Moisés Rockembach
Caterina Groposo Pavão

ARQUIVAMENTO DA WEB E PRESERVAÇÃO DIGITAL

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES
PPGCIN
pimenta cultural

Moisés Rockembach
Caterina Groposo Pavão

ARQUIVAMENTO DA WEB E PRESERVAÇÃO DIGITAL

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES
PPGCIN
I São Paulo I 2024 I
pimenta cultural

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)

R682a

Rockembach, Moisés.
Arquivamento da web e preservação digital / Moisés Rockembach e Caterina Groposo Pavão. – São Paulo: Pimenta Cultural, 2024.

Livro em PDF

ISBN 978-65-5939-811-9
DOI 10.31560/pimentacultural/2024.98119

1. Ciências da informação. 2. Preservação digital. 3. Arquivamento da web. 4. Arquivos digitais. 5. Bibliotecas digitais. I. Rockembach, Moisés. II. Pavão, Caterina Groposo. III. Título.

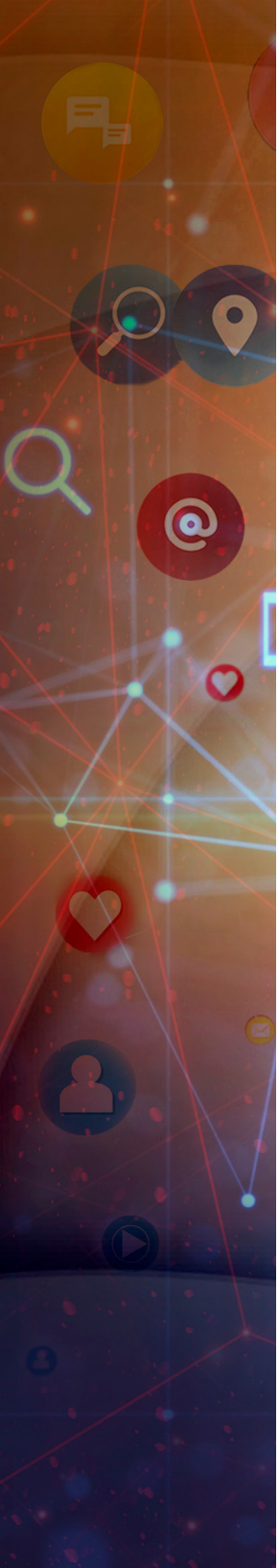
CDD 020

Índice para catálogo sistemático:
I. Biblioteconomia e ciências da informação.
Jéssica Oliveira • Bibliotecária • CRB-034/2023

Direção editorial	Patricia Bieging Raul Inácio Busarello
Editora executiva	Patricia Bieging
Coordenadora editorial	Landressa Rita Schiefelbein
Assistente editorial	Bianca Bieging
Diretor de criação	Raul Inácio Busarello
Assistente de arte	Naiara Von Groll
Editoração eletrônica	Andressa Karina Voltolini Potira Manoela de Moraes
Bibliotecária	Jéssica Castro Alves de Oliveira
Imagens da capa	Bizkette1, Your_Photo, Harryarts, Sketchepedia, Sitthiphong - Freepik.com
Tipografias	Acumin, Bebas Neue Pro, Rockwell
Revisão	Os autores
Autores	Moisés Rockembach Caterina Groposo Pavão

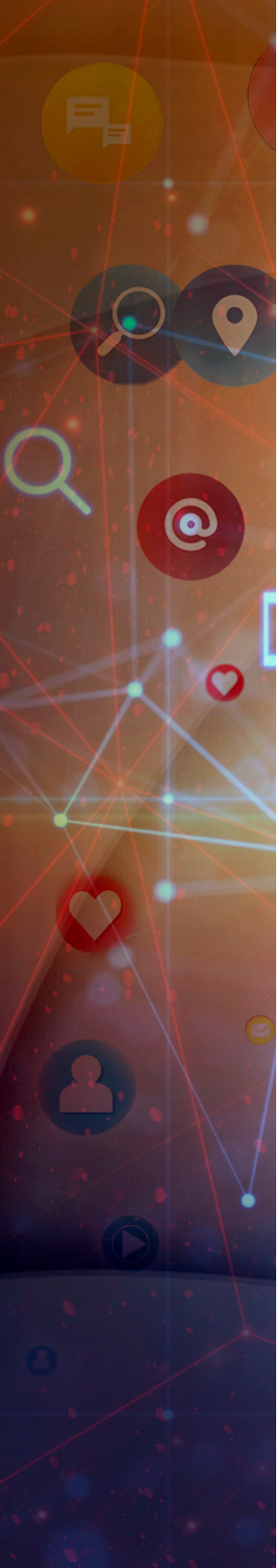
PIMENTA CULTURAL
São Paulo • SP
+55 (11) 96766 2200
livro@pimentacultural.com
www.pimentacultural.com

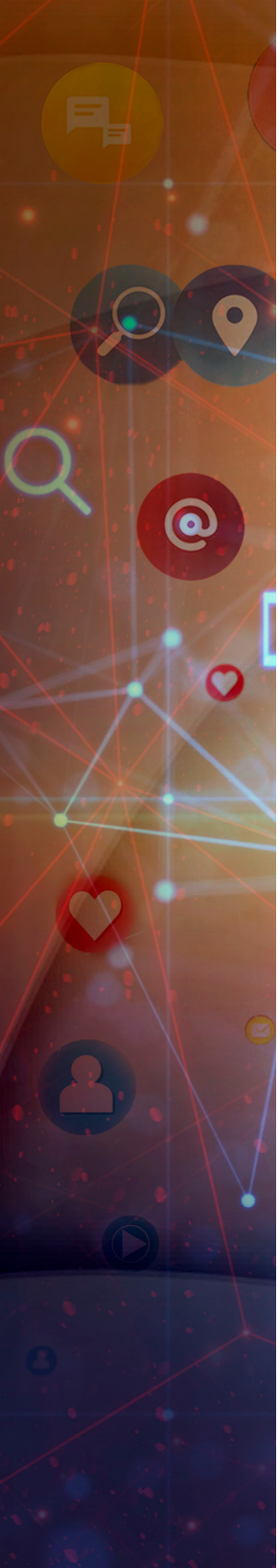
2024

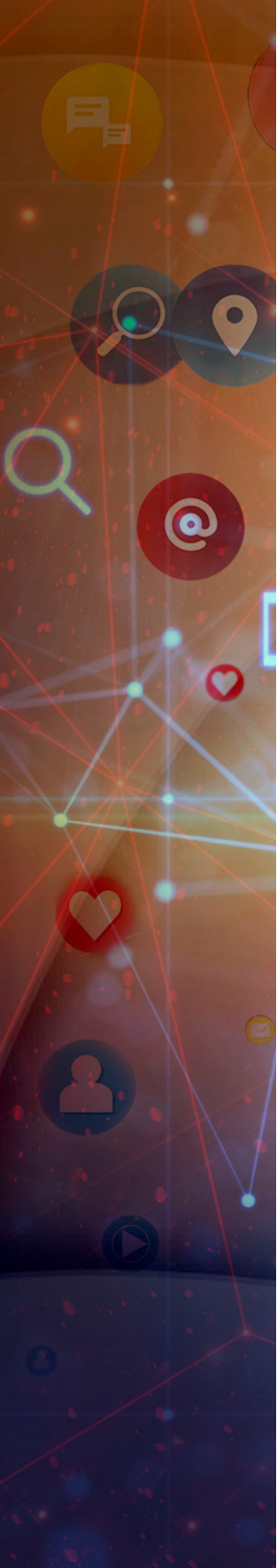


CONSELHO EDITORIAL CIENTÍFICO

Doutores e Doutoras







PARECERISTAS
E REVISORES(AS) POR PARES

Avaliadores e avaliadoras Ad-Hoc

Alessandra Figueiró Thornton
Universidade Luterana do Brasil, Brasil

Alexandre João Appio
Universidade do Vale do Rio dos Sinos, Brasil

Bianka de Abreu Severo
Universidade Federal de Santa Maria, Brasil

Carlos Eduardo Damian Leite
Universidade de São Paulo, Brasil

Catarina Prestes de Carvalho
Instituto Federal Sul-Rio-Grandense, Brasil

Elisiene Borges Leal
Universidade Federal do Piauí, Brasil

Elizabete de Paula Pacheco
Universidade Federal de Uberlândia, Brasil

Elton Simomukay
Universidade Estadual de Ponta Grossa, Brasil

Francisco Geová Goveia Silva Júnior
Universidade Potiguar, Brasil

Indiamaris Pereira
Universidade do Vale do Itajaí, Brasil

Jacqueline de Castro Rimá
Universidade Federal da Paraíba, Brasil

Lucimar Romeu Fernandes
Instituto Politécnico de Bragança, Brasil

Marcos de Souza Machado
Universidade Federal da Bahia, Brasil

Michele de Oliveira Sampaio
Universidade Federal do Espírito Santo, Brasil

Pedro Augusto Paula do Carmo
Universidade Paulista, Brasil

Samara Castro da Silva
Universidade de Caxias do Sul, Brasil

Thais Karina Souza do Nascimento
Instituto de Ciências das Artes, Brasil

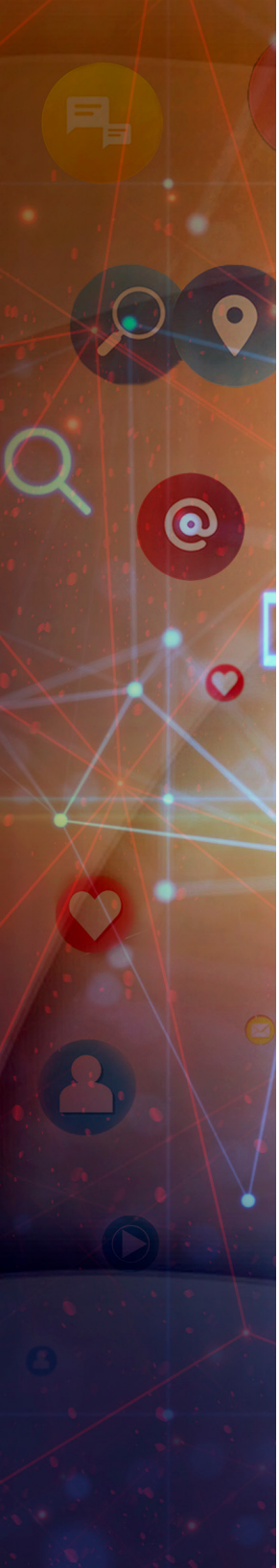
Viviane Gil da Silva Oliveira
Universidade Federal do Amazonas, Brasil

Weyber Rodrigues de Souza
Pontifícia Universidade Católica de Goiás, Brasil

William Roslindo Paranhos
Universidade Federal de Santa Catarina, Brasil

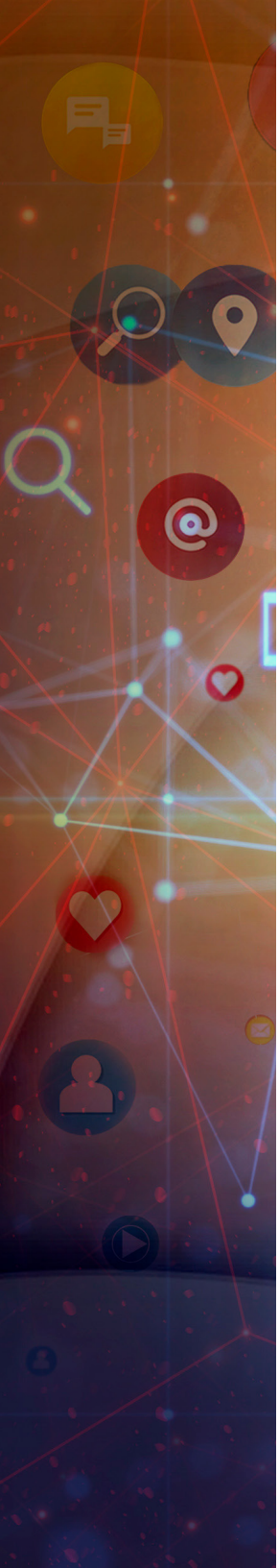
Parecer e revisão por pares

Os textos que compõem esta obra foram submetidos para avaliação do Conselho Editorial da Pimenta Cultural, bem como revisados por pares, sendo indicados para a publicação.

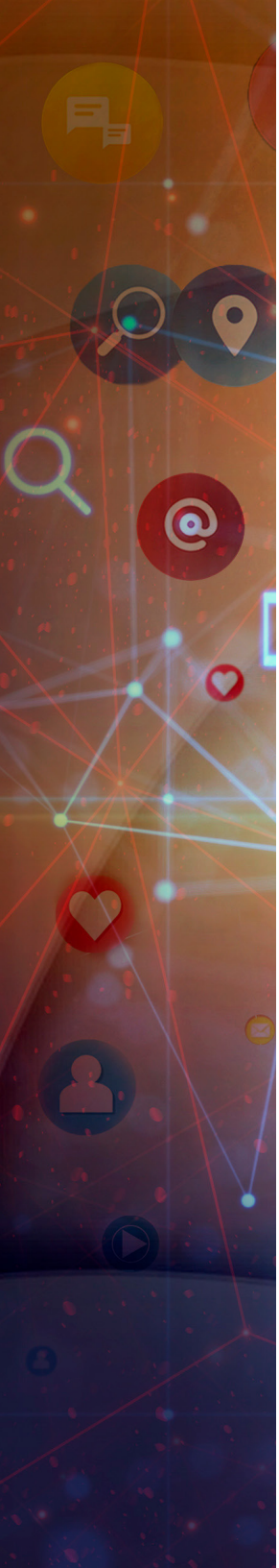


LISTA DE SIGLAS

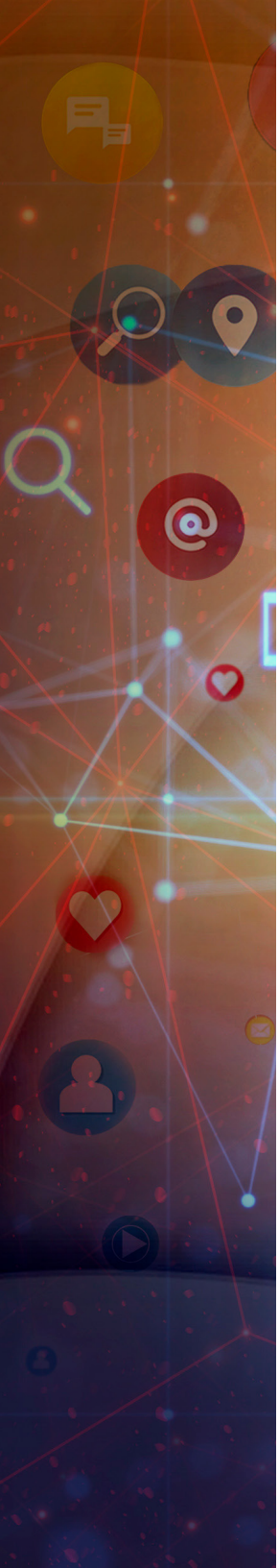
API	*Application Programming Interface*
ARC	*ARchive Container*
ARK	*Archival Resource Key*
BDM	Biblioteca Digital Mundial
CDD	Classificação Decimal Universal
CDL	*California Digital Library*
CERN	*Conseil Européen pour la Recherche Nucléaire*
CGI.br	Comitê Gestor da Internet no Brasil
CLEAR	*Credible Live Evaluation of Archive Readiness*
CONARQ	Conselho Nacional de Arquivos
CSS	Cascading Style Sheets
DACS	*Describing Archives: A Content Standard*
DC	*Dublin Core*
DCMI	*Dublin Core™ Metadata Initiative*
DOI	*Digital Object Identification*
EOT	*End of Term*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FBN	Fundação Biblioteca Nacional
FID	Federação Internacional de Informação e Documentação
GDPR	*General Data Protection Regulation*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*



IANA	*Internet Assigned Numbers Authority*
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
ICANN	*Internet Corporation for Assigned Names and Numbers*
IIPC	*International Internet Preservation Consortium*
IKS	*Indigenous Knowledge Systems*
IoT	*Internet of Things*
ISO	*International Organization for Standardization*
LAI	Lei de Acesso à Informação
LGPD	Lei Geral de Proteção de Dados Pessoais
MARC	*Machine-Readable Cataloging*
MODS	*Metadata Object Description Schema*
NDSA	*National Digital Stewardship Alliance*
NIC.br	Núcleo de Informação e Coordenação do Ponto BR
NLI	*National Library of Ireland*
NUAWEB	Núcleo de Pesquisa em Arquivamento da *web* e Preservação Digital
NYARC	*New York Art Resources Consortium*
OAI-PMH	*Open Archives Initiative Protocol for Metadata Harvesting*
OAIS	*Open Archival Information System*
OCLC	*On-line Computer Library Center*
PAI	Pacote de Arquivamento de Informação
PPAD	Política de Preservação de Acervos Digitais
PPDBN	Política de Preservação Digital da Biblioteca Nacional
PROMISE	*Preserving On-line Multiple Information: towards a Belgian strategy*
PURL	*Persistent Uniform Resource Locators*
PWA	Arquivo da *web* Portuguesa



RDA	*Resource Description and Access*
RNP	Rede Nacional de Ensino e Pesquisa
SAAI	Sistema Aberto de Arquivamento de Informação
TCP/IP	*Transmission Control Protocol/Internet Protocol*
TOI	Tecnologia e Organização da Informação
UFRGS	Universidade Federal do Rio Grande do Sul
UKWA	*UK web Archive*
URI	*Uniform Resource Identifier*
URN	*Uniform Resource Name*
WACZ	*Web ARCHive Collection Zipped*
WAM	W*eb Archiving Metadata Working Group*
WARC	*Web ARChive*
WCT	*Web Curator Tool*
WWW	*World Wide Web*



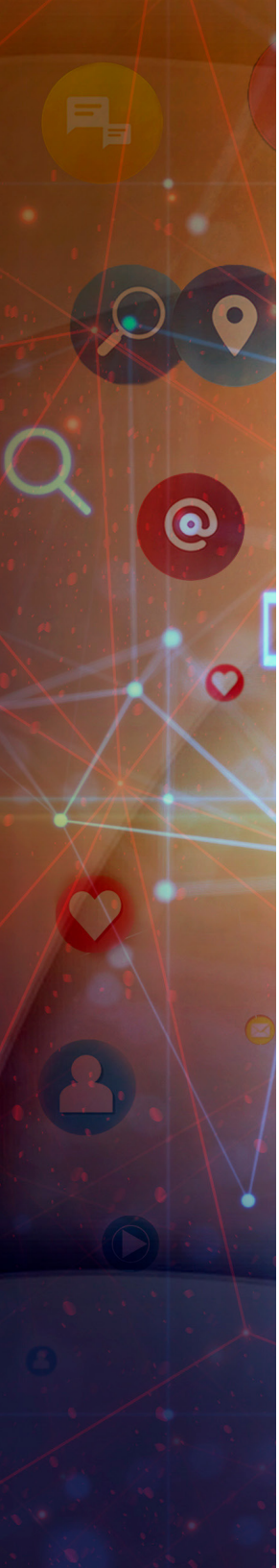
PREFACE

FINALLY BRAZIL GETS ITS BOOK ABOUT WEB ARCHIVING

I have been waiting for *Arquivamento da web e preservação digital* for almost a decade, and it is therefore a great privilege and honour to have the opportunity to preface it.

In April 2015 I was standing in the lobby at the Cardinal Hotel in Palo Alto in Silicon Valley, California, wanting to check in, a bit dizzy and jet lacked after having crossed the Atlantic. Next to me was a guy who I started talking with while waiting. He asked me what I was doing in Palo Alto, and I answered that I had come to attend a conference about *web* archiving (the IIPC 2015 General Assembly Open Conference, hosted by Stanford University and the Internet Archive), and to my big surprise his face immediately lit up, and he told me that he was a Brazilian journalist who had come to write about what was going on at the very same conference. This was my first encounter with *web* archiving in Brazil. At the time I had no idea whether an article was ever published by the journalist, and I have only sporadically kept track of what actually happened to the *web* archiving agenda in Brazil.

When preparing this preface I went through my own archive, and I found out more about the result of the journalist's visit. His name was Carlos Eduardo Entini, the piece he wrote was published 4 May 2015 in the digital journal *Estadão*, it is entitled "IIPC trabalha para salvar a memória da Internet", and, surprisingly, it is still *on-line*

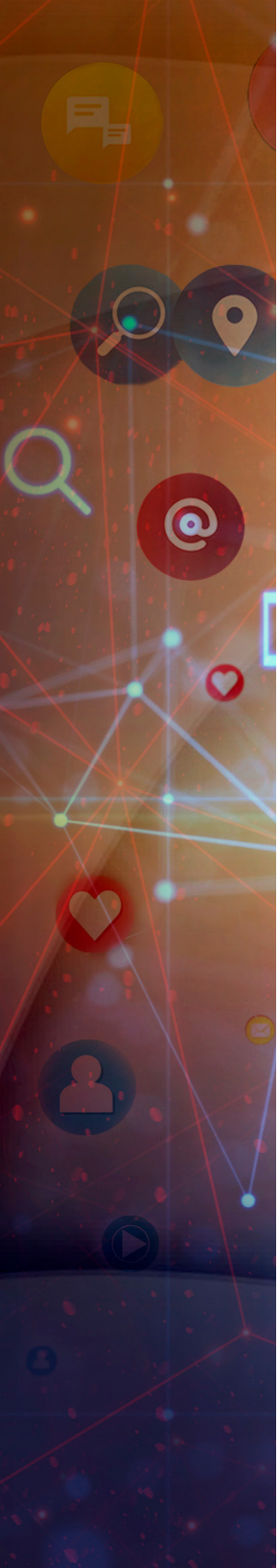


in the journal's archive[1]. Apparently, I did talk more with the journalist, or he just quickly picked up what I said in our chat in the lobby, at least a project of mine is quoted in the article.

Entini is referring his impressions and a number of facts from the conference, including why *web* archives are important to establish, quoting Vint Cerf, the Vice-President of Google whose keynote talk opened the conference: "Corremos o risco de perder muito da nossa história". Most importantly, Entine highlights that "O Brasil ainda está fora dessa nova realidade. Sem uma política oficial de preservação, muitos de seus antigos *site*s sobrevivem graças ao Internet Archive. [...] Em Portugal Mantido pela Fundação para a Computação Científica Nacional, órgão do governo, o Arquivo Português traz mais de 120 milhões de páginas pesquisáveis. É possível encontrar inclusive algumas do Brasil."

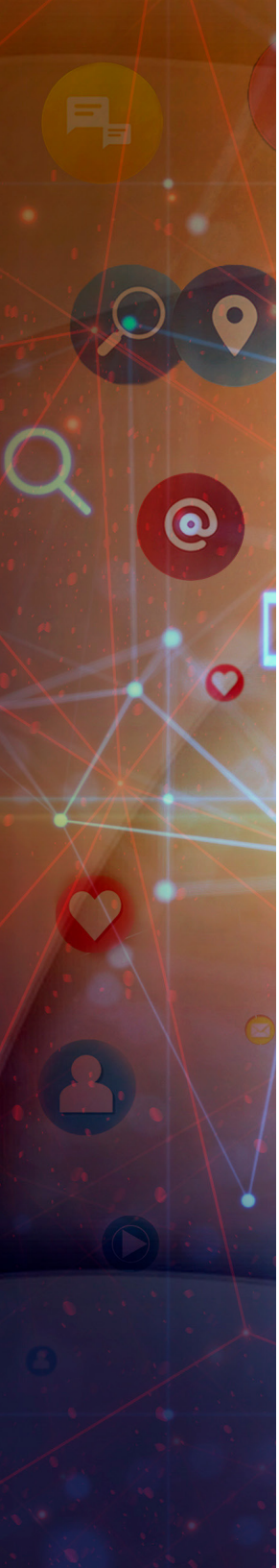
It is obvious that not having a comprehensive national *web* archiving initiative to preserve the cultural heritage and the public record of a nation challenges any scholar, for example future historians who wants to write *web* history understood as 'history with the *web*' as well as 'the history of the *web*' (Brügger, 2018, 5). Not having this important source type available makes it difficult, if not impossible, to study yesterday's and today's social and political movements, youth culture, or government practices in Brazil, just to mention a few examples of studies where the *web* would play a crucial role. As highlighted in the quote above the Internet Archive as well as Arquivo.pt, the Portuguese *web* archive, may be of some help, but scholars cannot rely on that these *web* archives will cover the area in a satisfactory manner — using only their holdings would be like having to write the history of Brazil based on material found in the US Library of Congress or in the Biblioteca Nacional de Portugal: such sources would be better than nothing, but probably not the ones a scholar would prefer, compared to a national Brazilian collection in its own right.

1 https://www.estadao.com.br/acervo/iipc-trabalha-para-salvar-a-memoria-da-Internet



However, the lack of a comprehensive national *web* archiving initiative with the remit of preserving the entire *on-line* record, irrespective of who communicates about what with whom, is only one side of the coin of what is needed to facilitate the writing of *web* history (in both senses of the word). What is evenly important is to raise awareness in the research communities about the archived *web* and its potential use as a source of study, including reflections on methods, and on what the barriers are for researchers who want to start using the archived *web* in their studies (see Winters, 2017).

It is therefore timely and promising that Moisés Rockembach and Caterina Marta Groposo Pavão with this book are now building on top of their previous works on matters related to *web* archiving and *web* archives. *Arquivamento da web e preservação digital* has a number of merits for the reader who wants to become familiar with the many topics related to *web* archiving. It provides useful insights into how the *web* works, and how it has previously worked, it gives a brilliant overview of the various *web* archiving initiatives that have taken place in Brazil throughout the years, as well as of the existing Brazilian literature related to *web* archives from 2014 onwards. In addition, the book provides solid reflections on the politics of and the politics involved in *web* archiving, including on the different archiving strategies and the role of curating, and the reader is informed about how *web* archiving is performed on a technical level, including reflections on quality assurance. Also, important topics such as metadata, storage and access, as well as ethical and legal questions are presented and debated. Finally, it is worth highlighting that although the authors are based in Brazil they are very well embedded in the international literature about *web* archiving, and they combine existing insights in new ways to push the agenda forward. All in all when reading this book a Brazilian scholar who would like to know more about this source type and its potential use is well equipped to explore this uncharted territory.



With *Arquivamento da web e preservação digital* the way is now paved for raising awareness in the Brazilian research communities on how *web* archives can be used as an important source type in various types of studies. This may help foster a research interest in *web* archives, which then, eventually, may have researchers and others push for the establishing of a national, systematic and comprehensive *web* archiving initiative in Brazil. Thus, hopefully Brazil will not remain "fora dessa nova realidade".

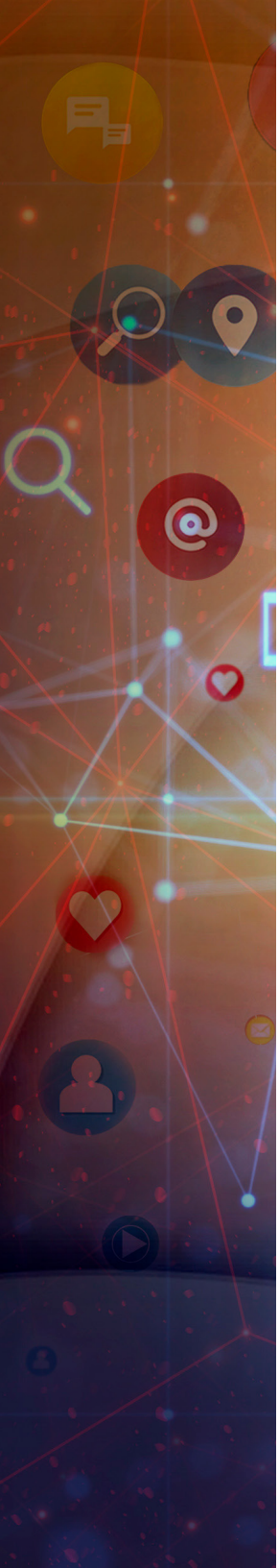
REFERENCES

Brügger, N. (2018). *The archived web: Doing history in the digital age*. Cambridge, MA: MIT Press.

Winters, J. (2017). Breaking in to the mainstream. *Internet Histories: Digital Technology, Culture and Society*, 1(1-2), 173–179.

Professor Dr. Niels Brügger

Aarhus University - Denmark



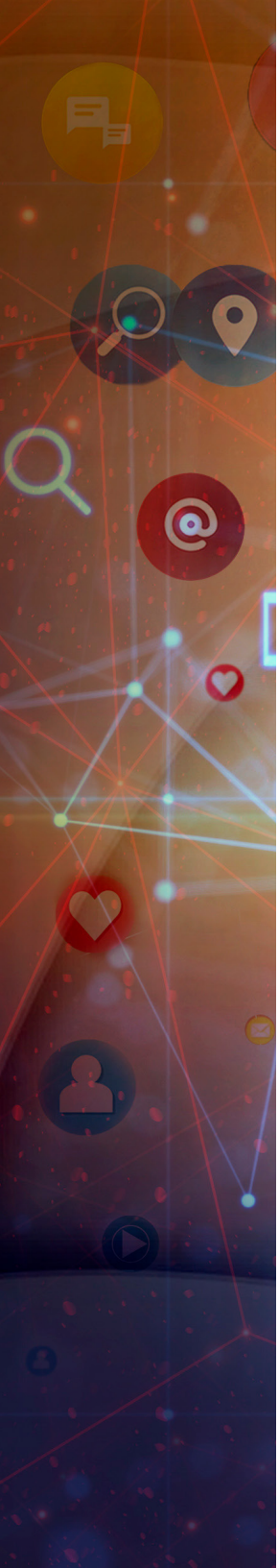
PREFÁCIO

FINALMENTE O BRASIL GANHA SEU LIVRO SOBRE ARQUIVAMENTO DA *WEB*

Há quase uma década que aguardo o Arquivamento da *web* e preservação digital e, portanto, é um grande privilégio e uma honra ter a oportunidade de prefaciá-lo.

Em abril de 2015, eu estava no saguão do Cardinal Hotel em Palo Alto, no Vale do Silício, Califórnia, querendo fazer o *check-in*, um pouco tonto e com *jet leg* depois de cruzar o Atlântico. Ao meu lado, estava uma pessoa com quem comecei a conversar enquanto esperava. Ele me perguntou o que eu estava fazendo em Palo Alto, e eu respondi que tinha vindo para participar de uma conferência sobre arquivamento da *web* (IIPC 2015 General Assembly Open Conference, organizada pela Stanford University e o Internet Archive), e, para minha grande surpresa, o seu o rosto imediatamente se iluminou e ele me disse que era um jornalista brasileiro que veio escrever sobre o que estava acontecendo na mesma conferência. Este foi meu primeiro encontro com o arquivamento da *web* no Brasil. Na época, eu não tinha ideia se algum artigo já havia sido publicado pelo jornalista e apenas esporadicamente acompanhei o que realmente aconteceu com a agenda de arquivamento da *web* no Brasil.

Ao preparar este prefácio, examinei meu próprio arquivo e descobri mais sobre o resultado da visita do jornalista. Seu nome era Carlos Eduardo Entini, e a matéria que escreveu foi publicada em 4 de maio de 2015 no jornal digital *Estadão*, tem como título

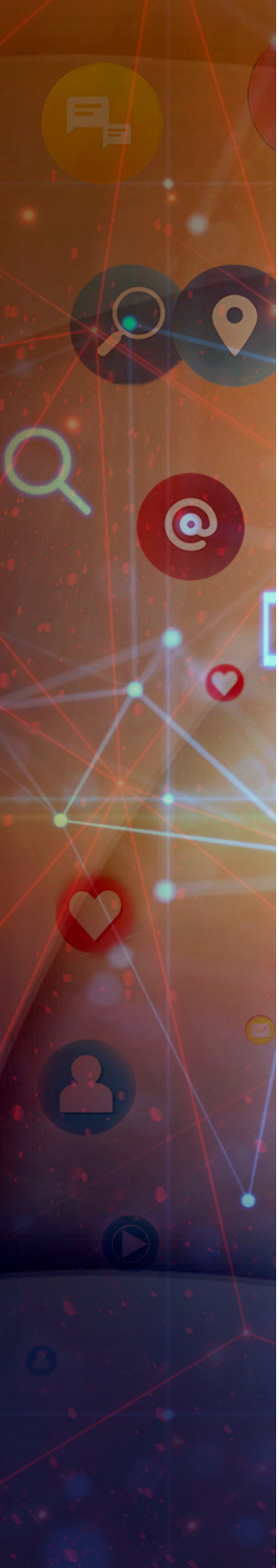


"IIPC trabalha para salvar a memória da Internet"[2] e, surpreendentemente, ainda está *on-line* no *site* do jornal. Aparentemente, conversei bastante com o jornalista, ou ele apenas anotou rapidamente o que eu disse em nosso bate-papo no saguão; pelo menos um projeto meu é citado na matéria.

Entini está se referindo às suas impressões e a uma série de fatos da conferência, incluindo por que os arquivos da *web* são importantes, citando Vint Cerf, o vice-presidente do Google cuja palestra abriu a conferência: "Corremos o risco de perder muito da nossa história". Mais importante ainda, Entine destaca que "O Brasil ainda está fora dessa nova realidade. Sem uma política oficial de preservação, muitos de seus antigos *site*s sobreviveram graças ao Internet Archive. [...] Em Portugal, mantido pela Fundação para a Computação Científica Nacional, órgão do governo, o Arquivo Português traz mais de 120 milhões de páginas pesquisáveis. É possível encontrar inclusive alguns do Brasil."

É óbvio que não ter uma iniciativa nacional abrangente de arquivamento da *web* para preservar o patrimônio cultural e o registro público de uma nação desafia qualquer estudioso, por exemplo, futuros historiadores que desejam escrever a história da *web* entendida como 'história com a *web*', bem como 'a história da *web*' (Brügger, 2018, p.5). A não disponibilidade deste importante tipo de fonte dificulta, senão impossibilita, estudar os movimentos sociais e políticos de ontem e de hoje, a cultura jovem ou as práticas governamentais no Brasil, apenas para citar alguns exemplos de estudos em que a *web* teria um papel crucial. Conforme destacado na citação acima, o Internet Archive, bem como o Arquivo.pt, o arquivo da *web* português, pode ser de alguma ajuda, mas os pesquisadores não podem confiar que esses arquivos da *web* cobrirão uma área de maneira satisfatória – usar apenas seus acervos seria como ter que escrever a história do Brasil com base em material encontrado

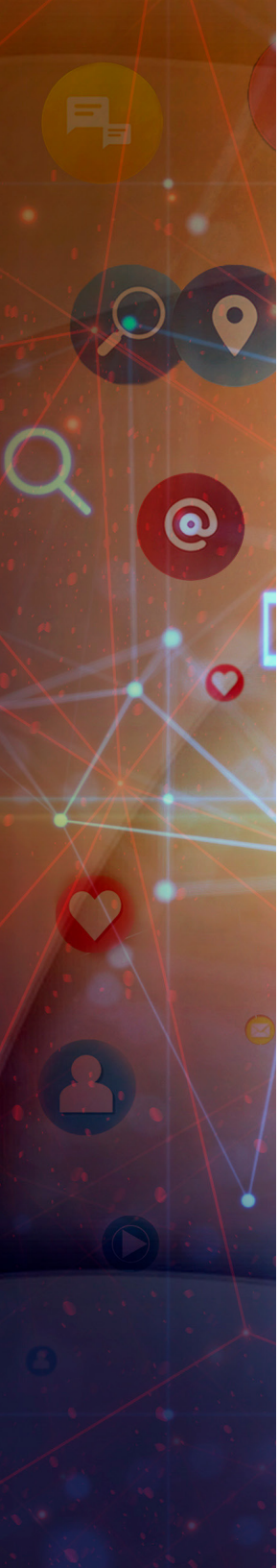
2 https://www.estadao.com.br/acervo/iipc-trabalha-para-salvar-a-memoria-da-Internet



na Biblioteca do Congresso dos Estados Unidos ou na Biblioteca Nacional de Portugal: tais fontes seriam melhores do que nada, mas provavelmente não as que um pesquisador preferiria, em comparação com uma coleção nacional brasileira por direito próprio.

