Universidade Federal de São Carlos
Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência da Informação

HELTON LUIZ DOS SANTOS GRACIANO

ScraperCI: um protótipo de Web scraper para coleta de dados

São Carlos
2022

HELTON LUIZ DOS SANTOS GRACIANO

ScraperCI: um protótipo de Web scraper para coleta de dados

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Universidade Federal de São Carlos, Campus de São Carlos, como requisito parcial para obtenção do título de Mestre em Ciência da Informação.

Área de Concentração: Conhecimento, Tecnologia e Inovação.

Orientador: Prof. Dr. Rogério Aparecido Sá Ramalho

São Carlos

2022

Graciano, Helton Luiz dos Santos

ScraperCI: um protótipo de Web scraper para coleta de dados / Helton Luiz dos Santos Graciano -- 2022.
79f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Rogério Aparecido Sá Ramalho
Banca Examinadora: Ana Carolina Simionato Arakaki, Ricardo César Gonçalves Sant'Ana
Bibliografia

1. Recuperação da informação. 2. Web Scraping. 3. Mecanismos de busca. I. Graciano, Helton Luiz dos Santos. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Ronildo Santos Prado - CRB/8 7325

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência da Informação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Helton Luiz dos Santos Graciano, realizada em 12/05/2022.

Comissão Julgadora:

Prof. Dr. Rogério Aparecido Sá Ramalho (UFSCar)

Profa. Dra. Ana Carolina Simionato Arakaki (UFSCar)

Prof. Dr. Ricardo César Gonçalves Sant'Ana (UNESP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Informação.

DEDICATÓRIA

Dedico à minha amada Raquel pelo apoio, dedicação e companheirismo em todos os momentos e aos nossos filhos Theo e Liz, com todo meu amor.

AGRADECIMENTO

A Deus por TUDO que tem me proporcionado, dando-me saúde e força para superar todos os obstáculos impostos pela vida.

Agradeço aos meus amados pais, Luiz Carlos Pereira Graciano e Ângela Maria dos Santos Graciano, por serem à base de tudo para mim e pelos ensinamentos que me orientaram para o caminho do bem.

À minha esposa Raquel, companheira nessa jornada e grande incentivadora na realização desse sonho.

Ao meu orientador Rogério Ramalho, pela paciência, dedicação e profissionalismo. Aos professores Ana Carolina Simionato e Ricardo Sant'Ana, pelo profissionalismo e contribuições para o enriquecimento da pesquisa.

A todos os professores do Programa de Pós-Graduação em Ciência da Informação (PPGCI), pelos conhecimentos e experiências compartilhadas nas disciplinas ministradas.

Ao meu amigo Paulo Martins por toda ajuda e incentivo, não somente no desenvolvimento dessa pesquisa, mas por apresentar e me fazer despertar o interesse pela Ciência da Informação.

Enfim, agradeço a todos aqueles que direta ou indiretamente contribuíram para a realização desta pesquisa. O meu MUITO OBRIGADO.

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*. O objetivo dessa pesquisa é contribuir para a ampliação da percepção das vantagens do uso de ferramentas de coleta de dados no processo de recuperação da informação, a partir do uso de *Web scrapers*. 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. Como resultado, foi elaborado na linguagem de programação *Python*, um protótipo de *Web scraper* e uma demonstração prática de sua utilização. Conclui-se que o uso do *Web scraper* pode favorecer a recuperação de informações, ampliando as possibilidades e trazendo maior produtividade no que tange a extração de recursos informacionais na *Web.* A presente pesquisa contribuirá para uma maior compreensão das potencialidades do uso de *Web scrapers* para a coleta de dados e servirá de estímulo aos profissionais da informação a desenvolver novas competências e possibilidades inovadoras de atuação profissional.

Palavras-chave: Recuperação da informação. Coleta de dados. *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. The objective of this research is to contribute to the expansion of the perception of the advantages of using data collection tools in the information retrieval process, using Web scrapers. 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 information retrieval process. As a result, a Web scraper prototype and a practical demonstration of its use were prepared in the Python programming language. It is concluded that the use of the Web scraper can favor the retrieval of information, expanding the possibilities and bringing greater productivity regarding the extraction of informational resources on the Web. This research will contribute to a greater understanding of the potential of using Web scrapers for data collection and will serve as a stimulus for information professionals to develop new skills and innovative possibilities for professional activity.

Keywords: Information retrieval. Data collection. Web Scraping. Search engines.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BRAPCI	Base de Dados em Ciência da Informação
CD-ROM	Compact-Disc Read-Only Memory
CI	Ciência da Informação
CSV	Comma Separated Values
CVD	Ciclo de Vida dos Dados
HTML	HyperText Markup Language
HTTP	Hyper Text Transfer Protocol
IA	Inteligência Artificial
IaaS	Infrastructure as a Service
ODT	Open Document Text
OPAC	Online Public Access Catalog
PaaS	Platform as a Service
PDF	Portable Document Format
RI	Recuperação da Informação
SGBD	Sistema de Gerenciamento de Banco de Dados
SRI	Sistema de Recuperação de Informação
TSV	Tab Separated Values
WWW	World Wide *Web*
XLS	Document Type Definition
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento tecnológico vivenciado nas últimas décadas, a popularização da *Web* e a produção massiva e exponencial de recursos informacionais tem proporcionado mudanças significativas que culminou em uma transformação sem igual do ambiente *Web*, tornando-o o maior repositório de informações existente na contemporaneidade.

Atualmente, temos uma extensa quantidade de dados sendo produzida, por pessoas, aplicativos, equipamentos e dispositivos. Tais dados, conforme sua natureza - física ou digital - são armazenados das mais distintas maneiras e locais. No caso de dados digitais, geralmente são mantidos em diretórios de rede, sistemas empresariais, *sites*, nuvens, e-mails, aplicativos *desktop* ou *mobile* e nas mais variadas formas de armazenamento.

Nesse cenário, observa-se que na rotina das organizações falta uma preocupação quanto a um melhor aproveitamento do tempo, bem como com a dificuldade de acesso às informações, vindo a culminar em uma grande ineficiência, que consequentemente, afeta a vantagem competitiva das empresas com o aumento dos custos e retardo na velocidade das tomadas de decisão.

Segundo Chowdhury (2010), métodos e sistemas, que permitam lidar e recuperar de maneira eficiente as informações, no momento em que são necessárias, são fundamentais para o desenvolvimento das mais diversas atividades, das mais complexas às mais triviais.

De acordo com o autor, recuperar uma mensagem ou e-mail recebido ou enviado em uma data específica, encontrar algo ou alguém na *Web* e pesquisar um livro em um catálogo de biblioteca online ou em uma biblioteca digital, são processos facilitados por tecnologias cada vez mais eficientes no processo de recuperação da informação (CHOWDHURY, 2010).

Essa grande diversidade de formas e meios de armazenamento aliada a falta de procedimentos e diretrizes estabelecendo boas práticas nesse processo, resultam em morosidade no resgate e muitas vezes até na perda definitiva da informação.

Como uma possível solução para os problemas relacionados à organização da informação foi proposto, por Vannevar Bush em 1945, o Memex. O objetivo da tecnologia, à época, era de que pudesse integrar profissionais e instrumentos computacionais capazes de tornar o armazenamento e a recuperação da informação

processos eficientes, mediante a associação de palavras para a recuperação eficaz de informações (BARRETO, 2002).

A proposta de Bush trouxe à luz uma série de debates no universo acadêmico quanto à correta forma e técnicas usadas, até então, para o armazenamento, a organização e a recuperação, uma vez que a informação é considerada um ativo de suma importância, sendo,

> [...] qualificada como instrumento modificador da consciência do homem. Quando adequadamente apropriada, produz conhecimento e modifica o estoque mental de saber do indivíduo; traz benefícios para seu desenvolvimento e para o bem-estar da sociedade em que ele vive (BARRETO, 2002, p. 70).

Segundo Chowdhury, (2010, p. 3, tradução nossa),

> [...] embora historicamente os sistemas de recuperação de informações tenham sido projetados para ajudar as pessoas a encontrar informações em bancos de dados bibliográficos e textuais, no mundo de hoje usamos sistemas de recuperação de informações em quase todos os aspectos de nossas vidas diárias.

Não obstante, mesmo após uma recuperação bem-sucedida dos dados e subsequente transformação em informação útil, ocorre que o produto gerado - traduzido em forma de apresentações, relatórios, infográficos, manuais, documentos, etc - é normalmente submetido às mesmas falhas no que tange aos processos de armazenamento e recuperação.

Devido a grande diversidade de fontes de informação, Feldman e Sherman (2001) destacam que se gasta muito mais tempo no processo de coleta do que na análise dos dados, ressaltando que a perda de tempo na busca por informações pode ser de até duas horas e meia diariamente, dependendo do ambiente ao qual o indivíduo esteja inserido.

Tal afirmação é corroborada por um estudo realizado por Chui et al. (2012) que mostrou que os indivíduos ainda gastavam 19% de seu tempo procurando e reunindo informações. Mais recentemente, uma pesquisa denominada “*The State of Data Discovery and Cataloging*”, mostrou que os profissionais da informação estão perdendo 50% do seu tempo todas as semanas - 30% pesquisando, controlando e preparando dados e mais 20% “reinventando a roda”, ou seja, elaborando ativos

informacionais que já existiam e que podem ser utilizados no formato que foram projetados (IDC, 2018).

Probstein (2019, “n.p”, tradução nossa), observa que:

> [...] vimos o surgimento da nuvem, da computação onipresente, da conectividade e de tudo o mais que era ficção científica quando éramos crianças se tornando uma realidade - incluindo o surgimento iminente da IA. Claramente, todos os avanços da tecnologia não mudaram o paradigma da produtividade; parece que ainda gastamos mais tempo procurando informações existentes do que analisando e criando novos conhecimentos.

Logo, explora-se a possibilidade de o desenvolvimento e a disponibilidade de recursos tecnológicos por si só não trazerem benefícios e / ou agregarem valor para as organizações. O que pode fazer a diferença de fato, é as pessoas estarem dispostas e preparadas para utilizarem tais tecnologias no seu dia a dia.

Diante do grande volume de recursos informacionais disponíveis, a grande problemática reside em como recuperar, de forma precisa, as informações necessárias que possam atender as demandas dos usuários. Para Feitosa (2006, p. 15),

> [...] a principal dificuldade do usuário na recuperação da informação é a falta de conhecimento ao elaborar questões que reflitam seus objetivos de busca, inabilidade de interpretar, classificar, priorizar ou filtrar as grandes quantidades de informações retornadas pelo sistema de busca.

Dito isso, Souza, Almeida e Baracho (2013, p. 162) salientam que, “[...] a organização de imensas massas de dados necessita de novas e criativas soluções; nunca se precisou tanto de uma Ciência da Informação (CI) para orquestrar estes esforços”.

A Ciência da Informação, como área do conhecimento, e seus mais diversos profissionais, seja no meio acadêmico ou corporativo, deveria exercer, neste processo de gerenciamento de informações, um papel proeminente, fundamentando toda a base que se segmenta relacionando-a com outras áreas incumbidas de lidar com partes específicas desse sistema complexo e dinâmico.

Para Borko (1968, p. 5, tradução nossa) a Ciência da Informação pode ser definida como:

> [...] uma ciência interdisciplinar que investiga as propriedades e o comportamento da informação, as forças que governam o fluxo e o uso da informação e as técnicas, tanto manuais quanto mecânicas, do processamento da informação para o ideal armazenamento, recuperação e disseminação.

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

Com base no processo de encontro ou descoberta, no que diz respeito as informações previamente armazenadas, a presente pesquisa adere como definição para o termo Recuperação da Informação (RI), a estabelecida por Mooers (1951).

> Recuperação de informação é o nome para o processo ou método pelo qual um usuário potencial de informação é capaz de converter sua necessidade de informação em uma lista real de citações de documentos armazenados contendo informações úteis para ele (MOOERS, 1951, p. 25, tradução nossa).

Nessa perspectiva, propõe-se nesta pesquisa, a realização de uma análise exploratória e descritiva que visa identificar as potencialidades da utilização de *Web scrapers* no processo de recuperação da informação.

> Um *scraper* é um software automatizado, usado para extrair 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 extraindo dados armazenados neles (UPADHYAY et al., 2017, p. 1, tradução nossa).

A utilização de tal ferramenta pode propiciar que grandes volumes de dados e informações sejam recuperados de forma estruturada e ágil, permitindo que os profissionais da informação possam empreender maior tempo nas atividades analíticas e criativas, gerando, dessa forma, mais valor ao conteúdo resultante das análises.

1.1 Problema de pesquisa

Diante da grande massa e volume de dados que são gerados atualmente nas mais diversas formas, é imprescindível que os mesmos sejam administrados de maneira que estejam disponíveis no momento, no formato e no lugar que forem necessários, sendo não somente desejado que cheguemos às conclusões certas, mas também que isso aconteça no menor tempo possível.

Apesar de todas as ferramentas e habilidades dos profissionais da informação, com o crescimento exponencial da quantidade de informações disponíveis, torna-se necessário o uso de ferramentas eficientes para a recuperação de dados em grandes volumes.

Nessa perspectiva, com enfoque especial na temática pertinente à coleta de dados em ambientes digitais e impacto dessas novas tendências tecnológicas, surge a seguinte questão de pesquisa: Como o uso de *Web scrapers*, como ferramenta de coleta de dados, pode beneficiar a Ciência da Informação como área e seus profissionais?

1.2 Proposição

A presente pesquisa parte da necessidade em se compreender o uso de técnicas e métodos mais eficientes para o processo de recuperação dos mais diversos recursos informacionais, que estão abundantemente disponíveis na *Web*. Nesse sentido, surge a proposta de se analisar como as ferramentas de recuperação como o *scraper* pode contribuir para a melhoria dos processos de recuperação e de que forma os profissionais podem aproveitar suas habilidades para análises mais assertivas dos conteúdos recuperados.

Para tanto, propõe-se a criação de um *scraper* a partir da linguagem *Python*, explicitando seus fundamentos e conceitos, bem como seus desafios e, principalmente, quais contribuições a ferramenta pode trazer para melhoria das habilidades dos profissionais da informação.

1.3 Justificativa

Devido ao valor intrínseco dos dados, tendo potencial para agregar valor às organizações quando analisados de diversas perspectivas, se faz necessário que os mesmos sejam explorados através de ferramentas de recuperação mais eficientes.

Seja para produzir um artigo, preparar uma palestra, escrever um e-mail, elaborar um relatório, a recuperação da informação é um dos grandes desafios, mesmo com os mecanismos de busca cada vez mais desenvolvidos e vastamente implementados nos ambientes virtuais. Isso porque, em geral, nessas pesquisas, recupera-se uma quantidade enorme de conteúdo, que posteriormente devem passar por uma estruturação manual, tabelados e somente a partir daí ficam aptos para as diversas análises.

Nesse contexto, destaca-se o processo inovador de coleta da informação por meio de *Web scrapers*, que são softwares capazes de resgatar informações em massa na *Web* em períodos de tempo muito curtos quando comparados a buscas manuais.

De posse das informações, elas são estruturadas e submetidas a análises resultando em conclusões diversas, possibilitando assim, as tomadas de decisão com agilidade, fundamentais para atender as demandas de um período tão desafiador para as organizações, que correm contra o tempo em um mundo dinâmico e incerto.

A relevância desse estudo encontra-se no processo de desenvolvimento de habilidades cada vez mais necessárias para os profissionais da informação que, ao recuperar uma quantidade enorme de conteúdo devem ter um mínimo de conhecimento para estruturar o conteúdo recuperado e, somente a partir daí ficam aptos para as diversas análises concernentes ao perfil profissional.

No âmbito da relevância social, destaca-se que este estudo pode favorecer o desenvolvimento de atividades que atendam de maneira mais eficiente as demandas atuais da sociedade no que tange a recuperação de informações em grandes volumes de dados.

Com isso, a presente pesquisa justifica-se uma vez que servirá de estímulo ao profissional de Ciência da Informação a enxergar possibilidades inovadoras em sua atuação através de ferramentas que tornam o resgate e análise de informações mais eficientes e alinhados com as demandas atuais e futuras dessa profissão.

1.4 Objetivos

1.4.1 Geral

O objetivo geral desta pesquisa é contribuir para a ampliação da percepção das vantagens do uso de ferramentas de coleta de dados no processo de recuperação da informação, a partir do uso de *Web scrapers*.

1.4.2 Específicos

- Apresentar conceitos sobre a temática de Recuperação da Informação e seus fundamentos na Ciência da Informação;
- Explanar as tecnologias de recuperação na *Web*, com foco em *Web scrapers*, trazendo reflexões sobre os profissionais da informação nesse contexto;
- Demonstrar o uso prático de um *Web scraper* para coleta da informação, trazendo as potencialidades da utilização dessa ferramenta.

1.5 Procedimentos metodológicos

O desenvolvimento de ferramentas tecnológicas com o objetivo de garantir uma recuperação mais eficiente de informações, é uma preocupação e um desafio cada vez mais crescente nos estudos desenvolvidos no campo da Ciência da Informação, os quais buscam construir métodos e processos que auxiliem de formas mais eficazes, os anseios dos usuários em suas demandas informacionais.

A fim de atingir os objetivos propostos realizou-se uma pesquisa de cunho qualitativa e de natureza aplicada na qual se procura ponderar e relacionar a questão de pesquisa aos objetivos específicos propostos. Segundo Triviños (1987), a pesquisa qualitativa, dentre outros fatores, busca analisar e compreender o contexto e anseios ao qual a pesquisa pretende atender. Dessa forma, considera-se

> [...] que há uma relação dinâmica entre o mundo real e o sujeito, isto é, um vínculo indissociável entre o mundo objetivo e a subjetividade do sujeito que não pode ser traduzido em números. A interpretação dos fenômenos e a atribuição de significados são básicas no processo de pesquisa qualitativa (SILVA; MENEZES, 2001, p. 20).

Caracteriza-se também como uma pesquisa exploratória e descritiva, pois buscou-se maior conhecimento, entendimento e familiaridade com o objeto a ser investigado proposto neste estudo (GONDIM; LIMA, 2010).

Com o objetivo de melhor detalhar a metodologia percorrida nesse estudo, dividiu-se os procedimentos metodológicos nas etapas que se segue:

Em uma primeira etapa, como caminho metodológico para atender o objetivo específico para a construção de um arcabouço de fundamentação teórica em que o autor se apoia para a construção de um *Web scraper*, fez-se um levantamento, nos idiomas português e inglês, de referencial teórico, bibliográfico e documental, em bases de dados, livros, periódicos e *sites* sobre os termos *Web scraping*, busca automática da informação, mecanismos de busca, coleta, extração e análise de dados na *Web* e profissional da informação, na Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (BRAPCI).

O levantamento bibliográfico se constitui uma etapa fundamental em uma pesquisa, pois é [...] a partir do levantamento de referências teóricas já analisadas e publicadas por meios escritos e eletrônicos, como livros, artigos científicos, páginas de *Web sites*, [...] que permite ao pesquisador conhecer o que já se estudou sobre o assunto (FONSECA, 2002, p. 32).

A segunda etapa se caracterizou pela delimitação dos dados selecionados para análise, leitura analítica dos artigos selecionados referentes ao tema proposto, *Web scraping*, e pertinência com a área de Ciência da Informação, na base de dados BRAPCI, com o objetivo de um maior aprofundamento do assunto e aplicação sistemática das decisões adotadas (BARDIN, 2011).

Na terceira e última etapa foi desenvolvido um *Web scraper* que permite buscas relacionadas e interligadas à base de dados BRAPCI, como ambiente experimental. A proposta da ferramenta foi de analisar de forma prática as contribuições de um *scraper* para a coleta de dados e sua aplicação por parte dos profissionais da informação, frente às demandas informacionais cada vez mais crescentes vivenciadas na Ciência da Informação.

1.6 Estrutura de pesquisa

O presente trabalho inicia-se com uma introdução, no Capítulo 1, acerca das questões iniciais e contextualização da temática proposta para esta pesquisa.

No Capítulo 2, é apresentado um breve referencial teórico sobre a temática de Organização e Recuperação da informação, contextualizando os pressupostos desta temática como fruto da interdisciplinaridade da Ciência da Informação. Nesse tópico, ainda são contemplados os principais conceitos, elementos, arquitetura, funcionamento e principais características concernentes aos sistemas de recuperação de informação.

No Capítulo 3, é apresentada uma abordagem da recuperação de informação com maior ênfase no ambiente *Web*, demonstrando as bases para a realização das buscas por meio de *Web scrapers*. Ainda nesse capítulo, é realizada a argumentação sobre os novos desafios da Ciência da Informação dado o cenário atual, que vivenciamos nas organizações, bem como a crescente e inevitável necessidade do profissional de CI estar aberto a se desenvolver, adotando abordagens inovadoras em seu repertório de habilidades.

