Encontros Bibli

SCRAPERCI: UM *WEB SCRAPER* PARA COLETA DE DADOS CIENTÍFICOS[1]

ScraperCI: a web scraper for scientific data collection

Helton Luiz dos Santos Graciano
Universidade Federal de São Carlos (UFSCar)
São Carlos, SP, Brasil
helton.graciano@estudante.ufscar.com
https://orcid.org/0000-0001-5372-7631

Rogério Aparecido Sá Ramalho
Universidade Federal de São Carlos (UFSCar)
São Carlos, SP, Brasil
ramalho@ufscar.br
https://orcid.org/0000-0002-8491-3514

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: O desenvolvimento tecnológico das últimas décadas tem impulsionado a produção massiva de recursos informacionais e mudanças significativas nos processos de coleta e gestão de dados em praticamente todas as áreas. Tal cenário não é diferente no âmbito científico, onde a coleta e tratamento adequado de dados tem se apresentado como um desafio para pesquisadores. A presente pesquisa teve como objetivo apresentar um protótipo de *Web scraper*, denominado como *ScraperCI*, e analisar as potencialidades da utilização de ferramentas computacionais como esta para a coleta em bases de dados disponíveis na *Web*.
Método: A pesquisa caracteriza-se como aplicada, de natureza exploratória e descritiva, com abordagem qualitativa que visa identificar as potencialidades da utilização de Web scrapers no processo de coleta de dados.
Resultado: Conclui-se que o protótipo desenvolvido possibilita avanços consideráveis no processo de automação da coleta de dados científicos e que tais ferramentas possibilitam a automatização de processos de recuperação, favorecendo maior produtividade no que tange a extração de recursos informacionais na Web.
Conclusões: Espera-se que esta pesquisa possa estimular os profissionais da informação a desenvolver novas competências e enxergar possibilidades inovadoras em suas áreas de atuação profissional, atuando com protagonismo nesse meio interdisciplinar.
PALAVRAS-CHAVE: Recuperação da informação. *Web scraping*. Mecanismos de busca. Gestão de dados.

ABSTRACT

Objective: The technological development of the last few decades has driven the massive production of informational resources and significant changes in data collection and management processes in practically all areas. This scenario is no different in the scientific field, where the collection and proper treatment of data has been a challenge for researchers. This research aimed to present a prototype of Web scraper, called ScraperCI, and to analyze the potential of using computational tools as it is for collection in databases available on the Web.
Methods: The research is characterized as applied, exploratory and descriptive in nature, with a qualitative approach that aims to identify the potential of using Web scrapers in the data collection process.
Results: It is concluded that the developed prototype enables considerable advances in the process of automating the collection of scientific data and that such tools enable the automation of retrieval processes, favoring greater productivity in terms of the extraction of informational resources on the Web.
Conclusions: It is hoped that this research can encourage information professionals to develop new skills and see innovative possibilities in their areas of professional activity, acting with protagonism in this interdisciplinary environment.
KEYWORDS: Information recovery. Web scraping. Search engines. Data management.

[1] A presente pesquisa foi realizada com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

1 INTRODUÇÃO

O desenvolvimento tecnológico vivenciado nas últimas décadas, a popularização da *Web* e a produção massiva e exponencial de recursos informacionais têm proporcionado mudanças significativas na forma como lidamos com os dados. Dessa maneira, a coleta de grandes volumes de dados necessita de novas e criativas soluções e a Ciência da Informação pode desempenhar um papel fundamental a partir do direcionamento de princípios teóricos e métodos para esse processo (SANT'ANA, 2016), uma vez que pode ser definida como uma área que se ocupa em estudar as propriedades e o comportamento da informação, seus fluxos e técnicas empregadas para o processo de armazenamento, recuperação e disseminação (BORKO, 1968).

Se por um lado temos um volume cada vez maior de dados aumentando vertiginosamente, do outro temos um tempo cada vez menor para transformá-los em informações úteis, no qual passa-se muito mais tempo coletando dados do que analisando. De acordo com a pesquisa denominada "*The State of Data Discovery and Cataloging*", os profissionais da informação gastam em média 50% de seu tempo em pesquisas e atividades redundantes, sendo 30% em atividades de pesquisa e 20% elaborando ativos informacionais existentes que poderiam ser reaproveitados (IDC, 2018).

Probstein (2019) destaca que, apesar dos avanços tecnológicos desenvolvidos nas últimas décadas para lidar com dados e informações, os usuários gastam mais tempo para recuperar informações existentes do que analisando e gerando novos conhecimentos.

O processo de recuperação tem impacto direto na eficiência da produção de novos recursos informacionais e otimizar esse processo é fundamental para que se possa empreender mais tempo nos trabalhos analíticos, que são os que norteiam as tomadas de decisão e de fato agregam valor.

Para este trabalho, considera-se recuperação como o processo, ou método, pelo qual um usuário em potencial é capaz de converter sua necessidade informacional em uma lista real de citações de documentos armazenados, contendo informações úteis para ele (MOOERS, 1951).

Nessa perspectiva, esta pesquisa tem como objetivo geral apresentar uma análise das potencialidades da utilização de *Web scrapers* no processo de coleta de dados em *sites* da *web* e, para atendê-lo buscou: a) descrever o processo de recuperação de dados e b) descrever as etapas da coleta de dados a partir de um protótipo de *Web scraper*.

Um *scraper* é um *software* usado para coletar dados de fontes direcionadas da *Web*. Em um nível fundamental, ele pode ser visto como um robô que imita as funções de um ser humano, interagindo com *sites* e coletando dados armazenados neles (UPADHYAY *et al*., 2017).

