UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
MESTRADO EM CIÊNCIA DA INFORMAÇÃO

FERNANDO FEIJÓ FERREIRA

CICLO DE VIDA DE HIPER*LINKS*:
um estudo sobre a persistência e perda de referências e conteúdos da web

Porto Alegre
2023

FERNANDO FEIJÓ FERREIRA

CICLO DE VIDA DE HIPER*LINKS*:
um estudo sobre a persistência e perda de referências e conteúdos da web

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio Grande do Sul como requisito para obtenção do título de Mestre em Ciência da Informação.

Orientador: Prof. Dr. Moisés Rockembach

Porto Alegre
2023

CIP - Catalogação na Publicação

Ferreira, Fernando Feijó
CICLO DE VIDA DE HIPERLINKS: um estudo sobre a persistência e perda de referências e conteúdos da web / Fernando Feijó Ferreira. -- 2023.
88 f.
Orientador: Moisés Rockembach.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Programa de Pós-Graduação em Ciência da Informação, Porto Alegre, BR-RS, 2023.

1. link rot. 2. reference rot. 3. hiperlink. 4. website. 5. preservação digital. I. Rockembach, Moisés, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Programa de Pós-Graduação em Ciência da Informação - PPGCIN
Rua Ramiro Barcelos, 2705
Porto Alegre, RS – CEP 90035-007
Telefone: (51) 3308.5067
E-mail: ppgcin@ufrgs.br

FERNANDO FEIJÓ FERREIRA

CICLO DE VIDA DE HIPER*LINKS*:

um estudo sobre a persistência e perda de referências e conteúdos da web

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio Grande do Sul como requisito para obtenção do título de Mestre em Ciência da Informação.

Orientador: Prof. Dr. Moisés Rockembach

COMISSÃO EXAMINADORA

Professor Dr. Moisés Rockembach (Orientador)
Universidade Federal do Rio Grande do Sul

Professora Dra. Ana Javes Andrade da Luz
Universidade Federal do Rio Grande do Sul

Professora Dra. Caterina Pavão
Universidade Federal do Rio Grande do Sul

Professora Dr. Fabiano Couto Corrêa da Silva
Universidade Federal do Rio Grande do Sul

Professora Dr. Rodrigo Silva Caxias de Sousa (Suplente)
Universidade Federal do Rio Grande do Sul

Porto Alegre, 27 de abril de 2023.

A você Sofia, que mesmo ainda não entendendo o que isso significa me deu a inspiração necessária para a conclusão dessa jornada.

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todas as pessoas que contribuíram para o desenvolvimento e conclusão desta pesquisa. Sem o apoio e colaboração de cada um de vocês, esse trabalho não seria possível.

Agradeço ao meu orientador Professor Dr. Moisés Rockembach, por sua orientação, incentivo e conhecimentos compartilhados ao longo deste processo. Agradeço por sua paciência, por sempre estar disponível para esclarecer minhas dúvidas e por me encorajar a superar os desafios encontrados.

Aos membros da banca examinadora, Dra. Ana Javes Andrade da Luz, Dra. Caterina Pavão e Dr. Fabiano Couto Corrêa da Silva, expresso meu profundo agradecimento pela disposição em avaliar e contribuir para este trabalho. Suas sugestões valiosas e críticas construtivas me ajudaram a aprimorar a qualidade desta dissertação.

À minha família, expresso minha imensa gratidão. Seu incentivo e compreensão durante todo o processo foram essenciais para minha motivação e sucesso.

Dedico um agradecimento especial à minha esposa, pois sem você, esta conquista não teria o mesmo valor. Seu apoio, amor incondicional e compreensão constante foram a base fundamental para o meu sucesso neste percurso acadêmico. Você é minha parceira de vida e sou imensamente grato por ter você ao meu lado.

RESUMO

Este estudo foi realizado para investigar as características e o grau de persistência dos *links* da *web*. Pesquisadores como Król e Zdonek (2020) demonstram diferentes graus de persistência de *links* da *web*, esse trabalho busca apontar os fenômenos como link rot e reference rot que causam falha no acesso do conteúdo direcionado por esses *links* da *web*, causando o seu desaparecimento. Koehler (2004) realizou uma das mais extensas pesquisas sobre o tema e apresentou um estudo comparando a persistência de recursos da web. Identificamos o ciclo de vidas dos *links* da web utilizados como referência bibliográfica em teses do Lume. Concluímos essa tarefa mapeando as teses do Lume, identificando páginas de referência bibliográfica nas teses e extraindo os *links* da web das páginas selecionadas. As etapas dessa pesquisa são guiadas pelas orientações encontradas em pesquisas correlatas. Criamos categorias para análise conforme orientações citadas por Dimitrova e Bugeja (2007), demonstrando dessa maneira a persistência dos *links* nas teses selecionadas. Na literatura científica brasileira identificamos uma escassez de informações sobre a persistência de *links* da web quando utilizados em referências bibliográficas. Para realização deste estudo, após a definição do Lume como nosso corpus de pesquisa, realizamos a extração dos *links* da web das referências bibliográficas, uma amostra de 368 teses entre os anos de 2012 e 2021. O conjunto continha 5582 *links* os quais foram testados para sua disponibilidade, sendo considerados acessíveis ou com falha. Os resultados apontam que apenas 48% dos *links* utilizados como referência bibliográfica em 2012 ainda estavam acessíveis, uma meia-vida de 8,02 anos foi estimada para o conjunto estudado. O uso de *links* persistentes e políticas de preservação da web podem contribuir para sua maior persistência.

Palavra-chave: link rot, reference rot, desaparecimento da web, persistência dos *links*, hiperlink.

ABSTRACT

This study was carried out to investigate the characteristics and degree of persistence of web links. Researchers such as Król and Zdonek (2020) demonstrate different degrees of persistence of web links. This work seeks to point out phenomena such as link rot and reference rot that cause failure to access content directed by these web links, causing their disappearance. Koehler (2004) carried out one of the most extensive studies on the subject and presented a study comparing the persistence of web resources. We identified the life cycle of web links used as bibliographic references in Lume theses. We completed this task by mapping Lume theses, identifying bibliographic reference pages in theses, and extracting web links from selected pages. The stages of this research are guided by the guidelines found in related research. We created categories for analysis according to guidelines cited by Dimitrova and Bugeja (2007), thus demonstrating the persistence of links in selected theses. In the Brazilian scientific literature, we identified a lack of information about the persistence of web links when used in bibliographic references. To carry out this study, after defining Lume as our research corpus, we extracted web links from bibliographic references, a sample of 368 theses between 2012 and 2021. The set contained 5582 links that were tested for their availability, being considered accessible or failing. The results indicate that only 48% of the links used as bibliographic references in 2012 were still accessible. A half-life of 8.02 years was estimated for the studied set. Persistent links and web preservation policies can contribute to its greater persistence.

Keywords: link rot, reference rot, web dissapear, links persistence, hiperlink.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS E QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ccTLD	Country Code Top Level Domain
CERN	Conseil Européen pour la Recherche Nucléaire
CGI.br	Comitê Gestor da Internet
CNRI	Corporation for National Research Initiatives
DARPA	Defense Advanced Research Projects Agency
DOI	Digital Object Identifier
Fapesp	Fundação de Amparo à Pesquisa do Estado de São Paulo
FTP	File Transfer Protocol
GPS	Sistema de Posicionamento Global
HNR	Handle.Net Registry
HTML	Hypertext Markup Language
IDF	International DOI Foundation
LANL	Los Alamos National Laboratory
LNCC	Laboratório Nacional de Computação Científica
MCA	Mestrado em Computação Aplicada
MCT	Ministério da Ciência e Tecnologia
RNP	Rede Nacional de Pesquisa
TLD	Top Level Domain
UFRGS	Universidade Federal do Rio Grande do Sul
UNIVALI	Universidade do Vale do Itajaí
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WWW	World Wide Web

SUMÁRIO

1 INTRODUÇÃO

A *world wide web*, ou simplesmente *web*, é uma das mais difundidas aplicações da internet, sendo que uma de suas interfaces é constituída de um protocolo denominado HTTP. Esse conjunto tecnológico permite que, através de um URL, consigamos acessar sites com mídias, informações, notícias, bem como realizar interações sociais. Assim se dá a comunicação pela *web*, uma ferramenta difundida durante os anos 1990, que continua crescendo em termos de usabilidade e abrangência até a atualidade.

A *web* modificou a forma como nos comunicamos, devido a sua capacidade de proporcionar novas interações e às funcionalidades que podem ser desenvolvidas por essa tecnologia. Essas são características que a tornam singular em termos de utilização. A integração com o cotidiano e o conjunto de avanços tecnológicos fazem da *web* uma peça central em nossa relação enquanto atores dentro da sociedade.

Essa dissertação pretende apresentar resultados capazes de fundamentar outras pesquisas, servindo como ponto de referência para o estudo da persistência de *links* da *web*, que em outras palavras é a capacidade de um *link* continuar funcionando ao longo do tempo. Atualmente a *web* tem se inserido como elemento de comunicação científica nos trabalhos acadêmicos, esse fato fica evidenciado ao se perceber que a maioria das publicações, da mais simples à mais complexa produção acadêmica, faz algum uso da *web* como referência bibliográfica.

O aumento do uso da *web* é resultado da sua acessibilidade e facilidade de navegação, permitindo-nos agora acessá-la através de uma variedade de dispositivos eletrônicos, incluindo smartphones, videogames e até mesmo eletrodomésticos. “No Brasil, algumas publicações na última década trouxeram importantes reflexões sobre os usos do digital, disponíveis na internet e na *web*, como fonte historiográfica” (RODRIGUES; ROCKEMBACH, 2023, p. 2). Além disso, a ampla disponibilidade de informações na *web* também contribui para o seu crescimento, fornecendo soluções para nossas necessidades diárias, encontramos uma ampla variedade de informações na *web*, por exemplo, se estamos planejando uma viagem, podemos encontrar informações sobre destinos turísticos, preços de

passagens e opções de hospedagem. Se estamos aprendendo um novo idioma, podemos acessar aulas online, tutoriais e ferramentas de prática. Se estamos realizando uma pesquisa, podemos acessar diversas bases de dados em diferentes locais do mundo. A *web* oferece uma ampla gama de recursos que podem ser úteis em muitas situações diferentes, com seus mecanismos de busca, a *web* tornou-se uma espécie de oráculo do século XXI.

Por outro lado, a *web* nem sempre está lá, estudos dos anos 2000, informam que as páginas na *web* podem sofrer alterações depois de publicadas, sendo que a vida média de um URL é em torno de 10 anos (MUSIANI *et al.*, 2019). Dessa maneira, o conteúdo que buscamos pode se perder, e simplesmente não estar mais no local onde o havíamos encontrado. Isso ocorre basicamente de duas maneiras: o conteúdo desaparece ou se modifica.

Além disso, essa preocupação se acentua quando estamos falando das referências bibliográficas, afinal de contas elas estão lá para garantir alguns elementos básicos na metodologia da pesquisa científica: a reprodutibilidade e a confiabilidade, pois não podemos garantir um bom trabalho sem as bases que o fizeram.

A *web* se tornou um recurso importante e talvez indispensável para pesquisas científicas, não apenas por sua conveniência na recuperação e compartilhamento rápido de informações, mas também por oferecer recursos que a mídia impressa não pode. No entanto, como demonstrado no presente estudo, a internet pode ser um ambiente desafiador para a realização de pesquisas baseadas na *web*, uma vez que as citações da *web* desaparecem rapidamente. Apesar desse desafio, o uso da Internet para identificar informações relevantes e oportunas é inevitável para a maioria dos cientistas e para o público em geral com acesso à *World Wide Web*, uma vez que trabalhos científicos e outros são criados e disponibilizados em formato digital na Internet todos os dias. Essa ideia é reforçada quando analisamos que a investigação na *web* é vital para estudos acadêmicos, pois ela serve como um armazém eletrônico conveniente de dados acessíveis a qualquer momento e em grandes quantidades, aumentando assim o escopo e a amplitude dos estudos realizados.

Existem outros usos dentro do universo acadêmico que essa inconstância da *web* pode afetar, como a publicação na ciência aberta, que realiza a divulgação de material advindo do desenvolvimento das pesquisas, de aplicações disponibilizadas na *web* que se encarregam de servir como apoio aos trabalhos realizados ou até mesmo o fruto resultante deste trabalho. Dessas constatações emerge o questionamento sobre a utilização de *links* da *web* como referências bibliográficas, pois a *web* pode não oferecer a condição necessária para salvaguardar essas publicações citadas para o futuro.

Os questionamentos levantados até aqui culminam no seguinte problema de pesquisa: **Quais são as características e o grau de persistência dos *links* disponibilizados em referências bibliográficas de teses depositadas no Repositório Digital da Universidade Federal do Rio Grande do Sul – Lume?**

Nesta dissertação buscamos compreender a durabilidade dos *links* ao longo do tempo, elencar as estratégias de preservação de *links* utilizadas e identificar possíveis desafios e soluções para garantir a persistência dos *links* e a integridade das referências em um ambiente digital em constante evolução.

Para elucidar essas questões foi realizado uma pesquisa nas bases de dados DOAJ, Scielo e Scopus, visando fundamentar a resposta para este problema. O início da ideia que aqui se apresenta se deu pelo estudo de Ntoulas, Cho, Olston em (2004), no qual os pesquisadores estimaram que “apenas 20% das páginas disponíveis disponível hoje ainda estará acessível após um ano” (NTOULAS; CHO; OLSTON, 2004, p. 2). Essa citação permanece atual como pode ser observado em Costa, Gomes e Silva (2017) que, em seu trabalho, utilizam os dados citados acima para fundamentação de seus estudos, ou na abordagem realizada por Masanès que apresenta a metodologia utilizada em 2004 como uma inovação (COSTA; GOMES; SILVA, 2017; MASANÈS, 2006).

A *web* é “[...] uma mídia penetrante e efêmera onde a cultura moderna em um sentido amplo encontra uma forma natural de expressão” (MASANÈS, 2006, p. 1). De fato, parte por sua natureza de criação, parte pela rapidez com que se dissemina e evolui, a *web* está em constante movimento.

Realizamos uma análise dos *links* da *web* e, para tanto utilizamos as teses publicadas no repositório da Universidade Federal do Rio Grande do Sul (UFRGS),

o Lume. Nas teses foi possível encontrar um corpus de *links* que uma vez estiveram acessíveis e então podemos testá-los para verificar sua persistência, também conhecida como *link rot*[1]. Atualmente, o Lume dispõe de um conjunto que contempla não apenas teses, como também, trabalhos de conclusão de curso, dissertações e trabalhos científicos publicados por alunos e pesquisadores vinculados à UFRGS.

Analisamos os trabalhos publicados a fim de extrair as referências bibliográficas, logo após, examinamos os resultados para criar listas com as referências que apresentaram endereços eletrônicos para direcionamento do conteúdo pesquisado, aferindo posteriormente a sua persistência no ambiente *web* e analisando as características informadas pelas ferramentas de análise utilizadas para testagem dos *links* da *web*.

O referencial teórico, aqui apresentado, foi formulado de modo a construir uma base de explicações dos termos utilizados e das relações que a comunicação na internet pode apresentar com o tema. Após o seu desenvolvimento, outros elementos se acrescentam ao conjunto teórico-metodológico que será aplicado na análise das referências bibliográficas das teses e a historicidade da internet, partes fundamentais para a união dos elementos.

1.1 JUSTIFICATIVA

Como mencionado, a pesquisa realizada por Ntoulas em 2004 na qual o pesquisador concluiu que “apenas 20% das páginas disponíveis hoje ainda estará acessível após um ano” (NTOULAS; CHO; OLSTON, 2004, p. 2) fez surgir o questionamento necessário para a construção do tema central desta dissertação de mestrado.

A internet tem adentrado nossas vidas nos últimos 30 anos, sendo que a criação da *web* no início dos anos 1990 (W3C, 2023) faz com que, em menos de uma década, ocorra um aumento informacional. De fato, nunca se produziu tanta

[1] Link rot, ou deterioração de links, é o fenômeno em que os links da web se tornam inativos ou inválidos ao longo do tempo.

informação como atualmente. Por isso, em 2010, um dos CEOs da Google declarou que dois dias eram suficientes para a produção de um volume de informações tal que a soma de toda a história da humanidade (GOMES *et al.*, 2021).

Essa onda massiva de conteúdo se torna um atrativo a mais para o uso da internet, inserindo-a, desta maneira, como uma revolução, historicamente falando, marcando-se a passagem para uma sociedade informacional como foi descrita por Manuel Castells (1999), na virada do milênio, em sua trilogia A Sociedade em Rede (CASTELLS, 1999). Não obstante, essa revolução global trouxe reflexos também para a vida acadêmica, “[...] o uso da internet para identificar informações valiosas e oportunas tornou-se inevitável para a maioria dos cientistas e do público com acesso à rede mundial de computadores (SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012, p. 1). A internet se tornou um recurso comum de citação em artigos de periódicos, teses e dissertações, fato que se concretiza sendo oportunamente incentivado pela facilidade de acesso e disseminação da informação proporcionada pela comunicação da *web* (KLEIN *et al*., 2014).

A efemeridade dos *links* nos faz cogitar que devemos encontrar meios de contabilizar e, assim, conseguir pesquisar soluções quando do desaparecimento do conteúdo da *web*, pois a *web* nem sempre dará conta de manter disponíveis as informações das páginas armazenadas. Conforme Luz (2022) existem “[...] limitações do ambiente e das ferramentas digitais disponíveis atualmente para promover a preservação de páginas *web*” (LUZ, 2022, p. 12). Brügger (2009) indica que, embora a *web* possa ser considerada um meio de armazenamento de nossa civilização, ela não se preserva para o futuro – ou seja, a velha *web* nem sempre pode ser encontrada na *web* (BRÜGGER, 2009).

> Em 2013, a vida útil média de um URL é de 9,3 anos; aqueles que não sobrevivem mantêm a "podridão do link" (a quebra dos *links*). Um "elo morto" é tanto mais prejudicial quanto pode servir de referência ou mesmo de garantia institucional. (MUSIANI *et al*., 2019, p. 45).

Destacamos aqui as teses e dissertações como uma parte significativa da produção científica, e a *web* está sendo utilizada como referência. Costumeiramente, esses trabalhos têm sido armazenados nos repositórios das instituições que realizam os programas de pós-graduação (MASSICOTTE;

BOTTER, 2017). Soma-se a isso o fato de que a crescente corrente de migração do material impresso para o digital faz com que o recolhimento proveniente da defesa desses trabalhos seja realizado de maneira virtual *(KLEIN et al*., 2014; MASSICOTTE; BOTTER, 2017).

Contudo, embora grande parte do material depositado nos repositórios institucionais possua estabilidade, pois estes fazem uso de identificadores permanentes, como DOI, Perma.cc, Handle, as citações e referências bibliográficas, quando endereçadas para páginas da *web*, podem não atender à mesma expectativa de persistência.

Na verdade, boa parte da *web* não está arquivada e não possui o tipo de ferramenta necessário para sua preservação ou localização. A manutenção da disponibilidade desse conteúdo da *web* se torna uma questão relevante, pois o aumento da incidência de citações da *web* em referências bibliográficas nos trazem a reflexão sobre o seu tempo de durabilidade. Mais do que isso, precisamos entender que existem cobertura distintas de preservação e aspectos políticos nos arquivos e dessa demanda ocorre o desdobramento dos conteúdos mais arquivados da *web* (ROCKEMBACH; SERRANO, 2021).

A despeito dos estudos prévios sobre a persistência de *links*, ainda existe uma lacuna notável na compreensão da extensão e das implicações do problema, especialmente no que diz respeito às referências bibliográficas de teses e dissertações. Almejamos, pois, contribuir com este campo de estudo, permitindo identificar as características das referências bibliográficas utilizadas nas teses publicadas pela Universidade Federal do Rio Grande do Sul.

1.2 OBJETIVOS

Para tanto, formulamos um objetivo geral de pesquisa e outros quatro objetivos específicos, os quais seguem.

1.2.1 Objetivo Geral

- Identificar o ciclo de vida dos *links* disponibilizados em referências bibliográficas das teses de doutorado armazenadas no Repositório Digital da Universidade Federal do Rio Grande do Sul – Lume. Esta pesquisa pretende entender a longevidade e a estabilidade dos *links* da *web* utilizados como referências nas teses acadêmicas.

1.2.2 Objetivos específicos

- Mapear as teses preservadas pelo Repositório Digital da Universidade Federal do Rio Grande do Sul – Lume: Este objetivo envolve catalogar e organizar as teses de doutorado disponíveis no repositório para análise.

- Identificar as fontes de informação da *web* a partir das referências nas teses selecionadas: Este objetivo busca avaliar o tipo de informações da *web* que os pesquisadores estão citando em suas teses.

- Categorizar as fontes de informação da *web* conforme seu conteúdo: Isso envolverá a categorização das fontes da *web* citadas em categorias relevantes com base em seu conteúdo, o que pode ajudar a entender a variedade de recursos da *web* que são usados em pesquisas acadêmicas.

- Demonstrar a persistência dos *links* da *web* das referências selecionadas: Este objetivo procura mostrar a extensão de *link rot* nas referências das teses, destacando a necessidade de melhores práticas para a citação de recursos da *web* em trabalhos acadêmicos.

