**XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB**

**ISSN 2177-3688**

**GT-8 – Informação e Tecnologia**

# *WEB SCRAPER*: POTENCIALIDADES COMO FERRAMENTA DE AUXILIO À PESQUISA

## *WEB SCRAPER: POTENTIALITIES AS A RESEARCH AID TOOL*

**Helton Luiz dos Santos Graciano.** UFSCar.

**Paulo George Miranda Martins.** UNESP.

**Rogério Aparecido Sá Ramalho.** UFSCar.

**Ricardo César Gonçalves Sant'Ana.** UNESP.

**Modalidade: Trabalho Completo**

**Resumo:** O desenvolvimento tecnológico vivenciado nas últimas décadas, a popularização da internet e a produção massiva de recursos informacionais dos mais variados tipos, tem proporcionado mudanças significativas que culminaram na transformação do ambiente *Web*. Nesse contexto, o principal objetivo deste trabalho foi demonstrar a utilização de um protótipo de *Web scraper* desenvolvido como ferramenta de coleta em bases de dados da *Web*. 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. Conclui-se que tais dispositivos são ferramentas viáveis para a coleta de dados, automatizando os processos de recuperação e ampliando as possibilidades, trazendo maior produtividade no que tange a extração de recursos informacionais na *Web*. Espera-se também, 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.

**Abstract:** The technological development experienced in the last decades, the popularization of the internet and the massive production of information resources of the most varied types, has provided significant changes that culminated in the transformation of the Web environment. In this context, the main objective of this work was to demonstrate the use of a Web scraper prototype developed as a collection tool in Web databases. The research is characterized as applied, exploratory and descriptive, with a qualitative approach that aims to identify the potential of using Web scrapers in the data collection process. It is concluded that such devices are viable tools for data collection, automating the recovery processes and expanding the possibilities, bringing greater productivity in terms of extracting information resources on the Web. It is also expected that this research can stimulate 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 Retrieval. Web Scraping. Search Engines.

## 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 (CI) 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, consideramos 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 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 CI "[...] 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 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.

## 2 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa caracteriza-se como uma pesquisa aplicada de natureza qualitativa, desenvolvida a partir de uma abordagem exploratória e 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 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*.

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

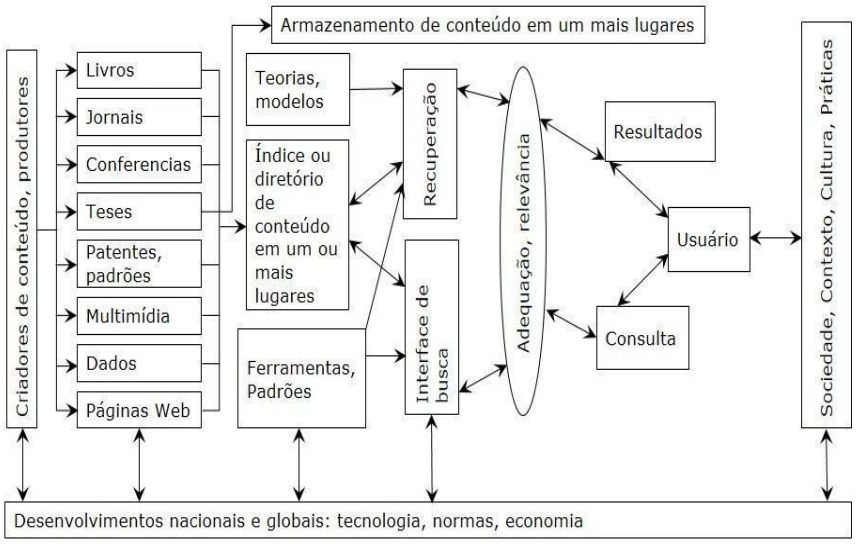
documentos, mas também como utilizá-la para decidir quanto à sua relevância, a qual exerce um papel central (BAEZA-YATES; RIBEIRO-NETO, 2013).

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*, 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*.

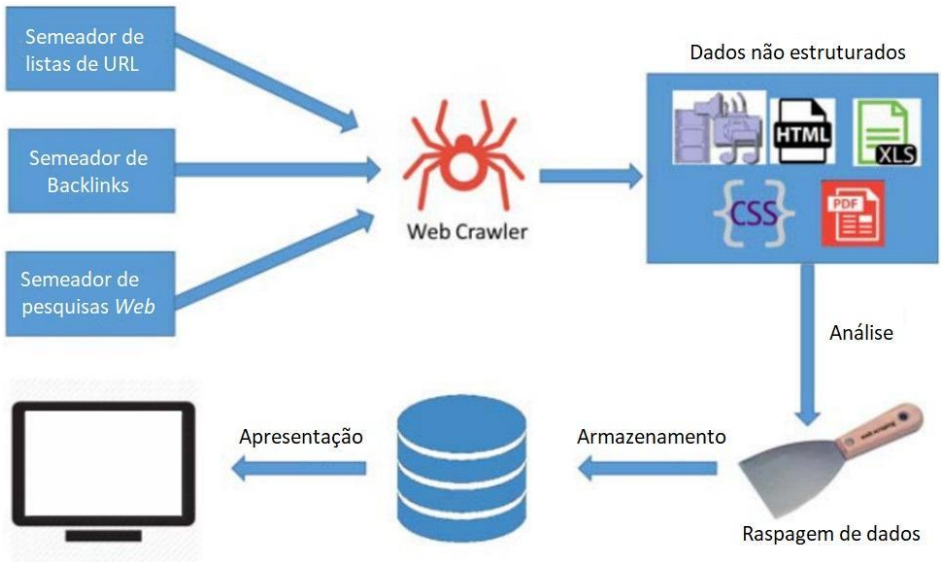
O rastreamento de conteúdo em um *site* a partir dessa técnica ocorre quando um robô passa por todas as páginas e *links* até a última linha do *site*, procurando determinados dados. Em síntese, esse processo consiste em visualizar uma página de um *site* como um todo e indexá-la, geralmente capturando 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 (SIRISURIYA, 2015; GHOSH DASTIDAR et al., 2016).