No aspecto da *práxis* e das competências exercidas pelos profissionais da informação no uso das aplicações tecnológicas emergentes, Souza, Almeida e Baracho (2013, p. 171) destacam que a interdisciplinaridade do campo da Ciência da Informação "[...] exorta o cientista da informação a navegar nos espaços teóricos, adaptar-se aos contextos tecnológicos e reinventar-se continuamente [...] ou assim deveríamos ser".

Dado a influência e o protagonismo que se espera dos profissionais da informação nesse meio interdisciplinar, como agente central de toda essa cadeia, faz-se necessário que esse profissional esteja aberto a se aproximar, entender e aplicar cada vez mais métodos inovadores de coleta, recuperação e análise de dados, em um contexto onde a velocidade e eficiência são exponencialmente demandados a cada dia.

Assim, a presente pesquisa justifica-se uma vez que busca incentivar os profissionais da Ciência da Informação a ampliarem suas perspectivas e explorarem novas possibilidades de atuação, por meio do uso de ferramentas modernas e eficientes para o resgate e análise de informações. Acreditamos que essa abordagem pode proporcionar um diferencial competitivo para os profissionais da área, ao mesmo tempo em que contribui para o desenvolvimento e aprimoramento contínuo desta importante profissão.

2 PROCEDIMENTOS METOLÓGICOS

A presente pesquisa caracteriza-se como uma pesquisa aplicada de natureza qualitativa, desenvolvida a partir de uma abordagem descritiva, com o objetivo de apresentar as potencialidades da utilização de *Web scrapers* no processo de recuperação, de modo que seus resultados possam "[...] gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos" (SILVEIRA; CÓRDOVA, 2009, p.35).

Como delimitação, a pesquisa tem como foco principal o escopo da coleta dos dados, sob a perspectiva dos usuários, de modo que se busca demonstrar a eficácia e representatividade dos dados coletados em um repositório e a partir "[...] das necessidades informacionais, passando pelo planejamento das ações, localização das fontes e culminando no acesso ao conteúdo desejado" (SANT'ANA, 2019, p. 119).

Para a realização da pesquisa foi desenvolvido um protótipo de *Web scraper,* implementado a partir da linguagem de programação *Python* e como fonte de coleta de dados utilizou-se o portal Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (BRAPCI), por ser um dos mais utilizados no âmbito de pesquisas nacionais na área de Ciência da Informação e pela quantidade de artigos indexados no formato *Open Source*.

Para desenvolver o *Web scraper*, realizou-se um levantamento bibliográfico e documental em bases de dados, livros, periódicos e sites em português e inglês sobre os termos *Web scraping*, busca automática da informação, mecanismos de busca, coleta, extração e análise de dados na *Web* e profissional da informação. Em seguida, delimitou-se os dados selecionados e realizou-se uma leitura analítica dos artigos selecionados referentes ao tema proposto, avaliando sua pertinência com a área de Ciência da Informação.

Buscando favorecer uma melhor compreensão do processo de recuperação, e descrever as etapas da coleta, foi utilizado, como exemplo, o termo “recuperação da informação”, considerando o período de 2002 a 2022, e posteriormente os dados coletados foram estruturados em um arquivo *Comma Separated Values* (CSV) para favorecer a análise e apresentação dos resultados.

3 COLETA DE DADOS NA WEB

Compreender os desafios e contribuições da aplicação de um *Web scraper* na coleta de dados é de suma importância para que se possa construir uma ferramenta capaz de recuperar não só quantidade, mas também conteúdos com qualidade. Para tanto, faz-se necessário uma abordagem sobre o processo de Recuperação da Informação (RI), os mecanismos de busca na *Web*, diferenças entre *crawlers e scrapers*, assim como características intrínsecas da estrutura *Web,* como sua capacidade semântica e distribuição em camadas.

Nesse cenário, os processos relativos ao controle dos ativos informacionais estão diretamente conectados à necessidade de transformá-los em conhecimentos que, em última instância, darão suporte para solução de demandas dos usuários que, no contexto de um sistema de RI, estão mais interessados em recuperar informações sobre determinado assunto do que em recuperar dados que satisfazem sua expressão de busca.

O processo de recuperação de dados consiste na identificação de quais documentos contêm as palavras-chave da consulta do usuário, fazendo com que, nem sempre, o resultado seja suficiente para satisfazer sua necessidade informacional (BAEZA-YATES; RIBEIRO-NETO, 2013).

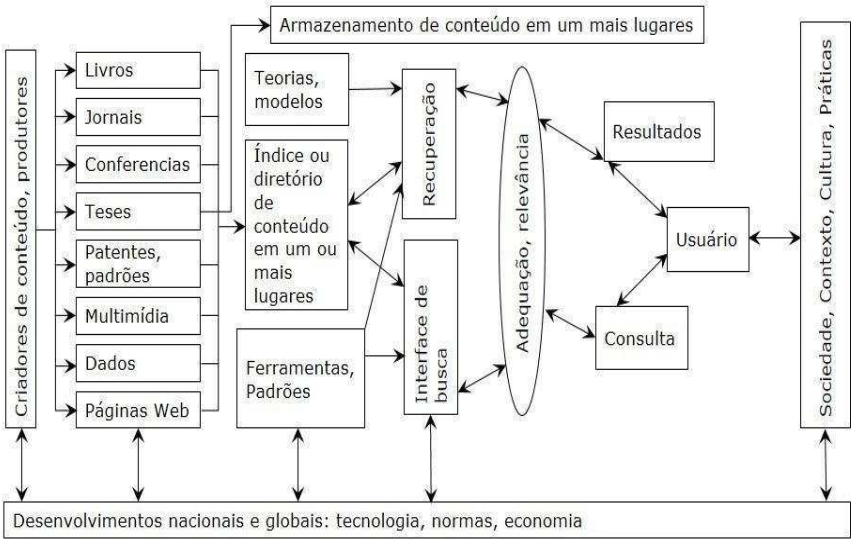
Baeza-Yates e Ribeiro-Neto (2013), destacam que para ser efetivo em sua tentativa de satisfazer a necessidade informacional do usuário, um Sistema de Recuperação de Informação (SRI) deve de alguma forma "interpretar" o conteúdo dos itens de informação, envolvendo a extração de informações sintáticas e semânticas dos textos, isto é, dos documentos de uma coleção e classificá-los de acordo com o grau de relevância à consulta do usuário. Ainda segundo os autores, a dificuldade não só está em saber como extrair a informação dos documentos, mas também como utilizá-la para decidir quanto à sua relevância, a qual exerce um papel central (BAEZA-YATES; RIBEIRO-NETO, 2013).