2 REFERENCIAL TEÓRICO

Como parte fundamental da pesquisa científica, a revisão da literatura é a etapa que inicia todo e qualquer trabalho científico. Nesse sentido, buscamos aqui realizar a integração das diferentes áreas que compõem esse projeto. No intuito de alcançar uma forte base conceitual, iniciamos com a revisão de fontes clássicas sobre o tema aqui apresentado, a durabilidade da *web*, e as bases epistemológicas que fazem da ciência da informação mais que apenas uma área de estudo, mas versam atualmente sobre a sociedade como um todo – uma sociedade informacional.

2.1 BREVE CONTEXTUALIZAÇÃO SOBRE A *WORLD WIDE WEB*

A *web* tem sua gênese no final da década de 1980, porém visando uma melhor compreensão dos fatos que culminaram nessa tecnologia, que hoje é considerada o maior aparato de comunicação da história humana, iremos realizar uma contextualização da sua criação.

No cenário após a segunda guerra, é crescente a disputa entre Estados Unidos da América e União Soviética, fase que conhecemos por Guerra Fria. Surge dessa tensão a necessidade cada vez maior de comunicação. Uma das preocupações dos governos, além do sigilo e criptografia da informação, é o da entrega da mensagem em si. Acreditava-se que um ataque nuclear pudesse interromper as comunicações e, dessa maneira, impedir uma contraofensiva.

Para satisfazer a esta última condição, foi idealizada uma maneira onde vários pontos de comunicação são ligados entre si, perfazendo assim uma rede interligada de vários pontos, ou como são usualmente chamados: nós. Desta forma seria possível fazer a informação chegar ao seu destino independente de algum intermediário estar indisponível (ABBATE, 2000).

Parte dessa ideia o conceito central da internet, a qual começa a ser desenvolvida em meados dos anos 1950, tendo como objetivo ser utilizada para uma rede de comunicação militar.

Esse projeto foi inicialmente criado por uma agência responsável por realizar pesquisas na área de novas tecnologias para defesa, conhecida como *Defense Advanced Research Projects Agency* – DARPA. Esta agência militar do governo dos Estados Unidos da América tem em seu rol de projetos criações de nosso cotidiano. Podemos citar, a título de exemplo, projetos como as tecnologias de reconhecimento de voz e tradução e o Sistema de Posicionamento Global (GPS), sendo que a sua necessidade de redução de tamanho dos equipamentos eletrônicos, acelerou o desenvolvimento de microchips em escala cada vez menor[2]. Neste trabalho nos atentaremos apenas ao desenvolvimento do projeto que viria a ser a internet. Inicialmente esse projeto era denominado de ARPANET (ABBATE, 2000).

No início, a internet foi utilizada quase que em sua totalidade por instituições públicas e de ensino. De fato, as suas atividades comerciais se iniciaram apenas no ano de 1988 (ABBATE, 2000). A internet pode ser definida como um “[...] sistema global de redes de computadores interconectados que trocam dados por comutação de pacotes usando um protocolo padronizado” (W3C, 2023, documento eletrônico).

No ano seguinte ao da abertura comercial da internet, Tim Berners-Lee, físico britânico que na época era contratado pelo CERN (*Organisation Européenne pour la Recherche Nucléaire*, antigo *Conseil Européen pour la Recherche Nucléaire*), com a ajuda de um estudante, escreve o projeto "*Information Management: A Proposal*". Neste documento seria proposta uma rede para gerenciamento de informações gerais sobre aceleradores e experimentos no CERN. Essa rede utilizaria a internet como via para sua propagação. Essa proposta foi um desdobramento do projeto de comunicação desenvolvido por Berner-Lee no ano de 1980. Nesse projeto inicial, a comunicação seria realizada utilizando o conceito de hipertexto para realizar o compartilhamento das informações (BERNERS-LEE, 1998b)

A pesquisa não foi conduzida de acordo com as tradições da ARPANET, mas foi influenciada pela cultura dos hackers dos anos 70. Os pesquisadores se basearam em parte no trabalho de Ted Nelson, que, em 1974 publicou um texto

[2] Disponível em: https://www.darpa.mil/about-us/about-darpa , acesso em 27/04/2023.

intitulado "*Computer Lib*", onde instigou as pessoas a utilizar o poder dos computadores para benefício próprio. Nelson propôs um novo sistema de organização de informações chamado "hipertexto", que consistia em *links* horizontais entre diferentes conteúdos, acompanhados de seus históricos de modificações e diversas versões, os quais seriam conectados através de *links*. Os criadores da *web* expandiram essa ideia pioneira de hiperlink, adicionando tecnologias multimídia para criar também uma linguagem audiovisual (CASTELLS, 1999)

Berners-Lee atribuía que a recuperação e circulação da informação poderiam melhorar se houvesse uma rede no qual um local ficaria responsável pelo armazenamento do conteúdo que, posteriormente, viria a ser compartilhado entre os usuários desse sistema. Este conteúdo compartilhado poderia contar com ligações entre outros locais de armazenamento e, consequentemente, os arquivos destes locais, tecendo dessa maneira uma conexão entre os objetos publicados (BERNERS-LEE, 1998b).

Nasceu assim a *World Wide Web* (www), ou simplesmente *web*. A *web* trata-se, resumidamente, de um espaço de compartilhamento de informação, com a utilização de hipertexto. Realizando, dessa maneira, a representação de informações estruturadas e com diferentes aplicações, que se utilizam de textos, vídeos, imagens ou sons, cuja forma de apresentação é conhecida como *websites* ou sites.

Para a utilização dessa rede se faz necessário alguns requisitos. Um deles é que o ponto onde a informação será depositada seja identificado por um endereço. Esse identificador é conhecido como *Uniform Resource Locator*, ou Localizador Padrão de Recursos, em português, sua sigla em inglês, URL, é comumente utilizada como sinônimo de nome, endereço ou domínio do *website* (BERNERS-LEE; MASINTER; MCCAHILL, 1994).

Outro tipo de identificador comumente associado a utilização da *web* é denominado *Uniform Resource Identifiers* – URI. Os URI são a chave de todo o processo, pois através deles conseguimos acessar os recursos disponibilizados na rede. A *World Wide Web* é um espaço de informação no qual os itens de interesse,

referidos como recursos, são identificados pelos chamados URI os identificadores globais as *Web* (W3C, 2023).

Basicamente o URL é um tipo específico de URI que fornece o local onde um recurso pode ser encontrado e a maneira de acessá-lo, ou seja, vincular a informação que está localizada no endereço indicado pelo URL a um documento (KLEIN *et al.*, 2014). Também chamamos esse texto utilizado no endereçamento URI/URL de link[3].

Um dos padrões básicos da *web* é o *Hypertext Markup Language* – ou HTML como é usualmente conhecido. Como citado anteriormente, a *web* se utiliza de hipertexto para a construção e apresentação da sua informação estruturada em páginas, como as conhecemos hoje (REHM, 2010). A ideia de hipertexto foi concebida mesmo antes da criação da internet e parte do princípio de um texto que pode se ligar a outros textos e materiais para sua referência (TSAY, 2009).

Berners-Lee reconhecia o valor da rede de contatos, mas via uma grande limitação no fato de que a maioria dos usos da Internet ainda estava limitada ao texto, apesar dos computadores pessoais estarem cada vez mais orientados para a imagem. Ele imaginou um sistema que permitisse aos cientistas colaborar e compartilhar dados multimídia de maneira mais fácil. Para isso, ele projetou um novo serviço que rodava os protocolos da Internet, aproveitando o TCP/IP que já havia sido adotado pelo CERN. Berners-Lee também criou um sistema de hipertexto que permitia a ligação de informações localizadas em computadores de todo o mundo, criando uma "rede mundial" de informações.

A ideia de hipertexto já havia surgido na contracultura hacker dos anos 60 e 70, e Ted Nelson já havia proposto um sistema de organização de informações que ele chamava de hipertexto. Berners-Lee e seus colaboradores enfrentaram alguns desafios técnicos para atingir seu objetivo de permitir o uso de documentos multimídia, como imagens, áudio e vídeos, mas conseguiram criar um formato único para documentos em hipertexto, o *Hypertext Markup Language* – HTML.

O HTML foi criado para orientar a troca de informações entre navegadores e servidores da *web*, permitindo que ambos localizassem informações na rede.

[3] Disponível em: https://www.nngroup.com/articles/url-as-ui/ , acesso em 27/04/2023.

Para isso, foi criado um formato padrão de endereço, o URL que especificava tanto o tipo de protocolo de aplicativo que estava sendo usado quanto o endereço do computador que possuía os dados desejados. O URL também permitia o uso de protocolos anteriores, como FTP, Gopher, WAIS e notícias da Usenet, possibilitando o acesso a serviços de Internet mais antigos.

Essa linguagem HTML é transmitida por meio do que se chama *Hypertext Transfer Protocol* – HTTP. O HTTP é um tipo de protocolo com o qual se constrói a interoperabilidade entre os servidores da *web* (SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012). Além dele, no universo da comunicação pela internet existem outros protocolos, com os quais podemos transferir dados entre usuários e servidores. Nestes casos, o que se altera são as suas linguagens e características de uso. Dentre esses diferentes protocolos, podemos citar como um dos mais difundidos o *File Transfer Protocol* – FTP, que serve principalmente para comunicação e troca de arquivos entre servidores (W3C, 2023).

Todos esses requisitos são padronizados, permitindo assim que seja implantada a ideia original da *web*, que consistia em um ambiente de comunicação interoperável. Sendo assim, em 1994, foi fundado o *World Wide Web Consortium* (W3C), uma comunidade internacional voltada para o desenvolvimento de protocolos e normativas que visam garantir o crescimento de longo prazo para a *Web*. Os padrões implantados pelo W3C definem as principais diretrizes do que faz a *World Wide Web* funcionar (W3C, 2023).

A *web* necessita da internet para o seu compartilhamento. Portanto, para falar da *web* precisamos também falar da internet. A internet no Brasil começou a ser desenvolvida em 1989, pela Rede Nacional de Pesquisa – RNP, vinculada ao então Ministério da Ciência e Tecnologia – MCT. Na sua criação a RNP tinha como função realizar o desenvolvimento de redes de internet no país, como foco principalmente em uma estrutura de internet para uso acadêmico.

No entanto, somente em 1992, uma rede de internet integrando 10 estados e o Distrito Federal é de fato implantada no Brasil pela RNP. Com o apoio de

grandes empresas, a RNP contribuiu nos anos posteriores para o desenvolvimento e regulamentação da internet[4].

Convém salientar que outras iniciativas ocorriam de maneira concomitante no Brasil. A Fundação de Amparo à Pesquisa do Estado de São Paulo – Fapesp, ainda em 1988, realizou "[...] a primeira conexão à rede através de uma parceria com o Fermilab, um dos mais importantes centros de pesquisa cientifica dos Estados Unidos" (VIEIRA, 2003, p. 8). Outras iniciativas como a do Laboratório Nacional de Computação Científica – LNCC também tiveram êxito em conexões na mesma época.

A *web* no Brasil é coordenada pelo Comitê Gestor da Internet (CGI.br). Esse comitê foi implementado no ano de 1995 pela portaria interministerial número 147, este órgão orienta o uso e a criação de sites e recursos que utilizam a internet, como, por exemplo, o registro dos nomes de domínios[5] (NIC.BR, [*s. d.*]). Ainda no ano de 1995 a internet começa a ser usada comercialmente, com apoio da RNP, "[...] estendendo seus serviços de acesso a todos os setores da sociedade"[6].

Com a desestatização das telecomunicações no Brasil e o surgimento dos primeiros protagonistas da *web*, como Yahoo e NetScape, o mundo vê o crescimento da *web* aumentar rapidamente (VIEIRA, 2003).

2.2 *PAGE NOT FOUND*: PERMANÊNCIA E EFEMERIDADE DA *WEB*

A *web* foi desenvolvida para ser um local de compartilhamento do conhecimento humano. Originalmente pensada para permitir o relacionamento entre projetos, sua ideia era de tornar acessível as informações sobre as pesquisas

[4] Disponível em: https://www.rnp.br/sobre/nossa-historia , acesso em 27/04/2023.
[5] Os nomes de domínios também têm sua organização própria, não sendo de uso exclusivo da Web, e estão estruturados globalmente em níveis hierárquicos. Os domínios de primeiro nível são chamados de TLDs, acrônimo de *Top Level Domains*; além disso, existem diversos tipos, por exemplo o ".net", o ".com", ".org", entre outros. Os domínios que identificam um determinado país de origem, como o ".br", por sua vez, são chamados de código de país ou *Country Code*. Por conta disso, o ".br" é considerado um *Country Code Top Level Domain* – ccTLD, o domínio de primeiro nível do Brasil.
[6] Disponível em: https://www.rnp.br/sobre/nossa-historia , acesso em 27/04/2023.

em andamentos em diferentes instituições interligadas a rede; criando um ambiente de comunicação e aprendizado.

Com o desenvolvimento da *web*, ela adquire uma natureza informacional para o qual não foi inicialmente pensada. As páginas da *web* passam a ter um caráter de arquivamento do seu conteúdo e, como todo suporte informacional, a *web* também está vulnerável ao desaparecimento, pois existem fenômenos que podem afetar a estabilidade e a durabilidade das informações publicadas na *web*.

> Todo o texto é efêmero. Alguns textos são mais efêmeros do que outros. A *Web* provou ser um dos veículos de informação mais efêmeros e mutáveis. Como as páginas da *Web* e os sites da *Web* são instáveis, os sites da *Web* sofrem erosão. (OGUZ; KOEHLER, 2011, p. 1).

Dessa maneira, a comunicação também é alterada na transformação de como a informação é partilhada. Assim, se antes as cartas demoravam semanas, hoje a comunicação se faz em frações de segundo, nossa memória social está sendo produzida como consequência desses novos aparatos (MELO; ROCKEMBACH, 2021). Todo o depósito de informações globais ancorado na *web* é, sem dúvida, passível de modificação, alteração e perda; nesse sentido, portanto, conhecer a expectativa de vida desse material pode auxiliar nas formações de políticas e orientações para a melhoria do uso da *web* como deposito de informação.

> É curioso como inquietações sobre a memória, manifestas há alguns séculos, continuam a embalar discussões na contemporaneidade. A virtualidade entendida no bojo da *World Wide Web* traz contribuições para o estudo da memória social e também lhe apresenta novas problemáticas. (DODEBEI, 2012, p. 9).

Entre essas novas problemáticas, está a estabilidade com que essa informação é armazenada, uma vez que o material publicado pode ser perdido facilmente. De fato, segundo matéria da Forbes de 2019[7], após uma semana de

[7] LEETARU, Kalev. Preserving Online News in an Ephemeral Web: A Look at Four Months of Global Digital Journalism. Forbes, 05 jan. 2019. Disponível em https://www.forbes.com/sites/kalevleetaru/2019/01/05/preserving-online-news-in-an-ephemeral-web-a-look-at-four-months-of-global-digital-journalism/?sh=6312b1552ce8. Acesso em 02/02/2021

publicação, quase 3% das notícias não estão mais disponíveis; já entre as que estão disponíveis, 93% tiveram ao menos uma alteração no HTML (código utilizado para a construção de páginas *web*) em apenas 24 horas. A mesma reportagem ainda cita que aproximadamente 20% desses artigos tiveram alteração no corpo do texto após transcorrida uma semana.

A utilização de formas de memória auxiliar tem sido verificada ao longo da história. Uma extensão de nossa memória fisiológica que, segundo Candau (2016), faz do homem um ser insatisfeito com o seu armazenamento de informação natural, o cérebro, e assim busca em outras formas o aumento da capacidade mnemônica, utilizando-se de formas auxiliares para gravação de informações. Como exemplo destas extensões, podemos citar a escrita, as inscrições rupestres, a imprensa e, atualmente, um infindo número de novas tecnologias capazes de gravar informações (CANDAU, 2016).

Algumas dessas extensões tendem a desaparecer por sua natureza efêmera, como já demonstrado por outros autores. Na verdade, “[...] vivemos na era da abundância de informações, mas da escassez de memória” (GOMES *et al.*, 2021, p. 2). Assim, mais que um debate, a discussão sobre os arquivos da *web*, sua preservação ou a nossa memória individual e coletiva, apresenta-se como uma necessidade, uma vez que hoje a *web* não é só um meio de comunicação, mas também realiza o depósito da informação que é comunicada.

Ainda nos anos 90 foi iniciado um debate sobre a estabilidade da informação publicada na *web*, o fenômeno do desaparecimento da *web* remonta aos seus primórdios, sendo seu incômodo já apontado por Mary P. Benbow (1998):

> Se as URLs armazenadas como *links* de hipertexto permanecerem as mesmas, esse sistema funcionará bem; no entanto, por vários motivos, as URLs de sites e documentos da *Web* mudam e o usuário é deixado na terra de ninguém cibernética. (MARY P. BENBOW, 1998, p. 248).

Questões sobre a permanência das informações para as futuras gerações ou mesmo os requisitos técnicos para evitar problemas de estabilidade da informação publicada na *web* são estudadas. A durabilidade das informações dispostas na *web* não era conhecida na época e, ainda hoje, não conseguimos estimá-la. Conforme o próprio criador da *web*, Tim Berners-Lee, existem questões

a serem levantadas sobre a possibilidade do desaparecimento das informações e de como as páginas da *web* são criadas e mantidas (BERNERS-LEE; MASINTER; MCCAHILL, 1994; CASTELLS, 1999).

A *web* pode desaparecer por muitos motivos, seja por resultado de arquivos ausentes, intenção dos proprietários ou administradores ou até por um simples código incorreto. Qualquer um desses motivos fará com que o acesso ao arquivo seja interrompido (FETTERLY *et al.*, 2004).

Berners-Lee (1998) esclarece que, para uma melhor utilização da *web*, os URIs devem ser pensados com uma lógica além do seu tempo. Segundo ele o URI deve manter uma estrutura que seja imutável ao longo dos anos, evitando dessa forma a sua modificação que causaria a perda do endereço original com o passar do tempo. Sobre a utilização da data de criação do documento como parte inicial para a estabilidade do URI, Berners-Lee ainda acrescenta: “Essa é uma coisa com a qual é bom iniciar um URI. Se um documento estiver datado de alguma forma, mesmo que seja de interesse por gerações, a data é um bom começo” (BERNERS-LEE, 1998a, documento eletrônico)

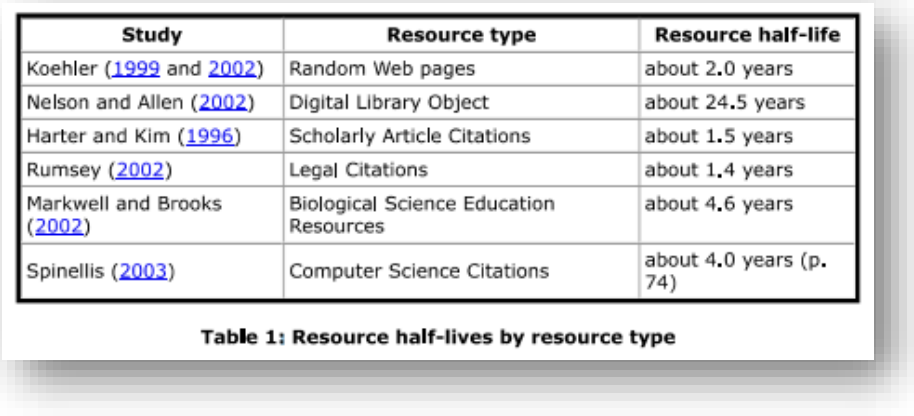
Uma vez que a *web* é considerada efêmera por vários autores (KLEIN *et al.*, 2014; KOEHLER, 2002; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012), podemos compreender, através das pesquisas e estudos desenvolvidos por eles, que a natureza da *web* não considera a preservação com um de seus pilares. A *web* foi desenvolvida como um sistema de comunicação para realizar a troca de informações entre os atores de sua utilização (KLEIN *et al.*, 2014).

Esse caráter instantâneo da *web* fica evidenciado por Koehler em um dos estudos mais extensos sobre a permanência das páginas da internet. De 1996 até 2003, Wallace Koehler monitorou um grupo de 361 URL’s. Os *links* desse conjunto foram acessados e salvos uma vez por semana durante todo o período da pesquisa. Obtendo dessa forma um total de 600 milhões de páginas (KOEHLER, 2004).

Koehler ainda apresenta outros estudos, onde foram constatados diferentes tempos de permanência, os quais ele denomina como meias-vidas, para cada tipo

de material da *web*. Conforme tabela apresentada em seu trabalho (KOEHLER, 2004), reproduzida a seguir na Figura 1[8]:

Figura 1 – Tabela publicada por Koehler (2004) sobre meia-vida de *links*



Study	Resource type	Resource half-life
Koehler (1999 and 2002)	Random Web pages	about 2.0 years
Nelson and Allen (2002)	Digital Library Object	about 24.5 years
Harter and Kim (1996)	Scholarly Article Citations	about 1.5 years
Rumsey (2002)	Legal Citations	about 1.4 years
Markwell and Brooks (2002)	Biological Science Education Resources	about 4.6 years
Spinellis (2003)	Computer Science Citations	about 4.0 years (p. 74)

Table 1: Resource half-lives by resource type

Fonte: KOEHLER (2004).