No entanto, a falta de uma iniciativa nacional abrangente de arquivamento da *web* com a missão de preservar todo o registro *on-line*, independentemente de quem se comunica sobre o quê, com quem, é apenas um lado da moeda do que é necessário para facilitar a escrita da história da *web* (em ambos os sentidos da palavra). O que é igualmente importante é conscientizar as comunidades de pesquisa sobre a *web* arquivada e seu uso potencial como fonte de estudo, incluindo reflexões sobre métodos e quais são as barreiras para os pesquisadores que desejam começar a usar a *web* arquivada em seus estudos (WINTERS, 2017).

É, portanto, oportuno e promissor que Moisés Rockembach e Caterina Marta Groposo Pavão estejam agora construindo este livro a partir de seus trabalhos anteriores em assuntos relacionados ao arquivamento da *web* e arquivos da *web*. O livro "Arquivamento da *web* e preservação digital" tem vários méritos para o leitor que deseja se familiarizar com os diversos tópicos relacionados ao arquivamento na *web*. Ele fornece informações úteis sobre como a *web* funciona e como funcionou, dá uma visão geral brilhante das várias iniciativas de arquivamento da *web* que ocorreram no Brasil ao longo dos anos, bem como da literatura brasileira existente relacionada a arquivos da *web* de 2014 em diante. Além disso, o livro fornece reflexões sólidas sobre as políticas envolvidas no arquivamento da *web*, inclusive sobre as diferentes estratégias de arquivamento e o papel da curadoria, e o leitor é informado sobre como o arquivamento da *web* é realizado em nível técnico, incluindo reflexões sobre controle de qualidade. Além disso, são apresentados e debatidos temas importantes como metadados, armazenamento e acesso, além de questões éticas e legais. Finalmente, vale a pena destacar que, embora os autores sejam baseados no Brasil, eles estão muito bem inseridos



na literatura internacional sobre arquivamento na *web* e combinam os conhecimentos existentes com novas maneiras para impulsionar a agenda. Em suma, ao ler este livro, um pesquisador brasileiro que gostaria de saber mais sobre esse tipo de fonte e seu uso potencial estará bem equipado para explorar esse território inexplorado.

Com "Arquivamento da *web* e preservação digital", está aberto o caminho para a conscientização das comunidades de pesquisa brasileiras sobre como os arquivos da *web* podem ser usados como um importante tipo de fonte em vários tipos de estudos. Isso pode ajudar a promover um interesse de pesquisa em arquivos da *web*, o que, eventualmente, pode levar pesquisadores e demais usuários a pressionarem pelo estabelecimento de uma iniciativa nacional, sistemática e abrangente de arquivamento da *web* no Brasil. Assim, esperamos que o Brasil não fique "fora dessa nova realidade".

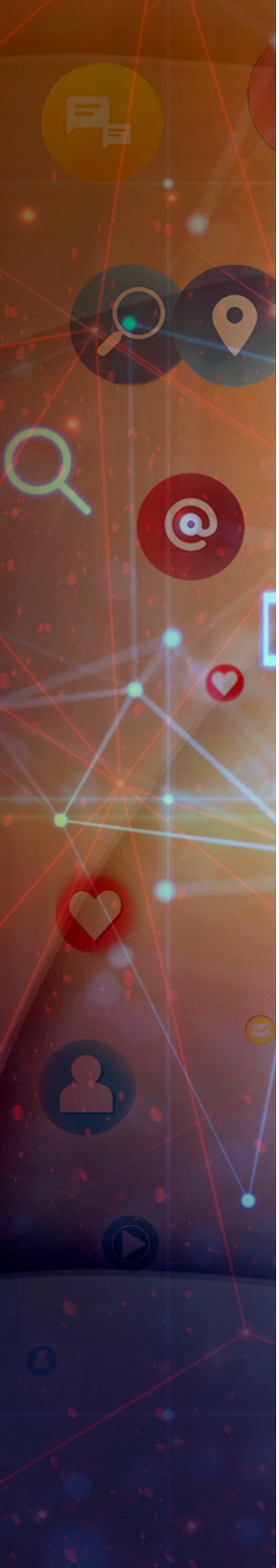
Professor Dr. Niels Brügger

Aarhus University - Dinamarca

REFERÊNCIAS

Brügger, N.. **The archived web**: Doing history in the digital age. Cambridge, MA: MIT Press, 2018.

Winters, J.. Breaking in to the mainstream. **Internet Histories**: Digital Technology, Culture and Society, 1(1-2), 173–179, 2017.

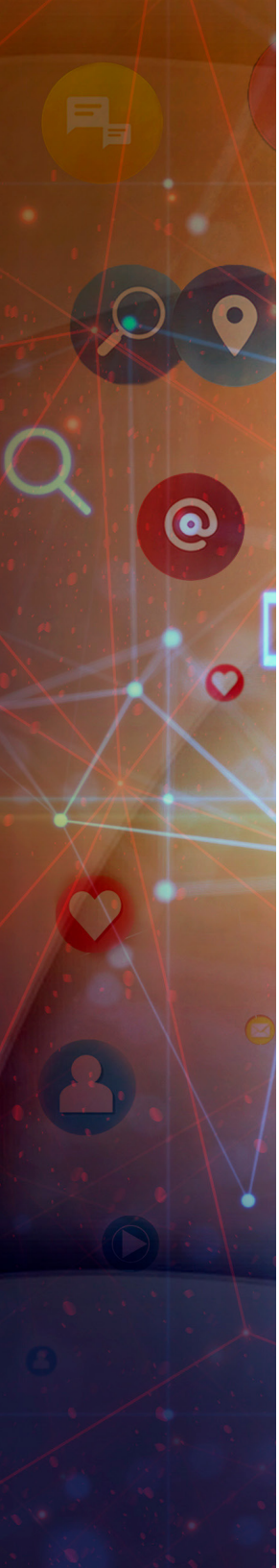


APRESENTAÇÃO

A Web é, sem dúvida, a aplicação sobre a Internet que mais sucesso teve e continua assim. Foi com a entrada da Web que se abriu a janela de oportunidade para que todos passassem a publicar quaisquer opiniões e conteúdos pessoais. A Web foi/é a alavanca para o que John Perry Barlow denominou "Ciberespaço" em sua "declaração de independência do Ciberespaço", de 1996. Diz Barlow: "O nosso é um mundo que está ao mesmo tempo em todos os lugares e em nenhum lugar, mas certamente ele não é onde as pessoas vivem. <...>. Estamos criando um mundo que todos poderão entrar sem privilégios ou preconceitos ligados a raça, poder econômico, força militar ou lugar de nascimento. Estamos criando um mundo onde qualquer um, em qualquer lugar, poderá expressar suas opiniões, não importando quão singulares elas sejam, sem temer que seja coagido ao silêncio ou conformidade." A declaração, que está claramente pautada no espírito inicial da Internet, hoje pode estar encontrando fortes e inesperadas dificuldades para sobrenadar...

Assim, a ideia de preservar a rede, sistema que o engenho humano construiu, e que foi rapidamente povoada com as contribuições de literalmente bilhões de indivíduos, é uma tarefa muito nobre e necessária, mas também ousada e que demanda recursos expressivos. E a coisa se torna ainda mais complexa se considerarmos os embates conceituais recentes, especialmente na área jurídica referente à proteção de dados privados e propriedade intelectual, além de iniciativas de alguns países que consagram o "direito ao esquecimento", ou debatem o eventual direito de "herança" dos dados já depositados na rede.

Do texto em mãos, cito textualmente: "A preservação digital é o conjunto coordenado e contínuo de processos e atividades que

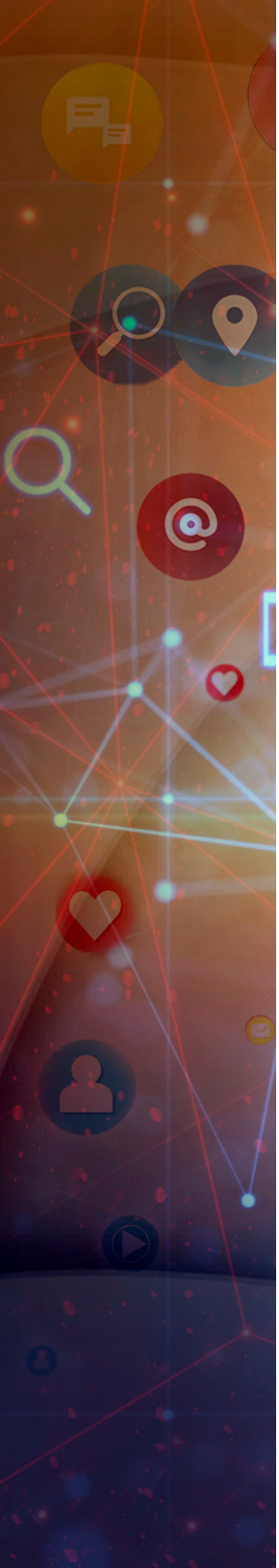


garantem o armazenamento de longo prazo e sem erros da informação digital, com meios para recuperação e interpretação, durante todo o período de tempo em que a informação é necessária", e "Os arquivistas da web desempenham um papel essencial na preservação do patrimônio cultural da internet, e seu trabalho é essencial para pesquisadores, historiadores e o público em geral que dependem do acesso à informação da web. Sem os arquivistas da web, muitas informações valiosas na internet seriam perdidas e as gerações futuras seriam incapazes de acessá-las e aprender com elas".

Há desde o início dos anos 2000 iniciativas para se preservarem historicamente conteúdos e sítios na Web, mas esse é um "alvo móvel", com complexidade novas e crescentes, e que a obra trata de abordar. Além das questões técnicas de como recolher automaticamente os dados que estão na Internet e, mais especificamente, na aplicação Web, há pontos delicados a tratar, por exemplo: o que guardar, o que ignorar, o que remover – e usando quais argumentos para tanto...

Também é igualmente importante conseguir catalogar as informações, de forma a que sua recuperação seja o mais racional possível. Uma alternativa proposta no caso foi o olhar a formas de identificação como o DNS. Explorar a ferramenta DNS pode auxiliar na classificação. São também abordadas tecnologias diversas da majoritariamente utilizada, que se baseia na suíte TCP/IP. O processo de busca e captura da informação descrito no texto, certamente poderá sofrer aportes futuros, especialmente ao se considerar o aporte de ferramentas novas, como as de Inteligência Artificial.

Finalmente, vem à mente uma analogia histórica importante: aquela com a biblioteca de Alexandria, que concentrava de forma organizada o antigo saber. Vint Cerf, pioneiro da Internet, num texto de 2005 comparava a necessidade de preservação dos dados na Internet com a riqueza de conhecimento armazenada em Alexandria até a destruição, por volta do ano 650, da biblioteca. E ele alerta para



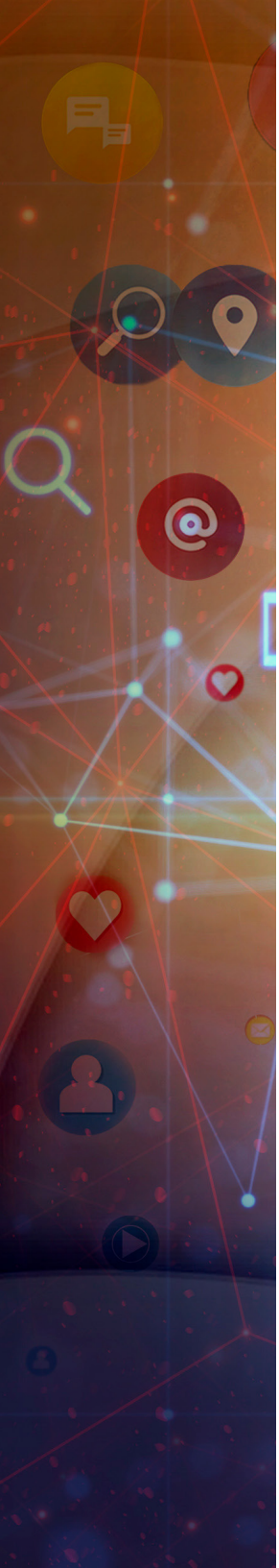
o fato de que não basta preservar a informação digital em si, se não forem também preservados os mecanismos e dispositivos que dão acesso a ela. Se velhos manuscritos em pergaminho sobreviveram milhares de anos e seguem legíveis, foi pela durabilidade do meio usado (pergaminho e outros) e pelo fato de continuarmos em condição de entender a escrita usada. Como se comportariam arquivos digitais armazenados, em meios que rapidamente se tornam obsoletos ou simplesmente desaparecem, acessados por programas que não mais rodam nos computadores atuais? Essa é uma questão crítica, a que não se pode deixar de atentar.

Finalmente, quanto ao que preservar e que descartar, de novo o exemplo da destruição da biblioteca de Alexandria pode ser um guia útil. Sem discutir outros eventos históricos que tenham contribuído na destruição parcial de seu magnífico arquivo, cito o que J. J. Rousseau escreveu numa nota ao seu "Discurso sobre as Ciências e as Artes", de 1750. Rousseau relata que "... dizem que o Califa Omar, quando perguntado sobre o que fazer com a biblioteca da Alexandria, teria respondido: 'se os livros contidos na biblioteca contêm temas opostos ao Corão, eles são nocivos e devem ser queimados. Se esses livros contêm apenas doutrinas alinhadas ao Corão, eles são supérfluos e devem ser queimados do mesmo jeito.'

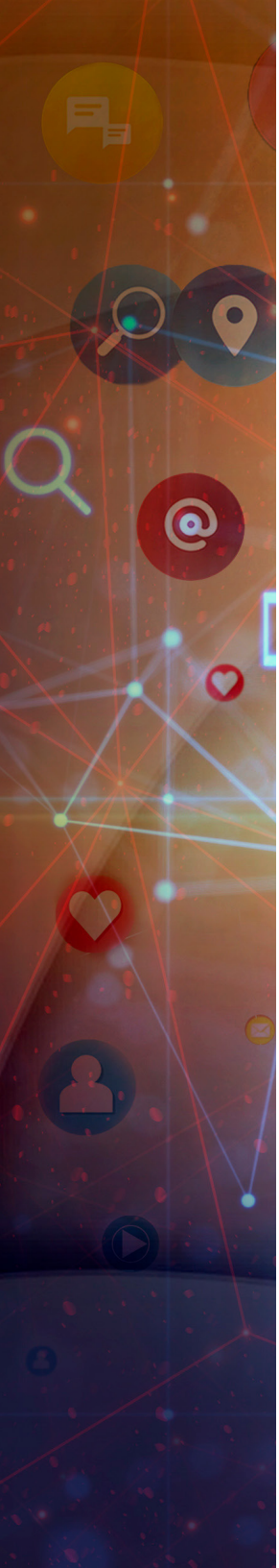
Que este importante livro que o leitor tem às mãos seja instrumental para que não se repita com a Internet o que ocorreu com os documentos da biblioteca de Alexandria. Boa leitura!

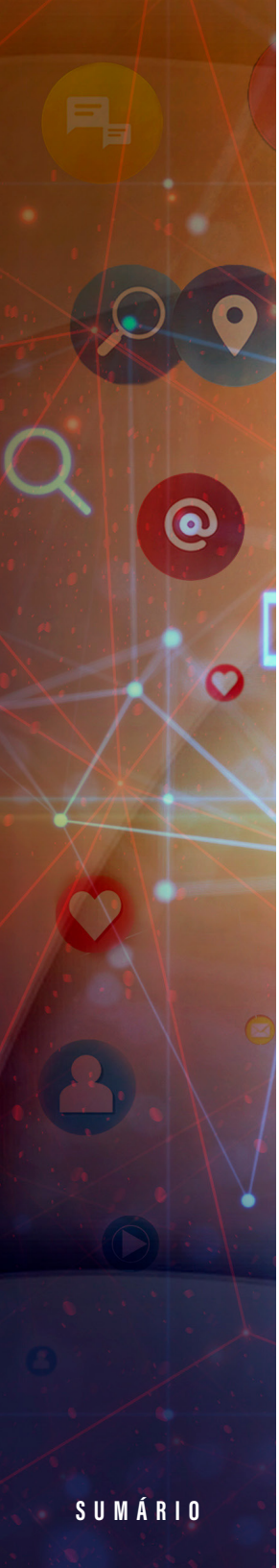
Professor Dr. Demi Getschko

Diretor-Presidente do Núcleo de Informação
e Coordenação do Ponto BR, NIC.BR



SUMÁRIO





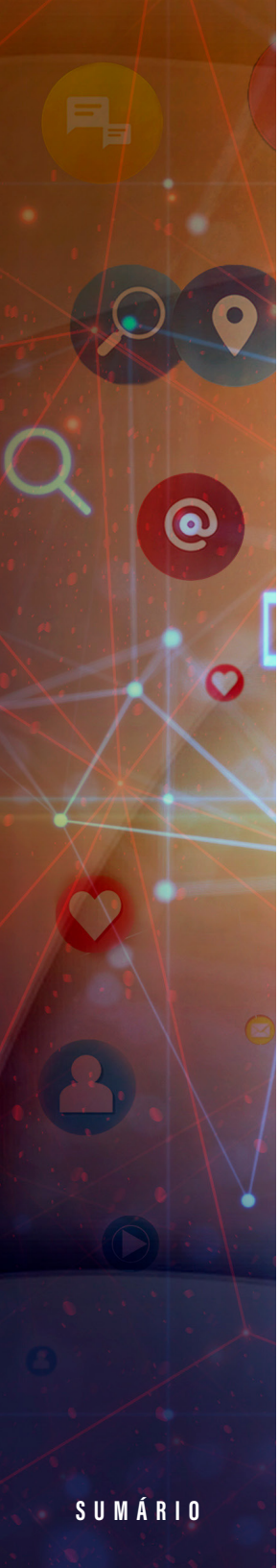
INTRODUÇÃO

O arquivamento da *web* é o processo de preservação do conteúdo da *World Wide Web* para futuros pesquisadores, o que inclui capturar e armazenar informações de *sites*, plataformas de mídia social e outros recursos *on-line*. O arquivamento da *web* nos permite preservar o histórico dos conteúdos *on-line* e garantir que informações valiosas não sejam perdidas à medida que *sites* e outros recursos *on-line* são atualizados ou desativados.

Muitos materiais que as instituições de memória costumam custodiar e armazenar, como documentos governamentais, publicações acadêmicas, correspondência e notícias, agora são encontrados exclusivamente na *web*. Isso criou uma série de desafios para as instituições encarregadas de preservar a cultura e o conhecimento contemporâneos.

Uma das questões cruciais sobre o arquivamento da *web* se refere à seleção e avaliação dos materiais a serem preservados, o que envolve decidir quais *site*s e recursos *on-line* são importantes o suficiente para serem incluídos no arquivo e avaliar seu valor, relevância, bem como os riscos de perda destes conteúdos. Esse processo geralmente é realizado por arquivistas, bibliotecários e outros profissionais especializados em identificar e analisar a importância de materiais *on-line* e elaborar políticas que orientem práticas e processos de preservação. Nesse sentido, também são analisadas a proveniência e a pertinência institucional das informações disponíveis na *web*.

As ferramentas de arquivamento da *web* são usadas para capturar e armazenar o conteúdo de *site*s e outros recursos *on-line*. Essas ferramentas variam de rastreadores da *web* simples que podem capturar a estrutura básica e o conteúdo de um *site*, até ferramentas

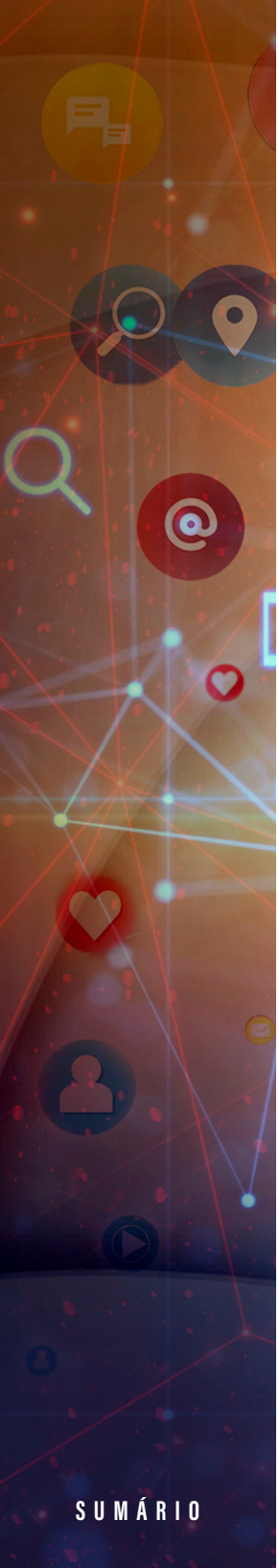


mais sofisticadas que podem capturar informações mais detalhadas, como código HTML e *links* entre páginas diferentes. Algumas plataformas de arquivamento da *web* também permitem que os usuários criem seus próprios arquivos de *site*s e recursos *on-line*, que podem ser compartilhados com outras pessoas para pesquisa ou outros fins.

Há também uma série de questões éticas e legais associadas ao arquivamento da *web*. Por exemplo, preocupações com a privacidade sobre a coleta e armazenamento de informações pessoais de *site*s e plataformas de mídia social. Além disso, pode haver questões legais envolvendo o uso de material protegido por direitos autorais em arquivos da *web*, principalmente se o material for usado para fins comerciais.

Existem vários modelos diferentes de arquivamento da *web*, cada um com seus pontos fortes e fracos. Alguns arquivos da *web* são mantidos por grandes instituições, como Arquivos, Bibliotecas, Museus e instituições de pesquisa e tecnologia, enquanto outros são criados e gerenciados por organizações menores ou usuários individuais. Alguns arquivos da *web* são focados em instituições, tópicos ou assuntos específicos, enquanto outros são mais abrangentes em escopo. Independentemente do modelo utilizado, o objetivo do arquivamento da *web* é preservar a história da Internet e disponibilizar informações importantes ao longo do tempo.

O fluxo do arquivamento da *web* compreende coletar, armazenar, preservar e disponibilizar a informação retrospectiva da *web* para futuros pesquisadores e para o público em geral. Este processo envolve iniciativas no mundo inteiro, algumas com abordagens globais, outras localizadas geograficamente, com foco em seus respectivos países, atributo identificado pelo domínio do endereço eletrônico ou a partir da verificação do produtor da informação e o contexto no qual se insere. O livro conta com capítulos, abordando: origens da *World Wide Web* e dos arquivos da *web*, as políticas de preservação da *web*, modelos e procedimentos técnicos no arqui-

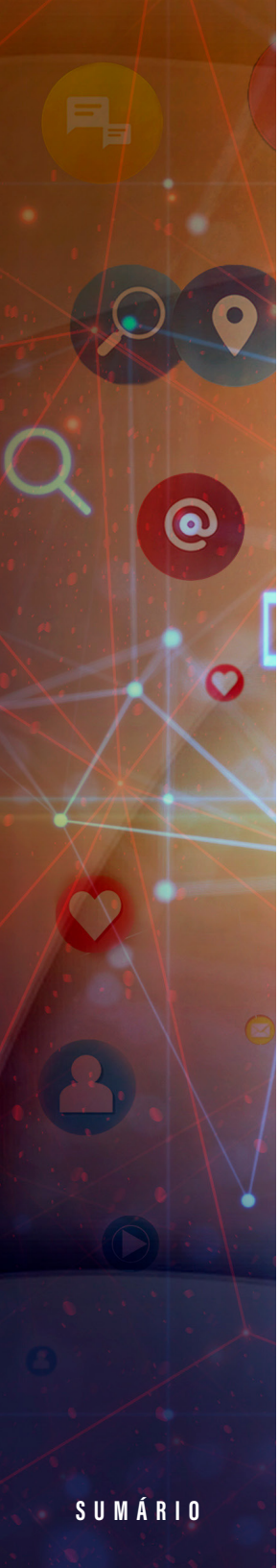


vamento da *web* e perspectivas de estudos e aplicações profissionais do arquivamento da *web*. Ao final, ainda traz um dicionário com os principais termos e referências neste campo de atuação. O conjunto de todos estes capítulos permite abordar as diversas etapas e facetas da captura, preservação e (re)uso dos arquivos da *web*, de forma a servir de material de referência para arquivistas, bibliotecários e profissionais da informação e tecnologia que desenvolvam projetos de preservação digital.

A preservação da *web* vem sendo realizada por muitas instituições internacionais desde 1996, dentre as iniciativas que iniciaram neste período constam o Internet Archive (Estados Unidos), Pandora (Austrália) e KulturaW3 (Suécia). Hoje, diversos países contam com seus arquivos da *web* preservados, sobretudo nos Estados Unidos e Europa, mas também com exemplos espalhados em todos os continentes e diversos países, incluindo o Brasil.

A preservação digital e a produção de arquivos da *web* são importantes por vários motivos. Em primeiro lugar, permitem manter registros da história da Internet, o que pode ser útil para entender como a *web* evoluiu ao longo do tempo. Além disso, os arquivos da *web* podem ser usados como um recurso para verificação de fatos. Por exemplo, se um *tweet* foi excluído ou se uma declaração em uma página da *web* foi alterada, os arquivos da *web* podem nos ajudar a rastrear essas alterações e verificar sua autenticidade. Também, privilegia a preservação da memória das instituições como uma forma de manter a história e fortalecer suas bases, na medida que organiza, registra, conserva e guarda os registros para futuras gerações.

Outro aspecto importante dos arquivos da *web* é que eles promovem a transparência e o acesso público. Muitos governos, por exemplo, costumam manter arquivos dos *site*s de cada administração, o que permite ao público acessar informações sobre políticas e ações passadas. Isso é crucial para manter a responsabilidade e a confiança nas instituições governamentais. Desta forma, os



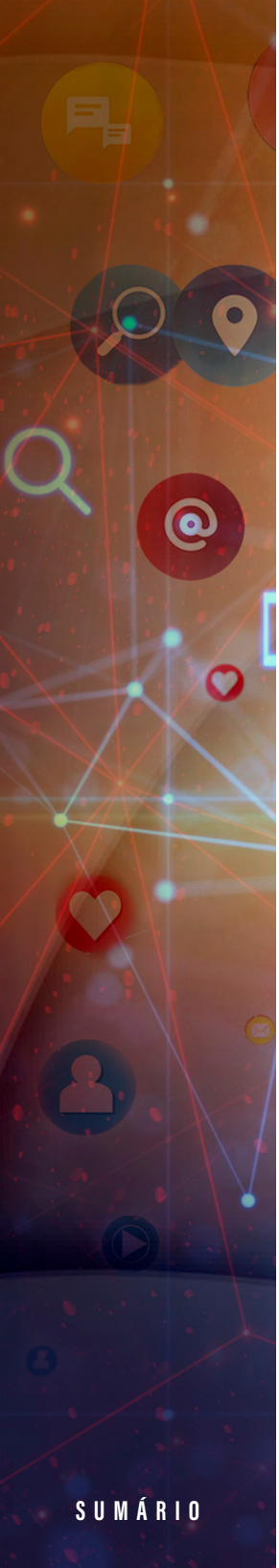
arquivos da *web* servem como uma ferramenta importante para preservar a integridade da informação e facilitar o processo de transparência pública e exercício da cidadania.

Moisés Rockembach

Caterina Groposo Pavão

1

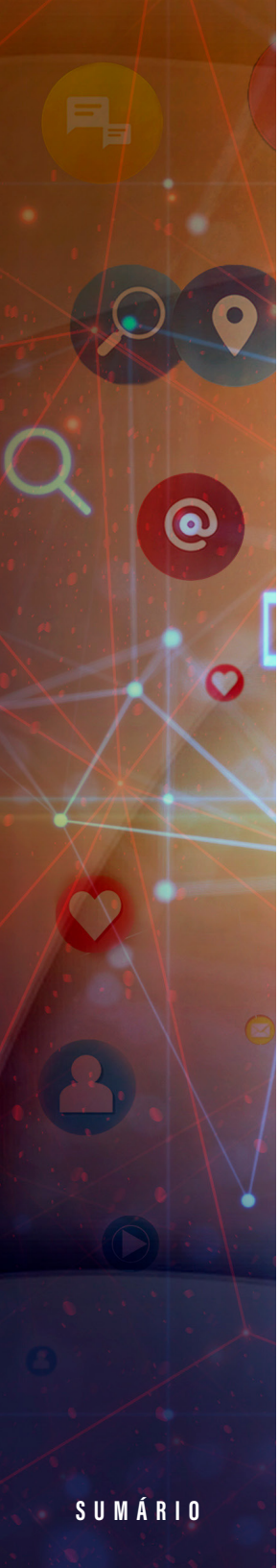
ORIGENS DA *WORLD WIDE WEB* E DOS ARQUIVOS DA *WEB*



A Internet e a *web* são termos frequentemente utilizados como sinônimos e de forma intercambiável, mas na verdade referem-se a conceitos diferentes. A Internet constitui-se na rede global de computadores conectados que permite a troca de informações e comunicação ao redor do mundo. Criada em meados dos anos 1960 como uma rede de computadores militares nos Estados Unidos, expandindo-se ao longo dos anos, revolucionou as formas de comunicação, o acesso e a troca de conhecimento, permitindo criar, organizar, manter, gerenciar, acessar, compartilhar e preservar conjuntos de documentos digitais.

A *Advanced Research Projects Agency* – ARPANET, rede única que conectava algumas dezenas de recursos, transformou-se na Internet, um sistema de muitas redes interconectadas capaz de expansão quase indefinida. A criação da Internet foi motivada por uma série de eventos imprevistos e representou uma nova abordagem para a rede. A arquitetura da Internet proposta por Robert Kahn e Vinton Cerf acabou sendo usada não apenas para construir a própria Internet, mas também como modelo para outras redes (ABBATE, 1999). Após separar os usuários militares e pesquisadores acadêmicos da ARPANET e comercializar a tecnologia da Internet, a disseminação dos protocolos ARPA tornou-se um padrão para redes. A Internet se afastou do projeto da ARPANET, resultando em um sistema que não era apenas maior, mas também mais flexível e descentralizado. Embora Cerf e Kahn fossem os principais arquitetos da Internet, eles contaram com vários colaboradores do grupo ARPANET e uma crescente comunidade internacional.

Já as origens da *World Wide Web* (WWW ou *web*) remontam ao final dos anos 1980, quando uma equipe de pesquisadores da Organização Europeia para Pesquisa Nuclear, do francês *Conseil Européen pour la Recherche Nucléaire* (CERN), começou a trabalhar em um sistema para compartilhar informações e colaborar em projetos de pesquisa.



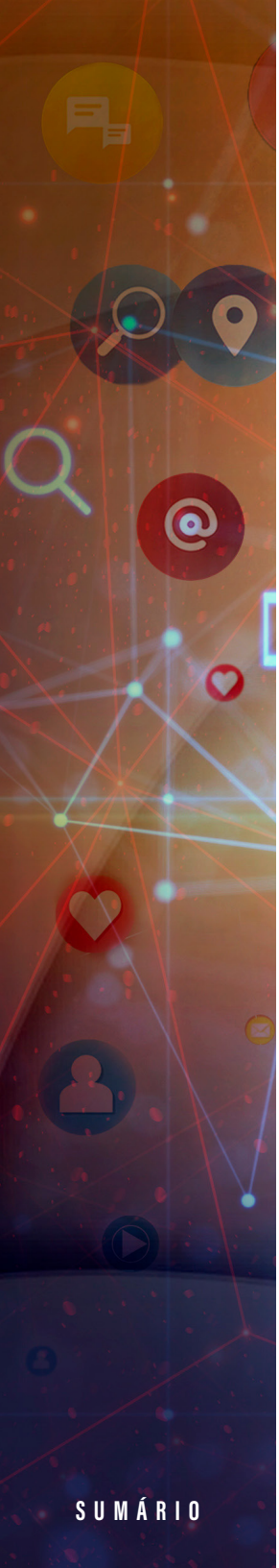
Uma das bases predecessoras da *web*, o hipertexto, é uma forma de linguagem que permite a organização de informações em uma rede não linear de conexões, permitindo a navegação livre entre diferentes partes do texto. O conceito de hipertexto não é novo e começa antes da *web* propriamente dita. Um dos exemplos do uso da escrita hipertextual remonta o século XV, quando Leonardo da Vinci realizava anotações nas margens das páginas de seus escritos[3].

A ideia de uma rede de conhecimento para preservar a produção bibliográfica mundial popularizou-se entre os séculos XIX e XX entre vários pensadores como Paul Otlet, com o Repertório Bibliográfico Universal ou *Bibliographia Universalis*; Henry La Fontaine fundador do *Centre Intellectuel Mondial*; Herbert George Wells, com sua visão do "cérebro do mundo"; Vannevar Bush, com a ideia do "memex"; Joseph Licklider, com seu sonho de uma rede de centros de pensamentos; Roberto Busa, criador do *Index Thomisticus* e Ted Nelson, que cunhou os termos *hipertexto* e *hipermídia*.

Paul Otlet (1868-1944) foi um pensador à frente de seu tempo, principalmente quanto a recolhimento, seleção, organização, armazenamento e disseminação do conhecimento em atividades definidas na concepção do *Palais Mondial*, em Bruxelas, em 1920, chamado mais tarde de *Mundaneu*m[4]. Embora o projeto *Mundaneum* não tenha se concretizado conforme o planejado, principalmente por conta da Segunda Guerra Mundial, grande parte de seus objetivos foram alcançados. O *Mundaneum* pode ser categorizado como uma rede de informação, compreendida como um corpo mundial onde se encontrariam registros da enciclopédia do conhecimento. Era um projeto que reuniria e disseminaria o conhecimento universal em um só lugar físico, com acesso universal, por meio de uma rede global,

3 Veneranda Biblioteca Ambrosiana | Codex Atlanticus é um projeto de design de informação que permite explorar a maior coleção existente de escritos e desenhos originais de Leonardo da Vinci - https://codex-atlanticus.ambrosiana.it/

4 *Centre d'archives* | *Mundaneum* - http://archives.mundaneum.org/fr



mediado por uma rede de comunicação, cooperação e intercâmbio. O projeto seria baseado nos princípios da totalidade, simultaneidade, gratuidade, voluntariedade, universalidade e mundialidade. O *Mundaneum* teve grande importância no desenvolvimento do hipertexto e da *web*, e inspirou muitos pensadores e inovadores ao longo do século XX e XXI.