Nesse contexto, a relevância de um documento é subjetiva e inerente ao julgamento do usuário, estando sujeita à interação do mesmo com o sistema e, sobretudo, ao que de fato ele espera recuperar em sua busca (SILVA; SANTOS; FERNEDA, 2013). Assim, o usuário especifica uma consulta que reflete sua necessidade de informação e a consulta é analisada sintaticamente e expandida com, por exemplo, variações das palavras da consulta.

Segundo Chowdhury (2010, p. 5, tradução nossa), as informações são geralmente recuperadas, "[...] na forma de documentos que contêm as informações necessárias, sempre que os termos da pesquisa correspondem aos termos do índice".

Para descrever um processo de recuperação de informações, foi escolhido o diagrama proposto por Chowdhury (2010), apresentado na Figura 1, que traz a visão conceitual de um SRI e as etapas que permeiam o processo como um todo.

Figura 1 – Extrato de um Sistema de Recuperação de Informação.



Fonte: adaptado de Chowdhury (2010, p. 4).

O sistema SRI apresentado na Figura 1 contempla vários tipos de documentos e recursos de multimídia, sendo que os processos de buscas e recuperação de dados são influenciados pelos conceitos de adequação e relevância. No entanto, por vezes os usuários não conseguem expressar suas necessidades de informações na forma de consultas e não podem passá-las para o sistema de pesquisa por meio de declarações de pesquisa apropriadas.

Devido à grande quantidade de informações disponíveis na *Web*, bem como o número de novos usuários inexperientes em buscas por informações, os mecanismos de pesquisa automatizados dependem da correspondência de palavras-chave e geralmente retornam muitas correspondências de baixa qualidade (BRIN; PAGE, 1998).

Brin e Page (1998) destacam que, a *Web* cria novos desafios para a recuperação de informações, sendo a criação de um mecanismo de busca que dimensione os eventos algo complexo, tornando certas tarefas cada vez mais difíceis à medida que a *Web* se desenvolve.

Para Upadhyay *et al*. (2017) ao realizar uma consulta em um mecanismo de pesquisa, um usuário, frequentemente, examina de três a quatro *links* principais para satisfazer seus requisitos de informação. Destaca-se que, o ato de enviar consultas manualmente e agrupar os dados é um processo complexo e uma solução automatizada seria bem-vinda.

Tal percepção é corroborada por Brin e Page (1998, p. 116) ao afirmarem que,

> [...] o maior problema que os usuários de mecanismos de busca na Web enfrentam [...] é a qualidade dos resultados que obtêm. Embora os resultados sejam geralmente divertidos e expandam os horizontes dos usuários, eles geralmente são frustrantes e consomem um tempo precioso.

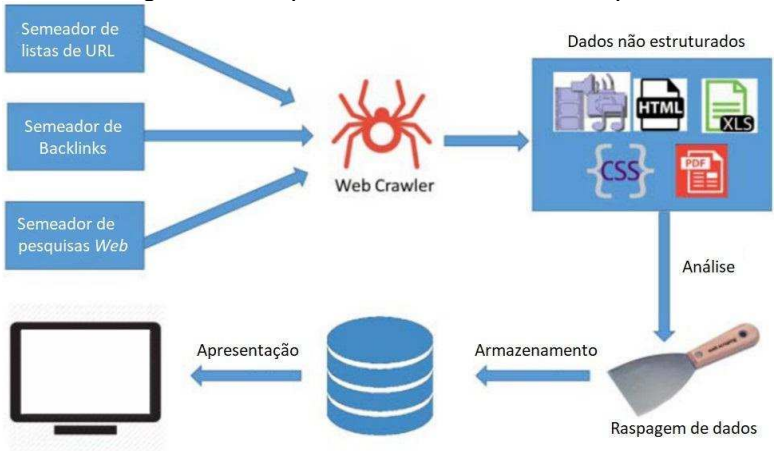
A técnica de *Web crawling*, segundo os autores, conhecida pelo uso de robôs, é empregada para indexar as informações em *sites*. Tal técnica de coleta de dados é essencialmente utilizada nos mecanismos de pesquisa como *Google, Bing, Yahoo*, agências estatísticas e grandes agregadores *online*.

O rastreamento de conteúdo em um *site* a partir dessa técnica ocorre quando um robô passa por todas as páginas e *links* até a última linha do *site*, procurando determinados dados. Em síntese, esse processo consiste em visualizar uma página de um *site* como um todo e indexá-la, geralmente capturando informações genéricas.

