destacado no guia do Comitê Gestor da Internet no Brasil CGI.br (s.d, *online*), "o objetivo dessa organização é que toda a sociedade que tem o acesso à informação seja capaz de interpretar os dados de maneira uniforme", a padronização e a documentação adequada dos elementos utilizados são cruciais para garantir que todos os usuários possam interpretar os dados de maneira uniforme. De acordo com a Open Knowledge Foundation (s.d, *online*), os dados abertos são informações que podem ser acessadas, utilizadas e redistribuídas livremente por qualquer pessoa. Para serem considerados verdadeiramente abertos, esses dados devem estar disponíveis em formatos legíveis por máquina e sob licenças que permitam sua reutilização.

No entanto, ao considerar a questão da abertura de dados no contexto privado, é importante reconhecer que as empresas frequentemente hesitam em compartilhar suas informações devido a preocupações com concorrência e segurança. Como observa Braga e Tuzzo (2017, p. 59), "sabemos que as empresas do setor privado não publicam seus dados abertamente por conta da concorrência e por segurança, com o objetivo de proteger informações". Essa resistência à abertura evidencia uma lacuna na compreensão do conceito de dados abertos. A prática de web scraping pode ser uma ferramenta valiosa na coleta desses dados abertos, mas a qualidade das informações obtidas depende fortemente da forma como estão estruturadas e disponibilizadas. Dados em formatos não estruturados ou semi-estruturados podem exigir processamento adicional para serem utilizados efetivamente. Além disso, a qualidade dos dados é diretamente impactada pela forma como são publicados; dados abertos, sejam eles governamentais ou privados, quando bem estruturados facilitam o scraping eficiente e garantem que as informações sejam precisas e utilizáveis.

A relação entre dados abertos e acessibilidade é uma necessidade crescente para qualquer tipo de negócio, pois as organizações podem ter demandas por informações provenientes de múltiplas fontes, sejam públicas ou privadas, incluindo dados acessíveis diretamente ou disponibilizados indiretamente através de relatórios e sites. Portanto, promover uma cultura de abertura no setor privado não apenas atende à demanda por transparência, mas também cria oportunidades para inovação e colaboração no mercado. Como afirmam Isotani e Bittencourt (2015, p. 42), "dados abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa", enfatizando a importância da acessibilidade na era digital.

2.1 Dado, Informação e Conhecimento

Os dados são a matéria-prima da informação, consistindo em fatos ou eventos isolados que, por si só, não transmitem significado. Segundo o dicionário Aurélio (FERREIRA, 1999) o dado é o elemento inicial de qualquer ato de conhecimento, essa definição ressalta a natureza bruta dos dados, que se tornam significativos apenas quando processados e organizados. Em um contexto de *web scraping*, a qualidade dos dados coletados é crucial; dados mal estruturados podem levar a análises incorretas e decisões equivocadas.

A informação é o resultado do processamento e da organização dos dados. Ela confere significado aos dados ao contextualizá-los, permitindo que sejam compreendidos e utilizados para a tomada de decisões. McGarry (1984) define informação como dar forma ou aparência aos dados brutos, enfatizando que ela é mais complexa do que os dados isolados. A informação é essencial nesse contexto, pois transforma dados brutos em insights úteis que podem ser aplicados em análises. Hoshovsky e Massey (1968) afirmam que a informação é o dado mais a avaliação para uso futuro antecipado, indicando que a avaliação crítica dos dados é fundamental para sua transformação em informações úteis.

O conhecimento vai além da informação ao incorporar experiência, contexto e interpretação pessoal. Davenport (1998) define o conhecimento como uma crença produzida por uma informação, destacando a relação intrínseca entre esses conceitos. O conhecimento pode ser dividido em duas categorias: tácito, que é difícil de transferir e está ligado à experiência pessoal; e explícito, que pode ser documentado e compartilhado formalmente. Essa distinção é importante porque o conhecimento tácito frequentemente enriquece o conhecimento explícito, proporcionando uma base sólida para a inovação e a resolução de problemas.

A relação entre dado, informação e conhecimento é vital para o sucesso em projetos de web scraping. Dados bem geridos são transformados em informações significativas que, por sua vez, alimentam o conhecimento coletivo da organização. Essa compreensão é essencial para garantir a qualidade dos dados coletados, assegurando que as informações extraídas sejam consistentes e integradas.