No Capítulo 4, é apresentado o caso prático do desenvolvimento de um *Web scraper* na linguagem de programação *Python*, a ser utilizado no portal BRAPCI, para resgate massivo, estruturação e análise de informações.

Finalmente, no Capítulo 5, apresenta-se os resultados e considerações da pesquisa, com discussões a respeito da ferramenta desenvolvida para as buscas informacionais, os desafios de seu uso e as possíveis contribuições para a CI e para formação de um perfil do profissional da informação que se adeque às novas demandas desse ofício.

2 ORGANIZAÇÃO E RECUPERAÇÃO DA INFORMAÇÃO

Desde os primórdios a humanidade busca organizar e armazenar informações com objetivo de proporcionar a geração de conhecimento a partir do processo de recuperação. Coleções de textos, registros e documentos eram armazenados em construções denominadas bibliotecas, através da organização e indexação de tabuletas de argila, hieróglifos, rolos de papiros e livros (BAEZA-YATES; RIBEIRO-NETO, 2013).

Neste capítulo, será discutido as bases que fundamentam a Recuperação da Informação (RI) com ênfase na perspectiva da Ciência da Informação (CI), sem deixar de mencionar, porém, alguns dos conceitos chave advindos da Ciência da Computação.

2.1 Bases históricas da organização da informação

A biblioteca mais antiga conhecida foi criada em Elba, no "Crescente Fértil", atual norte da Síria, entre 3000 e 2000 a.C. e a Biblioteca de Nínive, criada no século 7 a.C, continha mais de 30.000 tabuletas de argila até sua destruição em 612 a.C.

A mais famosa das bibliotecas do passado é a de Alexandria, criada em Alexandria na Macedônia em 300 a.C. em homenagem ao rei Alexandre o grande e que em conjunto com outras bibliotecas importantes fizeram de Alexandria a capital intelectual do mundo ocidental, por séculos (BAEZA-YATES; RIBEIRO-NETO, 2013).

Em constante ascensão, as bibliotecas tornaram-se largamente conhecidas e desenvolveram-se estando por toda a parte, fazendo parte da essência da memória coletiva da humanidade ao longo dos séculos.

Baeza-Yates e Ribeiro-Neto (2013) destacam que as bibliotecas estão entre as organizações pioneiras na implementação de Sistemas de Recuperação de Informação (SRI) para recuperar informações caracterizadas por três gerações:

- **primeira** - os sistemas eram baseados na automação de processos existentes, como a busca em catálogos de fichas, restritas ao nome do autor e ao título da obra.
- **segunda** - agregação de novas funções aos mecanismos de busca para abranger assuntos, palavras-chave e operadores de consulta.
- **terceira** - que está vigor na atualidade, tem convergido para o

aperfeiçoamento de interfaces gráficas, formulários eletrônicos, características de hipertexto e arquitetura de sistemas abertos (BAEZA-YATES; RIBEIRO-NETO, 2013).

Lesk (1996), baseando-se na divisão feita por Shakespeare (1599) nas sete fases - infância, escolaridade, idade adulta, maturidade, crise de meia idade, realização e reforma - procura verificar as relações de semelhança ou de disparidade entre a história da Recuperação da Informação (RI) e as etapas da vida de uma pessoa.

Para o autor, este ciclo inicia-se em 1945 com o artigo de Bush "As We May Think" e termina em 2010, com a reforma, sendo que, por volta de 2015, as atividades de pesquisa em geral seriam realizadas por meios digitais e não mais no papel, concluindo que o sonho de Vannevar Bush poderia ser alcançado em apenas um ciclo de vida de 65 anos, modificando o papel do bibliotecário, que num mar de informações, deixaria de ser o fornecedor de água para navegar o barco (LESK, 1996).

A descrição de cada uma dessas fases é apresentada no Quadro 1, a seguir:

Quadro 1 - Fases relacionais entre etapas da vida com a história da RI

Fases	Período	Descrição
Infância	1945 a 1955	Período em que surgiram os primeiros sistemas, como os índices KWIC - concordâncias usadas para recuperação de informações, por pesquisadores como Luhn. Também houve tentativas inovadoras de tipos alternativos de maquinário, como o uso de códigos sobrepostos em cartões com entalhes nas bordas de Calvin Mooers. A peça de equipamento mais famosa deste período foi o *WRU Searching Selector*, construído por Allen Kent na Western Reserve University.
Escolaridade	Anos 60	Embora sem muita aplicação prática, correspondeu a tempos de experimentação e aprendizagem sendo uma época de forte crescimento para a recuperação de informação. Os avanços tecnológicos contribuíram bastante com o desenvolvimento de grandes sistemas de informação como *Dialog* e *BRS* com destaque para temas e conceitos avançados como: bases de dados, pesquisa em texto-livre, surgimento dos 'thesauros', técnicas de avaliação (revocação e precisão), coleções para testes, '*relevance feedback*', recuperação multilíngue e linguagem natural.

Fases	Período	Descrição
Idade adulta	Anos 70	Tem-se o aparecimento de grandes quantidades de texto em forma legível por máquina e os sistemas de compartilhamento de tempo aceleraram a utilização prática dos sistemas. Devido a promessas não concretizadas e à perda de importância dentro dos departamentos, houve um declínio na pesquisa. O progresso deu-se, sobretudo na área da recuperação probabilística. Também nesta época, aconteceu a segregação entre a Recuperação da Informação (RI) e a Inteligência Artificial (IA), que procurou ampliar o escopo da pesquisa e focar aspectos mais fundamentais.
Maturidade	Anos 80	Surge com cada vez mais informação disponível em formato computadorizado. Deu-se uma dispersão da recuperação online pelo grande público, nomeadamente com os catálogos, revistas e jornais. Deu-se a popularização do CD-ROM, que encaixava perfeitamente com o modelo tradicional de publicação e distribuição de informação. Houve um enorme aumento das bases de dados online tanto em número quanto em variedade. A pesquisa reapareceu, mas as técnicas desenvolvidas não eram utilizadas na indústria.
Crise de meia idade	Anos 90	Nesta época, cada vez mais informação estava disponível online e acessível via algoritmos de pesquisa em texto integral. Contudo, esta era uma área que interessava quase unicamente a um grupo específico de especialistas. Os esforços em disseminar o uso pela população em geral tinham tido pouco sucesso. Com a revolução da Internet, a informação passa a ser distribuída por qualquer pessoa e não apenas pelas grandes empresas comerciais, através da proliferação das páginas pessoais. Lesk relaciona esta etapa com a visão de Bush, em que cada pessoa organiza a sua informação pessoal e a troca com outros. A disponibilização de imagens, com o aparecimento do navegador e do scanner, abre um novo conjunto de possibilidades. Começam a aparecer jornais disponíveis unicamente online. Os pesquisadores veem a aplicação prática de muitas das técnicas desenvolvidas e dá-se um forte avanço na área da avaliação dos sistemas.

Fases	Período	Descrição
Realização	Ano 2000	É previsto que as questões em sua maioria poderiam ser respondidas com acesso a materiais de referência online. Prevê a existência de consultorias online que ajudariam os usuários a encontrarem o que querem. Aponta como caminho para a pesquisa, o desenvolvimento de técnicas de pesquisa baseadas em imagens, sons e vídeo. No entanto, alerta que seria necessário um novo ciclo de redução de custos de armazenamento para que sejam viáveis as verdadeiras bibliotecas de vídeo digitalizado.
Reforma	Atualidade	65 anos após a publicação do artigo de Bush, a Recuperação da Informação (RI) atinge a **reforma**. O autor prevê que nesse período, o trabalho de conversão para o formato máquina esteja praticamente terminado. Dentre outras previsões destaca-se a de que a informação multimídia seria de acesso tão descomplicado como o texto, que a sintetização de voz seria o meio mais usado para acesso a conteúdo e enfatiza a importância de se abordar a internacionalização, resultante da Internet.

Fonte: Adaptado de Lesk (1996)

No final da Segunda Guerra Mundial, o presidente americano Fanklin Roosevelt solicitou recomendações sobre tecnologias aprendidas durante a guerra a Vannevar Bush, que ocupava uma alta posição no governo.

Bush inicialmente produziu um relatório intitulado Science, *The Endless Frontier* (Ciência, a fronteira sem fim) e, em seguida, ele escreveu o artigo *As We May Think* (Como podemos pensar), que discutia novas peças de hardware e software que poderiam ser inventadas nos anos seguintes (BAEZA-YATES; RIBEIRO-NETO, 2013).

De acordo com Bush (1945, p. 124, tradução nossa), “formas de enciclopédias inteiramente novas irão aparecer, contendo uma malha de trilhas associativas, prontas para serem colocadas no memex e lá ampliadas”.

À época, o texto *As We May Think* influenciou pessoas como Douglas Englebert, que, na *Fall Joint Computer Conference* ocorrida em São Francisco, em dezembro de 1968, fez uma demonstração na qual apresentou o primeiro mouse de computador, a videoconferência, a teleconferência e o hipertexto (BAEZA-YATES; RIBEIRO-NETO, 2013).

A partir de então novas tecnologias seriam desenvolvidas para ampliar e melhorar os processos de armazenamento e recuperação da informação, que será abordado no tópico a seguir.

2.2 Fundamentos da recuperação da informação

O volume de documentos e informações resultante do pós-Segunda Guerra se tornou um ponto chave que motivou diferentes campos da ciência, sobretudo a Ciência da Informação, a desenvolver métodos e técnicas mais eficientes de armazenamento, disseminação e recuperação da informação (BARRETO, 2002).

As tecnologias disponíveis, até então, tentavam solucionar a necessidade de armazenar grandes volumes de informações provenientes, sem se preocupar, em primeiro momento, com a organização do conteúdo e nem com a forma que tais informações fossem recuperadas.

A Ciência da Informação sempre teve papel importante nos processos de tratamento informacional e, diante do aumento exponencial de documentos nos últimos anos “[...] passou a ser uma instituição de reflexão da informação, como um campo, que estuda a ação mediadora entre informação e conhecimento acontecido no indivíduo” (BARRETO, 2002, p. 70).

Nesse cenário, os processos relativos ao controle dos ativos informacionais, estão diretamente conectados a necessidade de transformá-los em conhecimentos que, em última instancia, darão suporte para solução de demandas do nosso meio.

Barreto (2002) destaca que, a disseminação do conhecimento sempre foi uma necessidade humana, que ao longo das décadas buscou encontrar meios de tornar esse processo uma tarefa mais eficiente.

Dentre as inovações mostradas, a de maior interesse aqui é o hipertexto, que permite que o leitor salte de um documento eletrônico para outro, o que foi uma característica importante para endereçar problemas de compartilhamento de documentos enfrentados por Tim Berners-Lee em 1989, onde o mesmo observou que o hipertexto em rede, por meio da Internet, seria uma boa solução (BAEZA-YATES; RIBEIRO-NETO, 2013).

Tim Berners-Lee, começou então a trabalhar na implementação do hipertexto, culminado, em 1990, na escrita do protocolo HTTP, definição da linguagem HTML, escrita do primeiro navegador, ao qual chamou de "*World Wide Web*”, e codificação

do primeiro servidor *Web*. A *Web* nasceu em 1991 com a disponibilização, na Internet, do software do seu navegador e servidor (BAEZA-YATES; RIBEIRO-NETO, 2013).

Desde sua criação, a *Web* tornou-se um grande sucesso e mudou o mundo, devido

> [...] a simplicidade da linguagem de marcação HTML, os baixos custos de acesso, a disseminação do alcance da Internet, a interface de navegação interativa e as máquinas de busca, possíveis explicações para esse sucesso, contudo, apesar de terem provido a infraestrutura fundamental para a *Web*, essas tecnologias não são a razão de sua popularidade (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 11).

Segundo os autores, o que marca o nascimento da “Era da Publicação Eletrônica”, é o fato

> [...] das pessoas poderem publicar suas ideias e alcançar milhões de leitores da noite para o dia, sem pagar nada e sem precisar convencer uma grande editora. Isso significa que as restrições impostas pelas companhias de comunicação em massa e pelas barreiras geográficas naturais foram quase totalmente removidas pela invenção da *Web* (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 12).

A partir da rápida ascensão da *internet*, o termo Recuperação da Informação (RI), definido por Mooers em 1951, ganhou popularidade no meio acadêmico a partir de 1961, cuja objetivo principal era se ocupar de questões relacionadas à descrição da informação, especificar os termos de buscas e a localização da informação a partir do uso de técnicas e sistemas específicos para esse fim (MOOERS, 1951).

Para Chowdhury (2010, p. 1, tradução nossa), outro fator que deve ser levado em consideração é

> [...] a função organizadora da recuperação da informação que era vista, à época, como um grande avanço nas atividades das bibliotecas que não eram mais consideradas apenas depósitos de livros, mas como locais onde as informações ali armazenadas eram catalogadas e indexadas para facilitar o acesso.

No contexto da Ciência da Informação, Ferneda (2003, p. 14), pontua que

> [...] o termo “recuperação de informação” significa, para uns, a operação pela qual se seleciona documentos, a partir do acervo, em função da demanda do usuário. Para outros, “recuperação de

> informação” consiste no fornecimento, a partir de uma demanda definida pelo usuário, dos elementos de informação documentária correspondentes.

A Recuperação de Informação (RI) é “uma área abrangente da Ciência da Computação que se concentra principalmente em prover aos usuários o acesso fácil às informações de seu interesse” (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 1).

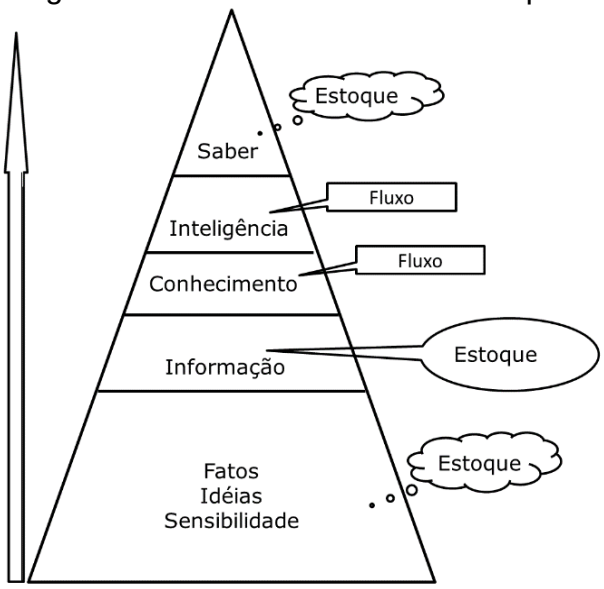
Tal mecanismo “possibilita que as pessoas interajam com um sistema ou serviço para encontrar informações - texto, imagens gráficas, gravações de som ou vídeo que atendam às suas necessidades específicas” (CHOWDHURY, 2010, p. 1, tradução nossa).

Os índices - estruturas de dados especializadas que possibilitam agilidade nas buscas - estão no cerne de todos os sistemas modernos de RI, de acordo com Baeza-Yates e Ribeiro-Neto (2013), uma vez que fornecem acesso rápido aos dados e aceleram o processamento das consultas.

Cada categoria no índice é composta tipicamente por rótulos, que identificam seus tópicos associados, e por ponteiros, para os documentos que discutem tais tópicos.

Com a finalidade de diferenciar na condição da informação seus estoques e fluxos, Barreto (2002) propõe uma estrutura piramidal passando uma qualificação, de valor subjetivo, em que menos é mais referente a quantidade maior na base e menor no topo, conforme pode ser observado na Figura 1.

Figura 1- Pirâmide de fluxos e estoques



Fonte: Barreto (2002, p. 68)

Santos e Sant’Ana (2002) apontam que, o conhecimento pode ser entendido como conhecimento tácito e conhecimento explícito sendo que o conhecimento tácito é aquele que não tem como ser totalmente convertido em um conjunto de códigos ou sinais, não permitindo sua transmissão completa e, portanto, não podendo ser registrado em meios artificiais, tendo sua existência, ligada diretamente às pessoas que o detêm.

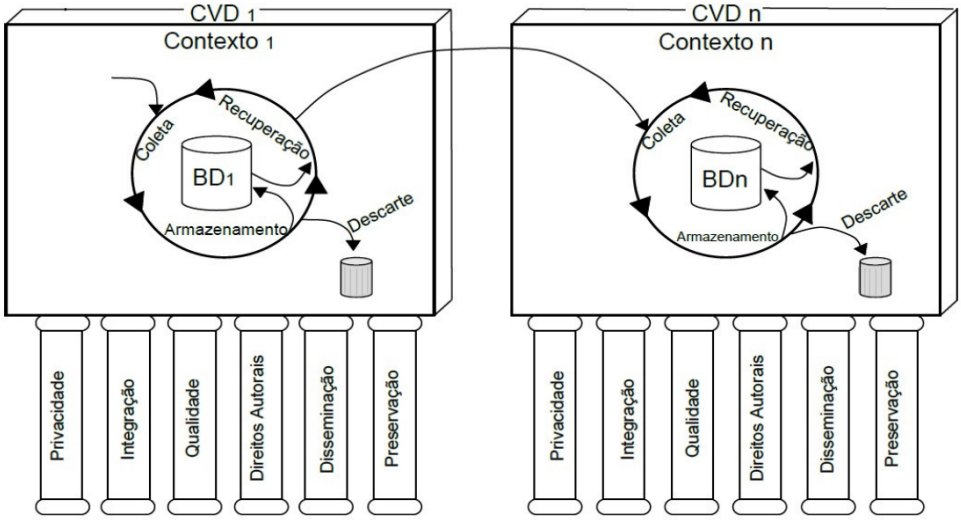
Todavia, o conhecimento explícito corresponde à parte do conhecimento que pode ser formalmente articulada de maneira mais precisa, uma vez que, pode ser convertido em um conjunto de informações e, portanto, ser transmitido e consequentemente, registrado em suportes artificiais (SANTOS; SANT’ANA, 2002).

Sendo a recuperação, uma das etapas do Ciclo de Vida dos Dados (CVD), conforme modelo proposto por Sant’Ana (2016), faz-se necessário uma visão geral desses conceitos para que possamos devidamente enquadrar a presente pesquisa nesse contexto.

Tendo os próprios dados como elemento central, Sant’Ana (2016), lança um novo olhar para o CVD, onde a Ciência da Informação (CI), contribui para o esse processo de maneira abrangente, ampliando as pontes entre os usuários e os dados que necessitam.

Na Figura 2, observa-se que o autor compõe este ciclo apresentando as fases - coleta, armazenamento, recuperação e descarte - permeadas por fatores transversais e presentes em todas elas que são privacidade, integração, qualidade, direito autoral, disseminação e preservação (SANT’ANA, 2016).

Figura 2 - Ciclo de Vida dos Dados para Ciência da Informação



Fonte: Sant’Ana (2016, p. 123)

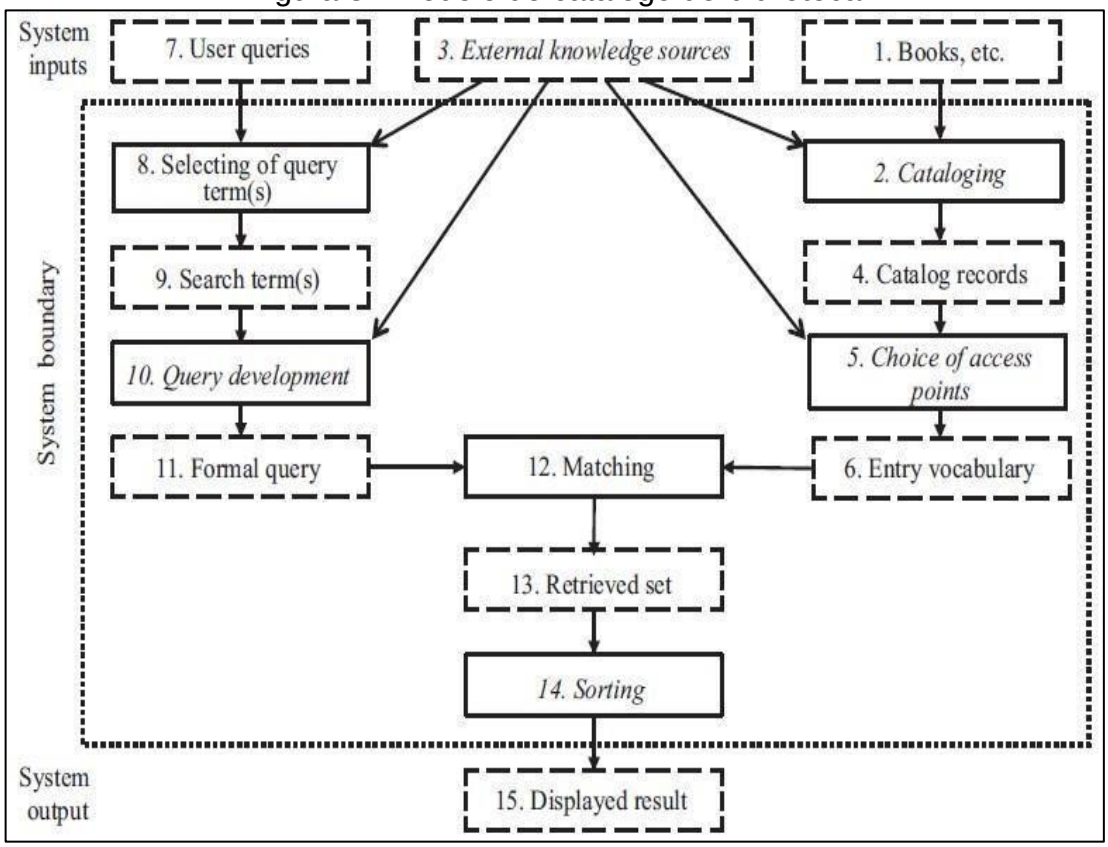
Adotando como base o modelo da Figura 2, trataremos neste trabalho, da recuperação de dados previamente coletados, armazenados e disponibilizados para acesso e uso, o que, sob a perspectiva dos usuários que acessarão estes dados, se configura a fase de coleta.