O projeto *Mundaneum* de Paul Otlet visava à centralização e disseminação do conhecimento. A utopia era que o *Mundaneum* se configurasse como o centro de informações das nações, reunindo toda produção intelectual do mundo. Assim, de acordo com Zafalon e Nóbrega de Sá (2019), pode-se comparar o *Mundaneum* e o protótipo da Biblioteca Digital Mundial (BDM), lançado na Conferência Geral da UNESCO de 2007, pois ambos os projetos possuem semelhanças nos seus objetivos e constatam-se pontos em comuns no sonho-utopia de Otlet e a BDM:

a. Atividades de recolhimento, seleção, organização, armazenamento e disseminação do conhecimento;

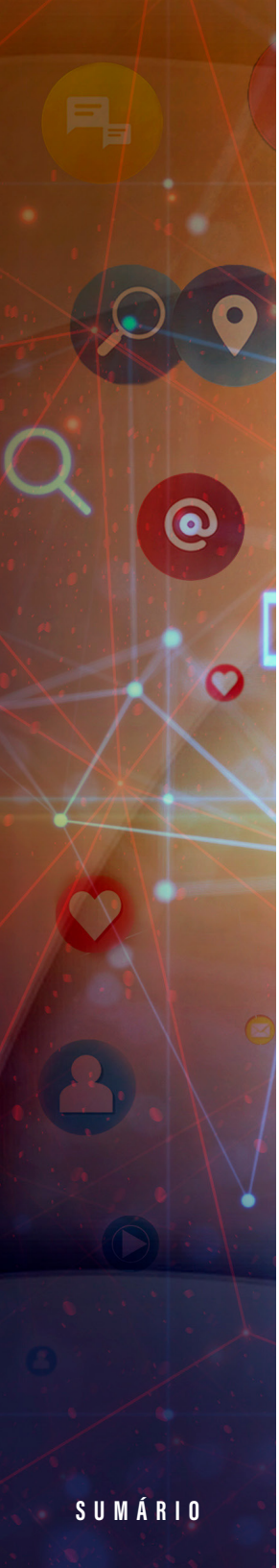
b. acesso onipresente: acesso em qualquer lugar do mundo ao mesmo tempo;

c. acesso rápido à informação: eficiência e rapidez;

d. apoio de instituições: não é composta somente de uma entidade/instituição;

e. caráter internacionalista: visando a disseminação do conhecimento por meio de informações que não se limitam a uma única nacionalidade, cooperação entre as nações;

f. variedade de recursos informacionais: manuscritos, mapas, livros raros, gravações, filmes, gravuras, fotografias, desenhos arquitetônicos e outros tipos de fontes básicas;



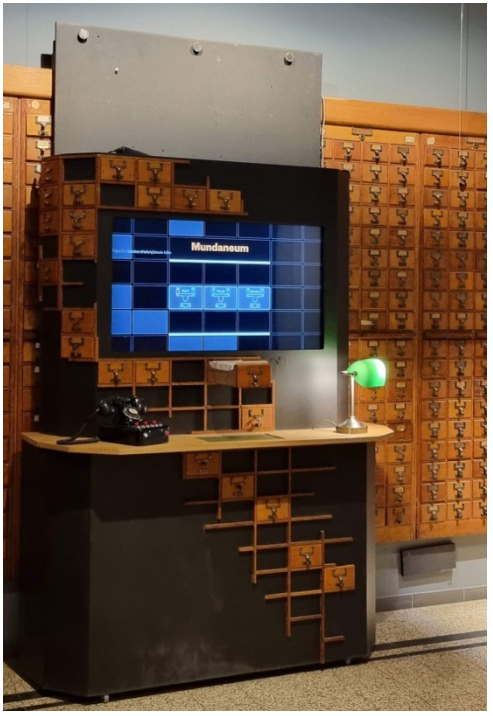
g. preservação da memória: disseminação, uso e preservação de dados de conteúdo cultural;

h. participação de diferentes esferas, desde a local até a internacional.

Henri La Fontaine (1854-1943) junto com Paul Otlet fundou a *Union of International Associations*. Ele também é o cofundador do *Institut International de Bibliographie* (que mais tarde se tornou a Federação Internacional de Informação e Documentação (FID) junto com Paul Otlet. Foi nessa função que ele e Otlet participaram do Congresso Mundial de Documentação Universal em 1937.

Além da contribuição com o desenvolvimento da ideia de hipertexto, Otlet e La Fontaine foram dois importantes pensadores e ativistas belgas. Trabalharam em cooperação na Seção Bibliográfica, da Sociedade de Estudos Sociais e Políticos de Bruxelas, que era coordenada por La Fontaine. Possuíam familiaridade com os diversos tipos de catálogos, índices e serviços de resumos de sua época, além de ter conhecimento dos problemas práticos da cooperação bibliográfica e da padronização de métodos e processos. De acordo com Simeão e Fontoura (2014) já se podia verificar o embrião dos projetos de cunho internacionalista que passaram a ser desenvolvidos por estes bibliógrafos e como a proposta de otimização e uso de repositórios universais estava relacionada com os interesses que Paul Otlet demonstrava ter na época. Em 1895, Otlet e La Fontaine obtêm acesso a um exemplar da *A Classification and subject index for cataloguing and arranging the books and pamphlets of a library* de Melvil Dewey e passam a estudá-lo em todos seus pormenores — assim foi criada a Classificação Decimal Universal (CDD), segundo os próprios autores, a primeira nomenclatura para todo o conhecimento humano, fixo, universal e próprio para ser expresso em uma linguagem internacional – a dos números.

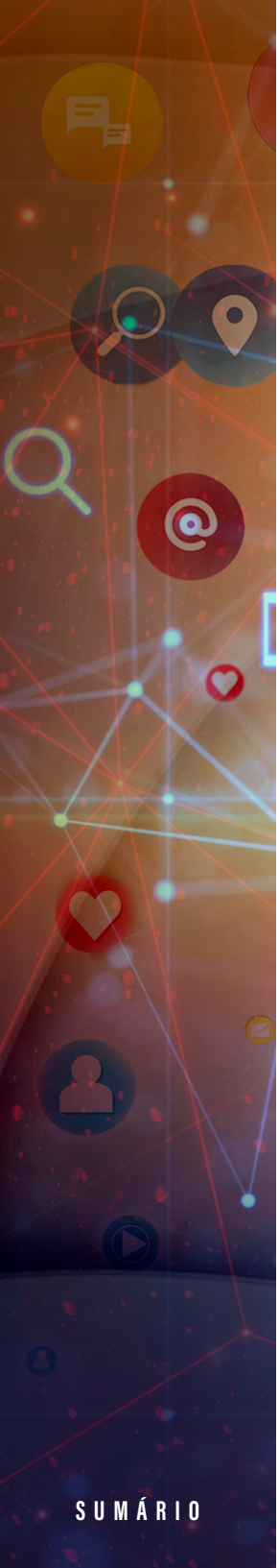
Figura 01 – Exposição do museu Mundaneum em Mons (Bélgica)



Fonte: Os autores (2023).

Herbert George Wells (1866-1946), em 1937, explora a ideia de uma "enciclopédia mundial permanente" que conteria toda a memória humana e que seria uma síntese mundial de bibliografia e documentação com os arquivos indexados do mundo todo. Este escritor britânico foi responsável por uma máquina que chamava de "cérebro do mundo", uma rede que congregaria todo o conhecimento humano – acessível e sem censura para alguns e devidamente filtrado para o restante da população. Acreditava que tal máquina seria possível ainda durante o século XX.

Vannevar Bush (1890-1974), foi um engenheiro, inventor e político estadunidense, conhecido pelo seu papel político no desenvolvimento



da bomba atômica, mas também conhecido pela ideia do "memex", visto como um conceito pioneiro, precursor da *web*. Bush, em seu famoso ensaio de 1945 "*As We May Think*", propôs um dispositivo chamado "memex" (*memory* + *index*), que permitiria a organização e acesso a informações em uma rede de conexões, semelhante ao que hoje conhecemos como hipertexto. O *site memex and beyond*[5] é um projeto que demonstra um panorama histórico das pesquisas sobre *hipermídia.* Bush imaginou uma máquina com o funcionamento da mente humana, que funciona sempre por associações e é capaz de estocar quantidades imensas de informações. O "memex" foi uma máquina visionária imaginada para auxiliar a memória e guardar conhecimentos.

Carl Robnett Licklider (1915-1990), em 1963, apresentou o conceito de uma "Rede intergaláctica de computadores". Essas ideias continham quase tudo que compõe a Internet contemporânea. Previu a computação interativa de estilo moderno e sua aplicação a todos os tipos de atividades, considerado um pioneiro da Internet, com a visão de uma rede mundial de computadores muito antes de ser construída. Ainda financiou pesquisas como a da ARPANET, a predecessora direta da Internet.

Outra inspiração vem de Jorge Luis Borges (1899-1986), que é conhecido por suas histórias que exploram a natureza da linguagem e da realidade. Na "A Biblioteca de Babel" (BORGES, 2000), ele imagina uma biblioteca infinita composta por todos os livros possíveis, organizados em uma rede de conexões que permite a livre navegação entre eles e, neste caso, representando a transfiguração do mundo real para o virtual (VIRGIL, 2007) Essa ideia antecipa a noção de hipertexto, que permite uma organização não linear de informações, assim como o seu conto "O jardim dos caminhos que se bifurcam" (BORGES, 2000).

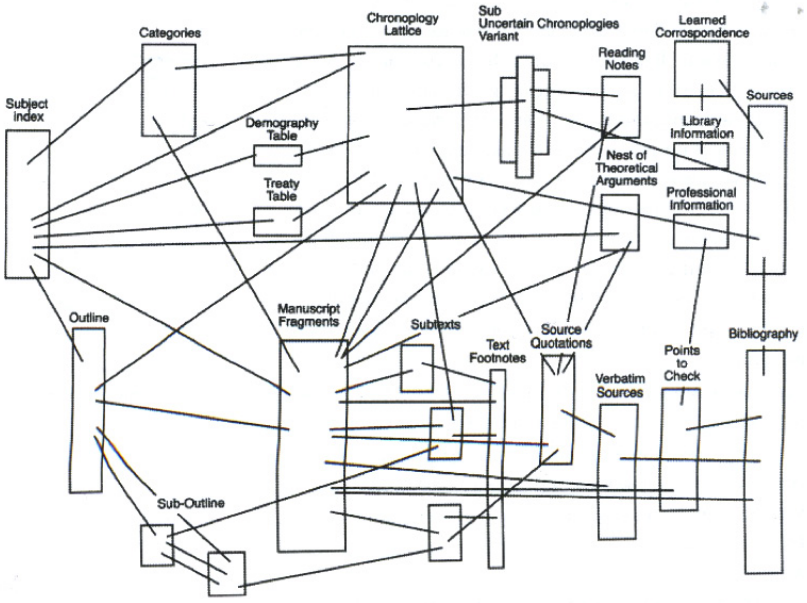
Roberto Busa (1913-2011) foi um jesuíta italiano que desenvolveu um sistema de indexação eletrônica da obra de Santo Tomás de Aquino, chamado *Index Thomisticus*, um trabalho de 50 anos com

5 https://cs.brown.edu/memex/

nove milhões de palavras (NYHAN; PASSAROTTI, 2019). Porém, os computadores IBM da época não conseguiam relacionar esses conteúdos. Busa insistiu para que se desenvolvesse um sistema para uni-los. Então, nasceu, em 1940, o projeto do hipertexto, ou seja, uma estrutura que nos permite compartilhar e relacionar informações entre fontes diferentes por meio de *links*.

Theodor Holm Nelson, mais conhecido como Ted Nelson, por sua vez, cunhou o termo "hipertexto" em 1965, em seu projeto Xanadu[6]. O Xanadu foi um dos primeiros sistemas hipertextuais, que permitia a criação e acesso a documentos em uma rede não linear de conexões (NELSON, 1965). O projeto Xanadu, no entanto, nunca foi concluído, mas influenciou fortemente o desenvolvimento posterior do hipertexto e da *World Wide Web*.

Figura 02 - Estrutura de documentos hipertextuais



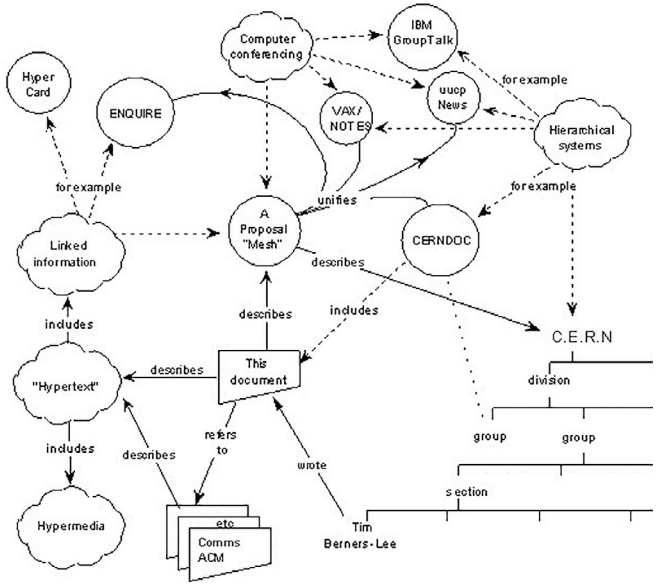
Fonte: NELSON (1965).

6 https://www.xanadu.com/

Um dos personagens mais importantes no contexto da *web* foi Tim Berners-Lee, um cientista da computação britânico, a quem se atribui a invenção da *World Wide Web*. Nascido em Londres, Inglaterra, em 1955, Berners-Lee estudou física no *Queen's College*, em Oxford, onde desenvolveu interesse por computadores e *software*. Em 1989, Tim Berners-Lee propôs o conceito da *World Wide Web*, que permitiria aos usuários acessar e vincular documentos na Internet usando hipertexto. Isso levou posteriormente ao desenvolvimento do primeiro navegador da *web*[7].

O artigo de Tim Berners-Lee, *Information Management* (BERNERS-LEE, 1998) pode ser considerado o marco conceitual da *web*, na medida que demonstra algumas das principais características da organização da informação seguindo uma estrutura não necessariamente hierárquica e o uso de *hiperlinks* como ilustrado na figura.

Figura 03 - Proposta de gestão da informação com o uso de *hiperlinks*

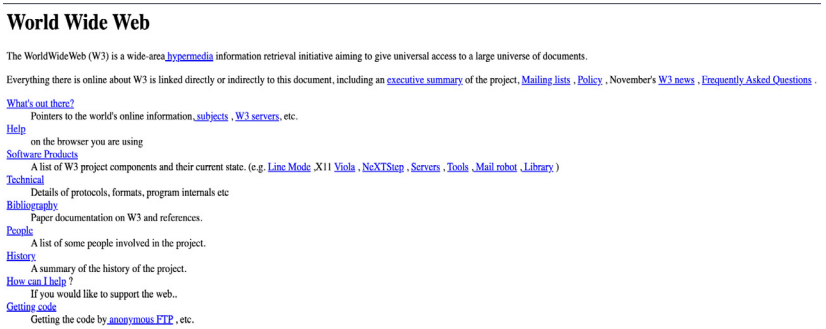


Fonte: Berners-Lee (1989).

7 https://worldwideweb.cern.ch/browser/

O primeiro *site* foi criado dois anos depois da concepção de Tim Berners-Lee. Em 6 de agosto de 1991 Berners-Lee publicou um *site* (fig. 04) que continha informações sobre o projeto *World Wide Web* que permitiria aos usuários visualizar e navegar em páginas da *web* e, assim, a *web* rapidamente se tornou um fenômeno global, com milhões de *site*s e usuários.

Figura 04 - A primeira página da *web*, com informações do projeto



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

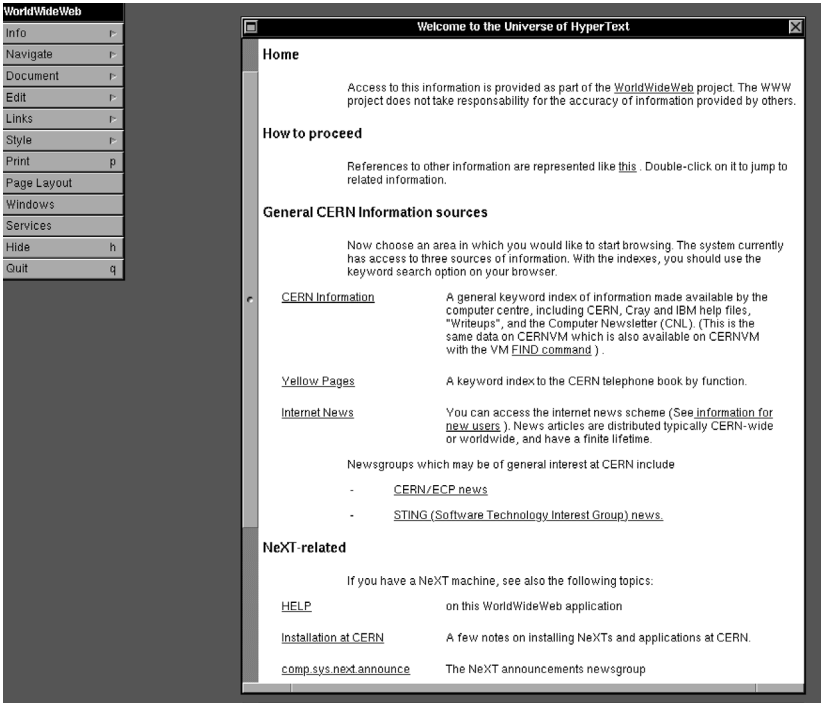
What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

Fonte: http://info.cern.ch/hypertext/WWW/TheProject.html. Acesso em: 10 mar. 2023.

O primeiro *website* (fig. 5) criado por Tim Berners-Lee ainda pode ser visualizado, graças a um recurso *on-line* que disponibiliza informações sobre o projeto[8].

8 http://info.cern.ch/hypertext/WWW/TheProject.html

Figura 05 - O primeiro navegador da *web, World Wide Web*



Fonte: https://worldwideweb.cern.ch/browser/. Acesso em: 10 mar. 2023.

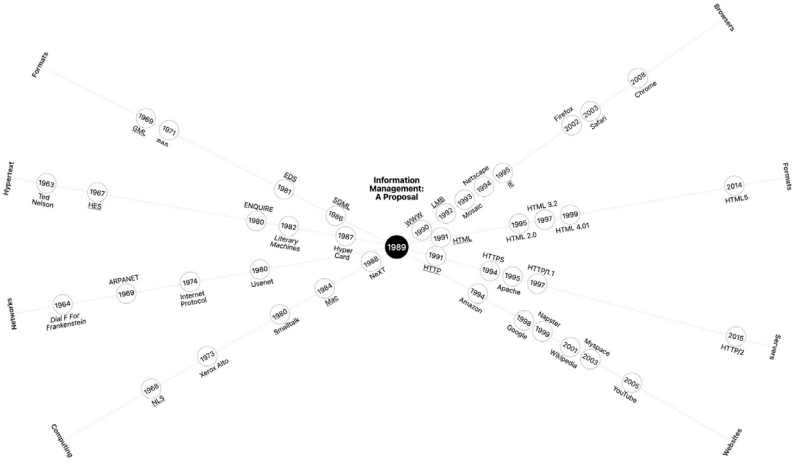
Enquanto trabalhava no CERN, Berners-Lee propôs uma rede de documentos interconectados e outros recursos que poderiam ser acessados e compartilhados por qualquer pessoa com um computador e conexão à Internet. Essa ideia acabou tornando-se a *World Wide Web*. Berners-Lee continuou trabalhando no desenvolvimento da *web* e, em dezembro de 1990, o primeiro navegador da *web* foi desenvolvido, o que permitia aos usuários acessar e navegar na *web* usando uma interface gráfica do usuário.

Berners-Lee é amplamente reconhecido por suas contribuições para o desenvolvimento da *web* e recebeu inúmeros prêmios e homenagens, incluindo o Prêmio Turing, a mais alta honraria em

Ciência da Computação, e o título de Cavaleiro da Rainha Elizabeth II. Tim Berners-Lee também é fundador, em companhia de Rosemary Leith, da *World Wide Web Foundation*[9], organização internacional criada em 2009 para promover a *web* aberta como um bem público e um direito básico, além da acessibilidade, neutralidade da rede, privacidade *on-line*, dados públicos e promoção da inovação local.

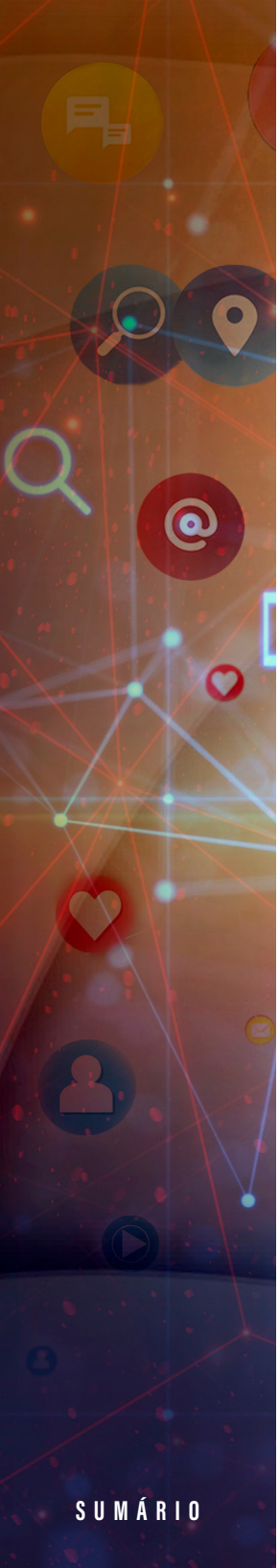
O projeto *World Wide Web* do CERN disponibiliza um *site* com diversas informações sobre a origem da *web*, uma delas é uma linha do tempo interativa, com os principais fatos relacionados, antes e depois de 1989. O gráfico em formato de linha de tempo, lançado em março de 2019, ano em que se celebrou os 30 anos da concepção da *web*, lembra também uma colisão de partículas, uma homenagem ao CERN, o local onde surgiu a *web* e onde fica o maior acelerador de partículas do mundo (fig. 06).

Figura 06 - *Timeline — World Wide Web NeXT Application*



Fonte: https://worldwideweb.cern.ch/timeline/. Acesso em: 10 mar. 2023.

9 https://webfoundation.org/

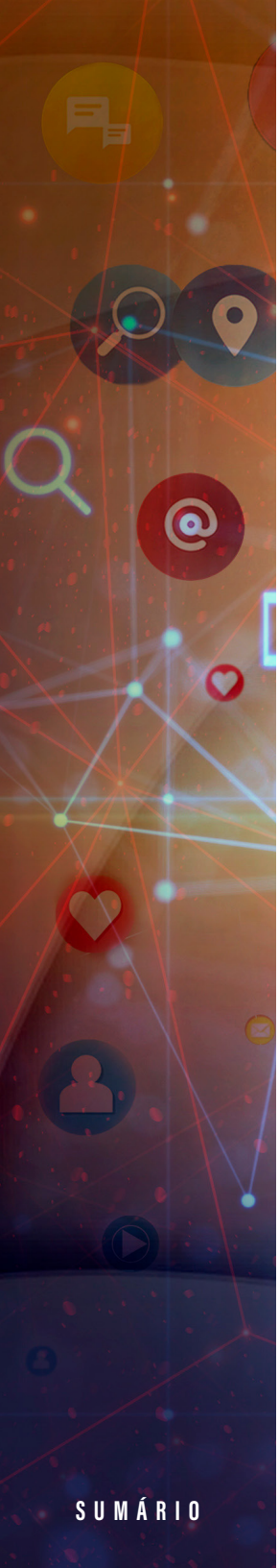


A história do surgimento da *web* no Brasil está diretamente ligada ao desenvolvimento das tecnologias de informação e comunicação em todo o mundo. Em 1988, foi possível observar a formação de redes independentes, conectando grandes universidades e centros de pesquisa do Rio de Janeiro, São Paulo e Porto Alegre aos Estados Unidos (AFONSO, 2000). Por meio de uma parceria entre a Embratel e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), criou-se a Rede Nacional de Ensino e Pesquisa (RNP), uma rede que permitia a conexão entre instituições acadêmicas e de pesquisa no país. A popularização da *web* nos Estados Unidos e em alguns países da Europa deu-se em meados dos anos 90 e em 1995 o Brasil conectou-se à *web*.

Em maio de 1995, os Ministérios das Comunicações e da Ciência e Tecnologia anunciaram a criação do Comitê Gestor da Internet no Brasil. O Comitê seria composto por representantes de vários setores, incluindo provedores de acesso, representantes de usuários e da comunidade acadêmica, com o objetivo de fomentar o desenvolvimento de serviços de Internet no país, coordenar a atribuição de endereços e registros de nomes de domínio e disseminar informações sobre serviços de Internet. Em 2003, o Decreto Federal nº 4.829[10] estabeleceu as normas de funcionamento e atribuições do Comitê Gestor da Internet no Brasil (CGI.br), que posteriormente foram regulamentadas em algumas áreas por portarias interministeriais.

O Núcleo de Informação e Coordenação do Ponto BR (NIC.br) foi criado para implementar decisões e projetos do CGI.br, que coordena a Internet no Brasil, e é responsável pelo registro e manutenção de nomes de domínios .br, distribuição de números de sistema autônomo e endereços IPv4 e IPv6, resposta a incidentes de segurança, projetos de infraestrutura de rede, produção de estatísticas sobre o desenvolvimento da Internet no país, recomendações

10 https://www.planalto.gov.br/ccivil_03/decreto/2003/D4829.htm



de normas e padrões de segurança e viabilização da participação do Brasil no desenvolvimento global da *web*.

Muitas pessoas contribuíram na disseminação da Internet e da *web* nacionalmente, Demi Getschko pode ser considerado uma das personalidades mais importantes no desenvolvimento da Internet no país. Getschko começou sua carreira na Embratel, onde participou da criação da RNP e da primeira conexão internacional do Brasil à Internet. Em 2005, ele se tornou diretor-presidente do NIC.br, organização responsável pelo registro de nomes de domínio e endereços IP no país, além de ter sido membro da diretoria da *Internet Corporation for Assigned Names and Numbers* (ICANN); em abril de 2014 foi homenageado com sua inclusão no *Hall* da fama da Internet[11], promovido pela *Internet Society*[12]. No contexto brasileiro, juntam-se também ao *Hall* da fama da Internet Tadao Takahashi (2017)[13], Michael Stanton (2019)[14], Liane Tarouco (2021)[15] e Carlos Afonso (2021)[16].

A *web*, como uma rede global de redes de computadores interconectados, que permite aos usuários acessar e compartilhar informações usando a Internet, funciona usando uma combinação de tecnologias, incluindo servidores, navegadores e protocolos *web*. Enquanto a Internet pode ser entendida como a infraestrutura que permite a conexão em rede, onde uma das bases é o protocolo *Transmission Control Protocol/Internet Protocol* (TCP/IP), a *web* constitui os recursos que são disponibilizados e serviços que são executados na Internet.

Os servidores da *web* constituem-se em computadores especializados que hospedam páginas da *web* e as disponibilizam

11 https://www.Internethalloffame.org/inductee/demi-getschko/

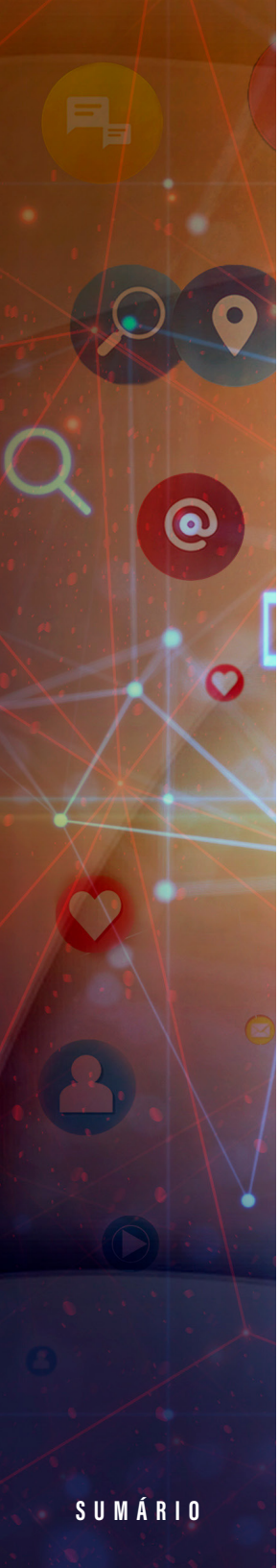
12 https://www.Internetsociety.org/

13 https://www.Internethalloffame.org/inductee/tadao-takahashi/

14 https://www.Internethalloffame.org/inductee/michael-stanton/

15 https://www.Internethalloffame.org/inductee/liane-tarouco/

16 https://www.Internethalloffame.org/inductee/carlos-afonso/



aos usuários na Internet. Quando um usuário insere um endereço da *web* em seu navegador da *web*, o navegador envia uma solicitação ao servidor da *web* para acessar a página da *web* correspondente.

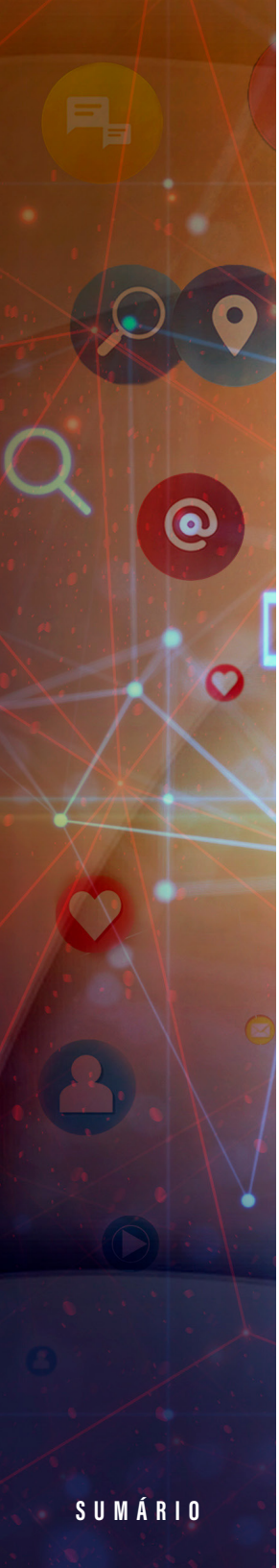
O servidor da *web* envia a página da *web* solicitada de volta ao navegador da *web*, junto a outras informações adicionais, como imagens e outras mídias. O navegador da *web* exibe a página da *web* no dispositivo do usuário, permitindo que ele visualize e interaja com o conteúdo.

Os protocolos da *web*, como o *Hypertext Transfer Protocol* (HTTP) ou Protocolo de Transferência de Hipertexto, são usados para estabelecer a comunicação entre os servidores da *web* e os navegadores da *web*. Esses protocolos definem as regras de como as informações são trocadas e formatadas na *web*.

Os códigos de status das respostas HTTP são utilizados para indicar se uma requisição HTTP foi corretamente concluída. Esses códigos são agrupados em cinco classes: respostas de informação (100-199), respostas de sucesso (200-299), redirecionamentos (300-399), erros do cliente (400-499) e erros do servidor (500-599). As respostas de informação são utilizadas para informar ao cliente que a requisição foi recebida e está sendo processada. As respostas de sucesso indicam que a requisição foi processada com sucesso. Já os redirecionamentos são usados para indicar que o cliente deve tomar alguma ação adicional para completar a requisição. Os erros do cliente indicam que houve um erro na requisição enviada pelo cliente. Por fim, os erros do servidor indicam que houve um problema no servidor que impediu que a requisição fosse processada com sucesso.

No *site* do W3C, o *World Wide Web Consortium* ou *Consórcio World Wide Web*[17], é possível encontrar a descrição de todos os

17 Disponível em: https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html. A lista também é acessível pelo site da *Internet Assigned Numbers Authority* (IANA): https://www.iana.org/assignments/http-status-codes/http-status-codes.xhtml



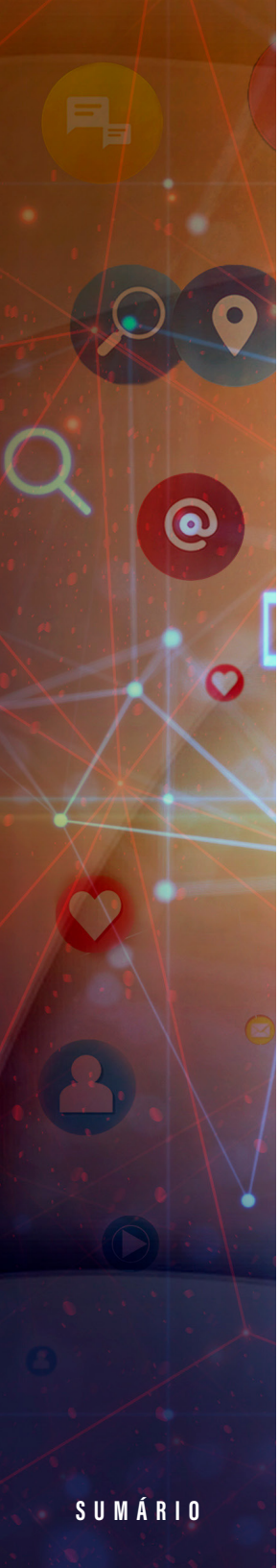
códigos e seus significados. Reproduzimos no Quadro 01 alguns dos principais códigos:

Quadro 01 - Códigos de requisição cliente - servidor

Código	Resposta
200	OK - A solicitação foi bem-sucedida e o servidor retornou os dados solicitados
201	Criado - A solicitação foi bem-sucedida e um novo recurso foi criado
204	Sem conteúdo - A solicitação foi bem-sucedida, mas não há dados para retornar
301	Movido Permanentemente - O recurso solicitado foi movido permanentemente para um novo URL
302	Encontrado/Redirecionamento temporário - O recurso solicitado foi movido temporariamente para uma nova URL
400	*Bad Request* - A solicitação era inválida ou malformada
401	Não autorizado - A solicitação requer autenticação ou o usuário não tem permissão para acessar o recurso solicitado
403	Proibido (*Forbidden*) - O servidor entendeu a solicitação, mas o cliente não tem direitos de acesso ao conteúdo solicitado
404	Não encontrado - O recurso solicitado não foi encontrado no servidor
500	Erro interno do servidor (*Internal Server Error*) - O servidor encontrou uma condição inesperada que o impediu de atender à solicitação

Fonte: W3C (tradução nossa).

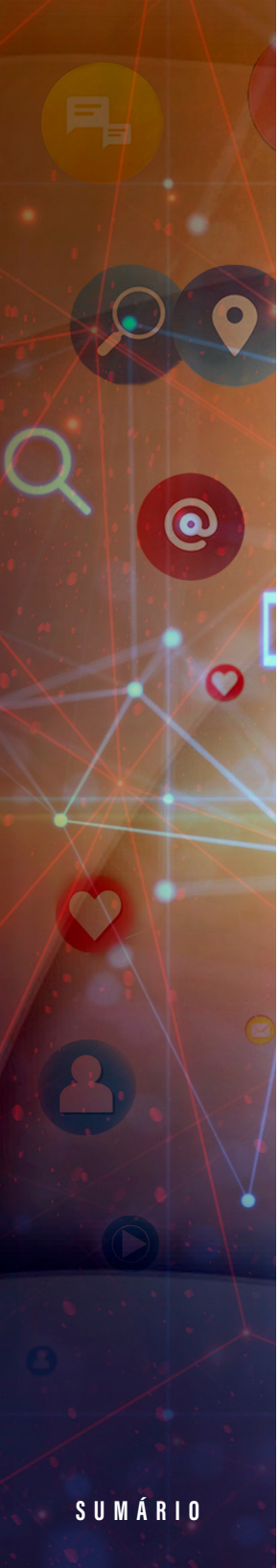
Como vemos acima, diversos códigos são utilizados para representar as respostas às requisições cliente-servidor na *web*, um destes códigos é especialmente importante no contexto da preservação da *web*. O código 404 é um código de *status* retornado por um servidor da *web* quando um cliente, geralmente por meio de um navegador da *web*, solicita um recurso, como uma página da *web*, imagem ou vídeo, que não pode ser encontrado no servidor. Quando um navegador da *web* envia uma solicitação a um servidor da *web* para um determinado recurso, o servidor procurará o recurso e tentará retorná-lo ao



navegador. Se o servidor não conseguir localizar o recurso, ele normalmente retornará um código de *status* 404 junto com uma mensagem de erro ou uma página padrão "404 não encontrado". O código de *status* 404 indica que o recurso solicitado não está disponível no servidor, o que pode acontecer por diversos motivos, como recurso excluído, movido ou renomeado, ou pode haver um problema com a URL inserida. Quando um usuário encontra um erro 404, isso geralmente significa que a página da *web* ou o recurso que ele estava procurando não está disponível e pode ser necessário verificar a URL ou tentar pesquisar o recurso de uma maneira diferente.