Aplicações tecnológicas com grande capacidade de coleta de dados como o *Web scraping* é uma forma de mineração de dados e o objetivo geral do processo de *scraping* é coletar dados e informações de *sites* e transformá-las em uma estrutura compreensível, como planilhas, banco de dados ou um arquivo *Comma Separated Values* (CSV) ou valores separados por vírgula (DASTIDAR; BANERJEE; SENGUPTA, 2016; SIRISURIYA, 2015).

Apresenta-se na Figura 2 uma arquitetura de um *Web scraper* no qual as palavras-chave são executadas em um mecanismo de pesquisa e, com base nas configurações de parâmetros, são produzidos cerca de oito a dez *links* da *Web* por palavra-chave. Os *links* são então enviados para o "Semeador de listas de URL", que usa um *Web crawler* para extrair o conteúdo dos *sites* visitados. O conteúdo coletado é então passado para um *scraper* e enviado na forma de arquivos de texto para uma estação de trabalho (UPADHYAY *et al.*, 2017).

Figura 2 - Arquitetura de um *Web scraper*



Fonte: adaptado de Upadhyay *et al.* (2017, p. 3)

Em um mundo orientado por dados, a técnica de *Web scraping* oferece uma abordagem inovadora para coleta de dados na *Web* e utilizá-los em um grande número de aplicativos de Ciência de Dados. Tal estrutura é geralmente disposta em páginas *Hypertext Markup Language* (HTML) e revela uma boa parte da intenção semântica, podendo ser usada em aplicativos de análise de dados.

O uso de tais tecnologias só é possível devido a *Web* atual incluir propriedades semânticas, sendo constituída por uma estrutura fundamentada em camadas e essa característica é responsável por possibilitar a aplicação das técnicas de raspagem de dados.

Com o intuito de proporcionar uma melhor compreensão e evitar o direcionamento do foco para pormenores técnicos, de maneira geral, a *Web* Semântica é composta pelas seguintes camadas: interface, lógica, semântica, sintática e estrutural. Segundo Ramalho e Ouchi (2011, p. 70), "[...] espera-se que a partir da camada de interface sejam desenvolvidos aplicativos que favoreçam a utilização das novas possibilidades oferecidas pelas Tecnologias Semânticas".

Dessa maneira, demonstra-se no tópico seguinte um protótipo de *Web scraping*, no qual se pode verificar a potencialização das atividades desempenhadas pelos profissionais da informação e o usuário de ambientes digitais na coleta de dados, em um portal da *Web*.

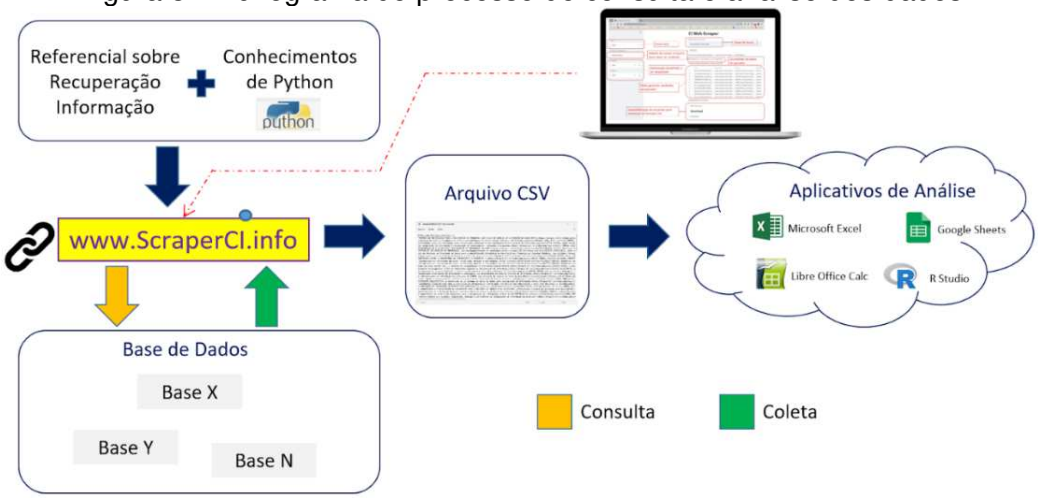
4 *SCRAPERCI* COMO PROTÓTIPO DE WEB SCRAPER PARA COLETA DE DADOS

Para atingir os objetivos propostos foi desenvolvido um protótipo de *Web scraper* denominado como *ScraperCI*, e disponibilizado publicamente a partir do endereço http://scraperci.info. Este protótipo, concebido com fins estritamente didáticos, foi proposto com o objetivo de favorecer uma maior compreensão prática das possibilidades oferecidas pela utilização em mecanismos de busca e coleta de dados na *Web*.

Simulando necessidades reais das organizações no processo de coleta de dados, o *Web scraper* foi desenvolvido na linguagem de programação *Python* (MITCHELL, 2018), propiciando a coleta de dados em bases de dados, em períodos de tempo curtos, quando comparados a buscas manuais realizadas nos *sites* eletrônicos usados nesta pesquisa. Ressalta-se que, nesta pesquisa optou-se por usar o *Python* devido sua vasta utilização no meio corporativo e acadêmico, mesmo em áreas do conhecimento não diretamente conectadas a informática.

Na Figura 3, apresenta-se o fluxo desenvolvido para o processo proposto o qual permite realizar consultas e, consequentemente, após a coleta e análise dos dados extrair as conclusões do conteúdo pesquisado.

Figura 3 - Fluxograma do processo de consulta e análise dos dados



Fonte: elaborado pelos autores.