Neste mesmo ponto, quando tomamos como referência, o ponto de vista dos que são responsáveis pela manutenção e disponibilização desses dados para coleta, estamos na fase de recuperação, uma vez que está se levando em consideração como foco, a base de dados (SANT'ANA, 2016).

Recuperação e coleta são, portanto, faces de uma mesma moeda e a perspectiva de onde se observa o fluxo de dados é fundamental para classificar a correta etapa no CVD.

No âmbito das bibliotecas online, Buckland (2017) ilustra o funcionamento de um sistema onde um banco de dados altamente estruturado retorna como saída o resultado de uma pesquisa. Na Figura 3, é ilustrado um modelo de um catálogo de biblioteca.

Figura 3 - Modelo de catálogo de biblioteca



Fonte: Buckland (2017, p. 185)

Na estrutura apresentada na Figura 3, as caixas sólidas contêm processos estruturados, já consolidados em um sistema informatizado. As caixas tracejadas contêm registros, consultas ou registros de catálogo. Itálico mostra componentes opcionais e setas indicam fluxos. Os documentos a catalogar são apresentados no canto superior direito (caixa 1). Um processo de catalogação (caixa 2) pode basear-se em regras de catalogação, vocabulários padrão e cópia do catálogo de outro lugar (caixa 3) e resultar em um conjunto de registros de catálogo (caixa 4).

Na prática, nem todas as partes dos registros do catálogo são pesquisáveis, então outro processo determina a escolha dos pontos de acesso (caixa 5), resultando no conjunto pesquisável de entradas de índice, também conhecido como vocabulário de entrada (caixa 6).

Os usuários da biblioteca têm suas consultas (caixa 7) e a expressão dessas consultas precisa ser adaptada à terminologia do sistema de recuperação (caixa 8) para selecionar um ou mais termos de pesquisa aceitáveis (caixa 9), que podem então ser formulados (caixa 10) em uma consulta formal (caixa 11) para correspondência (caixa 12) com termos pesquisáveis - “vocabulário de entrada” (caixa 6) - para derivar um resultado de pesquisa ou “conjunto recuperado” (caixa 13).

Normalmente, o conjunto inicialmente recuperado é classificado (caixa 14) para exibição do resultado da pesquisa (caixa 15).

Durante séculos, os índices foram criados de forma manual como conjuntos de categorias, porém, “o advento dos computadores modernos possibilitou a construção automática de índices de tamanho grande, o que acelerou o desenvolvimento da área de RI” (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 2).

Segundo Chowdhury (2010, p. 1, tradução nossa),

> [...] a partir do uso de computadores com grande capacidade de armazenamento e manuseio mais eficiente da informação, começam a surgir os bancos de dados contendo detalhes bibliográficos de documentos, partes de resumos e palavras-chave, fazendo com que o conceito de recuperação de informação passasse a significar a recuperação de informação bibliográfica de bancos de dados de documentos armazenados.

A introdução desses dispositivos na recuperação da informação, nas últimas décadas, e mais ainda, desde o advento da internet e da *Web*, os aspectos técnicos,

bem como os sociais e humanos dos sistemas de recuperação de informação, têm sido significativamente influenciados por fatores externos, incluindo desenvolvimentos nacionais e globais em tecnologia, regulamentos e economia (CHOWDHURY, 2010).

A Recuperação da Informação trata

> [...] da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas *Web*, catálogos online, registros estruturados e semiestruturados, objetos multimídia, etc. A representação e a organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 1).

Para Chowdhury (2010), um sistema de recuperação eficiente se preocupa com todas as atividades relacionadas à organização, processamento e acesso às informações em todas as formas e formatos, possibilitando ao usuário recuperar informações que atendam suas necessidades.

Tais prerrogativas desse processo são corroboradas por Souza (2006, p. 163), que destaca que tais definições

> [...] procuram apreender um fenômeno atemporal – as necessidades de informação – e as várias metodologias e tecnologias que, através dos tempos, foram engendradas para atender a essas necessidades, desde as atividades de organização de coleções de documentos em acervos bibliográficos, até os modernos sistemas informatizados que lidam com documentos em formato digital.

Atualmente, a pesquisa em RI inclui modelagem, classificação de textos, arquitetura de sistemas, interfaces de usuário, visualização de dados, filtragem e linguagens indo muito além dos seus objetivos iniciais de indexação de textos e de busca por documentos úteis em uma coleção levando a uma ampliação do escopo da área (BAEZA-YATES; RIBEIRO-NETO, 2013).

As ferramentas e sistemas de recuperação de informações emergiram para suprir a demanda de recuperar as informações, que são criadas continuamente de maneira cada vez mais intensas e velozes, nos espaços que se criaram no ambiente on-line, para os quais se verificou uma migração dos repositórios de informações geradas, no desempenho das inúmeras atividades humanas (SOUZA, 2006).

Em termos de pesquisa, a área pode ser estudada sob dois pontos de vista distintos e complementares: um centrado no computador e o outro centrado no usuário.

> Na visão centrada no computador, a RI consiste principalmente na construção de índices eficientes, no processamento de consultas com alto desempenho e no desenvolvimento de algoritmos de ranqueamento, a fim de melhorar os resultados. Na visão centrada no usuário, a RI consiste principalmente em estudar o comportamento do usuário, entender suas principais necessidades e determinar como esse entendimento afeta a organização e a operação do sistema de recuperação (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 1).

Os sistemas de recuperação de informações “são em sua maior parte, sistemas de recuperação de documentos, pois são projetados para recuperar informações sobre a existência (ou não) de documentos relevante para uma consulta do usuário” (CHOWDHURY, 2010, p. 1, tradução nossa).

Lancaster (1968) comenta que um sistema de recuperação de informação não informa (altera o conhecimento do) usuário sobre o assunto de sua consulta, mas apenas informa sobre a existência (ou inexistência) e o paradeiro de documentos relacionados à sua solicitação.

> Todavia, essa noção de recuperação de informação mudou desde a disponibilidade de documentos de texto completos em bancos de dados bibliográficos onde informações de itens bibliográficos ou o texto exato podem ser recuperadas caso correspondam aos critérios de pesquisa de um usuário em um banco de dados armazenado de textos completos de documentos (CHOWDHURY, 2010, p. 1, tradução nossa).

Segundo Chowdhury (2010), a recuperação de informação moderna lida com armazenamento, organização e acesso ao texto, bem como recursos de informação multimídia e, muito embora originalmente, sistemas de recuperação de informação significavam sistemas de recuperação de texto, uma vez que lidavam com documentos textuais, muitos sistemas modernos de recuperação de informação lidam com informação multimídia, compreendendo texto, áudio, imagens e vídeo e, embora com recursos aplicáveis a ambas modalidades, devido à natureza específica das informações de áudio, imagem e vídeo, houve o desenvolvimento de muitas novas ferramentas e técnicas de recuperação da informação.

Um sistema de recuperação de informações, segundo Chowdhury (2010, p. 1, tradução nossa), “visa coletar e organizar a informação em uma ou mais áreas temáticas sendo projetado para recuperar os documentos ou informações exigidas pela comunidade de usuários e deve disponibilizar as informações corretas para o usuário correto”.

Nesse contexto, é estabelecido que “o objetivo de um sistema de recuperação de informação é permitir que os usuários encontrem informações relevantes a partir de uma coleção organizada de documentos” (CHOWDHURY, 2010, p. 1, tradução nossa).

Segundo Baeza-Yates e Ribeiro-Neto (2013, p. 5), o usuário de um sistema de RI precisa

> [...] traduzir sua necessidade de informação em uma consulta na linguagem fornecida pelo sistema. Em um sistema de RI, como uma máquina de busca, isso geralmente implica na especificação de um conjunto de palavras que transmitam a semântica da necessidade de informação. Dizemos que o usuário está buscando ou consultando informações de seu interesse. Enquanto a busca por informações de interesse é a principal tarefa de recuperação na *Web*, a busca também pode ser utilizada para satisfazer outras necessidades do usuário, como a compra de produtos e a realização de reservas.

Baeza-Yates e Ribeiro-Neto (2013), observam que, usuários de sistemas modernos de RI, como os de máquinas de busca, necessitam de informação de diferentes complexidades e, por vezes, a descrição dada pelo usuário não provê a melhor redação de consulta para o sistema de RI. Logo, o usuário pode sumarizar sua necessidade de informação em uma consulta ou em uma sequência de consultas a serem submetidas ao sistema gerando uma série de palavras-chave, ou termos de indexação. Por este ponto de vista, o objetivo maior do sistema de RI é recuperar informações que sejam úteis ou relevantes para o usuário, estando a ênfase na recuperação de informação ao invés da recuperação de dados.

Souza (2006) enfatiza que, muito embora, apesar das diferenças entre os sistemas de recuperação de informações e sistemas de recuperação de dados, sendo ambos os termos utilizados para sistemas, há que se distinguirem os Sistemas de Recuperação de Informação (SRI) dos Sistemas de Gerenciamento de Bancos de Dados (SGBD).

Os sistemas convencionais de gerenciamento de banco de dados, de acordo com Chowdhury (2010, p. 2, tradução nossa), “como Access, Oracle, MySQL e assim por diante, lidam com dados estruturados, onde a organização ou estruturação dos dados ocorre dependendo dos atributos específicos dos elementos de dados”.

No que tange a recuperação de informação e a recuperação de dados, Baeza-Yates e Ribeiro-Neto (2013, p. 6), explicitam que “a recuperação de dados representa uma solução para o usuário de um sistema de banco de dados, contudo não resolve o problema de recuperar informações sobre um assunto ou tópico”. O objetivo particular desses bancos de dados é permitir ao usuário pesquisar registros específicos que atendam a uma ou mais condições ou critérios de pesquisa específicos (CHOWDHURY, 2010). Assim, dados são caracterizados como sequências de símbolos, para os quais são conferidos significados, sendo passíveis de codificação, interpretação e manipulação por programas de computador, podendo ser enviados através de redes e dispositivos de comunicação (SOUZA, 2006).

> Em sistemas gerenciadores de bancos de dados, os símbolos são armazenados em uma estrutura matricial em campos determinados, com metadados que lhes conferem certo sentido ontológico. Para recuperar dados específicos, basta especificar as restrições necessárias aos campos de pesquisa e codificá-las numa questão ou query (argumento de entrada no sistema) para que se tenha a resposta exata, fruto de busca completa e exaustiva (SOUZA, 2006, p. 163).

O usuário no contexto de um sistema de RI está mais interessado em recuperar informações sobre um assunto do que em recuperar dados que satisfazem sua expressão de busca, porém, a recuperação de dados, consiste na identificação de quais documentos da coleção contêm as palavras-chave da consulta do usuário, fazendo com que, frequentemente, o resultado não seja suficiente para satisfazer a necessidade de informação do usuário (BAEZA-YATES; RIBEIRO-NETO, 2013). Essa característica é o que, segundo Ferneda (2003, p. 15), diferencia os Sistemas de Recuperação de Informação (SRI) dos Sistemas de Gerenciamento de Bancos de Dados (SGBD), sendo que “a principal diferença está na natureza dos objetos tratados por estes dois tipos de sistema”.

O conceito de informação, para Souza (2006, p. 163), “carrega um grau maior de abstração e não prescinde do sujeito que a depreenda a partir dos dados, no ato conhecido como interpretação”.

> Ao contrário de um sistema de gerenciamento de banco de dados convencional, um sistema de recuperação de informações é projetado para lidar com dados não estruturados, sendo o principal objetivo de um sistema de recuperação de informações, recuperar as informações - sejam as informações reais ou os documentos que as contêm - que correspondam total ou parcialmente à consulta do usuário (CHOWDHURY, 2010, p. 2, tradução nossa).

Em um sistema de RI, que lida com texto em linguagem natural que não é bem estruturado, os objetos recuperados podem ser inexatos e pequenos erros podem passar despercebidos, o que não é aceitável em um sistema de recuperação de dados, como banco de dados relacional tratando-se de dados que possuem estrutura e semântica bem definidas, onde um único objeto incorreto em meio a milhares de objetos recuperados significa falha total (BAEZA-YATES; RIBEIRO-NETO, 2013).

Nessa perspectiva, Souza (2006, p. 163), lança uma importante reflexão:

> [...] no sentido estrito do conceito, nenhum programa de computador lida, sob o ponto de vista da máquina, com informações, a não ser que possua alguma capacidade de arrazoamento, e, assim mesmo, a utilização do termo dá margem a discussões.

Assim, com as necessidades do usuário em foco, o problema da RI é então definido:

> [...] o objetivo principal de um sistema de RI é recuperar todos os documentos que são relevantes à necessidade de informação do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 4).

“Alguns pesquisadores comentam que a recuperação de informação é um processo de comunicação uma vez que serve como uma ponte entre o mundo dos criadores ou geradores de informação e os usuários dessa informação” (CHOWDHURY, 2010, p. 6, tradução nossa).

A fim de ser efetivo em sua tentativa de satisfazer a necessidade de informação do usuário, Baeza-Yates e Ribeiro-Neto (2013), atestam que o sistema de RI 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. A

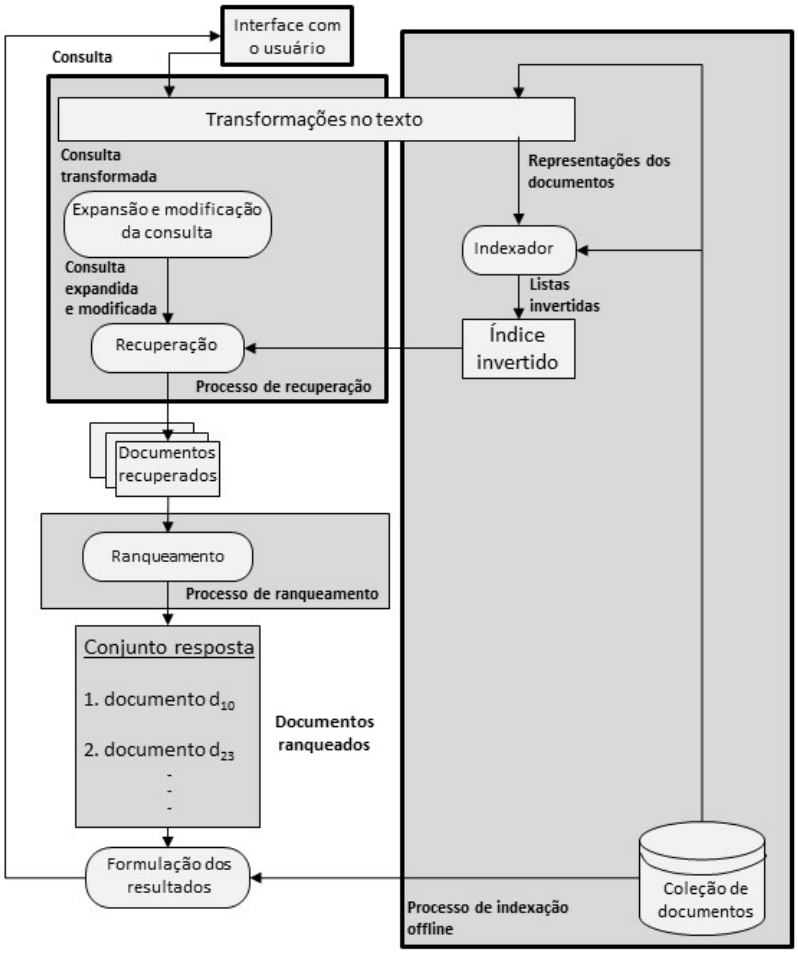
dificuldade não só está em saber como extrair a informação dos documentos, mas também como utilizá-la para decidir quanto à sua relevância que, segundo os autores, tem um papel central em RI.

Sendo assim, o sucesso de um Sistema de Recuperação de Informação (SRI) depende muito do julgamento do usuário sobre a relevância dos documentos recuperados para sua consulta, tornando as necessidades de informação vis-à-vis, formulação de consulta e julgamento de relevância uma área importante de estudo e pesquisa em recuperação de informação (CHOWDHURY, 2010).

> Logo, nenhum sistema de RI pode fornecer respostas perfeitas a todos os usuários continuadamente uma vez que a relevância é um julgamento pessoal que depende da tarefa a ser resolvida e de seu contexto, podendo mudar com o tempo (por exemplo, à medida que novas informações tornam-se disponíveis), com o local (por exemplo, a resposta mais relevante é a mais próxima), ou até mesmo com o dispositivo (por exemplo, a melhor resposta é um documento pequeno que seja mais fácil de baixar e visualizar) (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 4).

Para elucidar este ponto, faz-se necessário a introdução do conceito de ranqueamento e sua posição no contexto de um software de RI pode ser visualizada na Figura 4.

Figura 4 - Os processos de indexação, recuperação e ranqueamento



Fonte: Baeza-Yates e Ribeiro-Neto (2013, p. 9)

Na Figura 4, é demonstrado os processos de recuperação e ranqueamento, sendo que o procedimento de avaliação mais comum consiste em comparar o conjunto de resultados produzidos pelo sistema de RI com os resultados sugeridos pelos especialistas humanos, e afim de melhorar o *ranking*, podemos coletar *feedback* dos usuários e utilizar essas informações para alterar os resultados.

Na *Web*, a forma mais abundante de *feedback* são os cliques nos documentos do conjunto de resultados. Outra importante fonte de informações para o ranqueamento na *Web* são os *hiperlinks* entre as páginas que permitem a identificação de *sites* de maior autoridade.

O ranqueamento tem por finalidade identificar os documentos com maior possibilidade de serem considerados relevantes pelo usuário, configurando-se como a parte mais crítica de um sistema de RI (BAEZA-YATES; RIBEIRO-NETO, 2013).

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).

> Logo, avaliar a qualidade do conjunto-resposta é a chave para a melhoria do sistema de RI e através de um processo de avaliação sistemático pode-se sintonizar o algoritmo de ranqueamento e melhorar a qualidade dos resultados (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 7).

Sobre os documentos da coleção, primeiro aplicamos operações textuais como a eliminação de *stopwords*, radicalização (*stemming*) e a seleção de um subconjunto de termos para serem utilizados como termos de indexação.

Os termos de indexação são utilizados para compor a representação do documento, que pode ser menor do que o documento original (dependendo do subconjunto de termos de indexação selecionado).

A partir das representações dos documentos, é necessário criar um índice do texto. Diferentes estruturas de dados podem ser utilizadas, todavia, a mais popular é o índice invertido.

As etapas necessárias à geração do índice compõem o processo de indexação e precisam ser executadas *offline*, antes que o sistema esteja pronto para receber quaisquer consultas. Os recursos (tempo e espaço de armazenamento) necessários ao processo de indexação são amortizados pelas várias consultas feitas ao sistema de RI.

Os documentos no topo do *ranking* são então formatados e apresentados para o usuário. A formatação consiste em recuperar o título dos documentos e gerar *snippets* do texto, isto é, trechos que contenham os termos da consulta que serão mostrados ao usuário.