Existe também uma particularidade que se observou neste conjunto de *links* estudados, os sites são criados e desaparecem, mas em alguns casos esses sites podem voltar à vida (KOEHLER, 2002, 2004; OGUZ; KOEHLER, 2011). Isso significa que um determinado *link* pode não funcionar hoje e voltar a ser acessível daqui a um ano. Essa intermitência pode ter várias explicações, por exemplo, a troca da propriedade de um determinado domínio ou a falta de recursos para um projeto e, consequentemente, a retirada do ar dos *links* disponíveis, que podem retornar com um novo financiamento.

Esse estudo deixa claro que existe uma distinção entre o material que é publicado na *web* e o material para o qual a *web* serve como meio de acesso, conforme sugerido por Nelson e Allen (2002), eles indicam que existem meias-vidas longas para bancos de dados on-line, enquanto sugerem meias vidas muito mais curtas para recursos publicados na *web* (NELSON; ALLEN, 2002).

Os estudos sobre o desaparecimento da *web* têm mostrado que a *web* possui diferentes valores de meia-vida para cada variante. O desaparecimento do

[8] Os tipos de recursos (resource type) apresentados por Koehler(2004), podem ser traduzidos por “páginas aleatórias da web”, “objeto de biblioteca digital”, “citações em artigos acadêmicos”, “citações legais”, “recursos acadêmicos em ciências biológicas”, “citações da ciência da computação”, na ordem em que aparecem.

conteúdo da *web* é comumente descrito em valores de meia-vida[9], esse termo, também empregado em outras áreas do conhecimento, representa o tempo necessário para que metade de um grupo de amostras, no nosso caso os *links*, decaia ou não seja mais acessível, não correspondendo assim ao seu conteúdo original (GOMES *et al.*, 2021).

Essas variantes mencionadas acima podem ser de características próprias do conteúdo. Como verificado em estudos, sites voltados a notícias tendem a enfrentar mudanças mais rápido que sites voltados para outros assuntos (OGUZ; KOEHLER, 2011). Zittrain, Albert e Lessig (2014) descobriram que em, uma amostra da Suprema Corte dos EUA, 49,9% dos URL que foram utilizados como opiniões não direcionavam mais ao seu conteúdo original (ZITTRAIN; ALBERT; LESSIG, 2014).

2.3 OS ARQUIVOS DA *WEB*: DA COMPOSIÇÃO À PRESERVAÇÃO

Em uma definição ampla podemos entender por arquivos da *web*, toda e qualquer informação estruturada que esteja disponível através de um *link* dentro do espaço abstrato que é a *web*[10]. Seja ela um documento, sons, vídeos ou uma página, essa estrutura de dados forma um arquivo. “Quando falamos de arquivo da *web*, temos que imaginar um objeto singular, interativo, fluido e não fixo” (MUSIANI *et al.*, 2019, p. 10).

> A *web*, então, é feita de documentos e arquivos como a internet é feita de cabos e roteadores. Os documentos podem ser sobre qualquer coisa, então, quando passamos a falar sobre o conteúdo dos documentos, deixamos de falar sobre o espaço da informação e todo o universo do discurso humano - e da máquina - está aberto para nós. (W3C, 2023 Design Issues, documento eletrônico).

[9] A formula utilizada para calcular a meia vida é descrita como Nt=N0(½)^t/t ½, onde Nt é o que resta da amostra, N0 é a amostra original, t a variação do tempo ½ é uma constante e t ½ é o tempo de meia vida. Disponível em https://study.com/skill/learn/how-to-find-half-life-explanation.html .
[10] Disponível em (https://www.w3.org/People/Berners-Lee/FAQ.html#What1). Acesso em: 27 abril. 2023.

As inovações decorrentes da *web* fazem com que os seus arquivos tenham um ponto em comum “[...] o formato, já que grande parte das informações produzida na atualidade é nascida digital e difundida na internet” (LUZ, 2021, p. 91). Uma vez que o surgimento da *web* modificou a forma como pensamos a comunicação e a informação; de fato, nossa relação foi radicalmente alterada nos últimos 30 anos. Dentre as muitas transformações que a *web* nos trouxe, está incluída a forma como fazemos ciência (KLEIN *et al.*, 2014).

Considerando que atualmente a comunicação está intrínseca na sociedade, debater a utilização dos arquivos da *web* se torna imprescindível a fim de se evitar uma ruptura social entre o que se produz e o que se comunica – uma divergência da característica democrática da *web* (WOLTON, 2003).

> Os arquivos de páginas da *web* são sistemas informacionais que adquirem, armazenam e preservam conteúdos publicados na internet; contribuem para pesquisas e podem se consolidar como espaços fundamentais para a salvaguarda de informações de uma época. (MELO; NUNES; ROCKEMBACH, 2019, p. 2).

Os arquivos da *web* também podem ser entendidos como uma ferramenta cada vez mais importante para a preservação da história digital da nossa sociedade e cultura. Ao arquivar conteúdo da *web*, somos capazes de captar um instantâneo de um momento específico no tempo, preservando a memória de um momento da história (MILLIGAN, 2019). O arquivamento da *web* pode ser descrito “[...] como um processo que compreende coletar, armazenar e disponibilizar a informação retrospectiva da *World Wide Web* para futuros pesquisadores” (ROCKEMBACH, 2018, p. 9)

A comunicação sempre foi um dos objetivos da *web* e pode ser considerada um dos pilares da ciência. Dessa maneira, a comunicação cientifica tem se digitalizado, utilizando cada vez mais a *web* para sua disseminação (FERREIRA; MARTINS; ROCKEMBACH, 2018).

Mas as condições atuais podem apresentar para a comunicação cientifica um novo desafio. Afinal temos uma mudança no sistema de comunicação acadêmica, se antes existia uma produção quase que inteiramente em papel,

atualmente essa a produção acadêmica se dá em grande parte pela *web* (KLEIN et al., 2014).

Dentro da produção acadêmica observamos que o referenciar é uma parte essencial do trabalho, essas referências têm sido utilizadas com elementos da *web*, sendo que, em alguns casos pode não haver garantia da estabilidade da informação publicada.

> No outro lado, com o crescimento exponencial dos recursos de informação na era da *Web* 2.0 e e-Science, a acessibilidade e persistência de recursos online é uma questão crítica já que está crescendo em importância. A reconstrução, término, fusão, redirecionar e expandir sites da *web* pode significar uma inconsistência nos URLs da *web*. (SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012, p. 179).

Existe a expectativa de que os arquivos da *web*, principalmente os utilizados como fonte de referência ou no discurso acadêmico, possam ser verificados e acessados a qualquer momento (KLEIN *et al.*, 2014). Mas esse anseio pode não se concretizar, uma vez que os arquivos da *web* têm uma condição efêmera (DIMITROVA; BUGEJA, 2007).

Parte da construção da pesquisa científica visa garantir a sua reprodutibilidade e a garantia das fontes utilizadas. Esse vínculo se dá usualmente por meio das referências bibliográficas. Através delas, conseguimos reestabelecer o pensamento que originou a produção científica perfazendo o construto metodológico do autor. Por consequência disso, quando perdemos o acesso a estes materiais, não perdemos somente seu conteúdo, mas também parte do desenvolvimento acadêmico (MASSICOTTE; BOTTER, 2017; OGUZ; KOEHLER, 2011; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

Mas estudos apontam que, mesmo em repositórios, podemos sofrer com o desaparecimento da *web*, pois os *links* utilizados para referência, principalmente em níveis de pós-graduação, por conta de suas atividades de pesquisa e inovação, tendem a apontar para recursos da *web*. Para aqueles *links* que ainda podem ser acessados, mais de 75% levam a conteúdo diferente do originalmente referenciado (MASSICOTTE; BOTTER, 2017).

A *web* ainda carrega consigo a efemeridade de sua criação (KOEHLER, 2004), ao mesmo tempo que o crescente uso da internet como fonte de referência bibliográfica tornou-se inevitável no atual contexto social (SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012), sendo que atualmente não se imagina uma pesquisa bibliográfica sem o auxílio de *web*.

Por esse motivo, existem ferramentas que são utilizadas em publicações para atribuir estabilidade ao material depositado em repositórios e bases de dados; essas ferramentas são conhecidas como identificadores persistentes. Esses identificadores têm a tarefa de permitir que o conteúdo original seja sempre acessível pelo *link* ao qual estará fixado, através de um cadastro prévio em uma das instituições que realizam a custódia desse link, a instituição pode ainda armazenar o conteúdo da publicação da *web*, como o caso do Perma.CC[11], onde o identificador permanente irá apontar para o próprio servidor da instituição (DOI, 2019). As instituições que oferecem esse serviço, como DOI, Handle[12], Perma.cc, realizam a custódia do *link* persistente, comumente substituindo o *link* original por um identificador próprio.

A *International DOI Foundation* (IDF) é uma das instituições que realiza essa tarefa. Concebida em 1997 como uma parceria entre associações da indústria editorial, o *Digital Object Identifier* (DOI), em português Identificador de Objeto Digital, foi criado para realizar o gerenciamento de conteúdo em redes digitais (DOI, 2019). Basicamente, o que o DOI faz é estabelecer um número único para um determinado arquivo, garantindo acesso de longo prazo para esse objeto digital.

Mas para tornar possível essa missão de identificar os arquivos na *web*, a IDF constatou a necessidade de uma infraestrutura. Assim surge desde o início do DOI a parceria com o *Handle.Net Registry* (HNR). O HNR é responsável por desenvolver uma arquitetura de software, o Handle System, capaz de atender às exigências do DOI na forma das normas ISO. “O sistema DOI utiliza o HNR como um componente na construção de um aplicativo de valor agregado, para a identificação persistente e semanticamente interoperável de entidades de propriedade intelectual” (DOI, 2019, documento eletrônico).

[11] Disponível em https://perma.cc/ , acesso em 27/04/2023.
[12] Disponível em http://www.handle.net/ , acesso em 27/04/2023.

Um elemento interessante é que a origem da internet pela DARPA, nos anos 1960, gerou a preocupação com a identificação dos objetos dentro da rede. Assim a mesma DARPA financiou, através da *Corporation for National Research Initiatives* (CNRI), pesquisas que culminaram na criação do Global Handle Registry, atual *Handle.Net Registry*. Atualmente o Handle é administrado pela *DONA Foundation*, mas o CNRI ainda trabalha em conjunto com a *International DOI Foundation* (IDF).

Existem basicamente duas formas de criar *links* persistentes. Uma delas é associar o *link* a uma publicação, essa publicação será salva em um servidor para o qual o *link* direcionará. Esse é o modelo utilizado pelo DOI e Handle. Contudo, esse modelo identifica e cria um *link* persistente apenas para a publicação à qual está vinculado, não preservando as publicações que foram utilizadas nas referências bibliográficas ou os materiais suplementares dessa publicação original. Estando, assim, os *links* utilizados como referencial aptos ao desaparecimento.

Essa preocupação foi determinante na criação do Perma.CC, desenvolvido e mantido pela *Harvard Law School Library*. Nessa plataforma é possível realizar o arquivamento de todo esse conjunto. Na verdade, essa solução surge para resolver uma problemática identificada na suprema corte dos Estados Unidos da América. Em uma amostra de periódicos jurídicos, entre 1999 e 2011, foi constatada uma quebra de aproximadamente 70% nos *links* utilizados para embasamento das análises jurídicas (MASSICOTTE; BOTTER, 2017). Isso significa que os *links* encontrados como referencial não indicavam mais o seu conteúdo original. Essa evidencia ocasionou que as sentenças que continham *links* quebrados como embasamento pudessem perder sua fundamentação com o passar do tempo, pois as referências se tornariam inacessíveis[13].

Portanto, é como se depositássemos o material que se deseja arquivar no Perma.CC, pois, conforme a entidade, ele realiza o arquivamento do URL. Após ser salvo, esse material é identificado por um *link* persistente do Perma.CC, assim, quando o acessarmos, estaremos visitando uma cópia arquivada do original, encontrando o nosso arquivo de referência igual ao instante de seu arquivamento[14].

[13] Disponível em https://perma.cc/ , acesso em 27/04/2023
[14] Idem 4

Existiu ainda uma iniciativa denominada de Hiberlink, desenvolvida na Universidade de Edinburgh, ela foi incentivada pela constatação do desaparecimento dos *links* utilizados como referência em uma determinada base de dados[15] (MASSICOTTE; BOTTER, 2017). Trata-se de um estudo piloto do *Los Alamos National Laboratory* (LANL), que confirmou que até 30% dos *links* da *web* em uma seleção de 400.000 artigos do arXiv.org não funcionavam e que 65% dos restantes *links* referiam-se a um recurso que não foi arquivado e, portanto, em perigo de desaparecer sem deixar vestígios.

O projeto previa o desenvolvimento de um plugin para um gerenciador de referência, especificamente o Zotero, para servir como ferramenta no arquivamento das referências bibliográficas. Consistia em uma ferramenta que guardaria instantâneos dos materiais utilizados no momento em que foram citados, criando, dessa maneira, cópias das páginas ou de outros recursos publicados na *web*. A solução pretendia que, ao final do trabalho de pesquisa, todas as referências utilizadas fossem arquivadas em formato fixo com um arquivo apropriado[16] (MASSICOTTE; BOTTER, 2017). No entanto, o projeto Hiber*link* não foi levado adiante, restando apenas a documentação proveniente do seu desenvolvimento. Segundo o próprio site do projeto, ele foi descontinuado em agosto de 2017.

Com a crescente utilização da *web* como meio de informação, ocorre uma digitalização de nossas vidas. “Esses recursos são importantes tanto para profissionais e acadêmicos, quanto para atividades culturais e sociais” (OLIVEIRA; ROCKEMBACH, 2021, p. 189). Redes sociais, bancos, ensino, o governo, tudo está na *web*. Infelizmente, todas as facilidades de uso e disseminação podem se perder em instantes, seja por ataques hackers, quebra de *links* ou modificação de conteúdo (BARRETO, 2007). Em verdade, a *web* é dinâmica e assim deve ser considerada quando falamos de sua manutenção.

Essas garantias dos identificadores permanentes, as quais foram citadas anteriormente, são amplamente utilizadas em referências acadêmicas e algumas publicações editoriais. O nosso cotidiano ainda fica condicionado às diversas alternativas de preservação espalhadas ao redor do globo.

---

[15] Disponível em https://www.projects.ed.ac.uk/project/isi028 , acesso em 27/04/2023
[16] Disponível em https://www.projects.ed.ac.uk/project/isi028 , acesso em 27/04/2023

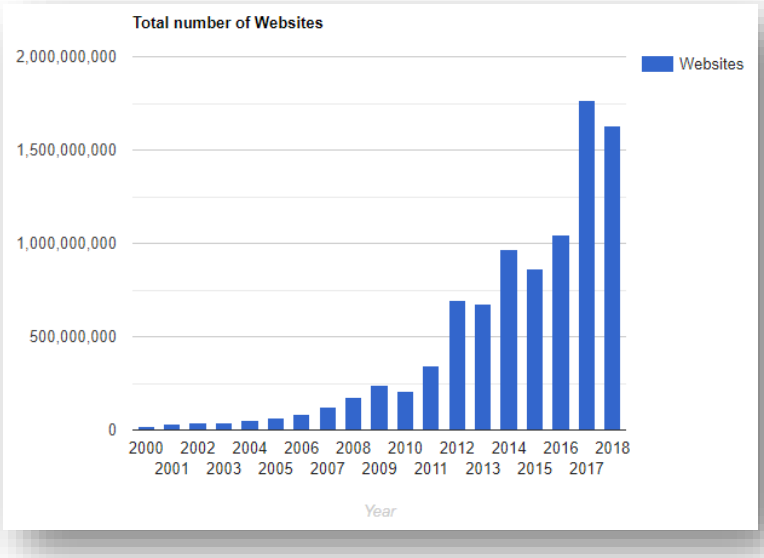
Preservar a *web* é uma tarefa complexa, mas que garantirá a possibilidade de entregar um futuro duradouro aos seus arquivos e "impactar o patrimônio cultural digital da sociedade" (ROCKEMBACH, 2017, p. 147). Nesse contexto, uma das principais dificuldades é a velocidade com que a informação é criada, publicada e transmitida, pois a tecnologia disponível atualmente pode não servir para arquivar a *web* de hoje (ANTRACOLI *et al.*, 2014). Uma das primeiras iniciativas de arquivamento da *web* é a instituição sem fins lucrativos *Internet Archive*, fundada em 1996, que atualmente acumula 735 bilhões de páginas da *web*[17], sendo realizada a captura de aproximadamente 150 milhões de páginas por dia. "Hoje, temos mais de 26 anos de história da *web* acessível por meio da *Wayback Machine* e trabalhamos com mais de 1.000 bibliotecas e outros parceiros por meio de nosso programa *Archive-It* para identificar páginas importantes da *web*" (INTERNET ARCHIVE, 2023, documento eletrônico).

Esses números muito mais que comporem um arquivo, fazem também um alerta sobre a necessidade de soluções que possam minimizar os danos causados pela perda da informação, pois todo esse conteúdo é apenas uma parte da *web*. Rockembach (2019) reforça esse entendimento quando fala da importância e urgência em realizar o arquivamento da *web*, pois vivenciamos um momento crucial em que podemos perder materiais significativos disponibilizados unicamente no formato *web* (ROCKEMBACH, 2019).

Em 2018, estimou-se que a *web* continha um número de quase 2 bilhões de sites, sendo que mais de 75% destes estão inativos[18] (Figura 2), por isso, é necessária nossa atenção imediata ao assunto. Faz-se imprescindível a adoção de políticas que sejam capazes de contribuir para a preservação da *web* e de seus arquivos.

[17] https://archive.org/about/
[18] https://www.internetlivestats.com/

Figura 2 – Total de *web*sites segundo Internetlivestats



Fonte: Internetlivestats (2023).

Em 2016, o número de sites quase dobrou: de 900 milhões para 1,7 bilhão. No entanto, a contagem de sites ativos mais confiáveis permaneceu estável, com cerca de 170 milhões ao longo do ano[19].

Esse dinamismo tecnológico, apesar de ser um dos maiores trunfos da *web*, carrega consigo alguns inconvenientes. De fato, como citado por Rockembach e Pavão, o próprio dinamismo intrínseco das páginas da *web* faz com que conteúdo seja criado e sobreposto a todo instante, e essas transformações, apesar de serem detectadas costumeiramente, constituem o desafio dos arquivos da *web*, uma vez que dificultam a recuperação de longo prazo dessa informação (ROCKEMBACH; PAVÃO, 2018).

Um ponto desfavorável que esse dinamismo pode acarretar é reduzir a chamada *Web Archivability* de uma página da *web*. Essa métrica tenta estipular a taxa em que uma página da *web* pode ser arquivada de uma maneira satisfatória, ou seja, na qual consiga reproduzir de forma idêntica as suas funcionalidades, igual ao momento de seu arquivamento (MELO; ROCKEMBACH, 2020).

[19] Disponível em: https://www.internetlivestats.com/total-number-of-websites/ , acesso em 27/04/2023.

De fato, os vários formatos, características e configurações da *web* tornam as tarefas de arquivamento complexas, portanto, existe a necessidade da aplicação de políticas e padrões para a concepção das páginas da *web* e seu posterior arquivamento (TERRADA, 2022). Os entraves podem se agravar se considerarmos o fator temporal no arquivamento de páginas da *web*. Holub e Rudomino afirmam que “Devido à natureza dinâmica da *web*, este conteúdo está em constante mudança, e um grande número de sites têm uma vida útil bastante curta, de aproximadamente 44 ou 75 dias” (HOLUB; RUDOMINO, 2014, p. 11).

Diante dessa realidade, alguns pesquisadores trazem uma ideia de perda orgânica, como se a *web* se autopreservasse[20], ou então afirmam que apenas as informações realmente importantes seriam preservadas. No entanto, isso não condiz com a realidade, e encontramos sérios impactos sociais advindos do desaparecimento da informação, seja ele intencional ou não.

Observamos a necessidade de se estudar a preservação da informação, mas não somente isso; há uma urgência em verificar os impactos do desaparecimento da *web* e onde podemos encontrar este fenômeno, se de fato ele existe, além de identificar quais são as iniciativas para contê-lo e, principalmente, como será a nossa relação com a memória, pois, do ponto de vista da construção social, é a partir dessa memória que ocorre a formação da nossa sociedade. Afinal, não existe um coletivo sem passado, sem memória, pois é com base nela que construímos nossos laços.

Rockembach e Pavão (2018) dizem que “caso não haja uma preservação digital dos conteúdos produzidos na *web*, muito do que foi desenvolvido neste meio se perderá para sempre” (ROCKEMBACH; PAVÃO, 2018, p. 169). Além disso,

> A preservação de filmes, por exemplo, também foi desconsiderada nos primórdios dos irmãos Lumière, e mesmo nas décadas seguintes. A diferença é que as sociedades não dependiam exclusivamente de filmes para se comunicar. (GOMES *et al.*, 2021 p.2).

[20] Esse termo foi utilizado como um alerta para as informações pessoais disponibilizadas na web, uma vez que elas podem acabar se replicando dentro de seu contexto, retirando dessa maneira o direito ao esquecimento. (GOMES *et al.*, 2021)

Isso significa que, atualmente, para realizar os serviços de nosso cotidiano, precisamos de alguma forma de interação com a *web*. Países de todo o mundo estão progressivamente aumentando a oferta de serviços digitais, de impostos a uma certidão de nascimento, a nossa vida passou a ser digital e, conforme os autores, esse é um caminho de mão única, não existira mais uma sociedade não digital (GOMES *et al.*, 2021).