Uma das preocupações que o profissional da informação deve ponderar, no exercício de suas atividades, é com a disseminação do conhecimento, bem como das ferramentas que podem contribuir para tal. Atualmente, a forma mais viável de se atingir esse objetivo é utilizar a própria *Web* para compartilhar conteúdos nos mais diversos formatos e tornar as ferramentas desenvolvidas compatíveis com a plataforma, podendo ser executadas em navegadores de computadores e dispositivos móveis.

Com esse intuito, o *Web scraper* desenvolvido foi disponibilizado no ambiente Web utilizando-se o *framework*[2] *Streamlit* para criar a interface gráfica Web e a plataforma *Heroku*, para hospedar a aplicação.

O *Streamlit* é um framework de código aberto que permite tornar interativo seu projeto de dados, transformando seu código *Python* em uma aplicação *Web* compartilhável e gratuita. Foi desenvolvido exatamente para ajudar cientistas de dados a colocarem em produção seus projetos sem a necessidade do conhecimento de ferramentas de *front-end*[3] ou de *deploy*[4] de aplicações.

Por meio desse framework é possível transformar um projeto de ciência de dados em uma aplicação interativa. Para essa aplicação é gerada uma URL pública que, ao ser

[2] Um framework em desenvolvimento de software, é uma abstração que une códigos comuns entre vários projetos de software provendo uma funcionalidade genérica.

[3] Front-End é tudo que envolve a parte visível de um site ou aplicação, com a qual os usuários podem interagir.

[4] Fazer um deploy, em termos práticos, significa colocar no ar alguma aplicação que teve seu desenvolvimento concluído.

compartilhada, permite que qualquer pessoa consiga acessar e usufruir sem necessariamente ter que conhecer o código que está por trás.

Considerando tais características do *Streamlit*, essa ferramenta se torna uma excelente forma de apresentar projetos técnicos para pessoas que são leigas na área, além de deixar a apresentação com uma aparência muito profissional.

Atualmente, a principal maneira de entrega e uso de uma aplicação é por meio da nuvem. É por lá que as aplicações são armazenadas e acessadas. Sendo assim, é imprescindível que uma aplicação seja armazenada em algum servidor e disponível pela internet.

O *Heroku* é uma plataforma de nuvem como serviço (PaaS) que suporta várias linguagens de programação. Uma das primeiras plataformas em nuvem, o *Heroku* está em desenvolvimento desde junho de 2007, quando suportava apenas a linguagem de programação *Ruby*, mas agora suporta *Java, Node.js, Scala, Clojure, Python*, PHP e Go.

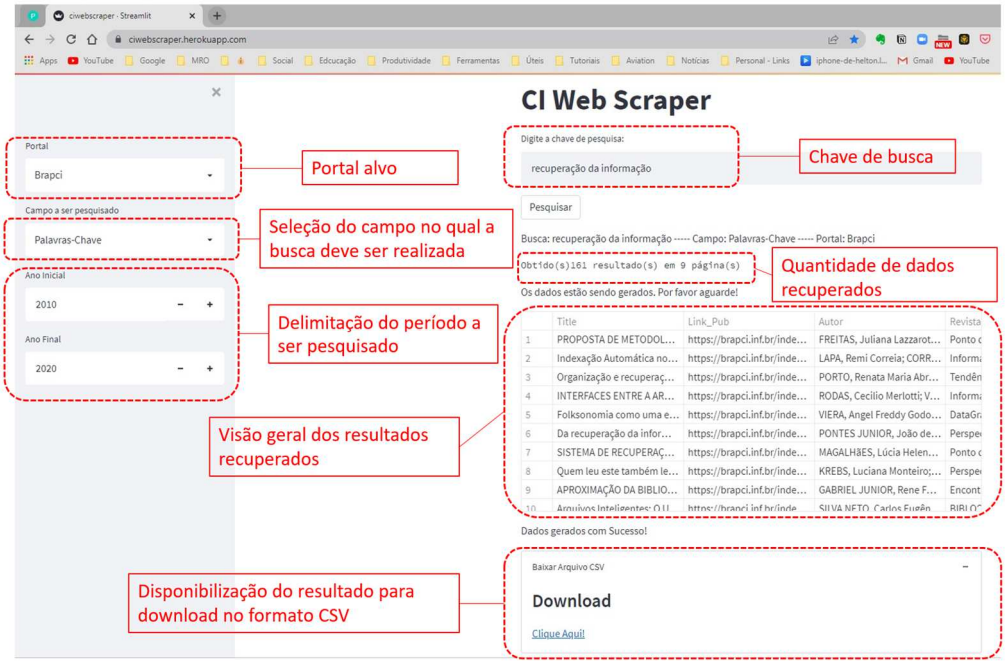
Para os desenvolvedores que querem se preocupar cada vez menos com infraestrutura e processos de *deploy*, concentrando-se apenas no desenvolvimento, o *Heroku* oferece um excelente serviço de hospedagem de aplicações com uma boa oferta de complementos simplificando o processo de escalar a aplicação.

Diferente do Infraestrutura como Serviço (IaaS), no qual o cliente contrata máquinas reais ou virtuais e é responsável pela instalação de bibliotecas, montagem das estruturas do sistema de arquivos, entre outros recursos, o PaaS é uma solução de alto nível que abstrai este tipo de preocupação.

O *Heroku*, assim como os demais serviços PaaS, disponibiliza um ambiente de execução de aplicações. Este tipo de solução abstrai do cliente detalhes do sistema operacional como bibliotecas, serviços de startup, gestão de memória, sistema de arquivos, entre outros, provendo uma maneira muito mais simples e prática de subir e escalar as aplicações.