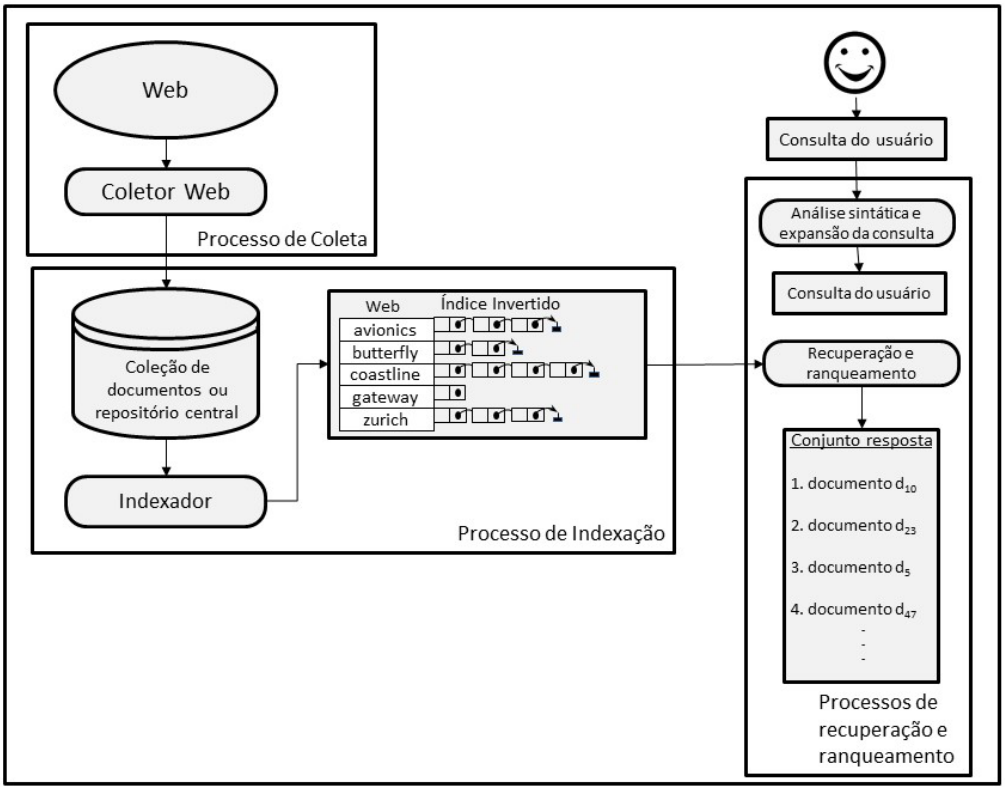
> Um sistema pode ser definido como um conjunto de componentes em interação, sob controle humano, operando em conjunto para atingir um propósito pretendido. Assim, um sistema realiza o processamento das entradas para produzir as saídas necessárias; os agentes desse processamento são pessoas e máquinas. O projeto do sistema pode ser visto como uma série de escolhas a partir das quais o projetista seleciona cada elemento e tenta ajustá-lo ao objetivo proposto do sistema (CHOWDHURY, 2010, p. 10, tradução nossa).

Liston e Schoene (1978, *apud* Chowdhury, 2010, p. 10, tradução nossa) sugerem que

> [...] um sistema eficaz de recuperação de informações deve ter estruturas para disseminação imediata de informações, filtragem de informações, fornecimento da quantidade certa de informações no momento certo, troca ativa de informações, recebimento de informações na forma desejada, navegação, obter informação de forma económica, literatura atual, disponibilizar acesso a outros sistemas de informação, comunicação interpessoal e oferta de ajuda personalizada. Também falam sobre como um sistema de recuperação de informação deve ser orientado para o usuário, dando ênfase principal à conveniência dos usuários.

Para descrever a arquitetura simples e genérica do software de um sistema de RI, utilizaremos a proposta de (BAEZA-YATES; RIBEIRO-NETO, 2013), conforme pode ser observado na Figura 5.

Figura 5 - Arquitetura de alto nível do software de um sistema RI



Fonte: Baeza-Yates e Ribeiro-Neto (2013, p. 7)

O primeiro passo na instalação de um sistema de RI é obter a coleção de documentos, que pode ser particular ou coletada a partir da *Web*. Um módulo coletor

é responsável pela obtenção dos documentos e um disco geralmente chamado de repositório central armazena a coleção de documentos.

Os documentos no repositório central são indexados para que a recuperação e o ranqueamento sejam efetuados em passo acelerado, neste caso, apresenta-se o índice invertido, forma mais empregada, que contempla todas as palavras distintas da coleção e, para cada palavra, a lista de documentos que a contém.

> Alguns sistemas de recuperação de informação, especialmente algumas ferramentas de pesquisa na *Web*, usam um diretório, que é como uma lista hierárquica de assuntos usada para mapear documentos em uma coleção, e que requerem que os usuários naveguem pelo diretório para identificar um termo ou conceito preferido e sigam os *links* de lá para acessar os documentos mapeados. No entanto, como nos livros, os documentos reais em um sistema de recuperação de informações são mantidos separadamente do índice e é o índice que é usado para uma pesquisa de informações (CHOWDHURY, 2010, p. 5, tradução nossa).

Com a coleção de documentos devidamente indexada, inicia-se o processo de recuperação, que consiste em recuperar documentos que satisfaçam uma consulta do usuário ou um clique em um *hiperlink*.

O usuário especifica uma consulta que reflete sua necessidade de informação, a consulta é analisada sintaticamente e expandida com, por exemplo, variações das palavras da consulta.

Essa consulta expandida é então processada, utilizando-se o índice para recuperar um subconjunto dos documentos. Em seguida, os documentos recuperados são ranqueados e aqueles que estão no topo do ranking são retornados ao usuário, sendo essa a etapa mais crítica, pois a qualidade do resultado percebida pelo usuário é fundamentalmente dependente do ranqueamento (BAEZA-YATES; RIBEIRO-NETO, 2013).

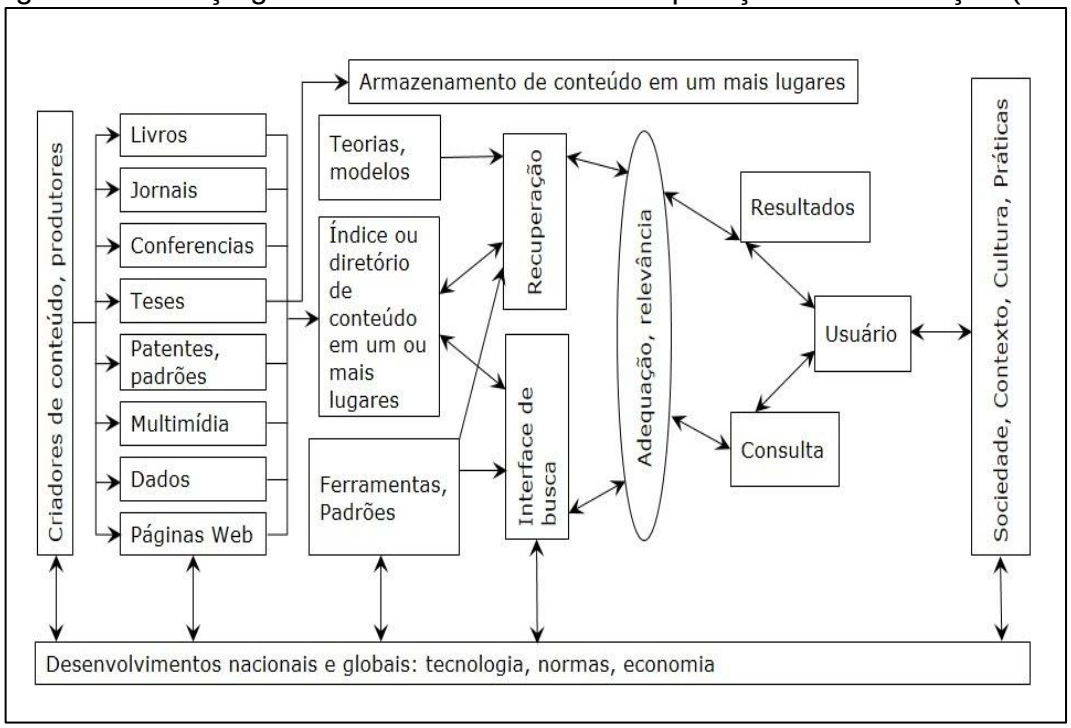
As informações são geralmente recuperadas, “na forma de documentos que contêm as informações necessárias, sempre que os termos da pesquisa correspondem aos termos do índice” (CHOWDHURY, 2010, p. 5, tradução nossa).

> Os usuários interagem com os sistemas de recuperação de informação por meio de uma interface, onde normalmente se espera que expressem suas necessidades de informação na forma de uma consulta, que é apresentada ao sistema de pesquisa por meio de uma expressão de pesquisa que pode conter um ou mais termos de pesquisa apresentados na forma de uma frase em linguagem natural

> ou em uma linguagem natural restrita em que os termos de pesquisa são vinculados a vários operadores de pesquisa (CHOWDHURY, 2010, p. 5, tradução nossa).

Dentre os diversos diagramas que descrevem o processo de recuperação de informações, foi escolhido o proposto por Chowdhury (2010), apresentado na Figura 6, que traz a visão conceitual de um Sistema de Recuperação de Informação (SRI).

Figura 6 - Esboço geral de um Sistema de Recuperação de Informação (SRI)



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

Nele, observa-se que um SRI pode contemplar um ou mais tipos diferentes de documentos e pode conter texto, bem como informações de multimídia. Todos os documentos são processados para criar um índice, que é pesquisado para recuperação de informações, sendo em sua forma mais simples, considerado como um índice de livro, mas na realidade é muito mais complexo do que isso.

Criar um índice, como podemos ver na Figura 6, é um processo bastante complexo, e várias ferramentas, técnicas e padrões são usados para esse propósito, sendo que, os documentos e o índice podem estar localizados em um ou mais locais para facilitar o acesso rápido e a fácil manutenção do documento e das bases de dados do índice.

Como pode ser observado na Figura 6, os processos de busca e recuperação de informações são muito influenciados pelos conceitos de adequação e relevância e

um dos problemas mais comuns enfrentados pelos SRI, com base na formulação de consultas, é que muitas vezes os usuários não podem 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.

Fatores muito importantes, que estão fora do domínio de um SRI específico, como sociedade, contexto, cultura e práticas, afetam os usuários e suas atividades de busca e recuperação de informações. Isso é traduzido na caixa à extrema direita da Figura 6 que mostra a influência de tais fatores nos usuários, suas necessidades de informações, comportamento de busca de informação, julgamentos de relevância e assim por diante.

Todas as tarefas mencionadas na Figura 6 podem ser agrupadas em dois grupos principais:

- Análise de assunto ou conteúdo - inclui as tarefas relacionadas à análise, organização e armazenamento de informações.
- Processo de busca e recuperação - inclui as tarefas de análise das consultas dos usuários, criação de uma fórmula de busca, a busca real e recuperação de informações.

> Um Sistema de Recuperação de Informação (SRI) lida com várias fontes de informação, por um lado, e os requisitos dos usuários, por outro, devendo, analisar o conteúdo das fontes de informação, bem como as consultas dos usuários, e então, compará-los para recuperar os itens que são relevantes (CHOWDHURY, 2010, p. 6, tradução nossa).

Segundo, Lancaster (1969, *apud* Chowdhury, 2010, p. 7, tradução nossa), um SRI “tem seis subsistemas principais: o subsistema de documentos, o subsistema de indexação, o subsistema de vocabulário, o subsistema de busca, a interface do sistema de usuário, o subsistema correspondente”.

Chowdhury (2010) pontua que os SRI podem ser categorizados de várias maneiras, sendo uma delas, agrupá-los em duas categorias: internos e online.

- Os sistemas internos de recuperação de informações são configurados por uma biblioteca ou centro de informações específico para atender principalmente os usuários da organização. Um tipo específico de banco de dados interno é o catálogo da biblioteca. Os OPACs fornecem facilidades para os usuários da

biblioteca realizarem pesquisas on-line no catálogo e, em seguida, verificarem a disponibilidade do item necessário.

- Por sistemas de recuperação de informações online, entendemos aqueles que foram projetados para fornecer acesso a um(s) banco(s) de dados remoto(s) para uma variedade de usuários. Esses são serviços principalmente comerciais e há vários fornecedores que os atendem.

Com o desenvolvimento da tecnologia de armazenamento óptico, outro tipo de SRI apareceu em *Compact-Disc Read-Only Memory* (CD-ROM), cujas técnicas básicas de pesquisa e recuperação de informações se assemelham as dos sistemas *on-line*, exceto pela vinculação dos usuários estarem a distância por meio da rede de comunicação eletrônica.

Outro, e talvez mais apropriado, agrupamento, segundo Chowdhury (2010), poderia ser feito com base no conteúdo, propósito e funções dos sistemas de recuperação de informação, e nesta abordagem, o autor classifica-os em quatro tipos distintos:

- **OPACs** - demonstram algumas características típicas e limitadas dos sistemas de recuperação de informações, que são projetados para manter em vista a natureza dos documentos que manipulam, bem como o usuário e sua finalidade específica para a pesquisa de informações.
- **Bancos de dados *on-line*** - Os sistemas de recuperação de informações online surgiram no início da era dos aplicativos de computador na recuperação de informações e, nas últimas cinco décadas, esses sistemas passaram por várias melhorias em seus recursos de pesquisa e recuperação. Uma característica única desses sistemas, é que eles são serviços pagos ou por assinatura e fornecem acesso a fontes de informação revisadas por pares e de qualidade, muitas vezes acadêmicas.
- **Bibliotecas digitais e serviços de informação baseados na *Web*** – Compõem-se de bibliotecas digitais e serviços de informação baseados na *Web* que podem ser acessados remotamente por meio de uma interface da *Web*. Esses sistemas de recuperação de informações são diferentes dos sistemas típicos de recuperação de informações *on-line*, descritos no parágrafo anterior, pois geralmente são gratuitos e podem ser acessados por praticamente qualquer pessoa por meio da *Web*.

- **Motores de busca na *Web*** - são os mecanismos de pesquisa da *Web* projetados para fornecer acesso a uma grande quantidade de recursos de informações da *Web*. Esses sistemas de recuperação de informações têm algumas características típicas - eles são robustos, projetados apenas para permitir que os usuários encontrem recursos da *Web*; eles não garantem o acesso aos recursos que recuperam, mas, talvez o mais importante, são gratuitos no ponto de uso.

Um fato extremamente relevante, mudou de uma vez por todas essas percepções e “quase da noite para o dia, a RI ganhou um lugar de destaque junto a outras tecnologias” (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 4).

Segundo Baeza-Yates e Ribeiro-Neto (2013, p. 3),

> [...] apesar de inegavelmente ter atingido um estado de maturidade, até recentemente a RI era vista como uma área de interesse restrita apenas a bibliotecários e a especialistas em informação, tendo essa visão sido predominante por muitos anos, a despeito da rápida disseminação, entre os usuários de computadores pessoais modernos, de ferramentas de RI para aplicações de multimídia e hipertexto.

O surgimento da *Web*, inventada em 1989 por Tim Berners-Lee, através de uma interface de usuário padrão que é sempre a mesma - não importando o ambiente computacional usado para executá-la - possibilitou a criação de bilhões de documentos por milhões de usuários, tornando-se o maior repositório universal da cultura e do conhecimento humano (BAEZA-YATES; RIBEIRO-NETO, 2013).

Nesse sentido, Chowdhury (2010, p. 11, tradução nossa), afirma que “as atividades de pesquisa e desenvolvimento de recuperação de informações avançaram rapidamente nos últimos anos como resultado do surgimento dos mecanismos de pesquisa e da *Web*”.

> [...] a recuperação de informações, agora é usada por todos para acessar informações na *Web* e o crédito deve ir para os motores de busca na *Web* por investir uma quantidade significativa de esforço e recursos disponíveis para desenvolver e melhorar os sistemas de recuperação de informação para que as pessoas possam ter acesso mais fácil e melhor às informações na *Web* Chowdhury (2010, p. 11, tradução nossa).

Encontrar informações úteis na *Web*, de acordo com Baeza-Yates e Ribeiro-Neto (2013, p. 3), “não é sempre uma tarefa simples e, normalmente, requer submissão de uma consulta a uma máquina de busca, a qual Recuperação de Informação tem tudo a ver com RI e suas tecnologias”.

Segundo Chowdhury (2010), tais buscadores na *Web* se tornaram menos complexos e as interfaces de busca de recuperação de informação que antes eram projetadas para usuários experientes e versados se tornaram muito mais simples e intuitivas, podendo ser usadas por qualquer pessoa sem nenhum conhecimento específico de técnicas de recuperação de informação ou domínio do assunto.

Dado o exposto acima, Baeza-Yates e Ribeiro-Neto (2013), concluem que a busca na *Web* é a aplicação mais relevante de RI e suas técnicas, e devido a isso, uma implicação imediata é o grande impacto que ela causou no desenvolvimento da RI, sendo os componentes de recuperação e indexação de qualquer máquina de busca fundamentalmente tecnologias de RI. Os autores elencam tais grandes impactos sobre a busca na *Web* em 5 categorias:

- **Primeiro** - está relacionado a natureza da coleção de documentos (ou páginas) na *Web*, distribuídos por milhões de *sites* conectados por *hiperlinks*, ou seja, elos que associam um trecho de texto em uma página a outras páginas da *Web*, fazendo com que seja necessário coletar os documentos e armazenar cópias deles em um repositório central antes de indexá-los.
- **Segundo** - tocante ao desempenho e escalabilidade dos sistemas de RI, com o crescimento exponencial do volume da coleção e de consultas de usuários submetidas diariamente, tornando tais características de sistemas de RI bastante críticas, muito mais do que eram antes da *Web*.
- **Terceiro** - atinente à dificuldade imposta de se prognosticar a relevância, dado o vasto tamanho da coleção de documentos, fazendo com que documentos que parecem estar relacionados à consulta, mas que, na verdade, não são relevantes, de acordo com o julgamento de uma grande parte dos usuários, sejam recuperados. Isso tem se agravado com o crescimento da *Web*, todavia tal problema está sendo endereçado, com a inclusão de novas fontes de evidência que não estão presentes nas coleções de documentos tradicionais - *hiperlinks* e os cliques dos usuários em documentos do conjunto resposta.
- **Quarto** - decorre do fato da *Web* ser, além de um repositório de documentos, ser um meio para a realização de negócios tornando a procura por informação

textual extrapolada para outras necessidades dos usuários, o que continuamente requer a identificação de dados estruturados associados ao objeto de interesse.

- **O quinto** - decorre do aumento de complexidade na busca por relevância, devido aos anúncios e de outros incentivos econômicos advindos do sucesso da *Web* como uma mídia interativa, os quais levaram também à disponibilidade abusiva de informações comerciais disfarçadas sob a aparência de conteúdo estritamente informacional, o que é denominado *Web* spam.

Assim, uma vez apresentados os conceitos que fundamentam as bases dos SRI, no capítulo seguinte, abordaremos a temática de recuperação de informações voltada mais especificamente para o ambiente *Web*, assim como em formas de se realizar essa atividade de maneira automática, por meio do uso de *Web scrapers*.

3 A TÉCNICA DE *SCRAPING* NA RECUPERAÇÃO DA INFORMAÇÃO

O grande volume de conteúdos disponíveis no ambiente *Web* o torna um terreno fértil para aplicação de tecnologias com grande capacidade na recuperação massiva de informações como *Web scrapers*, sendo essa uma ferramenta importante que pode contribuir nas atividades laborais dos profissionais da informação, no processo de recuperação.

Compreender os desafios e contribuições da aplicação de um *Web scraper* na recuperação da informação 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 os mecanismos de busca na *Web*, diferenças entre *crawlers* versus *scrapers* assim como características intrínsecas da estrutura *Web*, como sua capacidade semântica e distribuição em camadas.

A quantidade de informações na *Web* está crescendo rapidamente, assim como o número de novos usuários inexperientes na arte da pesquisa na *Web*. Dessa forma, os mecanismos de pesquisa automatizados que dependem da correspondência de palavras-chave geralmente retornam muitas correspondências de baixa qualidade (BRIN; PAGE, 1998).

De acordo com Ghosh Dastidar, Banerjee e Sengupta (2016), a geração de dados e a taxa de crescimento dos mesmos é um processo abrupto nos dias de hoje e crescerá exponencialmente a cada dia que passa. Os usuários da *Internet* podem desfrutar de serviços e informações abundantes em *sites* de comércio eletrônico, jornais eletrônicos, blogs e redes sociais.

Segundo Brin e Page (1998), a *Web* cria novos desafios para a recuperação de informações, tornando a criação de um mecanismo de busca que dimensione os eventos, algo complexo, até mesmo nos dias de hoje, tornando certas tarefas cada vez mais difíceis à medida que a *Web* cresce.

Para os autores, os desafios enfrentados nessa nova perspectiva, podem se dar da seguinte forma:

- A tecnologia de rastreamento rápido é necessária para reunir os documentos da *Web* e mantê-los atualizados;

- O espaço de armazenamento deve ser usado de forma eficiente para armazenar índices e, opcionalmente, os próprios documentos;
- O sistema de indexação deve processar centenas de gigabytes de dados com eficiência;
- As consultas devem ser tratadas rapidamente, a uma taxa de centenas a milhares por segundo (BRIN; PAGE, 1998).