Desta forma, a *web* funciona conectando computadores e dispositivos por meio de uma rede de servidores e protocolos. Quando um usuário insere um endereço de *site Uniform Resource Locators* (URL) em seu navegador da *web*, o navegador envia uma solicitação ao servidor da *web* que hospeda o *site*. O servidor da *web* responde enviando de volta o código HTML do *site*, acompanhado de todos os recursos associados, como imagens e vídeos. O navegador da *web* interpreta o código HTML e exibe o *site* na tela do usuário.

O navegador da *web* usa um conjunto de regras conhecidas como protocolos para comunicar-se com o servidor. A *web* é baseada em um conjunto de protocolos padronizados, como o mencionado HTTP (*Hypertext Transfer Protocol*) e HTTPS (*Hypertext Transfer Protocol Secure*), que definem como os navegadores e servidores da *web* se comunicam entre si. Esses protocolos permitem que os usuários acessem e naveguem pela grande quantidade de informações e recursos disponíveis na *web*, configurando a estrutura e o formato das mensagens que são enviadas entre clientes e servidores, permitindo a transferência de uma ampla variedade de conteúdo, incluindo texto, imagens, vídeos e outras mídias. Um navegador da *web* atua como um *gateway* para a Internet, permitindo que os usuários acessem e interajam com os recursos disponíveis *on-line*.



Além disso, a *web* conta com o *Domain Name System* (DNS), sistema hierárquico e distribuído de gestão de nomes para computadores, utilizado para traduzir endereços de *site*s legíveis por humanos em endereços IP (*Internet Protocol address*), rótulos numéricos atribuídos a servidores *web*. Isso permite que os usuários acessem *site*s usando URLs fáceis de lembrar em vez de endereços numéricos complexos, como o uso de domínios .com, .com.br, entre outros.

Depois que o servidor envia as informações solicitadas ao navegador da *web*, o navegador usa um mecanismo de renderização para exibir o conteúdo no dispositivo do usuário. O mecanismo de renderização interpreta a Linguagem de Marcação de Hipertexto (HTML), código usado para estruturar uma página *web* e seu conteúdo; o *Cascading Style Sheets* (CSS), mecanismo para adicionar estilos a uma página *web* aplicada diretamente nas *tags* HTML; e a linguagem de programação *Javascript*, associada ao *site*, e usa essas informações para renderizar a página na tela do usuário. Associado a HTML e CSS, o *Javascript* é uma das três principais tecnologias da *World Wide Web*.

Além de exibir o conteúdo de um *site*, um navegador da *web* também permite que os usuários interajam com o *site*, podendo incluir clicar em *link* para navegar para outras páginas, preencher e enviar formulários e usar vários *links* e complementos para aprimorar a experiência de navegação.

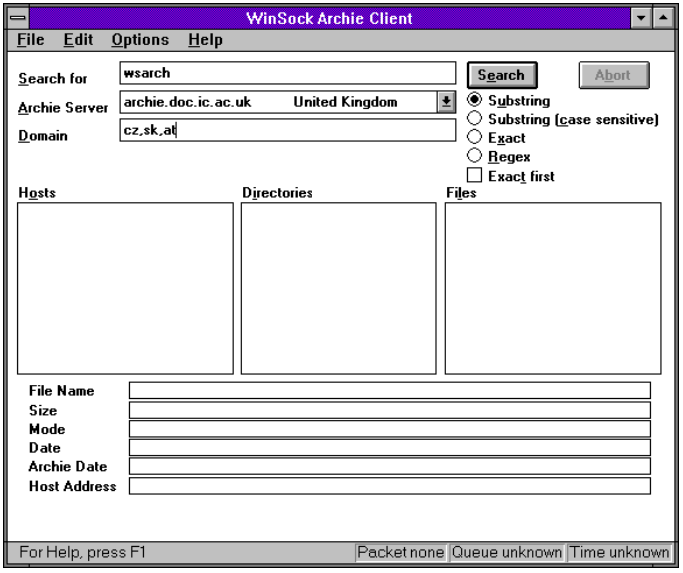
Deste modo, a *World Wide Web* forma uma rede global de documentos interconectados e outros recursos, vinculados por *hiperlinks* e URLs. A estrutura da *web* é descentralizada e distribuída, sem autoridade ou controle central. Esta estrutura também é baseada no modelo cliente-servidor e no uso de URLs, permitindo a distribuição eficiente e escalável de informações na *web*.

A partir do modelo cliente-servidor, os computadores clientes (como computadores *desktop*, *laptops* e dispositivos móveis) acessam e solicitam informações de computadores servidores (como servidores *web*), que hospedam e entregam as informações.

A *web* também se baseia no uso de URLs (*Uniform Resource Locators*), que são os endereços exclusivos que identificam recursos individuais na *web* e permitem que os usuários acessem e naveguem facilmente para diferentes recursos na *web*.

Outro ponto importante que permite a localização de páginas *web* e demais recursos eletrônicos é o uso de motores de busca, como o Google, surgido em 1998 é hoje parte do conglomerado chamado *Alphabet*, ou o *Bing*, lançado em 2009 como iniciativa da *Microsoft.* Porém, um dos primeiros buscadores, ainda que não possa ser considerado um motor de busca da *web*, lançado em 10 de setembro de 1990 foi o *Archie*. Este mecanismo de busca utilizou um sistema de indexação para ajudar os usuários a localizar e baixar os arquivos na Internet, especificamente de servidores FTP públicos. O nome *Archie* foi derivado da palavra "Arquivo" ou *Archive* com a letra "v" removida e é comumente reconhecido como um dos motores de busca inicial da Internet. No entanto, no final da década de 1990, *Archie* (fig. 07) parou de operar.

Figura 07 - O primeiro motor de busca, Archie, de 1995



Fonte: https://www.webdesignmuseum.org/web-design-history/archie-the-first-search-engine-1990

O *Aliweb* (acrônimo para *Archie-Like Indexing for the Web*) foi desenvolvido em 1993 e pode ser considerado um dos primeiros motores de busca da *web*[18]. No entanto, o *Aliweb* não era um mecanismo de pesquisa tradicional da forma que tratamos atualmente, pois ao invés de rastrear a *web* e indexar as páginas automaticamente, o *Aliweb* precisava contar com os usuários e proprietários dos *sites* para enviar manualmente as informações de seus *sites* ao banco de dados do *Aliweb*, o que fez com que relativamente poucas pessoas atualizassem os índices.

O *WebCrawler* pode ser considerado o primeiro motor de busca como conhecemos hoje (fig. 08), possibilitando a busca por texto e ainda se encontra ativo e disponível *on-line*[19].

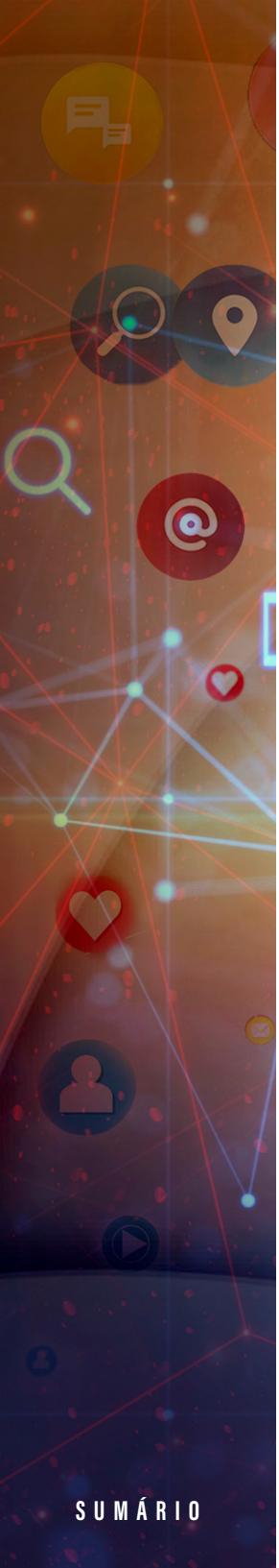
Figura 08 - O primeiro motor de busca da *web, WebCrawler*

Fonte: https://web.archive.org/web/19981212034012/https://www.webcrawler.com/

A *web* continua a evoluir e se expandir, com novas tecnologias e aplicativos sendo desenvolvidos todos os dias, tendo avançado em diversas gerações, *web* 1.0, 2.0. 3.0 e 4.0.

18 https://web.archive.org/web/19970618184044/http://www.nexor.com/public/aliweb/aliweb.html

19 https://www.webcrawler.com/



GERAÇÕES DA *WEB*:
WEB 1.0, 2.0, 3.0, 4.0 E ALÉM

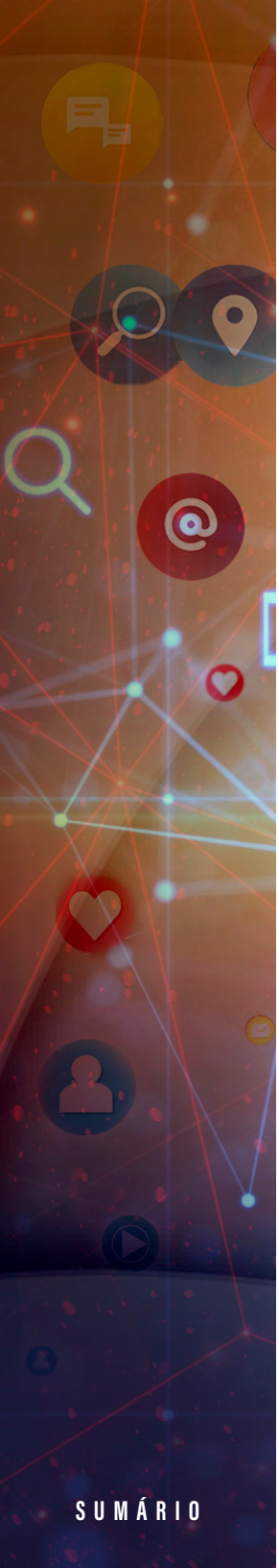
A *web* 1.0 é um termo que se refere à primeira geração da *World Wide Web*, surgiu no início dos anos 1990. A *web* 1.0 caracterizou-se por seu foco na disseminação de informações, com *site*s constituídos principalmente por páginas estáticas de texto e imagens.

As tecnologias e aplicativos da *web* 1.0 incluíam HTML, que era usado para criar e formatar páginas da *web*, e navegadores da *web*, que permitiam aos usuários acessar e visualizar páginas da *web*. A *web* 1.0 também introduziu o uso de URLs para identificar e localizar recursos na *web* e possibilitou um desenvolvimento significativo na história da Internet, pois viabilizou que as pessoas acessassem e compartilhassem informações em escala global. No entanto, era limitado por seu foco na disseminação de informações e não permitia muita interação do usuário ou a criação e compartilhamento de conteúdo gerado pelo usuário.

A *web* 1.0 foi substituída pela *web* 2.0, que introduziu novas tecnologias e aplicativos que permitiram aos usuários criar, compartilhar e colaborar em conteúdo e acessar informações.

A *web* 2.0 é um termo que se refere à segunda geração da *World Wide Web*, que surgiu no final dos anos 1990 e início dos anos 2000. Ao contrário da primeira geração da *web*, que era estática e focada na disseminação de informações, a *web* 2.0 é caracterizada por sua ênfase no conteúdo gerado pelo usuário, na colaboração e na capacidade de acessar e compartilhar informações de qualquer dispositivo com conexão à Internet.

Também conhecida como *web* participativa ou social, a *web* 2.0 revolucionou a tecnologia na internet, dando aos usuários a liberdade de controlar seus dados e enriquecer sua experiência, referin-

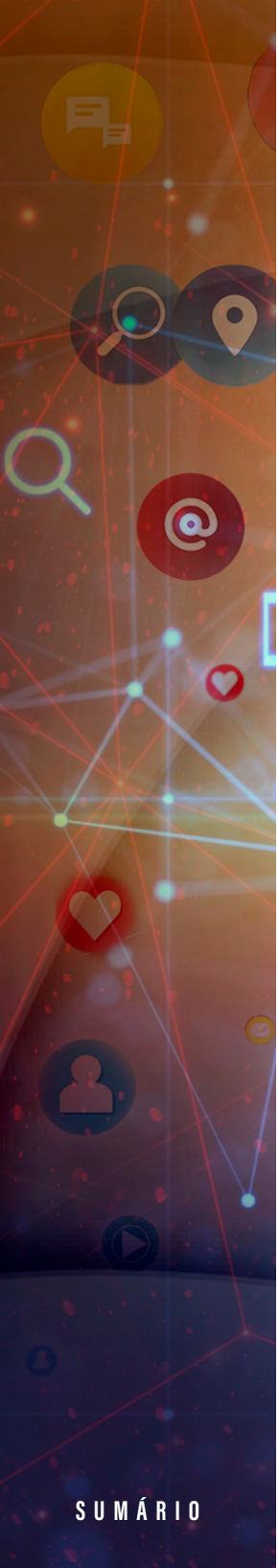


do-se a *site*s que valorizam o conteúdo criado pelo usuário, a facilidade de uso e a interação para o usuário final. O termo começou a ser usado em 1999 e popularizou-se com Tim O'Reilly (2007). Isto não significou uma alteração formal na natureza da *web*, mas sim uma mudança geral de *site*s estáticos para plataformas interativas e colaborativas. Exemplos desses *site*s incluem plataformas de mídia social como *Facebook*, *Instagram*, *blogs*, *wikis*, sistemas de marcação, *site*s de compartilhamento de vídeo e imagem, serviços hospedados, aplicativos *web* e plataformas de consumo colaborativo. Esses *site*s permitem que os usuários participem e colaborem ativamente uns com os outros por meio de diálogos de mídia social e conteúdo gerado.

A *web* 2.0 teve um impacto significativo na forma como as pessoas usam e interagem com a *web* e levou ao desenvolvimento de novos setores e modelos de negócios, como *marketing* de mídia social e possibilidades de colaboração *on-line*. Também, facilitou o surgimento da economia compartilhada, na qual indivíduos e organizações podem compartilhar recursos e serviços pela Internet. Reduziu o custo de entrega de informações e forneceu uma plataforma para participação por meio de colaboração, comunicação, comunidades pessoais e conectividade entre aplicativos.

Web 3.0 é um termo criado por Tim Berners Lee, se refere à terceira geração da *World Wide Web*, combina a *web* semântica, aplicativos da *web* 2.0 e inteligência artificial (IA). A *web* 3.0 pode ser caracterizada por seu foco na IA e no uso de algoritmos de aprendizado de máquina para permitir que a *web* entenda e interprete o significado das informações disponíveis nela, como mecanismos de pesquisa inteligentes que podem entender e interpretar as consultas dos usuários e sistemas de recomendação personalizados que podem fornecer aos usuários conteúdo individualizado com base em seus interesses e preferências.

O poder da *web* 3.0 reside na ligação de dados, tornando-os uma teia de dados, não apenas de máquinas, para criar uma teia de

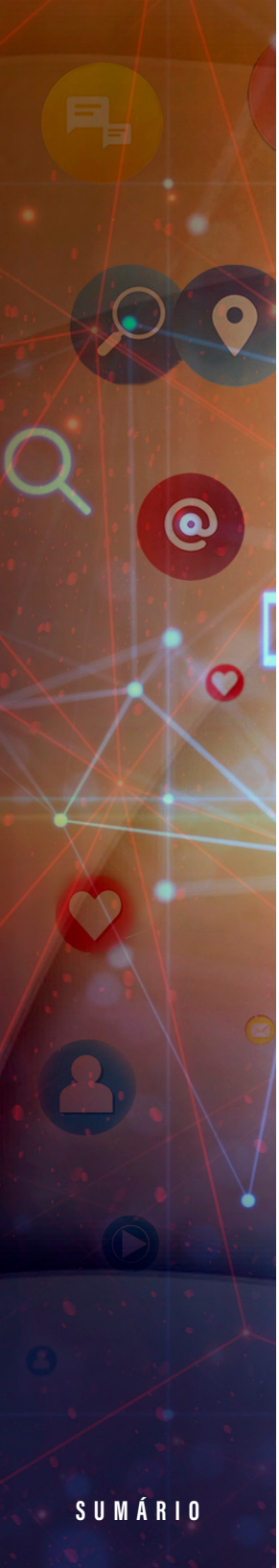


significados (semântica) ao invés da teia de *links* como nas versões anteriores. A *web* semântica baseada na integração de dados permite que o conteúdo da *web* seja exibido não apenas no formato de linguagem humana, mas também em um formato que pode ser lido por computador. Acontece quando as máquinas podem ler páginas da *web*, como nós, humanos, onde os mecanismos de pesquisa e *software* podem auxiliar a rede a encontrar o que estamos procurando. É a capacidade dos computadores de pesquisar informações facilmente usando inteligência artificial.

A *web* 3.0 utiliza *Resource Description Framework* (RDF) para descrever os recursos, diferente da linguagem de marcação XML e hipertexto HTML usadas na *web* 2.0 e *web* 1.0 respectivamente. O RDF permite atualizar bancos de dados automaticamente quando houver mudanças nos recursos de informação que os constituem e também, permite unificar as informações de diferentes fontes e formatos. Ainda, pode manter perfis *web* de cada usuário individual com base em seu histórico de navegação e usar os detalhes para personalizar a *web* de cada indivíduo baseado na sua experiência. Isso significa que se dois indivíduos realizaram pesquisas semelhantes na Internet usando ferramentas semelhantes, os resultados seriam diferentes, determinados por seus perfis.

As tecnologias e aplicações da *web* 3.0, com o uso de tecnologias da *web* semântica, permitirão a representação de dados e informações de forma que possam ser compreendidas pelas máquinas, de forma que a *web* entenda o significado e o contexto das informações disponíveis e forneça experiências mais relevantes e personalizadas para os usuários.

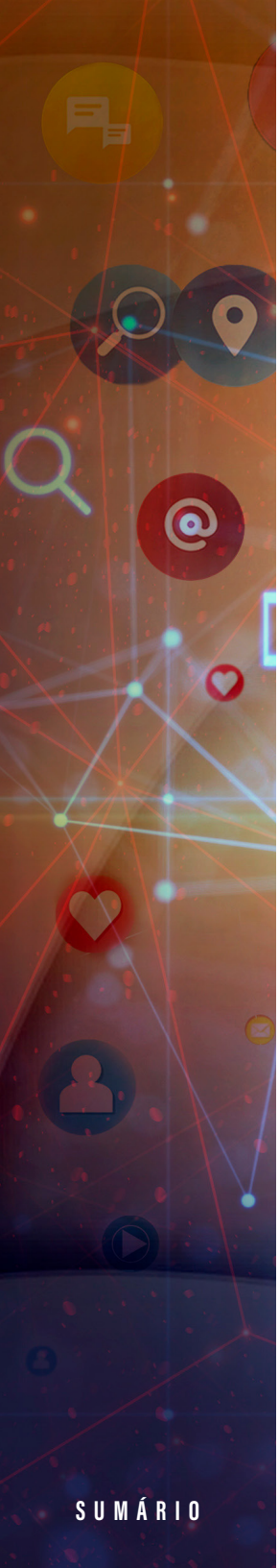
Se espera que a próxima evolução da Internet, a partir da *web* 3.0, resolva problemas anteriormente identificados na *web* 2.0 (KSHETRI, 2022). Acredita-se que a utilização de tecnologias descentralizadas, como *blockchain*, finanças descentralizadas, moedas digitais e *Non-fungible Token* (NFT), irá substituir as redes



sociais centralizadas e criar mais espaços abertos na *web* 3.0. O objetivo é que os indivíduos recuperem o controle sobre a Internet, suas informações pessoais e sua privacidade das grandes corporações e entidades governamentais.

Assim, chegamos a *web* 4.0 que se caracteriza por três condições que a constituem: onipresença, identidade e conexão. A onipresença ou ubiquidade é a disponibilidade a qualquer tempo e em qualquer lugar, onde a linha entre a vida *off-line* e *on-line* torna-se borrada e difusa. A identidade pressupõe a existência de protocolos para determinar, eficientemente, quem são os usuários, o que eles estão fazendo e que tipo de coisas precisam, possibilitando o fornecimento de serviços personalizados. Por sua vez, a conexão significa manter uma rede de usuários continuamente conectados. Ao contrário das versões da *web* no passado, onde os usuários vagavam em um mar de informações, a *web* 4.0 fornece apenas informações adequadas para os usuários, integrando e conectando todos os dados conhecidos sobre a identidade dos usuários. (NOH, 2015).

Considerada a *web* inteligente, será capaz de fazer ilações, usando a inteligência artificial para tomar decisões, usando inferência e conteúdo anteriormente pesquisado. Fará a interação entre usuários humanos e componentes da máquina assemelhando-se a uma relação simbiótica. A *web* 4.0 foi descrita como um tipo de *software* funcionando entre sistemas operacionais e o aplicativo utilizado em uma enorme teia de interações altamente inteligentes, parecidas com o cérebro humano. Na *web* 4.0 temos o surgimento da Internet das coisas, os dados são coletados e transmitidos para a "nuvem", onde são processados para que rapidamente o usuário obtenha o resultado esperado, por conta própria e em seu dispositivo. Ao mesmo tempo, o *big data*, caracterizado por grandes quantidades de dados que não podem ser coletados, armazenados, gerenciados ou analisados por dispositivos comuns, necessita de técnicas e tecnologias avançadas para coleta, armazenamento, distribuição, gerenciamento e análise. (NAZAROVETS; KULYK, 2017).



Por fim, voltamos às origens da *web*, pois um *startup* chamado *Inrupt*[20], fundada por Tim Berners-Lee, o criador da *World Wide Web*, pretende restaurar o controle dos dados e da *web* para os usuários. O *Inrupt* baseou-se no projeto *Solid*[21] de Berners-Lee, que permitiu que os usuários armazenassem seus dados em "*pods*" descentralizados e controlassem quem teria acesso a eles. A missão de descentralizar a *web* foi impulsionada por preocupações com a privacidade e segurança da Internet, à medida que os usuários se tornam cada vez mais expostos e conscientes dos riscos associados aos servidores centralizados e à exploração de dados pessoais. Desta forma, o objetivo é tornar a *web* mais democrática e equitativa, onde os indivíduos tenham maior controle sobre seus próprios dados e sejam menos vulneráveis aos governos ou corporações, tal como foi originalmente pensada por Berners-Lee.

O ARQUIVAMENTO DA *WEB*, DEFINIÇÕES E CONCEITOS

O arquivamento da *web* consiste em um processo que compreende coletar, armazenar e disponibilizar a informação retrospectiva da *World Wide Web* para futuros pesquisadores. Este processo envolve iniciativas no mundo inteiro, algumas com abordagens globais, outras localizadas geograficamente, com foco em seus respectivos países, atributo identificado pelo domínio do endereço eletrônico ou a partir da verificação do produtor da informação e o contexto no qual se insere.

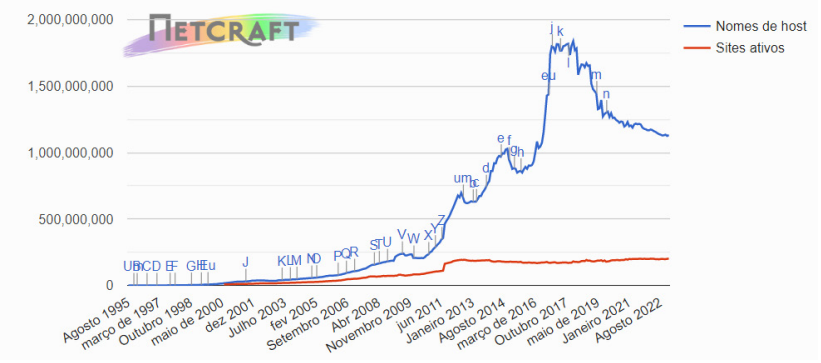
20 https://www.inrupt.com/

21 https://www.inrupt.com/solid

De acordo com os dados dos relatórios publicados pela empresa *Netcraft*[22], obtidos a partir de *surveys*, em fevereiro de 2021 obtiveram respostas de 1.204.252.411 *site*s em 263.042.054 domínios únicos e 10.766.606 de servidores *web*. Em dezembro de 2022 receberam respostas de 1.125.374.532 *site*s em 271.238.722 domínios únicos, de 12.234.425 servidores *web*. Isto reflete uma perda de 9,7 milhões de *site*s, 450.421 domínios únicos, e 72.200 servidores *web*. Na *survey* realizada em janeiro em 2023 receberam respostas de 1.132.268.801 *site*s, de 270.967.923 domínios únicos e de 12.156.700 servidores *web*, refletindo um ganho de 6.894.269 *site*s, mas uma perda de 270.799 domínios e 77.725 servidores *web*. Estes dados levaram a refletir sobre as razões para a diminuição do número de *site*s, domínios únicos e servidores *web*.

Os motivos para justificar esses números podem ser muitos, mas entre eles destacam-se principalmente os *site*s que são bloqueados pela maioria dos navegadores da *web* devido ao uso de protocolos desatualizados e vulnerabilidades para ataques e, ainda, *site*s atualizados, reestruturados e muitas vezes excluído assim, as informações são perdidas ou ficam inacessíveis. No Grafico 1 podemos verificar a quantidade de *site*s da *web* ativos e os domínios, no decorrer dos anos, em escala linear.

Gráfico 1 - *Sites* ativos e domínios, entre 1995 e 2022

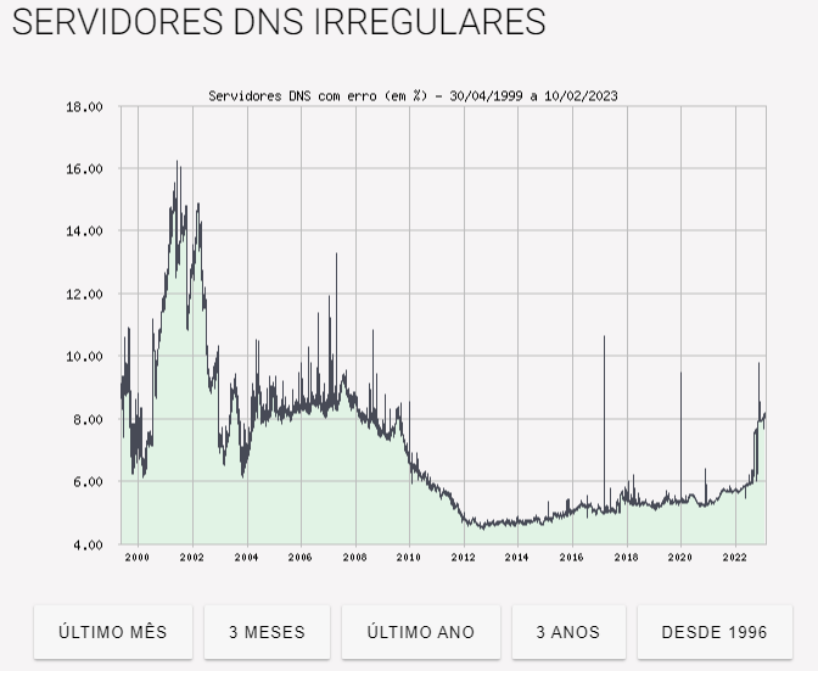


Fonte: Disponível em: https://news.netcraft.com/. Acesso em: 01 fev. 2023.

22 https://news.netcraft.com/

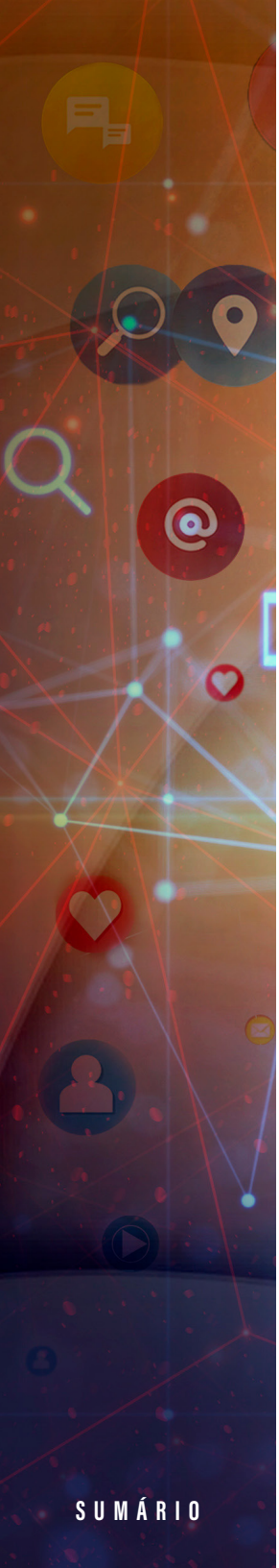
No Brasil a situação não é diferente, no Gráfico 2 verificamos a quantidade de Sistema de Nomes de Domínios (DNS) com algum tipo de dificuldade de acesso que pode ocorrer devido a conexão recusada, erro de DNS, domínio desconhecido, DNS desconhecido, pesquisa recusada, falha de DNS, entre outros.

Gráfico 02 - Percentual de servidores DNS com erros no período de 3 anos



Fonte: Disponível em: https://registro.br/dominio/estatisticas/. Acesso em: 11 fev. 2023.

A fim de preservar o conteúdo da *web* que pode representar grande interesse para as futuras gerações de internautas, diversas iniciativas de arquivamento da *web* em grande escala foram lançadas por várias instituições interessadas em preservar os recursos da *web*. O arquivamento da *web* compreende os processos de captura da *web*, preservando os conteúdos em formato de arquivo, possibilitando seu acesso e uso. Para atingir esse objetivo, utilizam-se



ferramentas, padrões, e tecnologias que permitem o gerenciamento dos arquivos da *web*.

A preservação da *web* ainda pode ser entendida como o rastreamento, gerenciamento e preservação de *site*s e recursos da *web*. Trata-se de "[...] qualquer forma de preservação deliberada e intencional de material da *web*" (BRÜGGER, 2011, p. 25, tradução nossa). A preservação da *web* deve ser uma atividade realizada do começo ao fim, o que significa da captura ao acesso, e abrangendo todo o ciclo de vida do recurso *web*.

De forma mais ampla, a preservação digital envolve, conforme Farrel, Aschley e Davis (2010), um conjunto de processos e atividades que garante o armazenamento, o acesso e a interpretação de informações digitais a longo prazo. Segundo os autores algumas considerações são importantes em relação à preservação:

1. para preservar, os recursos devem ser geridos de forma eficaz;

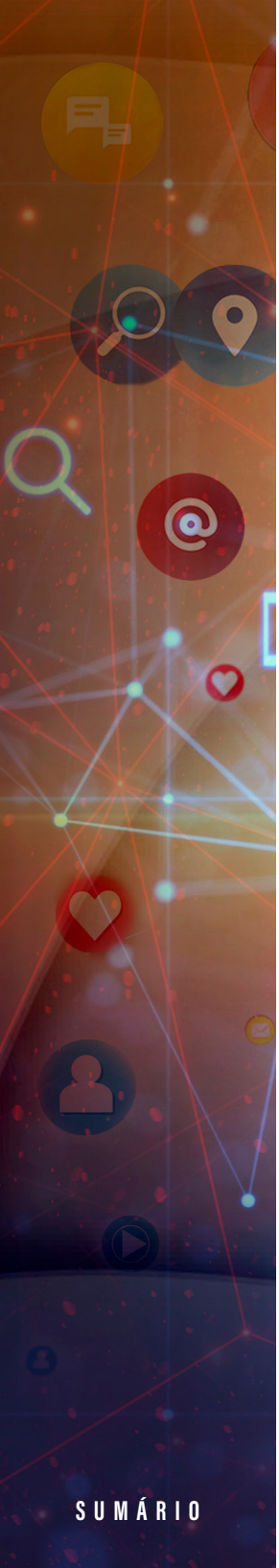
2. nem todos os recursos da *web* precisam ser preservados; uma abordagem seletiva é recomendada;

3. nem sempre é necessário preservar todas as versões de todos os recursos;

4. a preservação permanente, conforme definida pelo modelo *Open Archival Information System* (OAIS), não é a única opção, a proteção de curto prazo contra perdas ou danos também é uma forma válida de preservação e

5. os esforços de preservação não precisam resultar em uma solução "perfeita".

A preservação digital é o conjunto coordenado e contínuo de processos e atividades que garantem o armazenamento de longo prazo e sem erros da informação digital, com meios para recuperação e interpretação, durante todo o período de tempo em que a



informação é necessária. Atualmente, temos um cenário em que grandes quantidades de informações digitais são perdidas ou tornadas permanentemente irrecuperáveis, uma preocupação já demonstrada por um dos criadores do protocolo TCP/IP e considerado um dos pais da Internet, Vint Cerf[23], e por diversos autores desde os anos 1990 (KUNY, 1997; BRAND, 1999; PANOS, 2003), a chamada "*Digital Dark Age*" ou idade das trevas digital. Nesse contexto, podemos considerar duas razões críticas para desenvolver e implementar práticas de preservação digital: a) deterioração física das mídias e b) obsolescência tecnológica de *hardware* e *software*.

Uma das questões mais importantes para os estudiosos da Internet no futuro, de acordo com Brügger (2011), é possibilitar um contato com a Internet do passado e a Internet do presente. O autor coloca que há duas razões para isto: o arquivamento da *web* permite escrever a história da *web*, que é uma condição necessária para a compreensão da Internet do presente, bem como de novas formas emergentes de Internet. Além disso, o arquivamento da *web* nos proporciona documentar nossas descobertas quando estudamos a *web* de hoje, uma vez que, na prática, a maioria das pesquisas preserva conteúdos digitais de forma a ter um objeto para estudar e usar como referência. Portanto, as questões relacionadas ao arquivamento da *web* não são importantes apenas para os historiadores, também são relevantes para qualquer estudioso da *web* que pesquise, por exemplo, comunidades *on-line*, jogos *on-line*, *site*s de notícias, entre outros assuntos.

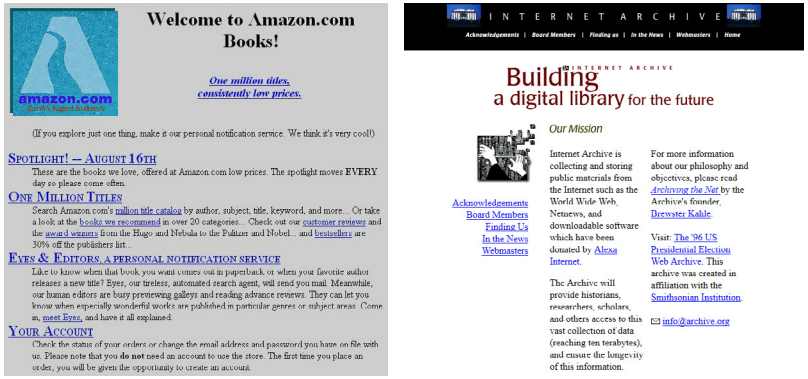
Podemos distinguir o arquivamento da *web* de duas formas: o micro e o macroarquivamento, independente do propósito e do tipo de material arquivado. O microarquivamento significa arquivamento realizado em pequena escala por indivíduos ou pequenos grupos, com base em uma necessidade imediata. Em contraste, o macroarquivamento é realizado em larga escala por instituições com profissionais

23 https://www.bbc.com/news/science-environment-31450389

e conhecimentos técnicos especializados com a finalidade de arquivar, por exemplo, o patrimônio digital em geral. O microarquivamento da *web* parte das necessidades de indivíduos, famílias, organizações e instituições em preservar o que tinham criado ou o que encontravam disponível, em sua maioria eram salvos arquivos HTML ou capturas da tela. Nessa modalidade, o ato de preservar é geralmente aleatório e baseado nas necessidades do momento, e não vem acompanhada de reflexões sobre o que já foi realizado em termos de arquivamento ou considerado como parte de um esforço maior e sistemático com o objetivo de preservar o patrimônio cultural.