2.2 Sistemas Gerenciadores de Banco de Dados

Um Sistema Gerenciador de Banco de Dados (SGBD) é uma tecnologia essencial que permite a construção, definição, manutenção e compartilhamento de dados armazenados em um banco de dados entre usuários e aplicações. Segundo Elmasri e Navathe (2011), um SGBD é responsável por receber solicitações, encaminhá-las ao banco de dados e reunir ou subdividir os dados conforme necessário. Além disso, ele atua no controle de redundância, mantém a consistência dos dados e gerencia o acesso e a concorrência dos usuários. Os dados podem ser obtidos por meio de um SGBD através de várias operações fundamentais. Primeiramente, o SGBD permite a definição do esquema do banco de dados, onde os tipos de dados e as estruturas são especificados. Em seguida, ocorre a construção, que envolve o armazenamento físico

dos dados. A manipulação dos dados é realizada por meio de comandos que permitem a recuperação, atualização, inclusão ou exclusão de informações. Por fim, o compartilhamento possibilita que múltiplos usuários acessem os dados simultaneamente, garantindo que todos possam interagir com as informações de maneira eficiente. Essas funcionalidades tornam os SGBDs ferramentas poderosas para empresas e organizações que precisam gerenciar grandes volumes de dados de forma eficaz, assegurando a consistência e a integridade das informações.

2.3 Obtenção de Dados

A obtenção de dados é uma necessidade fundamental em um mundo cada vez mais orientado por informações. Com o crescimento exponencial do volume de dados disponíveis, as organizações enfrentam o desafio de transformar esses dados em informações úteis que possam guiar a tomada de decisões estratégicas. De acordo com Westbrook (1994), a coleta de dados deve ser feita com um propósito claro, buscando não apenas dados brutos, mas também *insights* que agreguem valor à análise. Taylor (1982) complementa essa ideia ao discutir o conceito de informação com valor agregado, enfatizando que os processos de seleção e análise são essenciais para transformar dados em informações úteis.

Atualmente, a necessidade de obter dados se torna ainda mais evidente devido à abundância de informações disponíveis através das tecnologias da informação. Davis e Olson (1989) afirmam que os dados necessitam cada vez mais de um tratamento prático e de bom senso que os transformem em informação pertinente. Essa transformação é crucial para garantir que as informações obtidas sejam relevantes e aplicáveis nas decisões organizacionais. Assim, a obtenção de dados por meio do web scraping não é apenas uma questão técnica; envolve uma abordagem estratégica para garantir que as informações extraídas sejam consistentes e confiáveis.

2.3.1 Interface de Programação de Aplicativos

De acordo com Oliveira e Moura (2015) a Interface de Programação de Aplicativos (API, do inglês *Application Programming Interface*) reúne um conjunto de estruturas de dados, funções e protocolos que permitem que aplicações se comuniquem entre si, sem intervenção humana. Ele afirma que uma das grandes vantagens do uso de APIs é que as particularidades técnicas tanto dos provedores de serviço quanto dos consumidores são abstraídas. Os dados podem ser obtidos por meio de uma API em um processo estruturado. Primeiro, o cliente envia uma solicitação ao servidor da API usando um ponto de extremidade (*endpoint*) específico, que pode incluir parâmetros para definir quais dados são necessários. O servidor então recebe essa solicitação, processa o pedido e busca as informações em bancos de dados ou outros sistemas. Após isso, ele retorna uma resposta ao cliente, que contém os dados solicitados. A aplicação cliente interpreta esses dados e os apresenta ao usuário de forma compreensível.

2.3.2 Web Scraping

De acordo com Glez-Peña *et al.* (2014), web scraping pode ser definido como um processo para obtenção e combinação de dados a partir de rotinas sistematizadas e aplicadas em um ambiente Web. Entretanto, é interessante pontuar que, conforme descrito por Anish Chapagain (2023), as rotinas de web scraping se aplicam a ambientes digitais diversificados como, por exemplo, módulos de comunicação em sistemas de gerenciamento de dados. Técnicas de web scraping são aplicadas em estudos em diferentes áreas (CORDEIRO *et al*., 2022; CORREIA et al., 2023; CORDEIRO, 2024). Os dados são obtidos através de um processo que inclui o envio de uma solicitação ao servidor da plataforma, que então processa a requisição e retorna as informações desejadas. O web scraping depende da extração direta de dados visíveis em ambientes digitais que disponibilizam informações, frequentemente na forma de HTML. O processo geralmente envolve várias etapas: primeiro, uma solicitação é feita ao servidor, que responde com o conteúdo da página ou aplicação. Em seguida, esse conteúdo é analisado para identificar e extrair os dados relevantes, utilizando ferramentas como expressões regulares ou bibliotecas específicas. Além de sites, o web scraping pode ser aplicado em outros ambientes digitais, como plataformas que oferecem dados em formatos estruturados ou em sistemas que não permitem download direto. Nesses casos, técnicas mais avançadas são necessárias para acessar e extrair as informações desejadas