Para Upadhyay et al. (2017), um usuário envia uma consulta a um mecanismo de pesquisa e, em seguida, examina os três a quatro *links* principais para satisfazer seus requisitos de informação. Claramente, 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 hoje 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.

Os bancos de dados da *Web*, conforme observam Sufyan, Arjumand e Abdul Qayume (2016) contêm uma grande quantidade de dados estruturados que são facilmente obtidos apenas por meio de suas interfaces de consulta. Os resultados da consulta são apresentados em páginas da *Web* geradas dinamicamente, geralmente na forma de registros de dados, para uso humano.

Muito embora disponíveis para seu consumo pelos usuários, uma grande quantidade de tempo é gasta recuperando essas informações e processando-as. Além disso, o formato dos dados na forma de HTML e outras linguagens da *Web* não são adequados para agentes automatizados e programas de computador (GHOSH DASTIDAR; BANERJEE; SENGUPTA, 2016).

A única opção é copiar e colar manualmente os dados mostrados pelo *site* em um arquivo local em seu computador, sendo este, um trabalho muito tedioso que pode levar muito tempo (SIRISURIYA, 2015).

Quanto a essa abordagem manual, Upadhyay et al. (2017) afirmam não ser escalonável para a grande maioria dos aplicativos da vida real no nível corporativo e organizacional, sendo, portanto, muito interessante, uma estrutura robusta, automatizada e fácil de usar para extrair conteúdo da *Web* com um mínimo de esforço humano.

Nesse cenário, Sufyan, Arjumand e Abdul Qayume (2016, p. 65, tradução nossa)–atestam que, “[...] a extração automática de dados da *Web* é crítica na integração da *Web*”, e o *Web scraping* “é a técnica que visa solucionar esse problema” (SIRISURIYA, 2015, p. 135, tradução nossa).

Todavia, antes de falarmos das técnicas de *Web scraping*, faz-se necessário introduzir o conceito de *Web crawling* e as diferenças entre um e outro.

Usando robôs, conhecidos como *crawlers*, a técnica de *Web crawling* é empregada para indexar as informações no *site*. Essa técnica é essencialmente o que fazem os mecanismos de pesquisa como Google, Bing, Yahoo, agências estatísticas e grandes agregadores online.

Ao rastrear um *site*, tal robô passa por todas as páginas e *links* até a última linha do *site*, procurando qualquer informação. O processo consiste em visualizar uma página como um todo e indexá-la, geralmente capturando informações genéricas.

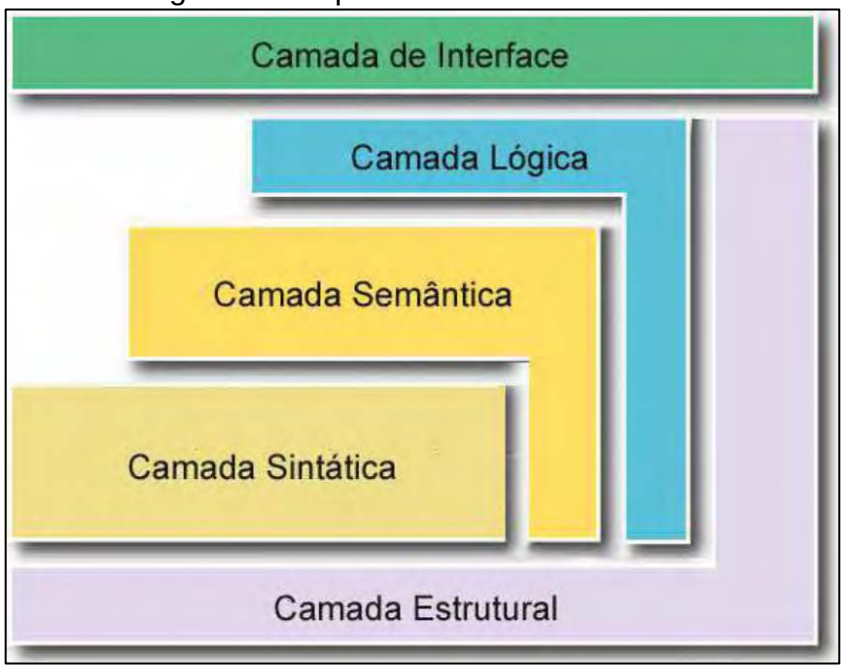
Já a técnica de *Web scraping*, também conhecida como “raspagem” de dados, é semelhante ao *Web crawling*, pois identifica e localiza os dados de destino das páginas da *Web*. É uma maneira de extrair conjuntos de dados usando robôs que também são conhecidos como *scrapers*. Na raspagem sabe-se o exato identificador do conjunto de dados, como por exemplo, elementos fixos da estrutura HTML das páginas da *Web* que estão sendo varridas, das quais os dados precisam ser extraídos.

A principal diferença é que o processo de *Web crawling* geralmente captura informações genéricas, enquanto o *Web scraping* recupera trechos de conjuntos de dados específicos.

A *Web* atual inclui 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 ou *Web Scraping*. Semântica é o estudo do significado das palavras considerado como o componente de sentido e de interpretação de sentenças e enunciados (CARVALHO; CARVALHO, 2018).

Com o intuito de proporcionar uma melhor compreensão e evitar o direcionamento do foco para pormenores técnicos, apresenta-se na Figura 7 uma visão geral da *Web* Semântica.

Figura 7 - Espectro da *Web* Semântica



Fonte: Ramalho (2010, p. 55)

As tecnologias semânticas caracterizam-se como linguagens que possibilitam ir além de representações sintáticas, descrevendo computacionalmente aspectos semânticos dos documentos, dando suporte à utilização de ontologias, que, no âmbito da ciência da informação, pode ser definida como um sistema de representação do conhecimento que possibilita descrever formalmente as propriedades e relacionamentos de um determinado modelo conceitual, favorecendo a realização de inferências automáticas (RAMALHO; OUCHI, 2011).

Para cada camada da *Web* semântica, fornecemos a descrição de sua função de sua função no Quadro 2.

Quadro 2 - Camadas da *Web* semântica e suas funções

Camada	Função
Estrutural	Constitui o alicerce para todas as demais camadas, possibilitando a identificação dos recursos de forma única e padronizada e fornecendo meios seguros para representação, armazenamento e transmissão das informações
Sintática	Fornece meios para a verificação da consistência dos recursos, por meio da definição e validação de regras sintáticas formalmente

	descritas, possibilitando a estruturação dos conteúdos associados a cada recurso
Semântica	Permite a criação de vocabulários para a descrição dos aspectos semânticos dos recursos e a definição das relações existentes entre estes, a partir de especificações formais, explícitas e compartilhadas de conceitos
Lógica	Define regras lógicas que possam ser verificadas computacionalmente, permitindo a realização de inferências automáticas e a verificação do nível de coerência lógica dos recursos
Interface	Possibilita a interligação de todas as camadas anteriores com aplicações desenvolvidas para propósitos específicos, favorecendo maior interoperabilidade e compatibilidade semântica entre sistemas

Fonte: Adaptado de Ramalho e Ouchi (2011)

Assim, 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”.

Nesse sentido, Ghosh Dastidar, Banerjee e Sengupta (2016, p. 26, tradução nossa) enfatizam que “[...] para extrair, categorizar, classificar e indexar as informações mais legítimas disponíveis na *Web* mundial de hoje para o usuário que as solicita, o *Web scraping* é uma ferramenta crucial”.

Para Sirisuriya (2015, p. 135, tradução nossa), “[...] tal técnica é usada para transformar dados não estruturados na *Web* em dados estruturados que podem ser armazenados e analisados em um banco de dados local central ou planilha”.

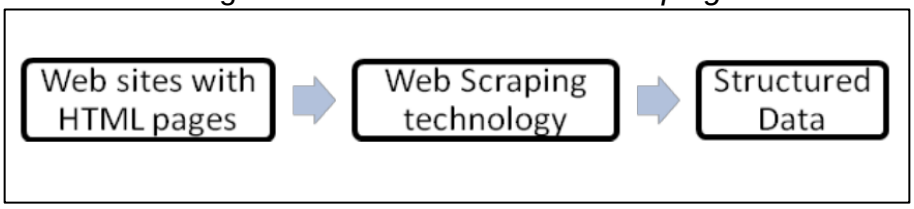
Ghosh Dastidar, Banerjee e Sengupta (2016, p. 25, tradução nossa) observam que esse “[...] é um processo de extração de informação útil de páginas HTML, que é a principal ferramenta de formatação de informações na WWW, e pode ser implementado em qualquer linguagem de programação”.

> Existem várias técnicas de *Web scraping*, incluindo copiar e colar tradicional, *Text grapping* e correspondência de expressão regular, programação HTTP, análise de HTML, análise de DOM, software de *Web scraping*, plataformas de agregação vertical, reconhecimento de

> anotação semântica e analisadores de página da *Web* por visão computacional (SIRISURIYA, 2015, p. 135, tradução nossa).

Web scraping é uma forma de mineração de dados e o objetivo geral do processo de *scraping* é extrair informações de *sites* e transformá-las em uma estrutura compreensível, como planilhas, banco de dados ou um arquivo de valores separados por vírgula (CSV), conforme mostrado na Figura 8 (SIRISURIYA, 2015).

Figura 8 - Estrutura do *Web Scraping*



Fonte: Sirisuriya (2015, p. 136)

Esses softwares são as ferramentas usadas para automatizar o trabalho manual de copiar e colar para coletar uma grande quantidade de dados de *sites* como *sites* de diretórios, *sites* de imóveis, *sites* classificados e anúncios de empregos.

Segundo Upadhyay et al. (2017), em termos gerais a função de um de um *Web scraper* é extrair e agrupar o conteúdo de uma maneira sistemática para facilitar a análise posterior dos dados, sendo que seu objetivo é coletar dados de *sites* identificados e convertê-los em arquivos de texto bruto. Ele imita as ações de navegação do usuário humano na *Web*, acessa os *sites* e extrai conteúdos relevantes para o usuário. Seu funcionamento pode ser resumido por meio de um processo de três etapas:

- Uma conexão é estabelecida com o *site* de destino por meio do protocolo HTTP.
- Em seguida, o documento relevante é recuperado pelo robô e o conteúdo é extraído com base em bibliotecas[1] de análise de HTML, expressões regulares, lógica de programação ou abordagens de aprendizado de máquina.

[1] Biblioteca é um conjunto de subprogramas ou funções, geralmente organizadas em classes, que podem ser usadas para a construção de um software.

- A etapa final envolve a transformação do conteúdo da *Web* em um formato adequado aos requisitos do aplicativo que solicita os serviços do *scraper* (UPADHYAY et al., 2017).

Utilizar softwares é a técnica de *scraping* mais fácil, pois todas as outras técnicas, exceto copiar e colar tradicionais, exigem alguma forma de conhecimento técnico. Existem centenas de softwares de *Web scraping* disponíveis hoje, a maioria deles desenvolvidos usando *Java*, *Python* e *Ruby*.

Há também alguns softwares de *Web scraping* de código aberto e também softwares comerciais. Softwares de *Web scraping* como *YahooPipes*, *Google Web Scrapers* e extensões *Outwit Firefox*, sendo as melhores ferramentas para iniciantes em *Web scraping*.

De acordo com Sirisuriya (2015), a maior parte dos softwares de *Web scraping* oferece suporte ao sistema operacional Windows. Arquivos Excel, CSV e XML são os formatos de exportação de dados mais comuns, vide Quadro 3.

Quadro 3 - Comparação de Softwares de *Web Scraping*

Web Scraping Software	Operating System	Data Export formats
Visual Web Ripper	Win	CSV, Excel, XML, SQL Server, MySQL, SQLite, Oracle and OleDB, Customized C# or VB script file output
Helium Scraper	Win	CSV, XML, MS Access database, MySQL script file
Screen Scraper	Win, Mac, Unix/Linux	Text. HTML, SQL Script File, MySQL Script File, XML file, HTTP submit form
OutWit Hub	Win, Mac OS-X, Linux,	CSV (TSV), HTML, Excel or SQL script
Mozenda	Win	CSV, TSV, or XML only.
WebSundew	Win	Text, CSV, Excel, XML; SQL Server, MySQL, Oracle and JDBC compatible DB (Pro and Enterprise edition)
Web Content Extractor	Win	Excel, text, HTML, MS Access DB, SQL Script File, MySQL Script File, XML file, HTTP submit form, ODBC Data source
Easy Web Extract	Win	Excel (CSV, TSV), text, HTML, MS Access DB, SQL Script File, MySQL Script File, XML file, HTTP submit form, ODBC Data source

Fonte: Sirisuriya, (2015, p. 136)

Na primeira coluna do Quadro 3 é apresentado a nomenclatura do *Web scraper*, na segunda, o tipo de sistema operacional que ele roda e na terceira, os formatos de arquivo que eles exportam para que possam ser abertos nas ferramentas de visualização e análise.

Outras ferramentas muito utilizadas por estudantes, profissionais de TI são elencadas por Ghosh Dastidar, Banerjee e Sengupta (2016, p. 28), como "*Diff bot, Scrappy, Selenium data scraping, Apache Camel, Archive.is, Jaxer e Import.io"*.

> Com poucas exceções, se você pode visualizar os dados em seu navegador, pode acessá-los por meio de um script *Python*. Se você puder acessá-lo em um script, poderá armazená-lo em um banco de dados. E se você pode armazená-los em um banco de dados, você pode fazer praticamente qualquer coisa com esses dados (RYAN MITCHELL, 2018, p. xi, , tradução nossa).

É possível desenvolver *Web scrapers* usando linguagens de programação populares como *Java*, *Python* e *Perl*, incorporando bibliotecas de terceiros uma vez que elas não estão presentes nativamente em tais linguagens. As bibliotecas fornecem acesso aos *sites* e os conteúdos são analisados usando ferramentas como expressões regulares.

Para o estudo de caso, a ser apresentado no próximo capítulo, no *Web scraper* utilizou-se a biblioteca *BeautifulSoup*. Sobre essa biblioteca, Ryan Mitchell (2018, p. 6, tradução nossa), observa que

> como seu homônimo do País das Maravilhas, *BeautifulSoup* tenta dar sentido ao absurdo; ajuda a formatar e organizar a bagunçada *Web* corrigindo HTML incorreto e nos apresentando objetos *Python* facilmente percorríveis que representam estruturas XML.

De acordo com Upadhyay et al. (2017), as bibliotecas normalmente têm dois componentes principais:

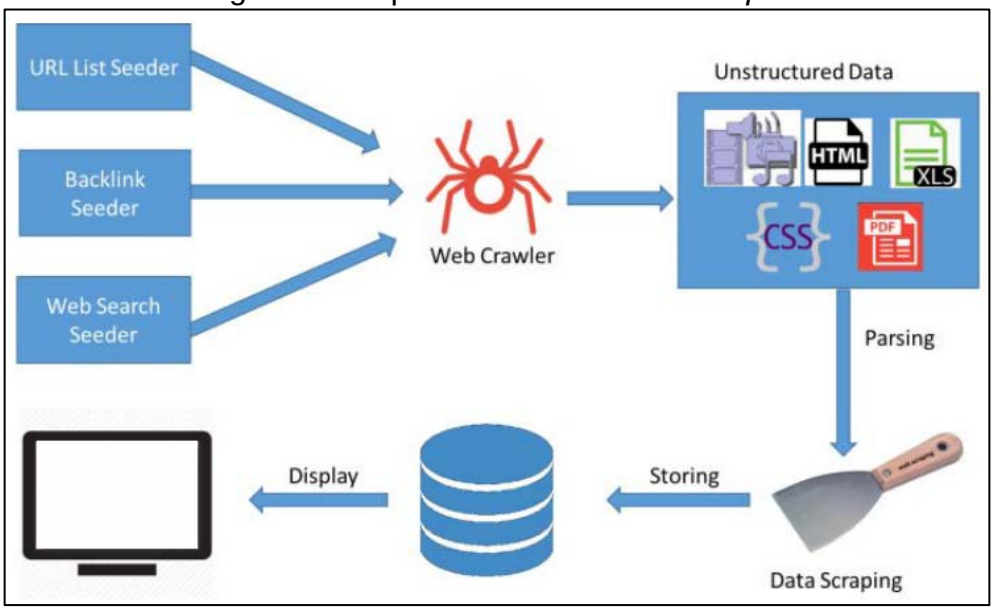
- O primeiro componente fornece suporte para os principais recursos HTTP, como autenticação, gerenciamento de histórico, gerenciamento de cookies e certificados SSL;
- O segundo componente fornece suporte de análise como correspondência XPath[2] e construção de árvore HTML.

[2] O XPath é uma linguagem de consulta que nos ajuda a navegar por documentos que usam marcadores, como os arquivos *Extensible Markup Language* (XML) e *HyperText Markup Language* (HTML)

Bibliotecas incorporadas em linguagens de programação apresentam certas desvantagens como a necessidade de fornecer acesso ao *site*, assim como a configuração de um componente é requerida para analisar o conteúdo da *Web*. No final de cada iteração, funções de análise operam no conteúdo extraído (UPADHYAY et al., 2017).

Na Figura 9, Upadhyay et al. (2017) apresentam uma arquitetura de um *Web scraper* onde 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. Esses *links* são então enviados para o "*Web Search Seeder*", que usa um *Web crawler* para extrair o conteúdo dos *sites* visados. Esse conteúdo extraído é então passado para um *scraper* para análise e o conteúdo raspado é enviado na forma de arquivos de texto para uma estação de trabalho.

Figura 9 - 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 extrair dados da *Web* e utilizá-los em um grande número de aplicativos de ciência de dados. Tal estruturação geralmente é estruturada em páginas HTML e revela uma boa parte da intenção semântica do autor, podendo ser usada em aplicativos de análise de dados.

> Tradicionalmente, um assistente de pesquisa gastaria milhares de horas de esforço manual para coletar dados e agrupá-los para análise,

> todavia, com a extração automatizada temos grandes vantagens de custo-tempo e mão de obra. Além disso, um grande volume de dados na *Web* permite uma análise comportamental em tempo real da dinâmica humana que pode nunca ter sido possível antes (UPADHYAY et al., 2017, p. 1, tradução nossa).

Na era dos dados, é uma ferramenta inestimável para organizações empresariais ou instituições de pesquisa, trazendo um enorme retorno sobre o investimento, fornecendo acesso sem precedentes a volumes de dados diversos, que de outra forma exigiriam tempo e recursos humanos consideráveis (UPADHYAY et al., 2017).

> [...] a simplicidade de operação, adaptabilidade a uma ampla gama de domínios, utilização ideal de recursos, aplicabilidade a uma ampla gama de formatos de arquivo populares (HTML, PDF, ODT, Word, XLS etc) e quase fornecimento instantâneo de dados, ao contrário dos mecanismos tradicionais, que podem levar de horas a dias para extrair o conteúdo (UPADHYAY et al., 2017, p. 3, tradução nossa).

Com o objetivo de situar a Ciência da Informação e, por consequência, o profissional da informação nesse importante contexto, aplica-se a afirmação de Souza, Almeida e Baracho (2013, p. 160) que destacam que,

> [...] dentre todos os campos científicos, e mesmo dentre aqueles mais recentes, pode-se seguramente apontar a Ciência da Informação (CI) como das mais introspectivas, no tocante às temáticas de pesquisa. As questões conceituais subjacentes à área são, por vezes, foco de reflexões tão apaixonadas e profundas, que os seus objetos de estudo ficam por vezes obnubilados, relegados a segundo plano.

Os autores, adicionalmente, apontam para um inexorável esvaziamento ou diminuição substancial da área como uma ciência autônoma, com a migração daqueles que poderiam ser considerados objetos legítimos e atavicamente ligados à Ciência da Informação para outras áreas do conhecimento (SOUZA; ALMEIDA; BARACHO, 2013).

Ironicamente, um agente catalisador para esse fenômeno tem sido a “ecologia técnica” que sempre favoreceu a área através da multiplicação dos problemas informacionais decorrentes, expandindo assim, as possibilidades para suas soluções (SOUZA; ALMEIDA; BARACHO, 2013).

> [...] em contrapartida, traz a possibilidade – e até mesmo a prerrogativa – de mediação dos diálogos disciplinares. Essa essência

> interdisciplinar exorta o cientista da informação a navegar nos espaços teóricos, adaptar-se aos contextos tecnológicos e reinventar-se continuamente. Ou assim deveríamos ser (SOUZA; ALMEIDA; BARACHO, 2013, p. 171).

Os estudos relacionados às métricas privilegiaram a “[...] produção científica e sempre desenvolveram pesquisas buscando medir índices, [...] com o objetivo de avaliação de instituições, de produtividade de autores e para ranqueamento de revistas, entre outros” (ARAÚJO, 2017, p. 12).