2.4 PARA ALÉM DO 404: COMO IDENTIFICAR OS PROBLEMAS NOS *LINKS*

Todo esse desenvolvimento tecnológico que a *web* nos trouxe levantou dúvidas sobre o tempo de vida que essa comunicação tem depois que é publicada na *web*. Desde 1996 Koehler estudava tais fenômenos e na virada dos anos 2000 publicou uma série de estudos sobre o tema. Na época, o problema da quebra de *links* na internet era novidade e tratado com certa naturalidade, visto que a *web* era considerada efêmera, ela era entendida como um “braço editorial da internet” (Koehler, 1999). Ainda assim, já existiam os questionamentos de quanto tempo duraria a “memória da *web*”, uma vez que um número cada vez maior de informações estava sendo comunicado por aquele meio (Koehler,1999).

Se Gutemberg permitiu a reprodução de maneira escalável, a *web* consolidou a possibilidade de que qualquer pessoa pudesse publicar e interagir com outras publicações ao redor do mundo, tudo isso de maneira instantânea (WOLTON, 2003).

Para acessar a *web*, utilizamos os *links*, esses *links* direcionam a comunicação para um tipo de protocolo, no caso, estamos interessados no HTTP, desse momento em diante, ocorre uma troca de informações entre o ponto de acesso e o que se deseja acessar. Quando digitamos o *link* e o enviamos ao servidor de destino, este servidor retorna com informações ao mesmo tempo em que solicita outras. Dessa maneira surgem diferentes interações, seja por preferências de usuários ou então alguma personalização do próprio servidor de destino (W3C, 2023 Protocols).

Entre os vários elementos nessa comunicação bilateral, um deles é um número conhecido por *Status-Code*. Esses números são respostas ao pedido de

acesso e informam se houve sucesso ou erro na comunicação, bem como qual o próximo passo a ser seguido para continuidade da comunicação (W3C, 2023 Protocols). As respostas são agrupadas em cinco classes diferentes, cada uma delas diz respeito a uma ação a ser seguida ou interrompida na comunicação pretendida. Na maior parte das vezes não enxergamos estes códigos, uma vez que, quando a conexão é estabelecida com êxito, o que nos é apresentado é a página que foi consultada.

Os números de resposta são formados por três dígitos, com o primeiro dígito indicando qual a sua classe e os dois seguintes a mensagem por ela atribuída. Essa nomenclatura é definida por organizações que realizam a padronização da internet e, consequentemente, da *web*, como o *World Wide Web Consortium* – W3C. A classe 100 é considerada informativa e orienta a continuidade do processo; na classe 200 se encontram as respostas de sucesso no estabelecimento da conexão; na classe 300 ficam os pedidos de redirecionamento, indicando que pode haver a necessidade de alguma tomada de ação; na classe 400 estão as respostas de erro de sintaxe ou falha por se tratar de uma solicitação inválida; a classe 500 é a última e nela obtemos informações sobre erros de serviços ou de acesso decorrentes de falhas em um pedido válido[21]. O Quadro 1 sistematiza esses códigos e suas mensagens de resposta.

Quadro 1 – Status-Code

Código do Status HTTP	Mensagem de resposta	Tradução da mensagem de resposta.
100	Continue	Continuar
101	Switching Protocols	Mudando Protocolos
102	Processing	Processando
200	Ok	Correto
201	Created	Criado
202	Accepted	Aceito
203	Non-Authoritative Information	Não autorizado
204	No Content	Nenhum Conteúdo
205	Reset Content	Resetar Conteúdo
206	Partial Content	Conteúdo Parcial
300	Multiple Choices	Múltipla Escolha
301	Moved Permanently	Movido Permanentemente
302	Found	Encontrado
303	See Other	Veja outro
304	Not Modified	Não modificado
305	Use Proxy	Use Proxy

[21] Disponível em: https://www.rfc-editor.org/rfc/rfc7231#section-6.5.10 , acesso em 27/04/2023.

306	Proxy Switch	Proxy Trocado
400	Bad Request	Solicitação Inválida
401	Unauthorized	Não autorizado
402	Payment Required	Pagamento necessário
403	Forbidden	Proibido
404	Not Found	Não encontrado
405	Method Not Allowed	Método não permitido
406	Not Acceptable	Não aceito
407	Proxy Authentication Required	Autenticação de Proxy Necessária
408	Request Time-out	Tempo de solicitação esgotado
409	Conflict	Conflito
410	Gone	Perdido
411	Length Required	Duração necessária
412	Precondition Failed	Falha de pré-condição
413	Request Entity Too Large	Solicitação da entidade muito extensa
414	Request-URL Too Large	Solicitação de URL muito longa
415	Unsupported Media Type	Tipo de mídia não suportado
416	Request Range Not Satisfiable	Solicitação de faixa não satisfatória
417	Expectation Failed	Falha na expectativa
500	Internal Server Error	Erro do Servidor Interno
501	Not Implemented	Não implementado
502	Bad Gateway	Porta de entrada ruim
503	Service Unavailable	Serviço Indisponível
504	Gateway Time-out	Tempo limite da Porta de Entrada
505	HTTP Version Not Supported	Versão HTTP não suportada

Fonte: RFC Editor.

O debate acerca da durabilidade das páginas e informações publicadas na internet apresenta algumas visões distintas em certos aspectos. Masanès (2006) expôs que alguns autores acreditam que a *web* se autopreservará, enquanto outros autores dizem que vivemos um presente rumo a um futuro sem memória (MASANÈS, 2006). Essa ideia de *web* autopreservada recai sobre o fato de que algumas informações podem enfrentar dificuldade para serem excluídas da *web*. Essa dificuldade tem relação com a multiplicação de arquivos dentro da *web*, indo assim contra o direito ao esquecimento garantido no regulamento europeu de proteção de dados pessoais (MASANÈS, 2006).

Ocorre que os códigos de status HTTP nos informam as respostas em êxito ou erro na conexão com o servidor de destino. As mensagens são simples e sem maiores detalhes, porém, no acesso aos *links*, podemos nos deparar com uma situação em que exista maior amplitude e complexidade no problema encontrado.

Dessa maneira podemos ter uma divisão entre os problemas no acesso aos *links* e os problemas relacionados com o seu conteúdo, formando assim dois grupos. Uma das causas é quando existe um problema com o acesso ao link, *persistence*, e a outra o problema que ocasiona a modificação no conteúdo indicado por um *link*, *constancy* (OGUZ; KOEHLER, 2011).

O primeiro pode ser compreendido como a durabilidade de um *link* com o passar do tempo, esse tempo pode ser de dias, meses, anos, mas convencionou-se na literatura da área a utilização do termo “meia-vida”. O tempo de durabilidade varia conforme o tipo de nível do URL, também sofrendo influência do seu tipo de conteúdo e da área do conhecimento à qual esse URL está vinculado (DIMITROVA; BUGEJA, 2007; OGUZ; KOEHLER, 2011; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

Quanto ao segundo grupo de problemas, Koehler realiza uma definição simples em que explica o que seria a taxa de alteração que um documento sofre no decorrer do tempo. Ou seja, essa taxa compara o conteúdo de uma página da *web* em dois momentos distintos, verificando se houve alguma modificação em sua forma ou conteúdo (OGUZ; KOEHLER, 2011).

Outros estudos citam essas características com outros nomes, como *persistence* e *decay* (DIMITROVA; BUGEJA, 2007), *web constancy* e *permanence* (TYLER; MCNEIL, 2003), *web site acessibility* e *persistence of URLs* (SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

A particularidade de cada fenômeno nos fez perceber que a simples divisão entre problemas nos *links* e problemas de conteúdo poderia ser vaga demais para uma correta compreensão do assunto.

Nos quadros apresentados a seguir está o agrupamento dos problemas de acesso com seus autores e o ano de citação. A terminologia dos problemas nos *links* de materiais da *web* foi dividida por fenômenos de *bit rot* e fenômenos de *link rot* (KRÓL; ZDONEK, 2019). O Quadro 2 exemplifica a nomenclatura por autor e data dos problemas provenientes de *bit rot*.

Quadro 2 – Nomenclatura utilizada na literatura para se referir ao fenômeno de bit rot

TERMINOLOGIA	**AUTOR**

Bit rot	Hayes (1998)
Bit loss	Hudgins (2011)
Bit decay, data rot, data decay and silent corruption	Baker et al. (2005), Bowers (2017), Rouse (2019)
Era of lost data, digital dark age Kuny (1998), Lyons (2016), Whitt (2016), Wernick (2018)	Kuny (1998), Lyons (2016), Whitt (2016), Wernick (2018)
Born-digital, digital-first	Lor and Britz (2012)
Data degradation	Baker et al. (2005)
Digital preservation, digital preservation strategies	Hedstrom (1997), Baker et al. (2005), Schlieder (2010), Deljanin (2012), Lor and Britz (2012)
Document persistence	Lor and Britz (2012), Koehler (2004)
Decaying digital artefact	Lyman and Kahle (1998), Tew (2005), Bowers (2017)
Software rot	Odersky and Moors (2009), Wernick (2018)
Software erosion, software entropy, software bloat	Odersky and Moors (2009), Hildenbrand (2017)
Digital heritage, digital archives, digital storage	Leung et al. (2001), Baker et al. (2005), Knight (2010), Schlieder (2010), Tait et al. (2013)
Digital vellum (by Vint Cerf)	Mottl (2015)
Digital obsolescence	Deljanin (2012)

Fonte: KRÓL; ZDONEK (2019).

Esses problemas não são os únicos a afetar os *links* e seu conteúdo, existem também os causados por *links* defeituosos e a mudança de conteúdo no arquivo para o qual o *link* direcionava. Assim, temos a tabela abaixo, que também relaciona os termos utilizados com seus autores situando a data de citação.

Quadro 3 – Nomenclatura utilizada na literatura para se referir ao fenômeno de *link rot*

TERMINOLOGIA	**AUTOR**
Link rot	Berners-Lee (1998), Nielsen (1998), Benbow (1998) Denemark (1996), Taylor and Hudson (2000)
Link decay	Goh and Ng (2007), Hennessey and Ge (2013)
Broken link, dead link, dangling link	Markwell and Brooks (2002), Kobayashi and Takeda (2002)
Links to nowhere (cyber no man's land)	Benbow (1998)
Decay and failures of Web references	Spinellis (2003)
Cool URI	Berners-Lee (1998)
A link is a promise	Pernice (2014)
Content drift	Burnhill et al. (2015), Zhou et al. (2015)
Never let any URL die	Nielsen (1998)
Reference rot, web references persistence, stability of Web sources	Taylor and Hudson (2000), Lawrence (2001), Rumsey (2002), Parker (2007), Zittrain et al. (2014), Burnhill et al. (2015)
Soft-404s, Web's decay	Bar-Yossef et al. (2004)
URL half-life	Markwell and Brooks (2003), Koehler (2004)

Ephemeral nature of wired progress	Merchant (2014)
Monument to the fragility of the internet, atrophy of links	Tew (2005)

Fonte: KRÓL; ZDONEK (2019).

Entre todos esses problemas que geram perda ou corrupção do conteúdo, os mais comuns são o *bit rot*, *content drift*, *link rot* e *reference rot*.

O *Bit Rot* pode ser um dos fenômenos mais complexos, pois ele muitas vezes trata de um nível de linguagem de máquina com o qual não temos interação direta. A degradação de um software com o passar do tempo ocorre de maneira organizada, pois, novas demandas vão surgindo do uso das tecnologias e do seu desenvolvimento. (KRÓL; ZDONEK, 2019). Os dados armazenados também tendem a sofrer uma espécie de erosão dos códigos binários utilizados na comunicação e no armazenamento da informação nos meios eletrônicos.

Como podemos observar no quadro, sobre os fenômenos de *bit rot*, vários autores debatem o tema. Os termos apresentam pequenas variações em seu significado, mas todos resultam no problema de perda da informação. Essas perdas podem ser ocasionadas por um software que foi descontinuado e por isso não abre mais um determinado formato de arquivo ou então por um erro de memória no qual se perdeu uma parte mínima da construção da página ocasionando a corrupção de seu conteúdo de maneira parcial ou total. Por sua natureza intrínseca aos arquivos e seus meios de acesso, os fenômenos categorizados como *bit rot* são, na sua maioria, irreversíveis (KOEHLER, 2004; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

A *web* pode ser considerada jovem, em um conjunto de amostras 20% das páginas da *web* tinha menos de 12 dias (KLEIN *et al.*, 2014). O fenômeno de *bit rot* pode fazer com que esses *links* não persistam o tempo que deveriam para manter a informação publicada. Concomitante a esses desafios está o fenômeno de *content drift,* o qual pode ser descrito como a alteração no conteúdo de uma página da *web*. Neste problema ocorre que uma página da *web* tem seu conteúdo alterado, resultando na modificação do conteúdo da página originalmente acessada pelo *link* (KLEIN *et al.*, 2014; OGUZ; KOEHLER, 2011). Essa alteração da informação publicada pode abranger fatores estéticos, como a cor da fonte utilizada, ou até a

completa dissolução do material anteriormente apresentado, resultando na perda completa da informação (KOEHLER, 2004; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

Existem páginas da *web* que permanecem disponíveis, mas cujos conteúdos sofreram alterações ao longo do tempo, essa alteração é uma ameaça para a integridade e persistência dos arquivos da *web* ao longo do tempo (KLEIN *et al.*, 2014; MASSICOTTE; BOTTER, 2017). Fetterly observou mais de 150 milhões de páginas da *web* durante um período de 11 semanas. Ele constatou em seu estudo que as páginas da *web* maiores, em termos de tamanho de bytes, tiveram uma taxa de mudança mais frequente e de forma mais significativa do que as páginas da *web* de tamanho menor.

Existe também uma associação da taxa de mudança com o tempo, páginas da *web* que tem alto grau de mudança em um determinado período acabam mantendo essa tendência, e o contrário também pode ser constatado. Essa natureza de estabilidade também pode ser verificada quando analisado o fenômeno de *link rot* (KLEIN *et al.*, 2014; KOEHLER, 2004; OGUZ; KOEHLER, 2011)

O *link rot* é um fenômeno conhecido desde os primeiros anos da *web*, e é visto como um desafio por parte dos pesquisadores. Seus primeiros apontamentos datam de 1998, quando Benbow constatou o fenômeno de *link rot* no acesso de *links* da *web* (KRÓL; ZDONEK, 2019), Tim Berners-Lee já tinha a preocupação com sua incidência sobre o material que era comunicado através da *web*.

Em um estudo chamado de *Cool URIs don't change*, Berner-Lee explica que a modificação dos endereços URI causa danos que podem superar a simples perda da informação, pois as referências são perdidas. “Manter os URIs para que eles ainda existam em 2, 20 ou 200 ou mesmo 2.000 anos claramente não é tão simples quanto parece” (BERNERS-LEE, 1998a, documento eletrônico). Sabendo disso ele indica várias maneiras de pensar o endereço da *web* como algo a ser duradouro e com lógica de longo prazo.

Os problemas de estabilidade da informação afetados pelo fenômeno *link rot*, nos quais o URI ou URL não retorna a um conteúdo, mas sim uma página de erro ou o redirecionamento para outras páginas, também são conhecidos como *broken link*. Em todos os casos, não conseguimos mais acessar o material que

anteriormente foi apontado pelo *link* que estamos utilizando (KOEHLER, 2004; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012). Esse tipo de fenômeno afeta grande parte da *web*, sendo amplamente estudado, conforme demonstra o levantamento de Klein *et al*. (2014), que demonstra que um grupo de pesquisadores extraiu 1.630 URIs de resumos MEDLINE em 2003 e 6.154 URIs únicos em um estudo de acompanhamento em 2007. Eles encontraram em ambos os estudos aproximadamente 20% das URIs inacessíveis. Outra pesquisa extraiu 1.616 URIs de artigos publicados em quatro periódicos da *Ecological Society of America* e descobriram que até 30% dos URIs eram inacessíveis (KLEIN *et al.*, 2014).

Dimitrova e Bugeja (2007) realizaram uma pesquisa em 5 periódicos de comunicação, realizando um recorte temporal de 4 anos, o resultado foi que 39% das referências que apontavam para a *web* não estavam mais disponíveis por problemas de *link rot*.

A crescente utilização de *links* da *web* em referências bibliográficas, principalmente em níveis de pós-graduação, é devida em parte a sua natureza especializada. Esses trabalhos comumente saem em busca de fontes da *web* como blogs, sites de notícias, sites de entidades diversas ficando vulneráveis a ocorrência de *link rot* ou *content drift*.

Quando um desses eventos ocorre com as informações referenciadas em um artigo, uma pesquisa científica ou então como embasamento para um parecer jurídico, é conhecido como *reference rot* (KLEIN *et al.*, 2014). A incidência deste evento acontece, como descrito acima, por uma corrupção do material endereçado, provocando a perda total da informação ou então a sua alteração – logo, apodrecendo as citações (KOEHLER, 2004; MASSICOTTE; BOTTER, 2017; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

O início dos periódicos é atribuído ao século XVII, quando as sociedades científicas começaram a criar suas publicações e difundi-las no meio acadêmico, sendo que este processo de difusão até então era feito por cartas trocadas entre cientistas e agremiações (FREITAS, 2006). Essas publicações permitiam uma circulação da informação produzida de maneira mais rápida e organizada.

A utilização da *web* como meio de comunicação aumentou a circulação de toda a produção cientifica, a reprodutibilidade é um dos pilares da ciência e deve ser garantida também para as fontes que fundamentaram os estudos, mas isso pode não estar acontecendo. Em um estudo realizado nas bases de dados arXiv, Elsevier e PubMed Central, foram extraídas mais de 1 milhão de referências da *web* dos artigos depositados nesses repositórios. A constatação foi que até 27% das referências sofriam de *reference rot* (KLEIN *et al.*, 2014).

A realização de pesquisa sobre o tema tem relevância não somente por permitir a elucidação de uma problemática cotidiana, mas certamente também busca alicerce na importância para o desenvolvimento da sociedade, pois as reformulações da comunidade, das organizações e de identidade têm interferência direta da internet, alterando a forma como se entrelaçam com a vida social (MILLIGAN, 2016).

Nos fenômenos observados pela Ciência da Informação, podemos ter a companhia de outras áreas como a Psicologia, a Filosofia, a Pedagogia, a História, o Direito; enfim, essa troca de saberes está interligada na formação e no desenvolvimento desses fenômenos. Em verdade, presenciamos em cada saber e em cada perspectiva uma ligação com a informação. Essa proximidade a outros campos aumenta a possibilidade de abordagens, visto que novos sentidos são formados na interação entre as questões propostas e suas reflexões quando perpassadas por outras áreas (MARQUES; GOMES, 2020).

3 PROCEDIMENTOS METODOLÓGICOS

No início da pesquisa procurávamos um corpus de pesquisa que servisse de fonte para a obtenção de *links* de páginas da *web*, o quais deveriam ter sido acessados anteriormente. Após a revisão realizada para o referencial teórico, descobrimos que muitas pesquisas se utilizam de repositórios e de bases de dados para a verificação da ocorrência do desaparecimento da *web*.

Atualmente, no Rio grande do Sul, contamos com uma referência internacional em repositório de pesquisa, o Repositório Digital da Universidade Federal do Rio Grande do Sul – Lume. O *"Transparent Ranking: Institutional Repositories by Google Scholar"*, que avalia vários tipos de repositórios como institucionais, de periódicos e de dados, nesse ano de 2023, atribuiu ao Lume a terceira posição entre os melhores repositórios institucionais. Esse ranking se refere ao nível de indexação de repositórios no Google Scholar[22].

A análise e compilação da literatura que encontramos nos fez moldar a metodologia que aqui se apresenta. Dessa maneira, escolhemos o repositório Lume/UFRGS para servir como estudo de caso.

O Lume conta com um conjunto de mais de 247 mil publicações, dentre elas 40.434 teses e dissertações, 34.502 trabalhos de conclusão de curso de graduação, centenas de livros e capítulos de livros, e, ainda, cerca de 75.033 trabalhos apresentados em eventos. O Lume é apresentado como um serviço de informação que realiza a administração da produção intelectual da UFRGS (UFRGS, 2021).

Sendo assim, todo o ciclo da informação depositada no Lume é pensado de modo a garantir a correta gestão do acervo. Desde o seu recolhimento, passando pelo correto armazenamento e organização, toda a sequência de etapas é cuidadosamente pensada, contribuindo, portanto, para a recuperação e o acesso a longo prazo desse material, uma vez que uma de suas principais atribuições é a disseminação do conhecimento (UFRGS, 2021).

[22] https://repositories.webometrics.info/en/node/32

Conforme observado, todo esse material se encontra online, disponível por meio do acesso via *web*. Os recursos pensados no Lume dispõem de cuidados e políticas de preservação, como já citado.

No entanto, nosso estudo pretende verificar se há um problema oculto no material publicado nesse repositório, a possibilidade de quebra dos *links* das referências bibliográficas que foram utilizadas no material depositado no Lume. Esse fenômeno de *reference rot* é observado em trabalhos publicados em repositórios e bases de dados ao redor do mundo. Conforme observado por Isfandyari e Saberi (2010), as citações de arquivos da *web* devem oferecer outros atributos além do seu conteúdo:

> Citar recursos da *web* adequadamente de acordo com um estilo estabelecido é importante na maioria dos campos de assunto e é diferente de citar recursos tradicionais. Além do estilo de citações da *web*, qualidade, autenticidade e sustentabilidade são as questões com documentos na *web*, exigindo a preocupação imediata do profissional da informação. (ISFANDYARI MOGHADDAM; SABERI; MOHAMMAD ESMAEEL, 2010, p. 58).