De acordo com o tipo de dados, pode ser interessante que os dados sejam organizados em formatos que aumentem sua eficácia e utilidade no contexto de web scraping. Os diferentes formatos de dados, como por exemplo os mais utilizados: XML, CSV e JSON, têm uma relação significativa com o *web scraping*, em especial na saída de dados. A escolha do formato adequado pode afetar a eficiência da extração, a interpretação dos dados e principalmente a facilidade de uso. Fernandes e Cordeiro (2016, p. 7) afirmam que “O XML (*Extensible Markup Language*) é um formato de arquivo amplamente usado para troca de informações por permitir diversas maneiras de manter a estrutura da informação e a forma que o arquivo é construído”, ele é particularmente útil quando se lida com estruturas complexas, permitindo que os dados sejam organizados de maneira que facilite a integração entre sistemas computacionais, isso é essencial em projetos de scraping onde a estrutura dos dados precisa ser preservada para análise futura. Por outro lado, o CSV oferece uma abordagem mais simples e compacta, ideal para grandes conjuntos de dados tabulados. No entanto, de acordo com Fernandes e Cordeiro (2016, p. 9) “é um formato tão rigoroso que o arquivo pode se tornar inútil caso não possua uma documentação explicando como os dados são separados” e isso pode resultar em erros na interpretação. Já o JSON, Fernandes e Cordeiro (2016, p. 10) descrevem que “é um formato de arquivo aberto e simples que pode ser facilmente lido por qualquer linguagem de programação”, permitindo uma comunicação eficiente entre cliente e servidor. Assim, fica evidente que a escolha do formato de saída é tão importante para o *web scraping* e deve considerar não apenas as características dos dados coletados, mas também as necessidades futuras da análise, garantindo que os dados possam ser utilizados de maneira eficaz.

3 Metodologia

Este trabalho se trata de uma pesquisa aplicada, uma vez que busca explorar de forma prática o uso de *web scraping* na avaliação da qualidade dos dados. A abordagem utilizada é mista, integrando elementos quantitativos e qualitativos: enquanto os dados

numéricos serão apresentados para quantificar a quantidade de artigos revisados, a análise qualitativa permitirá a identificação de significados implícitos nos resultados obtidos. Metodologicamente a pesquisa é bibliográfica, pois se baseia em um levantamento sistemático de literatura especializada que possibilita uma análise crítica e contextualizada de *web scraping*, com foco na identificação da qualidade dos dados coletados por meio dessa técnica. Para o desenvolvimento da pesquisa, considerando que é bibliográfica, foi realizada uma curadoria rigorosa dos resultados encontrados, com o objetivo de verificar a relevância dos trabalhos para a pesquisa. Para isso, foram consideradas diversas estratégias, como a pertinência temática, bases de referência, tipos de trabalhos acadêmicos, equações de busca, qualidade da publicação e a relevância das contribuições para o entendimento do uso de *web scraping* na avaliação da qualidade dos dados.

As bases de dados utilizadas para a busca e seleção dos trabalhos acadêmicos incluíram o Google Acadêmico, a Brapci (Base de Dados em Ciência da Informação), a BDTD (Biblioteca Digital Brasileira de Teses e Dissertações), e o Google. Por definição, "O Google Acadêmico é um subconjunto do índice de pesquisa maior do Google, que consiste em artigos de periódicos em texto completo, relatórios técnicos, preprints, teses, livros e outros documentos, incluindo páginas da Web selecionadas que são consideradas "acadêmicas"." Rita Vine (2006, p.97-99), ele é considerado uma grande base de dados, pois indexa uma ampla gama de trabalhos, a curadoria feita a partir desta fonte visou identificar publicações que se mostrassem pertinentes à temática em questão, ressaltando a importância da busca por citações relevantes em trabalhos reconhecidos. Embora não seja uma base de dados acadêmica tradicional, o Google foi utilizado para a busca de documentos complementares e publicações que pudessem enriquecer a discussão sobre *web scraping* e sua aplicação na qualidade dos dados.

Para a construção do referencial teórico, foram considerados diversos tipos de trabalhos acadêmicos, incluindo artigos, Trabalhos de Conclusão de Curso (TCC), monografias, teses e dissertações. Embora a literatura científica sobre *web scraping* não seja tão ampla, ela fornece percepções valiosas que foram essenciais para a construção deste trabalho, especialmente considerando que muitas dessas publicações são abordadas em TCCs e outros trabalhos acadêmicos. Além disso, foram utilizadas equações de busca que foram fundamentais para identificar a literatura relevante sobre o tema. As equações utilizadas foram: *"web scraping"* e "qualidade"; *"web scraping"* e "ética"; *"web scraping"* e "problemas"; e *"web scraping"* e "segurança". Cada uma dessas equações aborda aspectos que influenciam diretamente a qualidade dos dados obtidos por meio do *web scraping*.

A partir da identificação dos trabalhos relacionados durante a revisão bibliográfica, foi feita uma avaliação temática que permitiu organizar os estudos em categorias de similaridade. Essas categorias foram estabelecidas considerando aspectos como objetivos, metodologia e abordagem principal de cada trabalho. Esse processo teve como objetivo agrupar estudos com características semelhantes, facilitando a análise comparativa e a identificação de padrões relevantes. Como resultado dessa avaliação, se criou um quadro de categorização que busca organizar esses trabalhos em tópicos temáticos específicos, destacando suas contribuições para o entendimento do web scraping e da qualidade dos dados, o que é fundamental para direcionar futuras pesquisas nessa área.