Com a possibilidade de preservar páginas *web*, torna-se possível uma característica muito importante do arquivamento da *web*, que é o versionamento dos *website*s, isto é, navegar por diferentes versões de um mesmo *website*. Na imagem abaixo podemos identificar versões antigas de quatro grandes *site*s da Internet.

Figura 09 - Versões antigas da Amazon (1995), Internet Archive (1997), Google (1998) e Facebook (2004)

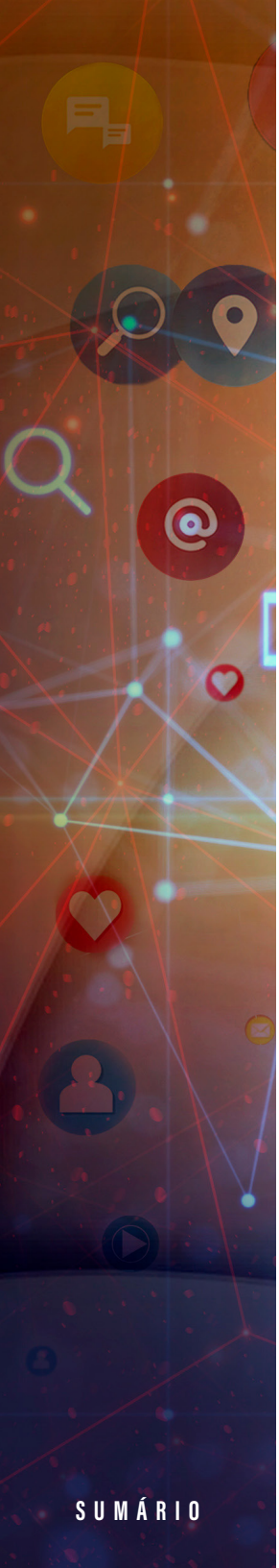


Fonte: https://archive.org/web/

Em outros casos, muitos *website*s e plataformas, após serem descontinuados, só sobrevivem nos arquivos da *web*. O *GeoCities* foi um serviço popular de hospedagem gratuita que permitia aos usuários criar *site*s pessoais na década de 1990 e ganhou popularidade à medida que os usuários criaram *site*s pessoais sobre si mesmos, seus *hobbies* e vidas, foi uma das primeiras plataformas de hospedagem de "conteúdo gerado pelo usuário". No entanto, depois que foi comprada pelo *Yahoo!*, em 1999, e mudou suas especificações de serviço, o número de usuários começou a cair e acabou sendo fechado em 2009. Como afirma Milligan (2017), a principal consequência foi que todos os arquivos de usuários do *GeoCities* desapareceram com a maioria dos *site*s e conteúdo multimídia sendo perdidos de forma irreversível. Porém, é possível recuperar parte do conteúdo por meio dos *site*s preservados. Segundo Milligan (2019), a partir destes exemplos de arquivos da *web* como o *GeoCities*, podemos encontrar histórias do passado e vozes de pessoas comuns que tinham um *site* no final dos anos 1990, e dessa forma puderam reconstruir o passado e compreender seu contexto.

O W3C E O IIPC

O *World Wide Web Consortium* (W3C) é uma comunidade internacional de organizações e indivíduos que trabalha para

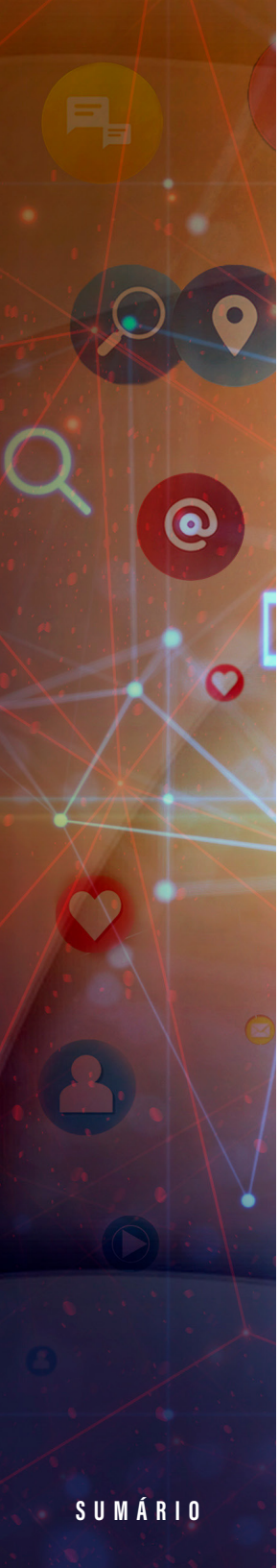


desenvolver, manter e promover os padrões da *web*. O W3C foi fundado em 1994 em Massachusetts, EUA, por Tim Berners-Lee, o inventor da *World Wide Web*. O W3C é uma organização que desempenha um papel fundamental no desenvolvimento e evolução da *World Wide Web*. Ao desenvolver e manter padrões da *web*, o W3C ajuda a garantir que esta seja uma plataforma consistente e interoperável para informações, comunicações, comércio e governo. O W3C é uma organização sem fins lucrativos e seus membros[24] incluem uma ampla gama de organizações e indivíduos de todo o mundo, incluindo desenvolvedores da *web*, pesquisadores, educadores e agências governamentais.

A missão do W3C é promover a *web* como plataforma informacional e comunicacional. Para cumprir esta missão, o W3C desenvolve e mantém padrões *web*, que são especificações técnicas que definem como os diferentes componentes da *web* devem funcionar. Esses padrões abrangem uma ampla gama de tópicos, incluindo HTML e CSS, usados para estruturar e estilizar o conteúdo da *web*, padrões XML e RDF para intercâmbio de dados e HTTP e HTTPS para comunicação na *web*. O W3C também desenvolve e promove padrões de acessibilidade na *web*, com o objetivo de garantir que a *web* seja acessível para pessoas com deficiência. O W3C fornece uma variedade de recursos e serviços para apoiar o desenvolvimento e a implementação de padrões da *web*. Esses recursos e serviços incluem materiais educacionais, ferramentas de *software* e fóruns para discussão e colaboração.

Além de desenvolver e manter padrões da *web*, o W3C também se envolve em atividades de pesquisa e desenvolvimento para promover o estado da arte em tecnologia da *web*. Isso inclui o trabalho em tecnologias emergentes, como a *web* semântica, que visa tornar a *web* mais orientada a dados e legível por máquina e a *Web*

24 https://netpreserve.org/about-us/members



das Coisas (*Web* of Things - WoT)[25], que visa estender a *web* ao reino dos objetos físicos.

Além de desenvolver e manter padrões da *web*, o W3C fornece recursos e ferramentas para desenvolvedores e outros membros da comunidade da *web*, incluindo materiais educacionais, ferramentas de *software* e fóruns para discussão e colaboração. O W3C é uma parte vital da comunidade da *web*, trabalhando para desenvolver e promover a *web* e fornecer recursos e suporte.

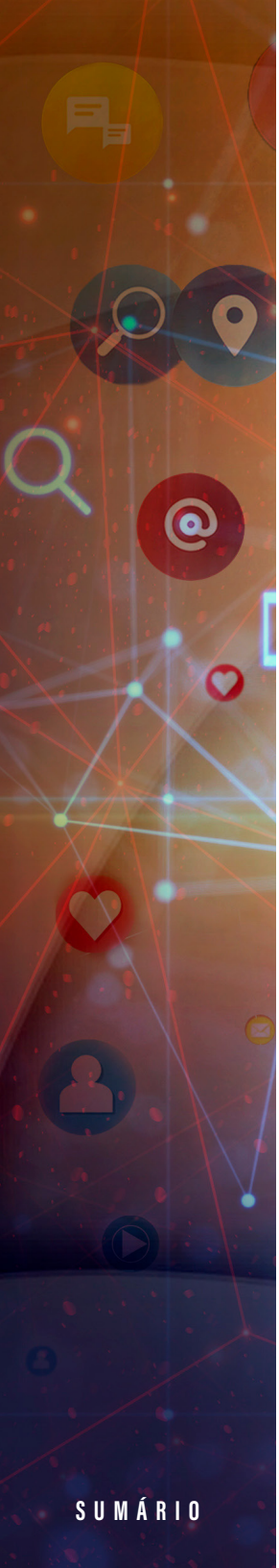
O *International Internet Preservation Consortium* (IIPC)[26] é uma organização internacional, sem fins lucrativos, que trabalha para preservar e fornecer acesso a arquivos da *web*. O IIPC foi fundado em 2003 e é composto por arquivos e bibliotecas nacionais, além de outras organizações de todo o mundo interessadas em arquivamento e preservação da *web*, o trabalho do IIPC é apoiado pelos seus membros e por subvenções de diversas organizações.

O IIPC trabalha para promover e apoiar o desenvolvimento de práticas e tecnologias de arquivamento e preservação da *web*, incluindo *site*s, documentos digitais e outros recursos *on-line* de importância cultural, histórica ou artística. Isso abrange fornecer educação e treinamento em arquivamento da *web*, desenvolver e promover padrões e melhores práticas para arquivamento da *web* e compartilhar informações e recursos entre seus membros.

O IIPC é uma parte importante da comunidade de arquivamento e preservação da *web*, trabalhando para promover e apoiar o desenvolvimento de práticas e tecnologias de arquivamento, preservação digital e acesso a vários arquivos da *web*. Uma das principais atividades do IIPC é a organização do *Web Archiving Conference (WAC)*, um evento anual que reúne arquivistas da *web* e outros

25 https://www.w3.org/WoT/

26 https://netpreserve.org/



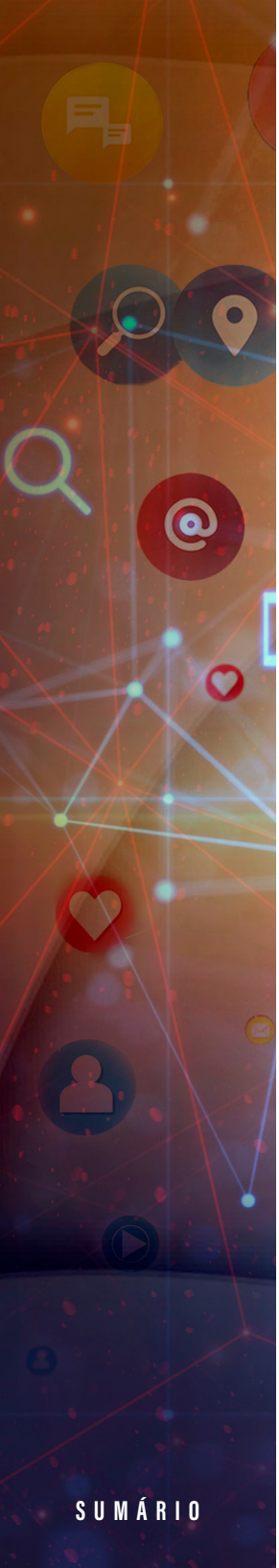
especialistas para discutir as melhores práticas de arquivamento da *web* e compartilhar experiências e conhecimentos. O IIPC também produz publicações e outros recursos para arquivistas da *web* e mantém vários projetos relacionados à preservação digital.

Nas conferências de arquivamento da *web* reúnem-se especialistas na área, estes eventos são normalmente organizados por instituições ou grupos que se dedicam ao arquivamento da *web*, como o *International Internet Preservation Consortium* (IIPC). A *Web Archiving Conference*[27] é uma conferência anual que reúne arquivistas da *web* e outros especialistas na área de arquivamento da *web*. A conferência é organizada pelo IIPC e normalmente é realizada em locais diferentes ao redor do mundo a cada vez que é realizada. Outro evento é a conferência *Research Infrastructure for the Study of Archived Web Materials* (RESAW)[28], com foco em promover infraestrutura de pesquisa europeia colaborativa para o estudo de materiais arquivados na *web*. As conferências de arquivamento da *web* geralmente incluem uma variedade de apresentações, *workshops* e discussões sobre tópicos relacionados à preservação digital, incluindo apresentações sobre os últimos desenvolvimentos em tecnologias e métodos de arquivamento da *web*, discussões sobre as melhores práticas para preservar e fornecer acesso a materiais de patrimônio cultural baseados na *web* e *workshops* sobre o uso de ferramentas e *softwares* específicos.

Além de apresentações e *workshops*, as conferências de arquivamento da *web* também podem incluir oportunidades de *networking*, fornecendo um fórum para compartilhar conhecimentos e experiências, aprender sobre novos desenvolvimentos em arquivamento da *web* e estabelecer contatos com outros profissionais da área.

27 https://netpreserve.org/general-assembly/

28 http://resaw.eu/events/



Outras organizações que trabalham com arquivamento da *web* e possuem recursos úteis também incluem materiais produzidos pela *Digital Preservation Coalition* (DPC)[29], pela *National Digital Stewardship Alliance* (NDSA)[30] e pela seção de arquivamento da *web* da *Society of American Archivists is North America's* (SAA)[31]

O ARQUIVAMENTO DA *WEB* NO BRASIL

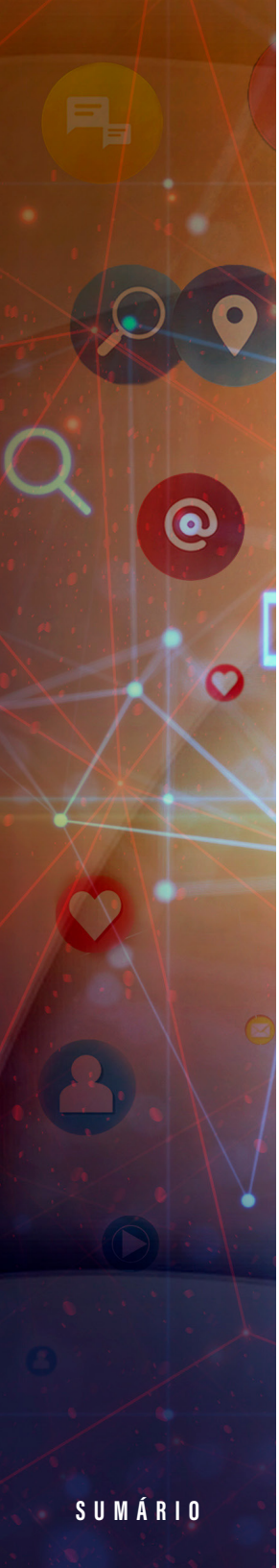
Pelas características que a *web* possui em relação ao uso de URL (*Uniform Resource Locator* ou Localizador Uniforme de Recursos) e uso de domínio para acessar uma página *web*, torna-se difícil identificar a região região à qual pertence um determinado *site*. Rogers (2016), quando menciona o uso dos arquivos da web em pesquisas utilizando Métodos Digitais, se pergunta: como seria demarcada uma *web* nacional? Naturalmente, os sites não são só demarcados pelo domínio de origem, mas também pelos conteúdos que abordam. Globalmente, os sete domínios de alto nível originais foram os .com, .org, .net, .int, .edu, .gov, .mil. Um *site* com o domínio .com.br, um dos mais comuns a serem registrados no Brasil, normalmente incluirá conteúdos brasileiros, pois o registro é realizado pelo Registro.br, que é um departamento do NIC.br (Núcleo de Informação e Coordenação do Ponto BR).

Assim como o .com.br, diversas outras categorias de domínios .br são oferecidas pelo Registro.br, tais como genéricos, onde inclui-se o .com.br (Atividades Comerciais), mas também outros exemplos, como o .net.br (Atividades Comerciais), ong.br (Atividades não governamentais individuais ou associativas), eco.br (Atividades com

29 https://www.dpconline.org/handbook/content-specific-preservation/web-archiving

30 https://ndsa.org/publications/

31 https://www2.archivists.org/groups/web-archiving-section



foco eco-ambiental) e o log.br (Transportes e Logística). Há também a possibilidade de outras categorias, como de pessoas físicas: blog.br (*web logs*), *wiki*.br (Páginas do tipo '*wiki*'); profissionais liberais: adv.br (Advogados), bib.br (Bibliotecários), med.br (Médicos), pro.br (Professores); pessoas jurídicas: inf.br (Meios de informação como rádios, jornais, bibliotecas), imb.br (Imobiliárias), ind.br (Indústrias); Cidades: floripa.br, jampa.br, niteroi.br, poa.br, salvador.br. O Registro.br disponibiliza uma página para uma lista completa e atualizada[32].

Domínios ainda podem ser classificados como extensões de topo genéricas, tais como .info, .com e extensões de topo específicas de cada país, como .br (Brasil), .ar (Argentina), .mx (México), entre outros.

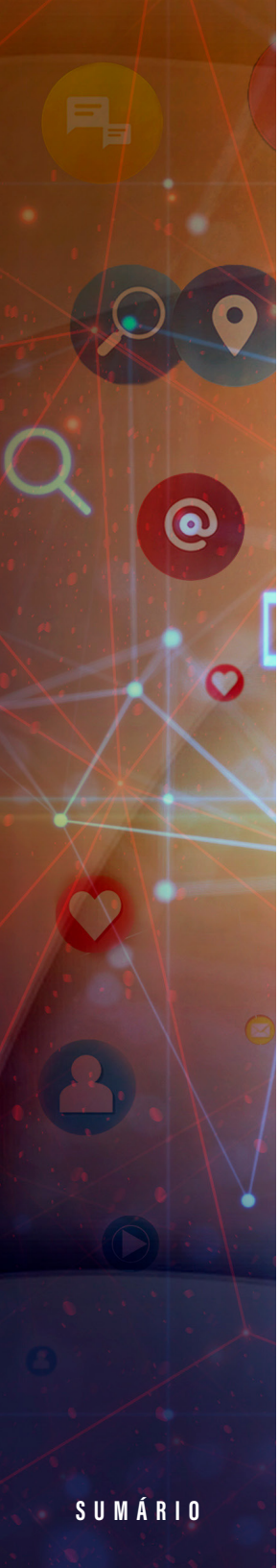
Além disso, é possível registrar domínios internacionais e hospedar conteúdo brasileiros, como o .com, .net, .org — estes considerados domínios de alto nível, *Top Level Domains* (TLD). Portanto, isso significa que o domínio da Internet não define a abrangência de conteúdo de um país, pois estes podem estar localizados em diferentes registros.

Rockembach e Pavão (2018) destacam a necessidade de desenvolver políticas de seleção com foco específico, tais como institucional, temático ou por domínio, e que acompanhado da administração do ciclo de vida do arquivamento da *web* e da adoção de tecnologias de código aberto, será possível não apenas salvaguardar a memória digital brasileira, mas também contribuir com a comunidade de arquivamento da *web* fora do Brasil.

Em 2017, foi fundado o Núcleo de Pesquisa em Arquivamento da W*eb* e Preservação Digital (NUA*WEB*)[33] na Universidade Federal do Rio Grande do Sul (UFRGS), com o propósito de investigar as características do arquivamento da *web*, analisando iniciativas

32 https://registro.br/dominio/categorias/

33 https://www.ufrgs.br/nuaweb/



nacionais e internacionais, bem como as políticas e tecnologias envolvidas no processo. O Núcleo aborda aspectos relacionados à preservação, uso e acesso ao longo do tempo de objetos digitais presentes na *web*, incluindo *website*s, áudio, imagens, vídeos, bancos de dados, dados de redes sociais e outros, como uma memória digital. O Núcleo conta com contribuições de várias áreas, como Arquivologia, Biblioteconomia, Ciência da Informação, Comunicação e Ciência da Computação. As atividades de preservação da *web* no Brasil realizadas pelo Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital são reconhecidas internacionalmente desde 2019, quando da participação na Conferência Internacional de Arquivamento da Web realizada em Zagreb, na Croácia (ROCKEMBACH; MELO, 2019), e pela participação constante nas conferências posteriores, em 2021 no formato *on-line* (ROCKEMBACH, 2021b) e 2023, em Hilversum, na Holanda (MELO; ROCKEMBACH, 2023).

A iniciativa arquivo.ong.br (fig. 10) destina-se a reunir uma comunidade de práticas e disseminação de ferramentas úteis na gestão de preservação de arquivos digitais, incluindo os arquivos da *web*. A plataforma captura e disponibiliza arquivos da *web* desde 2018, contendo *sites* das eleições presidenciais brasileiras de 2018, além de projetos sobre conteúdos da Covid, mudanças climáticas e arquivamento da *web* institucional. Vinculada ao Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital da Universidade Federal do Rio Grande do Sul, a iniciativa ainda envolve o fomento de discussões acerca de arquivos da *web* para a realidade brasileira, e também foi apresentada no Congresso Internacional de Arquivamento da *web* (*Web Archiving Conference* - WAC), promovido pelo *International Internet Preservation Consortium* (IIPC), realizado em 2019 na Croácia[34].

34 https://digital.library.unt.edu/ark:/67531/metadc1609000/

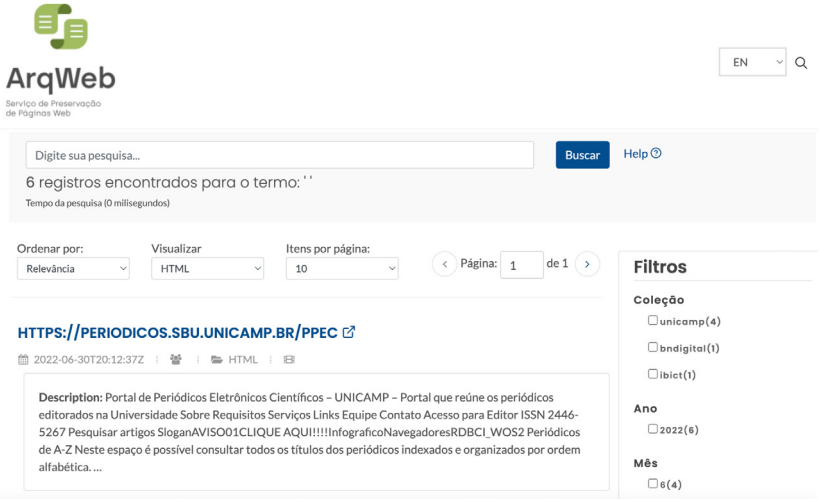
Figura 10 - Plataforma Arquivo.ong.br

Arquivo.ong.br: comunidade de práticas em arquivos da web

Fonte: http://arquivo.ong.br/

O Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) deu início em 2022 ao projeto piloto ARQWEB (BOERES; SAAD, 2023). Esse projeto tem como objetivo arquivar os *sites* da *web* das instituições parceiras da Rede Cariniana, bem como os *sites* governamentais, sem limitar-se apenas a essas categorias (fig. 11).

Figura 11 - Plataforma Arqweb



Fonte: https://arqweb.ibict.br

O Projeto Graúna[35], uma iniciativa do Instituto NUPEF (criado inicialmente como Núcleo de Pesquisa, Estudos e Formação),

35 https://nupef.org.br/grauna

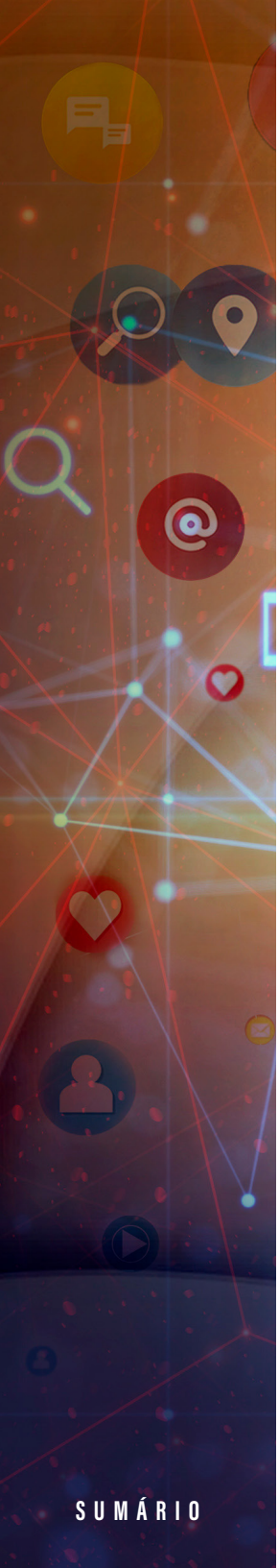
foi apresentado em 2022 na Reunião Brasileira de Ensino e Pesquisa em Arquivologia[36] e também em eventos realizados em maio de 2023[37], concentrando-se na preservação da memória na Internet. Com o objetivo de combater a impermanência dos conteúdos *on-line*, o projeto arquiva *websites* de interesse público relacionados aos direitos humanos, meio ambiente, cultura e saúde. A equipe do projeto identificou diversas ameaças, como ataques a *sites* de organizações da sociedade civil, remoção de informações por governos anteriores e falta de financiamento para projetos culturais, resultando na perda de publicações valiosas (fig. 12).

Figura 12 - Plataforma Graúna

Fonte: https://nupef.org.br/grauna/memoria

36 https://nupef.org.br/artigo/instituto-nupef-participa-da-reuniao-brasileira-de-ensino-e-pesquisa-em-arquivologia

37 https://nupef.org.br/artigo/apresentando-o-projeto-grauna-para-nossos-pares



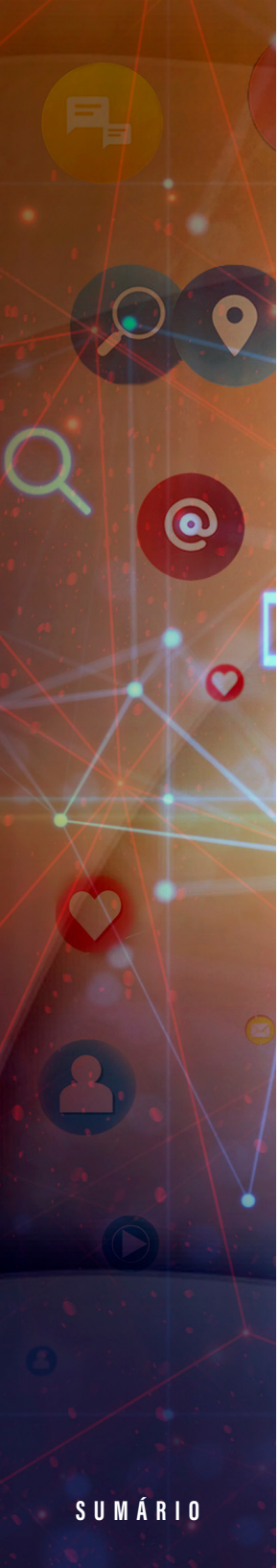
Conforme Melo, Oliveira e Rockembach (2023), distintas frentes, principalmente na Academia e nos Poderes Legislativo e Executivo brasileiro, caminham em direção a uma convergência de projetos e agendas em comum, tanto no desenvolvimento de plataformas, quanto em políticas e pesquisas teórico-aplicadas.

Farias e Bomfim (2019) destacam a produção científica sobre preservação de *website*s em língua portuguesa e o fato de que a garantia do acesso futuro a esses documentos tornou-se uma preocupação para muitas áreas que lidam com a gestão e produção de informação. Ainda, citando Bodê (2008, p. 44), afirmam a relevância das informações na *web*, com o mesmo status de relevância de "um livro de biblioteca, uma carta histórica ou um relatório financeiro contábil em papel de uma grande empresa".

Outra referência a ser destacada foi produzida pelo grupo de pesquisa Dríade: Estudos e Práticas de Preservação Digital, que realizou uma investigação de várias fontes bibliográficas relacionadas à preservação digital (MÁRDERO ARELLANO; SANTOS, 2021). O livro contém um capítulo exclusivo que se dedica ao arquivamento da *web*, abordando referências relevantes no tema.

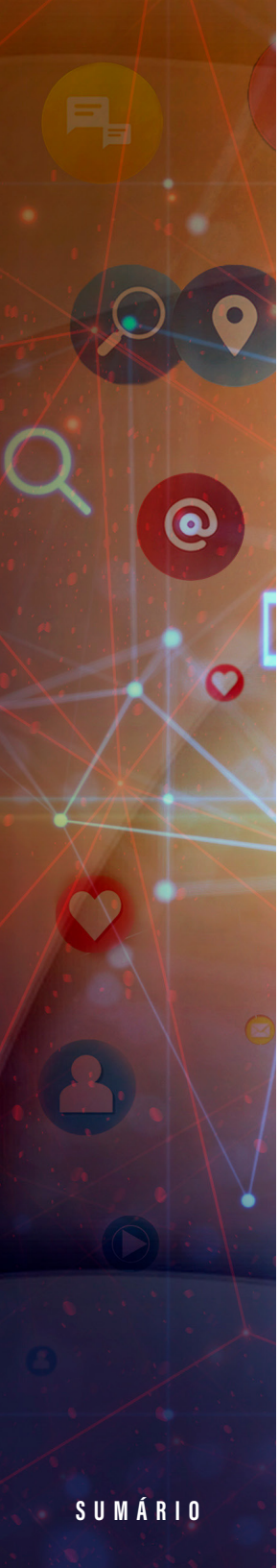
Reconhece-se, desde a reunião plenária do Conselho Nacional de Arquivos em 2004, e com a Carta para a Preservação do Patrimônio Arquivístico Digital (CONARQ, 2005), a importância de preservação e que tanto organizações públicas quanto privadas e indivíduos estão produzindo e transformando documentos arquivísticos em formato digital de maneira crescente, incluindo textos, bancos de dados, planilhas, mensagens eletrônicas, imagens estáticas e em movimento, gravações sonoras, material gráfico, *site*s da Internet e que a diversidade de formatos disponíveis para esses documentos aumenta constantemente.

Em 2010, o Arquivo Nacional do Brasil lançou o Programa Permanente de Preservação e Acesso a Documentos Arquivísticos Digitais, também conhecido como AN Digital. Desde então,



o programa tem disponibilizado em sua página *web* a Política de Preservação Digital, com versões de 2012 e 2016. Ambos os documentos destacam a importância da preservação de documentos digitais, mas afirmam que a preservação será focada em tipos documentais específicos, como texto estruturado com formatação, imagem matricial, imagem vetorial, áudio, audiovisual, mensagens de correio eletrônico, apresentação de *slides*, planilha e base de dados relacional. No entanto, os documentos afirmam que, em um futuro próximo, tipos de documentos digitais mais complexos, como multimídia e páginas *web*, também seriam contemplados (ARQUIVO NACIONAL, 2016)

Em 2021, o Conselho Nacional de Arquivos (CONARQ), por meio da Portaria nº 131, de 9 de novembro de 2021 (CONARQ, 2021a), instituiu a Câmara Técnica Consultiva (CTC Preservação de *website*s e mídias sociais) para definir diretrizes para a elaboração de estudos, proposições e soluções para a preservação de *website*s e mídias sociais. Foi elaborada uma proposta de resolução que estabelece a Política de Preservação de *website*s e Mídias Sociais no âmbito do Sistema Nacional de Arquivos – SINAR e proposta de resolução que define requisitos mínimos de preservação para *website*s e mídias sociais no âmbito do Sistema Nacional de Arquivos. Isso inclui princípios, diretrizes, objetivos e responsabilidades para a preservação e acesso a documentos digitais dinâmicos e complexos, servindo como respaldo para os gestores públicos responsáveis pela preservação do patrimônio arquivístico brasileiro. Como resultado, foram publicadas a Resolução nº 52 do Conselho Nacional de Arquivos (CONARQ) de 25 de agosto de 2023, que estabelece a Política de Preservação de Websites e Mídias Sociais no âmbito do Sistema Nacional de Arquivos - SINAR (CONSELHO NACIONAL DE ARQUIVOS, 2023a) e a Resolução nº 53 do Conselho Nacional de Arquivos (CONARQ) de 25 de agosto de 2023, que define requisitos mínimos de preservação para websites e mídias sociais no âmbito do Sistema Nacional de Arquivos - SINAR (CONSELHO NACIONAL DE ARQUIVOS, 2023b).

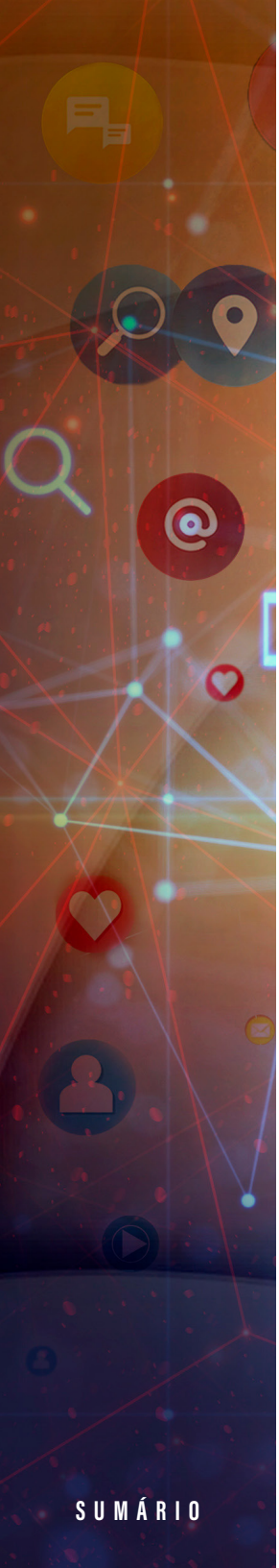


A Política de Preservação de *website*s e Mídias Sociais está alinhada às diretrizes nacionais para documentos digitais e com a política nacional de arquivos públicos e privados, desta forma, os órgãos integrantes do SINAR poderão apoiar a missão e visão do CONARQ por meio de um instrumento de gestão administrativa. Algumas das referências usadas como base foram a AN Digital - Política de Preservação Digital do Arquivo Nacional (2016), as Recomendações para Elaboração de Política de Preservação Digital do Arquivo Nacional (2019), a Política de Preservação Digital da Fundação Biblioteca Nacional (2020) e a Política de Preservação de Acervos Digitais da UFRGS da Universidade Federal do Rio Grande do Sul (2021).

Outro importante aspecto é a observação de pesquisas realizadas sobre o arquivamento da *web* no contexto brasileiro e, para isso, destacamos alguns dos estudos realizados nos últimos anos, em termos de pós-graduação.

O trabalho de Dantas (2014) explorou as práticas contemporâneas de arquivamento, preservação e acesso ao passado da *web*, utilizando a noção de criptografias da memória como uma maneira de analisar as operações mnemônicas em suporte digital. A pesquisa destacou a interdependência entre aspectos técnicos e culturais que moldam os modos de arquivamento e acesso ao passado recente da *web*. Dantas (2014) afirma que a confluência entre a noção de memória digital e acervo digitalização no Brasil têm impactos práticos, como a falta de coleções de páginas da *web* em instituições nacionais. Para abordar essa questão, a pesquisa criou a coleção Buscas.br, que inclui páginas de ferramentas de busca do Brasil de 1997 a 2013, permitindo o contato com a materialidade das fontes digitais.