Pensando nas questões de qualidade, autenticidade e sustentabilidade, este estudo busca alcançar os objetivos propostos para orientar soluções que sejam capazes de garantir o acesso as informações publicadas na *web*, pois, embora exista o risco de futura inacessibilidade das referências de arquivos da *web*, não podemos facilmente evitar seu uso nas publicações (ISFANDYARI MOGHADDAM; SABERI; MOHAMMAD ESMAEEL, 2010).

3.1 A COLETA DE DADOS

Os dados foram coletados entre os meses de novembro de 2022 e fevereiro de 2023, no endereço eletrônico do repositório Lume[23]. Para acesso ao site *web*, utilizamos o navegador Brave no sistema operacional Windows 11.

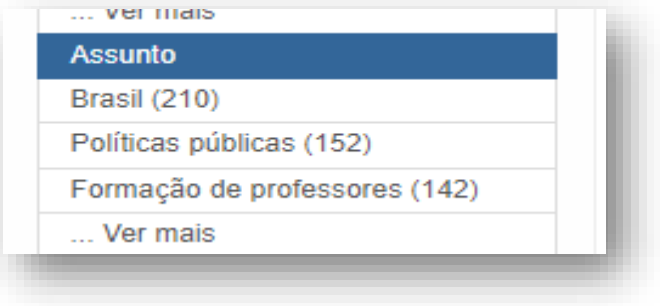
[23] Disponível em: https://www.lume.ufrgs.br/ , acesso em 27/04/2023.

Após acessar a página inicial, delimitamos nosso corpus de pesquisa acessando o menu “Teses e Dissertações”. A página apresentava um menu lateral esquerdo com filtros para os trabalhos. Selecionamos no filtro “Tipo”, o item “Tese”.

Como obtivemos erro nas primeiras tentativas para filtrar diretamente o ano, utilizando os filtros avançados da página do Lume, utilizamos o filtro padrão “Data de publicação”, no menu lateral. Após modificamos diretamente o endereço do *link*, para alterar a data de pesquisa[24]. No texto do *link* onde estão identificados o intervalo, em anos, que serão pesquisados foram modificados os valores para os anos correspondentes a nossa pesquisa[25].

Os filtros da página do Lume apresentaram problemas quando tentamos realizar a filtragem diretamente por eles, pois não conseguimos nenhuma ação da página da *web* ao clicar com o botão direito do mouse sobre “...Ver mais” (Figura 3).

Figura 3 – *Printscreen* “ver mais” do Lume



Fonte: Elaborado pelo autor.

Seguimos para a tentativa de realizar o filtro com as configurações de filtros avançados, como resultado não obtivemos o esperado uma vez que é ampla a lista de assuntos relacionados às teses, poderíamos dizer que correspondem mais a tags de pesquisa que propriamente a assuntos. Uma vez que não foi possível

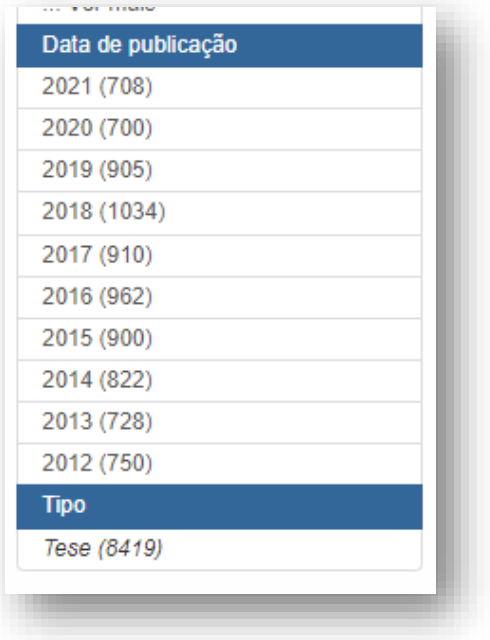
[24] Link original, a parte destacada está com os anos originais do filtro selecionado. “https://www.lume.ufrgs.br/handle/10183/1/discover?query=&filtertype_0=tipo&filter_relational_operator_0=equals&filter_0=Tese&select_ano_inicio_0=&select_mes_inicio_0=&select_dia_inicio_0=&select_ano_fim_0=&select_mes_fim_0=&select_dia_fim_0=&filtertype=dateIssued&filter_relational_operator=equals&filter=%5B**2010**+TO+**2019**%5D”

[25] Link modificado, a parte em destaque foi alterada para o período de nossa pesquisa. “https://www.lume.ufrgs.br/handle/10183/1/discover?query=&filtertype_0=tipo&filter_relational_operator_0=equals&filter_0=Tese&select_ano_inicio_0=&select_mes_inicio_0=&select_dia_inicio_0=&select_ano_fim_0=&select_mes_fim_0=&select_dia_fim_0=&filtertype=dateIssued&filter_relational_operator=equals&filter=%5B**2012**+TO+**2021**%5D”

realizar o filtro pelas áreas do conhecimento, descartamos essa classificação e seguimos com o trabalho de coleta dos dados.

Com os filtros “tese” e “2012 – 2021”, chegamos ao resultado de 8419 trabalhos distribuídos nos anos conforme demonstra a Figura 4 a seguir.

Figura 4 – *Printscreen* do filtro por ano no Lume



Fonte: Elaborado pelo autor.

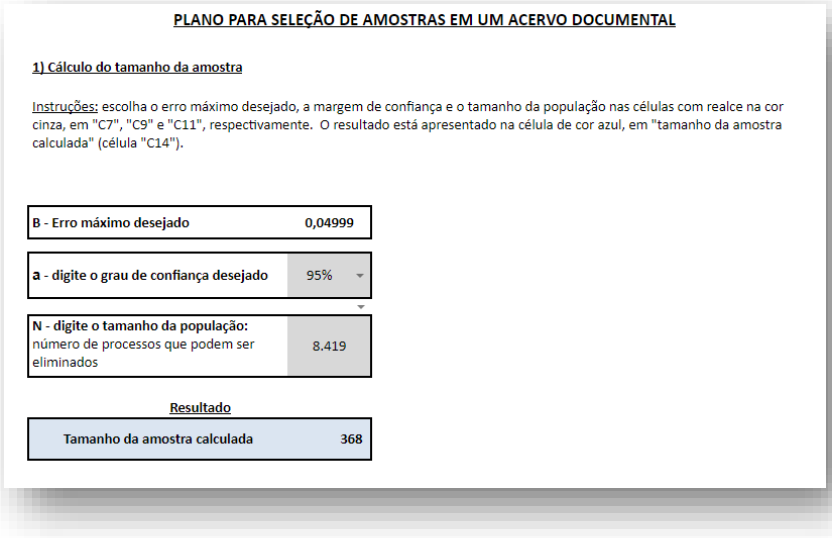
Notamos um acréscimo expressivo de publicações no ano de 2018, com uma diferença de mais de 10 pontos percentuais entre o ano anterior e posterior. O ano de 2020 foi o que teve menos teses publicadas, provavelmente em decorrência da pandemia de SARS-CoV-2 (Covid-19) que culminou na indisponibilidade e fechamento de diversas instituições, causando atrasos nas pesquisas e consequente nos prazos de defesa e publicação das teses.

O número total de teses é definido como o tamanho de nossa população amostral, pois é dentro desse total que encontraremos as amostras para nossa pesquisa (CNJ, 2021). O próximo passo foi encontrar a quantidade de teses que analisaríamos, procurando garantir um grau de confiança e um desvio padrão em que fosse refletido a realidade do todo. Para tanto utilizamos a ferramenta

disponibilizada pelo próprio CNJ para este tipo de cálculo[26]. A escolha das orientações do Conselho Nacional de Justiça para uma abordagem estatística se fez por garantir confiabilidade e praticidade uma vez que esses cálculos podem ser complexos para quem não está familiarizado.

Na planilha de dados, selecionamos um grau de confiança de 95%, com uma margem de erro no valor de 5%. Sendo o tamanho da população de 8419 teses, chegamos assim ao número de 368 teses. Esse número representa a amostra que será extraída do total.

Figura 5 – Plano para seleção de amostrar em um acervo documental

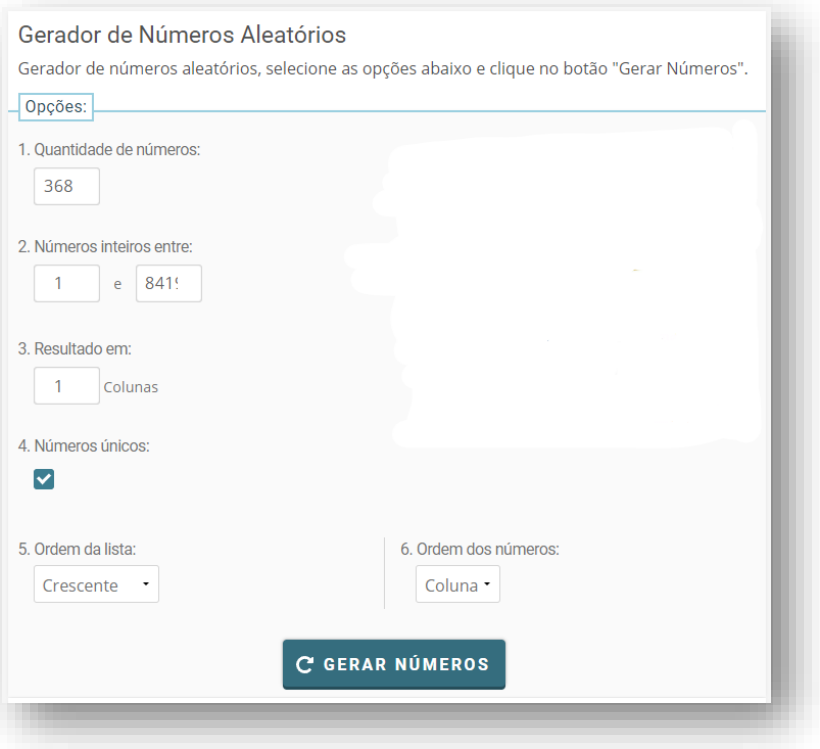


Fonte: Elaborado pelo autor.

Com a informação de quantas amostras seriam necessárias, precisávamos de uma aplicação capaz de gerar números aleatórios. No site 4Devs, encontramos a ferramenta “Sorteador de números aleatórios” (Figura 6). Preenchemos as informações básicas para geração dos números, inserimos as informações sobre a quantidade de números a serem sorteados e o intervalo de números que poderiam ser sorteados, e selecionamos para que números já sorteados não se repetissem.

[26] Disponível em: https://www.cnj.jus.br/wp-content/uploads/2014/02/planilha_calculo_amostra.xls#:~:text=Primeiro%20digite%20o%20n%C3%BAmero%20de,na%20coluna%20D%20ao%20lado.

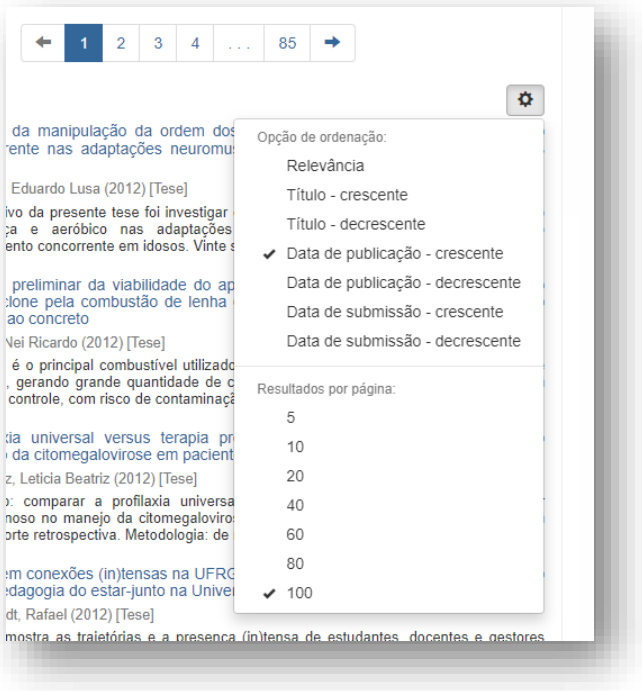
Figura 6 – *Printscreen* do gerador de números aleatórios



Fonte: Elaborado pelo autor.

O processo de geração dos números aleatórios resultou no conjunto que foi utilizado no Lume para coleta do corpo amostral. No Lume, ordenamos as teses localizadas anteriormente por “Data de publicação – crescente”. Dessa maneira, obtivemos as teses ordenadas cronologicamente. O número de resultados por página selecionado foi 100 (Figura 7).

Figura 7 – *Printscreen* dos resultados por página



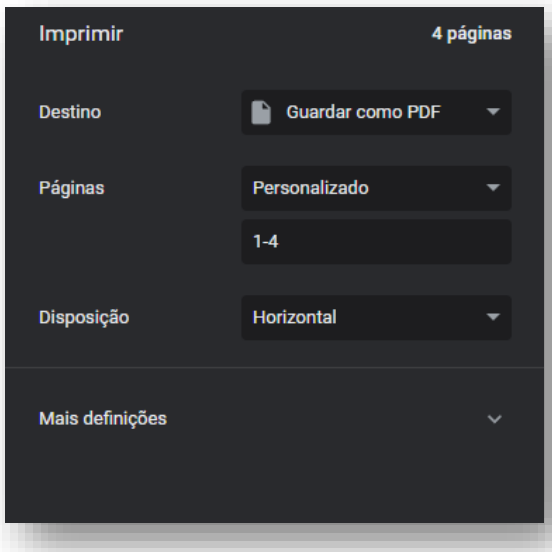
Fonte: Elaborado pelo autor.

Com a lista de números e a ordenação do Lume, iniciamos a coleta das teses e contamos o número da tese dentro da ordenação selecionada para sua escolha. Ao final de todas as 85 páginas, tínhamos realizado o download de 387 trabalhos. Um dos itens continha um erro, portanto, será contabilizado como "*reference rot*".

Grande parte dos trabalhos depositados no Lume são protegidos e, por isso, não permitem a extração de texto ou qualquer outra modificação no arquivo PDF coletado. Por esses motivos, optamos por realizar uma extração manual, salvando apenas as páginas cujas teses apresentavam as referências bibliográficas.

Essa etapa consistia em abrir o arquivo no navegador Google Chrome, localizar as referências bibliográficas no corpo do texto, clicar em "Imprimir", selecionar "Destino – guardar com PDF", na caixa de seleção, escolher "Páginas - personalizado", indicar o intervalo de páginas e depois salvá-las (Figura 8).

Figura 8 – *Printscreen* salvar em pdf



Fonte: Elaborado pelo autor.

Essa tarefa resultou em alguns dados, as teses não seguem uma estrutura de normatização, cada área e cada autor estabelece os elementos orientados por diferentes normas. As referências nem sempre estão descritas como “referencial teórico”, alguns trabalhos inclusive não têm essa divisão.

Dessa forma, encontramos teses com 8 intervalos de referencial teórico, selecionamos todos esses intervalos para incluir como o referencial do trabalho analisado. Outra particularidade é o número de páginas, entre 13 e 500, muitas estão com a informação de texto parcial, mas algumas só apresentam o referencial bibliográfico; outras têm um sumário indicando um número de páginas e efetivamente apresentam menos. Observamos ainda teses com páginas em branco no meio do texto. Ao final dessa etapa, contávamos com um número de 4791 páginas, separadas por ano.

Iniciamos o tratamento desses arquivos utilizando o software Adobe Acrobat Pro. Realizamos uma ação de combinar vários arquivos em um único documento PDF e criamos, para cada ano, um documento com todas as páginas extraídas das teses respectivas. A Tabela 1 sistematiza os dados.

Tabela 1 – Extração das páginas das teses

Ano	Número de teses	Número de páginas extraídas
2012	32	415
2013	36	502
2014	28	383
2015	36	455
2016	44	493
2017	39	549
2018	47	604
2019	43	587
2020	27	385
2021	36	418
Total	368	4791

Fonte: Elaborado pelo autor.

A variação da distribuição de amostras, número de páginas e total de teses por ano ficou relativamente estável entre si. Apesar da escolha aleatória das teses coletadas, esse resultado encontrado confere ao conjunto de amostras uma diferença menor do conjunto populacional original, resultando assim em uma análise mais próxima da realidade.

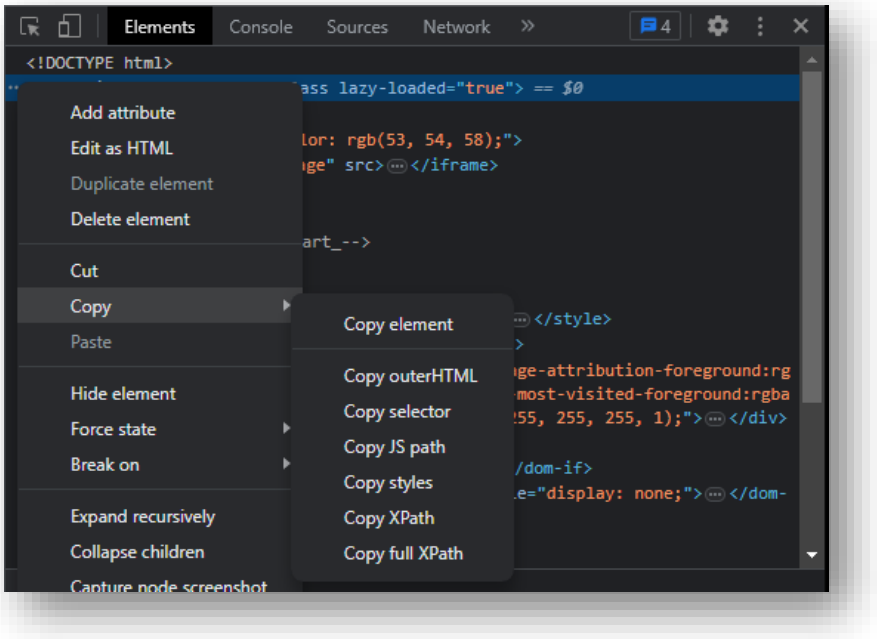
Para realizar a extração dos *links* dos arquivos PDF unificados utilizamos um tratamento denominado OCR, onde as imagens salvas são transformadas em texto, e obtivemos assim arquivos PDF pesquisáveis. O software Adobe Acrobat Pro ainda oferecia a opção de encontrar *hiperlinks* dentro do documento, que também foi utilizada. Por fim salvamos os arquivos em páginas HTML para permitir a utilização de uma ferramenta de extração de *links*.

Ao pesquisar maneiras de extrair os *links* dos arquivos, descobrimos que uma das mais difundidas é transformar o arquivo em um formato HTML. Existem inúmeros conversores de PDF para HTML, por já estar utilizando o software da Adobe conforme descrito acima, optamos por realizar essa conversão diretamente no Adobe Acrobat Pro.

Os arquivos resultantes da conversão foram abertos no navegador Google Chrome. Acessamos a área de Ferramentas de Desenvolvimento clicando com o botão esquerdo do mouse sobre a página aberta e selecionando “Inspecionar”, o que abre uma janela. Nessa janela se encontram as características da codificação

da página, incluindo o código HTML que foi copiado para a área de transferência do computador ao clicar em “*Copy>Copy outerHTML*.

Figura 9 – *Printscreen* da Ferramentas de Desenvolvimento do Google Chrome



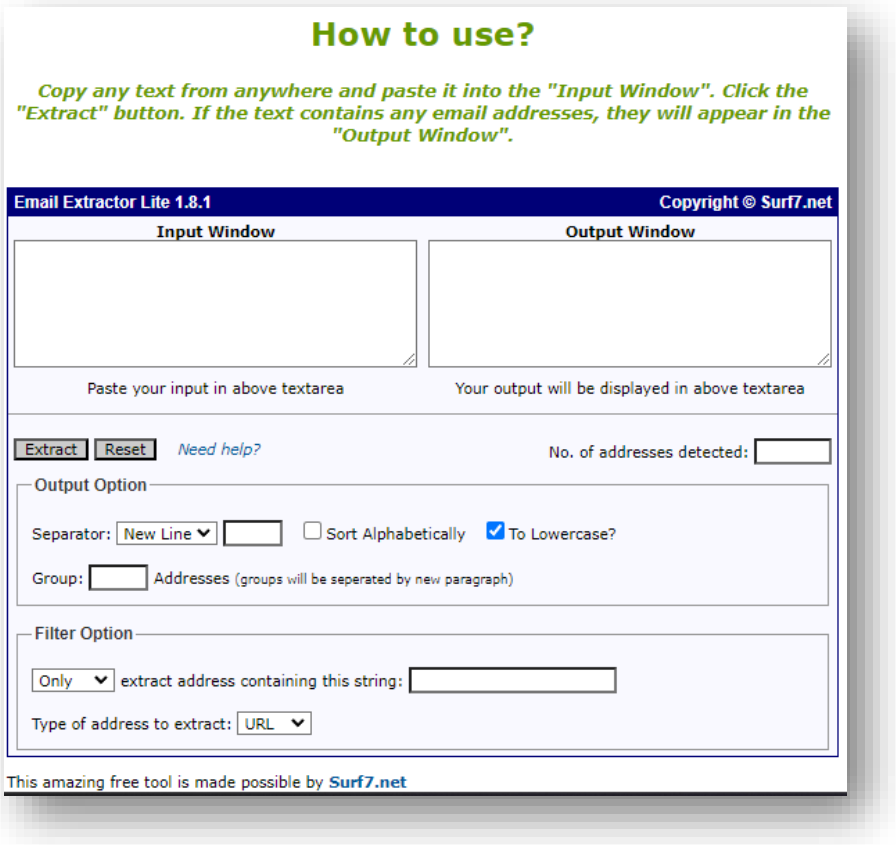
Fonte: Elaborado pelo autor.