Diante do que foi apresentado, no que tange as vantagens trazidas para os profissionais da informação, pela utilização de *Web scrapers*, e utilizando-se como aplicação prática a sugestão de Araújo (2017), descrita no parágrafo anterior, como proposta para a recuperação de informações sobre temática proposta, é demonstrado no capítulo seguinte, um protótipo onde pode-se verificar a potencialização das atividades desempenhadas por tais profissionais na recuperação de informações, em um portal da *Web*.

4 SCRAPERCI: UM PROTÓTIPO DE *WEB SCRAPER* PARA COLETA DE DADOS

Este capítulo descreve algumas informações prelimitares referente a elaboração de um protótipo de *Web Scraper,* batizado de ScraperCI, e pode ser utilizada por qualquer usuário na *internet* através do endereço http://scraperci.info. Este protótipo foi concebido com fins didáticos com o objetivo de ser uma possível solução para os problemas de ineficiência na utilização de mecanismos de busca e recuperação de informações na *Web*.

Diante do grande volume de informações disponíveis, destaca-se as grandes responsabilidades e desafios dos profissionais, tornando-se evidente a necessidade de uma maior familiarização com as novas tecnologias, para que as mesmas possam ser desenvolvidas a partir de princípios éticos e sociais, e não apenas através de conhecimentos e processos puramente técnicos como justificativa da escolha.

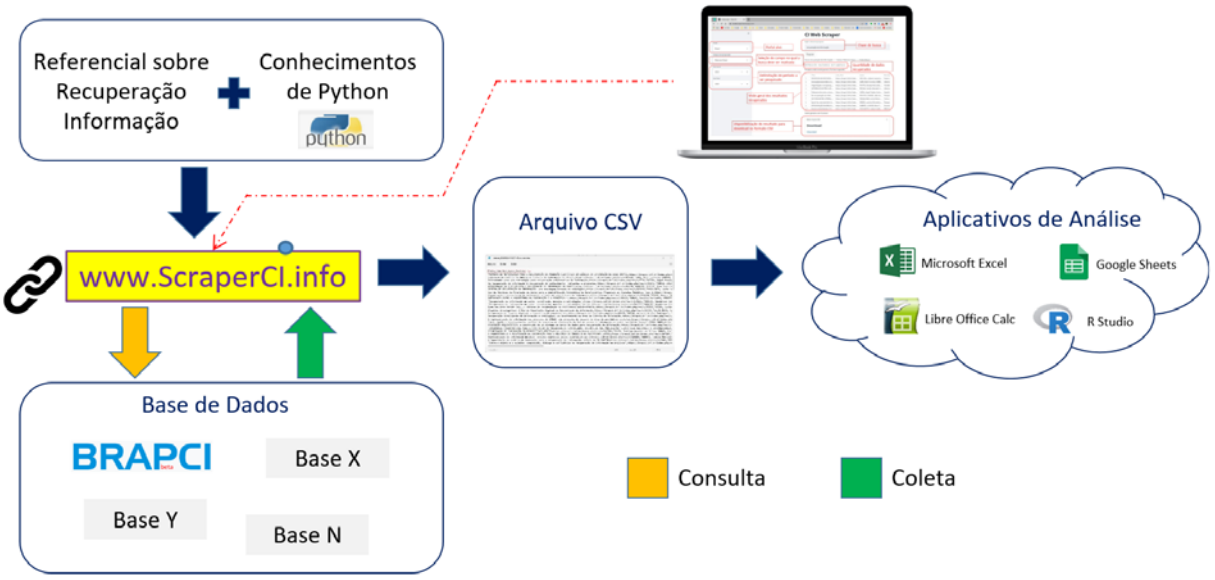
Tal afirmação, vai ao encontro das conclusões de Fernández Molina et al. (2005), quando comenta acerca das responsabilidades éticas dos profissionais da informação e do importante papel que desempenham no tratamento da informação no contexto das novas tecnologias, pois só assim os profissionais da informação estariam realmente assumindo e desempenhando seu verdadeiro papel como agentes sociais.

Destaca-se que, para este estudo foi utilizado a plataforma BRAPCI por ser umas das mais utilizadas no âmbito de pesquisas nacionais e pela quantidade de artigos indexados no formato *open source*.

Simulando necessidades reais das organizações, é proposto para o processo de coleta da informação o desenvolvimento de um *Web scraper* na linguagem de programação *Python*, propiciando o resgate de informações em massa no portal BRAPCI, em períodos de tempo muito curtos, quando comparados a buscas manuais.

Apresenta-se na Figura 10, o fluxo desenvolvido para o processo proposto, o qual permite realizar diversas consultas no portal BRAPCI, e consequentemente, extrair as conclusões do conteúdo pesquisado.

Figura 10 - Fluxograma do processo de consulta e análise dos dados



Fonte: Elaborado pelo autor

Com os dados tabulados no formato *Comma Separated Values* (CSV) diversas análises são realizadas através da importação do arquivo para uma ferramenta de análise como, por exemplo, o *Microsoft Excel, Google Sheets, Libre Office Calc* ou *R Studio*. Dessa forma, pode-se simular a realidade das grandes organizações, na qual decisões de extrema importância devem ser tomadas, frente a um expressivo volume de dados armazenados em diferentes plataformas.

Uma das preocupações que o profissional da informação também deve ponderar, no exercício de suas atividades, é com a disseminação do conhecimento assim como de ferramentas que contribuem para tal.

Atualmente, a forma mais viável de se atingir esse objetivo é utilizar a própria *Web* para compartilhar nossos conteúdos nos mais diversos formatos e tornar as ferramentas que desenvolvemos compatíveis com essa 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*[3] *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*

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

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*[4] ou de *deploy*[5] 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 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

---

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

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

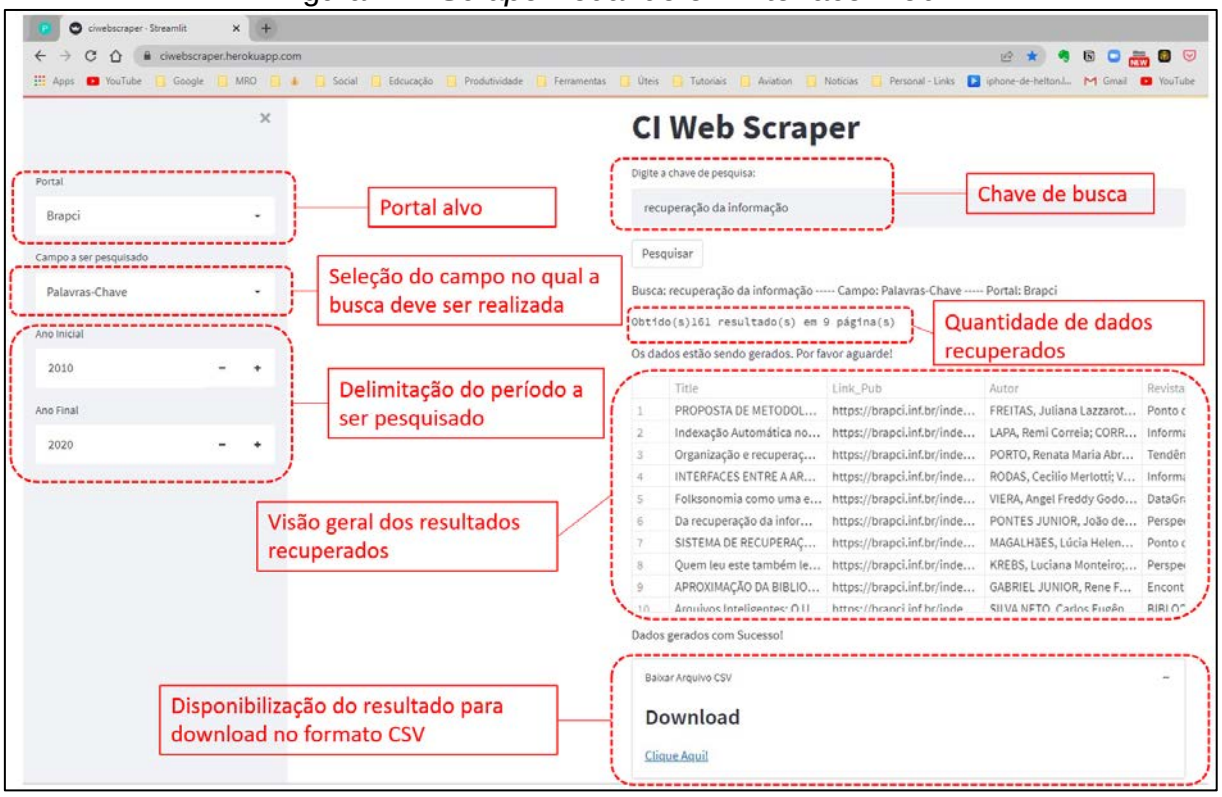
arquivos, entre outros, provendo uma maneira muito mais simples e prática de subir e escalar as aplicações.

A seguir será mostrado na prática, a simulação de uma possível necessidade do profissional da informação que por vezes tem que buscar e tomar decisões, sobre informações de determinado assunto, consultando os diversos portais e repositórios digitais.

A partir do uso do *Web scraper* desenvolvido, pesquisou-se pelo termo “recuperação da informação” no portal BRAPCI referente aos últimos dez anos (2010-2020), conforme fluxo demonstrado na Figura 10. O termo pesquisado e o espaço de tempo delimitado, foram definidos apenas a título de exemplo, todavia, isso pode ser expandido para qualquer outra demanda específica do usuário.

A Figura 11 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 11 - *Scraper* rodando em interface *Web*



Fonte: Elaborado pelo autor

Nos próximos parágrafos, de acordo com a Figura 11, 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

Para a consulta em questão, a fim de evitar a recuperação de dados espúrios, a pesquisa será focada no campo “Palavras-Chave”.

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. Conforme já mencionado, vamos delimitar a busca pelas publicações que atendem aos critérios de pesquisa, em um período de 10 anos (2010 a 2020), logo, digita-se 2010 e 2020 respectivamente.

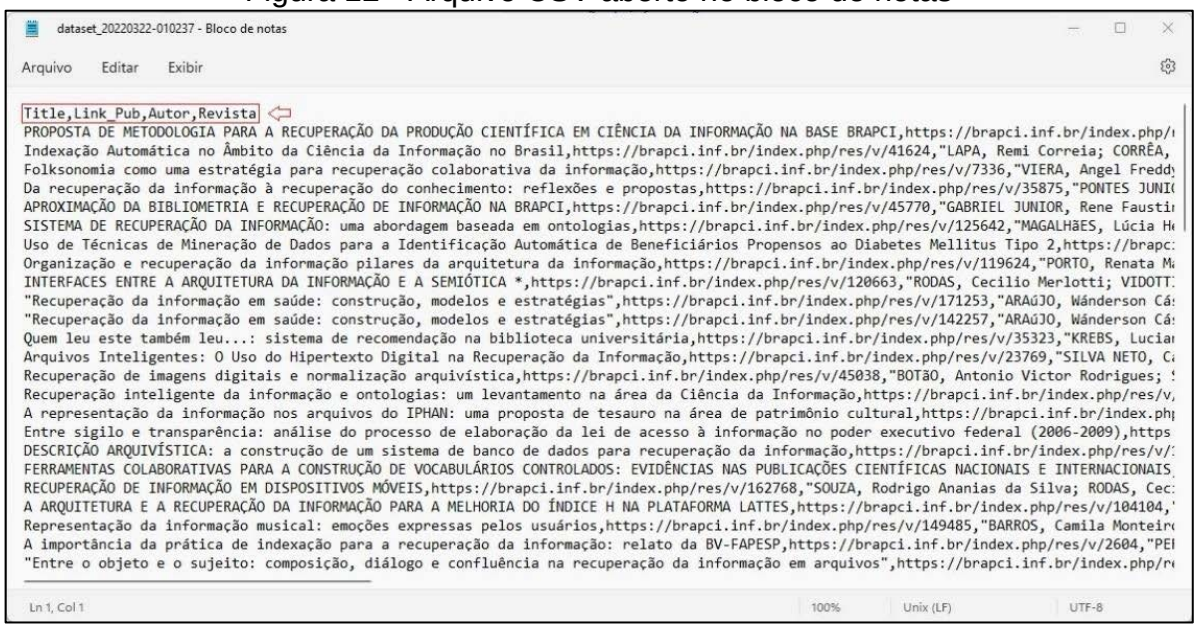
Clicando-se no botão “Pesquisar”, o software então, retorna o resultado da busca. Para essa consulta foi recuperado um total de 161 documentos distribuídos em 9 páginas do portal.

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.

O arquivo, tipo texto puro, pode ser aberto por um processador de texto como o bloco de notas do Windows, e observando sua estrutura na Figura 12, vemos os campos recuperados para cada documento, pela leitura da linha de cabeçalho, destacada pela caixa com bordas vermelhas:

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

Figura 12 - Arquivo CSV aberto no bloco de notas



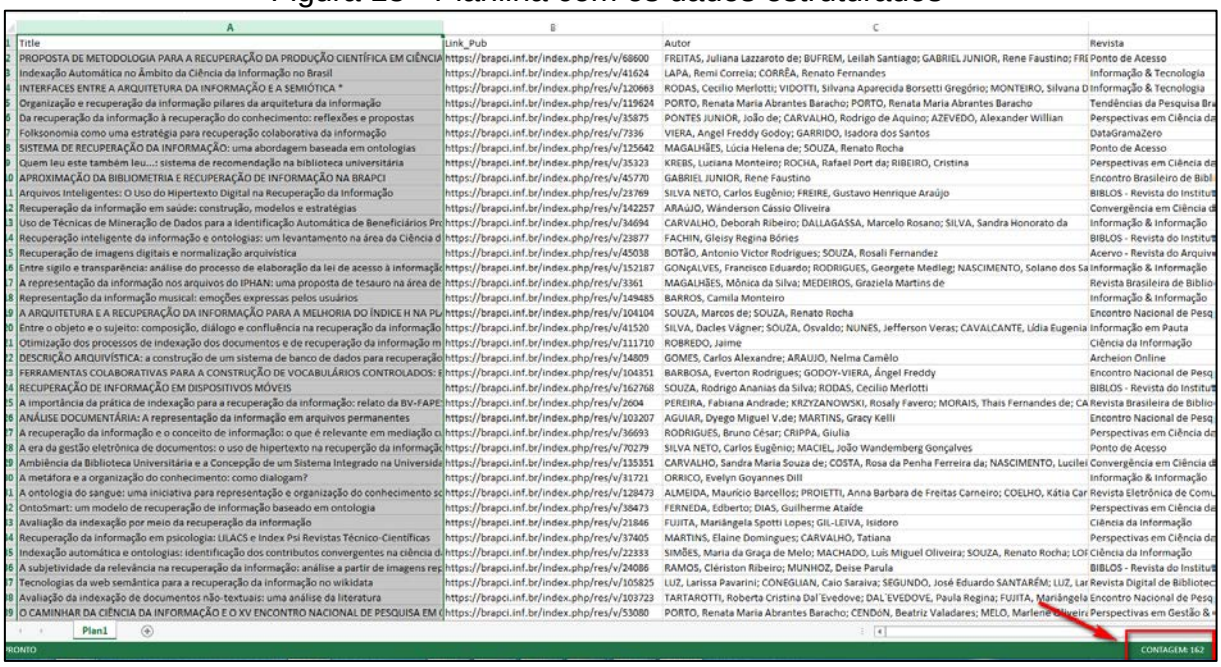
Fonte: Elaborado pelo autor

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. Um ponto a ser observado é que a quantidade de campos recuperados é diretamente proporcional ao tempo de processamento da consulta.

Em seguida, tal arquivo pode ser importado para qualquer software para análise dos dados. No APÊNDICE A é apresentado um passo a passo para importação no *Microsoft Excel,* assim como também é mostrado um exemplo de importação no *LibreOffice Calc*, no APÊNDICE B.

Uma vez na planilha, os dados poderão passar por diversas análises. Percebe-se, pelos pontos destacados na Figura 13 e Figura 14, respectivamente, que o campo contagem, mostra 162 linhas de dados, que se subtraindo 1 referente a linha de cabeçalho, corresponde ao número de 161 documentos recuperados pelo *scraper* (Figura 11).

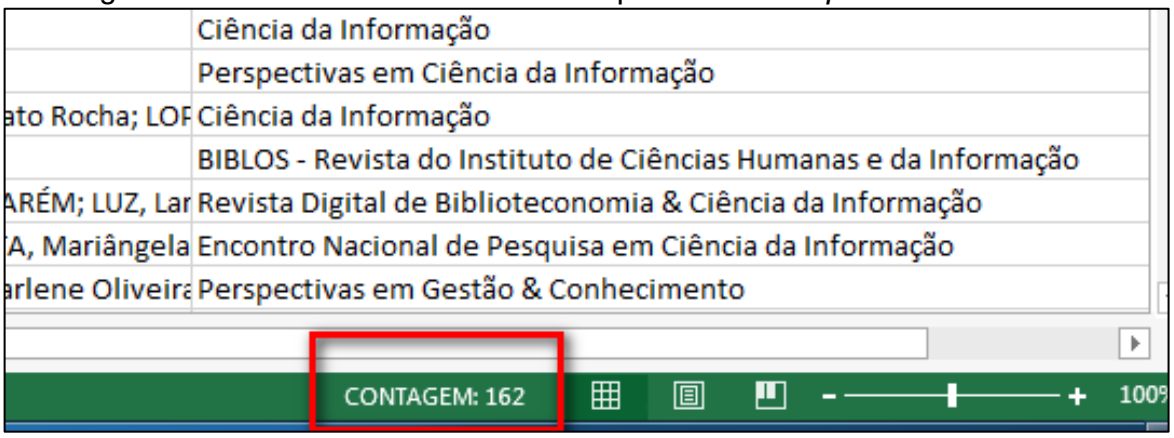
Figura 13 - Planilha com os dados estruturados