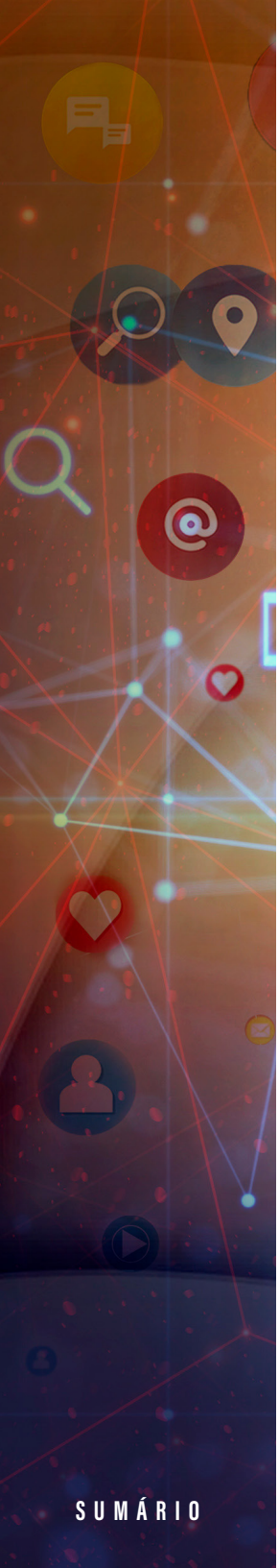
A pesquisa de Ferreira (2018) teve como objetivo preservar os conteúdos audiovisuais publicados pelos candidatos à Presidência em suas páginas do *Facebook* durante a campanha eleitoral de 2018 por meio do arquivamento da *web*, utilizando-se da revisão bibliográfica, análise documental e desenvolvimento de códigos



de programação para coletar e estruturar os dados. Nos resultados, constatou uma carência geral de arquivamento da *web* e mídias sociais de conteúdos brasileiros, com a maior parte dos vídeos não preservados pelas iniciativas analisadas. Dos 2.821 vídeos coletados, 32 não estavam mais disponíveis no formato original ou por restrições legais ou por indisponibilidades diversas. A autora conclui que a preservação de publicações públicas em mídias sociais deve ser fomentada no Brasil para minimizar a efemeridade dos registros.

Martins (2019), em seu estudo sobre mapeamento de públicos estratégicos, descreveu o fenômeno do arquivamento da *web* e as iniciativas das universidades de Columbia e Harvard. Ela também explorou o contexto da Universidade Federal do Rio Grande do Sul (UFRGS), com foco na Faculdade de Biblioteconomia e Comunicação e no Programa de Pós-graduação em Comunicação. A pesquisa foi realizada usando técnicas de análise de documentos oficiais das universidades e seus *site*s e concluiu-se que diferentes atores organizacionais influenciam as estruturas institucionais e que a comunicação é importante para a efetividade das iniciativas. As possibilidades de arquivamento analisadas pela autora incluem história organizacional e a tríade ensino, pesquisa e extensão universitária.

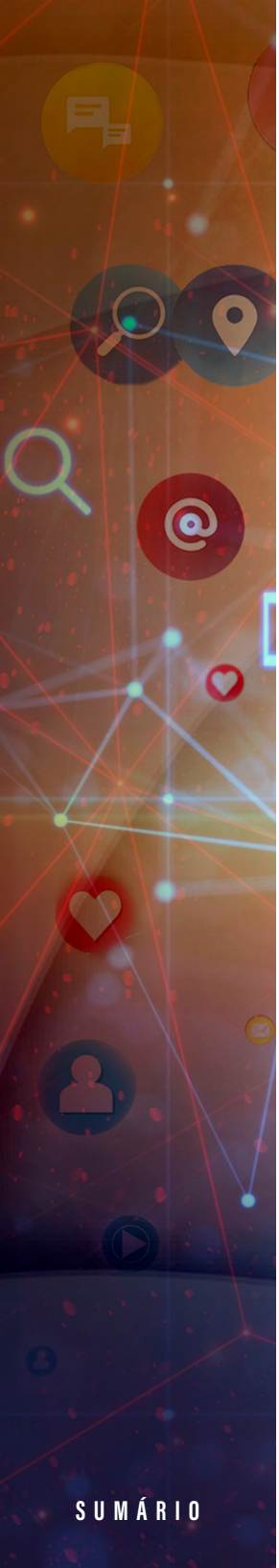
Melo (2020) investigou as possibilidades de arquivamento de *website*s do Governo Federal Brasileiro, com o objetivo de demonstrar as possibilidades de arquivamento a partir de um estudo de caso do domínio gov.br. Foram selecionados 23 *website*s governamentais e a pesquisa consistiu em verificar os recursos oferecidos por estes *website*s, arquivá-los com o uso de rastreador de páginas *web* automatizado Heritrix, reconstruí-los com o uso de *software* automatizado WABAC e comparar os recursos disponibilizados nas versões ao vivo e arquivadas. A recuperação dos *website*s arquivados foi considerada satisfatória, mas alguns recursos não foram recuperados e a permanência dos recursos dos *website*s após seu arquivamento são balizadores para definir a qualidade de uma coleta. Apontou ainda, a necessidade de políticas públicas para sistematizar o arquivamento dos *website*s governamentais.



O trabalho de Laitano (2021) dedicou-se a analisar dois tipos de fontes digitais: documentos digitalizados e arquivos nativos digitais, discutindo primeiro a prática de digitalizar acervos documentais e torná-los acessíveis em repositórios *on-line*, como forma de garantir sua preservação e acessibilidade, dando como exemplo o projeto Brasil: Nunca Mais Digit@l, que disponibiliza documentos relacionados à ditadura militar no Brasil, e em segundo lugar analisando o arquivo de registros nato digitais, como publicações em redes sociais, *site*s, *e-mails* e artigos científicos publicados exclusivamente na Internet. Neste caso, foi utilizado como exemplo o portal Memórias da Ditadura, que disponibiliza informações e conteúdos relacionados ao período da ditadura no Brasil. Em ambos os casos, é dada ênfase na relação entre as tecnologias digitais e a preservação e acessibilidade das fontes históricas, com especial atenção às possibilidades do arquivamento da *web*.

A abordagem de pesquisa de Nunes (2021) visou aos aspectos éticos e legais relacionados ao acesso e uso de informações de *website*s arquivados. O estudo identificou iniciativas internacionais de arquivamento da *web* e sistematizou os aspectos éticos e legais dos documentos jurídicos dessas plataformas, como "termos de uso" e "políticas de privacidade". A partir disto, foram analisados os documentos de 19 plataformas, selecionadas a partir de uma lista do *International Internet Preservation Consortium*, concluindo que o acesso ao universo informacional envolve diversos fatores e que é importante uma política de informação direcionada para decisões que beneficiem a sociedade como um todo. Os documentos jurídicos analisados por Nunes (2021) destacam a importância da educação do usuário e do uso ético, lícito e responsável das informações. A pesquisa ainda defende a importância do trabalho colaborativo e interdisciplinar entre as diferentes áreas.

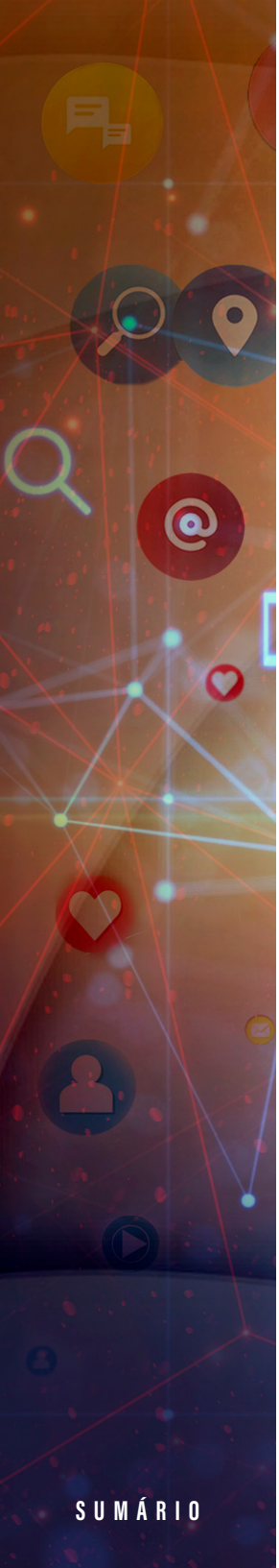
Luz (2021) investigou a relação entre comunicação e memória, propondo a hipótese da Memória Comunicacional. Para isso, realizou uma análise da comunicação dos mandatos da presidenta



Dilma Rousseff (2011–2014 e 2015–2016) disponibilizada no *site* oficial da Presidência da República do Brasil. A pesquisa demonstrou que os produtos da comunicação governamental registram de maneira única a política, a administração pública, a cultura e a sociabilidade de determinado momento histórico, permitindo constituir a Memória Comunicacional. Os resultados apontam para a negligência do Estado brasileiro quanto à preservação e ao acesso às informações oficiais, relacionando também a preservação e arquivamento da *web*, e concluiu que o *site* oficial da Presidência da República do Brasil não atende plenamente aos direitos à informação e à memória esperados nas democracias, reforçando a importância da adoção de uma política de Memória Comunicacional.

O estudo realizado por Fazano (2022) destacou a necessidade de diretrizes técnicas para orientar o processo de arquivamento da *web*, tornando-se ainda mais evidente quando comparado com políticas de instituições dos EUA, Inglaterra, França e outros países da Europa e da Ásia, que têm políticas de captura e preservação de conteúdo *on-line*. A metodologia do trabalho envolveu a análise de coleções existentes, desenvolvidas pela *Library of Congress*, incluindo as eleições presidenciais de 2018, uma coleção sobre literatura de cordel e os Jogos Olímpicos e Paralímpicos de 2016. Fazano (2022) ressalta que as coleções de arquivamento de *site*s diferem das coleções tradicionais de museus, arquivos ou bibliotecas devido à natureza do suporte em que o conteúdo é apresentado, sendo que os procedimentos de arquivamento precisam ser adaptados para garantir a captura adequada das informações, incluindo o número de capturas de determinada página e o nível de navegabilidade desejado, como exemplo, um artigo de jornal que pode ser capturado uma única vez, enquanto um portal sobre determinado assunto pode ter capturas mais abrangentes, incluindo todos os *links* disponíveis e capturas programadas para diferentes datas.

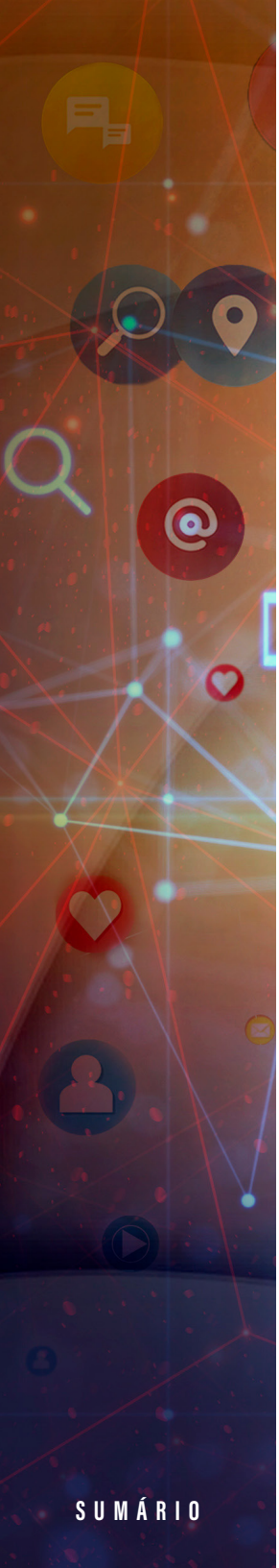
A pesquisa de Terrada (2022) teve como tema a preservação do patrimônio digital, especificamente, a preservação e o arquivamento



de páginas *web*, dada a urgência do desaparecimento destes conteúdos em curtos períodos após sua produção e publicação. O objetivo consistiu em estudar a preservação do patrimônio digital por instituições de guarda da memória das páginas *web*, que fazem parte do *International Internet Preservation Consortium*, com base na Carta sobre a Preservação do Patrimônio Digital da Unesco. Identificou que as bibliotecas nacionais localizadas no continente europeu lideram a tarefa de preservar a *web* dentro do escopo pesquisado e concluiu que as instituições de memória, como bibliotecas e arquivos, devem liderar esse movimento de preservação do patrimônio digital, pois esses objetos complexos fazem parte do patrimônio nacional.

Cabe destacar também o Projeto de Lei nº 2.431/2015, de autoria de Luizianne Lins (PT-CE), e que dispõe sobre o patrimônio público digital institucional inserido na rede mundial de computadores, propõe a caracterização como conduta ilícita da subtração de patrimônio digital em sua formatação original dos *site*s oficiais pelo gestor público, com o objetivo de proteger o acesso público às informações produzidas pelo Estado brasileiro.

O Projeto de Lei apresentado tinha o objetivo de proteger o acervo digital produzido pelo Poder Público, incluindo registros de imagens, vídeo, áudio e texto, a fim de preservar a história e a cultura da sociedade, evitando que gestores apaguem o acervo de seus antecessores, o que acarreta desperdício de dinheiro público e impede o acesso dos cidadãos a registros históricos importantes (LINS, 2015). Um *site* oficial é qualquer *site* de Internet vinculado a órgãos da administração pública direta ou indireta. O Projeto de Lei estipula que os chefes dos poderes públicos são responsáveis pela preservação e manutenção do conteúdo digital institucional. Embora a Lei N. 12.527/2011 Lei de Acesso à Informação (LAI) assegure o acesso à informação produzida pelo poder público, não há dispositivo legal que garanta a manutenção dessas informações em canais de livre acesso, como os *site*s institucionais. Os *site*s oficiais dos poderes públicos desempenham papel fundamental na comunicação entre o



governo e os cidadãos, além de facilitar o acesso às informações de interesse público e ajudar a constituir a memória das instituições. Entretanto, o desafio da preservação de documentos em HTML ou das páginas de *site*s na *web* é grande, já que não há uma legislação que assegure sua manutenção em canais de livre acesso.

Em estudo publicado por Luz (2016), nove das 27 capitais brasileiras tiveram seus *site*s apagados ou sem acesso público, o que evidencia a necessidade de proteger o acervo digital produzido pelo poder público. Portanto, a proposição busca amenizar um problema grave da atualidade, protegendo o patrimônio digital e garantindo o acesso aos registros históricos importantes para a sociedade.

Alguns eventos vêm discutindo o tema sobre o Arquivamento da *web* no Brasil nos últimos anos, como o Fórum da Internet do Brasil 2019 - 9ª edição - Preservação do Conteúdo *web* Brasileiro[38]; o I e II Simpósios Políticas e Estratégias de Preservação de Conteúdo na *web*, como parte do Congresso Internacional em Tecnologia e Organização da Informação (TOI), em 2020[39] e 2021[40]; o *web*inar Arquivamento e Preservação da *web* – Trocando Experiências[41], promovido pelo grupo de pesquisa DRÍADE – Estudos e Práticas de Preservação Digital, ligado à Rede Cariniana e ao Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e a Reunião Brasileira de Ensino e Pesquisa em Arquivologia VII REPARQ 2022 – Arquivamento da *web* e preservação digital: dos *website*s às redes sociais, promovida pelo Fórum Nacional de Ensino e Pesquisa em Arquivologia – FEPARQ[42].

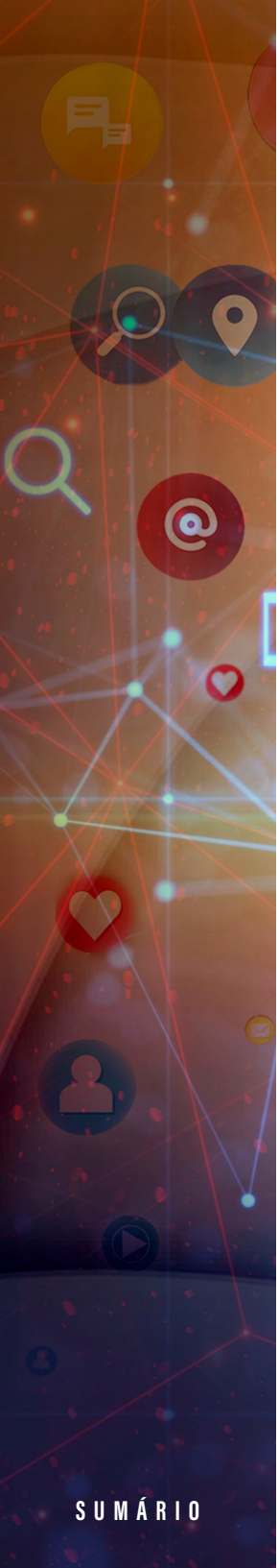
38 Programação disponível em: https://forumdaInternet.cgi.br/2019/programacao/detalhe/2/1577/ e vídeo do evento em: https://www.youtube.com/watch?v=y_k-kOc254Y

39 Vídeo do I Simpósio disponível em: https://www.youtube.com/watch?v=X6SmBmbD7fg

40 Vídeo do II Simpósio disponível em: https://www.youtube.com/watch?v=3KhAu7OczSA

41 Vídeo disponível em: https://www.youtube.com/watch?v=QPZuLwWR9y8

42 https://feparq.org



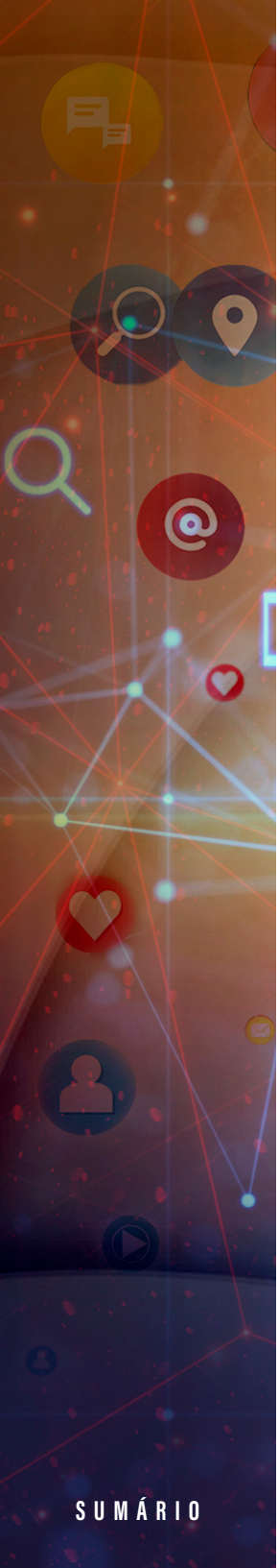
PROFISSIONAIS QUE TRABALHAM COM OS ARQUIVOS DA *WEB*

Como uma atividade complexa e que envolve diferentes habilidades, diversos profissionais podem fazer parte de uma equipe que desenvolve soluções e serviços em arquivamento da *web*, incluindo arquivistas, bibliotecários, profissionais de tecnologia da informação e curadoria digital, entre outros, corroborando com Márdero Arellano (2008) ao tratar de programas de preservação digital. Por sua vez, um arquivista da *web* é um profissional responsável por coletar, preservar e fornecer acesso às informações da Internet. Os arquivistas da *web* são importantes porque ajudam a garantir que a grande quantidade de informações na Internet não seja perdida e possa ser acessada no futuro.

O arquivamento da *web* envolve várias tarefas complexas, como identificar conteúdo relevante, coletar o conteúdo da *web*, armazená-lo de maneira segura e fornecer acesso ao conteúdo de maneira amigável e fácil de navegar. Os arquivistas da *web* devem ter uma compreensão profunda da tecnologia da informação e das ferramentas e técnicas usadas para coletar, preservar e fornecer acesso a informações baseadas na *web*.

Os arquivistas da *web* desempenham um papel fundamental na preservação do patrimônio cultural da Internet, e seu trabalho é essencial para pesquisadores, historiadores e o público em geral que dependem do acesso à informação da *web*. Sem os arquivistas da *web*, muitas informações valiosas na Internet seriam perdidas e as gerações futuras seriam incapazes de acessá-las e aprender com elas.

As tarefas e responsabilidades de um arquivista da *web* podem incluir, entre outras atividades:



a. identificar conteúdo relevante para arquivamento da *web*, avaliando, selecionando e utilizando de curadoria digital;

b. capturar o conteúdo da *web*, usando técnicas e ferramentas como rastreadores automatizados e captura manual;

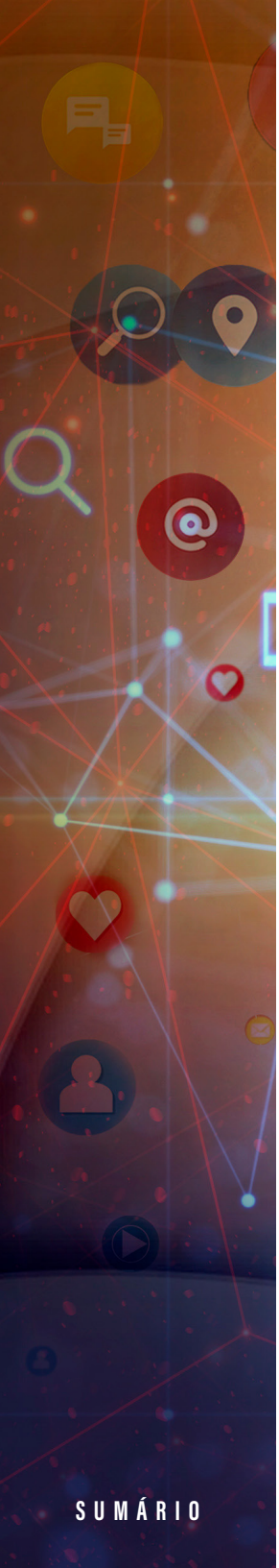
c. armazenar o conteúdo coletado de maneira segura, usando técnicas de segurança da informação e *backups* seguros;

d. fornecer acesso ao conteúdo arquivado, usando ferramentas como portais da *web* e mecanismos de pesquisa;

e. manter o arquivo da *web*, incluindo atualizações regulares, *backups* e verificações de segurança;

f. garantir a conformidade com as leis e regulamentações pertinentes, como por exemplo, o *General Data Protection Regulation* (GDPR) na União Europeia ou a Lei Geral de Proteção de Dados Pessoais (LGPD), no Brasil.

A relevância de um arquivista da *web* reside em sua capacidade de selecionar, capturar, preservar e fornecer acesso a informações valiosas da Internet. Ao realizar essas tarefas, os arquivistas da *web* ajudam a garantir que a herança cultural da Internet não seja perdida e que possa ser acessada pelas pessoas ao longo do tempo. É recomendável que as organizações tenham profissionais dedicados somente às atividades de arquivamento da web, estes profissionais também desempenham um papel crucial no apoio à pesquisa, educação e ao público em geral, preservando a *web* em relação aos seus conteúdos e contexto de produção.



OS ARQUIVOS DA *WEB*, USOS E USUÁRIOS

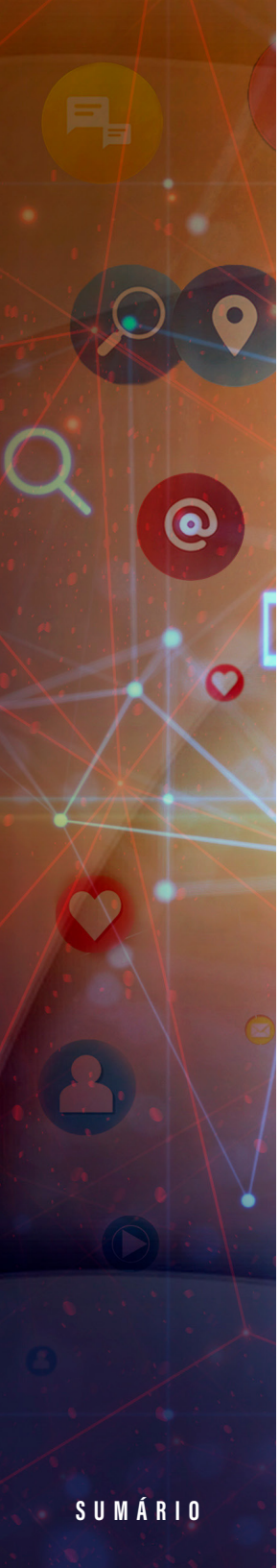
Antes de olharmos para os usuários dos arquivos da *web*, precisamos pensar na *web* capturada e os possíveis usos desses arquivos como fonte de informação. Do nascimento da *web* a partir do primeiro *website* em 1991, para a *web* 2.0 e demais gerações, muitas transformações se passaram. A cultura participativa e a convergência digital, propagada por Jenkins (2008), modificou a forma como nos comunicamos e produzimos informação e, como destacado por Castells e colaboradores (2009), tudo se torna potencializado com a onipresença da Internet a partir de dispositivos móveis, a conectividade sem fios e a Internet das coisas (IoT).

Os arquivos da *web* formam conjuntos de informações da Internet que foram capturadas, preservadas e organizadas para acesso e estudo e são recursos importantes que podem ser usados de várias maneiras, incluindo pesquisa, educação e público em geral.

1. Como ferramenta de pesquisa: o arquivamento da *web* permite que os pesquisadores acessem e estudem informações da *web* de maneira sistemática e confiável. Os arquivos da *web* fornecem uma riqueza de informações que podem ser usadas por pesquisadores em áreas como história, sociologia e humanidades digitais para estudar a evolução da Internet e seu impacto na sociedade.

2. Como meio de preservação: o arquivamento da *web* ajuda a garantir que a herança cultural da Internet seja preservada para as demais gerações.

3. Como um recurso para a educação: o arquivamento da *web* fornece uma relevante e única fonte de informações que pode ser usada por educadores para ensinar sobre a história e o impacto da Internet, evolução das tecnologias, de *web* design e arquitetura da informação. Possui ainda uma ampla



gama de informações históricas de diferentes comunidades virtuais, como importante recurso educacional.

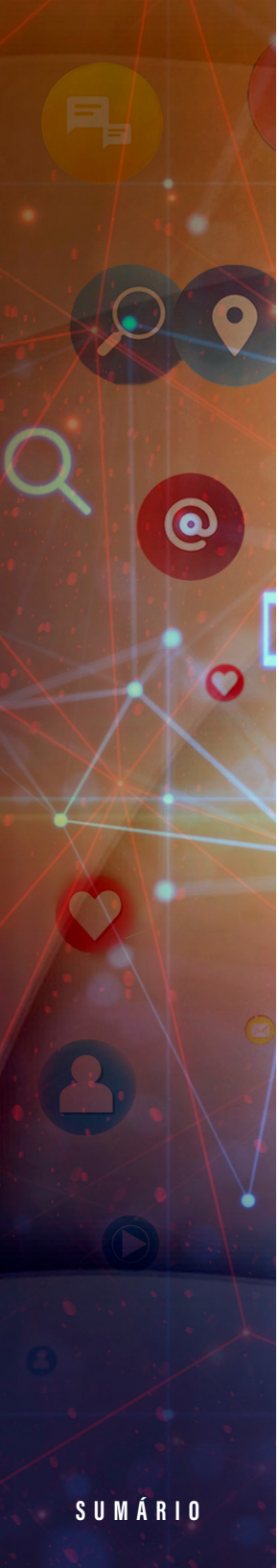
4. Acesso público: os arquivos da *web* fornecem acesso a informações da Internet que podem não estar disponíveis por outros meios, como *site*s que foram removidos da Internet ou informações que não são mais facilmente acessíveis.

Compreender o público, conforme Martins e Rockembach (2019, 2020) é de fundamental importância para a constituição de arquivos da *web*. Neste sentido, os estudos poderiam iniciar com a pergunta: Como é possível melhorar o engajamento do público com os arquivos da *web*? Desta forma, muitas outras subperguntas são provocadas: Qual é o seu público? Quais são seus interesses de pesquisa? Como isso pode se relacionar com a memória organizacional, individual e coletiva? Como proporcionar melhor acesso e melhores experiências do usuário às plataformas digitais?

As Humanidades digitais[43] constituem uma grande área que pode usufruir dos arquivos da *web*, enquanto fonte de informação sobre os últimos 30 anos no meio digital. As Humanidades Digitais formam, portanto, um campo que combina diferentes áreas das humanidades com tecnologia, possibilitando refletir sobre como o conhecimento é produzido e validado, incluindo as práticas, métodos, discursos, modelos e relações entre métodos computacionais e humanísticos.

Os métodos digitais, segundo Rogers (2016), permitem a análise de *hiperlinks*, a *web* preservada e novas perspectivas de pesquisa. Arquivos da *web* podem ser usados como uma ferramenta para pesquisas sociais, capturando históricos de *websites*,

43 Manifesto das Humanidades Digitais, versão em português – Disponível em: https://humanidades-digitais.org/manifesto-das-humanidades-digitais. No Brasil, a Rede Colaborativa para as Humanidades Digitais (Colab HD+) lançou em 2023 a Declaração de Pirenópolis para as Humanidades Digitais – Disponível em: https://zenodo.org/record/8030170

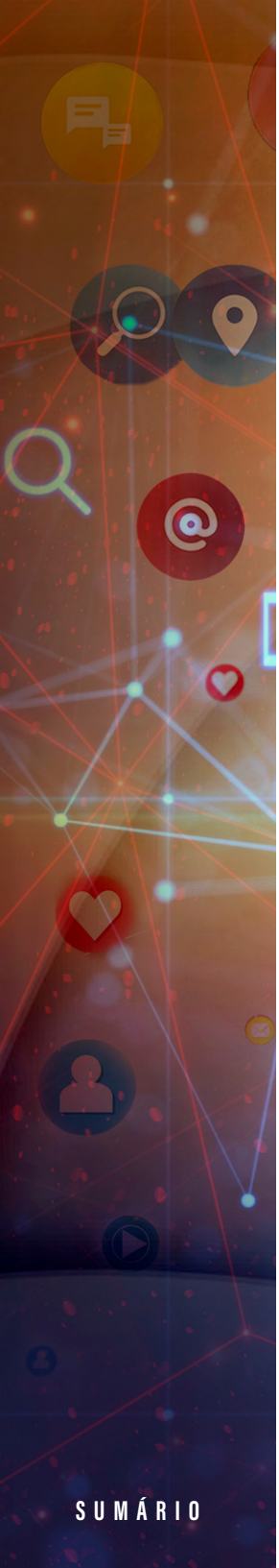


seu versionamento, e revelando mudanças na credibilidade e dependência de fontes na *web*, também sendo possível acompanhar a evolução de páginas individuais ou múltiplas ao longo do tempo.

Além disso, um ponto importante a ser destacado é a relação entre a preservação digital e o planejamento colaborativo para garantir o acesso a informações digitais para pesquisas em Humanidades Digitais, conforme assinala Rockembach (2019). Especialmente no campo informacional, conforme Pimenta (2016), a Ciência da Informação e as Humanidades Digitais se conectam pela importância do espaço criado pela experiência digital, que oferece possibilidades inéditas de comunicação, produção e distribuição de informação. Essas duas áreas se complementam na busca de novas formas de reflexão sobre os efeitos das tecnologias e suas implicações éticas, políticas e sociais, criando uma zona de negociação para a inovação e a reflexão.

Existem três perspectivas em Humanidades Digitais que podem ser identificadas, como mencionado por Brügger e Finnemann (2013). A primeira surgiu nas décadas de 1950 e 1960, focando a digitalização de materiais e a utilização de métodos computacionais, especialmente em mainframes. A segunda perspectiva surgiu na década de 1980, com foco nos estudos de interação humano-computador, e na relação com computadores pessoais. Além dessas duas abordagens, há uma terceira perspectiva que surge na década de 1990 em torno da mídia e dispositivos digitais. Nessa abordagem, a *web* ainda era vista como um ambiente para distribuição, apresentação e comunicação, e não como um objeto de pesquisa em si mesmo, como se torna possível com os arquivos da *web*.

O arquivamento da *web* é uma ferramenta importante para as Humanidades Digitais, que compreende o estudo da herança cultural da Internet e seu impacto na sociedade. A partir disto, faz-se necessário pensar no planejamento de estruturas e procedimentos de preservação digital que possibilitem o acesso à informação



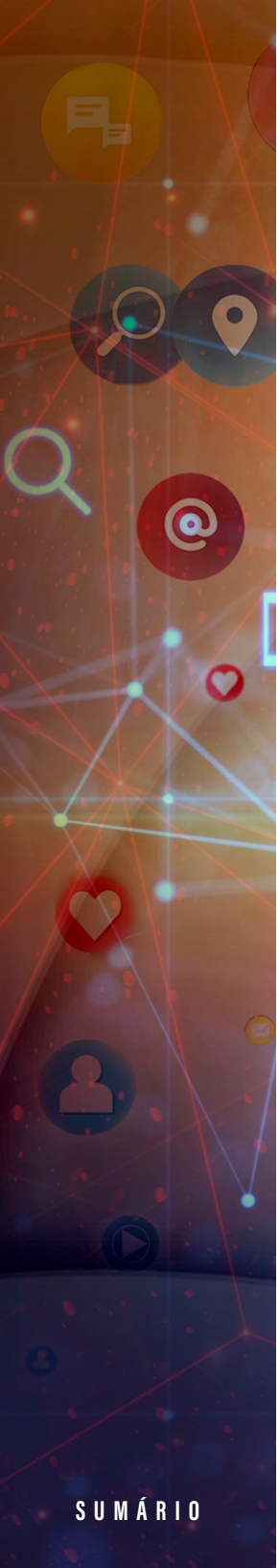
(ROCKEMBACH, 2022). Nicodemo, Rota e Marino (2022) questionam como a presença das "nem tão novas" tecnologias podem afetar não só a produção e difusão do conhecimento e seus artefatos, mas também mudanças epistemológicas, teóricas e metodológicas. A produção e disseminação dos arquivos da *web* resultam em uma diversificação de fontes de informação e o arquivamento da *web* permite que os pesquisadores em Humanidades Digitais acessem e estudem a vasta quantidade de informações na Internet, incluindo *site*s, postagens em mídias sociais e outros conteúdos *on-line*.

Existem várias maneiras pelas quais o arquivamento da *web* pode ser usado em Humanidades Digitais, sendo um de seus aspectos principais o uso como fonte de dados. O arquivamento da *web* fornece uma riqueza de dados que pode ser usada por pesquisadores em Humanidades Digitais para estudar a evolução da Internet e seu impacto na sociedade. Abordagens etnográficas[44] ou netnográficas também podem ser impulsionadas com o uso dos arquivos da *web* como fonte de informação, bem como análises de conteúdo[45].

A conexão entre tecnologia e humanidades é incontornável na sociedade da informação atual. As Humanidades Digitais formam uma comunidade de práticas, que trabalha no cruzamento da computação e das humanidades (TERRAS, 2006; ALVES, 2016). A importância da comunidade é evidente nas Humanidades Digitais e preservação digital, pois desempenha um papel crucial no desenvolvimento de boas práticas e pesquisas sobre o arquivamento da *web*, bem como no aprimoramento das tecnologias relacionadas. A colaboração pode ocorrer em nível internacional ou em pesquisas locais que contribuam para a preservação de conteúdos digitais e tenham impacto na investigação científica, ensino e prática profissional.

44 Para estudos etnográficos em arquivos da web, ver Ogden, Halford e Carr (2017)

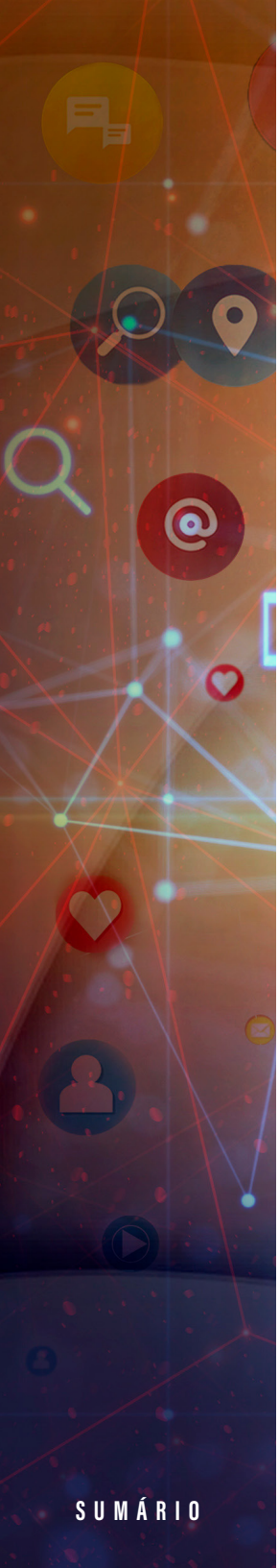
45 Para análise de conteúdo ver https://www.york.ac.uk/res/e-society/projects/28/Qualitative.pdf#page=6



Brügger (2021) também argumenta a importância de arquivos da *web* disponibilizarem seus acervos como dados para pesquisas em Humanidades Digitais. Atualmente, a maioria dos arquivos da *web* oferecem apenas acesso limitado aos seus arquivos, mas combinar metadados e dados de *hiperlink* de um arquivo da *web* poderia fornecer conhecimento relevante sobre a estrutura de um domínio da *web*. O autor argumenta que os arquivos da *web* devem estar preparados para a aplicação de métodos digitais e que as informações sobre a proveniência dos arquivos da *web* devem ser transparentes para melhor compreensão dos resultados da análise.