4 RESULTADOS

A partir da obtenção dos trabalhos, foi feita uma observação da similaridade entre os artigos, levando em consideração os aspectos metodológicos apresentados. Com base nessa análise, foi levantado um grupo de 40 artigos, publicados entre 2015 e 2024, que foram verificados em termos de conteúdo e tema. O critério de inclusão foi a relevância para o tema e a aplicação prática de soluções de controle de qualidade em processos de *web scraping*. Os temas abordados incluem, entre outros, Desenvolvimento de software para Web Scraping, Ética em Web Scraping, Coleta de dados com Web Scraping e Questões teóricas/conceituais sobre Web Scraping. Cada categoria reúne os estudos relevantes que abordam diferentes aspectos do web scraping, desde a implementação técnica e as questões éticas até os métodos de coleta de dados e os fundamentos teóricos sobre a prática. O Quadro 1 apresenta a categorização desses artigos, para facilitar a identificação das principais áreas de pesquisa e contribuir para uma melhor compreensão dos tópicos que envolvem a garantia da qualidade dos dados extraídos por meio dessa técnica.

Quadro 1 - Categorização

Item descritivo	Trabalhos relacionados
Desenvolvimento de software para Web Scraping	Farias *et al.*(2021); Soares *et al*. (2022); Graciano; Ramalho; (2021); Castanha (2024); Cavalheiro (2023); Medeiros; Manheze; Cordeiro (2023); Lessa (2022); Lima; Ribeiro (2021); Teixeira *et Al*. (2018); Paulino; Mariano (2022); Trenti (2023); Santos (2015); Presser; Dagnino; Weber (2021)
Ética em Web Scraping	Oliveira *et al.* (2023); Krotov; Redd; Silva (2020); Logos *et al.* (2023); Gonçalves (2021); Munger (2023)
Coleta de dados com Web Scraping	Gheorghe; Mihai; Dârdala (2020); Martins De Paula; Marynowski; Feger (2024); Assis (2021);

	Marta-Lazo; González; Herrero (2021); Santos; Carrijo (2021); Teles; Silva (2021); Mathias (2017); Fagundes; Macedo. Freund (2018); Gualdani *et al.* (2022); Lüdtke Espíndola *et al.* (2018); Piccolo (2018)
Questões teóricas/conceituais sobre Web Scraping	Milev (2017); Khder (2021); Dogucu; Çetinkaya-Rundel (2021); Matiquite *et al.* (2023); Mello; Paiva (2023); Alves (2024); Oliveira (2023); Ioscote (2023); Lopes; Candeia (2023); Benedito (2022)

Fonte: autores.

A categoria Desenvolvimento de software para Web Scraping reúne estudos que abordam a criação e a melhoria de ferramentas e softwares para a extração de dados da web. Muitos desses artigos, como Farias et al. (2021) e Soares et al. (2022), destacam-se pela busca por maior eficiência e automação no processo de scraping, com ênfase em algoritmos mais rápidos e escaláveis. Graciano e Ramalho (2021), discutem a criação de bibliotecas abertas que promovem o acesso e a disseminação de ferramentas para a comunidade. Outros estudos, como Castanha (2024) e Cavalheiro (2023), focam no desenvolvimento de softwares específicos adaptados a nichos de mercado, o que torna essas ferramentas mais personalizáveis. Além disso, Medeiros, Manheze e Cordeiro (2023) e Lessa (2022) exploram como integrar web scraping com outras tecnologias emergentes, como Big Data e Machine Learning, ampliando as possibilidades de uso dessa técnica em diferentes contextos. Em conjunto, esses artigos refletem um crescente aprimoramento das ferramentas de web scraping, com foco na eficiência, personalização e integração com outras tecnologias.

Por outro lado, a categoria Ética em Web Scraping aborda as questões legais e morais envolvidas na prática de extração de dados. Oliveira et al. (2023) e Krotov, Redd e Silva (2020) discutem a regulamentação do uso do web scraping, destacando os desafios jurídicos relacionados ao acesso e à utilização de dados sem o consentimento dos proprietários dos sites. Logos et al. (2023) e Gonçalves (2021) exploram a proteção de dados pessoais e os direitos à privacidade, especialmente à luz de legislações como a LGPD e o GDPR. Já Munger (2023) oferece uma análise crítica sobre os limites éticos do uso do web scraping, questionando o impacto que a raspagem de dados pode ter sobre os indivíduos e as organizações. Esses estudos enfatizam a importância de

práticas éticas, regulamentações e a proteção dos direitos digitais, destacando as questões legais e sociais envolvidas na utilização dessa técnica.