Como pode ser observado, no código HTML é apresentado o estilo da fonte, as posições dos elementos na tela e as cores e imagens que devem ser exibidas. Esse código copiado foi colado na página da *web*[27], onde utilizamos a ferramenta *Email Extractor Lite* (Figura 10) para obter os *links* dentro desse código HTML.

Nas opções de saída dessa ferramenta selecionamos como separador uma nova linha e, como filtro, apenas endereços URL. A cada extração, os *links* eram copiados e colados em um documento de texto em formato “.txt”. Esse tipo de arquivo pode ser facilmente modificado para o formato de planilha de dados, além de assegurar um texto mais limpo e sem formatação, o que ocasionalmente poderia interferir nos resultados.

[27] Disponível em: http://eel.surf7.net.my/ , acesso em 27/04/2023.

Figura 10 – *Printscreen* do Email Extractor Lite



Fonte: Elaborado pelo autor.

Uma vez obtidos os *links*, era preciso garantir uma padronização e a não duplicidade para os testes, denominamos essa etapa de normalização. Para realizar essa tarefa, alimentamos uma planilha de dados no software Microsoft Excel, retiramos as duplicidades de *links* e corrigimos os erros de *links* com quebra de página ou espaços em branco. No final desta etapa, copiamos e salvamos os *links* novamente no formato de documento de texto (.txt). Assim chegamos ao nosso corpus de *links* para testagem.

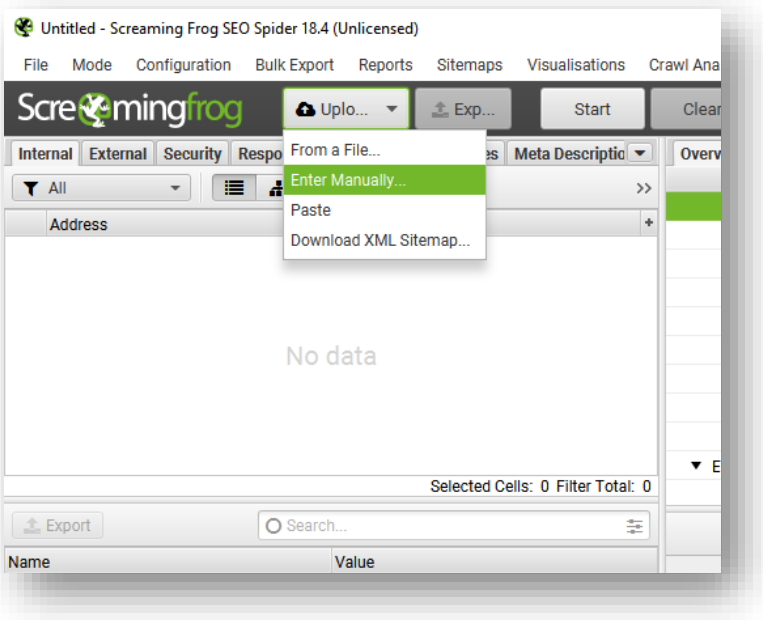
3.2 O TESTE DOS *LINKS*

Para realizar a testagem dos *links*, utilizamos dois programas, o *Screaming Frog*[28] e o *Xenu Link*[29]. O primeiro é desenvolvido por uma empresa inglesa que realiza tratamento de dados da *web* para fins de marketing, como um dos únicos a ofertarem uma versão gratuita, optamos por sua utilização.

O segundo software foi desenvolvido para uso gratuito, justamente para teste de *links* com defeito. A empresa na qual o desenvolvedor do *Xenu Link* trabalha foi citada nos trabalhos de Wallace Koehler como uma referência para teste de *links* defeituosos (KOEHLER, 1999).

Foi criado um fluxo de trabalho no qual primeiramente copiávamos os *links*, normalizados no Microsoft Excel, dos arquivos salvos para um documento de texto com o comando CTRL+A e CTRL+V. Com o programa *Screaming Frog* já iniciado, selecionamos no menu superior o modo LIST. Para inserir o conteúdo copiado, acessamos o menu suspenso no botão "*Upload*" e selecionamos a opção "*Enter Manually*" (Figura 11).

Figura 11 – *Printscreen* do *Screaming Frog*

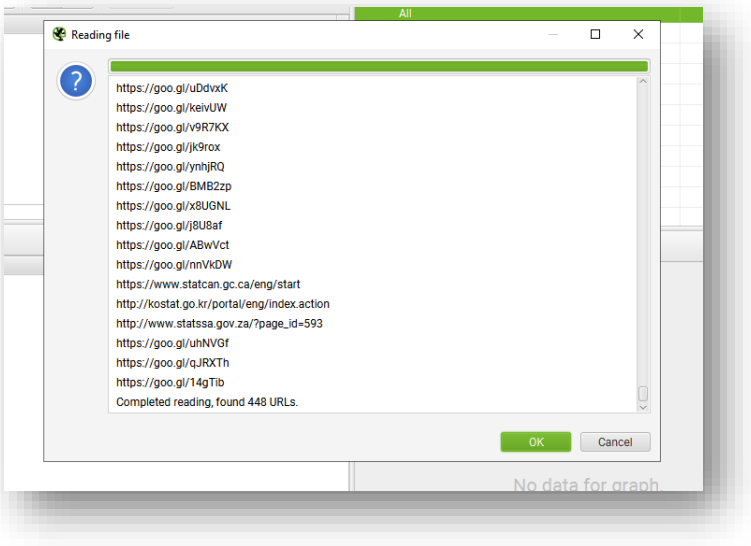


Fonte: Elaborado pelo autor.

[28] Disponível em: https://www.screamingfrog.co.uk/
[29] Disponível em: https://home.snafu.de/tilman/xenulink.html#Description

Isso abriu uma nova janela, onde colamos os *links* e clicamos em prosseguir. Nesse momento, o programa realiza uma validação prévia dos *links*, excluindo os que tenham problemas de digitação ou endereçamento. Os *links* resultantes dessa verificação são exibidos em uma nova tela (Figura 12).

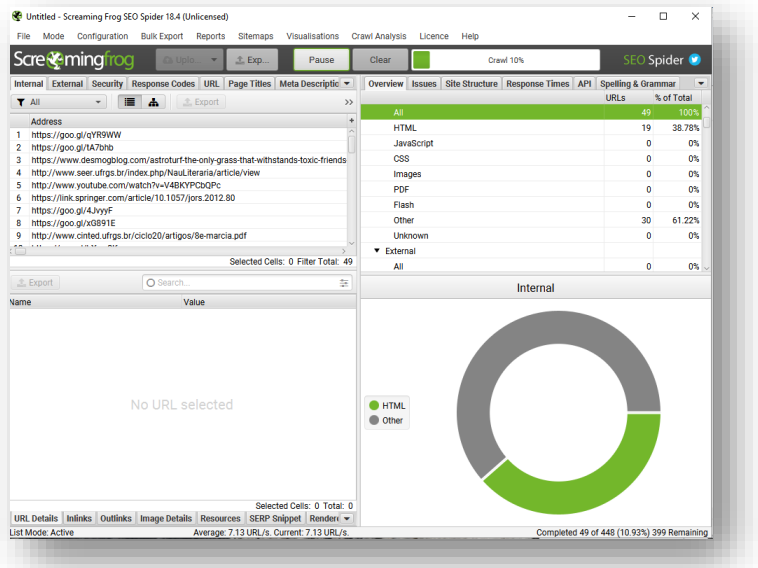
Figura 12 – *Printscreen* do *Screaming Frog* 2



Fonte: Elaborado pelo autor.

Essa verificação acaba por criar mais uma camada de filtro, tornando-se efetiva na remoção de *links* com erros que poderiam indicar um falso positivo para falhas no seu acesso. Iniciamos a testagem dos *links*, conhecida também como *web crawling*. O software realiza a verificação automatizada, não sendo necessária a interação do usuário. Um inconveniente no uso da ferramenta *Screaming Frog* foi a necessidade de dividir os arquivos de documento de texto com mais de 500 *links*, pois esse era o limite para o modo gratuito disponibilizado.

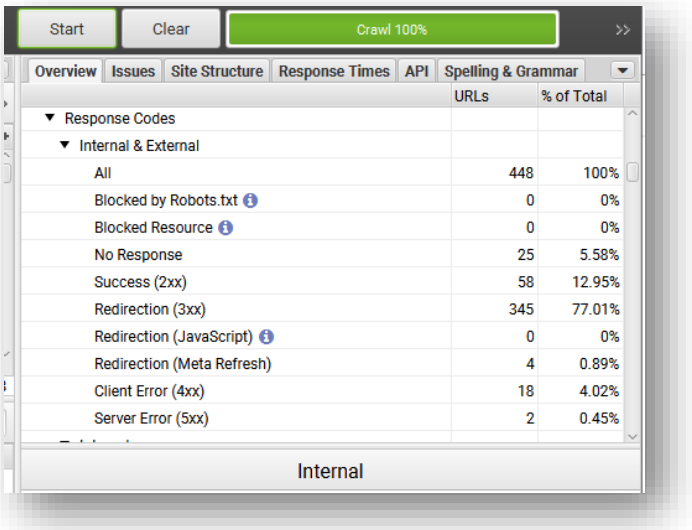
Figura 13 – *Printscreen* do *Screaming Frog* 3



Fonte: Elaborado pelo autor.

Ao final do processo, as abas disponíveis na tela inicial são preenchidas com as informações do teste. Como resultado, obtivemos um resumo do acesso aos *links*. Para a nossa pesquisa, o foco é o código de resposta HTML.

Figura 14 – *Printscreen* do *Screaming Frog* 4



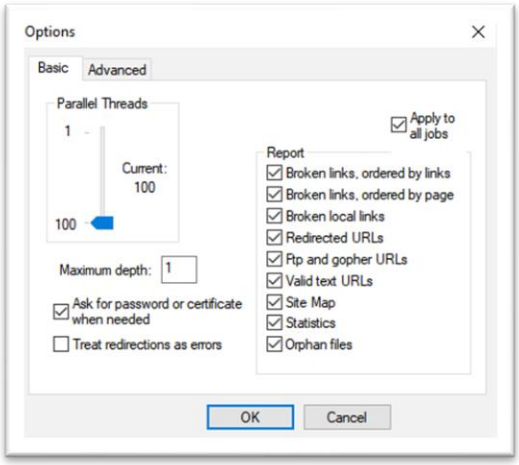
Fonte: Elaborado pelo autor.

Exportamos essas informações em uma planilha de dados e identificamos por ano, utilizamos esse processo para realizar o tratamento dos dados recuperados do teste dos *links*. Identificamos, durante a fase dos testes preliminares, que o programa *Screaming Frog* faz uma tentativa de acesso e aguarda o retorno do servidor. Se o servidor redireciona o *link*, este *link* entra na classificação de código de resposta HTML 300, sendo tratado como um *link* funcional. Por esse motivo utilizamos o software *Xenu Link*, constatamos que seria necessária uma retestagem dos *links* que foram classificados com um dos códigos HTML 3XX, pois esses mascaravam os erros de acesso, indicando apenas o redirecionamento.

O *Xenu Link* realiza uma varredura de modo a investigar os *links* dentro dos *links*, se ele fosse utilizado para todos os testes, as páginas com código HTML 2XX, também teriam os *links* internos testados, provocando um aumento no número de *links* e, consequentemente, um erro na amostra. Como os *links* do código HTML 3XX já fazem um redirecionamento, o *Xenu Link* tratou essa instrução para redirecionamento como o acesso de nível 0, e o *link* ao qual foi redirecionado como o de nível 1. Dessa maneira foi possível testar se o redirecionamento estava realmente indicando um *link* ativo ou com erro.

Os dados resultantes do *Screaming Frog* foram tratados no Microsoft Excel. Dividimos os *links* pelo código HTML, salvamos o conjunto de *links* com códigos HTML 3XX como documento de texto (.txt), esses *links* foram inseridos no *Xenu Link* na opção *file>open list*. O programa inicia automaticamente após a inserção dos dados. Para esse processo de “*web crawling*” configuramos o *Xenu Link* para não tratar os redirecionamentos como erro e prosseguir para o próximo nível, ou seja, seguir o redirecionamento para acessar o novo *link* (Figura 15).

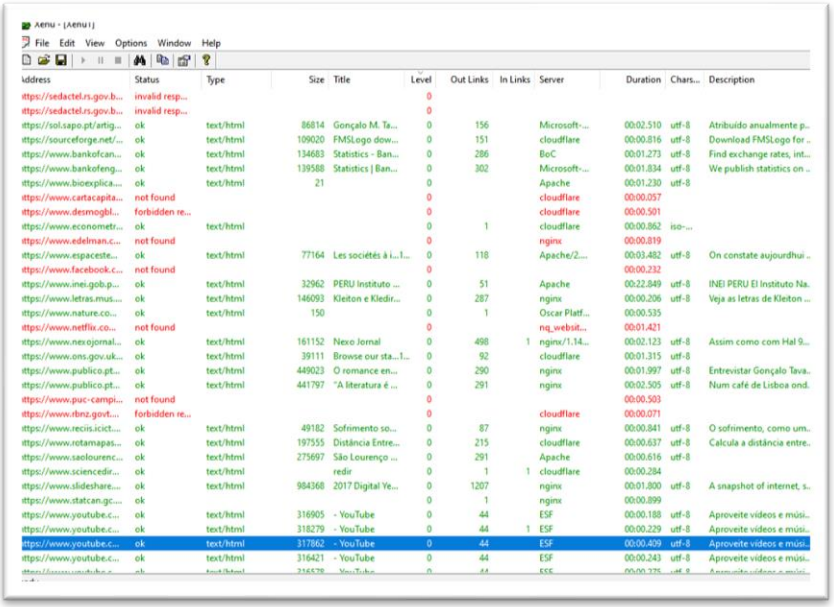
Figura 15 – *Printscreen* do *Xenu Link*



Fonte: Elaborado pelo autor.

Finalizado o processo de verificação dos *links*, o *Xenu Link* exibe a tela de resultado. Para realizar o tratamento desses dados, foi salvo um documento de texto (.txt) separado por virgula, o que nos permitiu alimentar facilmente uma planilha de dados.

Figura 16 – *Printscreen* do *Xenu Link* 2



Fonte: Elaborado pelo autor.

Logo após as etapas realizadas iniciamos a tabulação dos dados recuperados. Utilizamos o programa Microsoft Excel para realizar o cruzamento das

informações e, consequentemente, alcançar os objetivos dessa pesquisa. Criamos uma Tabela com o número total de *links*, distribuídos por ano (Tabela 2).

Tabela 2 – Números de teses e *links* distribuídos por ano

Ano	**N teses**	**N *links***
2012		
2013		
2014		
2015		
2016		
2017		
2018		
2019		
2020		
2021		
Total		

Fonte: Elaborado pelo autor.

A letra “N” maiúscula é utilizada para representar a palavra “número”. A palavra “tese” na segunda coluna se refere aos trabalhos que foram selecionados de forma aleatória para compor a pesquisa. Os “*links*” na terceira coluna são o número total de *links* extraídos das teses antes da normalização.

Ainda para esses dados, criamos outras duas tabelas, uma com o número de páginas de referências extraídas das teses agrupadas por ano (Tabela 3).

Tabela 3: Números de páginas e *links* por página distribuídas por ano

Ano	**N páginas teses**	***links* por páginas teses**
2012		
2013		
2014		
2015		
2016		
2017		
2018		
2019		
2020		
2021		
Total		

Fonte: Elaborado pelo autor.

E outra tabela relacionando a nossa amostra com o número total da população da pesquisa de teses no Lume separadas por ano (Tabela 4). Esse cruzamento de dados mostrará a distribuição das teses no nosso corpo amostral e o total publicado.

Tabela 4 – Relação entre corpus e população distribuída por ano

Ano	**N teses corpus**	**N teses população amostral**	**Distribuição corpus (%)**	**Distribuição população (%)**
2012				
2013				
2014				
2015				
2016				
2017				
2018				
2019				
2020				
2021				
Total				

Fonte: Elaborado pelo autor.

Dessa maneira, conseguimos observar a relação de distribuição entre a nossa amostra e o total. Os dados sobre os *links* extraídos foram explicitados da seguinte maneira, também em forma de planilha (Tabela 5)

Tabela 5 – Tipo de arquivo acessado pelo *link* distribuído por ano

Ano	**N *links***	**html**	**pdf**	**outros**
2012				
2013				
2014				
2015				
2016				
2017				
2018				
2019				
2020				

2021				
Total				

Fonte: Elaborado pelo autor.

Nessa apresentação, conseguimos verificar os tipos de arquivos para os quais os *links* direcionavam. Esses dados foram obtidos a partir do *Screaming Frog*. Formulamos outras duas tabelas, uma para inserir os dados referentes aos códigos HTML recuperados e outra para inserir os principais gTLD[30] encontrados nos *links* testados (Tabela 6).

Tabela 6 – TLD'S do corpo amostral distribuídos por ano

	Total									
gov										
org										
com										
net										
doi										
handle										

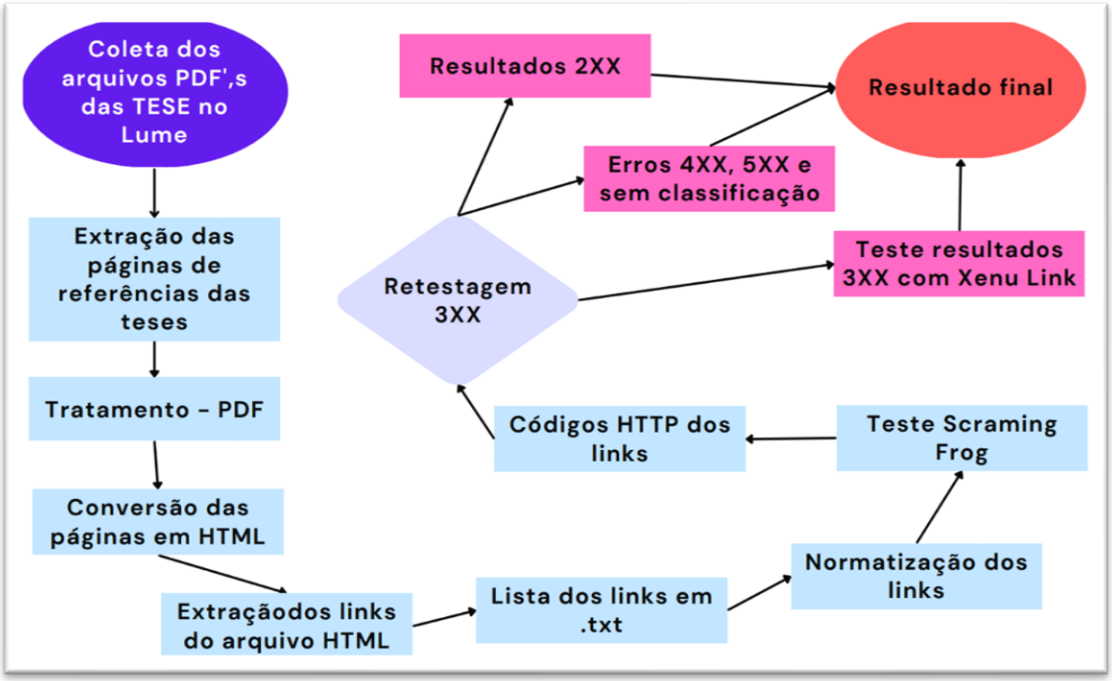
Fonte: Elaborado pelo autor.

O último passo a ser realizado foi a inserção dos dados de número total de *links* por ano e de *links* que continuavam ativos e funcionando na fórmula para cálculo da meia-vida dos dados após os testes.

Para melhor compreensão da sequência de procedimentos que compreende desde a coleta dos arquivos em PDF no Lume até o resultado final da análise, foi elaborado um Fluxograma que sistematiza os procedimentos (Figura 17).

[30] Falaremos sobre esse conceito na subseção deste trabalho "4.1 A categorização dos resultados".

Figura 17 – Fluxograma do processo de testagem dos *links* da *web*



Fonte: Elaborado pelo autor.

Com base nos procedimentos metodológicos demonstrados, apresentaremos a seguir os resultados de pesquisa a partir da análise das amostras do Lume.

4 RESULTADOS

Extraímos 368 amostras das 8419 teses identificadas no Lume utilizando os filtros de pesquisa. Esse conjunto rendeu um total de 5625 *links* que, após serem normalizados, resultaram em 5582 amostras, as quais foram testadas para sua persistência e cujos resultados são aqui apresentados.

4.1 A CATEGORIZAÇÃO DOS RESULTADOS

Como ponto de partida para a análise dos resultados, separamos os dados obtidos de modo a permitir a observação dos diferentes aspectos sobre o desaparecimento dos *links*. O balizamento que definiu as categorias foi fundamento nos estudos sobre o tema que analisamos no referencial teórico. Além dos dados categorizados, são apresentadas outras informações resultantes das análises e que podem servir para pesquisas futuras, mas que não serão aprofundados aqui.