Já ao trabalhar com arquivos da *web* na historiografia, é importante considerar fatores como escala e expectativas de privacidade. Os arquivos da *web* oferecem o poder de reconstruir vidas *on-line* e fornecer informações sobre os pensamentos de pessoas do passado (RODRIGUES, ROCKEMBACH, 2021). Para usar efetivamente os arquivos da *web*, é importante considerar o papel das plataformas, colaborar com arquivistas e bibliotecários. Sem isso, pode ser difícil recuperar histórias precisas do início da era da *web* e, ao mesmo tempo, sem estes recursos digitais, não seremos capazes de escrever a história a partir dos anos 1990 ou posterior.

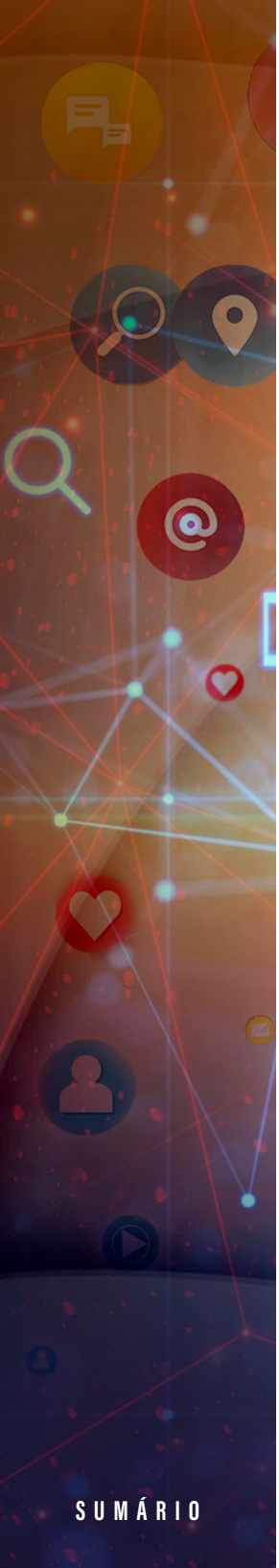
Em uma era de abundância de informações, os métodos e tecnologias empregados podem fazer a diferença nas análises dos materiais digitais. Pimenta (2022) argumenta como a mudança em andamento afeta a forma como os historiadores lidam com a leitura e a escrita e que a prática linear e textual já não é suficiente, pois a leitura e a escrita também são imagéticas, não lineares, multidimensionais e mosaicas. A "virada digital" transformou o campo da História e demanda do historiador contemporâneo habilidades para lidar com diversas fontes digitais e transversalizar diferentes registros de informação, sendo necessário distanciar-se do paradigma bidimensional de uma História escrita apenas em papel ou processadores de texto.



No contexto da pesquisa científica, os arquivos da *web* podem auxiliar cientistas que estudam a evolução da Internet e seu impacto na sociedade. A Historiografia é o estudo dos métodos e princípios da história e da escrita histórica e no contexto dos arquivos da *web*, pode ser uma ferramenta valiosa sobre a história da Internet. Com a grande quantidade de informações sobre a história da Internet, incluindo *site*s, publicações em redes sociais e outros conteúdos *on-line*, os historiógrafos podem obter *insights* sobre a Internet, seu desenvolvimento e impacto na sociedade. Por exemplo, eles podem estudar o crescimento e a evolução da Internet ao longo do tempo, a adoção de novas tecnologias e o impacto da Internet no discurso político, cultural e social analisando postagens de mídia social relacionadas a um evento específico. No contexto da cidadania, os arquivos da *web* podem ser um recurso relevante para entender o papel da Internet na formação do discurso público e do engajamento político.

Ainda, um cientista que estuda o crescimento da Internet pode usar um arquivo da *web* para analisar o número e os tipos de *site*s que foram criados ao longo do tempo ou para estudar como o uso das mídias sociais mudou ao longo dos anos. Além de fornecer dados para pesquisa, os arquivos da *web* podem ser usados por cientistas como uma ferramenta para conduzir pesquisas (SCHNEIDER; FOOT; WOUTERS, 2010).

Como afirmado por Hockx-YU (2014), as fontes acadêmicas devem ser de fácil acesso e de boa qualidade, sendo a integridade da captura, o conteúdo intelectual, o comportamento e a aparência atributos importantes para o arquivamento da *web*. Além de textos e conteúdos multimídia, os pesquisadores também estão interessados em paratextos, que incluem elementos que cercam ou prolongam o texto em um *site*, como cabeçalhos, rodapés, palavras ou frases de referência e mapas de *site*s.



Assim como ilustrado por Gomes (2010), alguns exemplos de perfis demonstram diferentes usos, incluindo *webmasters* ou administradores que procuram uma versão anterior dos *site*s e podem usar os arquivos da web, como o Internet Archive, para reconstruir estes *sites*, baseado em aplicações de terceiros, tais como o *Wayback Rebuilder*[46], *Wayback Machine Downloader*[47] e o *Wayback Downloads*[48].

Ainda destacamos outros usos, como daqueles que precisam da obtenção de provas por profissionais do direito e usuários que buscam *site*s a partir de *links* quebrados nos seus favoritos. Outros perfis incluem jornalistas procurando informações que não encontram em outros meios e fontes. Neste sentido, os arquivos da *web* podem servir como instrumento para verificação das fontes e de *fake news*. O projeto "Fake Dói"[49], do Instituto Vero[50], é um movimento que promove a verificação de informações na Internet, visando combater as notícias falsas. Através de dicas e vídeos acessíveis, o projeto busca desenvolver habilidades de checagem e conscientizar sobre a responsabilidade de compartilhar conteúdo. Além disso, salientamos a importância dos arquivos da *web* para revisitar o passado e verificar informações antigas.

Algumas ferramentas dos arquivos da *web*, como a funcionalidade "*changes*" do *Wayback Machine (Internet Archive)*, podem ser significativamente importantes no estudo das mudanças dos *site*s e do versionamento. Comparando, por exemplo, as mudanças implementadas em políticas e termos de uso de plataformas como o *Facebook* com a funcionalidade *changes*[51], é possível verificar

46 https://waybackrebuilder.com

47 https://www.waybackmachinedownloader.com

48 https://waybackdownloads.com/

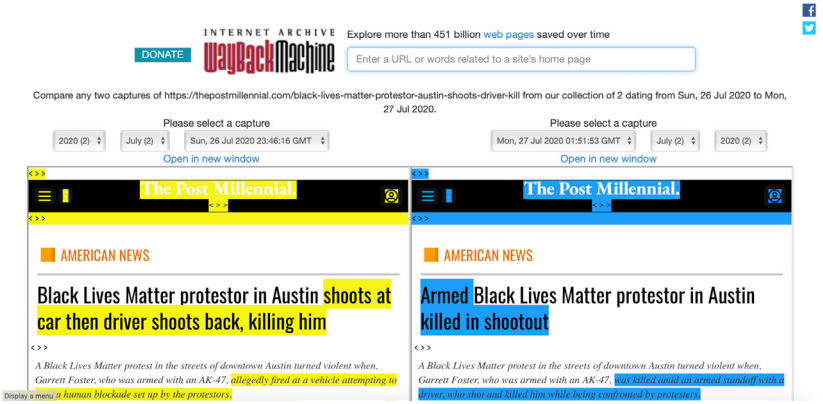
49 https://www.vero.org.br/fakedoi/a-maquina-do-tempo-na-internet

50 https://www.vero.org.br/

51 https://web.archive.org/web/changes/https://www.facebook.com/policy.php

quais pontos foram realmente editados nesses documentos *on-line* no passar dos anos[52]. O uso dos arquivos da *web* pode se aplicar a diversos outros casos, inclusive na checagem de fatos e no combate à desinformação.

Figura 13 - Ferramenta *changes* do *Wayback Machine*

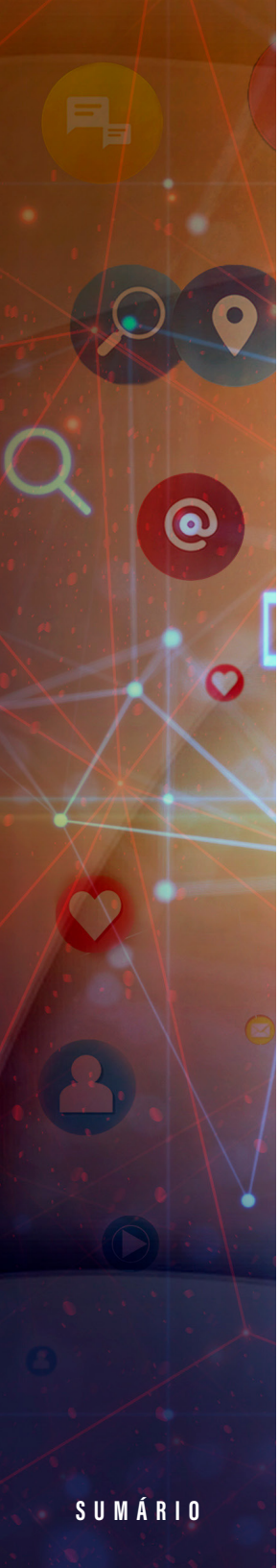


Fonte: https://twitter.com/waybackmachine/status/1287851081389621248

No exemplo da imagem no *X.com* (antigo *Twitter*) do *Wayback Machine*, é possível perceber as mudanças no texto e na notícia veiculada, mesmo que o *link* original não tenha mudado. Desta forma, é possível controlar o versionamento das páginas e também perceber o que foi acrescentado, retirado ou editado.

Os arquivos da *web* fornecem uma riqueza de informações sobre o uso da Internet por indivíduos e grupos para fins políticos e cívicos, incluindo *site*s, publicações em redes sociais e outros conteúdos *on-line*. Ao estudar essas informações, os pesquisadores podem obter conhecimento sobre as maneiras pelas quais a Internet mudou a natureza do engajamento político e cívico e o impacto da Internet na formação da opinião pública e da ação política.

52 https://thenextweb.com/news/the-wayback-machine-now-lets-you-track-changes-to-web-pages



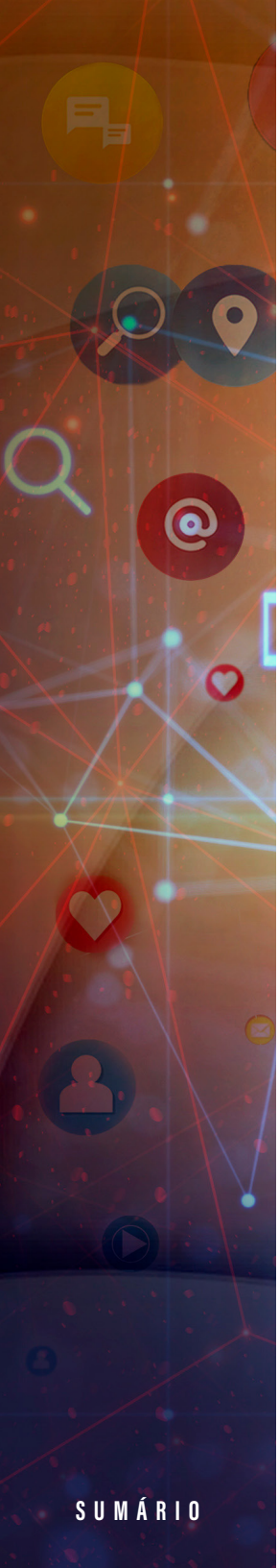
Além de fornecer dados para pesquisa, os arquivos da *web* também podem ser usados por indivíduos e grupos como uma ferramenta para conduzir suas próprias pesquisas e análises. Por exemplo, um grupo de cidadãos interessados em estudar o impacto da Internet no discurso político poderia usar um arquivo da *web* para coletar e analisar postagens de mídia social relacionadas a um evento político específico.

Segundo Maemura e colaboradores (2018) os pesquisadores devem entender o processo de criação de um arquivo da *web* para confiar na validade de suas descobertas. Ainda, os autores colocam que algumas das pesquisas realizadas utilizando arquivos da *web* apontam que os esforços de desenvolvimento de ferramentas se concentram no estudo de arquivos em escala, analisando os *web ARChive* (WARCs), o que combina múltiplos recursos digitais em um arquivo agregado junto com as informações relacionadas, e usando esses conjuntos de dados como objeto de estudo. Além disso, o desenvolvimento de ferramentas está focado em fornecer interfaces para análise e visualização de dados, bem como APIs (*Application Programming Interface*) para arquivos da *web* e para dar suporte ao uso em pesquisas.

O uso do arquivamento da *web* nas escolas pode dar novas perspectivas na compreensão de como os estudantes enxergam a produção de documentos públicos, mas sobretudo documentos pessoais, uma oportunidade educacional que vai além daquela aplicada à arquivistas e bibliotecários. Como destacado por Harris, Beis e Shreffler (2021), aprender sobre arquivamento da *web* ensina aos alunos a importância dos registros *on-line* como documentação histórica, eles podem apreciar o valor de suas próprias contribuições, incluindo *postagens* de mídia social, *blogs* e projetos em grupo. O chamado "arquivamento cidadão na *web*" destaca a responsabilidade cívica e a necessidade de produzir um registro histórico diversificado.

2

POLÍTICAS DE PRESERVAÇÃO DA *WEB*



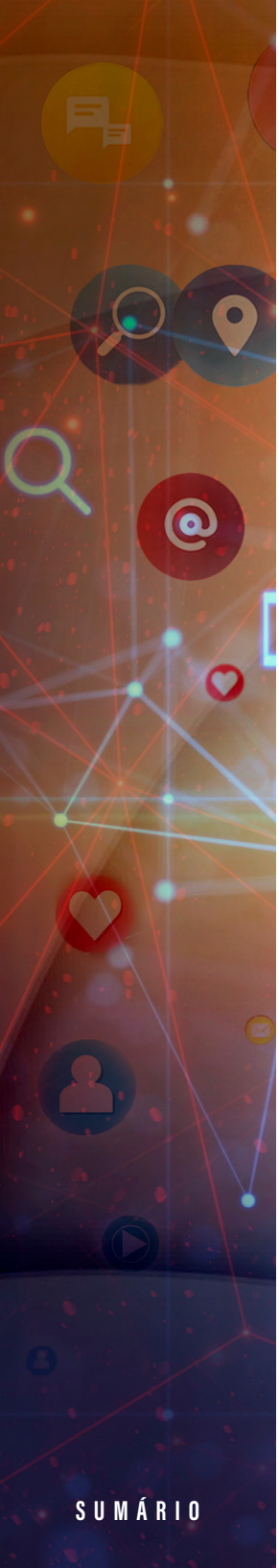
As políticas de preservação da *web* referem-se ao conjunto de diretrizes e práticas que as organizações e os indivíduos seguem para garantir que as informações e o conteúdo da *web* sejam preservados para acesso futuro. Elas descrevem os métodos e estratégias que as organizações seguem para manter e preservar o conteúdo baseado na *web*. Essas políticas são importantes porque as informações na *web* estão em constante mudança e podem ser facilmente perdidas se não houver uma política para garantir que sejam preservadas adequadamente.

As políticas de preservação da *web* geralmente incluem diretrizes para fazer *backup* regular do conteúdo da *web*, armazená-lo em vários locais e usar ferramentas de arquivamento da *web* para capturar e preservar o conteúdo dinâmico da *web*. Elas também podem incluir diretrizes para garantir a acessibilidade de longo prazo do conteúdo da *web*, como o uso de formatos de arquivo e metadados padronizados.

As organizações responsáveis pela preservação de importantes informações históricas ou culturais na *web*, como arquivos, bibliotecas e museus, geralmente possuem políticas de preservação da *web* mais abrangentes. Essas políticas podem incluir diretrizes específicas para digitalização e armazenamento de artefatos físicos, bem como procedimentos para curadoria e organização de conteúdo da *web*.

As políticas de preservação da *web* são críticas para garantir que a riqueza de informações na *web* não seja perdida e permaneça acessível para as gerações futuras. Sem essas políticas em vigor, grande parte das informações valiosas na *web* correriam o risco de se perder para sempre.

Um dos principais componentes das políticas de preservação da *web* é o uso de ferramentas e técnicas de arquivamento. Essas ferramentas capturam, armazenam e gerenciam o conteúdo da *web* ao



longo do tempo, permitindo que as organizações preservem informações valiosas e as disponibilizem para pesquisadores e o público.

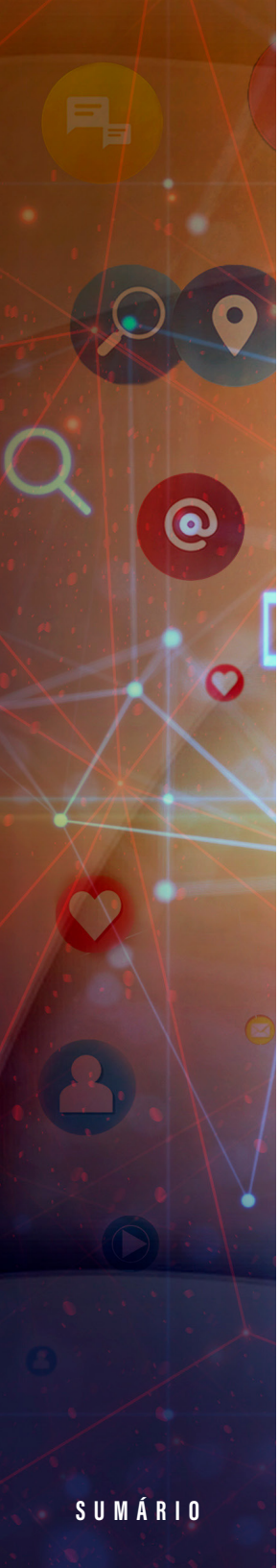
As políticas de preservação da *web*, geralmente, incluem diretrizes para a seleção e priorização do conteúdo a ser preservado. As organizações devem considerar, cuidadosamente, qual conteúdo baseado na *web* é mais valioso e digno de preservação e priorizar esse conteúdo para arquivamento e acesso de longo prazo.

Outro aspecto importante das políticas de preservação da *web* é o uso de estratégias e tecnologias de preservação digital. Essas estratégias e tecnologias ajudam a garantir que o conteúdo da *web* seja preservado em um formato acessível e sustentável ao longo do tempo. Isso pode incluir o uso de formatos e padrões de preservação digital, bem como a implementação de sistemas robustos de preservação digital. Ao delinear estratégias e técnicas para arquivamento da *web*, seleção e priorização de conteúdo e preservação digital, essas políticas ajudam as organizações a gerenciar e manter com eficácia seu conteúdo baseado na *web*.

Também, devemos levar em conta, nas políticas de preservação da *web*, a identificação e preservação de importantes materiais culturais e históricos baseados na *web*, podendo incluir a coleta e preservação de *site*s, postagens de mídia social e outros conteúdos digitais que tenham valor histórico ou cultural significativo.

Um dos principais desafios na preservação de *site*s é o ritmo acelerado das mudanças tecnológicas. À medida que novas tecnologias são desenvolvidas e adotadas, *site*s mais antigos podem se tornar obsoletos e difíceis ou impossíveis de acessar, podendo dificultar a preservação do conteúdo desses *site*s ao longo do tempo.

O grande volume de informações na *web* é outra questão a ser considerada. Existem bilhões de *site*s na Internet, cada um contendo uma grande quantidade de informações. Isso dificulta a identificação de quais *site*s valem a pena preservar, como gerenciar



o armazenamento e disponibilizar mecanismos necessários para a recuperação dessas informações. Além disso, muitos *site*s são dinâmicos e mudam constantemente, o que pode dificultar a captura e preservação de um instantâneo do *site* em um determinado momento e ainda, torna difícil garantir que todas as informações sejam preservadas e permaneçam acessíveis ao longo do tempo.

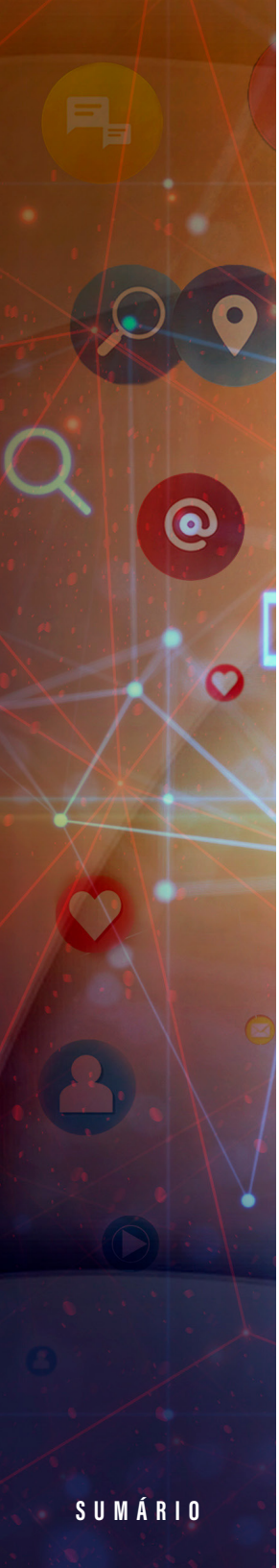
A propriedade e o controle de *site*s também podem ser um problema. Em alguns casos, o criador original de um *site* pode não estar mais disponível para dar permissão para que o *site* seja preservado, ou a propriedade do *site* pode não ser clara, dificultando a obtenção dos direitos necessários para preservar o conteúdo.

A preservação de *site*s apresenta uma série de desafios, mas é uma tarefa importante para garantir que informações e recursos inestimáveis, disponíveis na *web*, não sejam perdidos.

Preservar *site*s pode ser uma tarefa desafiadora por vários motivos. Um dos principais desafios é o fato de que a tecnologia utilizada para criar e hospedar *site*s está em constante mudança. À medida que novas tecnologias são desenvolvidas, as mais antigas se tornam obsoletas, o que pode dificultar a manutenção e o acesso a *site*s criados com métodos mais antigos.

Outro desafio é o fato de muitos *site*s serem construídos usando tecnologias proprietárias ou formatos proprietários. Isso pode dificultar o acesso e a preservação do conteúdo desses *site*s, pois a tecnologia ou formato proprietário pode não ser amplamente suportado ou pode exigir ferramentas especializadas para acesso.

Finalmente, a natureza efêmera da *web* também representa desafios para a preservação de *web*. Os *site*s estão em constante atualização, com novos conteúdos sendo adicionados e conteúdos antigos sendo removidos, tornando difícil a preservação de um instantâneo de um *site* em um ponto específico no tempo, pois o conteúdo do *site* pode ter mudado no momento que é preservado.

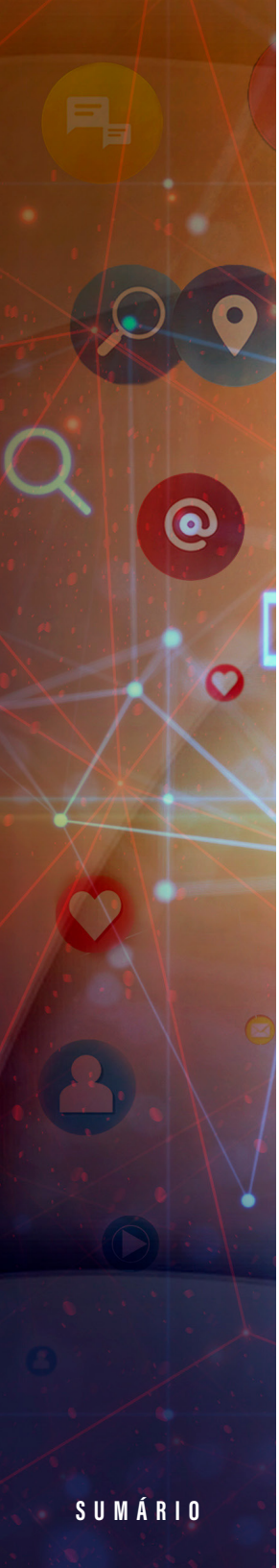


Os arquivos da *web* envolvem diversas áreas, como biblioteconomia, arquivologia, gestão da informação e tecnologia, e seus respectivos profissionais. Os conteúdos podem ser organizados por coleções, mas também por sua proveniência, para fins de memória institucional e organizacional. Os *website*s podem ser considerados publicações, mas também documentos que fornecem informações orgânicas e únicas. A natureza multifacetada da *web* e seus objetivos requerem uma abordagem interdisciplinar e a colaboração entre diferentes profissionais é essencial para projetos de arquivamento da *web*.

JUSTIFICATIVA DA NECESSIDADE DA CRIAÇÃO DE NOVOS ARQUIVOS DA *WEB*

Os dados digitais gerados hoje podem não ser legíveis por dispositivos e *software*s no futuro, levando a preocupações sobre a preservação de nossa história digital e conteúdos *on-line* podem tornar-se impossíveis de recuperar. Esse processo de decadência dos dados está em andamento, com muitos disquetes do início da era digital não sendo mais legíveis, da mesma forma, não se espera que os CDs durem mais do que algumas décadas. A partir do uso da *web* como forma de produção e acesso à informação *on-line*, com a estrutura cliente-servidor, o foco da preservação digital desloca-se do suporte para a rede, procurando manter a relação entre diferentes *hiperlinks* que dão contexto a informação produzida e aos recursos multimídia e multiplataformas utilizados.

Um dos mais famosos arquivos da *web* do mundo é o *Internet Archive,* uma organização sem fins lucrativos, fundada em 1996, que visa preservar *site*s da Internet e outros artefatos culturais em formato digital. Sua missão é fornecer acesso universal a todo o conhecimento, fornecendo acesso gratuito a pesquisadores, historiadores, acadêmicos e ao público em geral. O *Internet Archive* começou



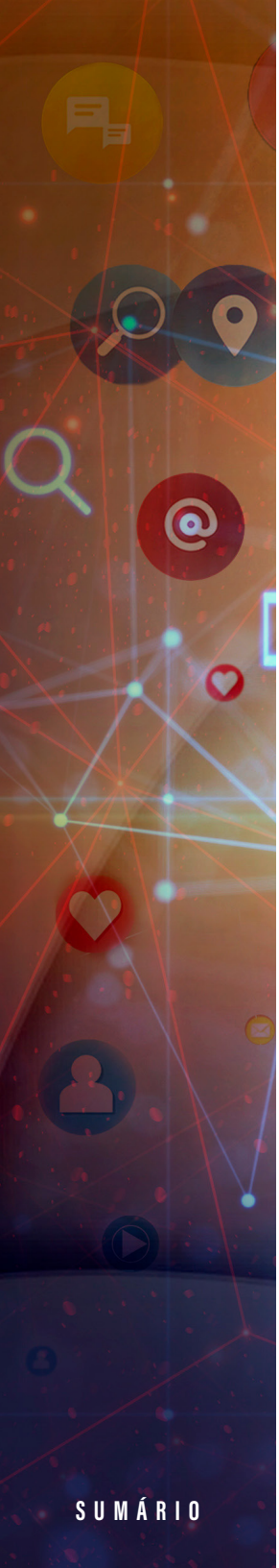
arquivando a Internet nos anos 1990 e, desde então, preservaram mais de 735 bilhões de páginas da *web*, 41 milhões de livros e textos, 14,7 milhões de gravações de áudio, 8,4 milhões de vídeos, 4,4 milhões de imagens e 890.000 programas de *software*.[53] A Organização trabalha com diversos parceiros ao redor do mundo visando a preservação digital, e qualquer pessoa com uma conta gratuita pode fazer *upload* de mídia para o *Internet Archive*. Alguns de seus valores incluem a privacidade de seus usuários, evitar manter os endereços IP de seus leitores e oferecer seu *site* no protocolo seguro (https).

Milligan, Ruest e Lin (2016) recomendam uma estratégia híbrida na cobertura de determinados eventos, onde a curadoria cruza diferentes métodos, como a captura de *websites* com rastreadores automatizados, uso de fontes capturadas pelo *Internet Archive* e captura de redes sociais como o *X.com* (antigo *Twitter*). Os autores afirmam que não é possível confiar somente na captura do *Internet Archive* como uma substituição a rastreamentos e curadorias profissionais e especializadas. O *Internet Archive*, apesar de realizar uma ampla cobertura, pode por vezes não ser suficiente em termos de profundidade de captura e não alcançar determinados *sites* e respectivas versões.

Shiozaki e Eisenschitz (2009) argumentam que os arquivos da *web* proporcionam novas possibilidades de fontes de informação e a captura de uma preciosa história social que não se encontra em papel, construindo arquivos e coleções mais abrangentes, a partir da estrutura e contexto único da *web*. Nos casos de captura de domínios nacionais completos, a justificativa também recai sobre outros *site*s além dos públicos, pois as organizações do setor privado e os proprietários de *site*s individuais não oferecem preservação da *web* a longo prazo.

Uma das principais justificativas é a efemeridade das informações produzidas na *web*. A quebra de *links* é considerada um grande problema para localizar os recursos disponíveis *on-line*, mas existem também outros fatores que devemos levar em consideração.

53 https://archive.org/about/

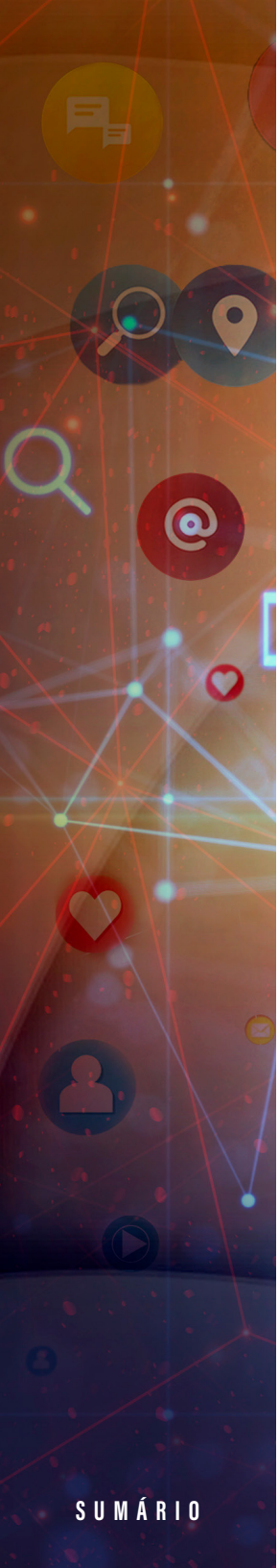


Quando um usuário tenta acessar um recurso na *web* e não consegue, isso pode ocorrer por diversos motivos, incluindo quebra de *link*, também conhecida como *link rot*, um fenômeno da Internet bastante estudado e documentado. O conteúdo também pode ser deletado do servidor, gerando uma falta de controle sobre o apagamento destas informações. Como os recursos da *web* são encontrados por meio de *hiperlinks*, o recurso também pode ter sido movido para outro endereço ou renomeado sem que os *links* que apontavam para ele tenham sido atualizados. Além disso, problemas de conectividade ou configuração do servidor também podem fazer com que o conteúdo não seja recuperado.

Outro motivo comum é a falta de manutenção e atualização do *site* ou do servidor. Muitas vezes, os *site*s são criados sem um planejamento adequado de manutenção e, com o tempo, os *links* tornam-se obsoletos e o conteúdo desatualizado. Outro ponto são as tecnologias utilizadas na criação do *site* que desatualizam rapidamente, o que pode levar a problemas de compatibilidade com navegadores mais recentes.

Outro problema se relaciona com o chamado *content drift*, quando o conteúdo original de uma página é modificado ao longo do tempo sem acompanhamento e quando há mudanças no conteúdo, sem a possibilidade de identificar as mudanças e sem controle de versionamento. Ao mesmo tempo em que milhões de informações são produzidas, outras são sobrepostas, dificultando a recuperação destas informações com o passar do tempo.

Os chamados *sites* em fim de vida, ou *end-of-life*, também podem ser motivo de preocupação, podendo ocorrer porque o proprietário do *site* decidiu retirá-lo do ar, porque a organização pública ou privada foi dissolvida e não previu a preservação de seus *websites*, porque o registro do domínio expirou ou ainda porque o *site* foi encerrado por outros motivos não identificados. Um exemplo é o acontece quando uma plataforma é desativada, como o caso da rede social *Orkut*.



Quando acontece a troca de governos e ocorre a necessidade de produzir novas informações e novos *sites*, é necessário pensar na preservação digital, pois quando não há uma preocupação em preservar ou não há um responsável pela preservação destes conteúdos a história se perde. É o que acontece com campanhas eleitorais ou eventos pontuais, ou ainda em casos de guerras e catástrofes, como a guerra na Ucrânia e a iniciativa Salvando o patrimônio cultural ucraniano *on-line*, *Saving Ukrainian Cultural Heritage On-line* (SUCHO)[54].

Os *sites* em fim de vida correm o risco de serem perdidos para sempre, a menos que sejam capturados e preservados por meio de arquivamento da *web*. Os esforços de arquivamento da *web* geralmente se concentram na captura e preservação de *sites* em fim de vida útil para garantir que permaneçam acessíveis às gerações futuras para pesquisa, histórico e acesso à informação.

Major (2021), observa que plataformas e empresas encerram suas atividades, gerando a perda de informações e até mesmo de *sites* populares, como o *Flickr*, podem alterar suas políticas, resultando na perda de conteúdo do usuário. Dentre os motivos para encerramento de atividades incluem-se decisões de empresas para tornar os serviços mais lucrativos, falta de financiamento, problemas técnicos ou alterações no objetivo de um *site*, mudanças na propriedade e nos modelos de negócios, diminuição do envolvimento do usuário e obsolescência tecnológica. Este fechamento de *sites*, que leva à perda de conteúdo, pode ser mitigado pelo arquivamento da *web* e esforços individuais para baixar e salvar dados.

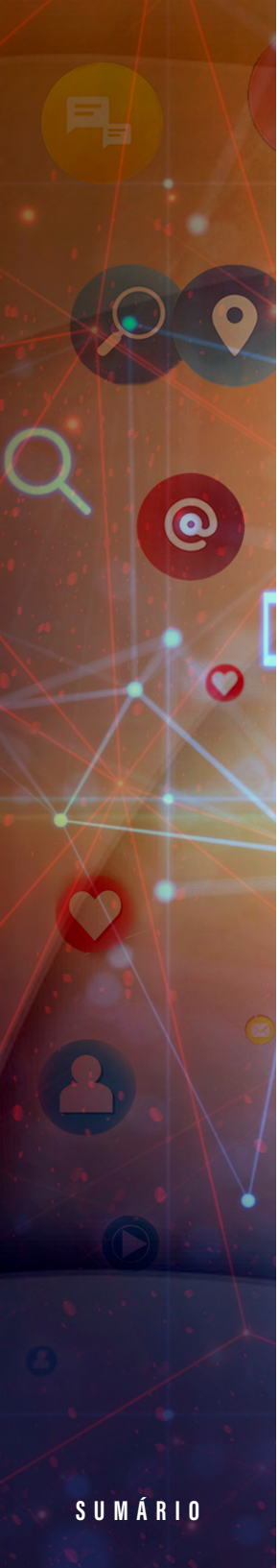
Muitos termos podem ser utilizados para referir-se ao fenômeno da quebra de *links* como obsolescência, persistência (ou a falta de persistência dos *links*), *links* podres (*links rot*), *links* descaídos (*link decay*), idade das trevas digital (D*igital Dark Age*) e pergaminho digital (*digital vellum*). Estes termos representam uma parte

54 www.sucho.org

da variedade de conceitos encontrados em artigos. Como mencionado por Król e Zdonek (2020), a quebra de *links* e a perda de dados de maneira geral, é amplamente discutida na literatura científica internacional, podendo ser identificado por diferentes nomes, tais como *link rot* (BERNERS-LEE, 1998; NIELSEN, 1998; BENBOW, 1998; DENEMARK, 1996; TAYLOR; HUDSON, 2000), *link decay* (GOH; NG, 2007; HENNESSEY; GE, 2013), *Broken link, dead link, dangling link* (MARKWELL; BROOKS, 2002; KOBAYASHI; TAKEDA, 2002), *Content drift* (BURNHILL *et al*., 2015; ZHOU *et al*., 2015), entre outros termos. A figura abaixo, em formato de disquete, ilustra os diversos termos encontrados na literatura relacionados à quebra de *links* e perda de dados.