A Figura 4 mostra o funcionamento do scraper em atividade na internet, apresentando a entrada dos parâmetros para execução da consulta proposta assim como os respectivos resultados obtidos.

Figura 4 - Scraper rodando em interface Web



Fonte: elaborado pelos autores

Nos próximos parágrafos, de acordo com a Figura 4, serão explanados os campos de entrada e saída da consulta realizada.

Ao se acessar o scraper em seu endereço Web, verifica-se que o campo Portal, na versão atual, tem como única possibilidade de seleção o BRAPCI, porém no futuro, pretende-se expandir a gama de fontes de extração de dados para contemplar outros repositórios.

Seleciona-se o tipo de pesquisa que o usuário pretende realizar. Isso refere-se aos campos dos documentos armazenados no portal BRAPCI, onde a chave de busca será pesquisada, conforme abaixo:

1. Todos
2. Autores
3. Título
4. Palavras-chave
5. Resumo
6. Texto completo

No passo seguinte, realiza-se a entrada da chave de pesquisa. Na sequência, delimita-se o período a ser pesquisado, com a entrada do Ano inicial e Ano final.

Clicando-se no botão “Pesquisar”, o *software* então, retorna o resultado da busca.

Após processamento, é mostrado um painel com uma visão geral dos resultados recuperados e um arquivo do tipo CSV é gerado e disponibilizado para download contendo os seguintes campos:

- Title – Título da Publicação
- Link Pub – Link direto para o documento no portal
- Autor – Autor da obra
- Revista – Instituição que realizou a publicação

A recuperação destes campos específicos, foi previamente definida durante o desenvolvimento do *scraper*, entretanto, outros itens disponibilizados na estrutura do BRAPCI, também poderiam ser resgatados.

Com o objetivo de favorecer uma melhor compreensão do processo de recuperação, e descrever as etapas da coleta, foi utilizado, como exemplo, o termo “recuperação da informação”, no campo de busca do *ScraperCI.* Para evitar a recuperação de dados não correspondentes, foi selecionado o campo “Palavra-chave” como critério de resultado e delimitado o período de 20 anos, de 2002 a 2020, o que resultou em um total de 158 documentos distribuídos em 9 páginas.

Após o processamento dos dados coletados, o arquivo tipo texto puro pode ser importado para qualquer *software* para análise dos dados e, neste trabalho, os dados foram tabulados no *Google Sheets*, sendo possível realizar de maneira eficiente múltiplas análises e obter conclusões.

Na Figura 5, demonstra-se que as 158 publicações recuperadas foram publicadas em 37 instituições, sendo que 6 delas foram responsáveis por 83 publicações (52% do total). Com essa verificação, o profissional da informação, poderia por exemplo, endereçar o artigo que deseja publicar, para as instituições onde se tem maior probabilidade de aceite dada a afinidade que elas têm com o tema alvo.

Figura 5 - Publicações por instituição

Revista	Quantidade	Percentual	Sparkline
Informação & Informação	20	12,7%	
Ciência da Informação	17	10,8%	
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação	15	9,5%	
Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação	11	7,0%	
Revista Digital de Biblioteconomia & Ciência da Informação	10	6,3%	
Perspectivas em Ciência da Informação	10	6,3%	
Em Questão	8	5,1%	
AtoZ: Novas Práticas em Informação e Conhecimento	7	4,4%	
Ponto de Acesso	6	3,8%	
Revista Brasileira de Biblioteconomia e Documentação	5	3,2%	
Revista Bibliomar	4	2,5%	
Informação & Sociedade: Estudos	4	2,5%	
Tendências da Pesquisa Brasileira em Ciência da Informação	3	1,9%	
Seminário Nacional de Gestão da Informação e do Conhecimento	3	1,9%	
DataGramaZero	3	1,9%	
Congresso de Gestão Estratégica da Informação, Empreendedorismo e Inovação	3	1,9%	
Ágora	3	1,9%	
Revista Fontes Documentais	2	1,3%	
Revista Folha de Rosto	2	1,3%	
Revista ACB: Biblioteconomia em Santa Catarina	2	1,3%	
Logeion: filosofia da informação	2	1,3%	
Informação@Profissões	2	1,3%	
Informação & Tecnologia	2	1,3%	
Revista Eletrônica de Comunicação, Informação e Inovação em Saúde	1	0,6%	
Revista Conhecimento em Ação	1	0,6%	
Perspectivas em Gestão & Conhecimento	1	0,6%	
Múltiplos Olhares em Ciência da Informação	1	0,6%	
Memória e Informação	1	0,6%	
Informação em Pauta	1	0,6%	
Informação Arquivística	1	0,6%	
InCID: Revista de Ciência da Informação e Documentação	1	0,6%	
Convergência em Ciência da Informação	1	0,6%	
Ciência da Informação em Revista	1	0,6%	
Brazilian Journal of Information Science	1	0,6%	
BIBLOS - Revista do Instituto de Ciências Humanas e da Informação	1	0,6%	
Biblios (Peru)	1	0,6%	
Archeion Online	1	0,6%	

Fonte: elaborado pelos autores.

Observou-se que, 252 autores estão relacionados aos 158 documentos recuperados, todavia, como apresentado na Figura 6, um total de 22 autores (8,7% do total) tiveram recorrência de 3 ou mais documentos publicados relacionados ao tema, no período. Esse *ranking* poderia ser utilizado, por exemplo, como ponto de partida para a pesquisa sobre representação da informação, onde se selecionaria os maiores especialistas no assunto.