A primeira categoria de análise foi pensada para conseguir estabelecer o quanto nossa amostra divergia da população amostral em termos de número total por ano. Assim, foi possível estabelecer qual a aproximação da amostra com a realidade, definindo um peso diferente para cada ano se necessário.

A segunda categoria faz menção aos códigos HTML, dessa maneira, conseguimos observar a resposta dos *links* da *web* quando realizada a tentativa de acesso a eles. O código de status HTTP orientou os próximos passos da pesquisa, identificando os erros de acesso e quais *links* da *web* devem ser testados novamente.

Na terceira categoria, trazemos o conceito de *Top Level Domain* – TLD. O termo TLD se refere ao tipo original de extensão de domínio, são as letras após o nome da página da *web*, o primeiro TLD criado foi o “.com”, quando o TLD não é conjugado com um código de pais ou território, é denominado genérica – gTLD.

Entre as principais extensões genéricas, estão incluídas .com, que é a mais comum em hospedagens; .gov, usada para sites governamentais; .net, uma

extensão conhecida, mas menos usada atualmente e .edu, que geralmente indica sites de instituições educacionais.

Atualmente os gTLDs são chamados de "Internacionais", pois geralmente indicam sites que podem ser acessados em todo o mundo. O registro dos gTLDs é efetuado pela instituição ICANN[31].

O código de duas letras após o gTLD, atribuído a todos os países ou territórios do mundo é conhecido como Código de País. O *Country Code Top Level Domain* – ccTLD são domínios geralmente registrados por empresas ou organizações, que estão diretamente relacionados com o país correspondente. No Brasil, o Registro.br realiza o trabalho de registro e manutenção dos domínios “.br” (registro br). Alguns exemplos de Códigos de País são .br (Brasil), .ar (Argentina), .uy (Uruguai), .ca (Canadá) e .in (Índia).

Autores indicam que existe uma conexão entre a persistência dos *links* da *web* e o seu TLD correspondente, conforme esses estudos, existe uma predisposição para domínios .gov e .org serem mais estáveis, enquanto outros, como .net ou .int, têm uma volatilidade maior (DIMITROVA; BUGEJA, 2007; KOEHLER, 2004).

O último estudo nos *links*, extraídos das teses para essa pesquisa, foi categorizado como a quantificação da sua meia-vida distribuída por ano, e o cálculo da mediana de todos os 10 anos estudados no corpus de pesquisa. A mediana é uma taxa média na qual se exclui o valor mais alto e o valor mais baixo do conjunto, para que seja obtida uma média com um desvio menor em relação a uma média simples.

4.2 APRESENTAÇÃO DOS DADOS

Demonstraremos aqui os dados obtidos nas análises e testes efetuados durante a pesquisa. As quatro categorias de dados analisados do conjunto são

[31] https://www.icann.org/history

apresentadas em formato de tabelas, em conjunto com informações resultantes do processo de pesquisa.

A distribuição de *links* por ano nos conjuntos corpo amostral (teses amostra na tabela) e população amostral (total de teses na tabela) ficou relativamente estável, conforme podemos verificar. Não houve uma variação significativa nos números encontrados, conforme verificamos na terceira e quarta coluna de dados (Tabela 7).

Tabela 7 – Resultado da extração das páginas das teses

Ano	**Teses amostra por ano**	**Teses total por ano**	**Distribuição amostra por ano**	**Distribuição total por ano**
2012	32	750	9%	9%
2013	36	728	10%	9%
2014	28	822	8%	10%
2015	36	900	10%	11%
2016	44	962	12%	11%
2017	39	910	11%	11%
2018	47	1034	13%	12%
2019	43	905	12%	11%
2020	27	700	7%	8%
2021	36	708	10%	8%
Total	368	8419	100%	100%

Fonte: Elaborado pelo autor.

Na distribuição de *links* por teses e *links* por páginas de referência extraídas das teses, podemos observar um aumento na densidade de *links* por tese e seu uso por páginas no decorrer dos anos.

Tabela 8 – Resultado relação entre corpus e população distribuída por ano

Ano	**N teses**	**N *links***	***links* por teses**	**N páginas teses**	***links* por páginas teses**
2012	32	108	3,38	415	0,26
2013	36	434	12,06	502	0,86
2014	28	219	7,82	383	0,57
2015	36	347	9,64	455	0,76
2016	44	616	14,00	493	1,25
2017	39	667	17,10	549	1,21
2018	47	961	20,45	604	1,59

2019	43	759	17,65	587	1,29
2020	27	698	25,85	385	1,81
2021	36	816	22,67	418	1,95
Total	368	5625	15,29	4791	1,17

Fonte: Elaborado pelo autor.

Em 2019, notamos um decréscimo no uso de *links* como referência (Tabela 9), quando analisamos o tamanho do corpo amostral, é mais aceitável que esse número tenha sido influenciado por características intrínsecas das teses extraídas e esse dado seja resultante de um desvio, assim como notamos um aumento do número de *links* na amostra no ano de 2013.

Um dos resultados obtidos no software *Screaming Frog* foi o tipo de arquivo acessado diretamente pelo link. Essa informação pode ser relevante quando pensamos em termos de preservação da *web*.

Tabela 9 – Resultado tipo de arquivo acessado pelo *link* distribuído por ano

Ano	**html**	**pdf**	**outros**
2012	55%	1%	44%
2013	58%	1%	42%
2014	57%	3%	40%
2015	42%	4%	54%
2016	54%	2%	44%
2017	61%	6%	33%
2018	50%	3%	47%
2019	66%	6%	28%
2020	81%	2%	17%
2021	79%	5%	16%
Total	60%	3%	36%

Fonte: Elaborado pelo autor.

Para a verificação dos códigos HTML, realizamos a divisão utilizando as classes dos próprios códigos, apresentando os resultados divididos como segue.

Tabela 10 – Resultado código de status HTTP distribuído por ano

Ano	**5XX**	**4XX**	**3XX**	**2XX**	**Outros erros sem acesso**
2012	3	30	38	13	23

2013	9	117	164	48	93
2014	2	65	82	23	42
2015	0	101	175	31	32
2016	6	114	395	54	42
2017	7	158	294	144	63
2018	5	310	367	197	77
2019	6	139	324	210	75
2020	2	192	416	70	14
2021	2	152	462	177	17
Total	42	1378	2717	967	478

Fonte: Elaborado pelo autor.

Observamos uma predominância de erros (Tabela 10) de classe 4XX, o que condiz com os estudos encontrados. As classes 2XX e 3XX são as páginas que estão ativas e obtiveram ou um acesso direto e bem-sucedido, 2XX, ou um redirecionamento para uma página ativa, 3XX. A coluna “outros erros sem acesso” diz respeito a falhas relacionas a problemas de servidor não relacionados na classe 5XX, problemas de resolução de *proxy* no acesso ao servidor e outras falhas não enquadradas nas classes 4XX e 5XX.

Para construção da tabela (Tabela 11) comparando os TLDs, utilizamos as categorias do Registro.br para verificar os TLDs mais usados. Dessa maneira, trabalhamos com os dois maiores TLDs entre as 3 maiores categorias (pessoas jurídicas, genéricos, profissionais liberais). Notamos que os TLDs mais comuns no Brasil também figuram como os mais comuns no mundo (DIMITROVA; BUGEJA, 2007; KOEHLER, 2004).

Inserimos na busca pelos TLDs os termos doi e handle, o identificador DOI utiliza o TLD .org nos seus *links*, já para o Handle, encontramos frequentemente o TLD .net. Após teste, decidimos utilizar somente as categorias pessoas jurídicas e genéricos para busca dos TLDs, pois a categoria “profissionais liberais” não apresentou resultados; da mesma forma, o TLD “.edu” não foi estudado por não ter número significativo dentro do conjunto da amostra. Optamos por retirar o código de país para a busca, dessa forma os TLDs podem ser de territórios ou internacionais.

Tabela 11 – TLD's do corpus distribuídos por ano

	2012	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
gov	22	90	62	34	158	99	87	130	41	138
org	20	94	45	141	222	143	182	240	549	432
com	19	107	34	76	117	159	324	190	36	70
net	1	5	3	3	14	23	17	17	4	5
doi	2	0	3	74	13	19	5	34	90	32
handle	1	3	2	3	3	9	1	5	0	3

Fonte: Elaborado pelo autor.

Essa mesma busca foi realizada nos *links* que apresentaram erro, ou seja, foram enquadrados nas classificações 5XX, 4XX ou “outros erros sem acesso” (Tabela 12).

Tabela 12 – TLD's dos *links* com falha de acesso distribuídos por ano

	2012	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
gov	11	34	27	29	17	25	30	36	16	34
org	10	53	22	30	48	49	65	60	125	46
com	7	54	13	30	47	57	99	57	40	36
net	1	1			5	4	8	6	3	1
doi	0	0	0	4	1	0	0	0	5	3
handle	0	0	1	0	0	0	0	0	0	0

Fonte: Elaborado pelo autor.

A nossa última categoria de análise é o tempo de meia-vida nos conjuntos de *links* testados. Após inserir e ajustar as fórmulas na planilha de dados do Microsoft Excel, obtivemos os seguintes resultados (Tabela 13).

Tabela 13 – Tempo de meia-vida

Ano	**meia-vida (anos)**
2012	9,35
2013	8,79
2014	7,79
2015	9,74
2016	13,50
2017	8,27
2018	5,25
2019	6,03

2020		3,89
2021		2,92
	Mediana	8,03

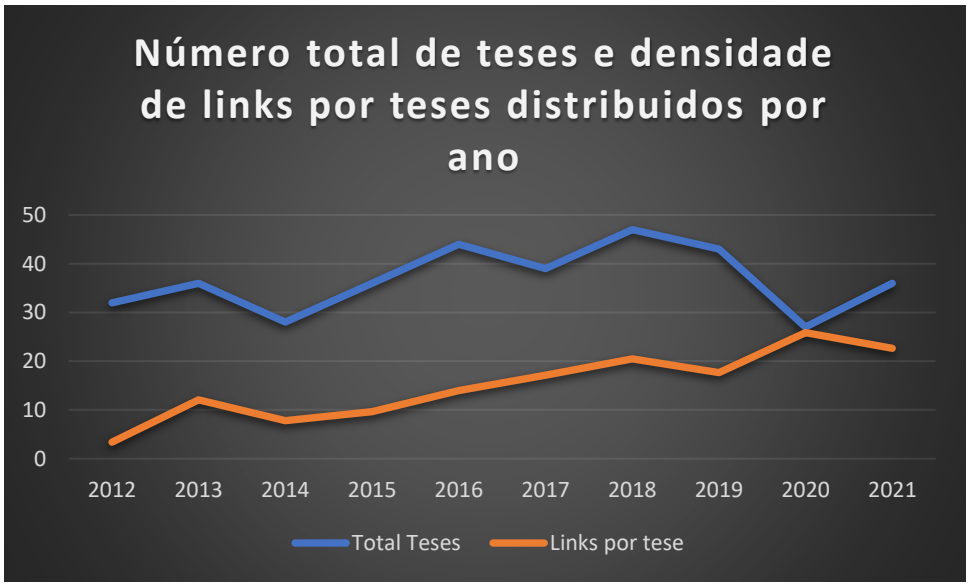
Fonte: Elaborado pelo autor.

Concluímos, assim, a apresentação dos dados referentes às análises e aos testes realizados durante a pesquisa. Salientamos que estas informações são de um pequeno grupo e podem apresentar divergências quando extrapoladas para um conjunto maior.

5 DISCUSSÃO

Conforme a distribuição dos *links* testados por ano, notamos o aumento no uso de *links* da *web* como referência, essa é uma consequência da digitalização da sociedade e, principalmente, do modo como a produção acadêmica tem aderido ao meio eletrônico para publicação (FERREIRA; MARTINS; ROCKEMBACH, 2018). Observamos a queda no número de publicações em 2020, como relatado, provavelmente um reflexo decorrente da pandemia de SARS-CoV-2, porém, houve um aumento na densidade de *links* por tese nesse mesmo ano. Esse resultado evidencia uma maior utilização da *web* como fonte de pesquisa (Figura 18).

Figura 18 – Número total de teses e densidade de *links* por teses distribuídas por ano

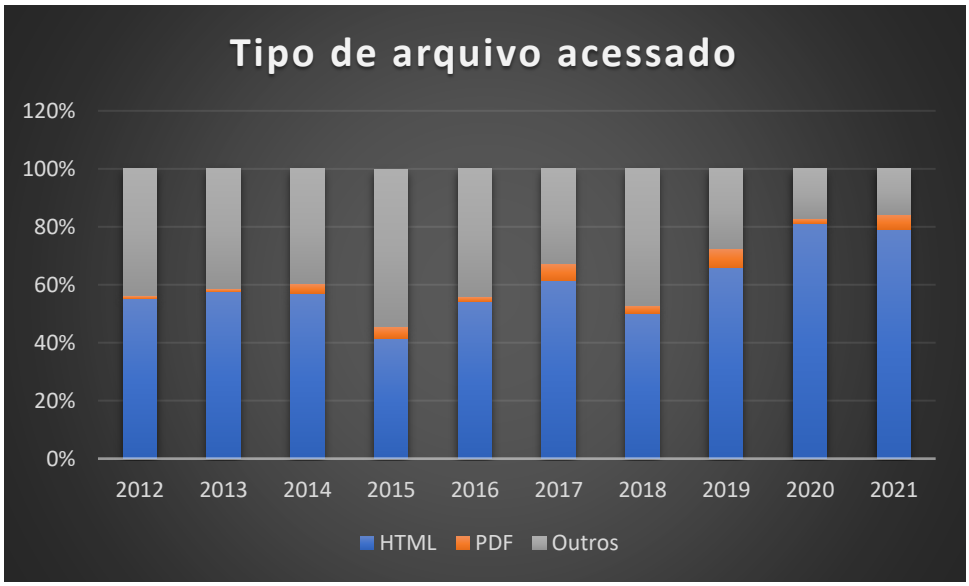


Fonte: Elaborado pelo autor.

A investigação dos códigos de resposta HTTP foi pensada de modo a guiar as outras etapas, pois o resultado proveniente dessa investigação permitirá verificar a persistência e a meia-vida dos *links* da *web* utilizados como referências nas teses de doutorado armazenadas no Lume. Encontramos predominantemente páginas em formato HTML na amostra dos *links* da *web* (Figura 19), documentos PDF representavam uma minoria, enquanto outras formas de arquivo, como os

desenvolvidos em Java, Javascript, Python, PHP, C#, SQL e outras linguagens, somam o restante deste corpus.

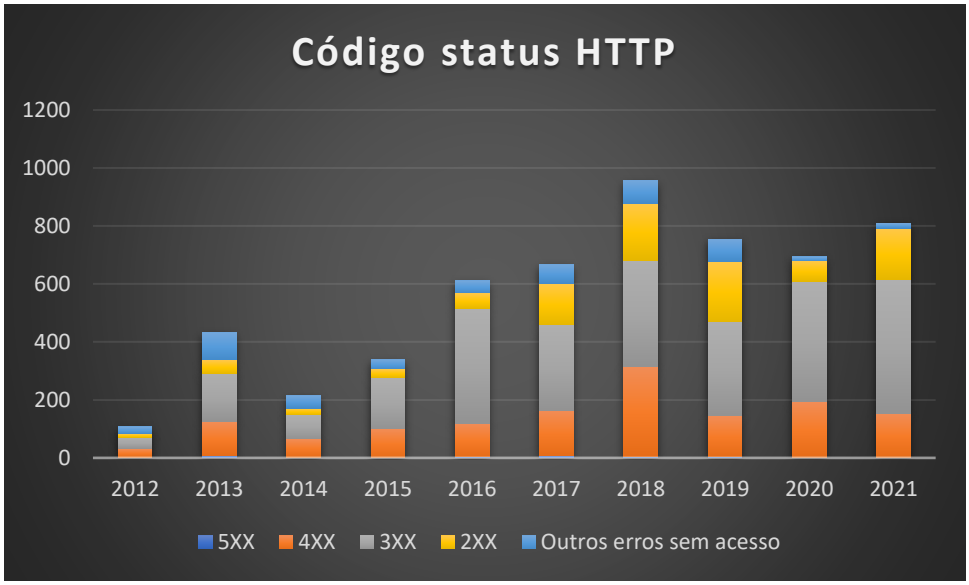
Figura 19 – Gráfico do tipo de arquivo acessado



Fonte: Elaborado pelo autor.

Os códigos de status HTTP quando comparada a amostra total (Figura 20), depois da verificação dos *links* no software *Xenu Link*, indicam muitos redirecionamentos (3XX), cerca de 49% da amostra, contra apenas 17% de códigos 2XX. Tradicionalmente, a recuperação malsucedida de conteúdo é chamada de "códigos de erro". Por exemplo, erro de "404: Não Encontrado" significa que o servidor não encontrou nada que corresponda ao URI de Solicitação" (SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012, p. 2). Os erros de nossa amostra somam juntos 34% do total testado, sendo que 25% são erros de classe 4XX, na sua maioria 404, situação semelhante ao demonstrado por Koehler (2004), Dimitrova e Bugeja (2007) e Sadat-Moosavi, Isfandyari-Moghaddam e Tajeddini (2012). Nestes estudos, a maioria dos erros foram decorrentes de problemas relacionados ao cliente, tais como filtragem, falhas de conexão e mau funcionamento do navegador devido aos proxies utilizados (DIMITROVA; BUGEJA, 2007; KOEHLER, 2004; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

Figura 20 – Gráfico do código status HTTP



Fonte: Elaborado pelo autor.

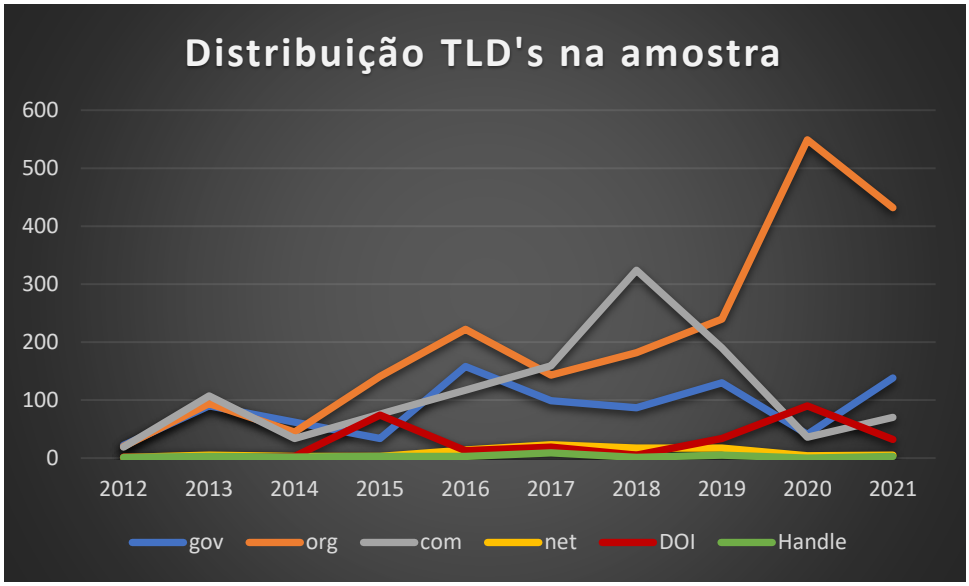
O estudo dos TLDs permitiu uma comparação da afirmação realizada por Koehler (2004), e corroborada por Dimitrova e Bugeja (2007) (Figura 21). Existe uma variação na persistência dos *links* da *web* para os diferentes TLDs. Essa característica é apontada como intrínseca à sua criação, pois o TLD “.com” muitas vezes é utilizado para navegação, enquanto páginas com TLD “.edu” têm uma característica voltada para publicação (DIMITROVA; BUGEJA, 2007; KOEHLER, 2002).

Ao comparar os nossos resultados, encontramos uma menor persistência nos *links* da *web* com TLDs “.org” e “.com”, assim como uma maior persistência nos TLDs “.gov” e “.net”. Essa característica difere de outros estudos onde o TLD “.org” figura entre os mais estáveis; porém, como mencionado no estudo realizado por Sadat-Moosavi, Isfandyari-Moghaddam e Tajeddini (2012), na comparação entre diferentes amostras pode haver variação da persistência de *links* com as mesmas características (DIMITROVA; BUGEJA, 2007; KOEHLER, 2004; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012).

O uso de identificadores permanentes, como DOI, oscila no conjunto amostral, observamos uma grande persistência para esse tipo de *link* da *web* na amostra. Podemos verificar que o identificador Handle tem um emprego menor que

o DOI, mas uma maior persistência, o Handle apresenta 3% de erro nos *links* persistentes da *web*, enquanto constatamos 5% de erro nos *links* da *web* atribuídos ao identificador persistente DOI.

Figura 21 – Gráfico da distribuição TLD's na amostra



Fonte: Elaborado pelo autor.

Na distribuição do resultado da pesquisa, notamos um decaimento da persistência conforme o envelhecimento (Figura 22), efeito condizente com o que foi concluído por Koehler (2004) e Dimitrova e Bugeja (2007), eles constataram que em conjuntos de *links* da *web* existe uma tendência à estabilização da amostra após um decaimento inicial mais acentuado. De acordo com Koehler (2011), essas estabilizações da amostra de *links* após um determinado decaimento, "[...] sugerem que, à medida que uma coleção de URL estática envelhece, ela pode se tornar relativamente estável com o tempo" (KOEHLER, 2004, p. 6).