A categoria Coleta de dados com Web Scraping se concentra nas metodologias e desafios relacionados à extração de dados da web. Gheorghe, Mihai e Dârdala (2020) e Martins de Paula, Marynowski e Feger (2024) abordam a aplicação de Web Scraping em contextos comerciais, como o monitoramento de preços e coleta de dados de produtos, destacando a utilização de scraping em larga escala para geração de insights estratégicos. Assis (2021) e Marta-Lazo, González e Herrero (2021) investigam como a técnica pode ser aplicada em pesquisas acadêmicas, como na coleta de informações científicas. Além disso, Santos e Carriço (2021) analisam as vantagens e dificuldades da raspagem em tempo real, quando é necessário atualizar constantemente os dados extraídos. Outros estudos, como Mathias (2017) e Fagundes, Macedo e Freund (2018), discutem a qualidade dos dados, abordando a filtragem e organização das informações extraídas para garantir sua consistência. Esses trabalhos demonstram a versatilidade e os desafios associados à coleta de dados utilizando web scraping, que envolve não apenas aspectos técnicos, mas também a gestão de grandes volumes de dados e a adaptação às mudanças nos sites alvo.

Por fim, a categoria Questões teóricas/conceituais sobre Web Scraping aborda os aspectos conceituais e teóricos que sustentam o uso de Web Scraping. Milev (2017) e Khder (2021) exploram os fundamentos computacionais do Web Scraping, discutindo as abordagens algorítmicas e matemáticas que permitem a extração eficiente de dados. Dogucu e Çetinkaya-Rundel (2021) focam na aplicação da teoria estatística para garantir a precisão dos dados coletados, propondo métodos que combinam análise quantitativa e qualitativa. Outros estudos, como Matiquite et al. (2023) e Mello e Paiva (2023), discutem a evolução do conceito de Web Scraping, esclarecendo o que define essa técnica e como ela se diferencia de outras abordagens de coleta de dados. Alves (2024) também aborda as implicações da teoria dos dados e das metodologias computacionais no desenvolvimento das ferramentas de Web Scraping. Essa categoria destaca a importância de uma compreensão teórica clara para a aplicação eficaz e ética da técnica.

5 Conclusões

A pesquisa realizada sobre web scraping e a qualidade dos dados coletados revelou que, embora essa técnica ofereça uma solução eficaz para a extração de dados, ela enfrenta desafios significativos relacionados à confiabilidade e à integridade dos dados. As análises indicaram que a qualidade dos dados obtidos por meio dessa técnica pode ser comprometida por fatores como a estrutura dinâmica das páginas web, a possibilidade de extração de dados incompletos ou desatualizados, e a introdução de "sujeira" nos dados. Assim, surge a necessidade de maior documentação e padronização no desenvolvimento de web scraping, além de uma abordagem mais detalhada sobre os limites éticos e legais dessa prática.

Diante do exposto, pode-se concluir que o web scraping, quando utilizado de forma planejada e ética, pode ser uma ferramenta essencial para a coleta de dados no contexto da era digital. No entanto, o sucesso de sua aplicação depende de práticas de controle de qualidade, planejamento técnico e respeito às regulamentações vigentes. Os resultados deste trabalho podem contribuir para orientar pesquisas futuras que busquem

o desenvolvimento de abordagens mais eficientes, seguras e éticas para o uso do web scraping.

Referências

ASSIS, Wendel Vilaça de. **Chat Bot Sumé**: Web Scraping em dados governamentais para consulta de gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte. Dissertação apresentada à Universidade FUMEC Faculdade de Ciências Empresariais, Belo Horizonte-MG, 2021.

BERNERS-LEE, Tim. **Five stars for open data**. 2010. Disponível em: https://5stardata.info/en/. Acesso em: 17 dez. 2024.

BRAGA, Claudomilson Fernandes; TUZZO, Simone Antoniaci. Dados abertos à brasileira: aspecto de uma cidadania denegada. **Comunicação & Inovação**, v. 18, n. 37, p. 48–65, 2017. DOI: https://doi.org/10.13037/ci.vol18n37.4359.

CASTANHA, Rafael Gutierres. Visualização de redes de coautoria como insumo bibliométrico às revistas científicas: uma proposta via web scraping para os periódicos Em Questão e Encontros Bibli. **Encontros Bibli**, v. 29, 2024. DOI: https://doi.org/10.5007/1518-2924.2024.e96589.

CHAPAGAIN, Anish. **Hands-on web scraping with Python**: extract quality data using Python techniques. United Kingdom: Packt Publishing, 2023.

COMITÊ GESTOR DA INTERNET NO BRASIL (CGI). **Dados abertos**: capítulo 16. 2024. Disponível em: https://ceweb.br/guias/dados-abertos/capitulo-16/. Acesso em: 17 dez. 2024.

CORDEIRO, D. F. Perspectivas en contraste: análisis comparativo cuantitativo España y Brasil de la cobertura del conflicto israelí-palestino en Google News. **Documentación de las Ciencias de la Información**, v. 47, p. 15–25, 2024. DOI: https://doi.org/10.5209/dcin.92187.

CORDEIRO, D. F.; LOPEZOSA, C.; GUALLAR, J.; VÁLLEZ, M. Análisis de la cobertura de Google news: un estudio comparativo de Brasil, Colombia, México, Portugal y España. **Contratexto**, n. 42, p. 177-208, 2024. DOI: https://doi.org/10.26439/contratexto2024.n42.7212.