Caso o processo de consulta fosse realizado sem o uso da tecnologia aqui proposta, e com o buscador padrão do *site*, o usuário teria que compilar os dados manualmente, registrando os dados de 158 documentos e navegando por 8 páginas diferentes. Esse processo, embora possa levar aos mesmos resultados, seria moroso e mais propenso a ocorrências de erros.

Figura 6 - Autores que publicaram 3 ou mais documentos no período

Autores	Quantidade	Sparkline
RAMALHO, Rogério Aparecido de Sá	10	
SIMIONATO, Ana Carolina	8	
CASTRO, Fabiano Ferreira de	7	
MARTINS, Paulo George Miranda	6	
FUJITA, Mariângela Spotti Lopes	6	
ALBUQUERQUE, Maria Elisabeth Baltar Carneiro de	6	
SOUSA, Janailton Lopes	5	
MACULAN, Benildes Coura Moreira dos Santos	5	
LIMA, Gercina Ângela Borém de Oliveira	5	
BRÄSCHER, Marisa	5	
BARROS, Camila Monteiro	5	
VITAL, Luciane Paula	4	
SANTOS, Plácida Leopoldina Ventura Amorim da Costa	4	
ARAúJO JUNIOR, Rogério Henrique	4	
SOUZA, Rosali Fernandez	3	
SOUSA, Renato Tarciso Barbosa	3	
RAUTENBERG, Sandro	3	
PINHO, Fabio Assis	3	
MOREIRA, Walter	3	
CERVANTES, Brígida Maria Nogueira	3	
CAFé, Lígia	3	
ALBUQUERQUE, Ana Cristina	3	

Fonte: elaborado pelos autores

Tal fato denota que, o processo automatizado de busca pelo uso do *Web scraper* traz inúmeros benefícios aos usuários de ambientes digitais e para profissionais da informação, que podem dispor de mais tempo para o exercício de atividades relacionadas a análises de conteúdos e endereçamento de soluções. Com isso, tais atividades podem agregar maior valor as organizações, uma vez que os profissionais da informação não necessitam empreender esforços em processos manuais e repetitivos de coleta, preparação e estruturação dos dados, para só depois pôr em prática seu trabalho analítico.

Apesar das funcionalidades disponibilizadas pelos repositórios informacionais, incluindo o BRAPCI, com o objetivo de simplificar a recuperação da informação, estas nem sempre são de fácil compreensão e muitas vezes estão restritas ao acervo do próprio repositório. Nesse sentido, ferramentas como o *Scraperci* apresentam uma vantagem significativa, uma vez que podem ser configuradas para extrair dados de vários portais, além de possibilitar a personalização do resgate de campos específicos de acordo com a relevância para o usuário, bem como o cruzamento de informações.

Uma desvantagem associada ao uso da ferramenta em questão é a necessidade de configuração individual para cada novo portal que se deseja incluir em sua gama de possibilidades de busca. Isso se dá em razão da estrutura específica de cada site, que deve

ser contemplada na ferramenta para que a funcionalidade de buscas funcione adequadamente. Tal fato pode gerar um esforço adicional por parte do usuário, que precisará dispor de tempo e recursos para efetuar a configuração necessária. Entretanto, uma vez que a ferramenta esteja devidamente configurada para o site em questão, é possível extrair informações relevantes de forma eficiente e precisa, o que pode compensar o esforço inicial despendido.

Acredita-se que a ferramenta em questão possa ser de grande auxílio para pesquisadores que utilizam dados coletados na web em suas pesquisas, incluindo estudos métricos, tendo em vista a atual necessidade de disponibilização dos dados brutos utilizados em pesquisas juntamente com os artigos resultantes das pesquisas.

Nessa perspectiva, a discussão em torno da relevância e utilidade dessa ferramenta pode contribuir significativamente para o desenvolvimento de estudos mais precisos e detalhados no campo da Ciência da Informação.

5 CONSIDERAÇÕES FINAIS

A partir dos resultados apresentados foi possível demonstrar como a aproximação entre conhecimentos teóricos, relacionados à temática de recuperação da informação, e aspectos práticos relacionados a utilização de linguagens de programação podem favorecer o desenvolvimento de ferramentas capazes de possibilitar processos de recuperação mais eficientes, e melhoras significativas nas atividades desenvolvidas por profissionais da informação. Dessa forma, a pesquisa alcançou tanto o objetivo geral quanto os objetivos específicos, embora tenha sido conduzida em um único repositório como ambiente de teste e com recursos computacionais limitados.

Apesar das limitações do protótipo *ScraperCI*, verificou-se que o uso de *Web scrapers* favorece a automatização de processos de coleta de dados, ampliando as possibilidades e trazendo maior produtividade no que tange a extração de recursos informacionais na *Web*, apresentando-se como uma alternativa viável a ser explorada pelos profissionais da informação.

Tal ferramenta apresenta várias potencialidades no processo de coleta de dados em sites da web, tais como a automatização do processo, que possibilita a obtenção rápida e eficiente de grandes volumes de informações, a flexibilidade na escolha dos dados a serem coletados, a customização dos métodos de busca, redução de erros de coleta de dados em

relação à coleta manual, a potencialização da análise de dados, e redução de custos em relação à coleta manual de dados.

Espera-se, que esta pesquisa possa despertar o interesse de outros pesquisadores interessados nesta temática, contribuindo para uma maior disseminação de pesquisas sobre o uso de *Web scrapers* na área de Ciência da Informação.