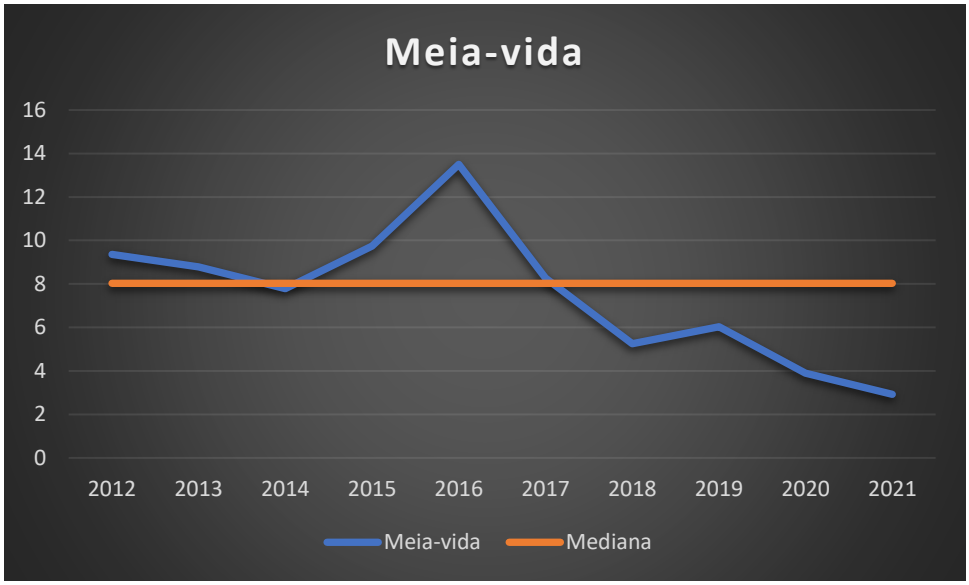
Figura 22 – Gráfico da persistência dos *links* (números absolutos)

Fonte: Elaborado pelo autor.

A persistência encontrada no conjunto (Figura 22) é de cerca de 50% para as amostras mais antigas (2012 e 2013), diminuindo gradativamente até os 21% encontrados no ano de 2021.

Os cálculos de meia-vida (Figura 23) provaram ser condizentes com as pesquisas realizadas por Dimitrova e Bugeja (2007), Klein, Van de Sompel, Sanderson, Shankar e Balakireva (2014), Koehler (2004), Oguz e Koehler (2011), que nos mostram que existe uma meia-vida de aproximadamente 10 anos. Em nossa amostra, a mediana das meias-vidas no período 2012-2021 foi de 8,02 anos, isso significa que existe uma expectativa de que apenas metade de uma amostra se mantenha persistente após 8 anos (DIMITROVA; BUGEJA, 2007; KLEIN *et al.*, 2014; KOEHLER, 2004; OGUZ; KOEHLER, 2011).

Figura 23 – Gráfico da meia-vida



Fonte: Elaborado pelo autor.

Os resultados sobre a persistência dos *links* da *web* obtidos nos estudos de Koehler e Oguz (2011), Klein, Van de Sompel, Sanderson, Shankar e Balakireva (2014) e Massicote e Botter (2017), diferem no tempo de persistência de um *link* da *web*, porém, todos indicam que a *web* está desaparecendo (Tabela 14). A mediana da meia-vida de nossa amostra foi calculada em 8,02 anos, apresentamos a tabela a seguir com alguns tempos de meia-vida em outros estudos revisados neste trabalho.

Tabela 14 – Tempo de meia-vida para diferentes amostras

Autor	**Amostra**	**Meia-vida**	**Ano**
Harter e Kim	Citações em artigos acadêmicos	1,5	1996
Rumsey	Citações legais	1,4	2002
Koehler	Páginas aleatórias	2	2002
Markwell e Brooks	Pesquisas em biologia, ciência e educação	4,6	2002
Nelson e Allen	Objetos digitais de bibliotecas	24,5	2002
Spinellis	Citações em ciência da computação	4	2003
Dimitrova e Bugeja	Periódicos de comunicação	3,17	2007

Goh e Ng	Periódicos de ciência da informação	5	2007
Moghaddam, Saberi e Esmaeel	Revista Information Research	14	2010
Kumar e Kumar	Periódicos de acesso aberto	11,5	2012
Jalalifard, Norouzi e Isfandyari-Moghaddam	Revistas médicas	1,02	2013
Sife e Bernanrd	Teses e dissertações	2,5	2013
Ferreira	Teses do Lume	8,02	2023

Fonte: Elaborado pelo autor.

A partir dos resultados apontados podemos verificar que os *links* da *web* apresentam uma tendência ao desaparecimento. Apesar de todo o esforço no uso de identificadores permanentes, muitos *links* utilizados nos trabalhos acadêmicos continuam a apontar para a *web* de maneira direta, ou seja, utilizam o *link* de acesso sem a garantia de persistência de que precisam para servir como referencial do trabalho ao qual foram vinculados (KLEIN *et al.*, 2014).

6 CONSIDERAÇÕES FINAIS

O trabalho consistiu na revisão de abordagens e diretrizes para testagem dos *links* da *web*, embora não seja nossa pretensão elaborar uma ferramenta para verificação da persistência de *links* da *web*, utilizamos em nossa pesquisa um modelo que pode ser reproduzido para testagem de outras amostras.

Este estudo constata que muitos recursos da *web* mencionados como referência bibliográfica em teses de doutorado disponíveis no Repositório Digital da Universidade Federal do Rio Grande do Sul – Lume desapareceram de seus locais originais. O código de *status* HTTP 4XX foi o erro mais comum encontrado, e o domínio de nível superior ".com" teve o maior número de *links* da *web* com falha. Os resultados mostram que metade das citações da *web* dessas teses desapareceram em 2012 e 2013. Isso sugere que a disponibilidade a longo prazo de recursos de informação online não pode ser garantida, o que levanta questões sobre como realizar a inclusão de *links* da *web* como parte dos artefatos bibliográficos nas listas de referências.

Conforme descrito na apresentação dos resultados, as características que fazem um *link* persistir não são objetivas, as amostras se comportam de maneira única, mantendo-se a confirmação do desaparecimento da *web* como o ponto em comum entre todos os estudos revisados. Pudemos, em nosso estudo, observar os tipos de arquivos apontados pelos *links* da *web*, assim como os TLDs mais encontrados, esses dados podem ajudar a prever o tempo de persistência dos *links* do nosso corpus.

O nosso primeiro objetivo foi concluído com a ajuda dos filtros do Lume, tivemos algumas dificuldades com a utilização dos filtros simples, mas conseguimos mapear as teses necessárias para a construção de nosso corpus de pesquisa. O Lume apresentou-se indisponível para exibição do arquivo de uma das teses selecionadas para amostragem, para o qual o servidor retornou uma mensagem de erro.

Identificamos que são comuns os programas de depósito legal, porém, muitas vezes, estes programas têm foco na infraestrutura de interoperabilidade,

nas suas políticas, nas conformidades de padrões ou então na adequação de metadados, mas não acompanham a qualidade do conteúdo nos *links* citados ou arquivados. Entendemos isso como um problema a ser encarado, no sentido que o controle do conteúdo dos repositórios e bases de dados é tão importante quanto a sua coleta.

É necessário um esforço conjunto de autores, editores, bibliotecários, gerentes de *web* e profissionais de TIC para enfrentar o problema do desaparecimento de URLs. Os autores devem ter cuidado ao digitar as URLs e os editores precisam se tornar mais proativos em suas funções de controle de qualidade, existe a necessidade de treinamento e orientação quanto as políticas e procedimentos que propiciem um ambiente de preservação para as publicações.

Para identificação das fontes de informação da *web*, nas teses da amostra, conseguimos realizar a leitura e consequente formatação dos *links* da *web* extraídos por meio de uma sequência de operações técnicas. Após a realização da primeira testagem, pudemos concluir os tipos de arquivos da *web* mais comuns, sendo predominante o uso da linguagem HTML nos *links* da *web* estudados.

A persistência e o tempo de meia-vida resultaram da dupla testagem dos *links* da *web*, seguindo os procedimentos descritos em estudos anteriores. As estimativas de persistência e a meia-vida dos *links* da *web* variam para diferentes anos, portanto, utilizamos um intervalo de amostra de 10 anos, que torna-se mais confiável para ser utilizado por pesquisadores no futuro, como fonte de referência no estudo da *web*. A meia-vida estimada no conjunto de 8,02 anos salienta a necessidade de intervenções com o intuito de preservação das informações publicadas na *web*.

Os fenômenos de *link rot* e *reference rot* têm consequências negativas em várias áreas, incluindo a pesquisa acadêmica, as referências bibliográficas, o jornalismo online e o acesso as informações históricas. Quando um link se torna inválido, inacessível ou tem seu conteúdo modificado, pode ser difícil para os leitores recuperarem o conteúdo originalmente publicado, resultando em lacunas no conhecimento e na falta de acesso a importantes recursos.

As mudanças tecnológicas, como a instalação de sistemas de gerenciamento de conteúdo e troca de servidores, podem causar os fenômenos como *link rot* e *reference rot*. É importante que os administradores e editores de sites se concentrem na manutenção da disponibilidade de recursos e oportunidades de *links* da *web*, como realizando a utilização de serviços de arquivamento da *web*, que capturam e preservam cópias de páginas da *web* ao longo do tempo, assim como que os atores da *web* pensem na expansão da sua preservação por meio de armazenamento e utilização de *links* persistentes para as referências bibliográficas utilizadas nas pesquisa.

REFERÊNCIAS

ABBATE, J. **Inventing the internet**. [*S. l.*]: MIT press, 2000.

ANTRACOLI, A. *et al.* Capture All the URLs: First Steps in Web Archiving. **Pennsylvania Libraries: Research & Practice**, [*s. l.*], v. 2, n. 2, p. 155–170, 2014.

BARRETO, A. de A. Uma história da ciência da informação. **Para entender a ciência da informação. Salvador: EDUFBA**, [*s. l.*], p. 13–34, 2007.

BERNERS-LEE, T. Cool URIs don't change. **http://www. w3. org/Provider/Style/URI**, [*s. l.*], 1998a.

BERNERS-LEE, T. **The original proposal of the WWW, HTMLized**. [*S. l.*], 1998b. Disponível em: https://www.w3.org/History/1989/proposal.html. Acesso em: 21 set. 2022.

BERNERS-LEE, T.; MASINTER, L.; MCCAHILL, M. **Uniform Resource Locators (URL) - rfc1738**. , 1994. Disponível em: https://www.ietf.org/rfc/rfc1738.txt. Acesso em: 29 set. 2022.

BRÜGGER, N. Website history and the website as an object of study. **New Media & Society**, [*s. l.*], v. 11, n. 1–2, p. 115–132, 2009.

CANDAU, J. **Memoria e Identidade**. 1. ed. [*S. l.: s. n.*], 2016.

CASTELLS, M. A Era da Informação: Economia, Sociedade e Cultura Vol. 1-A Sociedade em Rede. São Paulo, Ed. Paz e Terra, 1999. **Sociologias**, [*s. l.*], v. 1, n. 2, 1999.

CNJ, C. N. de J. Estatística Aplicada a Dados - Atos normativos estruturantes. **Estatística Aplicada a Dados - Atos normativos estruturantes**, [*s. l.*], p. 40, 2021.

COSTA, M.; GOMES, D.; SILVA, M. J. The evolution of web archiving. **International Journal on Digital Libraries**, [*s. l.*], v. 18, p. 191–205, 2017.

DIMITROVA, D. V.; BUGEJA, M. The half-life of internet references cited in communication journals. **New Media & Society**, [*s. l.*], v. 9, n. 5, p. 811–826, 2007.

DODEBEI, V. L. D. L. de M. Contribuições das teorias da memória para o estudo do patrimônio na web. [*s. l.*], 2012. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/494. Acesso em: 24 ago. 2021.

DOI, F. **The DOI® handbook**. [*S. l.*], 2019. Disponível em: https://www.doi.org/the-identifier/resources/handbook/. .

FERREIRA, L. B.; MARTINS, M. R.; ROCKEMBACH, M. Usos do Arquivamento da Web na Comunicação Científica. **PRISMA.COM**, [*s. l.*], v. 0, n. 36, p. 78–98, 2018.

FETTERLY, D. *et al.* A large-scale study of the evolution of Web pages. **Software: Practice and Experience**, [*s. l.*], v. 34, n. 2, p. 213–237, 2004.

FREITAS, M. H. Considerações acerca dos primeiros periódicos científicos brasileiros. **Ciência da Informação**, [*s. l.*], v. 35, n. 3, p. 54–66, 2006.

GOMES, D. *et al.* (org.). **The Past Web: Exploring Web Archives**. Cham: Springer International Publishing, 2021. *E-book*. Disponível em: https://link.springer.com/10.1007/978-3-030-63291-5. Acesso em: 13 set. 2022.

HOLUB, K.; RUDOMINO, I. Croatian Web Archive: an overview. **Review of the National Center for Digitization**, [*s. l.*], n. 25, 2014.

INTERNET ARCHIVE. **Internet Archive: About IA**. [*S. l.*], 2023. Disponível em: https://archive.org/about/. Acesso em: 23 abr. 2023.

ISFANDYARI MOGHADDAM, A.; SABERI, M. K.; MOHAMMAD ESMAEEL, S. Availability and Half-life of Web References Cited in Information Research Journal: A Citation Study. **International Journal of Information Science and Management (IJISM)**, [*s. l.*], v. 8, n. 2, p. 57–75, 2010.

KLEIN, M. *et al.* Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. **PLoS ONE**, [*s. l.*], v. 9, n. 12, p. e115253, 2014.

KOEHLER, W. A longitudinal study of Web pages continued: a consideration of document persistence. **Information Research**, [*s. l.*], v. 9, n. 2, p. 9–2, 2004.

KOEHLER, W. **Digital libraries and World Wide Web sites and page persistence.** [*S. l.*], 1999. Disponível em: https://informationr.net/ir/4-4/paper60.html. Acesso em: 13 abr. 2023.

KOEHLER, W. Web page change and persistence—A four-year longitudinal study. **Journal of the American Society for Information Science and Technology**, [*s. l.*], v. 53, n. 2, p. 162–171, 2002.

KRÓL, K.; ZDONEK, D. Peculiarity of the bit rot and link rot phenomena. **Global Knowledge, Memory and Communication**, [*s. l.*], 2019.

LUZ, A. J. A. da. Comunicação pública e memória comunicacional: revelações e apagamentos sobre o governo da presidenta Dilma Rousseff. [*s. l.*], 2021.

LUZ, A. J. Preservação de sites oficiais: exemplos internacionais e o desafio brasileiro. **Revista Brasileira de Preservação Digital**, [*s. l.*], v. 3, p. e022010–e022010, 2022.

MARQUES, M. B.; GOMES, L. E. **Ciência da Informação: visões e tendências**. [*S. l.*]: Imprensa da Universidade de Coimbra / Coimbra University Press, 2020.

MARY P. BENBOW, S. File not found: the problems of changing URLs for the World Wide Web. **Internet Research**, [*s. l.*], v. 8, n. 3, p. 247–250, 1998.

MASANÈS, J. (org.). **Web archiving**. Berlin ; New York: Springer, 2006.

MASSICOTTE, M.; BOTTER, K. Reference Rot in the Repository: A Case Study of Electronic Theses and Dissertations (ETDs) in an Academic Library. **Information Technology and Libraries**, [*s. l.*], v. 36, n. 1, p. 11, 2017.

MELO, J. F.; NUNES, L. A. N. de O.; ROCKEMBACH, M. Preservação de websites governamentais a partir do arquivamento da web: abordagens e metodologias. **Encontro Nacional de Pesquisa em Ciência da Informação (20.: 2019 out. 21-25: Florianópolis, SC). Anais [recurso eletrônico]. Florianópolis: ANCIB 2019.**, [*s. l.*], 2019.

MELO, J. F.; ROCKEMBACH, M. Arquivabilidade de websites para preservação digital: estudo a partir da área da saúde. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, [*s. l.*], v. 14, n. 3, 2020.

MELO, J. F.; ROCKEMBACH, M. International Initiatives and Advances in Brazil for Government Web Archiving. *Em*: , 2021. **Data and Information in Online Environments: Second EAI International Conference, DIONE 2021, Virtual Event, March 10–12, 2021, Proceedings**. [*S. l.*]: Springer, 2021. p. 83–95.

MILLIGAN, I. **History in the Age of Abundance: How the Web is Transforming Historical Research**. [*S. l.*], 2019. Disponível em: https://www.ianmilligan.ca/publication/history-in-the-age-of-abundance/. Acesso em: 21 abr. 2023.

MUSIANI, F. *et al.* **Qu'est-ce qu'une archive du Web?** [*S. l.*]: OpenEdition Press, 2019.

NELSON, M. L.; ALLEN, B. D. Object persistence and availability in digital libraries. **D-Lib magazine**, [*s. l.*], v. 8, n. 1, 2002.

NIC.BR. **CGI.br - Comitê Gestor da Internet no Brasil**. [*S. l.*], [*s. d.*]. Disponível em: https://cgi.br. Acesso em: 22 set. 2022.

NTOULAS, A.; CHO, J.; OLSTON, C. What's new on the Web? The evolution of the Web from a search engine perspective. *Em*: , 2004. **Proceedings of the 13th international conference on World Wide Web**. [*S. l.: s. n.*], 2004. p. 1–12.

OGUZ, F.; KOEHLER, W. C. Document constancy and persistence: A study of Web pages in library and information science domain. **Proceedings of the American Society for Information Science and Technology**, [*s. l.*], v. 48, n. 1, p. 1–9, 2011.

OLIVEIRA, L.; ROCKEMBACH, M. Análise de conteúdo de termos de uso e políticas de privacidade de arquivos da web. **Fórum de Estudos em Informação, Sociedade e Ciência**, [*s. l.*], v. 4, p. 188–195, 2021.

REHM, G. Hypertext Types and Markup Languages. *Em*: WITT, A.; METZING, D. (org.). **Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology**. Dordrecht: Springer Netherlands, 2010. (Text, Speech and Language Technology). p. 143–164. *E-book*. Disponível em: https://doi.org/10.1007/978-90-481-3331-4_8. Acesso em: 29 set. 2022.

ROCKEMBACH, M. Arquivamento da Web: estudos de caso internacionais e o caso brasileiro. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, [*s. l.*], v. 16, n. 1, p. 7–24, 2018.

ROCKEMBACH, M. Arquivamento da Web no contexto das Humanidades Digitais: da produção a preservação da informação digital | Web archiving in the context of digital humanities: from production to preservation of digital information. **Liinc em Revista**, [*s. l.*], v. 15, n. 1, 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4578. Acesso em: 20 ago. 2021.

ROCKEMBACH, M. Inequalities in digital memory: ethical and geographical aspects of web archiving. **The International Review of Information Ethics**, [*s. l.*], v. 26, 2017.

ROCKEMBACH, M.; PAVÃO, C. M. G. Políticas E Tecnologias De Preservação Digital No Arquivamento Da Web. **Revista Ibero-Americana de Ciência da Informação**, [*s. l.*], v. 11, n. 1, p. 168–182, 2018.

ROCKEMBACH, M.; SERRANO, A. Climate change and web archives: an Ibero-American study based on the Portuguese and Brazilian contexts. **Records Management Journal**, [*s. l.*], 2021.

RODRIGUES, V. L. D.; ROCKEMBACH, M. **Arquivos da web como fonte historiográfica**. SciELO Brasil, , 2023.

SADAT-MOOSAVI, A.; ISFANDYARI-MOGHADDAM, A.; TAJEDDINI, O. Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI-ranked journals. **Aslib Proceedings**, [*s. l.*], v. 64, n. 2, p. 178–192, 2012.

TERRADA, G. A. F. Preservação digital da web: uma reflexão sobre políticas e práticas. [*s. l.*], 2022.

TSAY, M. **Citation analysis of Ted Nelson's works and his influence on hypertext concept**. [*S. l.*], 2009. Disponível em: https://link.springer.com/article/10.1007/s11192-008-1641-7. Acesso em: 29 set. 2022.

TYLER, D. C.; MCNEIL, B. Librarians and link rot: a comparative analysis with some methodological considerations. **portal: Libraries and the Academy**, [*s. l.*], v. 3, n. 4, p. 615–632, 2003.

UFRGS. **Bibliotecas da UFRGS. LUME - o que e como pesquisar no repositório digital de acesso aberto da UFRGS.** Porto Alegre: [*s. n.*], 2021. Disponível em: https://www.ufrgs.br/super8/wp-content/uploads/apresentacoes/lume.pdf.

VIEIRA, E. **Os bastidores da Internet no Brasil**. [*S. l.*]: Editora Manole Ltda, 2003.

W3C, W. W. W. C. **World Wide Web Consortium**. [*S. l.*], 2023. Disponível em: https://www.w3c.br. Acesso em: 21 set. 2022.

WOLTON, D. **Internet, e depois?: uma teoria crítica das novas mídias**. [*S. l.*]: Editora Sulina, 2003.

ZITTRAIN, J.; ALBERT, K.; LESSIG, L. Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations. **Legal Information Management**, [*s. l.*], v. 14, n. 2, p. 88–99, 2014.