CORDEIRO, D. F.; LEAL, M. R. C.; VIEIRA, L. M.; DA SILVA, N. R. Cartografando comentários e sentimentos no perfil de Jair Bolsonaro no Instagram acerca da Covid-19. **Galáxia**, v. 47, e56929, 2022.

CORREIA, G. P.; MENDONÇA, R. F.; BORGES, R. M. R.; CORDEIRO, D. F. Saúde mental no Twitter: análise de manifestações por meio de mineração de dados. **Novos Olhares**, v. 12, n. 2, p. 116–130, 2023. https://doi.org/10.11606/issn.2238-7714.no.2023.210801.

DAVENPORT, Thomas H. **Ecologia da informação**: porque só tecnologia não basta para o sucesso na era da informação. 4. ed. São Paulo: Futura, 1998.

DAVIS, Gordon B.; OLSON, Margrethe H. **Management information systems**: conceptual foundations, structure, and development. 2. ed. New York: McGraw-Hill, 1985.

DE PAULA, João Pedro Martins; MARYNOWSKI, João Eugenio; FEGER, José Elmar. Tourist data collection with web scraping: problems, solutions and optimizations. **Revista Brasileira de Sistemas de Informação**, v. 17, n. 1, p. 8:1-8, 2024. DOI: https://doi.org/10.5753/isys.2024.3644.

DOGUCU, Mine; ÇETINKAYA-RUNDEL, Mine. Web scraping in the statistics and data science curriculum: challenges and opportunities. **Journal of Statistics and Data Science Education**, v. 29, n. S1, p. 112-S122, 2020. Disponível em: https://doi.org/10.1080/10691898.2020.1787116.

DOS SANTOS, Vitor C. **Bem-Falado**: Um Sistema de Coleta e Análise de Avaliações de Produtos com Técnicas de Web Scraping. Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação. Porto Alegre: Universidade Federal do Rio Grande do Sul, Instituto de Informática, 2023.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 6. ed. São Paulo: Pearson Education, 2011.

ESPÍNDOLA, Priscilla Lüdtke; SALM JUNIOR, José Francisco; ROSA, Francisco; JULIANI, Jordan Paulesky. **RDBCI: Revista Digital Biblioteconomia e Ciência da Informação**, v. 16, n. 3, p. 274-298, 2018. DOI: https://doi.org/10.20396/rdbci.v16i3.8651080.

FAGUNDES, Priscila Basto; MACEDO, Douglas Dyllon Jeronimo de; FREUND, Gislaine Parra. A produção científica sobre qualidade de dados em big data: um estudo na base de dados Web of Science. **RDBCI: Revista Digital Biblioteconomia e Ciência da Informação,** v. 16, n. 1, p. 194-210, 2018. DOI: 10.20396/rdbci.v16i1.8650412.

FARIAS, Jorge Luiz F.; CORDEIRO, Douglas F. Avaliação de formatos de publicação de dados abertos governamentais através de indicadores de usabilidade. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n. 1, 2016.

FARIAS, Marcello Tenorio de; ANGELUCI, Alan César Belo; PASSARELLI, Brasilina. Web scraping e ciência de dados na pesquisa aplicada em comunicação: um estudo sobre avaliações online. **Revista Observatório**, v. 7, n. 3, p. 1-21, 2021. Disponível em: http://dx.doi.org/10.20873/uft.2447-4266.2021v7n3a1pt.

FERREIRA, Aurélio Buarque de Holanda. **Dicionário Eletrônico Aurélio Século XXI**. Rio de Janeiro: Editora Nova Fronteira e Lexikon Informática, 1999.

GHEORGHE, Mihai; MIHAI, Florin-Cristian; DÂRDALĂ, Marian. Modern techniques of web scraping for data scientists. **Revista Romana de Interactiune Om-Calculator**, v. 11, n. 1, p. 63-75, 2018.

GLEZ-PEÑA, Daniel; LOURENÇO, Anália; LÓPEZ-FERNÁNDEZ, Hugo; REBOIRO-JATO, Miguel; FDEZ-RIVEROLA, Florentino. Web scraping technologies in an API world. **Briefings in Bioinformatics**, v. 15, n. 5, p. 788-797, 2014. Disponível em: https://doi.org/10.1093/bib/bbt026.

GRACIANO, Helton Luiz dos Santos. **ScraperCI**: um protótipo de Web scraper para coleta de dados. Dissertação (Mestrado) — Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil, 2022. DOI: https://doi.org/10.5007/1518-2924.2023.e92471.

GUALDANI, Fabrício Amadeu; MARCHI, Késsia Rita da Costa; ALVES, Fábio Henrique; BOTEGA, Leonardo Castro. Critérios de qualidade de dados em saúde: uma

análise quantitativa. **Informação & Informação**, v. 27, n. 2, p. 466-490, 2022. DOI: https://doi.org/10.5433/1981-8920.2022v27n2p466.