	A	B	C	
1	Title	Link_Pub	Autor	Revista
2	PROPOSTA DE METODOLOGIA PARA A RECUPERAÇÃO DA PRODUÇÃO CIENTÍFICA EM CIÊNCIA	https://brapci.inf.br/index.php/res/v/68600	FREITAS, Juliana Lazzaroto de; BUFREM, Leilah Santiago; GABRIEL JUNIOR, Rene Faustino; FRE	Ponto de Acesso
3	Indexação Automática no Âmbito da Ciência da Informação no Brasil	https://brapci.inf.br/index.php/res/v/41624	LAPA, Remi Correia; CORRÊA, Renato Fernandes	Informação & Tecnologia
4	INTERFACES ENTRE A ARQUITETURA DA INFORMAÇÃO E A SEMIÓTICA *	https://brapci.inf.br/index.php/res/v/120663	RODAS, Cecilio Merlotti; VIDOTTI, Silvana Aparecida Borsetti Gregório; MONTEIRO, Silvana D	Informação & Tecnologia
5	Organização e recuperação da informação pilares da arquitetura da informação	https://brapci.inf.br/index.php/res/v/119624	PORTO, Renata Maria Abrantes Baracho; PORTO, Renata Maria Abrantes Baracho	Tendências da Pesquisa Bra
6	Da recuperação da informação à recuperação do conhecimento: reflexões e propostas	https://brapci.inf.br/index.php/res/v/35875	PONTES JUNIOR, João de; CARVALHO, Rodrigo de Aquino; AZEVEDO, Alexander Willian	Perspectivas em Ciência da
7	Folksonomia como uma estratégia para recuperação colaborativa da informação	https://brapci.inf.br/index.php/res/v/7336	VIERA, Angel Freddy Godoy; GARRIDO, Isadora dos Santos	DataGramaZero
8	SISTEMA DE RECUPERAÇÃO DA INFORMAÇÃO: uma abordagem baseada em ontologias	https://brapci.inf.br/index.php/res/v/125642	MAGALHÃES, Lúcia Helena de; SOUZA, Renato Rocha	Ponto de Acesso
9	Quem leu este também leu...: sistema de recomendação na biblioteca universitária	https://brapci.inf.br/index.php/res/v/35323	KREBS, Luciana Monteiro; ROCHA, Rafael Port da; RIBEIRO, Cristina	Perspectivas em Ciência da
10	APROXIMAÇÃO DA BIBLIOMETRIA E RECUPERAÇÃO DE INFORMAÇÃO NA BRAPCI	https://brapci.inf.br/index.php/res/v/45770	GABRIEL JUNIOR, Rene Faustino	Encontro Brasileiro de Bibl
11	Arquivos Inteligentes: O Uso do Hipertexto Digital na Recuperação da Informação	https://brapci.inf.br/index.php/res/v/23769	SILVA NETO, Carlos Eugênio; FREIRE, Gustavo Henrique Araújo	BIBLOS - Revista do Institut
12	Recuperação da informação em saúde: construção, modelos e estratégias	https://brapci.inf.br/index.php/res/v/142257	ARAÚJO, Wânderson Cássio Oliveira	Convergência em Ciência d
13	Uso de Técnicas de Mineração de Dados para a Identificação Automática de Beneficiários Pr	https://brapci.inf.br/index.php/res/v/34694	CARVALHO, Deborah Ribeiro; DALLAGASSA, Marcelo Rosano; SILVA, Sandra Honorato da	Informação & Informação
14	Recuperação inteligente da informação e ontologias: um levantamento na área da Ciência d	https://brapci.inf.br/index.php/res/v/23877	FACHIN, Gleisy Regina Bóries	BIBLOS - Revista do Institut
15	Recuperação de imagens digitais e normalização arquivística	https://brapci.inf.br/index.php/res/v/45038	BOTÃO, Antonio Victor Rodrigues; SOUZA, Rosali Fernandez	Acervo - Revista do Arquiv
16	Entre sigilo e transparência: análise do processo de elaboração da lei de acesso à informaçã	https://brapci.inf.br/index.php/res/v/152187	GONÇALVES, Francisco Eduardo; RODRIGUES, Georgete Medleg; NASCIMENTO, Solano dos Sa	Informação & Informação
17	A representação da informação nos arquivos do IPHAN: uma proposta de tesauro na área de	https://brapci.inf.br/index.php/res/v/3361	MAGALHÃES, Mônica da Silva; MEDEIROS, Graziela Martins de	Revista Brasileira de Biblio
18	Representação da informação musical: emoções expressas pelos usuários	https://brapci.inf.br/index.php/res/v/149485	BARROS, Camila Monteiro	Informação & Informação
19	A ARQUITETURA E A RECUPERAÇÃO DA INFORMAÇÃO PARA A MELHORIA DO ÍNDICE H NA PL	https://brapci.inf.br/index.php/res/v/104104	SOUZA, Marcos de; SOUZA, Renato Rocha	Encontro Nacional de Pesq
20	Entre o objeto e o sujeito: composição, diálogo e confluência na recuperação da informação	https://brapci.inf.br/index.php/res/v/41520	SILVA, Dacles Vágner; SOUZA, Osvaldo; NUNES, Jefferson Veras; CAVALCANTE, Lídia Eugenia	Informação em Pauta
21	Otimização dos processos de indexação dos documentos e de recuperação da informação m	https://brapci.inf.br/index.php/res/v/111710	ROBREDO, Jaime	Ciência da Informação
22	DESCRIÇÃO ARQUIVÍSTICA: a construção de um sistema de banco de dados para recuperação	https://brapci.inf.br/index.php/res/v/14809	GOMES, Carlos Alexandre; ARAUJO, Nelma Camêlo	Archeion Online
23	FERRAMENTAS COLABORATIVAS PARA A CONSTRUÇÃO DE VOCABULÁRIOS CONTROLADOS: E	https://brapci.inf.br/index.php/res/v/104351	BARBOSA, Everton Rodrigues; GODOY-VIERA, Ángel Freddy	Encontro Nacional de Pesq
24	RECUPERAÇÃO DE INFORMAÇÃO EM DISPOSITIVOS MÓVEIS	https://brapci.inf.br/index.php/res/v/162768	SOUZA, Rodrigo Ananias da Silva; RODAS, Cecilio Merlotti	BIBLOS - Revista do Institut
25	A importância da prática de indexação para a recuperação da informação: relato da BV-FAPE	https://brapci.inf.br/index.php/res/v/2604	PEREIRA, Fabiana Andrade; KRZYZANOWSKI, Rosaly Favero; MORAIS, Thais Fernandes de; CA	Revista Brasileira de Biblio
26	ANÁLISE DOCUMENTÁRIA: A representação da informação em arquivos permanentes	https://brapci.inf.br/index.php/res/v/103207	AGUIAR, Dyego Miguel V.de; MARTINS, Gracy Kelli	Encontro Nacional de Pesq
27	A recuperação da informação e o conceito de informação: o que é relevante em mediação c	https://brapci.inf.br/index.php/res/v/36693	RODRIGUES, Bruno César; CRIPPA, Giulia	Perspectivas em Ciência da
28	A era da gestão eletrônica de documentos: o uso de hipertexto na recuperação da informaçã	https://brapci.inf.br/index.php/res/v/70279	SILVA NETO, Carlos Eugênio; MACIEL, João Wandemberg Gonçalves	Ponto de Acesso
29	Ambiência da Biblioteca Universitária e a Concepção de um Sistema Integrado na Universida	https://brapci.inf.br/index.php/res/v/135351	CARVALHO, Sandra Maria Souza de; COSTA, Rosa da Penha Ferreira da; NASCIMENTO, Lucilei	Convergência em Ciência d
30	A metáfora e a organização do conhecimento: como dialogam?	https://brapci.inf.br/index.php/res/v/31721	ORRICO, Evelyn Goyannes Dill	Informação & Informação
31	A ontologia do sangue: uma iniciativa para representação e organização do conhecimento so	https://brapci.inf.br/index.php/res/v/128473	ALMEIDA, Maurício Barcellos; PROIETTI, Anna Barbara de Freitas Carneiro; COELHO, Kátia Car	Revista Eletrônica de Comu
32	OntoSmart: um modelo de recuperação de informação baseado em ontologia	https://brapci.inf.br/index.php/res/v/38473	FERNEDA, Edberto; DIAS, Guilherme Ataíde	Perspectivas em Ciência de
33	Avaliação da indexação por meio da recuperação da informação	https://brapci.inf.br/index.php/res/v/21846	FUJITA, Mariângela Spotti Lopes; GIL-LEIVA, Isidoro	Ciência da Informação
34	Recuperação da informação em psicologia: LILACS e Index Psi Revistas Técnico-Científicas	https://brapci.inf.br/index.php/res/v/37405	MARTINS, Elaine Domingues; CARVALHO, Tatiana	Perspectivas em Ciência da
35	Indexação automática e ontologias: identificação dos contributos convergentes na ciência d	https://brapci.inf.br/index.php/res/v/22333	SIMÕES, Maria da Graça de Melo; MACHADO, Luis Miguel Oliveira; SOUZA, Renato Rocha; LOF	Ciência da Informação
36	A subjetividade da relevância na recuperação da informação: análise a partir de imagens rep	https://brapci.inf.br/index.php/res/v/24086	RAMOS, Clériston Ribeiro; MUNHOZ, Deise Parula	BIBLOS - Revista do Institut
37	Tecnologias da web semântica para a recuperação da informação no wikidata	https://brapci.inf.br/index.php/res/v/105825	LUZ, Larissa Pavarini; CONEGLIAN, Caio Saraiva; SEGUNDO, José Eduardo SANTARÉM; LUZ, Lar	Revista Digital de Biblioteco
38	Avaliação da indexação de documentos não-textuais: uma análise da literatura	https://brapci.inf.br/index.php/res/v/103723	TARTAROTTI, Roberta Cristina Dal´Evedove; DAL´EVEDOVE, Paula Regina; FUJITA, Mariângela	Encontro Nacional de Pesq
39	O CAMINHAR DA CIÊNCIA DA INFORMAÇÃO E O XV ENCONTRO NACIONAL DE PESQUISA EM C	https://brapci.inf.br/index.php/res/v/53080	PORTO, Renata Maria Abrantes Baracho; CENDON, Beatriz Valadares; MELO, Marlene Oliveira	Perspectivas em Gestão &

Fonte: Elaborado pelo autor

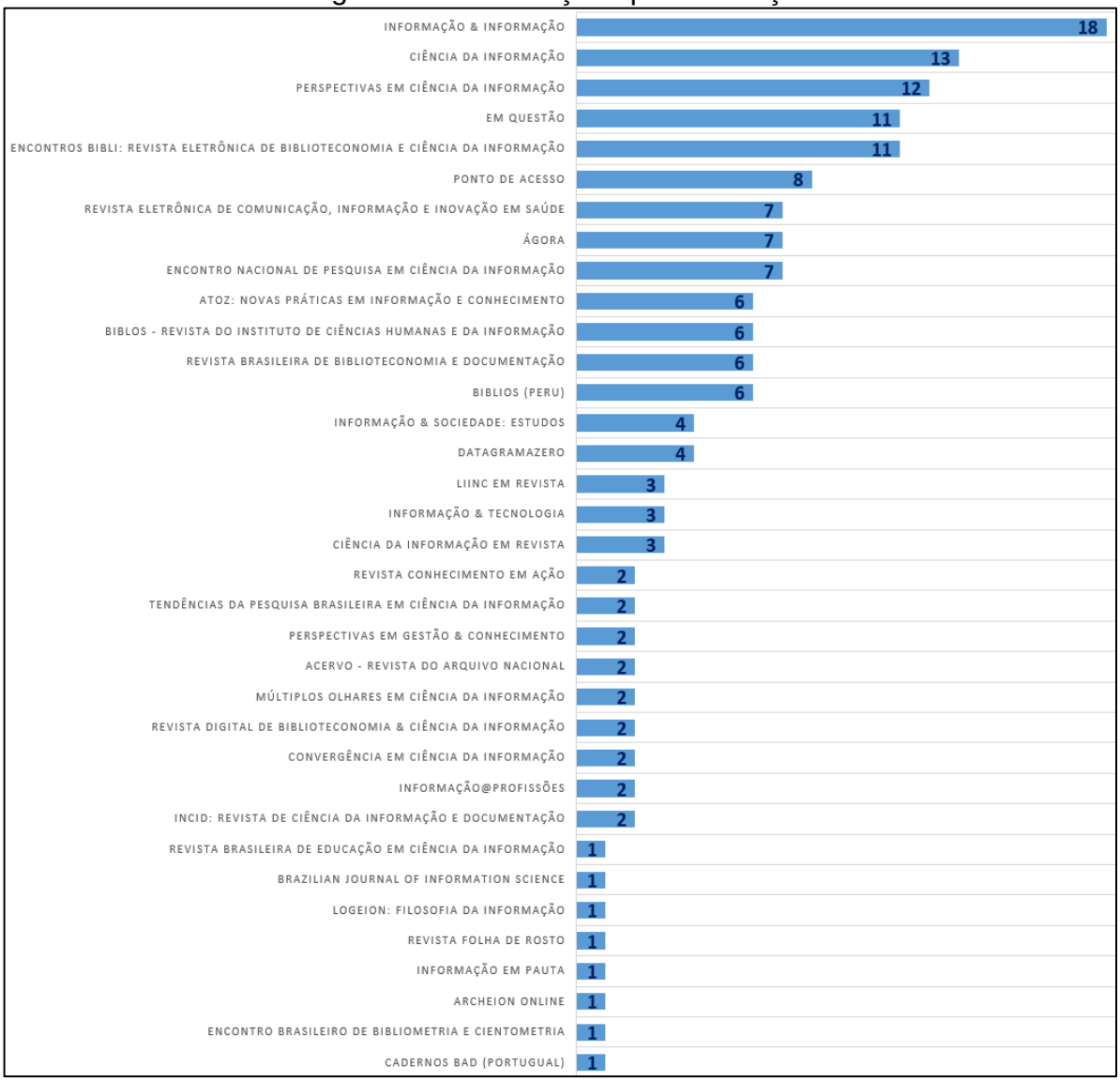
Figura 14 - Conferência dos itens recuperados - *Scraper* versus Planilha



Fonte: Elaborado pelo autor

Pode-se então, de maneira eficiente, fazer diversas análises e tirar conclusões. Por exemplo, verificamos pela análise da Figura 15, que as 161 publicações recuperadas foram publicadas em 35 instituições, sendo que 7 delas foram responsáveis por 80 publicações (50% 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 15 - Publicações por instituição



Fonte: Elaborado pelo autor

Observou-se que, 293 autores estão relacionados aos 161 documentos recuperados, todavia, vemos no Quadro 4 os 19 (6,4% do total), que tiveram recorrência de 3 ou mais documentos publicados relacionado ao tema, no período. Esse ranking poderia ser utilizado, por exemplo, como ponto de partida para uma pesquisa sobre recuperação da informação, onde se selecionaria os maiores especialistas no assunto.

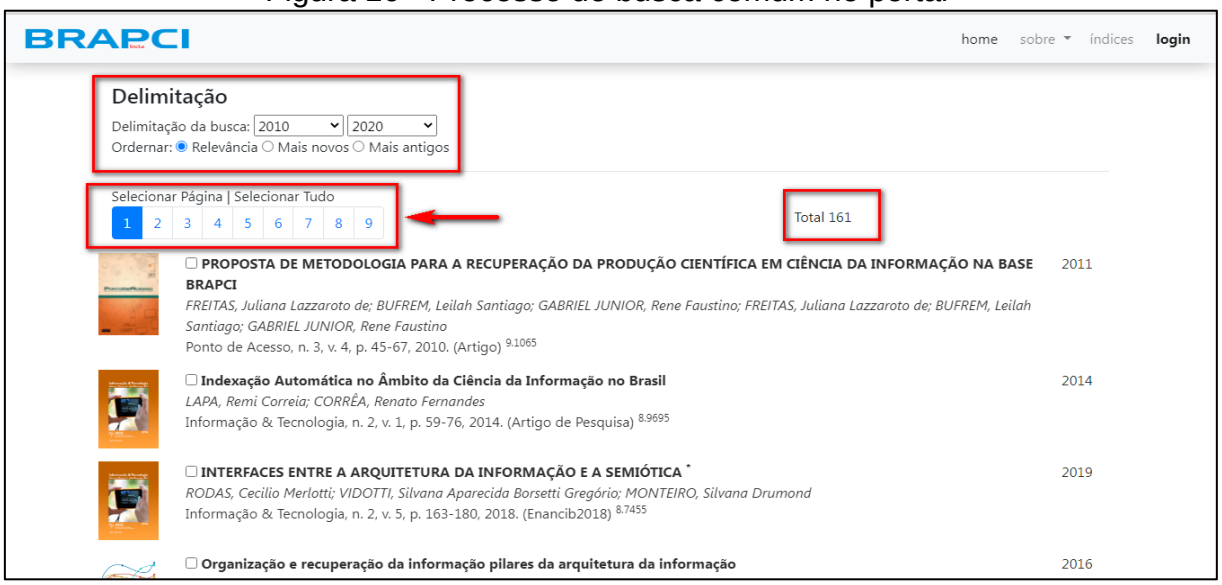
Quadro 4 - Autores que publicaram 3 ou mais documentos no período

Autores	Quantidade
PORTO, Renata Maria Abrantes Baracho	8
CORRÊA, Renato Fernandes	7
VIDOTTI, Silvana Aparecida Borsetti Gregório	5
VIERA, Angel Freddy Godoy	5
SOUZA, Renato Rocha	4
ALMEIDA, Maurício Barcellos	4
FERNEDA, Edberto	4
SANTARÉM SEGUNDO, José Eduardo	4
DIAS, Thiago Magela Rodrigues	4
GABRIEL JUNIOR, Rene Faustino	3
RODAS, Cecilio Merlotti	3
MONTEIRO, Silvana Drumond	3
FUJITA, Mariângela Spotti Lopes	3
VIEIRA, Jessica Monique de Lira	3
MOREIRA, Fábio Mosso	3
SANTANA, Ricardo César Gonçalves	3
SANTOS, Luana Carla de Moura dos	3
BRÄSCHER, Marisa	3
MOITA, Gray Farias	3

Fonte: Elaborado pelo autor

Caso, o processo de consulta fosse realizado sem o uso da tecnologia, com o buscador padrão do *site* (vide Figura 16), o usuário teria que compilar os dados manualmente, registrando os dados de 161 documentos e navegando por 9 páginas diferentes. Esse processo, embora possa levar aos mesmos resultados, seria moroso e ineficiente, além de estar mais propenso a ocorrências de erros.

Figura 16 - Processo de busca comum no portal



Fonte: Elaborado pelo autor

Tal fato denota que, o processo de coleta pelo uso do *Web scraper,* traz inúmeros benefícios ao profissional da informação, que pode passar maior tempo em atividades nobres, realizando análises, tirando conclusões, endereçando soluções, e com isso, agregar valor de fato as organizações, do que gastar energia em processos manuais e repetitivos de coleta, preparação e estruturação dos dados, para só depois pôr em prática seu trabalho analítico.

Demonstra-se também, outro tipo de contribuição, que como profissionais da informação, podemos trazer ao meio em que atuamos, onde de forma colaborativa, multiplicamos as ferramentas disponíveis aos nossos pares. Tecnologias e ou ferramentas disponíveis a um número restrito de pessoas passam a ganhar escala, com potencial para solucionar problemas de outros profissionais, que podem, a partir disso, propor melhorias e até mesmo se inspirar na criação de novas soluções.

5 CONSIDERAÇÕES FINAIS

A recuperação de grandes quantidades de dados estruturados ou não estruturados e seu uso sistemático para tomadas de decisão, depende do uso de aplicações tecnológicas que possibilitem a recuperação de recursos informacionais das mais diversas fontes disponíveis em ambientes digitais e *Web*.

O caso prático de utilização do *Web scraper* desenvolvido, e testado no portal BRAPCI, despontou que o uso dessa tecnologia se mostrou eficiente na coleta de dados, ampliando as possibilidades e trazendo maior produtividade, no que tange a extração de recursos informacionais na *Web*, se mostrando como uma possibilidade viável a ser explorada pelo profissional da informação, que está no cerne desse processo de transformação digital que estamos vivenciando.

Por conseguinte, é salutar que o profissional da Informação esteja familiarizado com o funcionamento dos sistemas de recuperação de informações e se atualizem munindo-se de ferramentas e habilidades que o auxiliem a explorar as bases de dados, coletando a informação de maneira mais precisa e eficiente.

Com isso, espera-se que esta pesquisa possa contribuir para uma maior compreensão das potencialidades do uso dessas ferramentas e possa estimular os profissionais da informação a desenvolver novas competências e possibilidades inovadoras de atuação profissional ao propor a criação de novos métodos de recuperação informacional.

Nessa perspectiva, constatou-se o cenário desafiador para CI no contexto contemporâneo uma vez que, com o efeito da globalização do conhecimento e o compartilhamento em massa de grandes quantidades de conteúdos informacionais, a demanda por informação de qualidade, que geram valor e tenham potencial inovador é cada vez maior para as tomadas de decisão sejam elas individuais, sejam corporativas.

O presente trabalho buscou fomentar uma discussão na área da Ciência da informação, acerca do uso de *Web scrapers* para coleta de dados, abordando características conceituais e práticas dessa tecnologia.

Ainda que o grande volume de conteúdos disponíveis na *Web* seja algo desafiador no que diz respeito à eficiência no processo de recuperação, é inegável o fato de que o uso de *Web scrapers* possa contribuir para a coleta dos mais variados conteúdos de forma rápida, sistêmica e padronizada na *internet*.

Assim, em pesquisas futuras, temos a possibilidade de expandir o escopo da ferramenta desenvolvida, agregando funcionalidades que a permitam recuperar informações em outros repositórios assim como aperfeiçoar seus métodos de busca melhorando continuamente tanto o tempo de resposta quanto a qualidade do resultado para o usuário.

Também há possibilidade, no futuro, de se explorar as tendências atuais de disponibilização, representação e recuperação dos recursos informacionais em ambientes digitais, assim como a crescente necessidade da Ciência da Informação em adaptar-se a essa realidade.

Outras análises poderão ser realizadas não somente quanto aos aspectos técnicos e produtivos, mas também em relação aos impactos sociais do uso desse tipo de tecnologia.

Por fim, ressalta-se que frente as crescentes demandas informacionais, pesquisas ainda são necessárias no sentido de propor um maior aprofundamento e compreensão nessa temática.

6 REFERÊNCIAS

ARAÚJO, C. A. Á. Teorias e Tendências Contemporâneas da Ciência da Informação. **Inf. Pauta**, v. 2, n. 2, p. 9–34, 2017.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca**. 2. ed. Porto Alegre: Bookman, 2013.

BARDIN, L. **Análise De Conteúdo**. 1ª Ed. ed. São Paulo: Edições 70, 2011.

BARRETO, A. DE A. A condição da informação. **São Paulo em Perspectiva**, v. 16, n. 3, p. 67–74, jul. 2002.

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.

BUCKLAND, M. K. **Information and Society**. Cambridge, United States: MIT Press Ltd, 2017.

BUSH, V. As We May Think. **Atlantic Monthly**, v. 176, p. 112–124, 1 fev. 1945.

CARVALHO, A. DE O.; CARVALHO, M. B. P. DE. A semântica e a Classificação Decimal Universal. **Ciência da Informação**, v. 24, n. 2, 19 abr. 2018.

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

CHUI, M. et al. The social economy: Unlocking value and productivity through social technologies. **McKinsey Global Institute**, p. 1–184, 2012.

FEITOSA, A. **ORGANIZAÇAO DA INFORMAÇAO NA *WEB*: DAS TAGS A *WEB* SEMANTICA**. 1ª ed. Brasília: Thesaurus, 2006.

FELDMAN, S.; SHERMAN, C. The High Cost of Not Finding Information. **IDC White Paper**, p. 10, 2001.