Em estudos futuros, há a perspectiva de ampliar a abrangência da ferramenta desenvolvida, adicionando recursos que permitam a recuperação de informações em diferentes repositórios, bem como aprimorar seus métodos de busca, a fim de otimizar tanto o tempo de resposta quanto a qualidade dos resultados para o usuário. Ademais, é possível realizar análises não apenas dos aspectos técnicos e produtivos, mas também dos impactos sociais decorrentes do uso dessa tecnologia.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação**: conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

BORKO, H. Information science: What is it? **American Documentation**, [*s.l.*], v. 19, n. 1, p. 3-5, 1968.

CHOWDHURY, G. G. **Introduction to modern information retrieval.** 3. ed. New York: Neal-Schuman Publishers, 2010.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. **Computer Networks and ISDN Systems**, [*s.l.*], v. 30, n. 1-7, p. 107-117, 1998. Disponível em: https://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf. Acesso em: 25 fev. 2022.

DASTIDAR, B. G.; BANERJEE, D.; SENGUPTA, S. An Intelligent Survey of Personalized Information Retrieval using *Web Scraper*. **International Journal of Education and Management Engineering**, [*s.l.*], v. 6, n. 5, p. 24-31, 2016. Disponível em: https://www.mecs-press.org/ijeme/ijeme-v6-n5/IJEME-V6-N5-3.pdf. Acesso em: 25 fev. 2022.

IDC. **The State of Data Discovery and Cataloging**. IDC White Paper, 2018. Disponível em: https://www.datateam.mx/downloads/alteryx/The_State_of_Data_Discovery__Cataloging.pdf. Acesso em: 25 fev. 2022.

MITCHELL, R. ***Web* Scraping with Python**: collecting more data from the modern *web*. 2nd ed. [*S.l.*]: O'Reilly Media, 2018.

MOOERS, C. N. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, [*s.l.*], v. 2, n. 1, p. 20-32, 1951. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090020107. Acesso em: 25 fev. 2022.

PROBSTEIN, S. Reality check: still spending more time gathering instead of analyzing. **Forbes Technology Council**, 2019. Disponível em: https://www.forbes.com/sites/forbestechcouncil/2019/12/17/reality-check-still-spending-more-time-gathering-instead-of-analyzing. Acesso em: 25 fev. 2022.

RAMALHO, R. A. S.; OUCHI, M. T. Tecnologias Semânticas: novas perspectivas para a representação de recursos informacionais. **Informação & Informação**, Londrina, v. 16, n. 3, p. 75-60, 2011. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/9829. Acesso em: 25 fev. 2022.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116-142, 2016. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27940. Acesso em: 25 fev. 2022.

SANT'ANA, R.C.G. Transdução informacional: impactos do controle sobre os dados. *In*: MARTÍNEZ-ÁVILA, D.; SOUZA, E.A.; GONZALEZ, M.E.Q. (ed.). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução? Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019, p. 117-128. Disponível em: http://books.scielo.org/id/gfrbh/pdf/martinez-9788572490559-09.pdf. Acesso em: 25 fev. 2022.

SILVEIRA, D. T.; CÓRDOVA, F. P. A pesquisa científica. *In*: GERHARDT, T. E., SILVEIRA, D. T. (orgs.). **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009. Disponível em: http://hdl.handle.net/10183/52806. Acesso em: 25 fev. 2022.

SIRISURIYA, S. A. Comparative study on web scraping. *In*: INTERNATIONAL RESEARCH CONFERENCE, 8., 2015, KDU. **Proceedings** […]. [*S.l.: s.n.*], 2015. Disponível em: http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf. Acesso em: 25 fev. 2022.

SOUZA, R. R.; ALMEIDA, M. B.; BARACHO, R. M. A. Ciência da informação em transformação: Big Data, nuvens, redes sociais e Web Semântica. **Ciência da Informação**, Brasília, v. 42, n. 2, p. 159-173, 2013. Disponível em: https://revista.ibict.br/ciinf/article/view/1379. Acesso em: 25 fev. 2022.

SILVA, R. E. DA; SANTOS, P. L. V. A. DA C.; FERNEDA, E. Modelos de recuperação de informação e web semântica: a questão da relevância. **Informação & Informação**, Londrina v. 18, n. 3, p. 27, 2013. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/12822. Acesso em: 25 fev. 2022..

UPADHYAY. S. *et al*. Articulating the construction of a Web scraper for massive data extraction. *In*: INTERNATIONAL CONFERENCE ON ELECTRICAL, COMPUTER AND COMMUNICATION TECHNOLOGIES (ICECCT), 2., 2017, Coimbatore, India. **Proceedings** […]. [*S.l.: s.n.*], 2017. Disponível em: https://ieeexplore.ieee.org/document/8117827. Acesso em: 22 jan. 2022.

NOTAS

AGRADECIMENTOS
A presente pesquisa foi realizada com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

CONTRIBUIÇÃO DE AUTORIA
Os papéis descrevem a contribuição específica de cada colaborador para a produção acadêmica inserir os dados dos autores conforme exemplo, excluindo o que não for aplicável. Iniciais dos primeiros nomes acrescidas com o último Sobrenome, conforme exemplo.

Concepção e elaboração do manuscrito: H. L. S. Graciano
Coleta de dados: H. L. S. Graciano
Análise de dados: H. L. S. Graciano
Discussão dos resultados: H. L. S. Graciano, R. A. S. Ramalho
Revisão e aprovação: R. A. S. Ramalho, H. L. S. Graciano

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

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Mayara Madeira Trevisol, Jônatas Edison da Silva, Camila Letícia Melo Furtado e Beatriz Tarré Alonso.

HISTÓRICO
Recebido em: 10-01-2023 – Aprovado em: 17-04-2023 - Publicado em: 11-05-2023.