HOSHOVSKY, Alexander G. MASSEY, Robert J. Information Science: its ends, means & opportunities. In: PLATAU, Gerard O. (Ed.). **Information transfer**. Proceedings of the Annual Meeting of the ASIS, 1968, October, 20-24. Columbus: Ohio, DC: ASIS, 1968. v.5 p.47-55.

IOSCOTE, Fabia Cristiane. Jornalismo e IA: tendências globais e latino-americanas em artigos científicos (2018-2022). **Revista Estudos em Jornalismo e Mídia**, v. 20 n. 2, p. 06-21, 2023. DOI: https://doi.org/10.5007/1984-6924.2023.95335.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados Abertos Conectados**: Em busca da Web do Conhecimento. São Paulo: Novatec Editora, 2015. Capítulo 2, p. 42.

KHDER, Moaiad A. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. **Int. J. Advance Soft Compu. Appl**., v. 13, n. 3, p. 144-168, 2021. DOI: https://doi.org/10.15849/IJASCA.211128.11.

KROTOFF, Vlad; JOHNSON, Leigh; SILVA, Leiser. Tutorial: Legality and Ethics of Web Scraping. **Communications of the Association for Information Systems**, v. 47, n. 24, p. 540-563, 2020. DOI: https://doi.org/10.17705/1CAIS.04724.

LESSA, Marcos Aurélio. **CrawlEX**: uma ferramenta para extração de dados na web configurável através de exemplos. Trabalho de Conclusão de Curso, Universidade Federal de Santa Catarina - UFSC, Florianópolis– SC, 2022.

LIMA, Ágata Raiza; RIBEIRO, Renato. **Comparador de Preços – Web Crawler**. Trabalho de Conclusão de Curso — Universidade Presbiteriana Mackenzie, São Paulo – SP, Brasil, 2021.

LOGOS, Katie; BREWER, Russell; LANGOS, Colette; WESTLAKE, Bryce. Establishing a framework for the ethical and legal use of web scrapers by cybercrime and cybersecurity researchers: learnings from a systematic review of Australian research. **International Journal of Law and Information Technology**, v. 31, n. 3, p. 186–212, 2023. DOI: https://doi.org/10.1093/ijlit/eaad023.

LOPES, Camilo da Mota; CANDEIA, Allan Stieg. Automatização Robótica de Processos (RPA) no Desenvolvimento e Criação de Empresas. **UNESC em Revista**, v. 7, n. 1, p. 52-66, 2023. DOI: https://doi.org/10.54578/unesc.v7i1.387.

MATHIAS, Gilney Nathanael. **qFex: um crawler para busca e extração de questionários de pesquisa em documentos HTML**. Trabalho de Conclusão de Curso — Universidade Federal de Santa Catarina, Florianópolis, SC, 2017.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data: The Management Revolution. Harvard Business Review, 2012. Disponível em: https://hbr.org/2012/10/big-data-the-management-revolution. Acesso em: 17 dez. 2024.

MCGARRY, K. J. **Da documentação à informação**: um conceito em evolução. Lisboa: Editorial Presença, 1984.

MEDEIROS, Jaqueline Souza; MANHEZE, Vitória Nery Lopes; CORDEIRO, Douglas Farias. Ciência de dados aplicada à extração e análise do programa Nota Fiscal Goiana. **Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica**, v. 13, n. 2, p. 109-120, 2023.

MELO, Katia Isabelli. B. B.; PAIVA, Douglas. Web documentação: uma experiência da base de dados na construção de um conjunto de documentos interligados. **Revista Fontes Documentais**, v. 6, n. 1, p. 134-150, 2023.

MILEV, Plamen. Conceptual Approach for Development of Web Scraping Application for Tracking Information. **Economic Alternatives**, v. 3, p. 475-485, 2017.

MUNGER, Kevin. Temporal validity as meta-science. **Research & Politics**, v. 10, n. 3, p. 1-10, 2023. DOI: https://doi.org/10.1177/20531680231187271.

NETO, Manoel Benedito. **Apache Kafka**: Implementação da Técnica de Replicação de Banco de Dados Baseada em Middleware para o Contexto de Raspagem de Dados. Monografia (graduação) – Universidade Federal do Rio Grande do Norte, Natal, RN, 2022.

OLIVEIRA, Alyfer Ricardo Sousa de; CHIQUINI, Andressa de Lima; LIMA, Thais Saturnino de. **Web Scraping Aplicado à Cibersegurança**. Trabalho de conclusão de curso, Faculdade de Tecnologia de Jundiaí - FATEC, Jundiaí-SP, 2023.

OLIVEIRA, Cláudio de; MOURA, Samuel Pedrosa. TIC's na educação: a utilização das tecnologias da informação e comunicação na aprendizagem do aluno. **Pedagogia em Ação**, v. 7, n. 1, p. 75-94, 2015.