FERNEDA, E. **Recuperação de informação: análise sobre a contribuição da ciência da computação para a ciência da informação**. São Paulo: Universidade de São Paulo, 15 dez. 2003.

FONSECA, J. J. S. **Metodologia da Pesquisa Científica**. [s.l.] UECE, 2002.

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.

GONDIM, L. M. P.; LIMA, J. C. **A Pesquisa Como Artesanato Intelectual: Considerações Sobre Método E Bom Senso**. São Carlos: EdUFSCar, 2010. v. 6

LANCASTER, F. W. **Information retrieval systems; : characteristics, testing, and evaluation**. New York: John Wiley, 1968.

LESK, M. The Seven Ages of Information Retrieval. **International Federation of Library Associations and Institutions**, 1996.

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

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

RAMALHO, R. A. S. **Desenvolvimento e utilização de ontologias em bibliotecas digitais: uma proposta de aplicação**. [s.l.] Universidade Estadual Paulista (Unesp), 25 mar. 2010.

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, 8 maio 2011.

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

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, 20 dez. 2016.

SANTOS, P. L. V. A. DA C.; SANT'ANA, R. C. G. Transferência da Informação: análise para valoração de unidades de conhecimento. **Ciência da Informação**, v. 3, n. 2, p. 1–21, 2002.

SILVA, E. L. DA; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 3. ed. Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001.

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.

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

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na *Web*: panorama atual e tendências. **Perspectivas em Ciência da Informação**, v. 11, n. 2, p. 161–173, 2006.

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.

SUFYAN, D.; ARJUMAND, M.; ABDUL QAYUME, K. *Web* Scrapper Tool for Data Extraction. **IJSTE-International Journal of Science Technology & Engineering |**, v. 2, n. 12, p. 64–71, 2016.

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

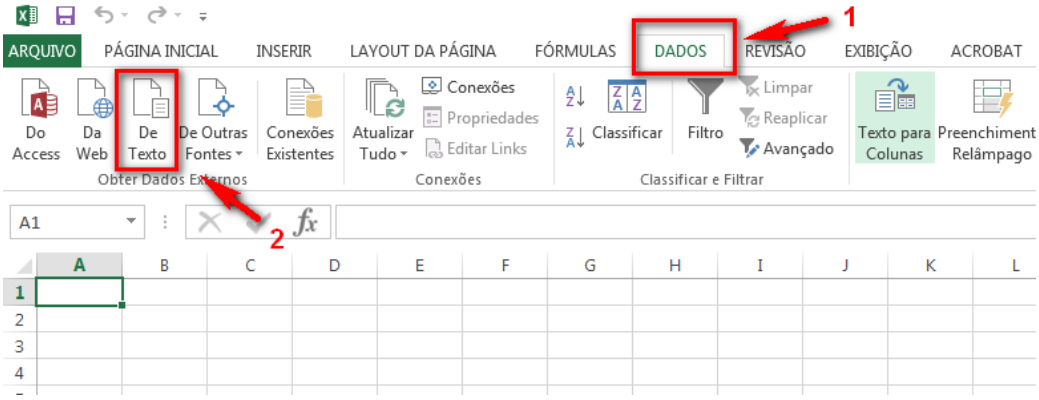
TRIVIÑOS, A. N. S. **Introdução à Pesquisa em Ciências Sociais: a Pesquisa Qualitativa em Educação – O Positivismo, A Fenomenologia, O Marxismo**. SÃO PAULO: EDITORA ATLAS S.A, 1987.

UPADHYAY, S. *et. al* 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: http://ieeexplore.ieee.org/document/8117827/. Acesso em: 22 jan. 2022.

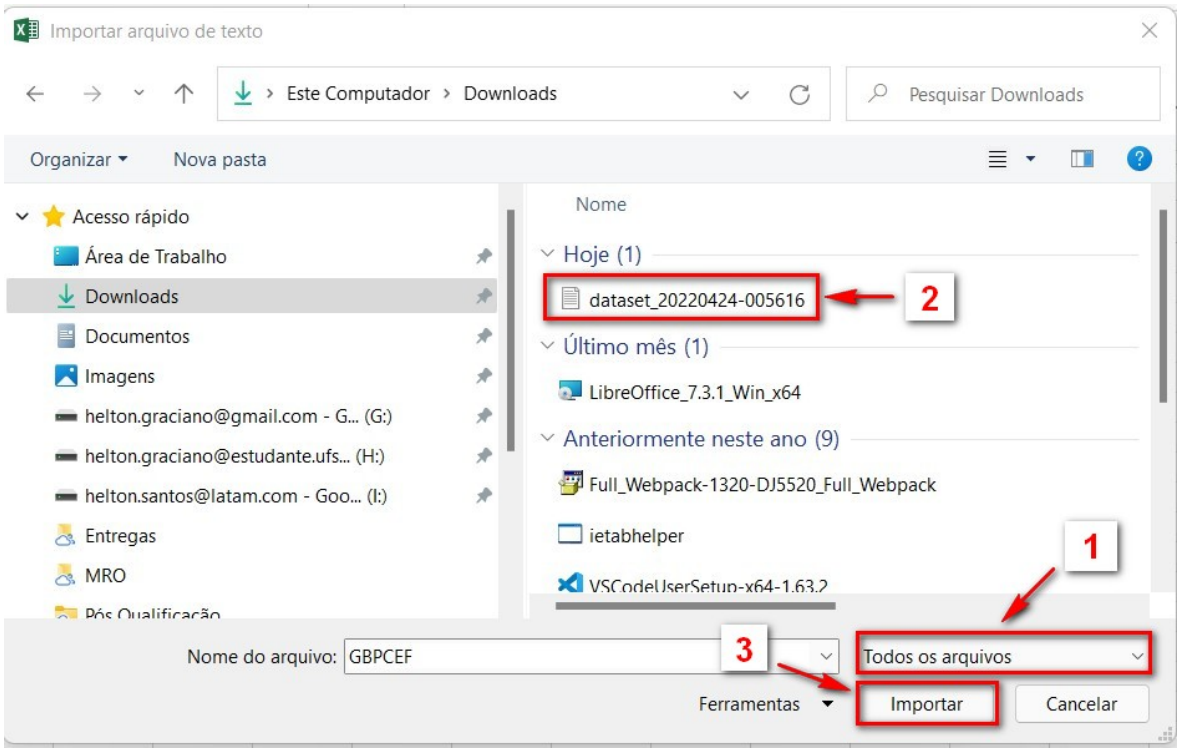
7 APÊNDICES

APÊNDICE A - Importação do arquivo CSV para o Excel

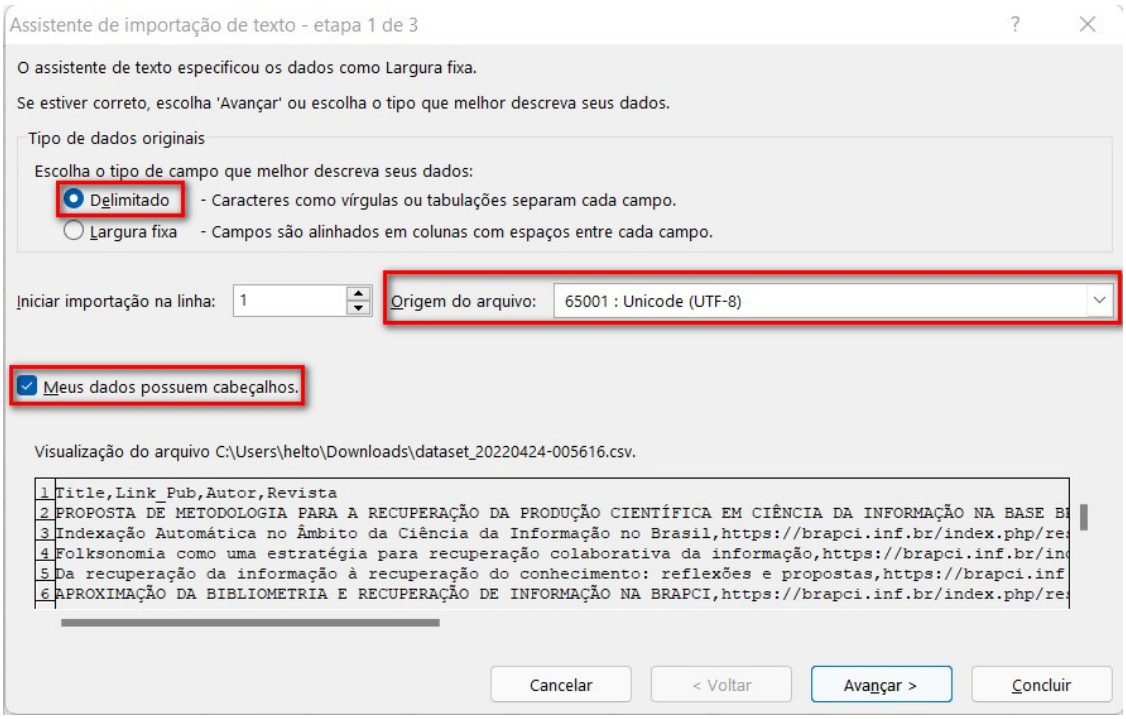
Demonstra-se o processo de importação através da ferramenta Microsoft Excel, dado sua popularidade e facilidade de uso. Conforme demonstrado, na guia “Dados” seleciona-se “De Texto”.



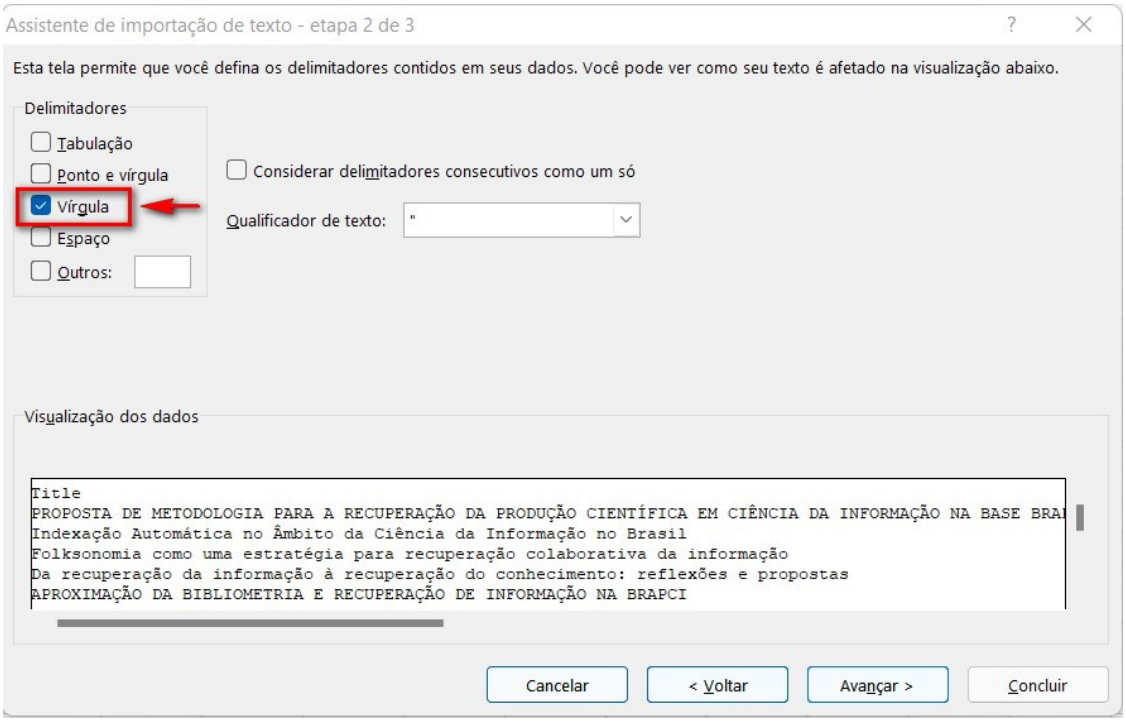
Na caixa de diálogo que se abrirá, seleciona-se “Todos os arquivos” para que o arquivo CSV fique visível. Em seguida, deve-se selecionar o arquivo e clicar em “Importar”.



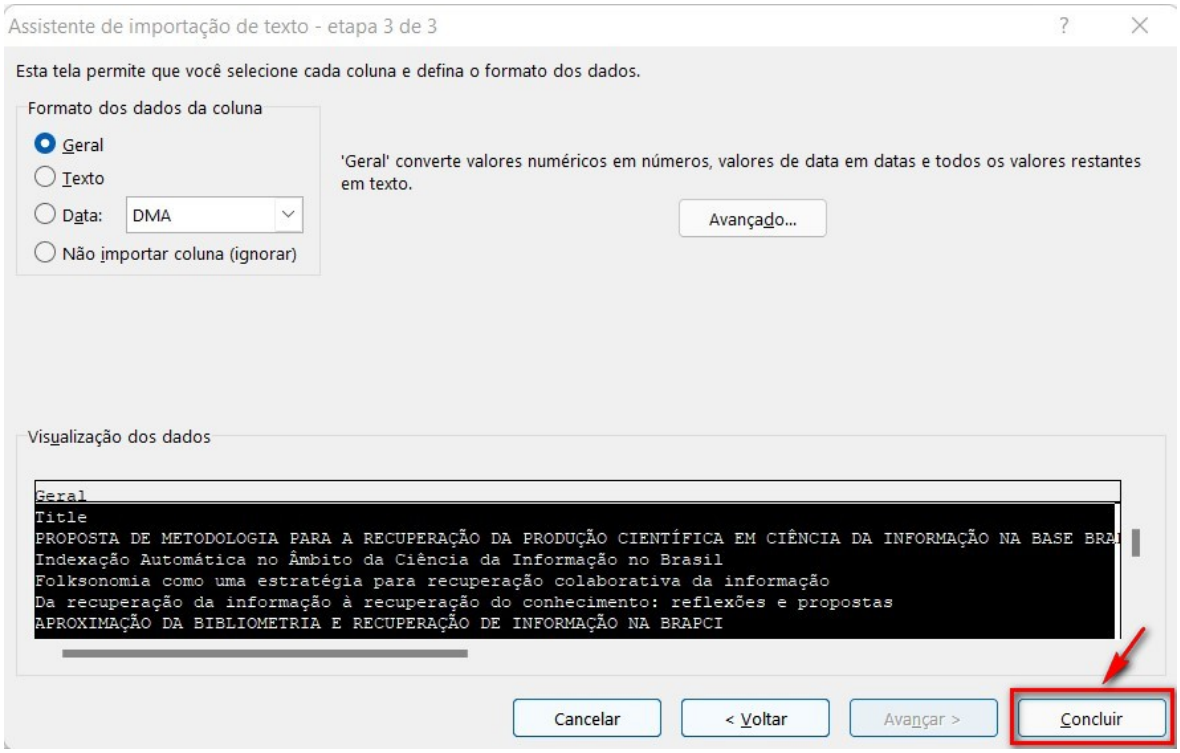
Segue-se então com as etapas de importação e, na caixa de diálogo mostrada abaixo, deve-se selecionar as informações conforme os campos destacados, para que os dados sejam importados adequadamente.



Na próxima tela, utiliza-se como delimitador a “Vírgula” (vide campo destacado), uma vez que o arquivo a ser importado foi gerado no formato CSV.



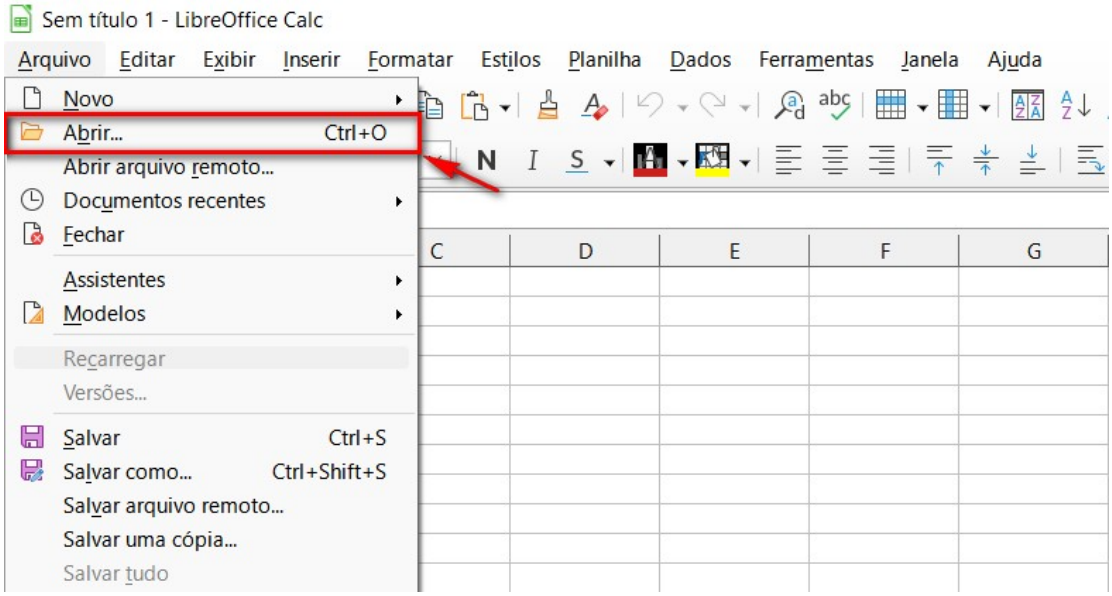
No penúltimo passo, finaliza-se a importação do arquivo clicando-se em “Concluir”.



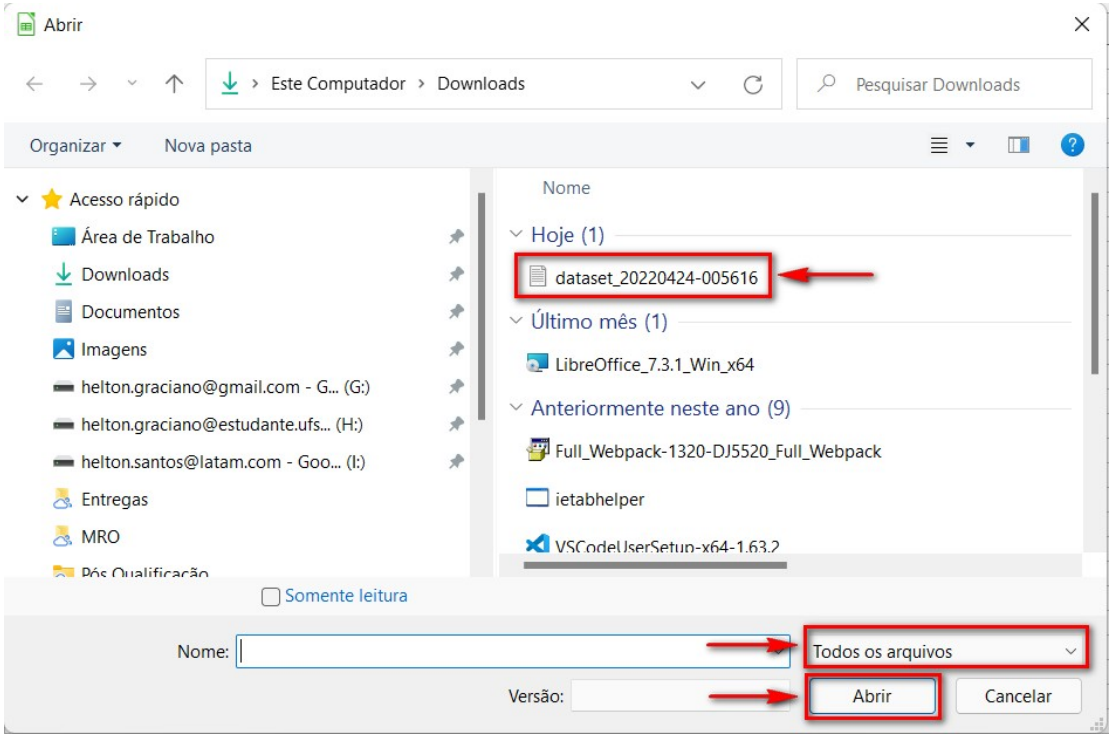
O processo de importação é concluído selecionando-se a aba e a célula nas quais os dados importados serão inseridos.

APÊNDICE B - Importação do arquivo CSV para o LibreOffice Calc

Como exemplo de importação em um software não proprietário, apresenta-se as etapas a serem seguidas no LibreOffice Calc. Primeiramente, no menu “Arquivo”, clica-se em “Abrir”.



Com a opção “Todos os Arquivos” selecionada, escolhe-se o arquivo CSV e clica-se em “Avançar”.



Para que os dados sejam importados adequadamente, na próxima caixa de diálogo, seleciona-se o padrão “Unicode (UTF-8)” utilizando-se como delimitador a “Vírgula” e clica-se em “OK”.