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*.

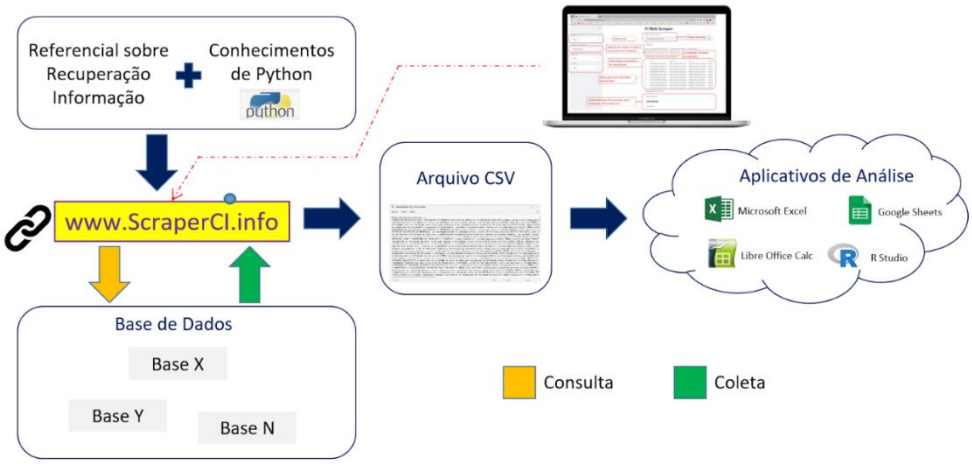
## 4 *SCRAPERCI* COMO PROTÓTIPO DE *WEB SCRAPER* PARA COLETA DE DADOS

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 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, um painel geral dos resultados foi estruturado e um arquivo do tipo CSV foi gerado e disponibilizado para *download*. 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 Tabela 1, demonstra-se que as 158 publicações recuperadas foram publicadas em 37 instituições, sendo que 6 delas foram responsáveis por 83 publicações (52% do total).

**Tabela 1 - 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 Tabela 2, 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.

**Tabela 2 - 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.

Nessa perspectiva, as tecnologias e ou ferramentas disponíveis a um número restrito de pessoas passam a ganhar escala com potencial para solucionar problemas de outros profissionais e áreas que podem, a partir disso, propor melhorias e até mesmo se inspirar na criação de novas soluções.

## 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.

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.

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.

**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, v. 19, n. 1, p. 3-5, 1968.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. **Computer Networks and ISDN Systems**, v. 30, n. 1–7, p. 107–117, 1998.

GHOSH DASTIDAR, B.; BANERJEE, D.; SENGUPTA, S. An Intelligent Survey of Personalized Information Retrieval using *Web Scraper*. **International Journal of Education and Management Engineering**, v. 6, n. 5, p. 24-31, 2016.

MITCHELL, R. ***Web* Scraping with Python: Collecting More Data from the Modern *Web***. 2nd Editio ed. [s.l.] O'Reilly Media, 2018.

MOOERS, C. N. **Zatocoding applied to mechanical organization of knowledge**. American Documentation, v. 2, n. 1, p. 20-32, 1951.

PROBSTEIN, S. Reality Check: Still Spending More Time Gathering Instead Of Analyzing. **Forbes Technology Council**, 2019.

RAMALHO, R. A. S.; OUCHI, M. T. Tecnologias Semânticas: Novas Perspectivas para a Representação de Recursos Informacionais. **Informação & Informação**, v. 16, n. 3, p. 75-60, 2011.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, v. 21, n. 2, p. 116-142, 2016.

SANT'ANA, R.C.G. **Transdução informacional: impactos do controle sobre os dados**. In: MARTÍNEZ-ÁVILA, D., SOUZA, E.A., and GONZALEZ, M.E.Q., eds. Informação, conhecimento, ação autônoma e big data: continuidade ou revolução? [online]. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica; FiloCzar, 2019, p. 117-128. ISBN: 978-85-7249-055-9. Disponível em: http://books.scielo.org/id/gfrbh/pdf/martinez-9788572490559-09.pdf. Acesso em: 25 fev. 2022.

SILVEIRA, D. T., & Córdova, F. P. (2009). **A pesquisa científica**. In: Gerhardt, T. E., & Silveira, D. T. (Orgs.). Métodos de pesquisa. Porto Alegre: Editora da UFRGS.

SIRISURIYA, S. A **Comparative Study on Web Scraping**. 8th International Research Conference, KDU, n. November, p. 135-140, 2015.

SOUZA, R. R.; ALMEIDA, M. B.; BARACHO, R. M. A. Ciência da informação em transformação: Big Data, nuvens, redes sociais e Web Semântica. **Ciencia da Informacao**, v. 42, n. 2, p. 159-173, 2013.

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; Los Modelos de recuperación de la información y la web semántica: la cuestión de la pertinencia. **Informação & Informação**, v. 18, n. 3, p. 27, 2013.

IDC. **The State of Data Discovery and Cataloging**. IDC White Paper, 2018.

UPADHYAY. S.; PANT, P.; BHASIN, S.; PATTANSHETTI, M. K. Articulating the construction of a Web scraper for massive data extraction. Proceedings of the 2017 2nd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT, 2017. **Anais**...IEEE, fev. 2017. Disponível em: https://ieeexplore.ieee.org/document/8117827. Acesso em: 22 jan. 2022.