OLIVEIRA, Jéssica Sousa. **Web scraping na extração e combinação sistemática de conteúdos**: ferramenta auxiliar em processos de pesquisa, desenvolvimento e inovação (PD&I). Dissertação apresentada à Universidade de Brasília – UNB Campus Gama, Brasília/DF, 2023.

OPEN KNOWLEDGE FOUNDATION. **What is Open**? Disponível em: https://okfn.org/en/library/what-is-open/. Acesso em: 17 dez. 2024.

PAULINO, Daniele; MARIANO, Matheus. **RPA e Python**: Um processo Otimizado de Webscraping. Trabalho de Conclusão de Curso – Faculdade de Tecnologia de São José do Rio Preto, São José do Rio Preto, 2022.

PEREIRA, Maria João Gonçalves. **Análise e Mecanismos de Prevenção de Web Scraping**. Dissertação apresentada à Faculdade de Ciências da Universidade do Porto, Porto - Portugal, 2021.

PICCOLO, Daiane Marcela. Qualidade de dados dos sistemas de informação do DATASUS: análise crítica da literatura. **Ciência da Informação em Revista**, v. 5, n. 3, p. 13-19, 2018. DOI: https://doi.org/10.28998/cirev.%25y513-19.

PRESSER, Martim Kowalczuk; DAGNINO, Ricardo de Sampaio; WEBER, Eliseu José. Automatização da aquisição de dados da COVID-19 por web scraping e atualização de mapas ArcGIS Online utilizando Python. **Anais da 10ª Mostra de Ensino, Extensão e Pesquisa**, IFRS Campus Osório, Rio Grande do Sul, 2021.

SANTOS, Eliseu Xavier Bernardo; CARRIJO, Pedro Felipe de Moraes. **Comparador de preços do mercado de skins do CS GO**. Trabalho de Conclusão de Curso, UniEVANGÉLICA, Anápolis - GO, 2021.

SANTOS JR, Marcelo Alves Dos. Estudo exploratório do financiamento da desinformação na web: fraudes, apostas, trading e clickbaits. **Contracampo**, v. 43, n. 1, p. 1-18, 2024. https://doi.org/10.22409/contracampo.v43i1.56987.

SANTOS, Márcio Carneiro dos. Métodos digitais e a memória acessada por APIs: desenvolvimento de ferramenta para extração de dados de portais jornalísticos a partir

da WayBack Machine. **Revista Observatório**, v. 1, n. 2, p. 207-228, 2015. DOI: https://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23.

SEMELER, Alexandre Ribas; OLIVEIRA, Arthur Longoni; PEREIRA, Fabiana Andrade; MATIQUITE, Policarpo. Python scripts para o web scraping de metadados das descrições sobre os conjuntos de dados do cenário internacional de repositórios de dados de pesquisa. **Encontros Bibli**, v. 28, p. 1–8, 2023. DOI: https://doi.org/10.5007/1518-2924.2023.e94877.

SOARES, Renan Moreira; CAMARGO, Guilherme Rezende Pereira; OLIVEIRA, Marcelo Escobar de; MARQUES, Leonardo Garcia. Extração de dados via web scraping como suporte em análises envolvendo a geração distribuída. **Sociedade Brasileira de Automática (SBA)**, v. 2, n. 1, p. 398-404, 2022. DOI: https://doi.org/10.20906/sbse.v2i1.2932.

TAYLOR, Robert S. Value-Added Processes in Information Systems. **Journal of the American Society for Information Science**, v. 33, n. 5, p. 341-346, 1982. DOI: https://doi.org/10.1002/asi.4630330517.

TEIXEIRA, Marcelo Augusto Muniz; LOBATO, Fábio Manoel França; CHAGAS, Beatriz Nery Rodrigues; JACOB JUNIOR, Antonio Fernando Lavareda. Um Sistema de Aquisição e Análise de Dados para Extração de Conhecimento da Plataforma Ebit. **ResearchGate** [preprint], 2018. https://doi.org/10.13140/RG.2.2.33015.32162.

TELES, Caio Matheus; SILVA, Lucas Alves da. **Jambo**: coleta de dados com Web scraping. Trabalho de Conclusão de Curso, Centro Paula Souza – FATEC Faculdade de Tecnologia de Americana - SP, 2021.

VIEIRA, L. M.; CORDEIRO, D. F. The dark side of anti-vaccination: analysis of a brazilian anti-vaccine Facebook group. Famecos, v. 31, n. 1, e43710, 2023. DOI: https://doi.org/10.15448/1980-3729.2023.1.43710.

VINE, Rita. Google Scholar. **Journal of the Medical Library Association**, v. 94, n. 1, p. 97-99, 2006.

WESTBROOK, Lynn. Qualitative research methods: A review of major stages, data analysis techniques, and quality controls. **Library & Information Science Research**, v. 16, n. 3, p. 241-254, 1994. DOI: https://doi.org/10.1016/0740-8188(94)90026-4.