Figura 14 - Nuvem de palavras sobre a quebra de *links* e perda de dados

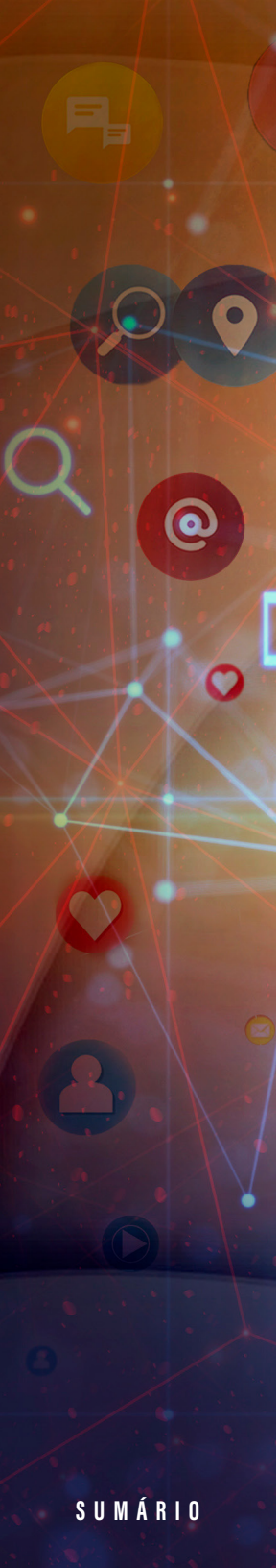
Fonte: os autores, baseado em Król e Zdonek (2020).



Como mencionado por Costa, Gomes e Silva (2016), muitos estudos demonstram dados relacionados à vida útil dos *links.* Ntoulas, Cho e Olston (2004) constataram que 80% das páginas *web* não estavam disponíveis em sua forma original após um ano da publicação. Outros estudos como o de Dellavalle e colaboradores (2003) afirmam que 13% das referências da *web* em artigos acadêmicos desaparecem após 27 meses; Salaheldeen e Nelson (2012) destacam que 11% dos recursos de mídia social, como os postados no *Twitter*, agora *X.com*, são perdidos após um ano e de acordo com Evangelou, Trikalinos e Loannidis (2005) a perda relativa aos materiais suplementares dos artigos científicos chega a 9,6% após cinco anos da publicação.

Um estudo que analisou 1 milhão de *links* extraídos de 3.5 milhões de artigos (KLEIN *et al*. 2014), descobriu que 13% dos *links* em artigos *arXiv* e 22% dos *links* em artigos de periódicos da Elsevier estavam quebrados e, no geral, 75% dos *links* não foram armazenados em *cache* em nenhum *site* de arquivamento da *web* em um período de duas semanas após a publicação. Foram considerados somente *links* que apontam para conteúdo da *web* em geral e não para outros artigos científicos. Apesar de apenas 25% dos artigos acadêmicos conterem *links* para o conteúdo da *web* em geral, cerca de 80% dos artigos que tinham pelo menos um *link* para conteúdo da *web* estava com problemas (quebra de *link* ou mudança do conteúdo original), o que significa que, pelo menos, uma referência ao conteúdo da *web* em geral estava *off-line* ou não tinha sido arquivada. Isso é problemático porque o conteúdo dos servidores pode mudar, desaparecer ou ser transferido para outras mãos, deixando os pesquisadores sem acesso aos recursos que foram utilizados em suas pesquisas.

De acordo com uma pesquisa realizada por Brunelle e colaboradores (2015), algumas redes sociais, como o *X.com* (antigo *Twitter*), apresentam problemas de arquivamento de conteúdo. Os resultados dos testes realizados indicam que apenas 4,2% dos conteúdos foram arquivados corretamente. Além disso, os autores

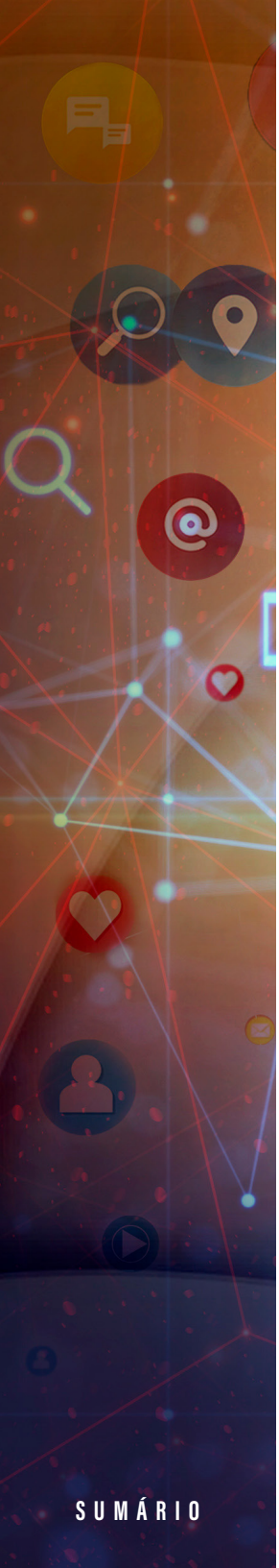


apontam a linguagem *Javascript* como uma das principais dificuldades para a automação da coleta de dados na *web*. Isso ocorre porque o *Javascript* utiliza *scripts* executados do lado do cliente, o que pode carregar dados sem alteração do *Uniform Resource Identifier* (URI), em português, Identificador Uniforme de Recurso ou exigir interação do usuário, o que dificulta a coleta automatizada de dados.

Zittrain, Albert e Lessig (2014) relataram, em estudo a partir de uma amostra de periódicos jurídicos, que 70% dos *links* publicados entre 1999 e 2011 e 50% dos *links* em pareceres da Suprema Corte dos Estados Unidos estão quebrados e, portanto, não são mais acessíveis.

Já os dados do estudo de Sife e Bernard (2013) sobre referências utilizadas em teses e dissertações indicam que 58% das referências *web* não estavam disponíveis. A maioria dos casos (92,7%) resultou em uma mensagem de erro «404 Arquivo Não Encontrado», sendo que o domínio .com representou a maior parcela de URLs indisponíveis (28,2%). A duração média das URLs citadas foi de 2,5 anos.

Na pesquisa de Goh e Ng (2007), foi explorado o fenômeno de degradação de *links* em três periódicos de Ciência da Informação de grande relevância. Realizaram o *download* de artigos publicados em um intervalo de sete anos (1997 a 2003) e os *links* presentes foram analisados. Por meio da análise, foi possível calcular a meia-vida dos *links* em cerca de cinco anos, o que se compara favoravelmente com outras áreas de estudo (1,4 a 4,8 anos). Além disso, foram identificados os tipos de problemas de acessibilidade encontrados nos *links*, bem como as características que podem estar associadas à deterioração. Verificaram que cerca de 31% das citações não estavam acessíveis durante o período de teste e a maioria dos erros foi causada pela ausência de conteúdo (com o código de erro HTTP 404). O domínio .edu apresentou as maiores taxas de falha, chegando a 36%, quando comparado com outros domínios populares de nível superior.



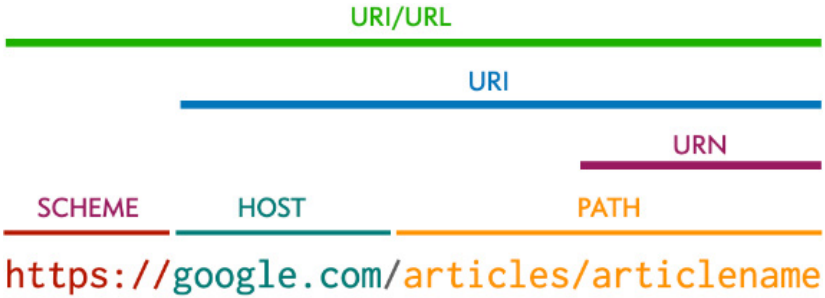
Um estudo de Aronsky e colaboradores (2007) aponta que um total de 11,9% das URLs encontradas nas referências não funcionavam apenas dois dias após a inclusão na base de dados *PubMed*, e os motivos mais comuns de inacessibilidade foram URLs não encontrados e URLs com tempo limite esgotado. O estudo demonstrou que a taxa de URLs inacessíveis e imprecisas no momento da publicação ainda é alta, e autores e editores devem verificar a precisão e acessibilidade das referências de URL.

Além do conceito de URL, ainda temos os conceitos de *Uniform Resource Name* (URN)[55] ou Nome Uniforme do Recurso e *Uniform Resource Identifier* (URI)[56] ou Identificador de Recurso Uniforme. Um URI é uma sequência de caracteres que identifica exclusivamente recursos em uma rede, enquanto um URL geralmente se refere a recursos que podem variar ao longo do tempo. Um URI pode ser um URL, um URN ou ambos e consiste em várias partes, incluindo um esquema, autoridade, caminho, consulta e fragmento. O esquema identifica a especificação do identificador e o protocolo de acesso para o recurso, enquanto a autoridade identifica a autoridade de nomeação hierárquica. Já o caminho identifica o recurso dentro do escopo do esquema de URI e da autoridade de nomenclatura, e a consulta fornece informações não hierárquicas sobre o recurso. Por fim, o fragmento permite identificar uma parte do recurso principal ou visualizar sua representação. Embora as pessoas geralmente usem URL para endereços da *web*, o URI é um identificador mais abrangente e recomendado para uso em vez de URL (fig. 15). O termo URL, entretanto, é mais disseminado entre usuários e público em geral.

55 https://www.rfc-editor.org/rfc/rfc8141

56 https://www.rfc-editor.org/rfc/rfc3986

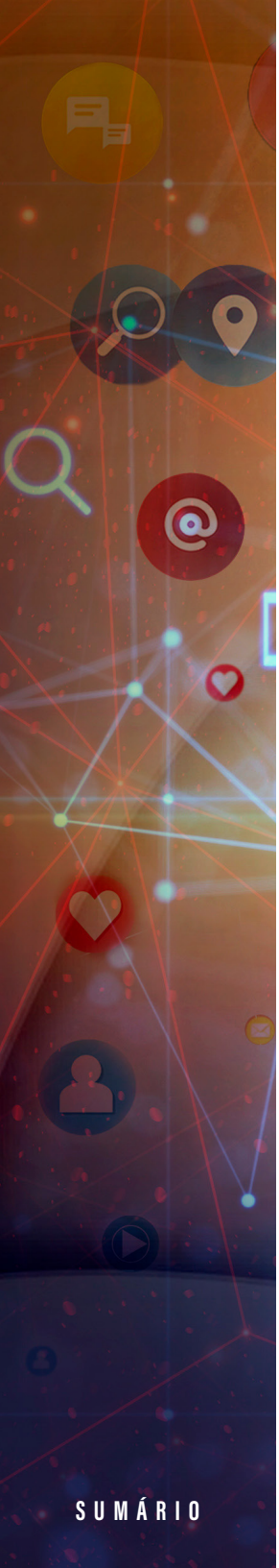
Figura 15 – Estrutura de uma URL



Fonte: Miessler (2022).

Uma forma de contornar o problema dos *links* inacessíveis é a utilização de identificadores persistentes que podem ser citados para recuperar objetos digitais por um longo período de tempo. Permitem uma identificação precisa dos recursos, mediante uma identificação publicamente visível, remetendo para os metadados, o conteúdo de um repositório digital ou de uma base de dados que dará o endereço correspondente ao identificador, por meio de um serviço de resolução de *links*, mesmo que o endereço mude com o tempo.

A atribuição de identificadores únicos e persistentes às entidades descritas, qualquer que seja a sua natureza, de acordo com Bermès (2006a), é absolutamente necessária para assegurar a correta gestão, acessibilidade e reuso dos dados e metadados que serão produzidos. A questão de identificadores persistentes para citações e *links* vem se tornando cada vez mais crucial em nível internacional revelando não só a sua importância, mas também seu potencial. Cada vez mais, um recurso que não pode ser citado perde sua visibilidade. As tecnologias semânticas da *web* como o RDF e as ontologias são baseadas inteiramente na noção de URIs, que já não identificam apenas recursos, mas também pessoas, conceitos etc. Neste universo, não ter um identificador persistente será equivalente a não existir.



Alguns sistemas de identificação persistentes requerem ou recomendam a captura de metadados em conjunto com o registo do recurso. Este é o caso, por exemplo, do *Digital Object Identification* (DOI) e do *Archival Resource Key* (ARK). Outros sistemas são dedicados ao intercâmbio de metadados, mas incluem ou exigem um sistema de identificação persistente para cumprirem o seu papel, que é o de facultar o acesso ao próprio recurso, e necessitam para seu funcionamento a incorporação do *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), protocolo desenvolvido pela *Open Archives Initiative*[57] que define um mecanismo para coleta de registros de metadados em repositórios.

O DOI é um sistema de identificação de objetos num ambiente digital, baseado no sistema *Handle* e gerido pela *International DOI Foundation*[58]. Permite o monitoramento de um recurso, desde o compartilhamento e o acesso às informações *on-line* visto que não é suficiente identificar um recurso quando ele é criado; ele deve ser rastreável e acessível também ao longo do tempo.

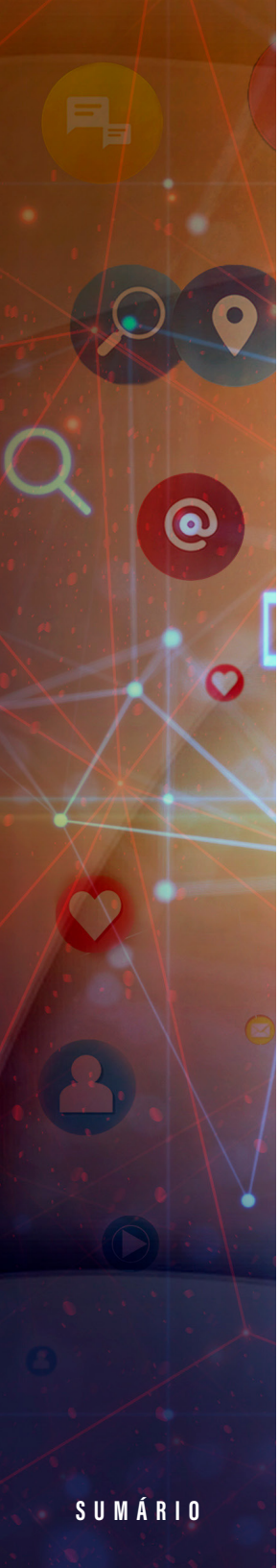
O ARK é um sistema de identificação que, entre outras coisas, pretende ser um endereço da *web* (URL) que não retorna o erro *404 Page Not Found*. A *ARK Alliance*[59] é uma comunidade global aberta que apoia a infraestrutura ARK. São identificadores persistentes abertos, correntes, não pagos e descentralizados, que podem ser criados em menos de 48 horas depois de contratado. Eles identificam qualquer objeto digital, físico, ou abstrato e são semelhantes a DOIs, URNs e *Handles*. Para começar a criar ARKs, basta preencher um formulário de solicitação para a organização interessada.

De acordo com Bermès (2006b), os identificadores persistentes têm uma sintaxe comum que se baseia nas especificações do W3C e a sintaxe desses identificadores é composta por três partes:

57 https://www.openarchives.org/

58 https://www.doi.org/

59 https://arks.org/



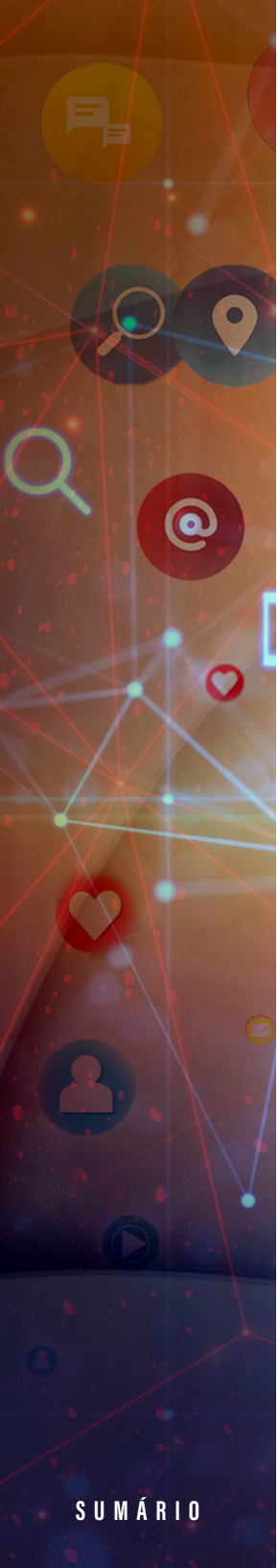
um prefixo que indica o contexto em que o identificador é atribuído (http:, ftp:, urn, etc.); um elemento que designa a autoridade dentro desse sistema, que pode ser designada pelo nome, por um código ou nome codificado, como um nome de domínio e por último o próprio "nome", ou seja, uma cadeia de caracteres que identifica de forma única o recurso, dentro desse sistema e para essa autoridade.

O conceito *Persistent Uniform Resource Locators* (PURL)[60] foi desenvolvido por Stuart Weibel e Erik Jul na *On-line Computer Library Center, Inc.* (OCLC), em 1995. Os PURLs implementam uma forma de identificador persistente para recursos digitais. São genericamente URIs http cujos serviços redirecionam os pedidos de GET (usado quando o cliente deseja obter recursos do servidor) para outro URI baseado em localização, com a promessa adicional de persistência a longo prazo. PURL está hospedado desde 2016 no *Internet Archive*, após ter sido transferido da OCLC. O fato de a propriedade ter sido transferida para o *Internet Archive* salienta que foi tomado especial cuidado em preservar e continuar a operar esta infraestrutura identificadora e os seus registros. São endereços da *web* que atuam como identificadores permanentes em face de uma infraestrutura da *web* dinâmica e em constante mudança.

O *Perma.cc*[61] é um serviço oferecido pela Escola de Direito de Harvard e procura oferecer soluções para o problema de *link* "podre" (*link rot*) fornecendo a estudantes e pesquisadores um URL permanente, conhecido como "*link Perma.cc*", para materiais da *web* citados em seus artigos. Evita o fenômeno em que as citações em periódicos acadêmicos para materiais da *web* desaparecem com o passar do tempo, resultando em "*links* quebrados". Isso permite que os usuários garantam que os *links* sejam preservados indefinidamente para fins acadêmicos. *Perma.cc* pode armazenar uma variedade de tipos de arquivos, incluindo PDFs, documentos do *Word*, planilhas,

60 http://www.archive.org/services/purl

61 https://guides.library.harvard.edu/perma

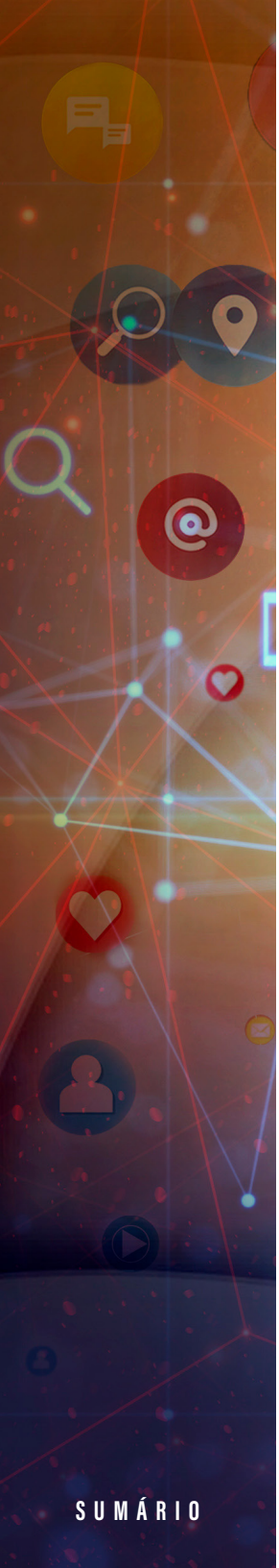


PowerPoints e arquivos de vídeo. Geralmente, todos os *links Perma* são públicos, mas poderá haver *links* privados, se necessário, ou quando o proprietário do *site* codificou o html como <*no archive*> ou o *site* está protegido por um acesso pago ou assinatura.

Trabalhos acadêmicos há muito citam fontes primárias ou outros trabalhos científicos para fornecer fontes de fatos, incorporar estudos anteriores e reforçar argumentos. A citação ideal deve conectar o leitor às obras às quais o autor do texto faz referência, tornando mais fácil rastrear, verificar e aprender mais com as fontes indicadas. À medida que as fontes citadas se movem para a *web*, essa ligação deveria se tornar mais fácil, evitando o deslocamento do leitor até uma biblioteca ou arquivo, permitindo recuperar o material citado imediatamente com um único clique. Mas o *link*, uma URL, que aponta para um recurso hospedado por terceiros, só sobreviverá enquanto este terceiro o preservar. E conforme os *sites* evoluem, nem todos terão interesse suficiente em preservar esses *links*. Este problema não existe para fontes impressas, ou, pelo menos, não da mesma forma.

Para serem "Encontráveis" (*findable*), os objetos digitais devem ser exclusivamente identificáveis, ou seja, devem ter nomes exclusivos e devem ser localizáveis, por associação do nome único com um protocolo de recuperação. Eles também devem ter, pelo menos, alguns metadados descritivos associados para ajudar na descoberta e verificação desse objeto quando o identificador não é conhecido *a priori*. Para Juty e colaboradores (2020) um identificador é um nome único dado a um objeto. Identificadores digitais são uma sequência ordenada de caracteres para nomear exclusivamente recursos digitais. Às vezes, um identificador pode ser "lido", como se sua *string* revelasse uma semântica significativa, que é o recurso ao qual se refere.

No estudo de Zittrain, Albert e Lessig (2014) ressalta-se que a ascensão da *web* permitiu a criação e troca de conhecimento acadêmico e as fontes nas quais ele se baseia e, ao menos que sejam

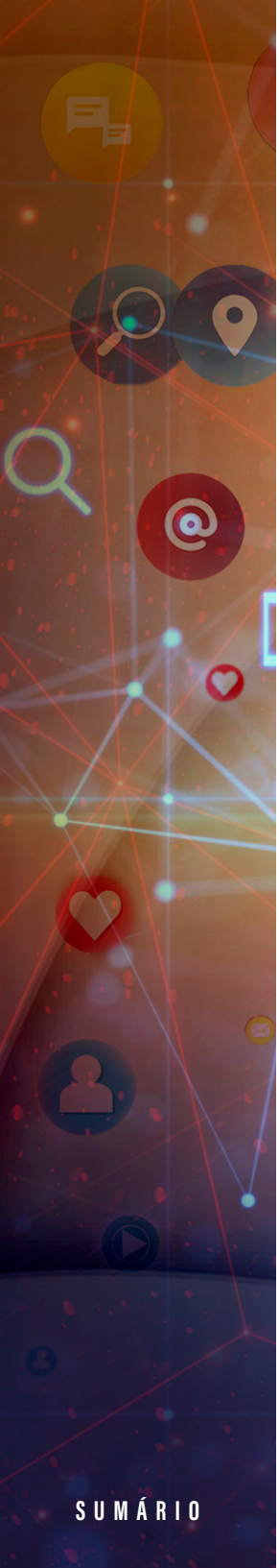


tomadas medidas para arquivar esse tipo de informação, os futuros leitores não conseguirão obter as fontes nas quais os autores confiaram. Para que a integridade da erudição não sofresse com essas perdas, sugeriram um sistema distribuído para unir as fontes e os autores e assim restaurar a integridade, garantindo que as fontes fossem adequadamente preservadas para a posteridade. Mas ainda hoje, verificamos em diversos estudos que o problema persiste sem uma solução consensual e aplicada em todos os níveis da preservação de *links* da *web*.

Em outro estudo do mesmo ano de Agata e colaboradores (2014) foram analisados um conjunto de 10 milhões de páginas da *web* coletadas em 2001 e os resultados da pesquisa indicaram que mais de 90% dessas páginas desapareceram em um período de 12 anos e a análise da vida útil revelou que a média de vida de uma página da *web* é de 1.132,1 dias ou aproximadamente 37 meses.

O *hiperlink* é a conexão entre o usuário e o recurso *on-line* que ele procura. Dito isso, os percentuais da vida útil dos *links* podem variar conforme o contexto escolhido, o tipo de conteúdo, os responsáveis pela produção e manutenção dos *links* e seus respectivos *site*s, mas podemos observar pelos diversos estudos que a garantia de acesso ao longo do tempo dos conteúdos é bastante frágil e necessita de medidas de preservação digital.

Costa, Gomes e Silva (2016) discutem o arquivamento da *web* e as políticas que regem quais partes da *web* são preservadas. A maioria das iniciativas de arquivamento da *web* escolhem seletivamente os *site*s para preservação com base em fatores como relevância e consentimento dos autores, sendo que muitas iniciativas possuem conteúdo exclusivamente relacionado ao país, região ou instituição que os hospeda. Os autores apontam que o tamanho das coleções arquivadas aumentou desde 2010, com mais iniciativas preservando coleções entre 10 e 100 TB e que o *Internet Archive* é o

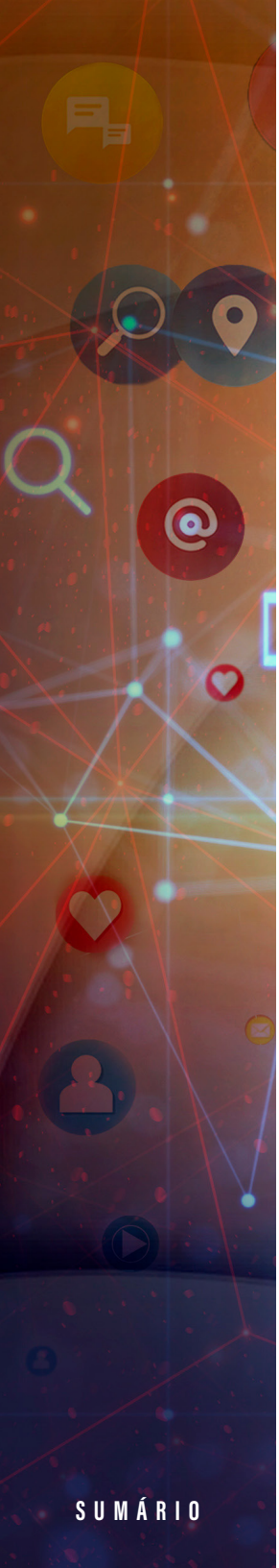


maior arquivo da *web*, no entanto, algumas iniciativas podem replicar o conteúdo arquivado pelo *Internet Archive*.

A cobertura do arquivamento da *web* é a medida em que os materiais baseados na *web* são preservados e disponibilizados por meio do arquivamento da *web*. Desigualdades na cobertura de arquivamento da *web* podem ocorrer quando certos grupos ou regiões estão sub-representados nos esforços de arquivamento da *web*, levando a lacunas na preservação e acessibilidade de materiais baseados na *web*.

Uma das principais causas das desigualdades na cobertura do arquivamento da *web* é a distribuição distinta dos recursos de arquivamento da *web*. O arquivamento da *web* requer recursos específicos, incluindo financiamento, equipe e infraestrutura técnica, e esses recursos muitas vezes não se encontram distribuídos uniformemente entre as diferentes regiões e países. Como resultado, algumas regiões podem ter acesso a mais recursos de arquivamento na *web* do que outras, levando a uma cobertura desigual na captura de materiais baseados na *web*. As políticas de preservação digital, ou a inexistência de políticas, e o quanto se investe nestas atividades tem relação direta com acervos preservados e disponibilizados.

Além disso, as desigualdades na cobertura e preservação também podem ser causadas por barreiras culturais e linguísticas, quando são parte das políticas de captura e preservação digital. Os esforços de arquivamento da *web* podem concentrar-se na preservação de materiais baseados na *web* em determinados idiomas ou contextos culturais, levando a lacunas na preservação e acessibilidade de materiais em outros idiomas ou contextos culturais. Esforços para lidar com essas desigualdades, como aumentar o financiamento e os recursos para arquivamento da *web* e promover o arquivamento da *web* em uma ampla gama de idiomas e contextos culturais, podem ajudar a garantir que os materiais baseados na *web* sejam preservados e acessíveis a todos.



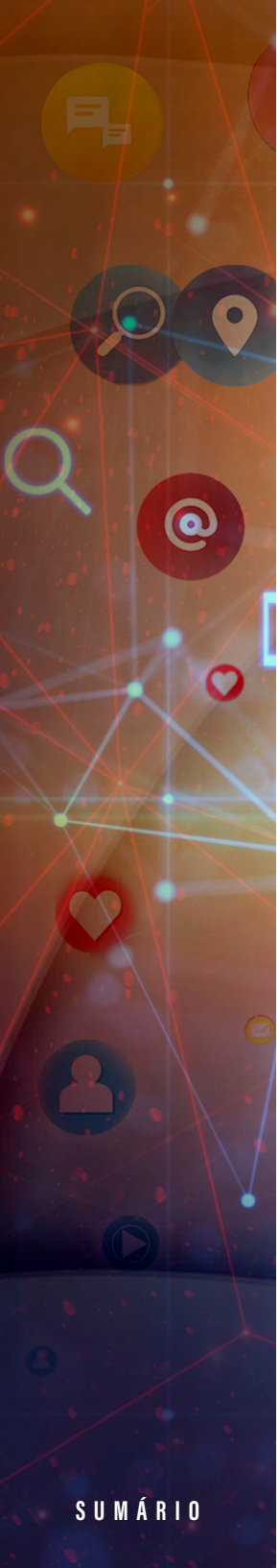
As iniciativas nacionais e outras de alcance mais amplo, como a *Internet Archive*, acabam capturando conteúdo relacionado ao Brasil e aos domínios .br, mas isso não ocorre de uma forma similar com outros países. Há uma grande desigualdade geográfica na preservação de recursos e conteúdo da *web* (LEETARU, 2015; ROCKEMBACH, 2017), devido ao fato de que os domínios e conteúdo do Brasil não estão dentro do escopo principal dessas iniciativas de captura, provocando uma lacuna na cobertura do conteúdo brasileiro de forma sistemática.

O *Internet Archive* não cobre todas as regiões igualmente porque é uma organização sem fins lucrativos que depende de voluntários e doações para coletar e preservar o conteúdo da *web*. Como tal, a cobertura do *Internet Archive* é limitada pelos recursos e esforços disponíveis para a organização.

Além disso, o *Internet Archive* concentra-se na coleta de conteúdo da *web* de *site*s acessíveis publicamente, o que significa que pode não ter acesso ao conteúdo de *site*s privados ou restritos. Isso pode limitar a cobertura do *Internet Archive* em determinadas regiões, pois algumas áreas podem ter uma proporção maior de *site*s privados ou restritos.

Também, o *Internet Archive* pode enfrentar desafios na coleta e preservação de conteúdo da *web* de certas regiões devido a problemas técnicos ou logísticos. Por exemplo, algumas regiões podem ter conectividade ou infraestrutura de Internet limitada, o que pode dificultar a captura e armazenamento de conteúdo da *web* dessas áreas.

Ao fim, é um esforço coletivo e colaborativo. Instituições que colaboram entre si para um propósito comum podem ser uma força poderosa para atingir metas e objetivos compartilhados. A colaboração entre instituições pode assumir várias formas, como parcerias, alianças ou redes, e pode envolver diferentes tipos de instituições, como governos, empresas, organizações sem fins lucrativos ou instituições acadêmicas.

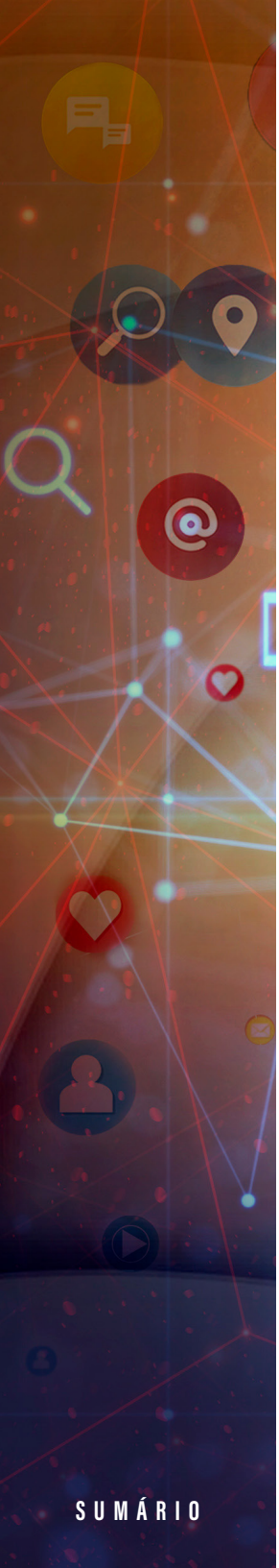


Um dos principais benefícios da colaboração institucional é que ela permite que as organizações reúnam seus recursos, conhecimentos e redes para obter maior impacto e sucesso. Ao trabalharem juntas, as instituições podem alavancar suas forças e capacidades coletivas para enfrentar desafios complexos e oportunidades que podem estar além do escopo de qualquer organização individual.

A colaboração institucional também pode ajudar a fomentar a inovação e a criatividade, pois permite que as organizações aprendam e desenvolvam ideias e abordagens beneficiando-se umas das outras. Isso pode levar ao desenvolvimento de novas soluções e avanços, auxílio para enfrentar desafios e podem promover inovação, criatividade e transparência que não seriam possíveis sem a colaboração.

Além disso, a colaboração institucional pode promover transparência, responsabilidade e inclusão, pois muitas vezes envolve o compartilhamento de informações, dados e recursos entre as organizações participantes, possibilitando construir confiança e apoio entre as instituições e aumentando a credibilidade e eficácia de seus esforços coletivos.

A *web* não pode ser delimitada geograficamente porque é uma rede global de sistemas e dispositivos de computadores interconectados. A *web* não está confinada a um local ou região em particular e pode ser acessada de qualquer lugar do mundo com uma conexão à Internet. Da mesma maneira, o conteúdo da *web* não está necessariamente limitado a um determinado local ou região. *Site*s e páginas da *web* podem ser criados e acessados por qualquer pessoa, em qualquer lugar, e podem conter informações e recursos de todo o mundo. Isso torna difícil delimitar a *web* geograficamente, pois é um fenômeno verdadeiramente global. A *web* está em constante evolução e mudança, com novos conteúdos e recursos sendo adicionados e atualizados o tempo todo. Isso torna ainda mais difícil a delimitação geográfica da *web*, pois as fronteiras estão sempre mudando e movendo-se.



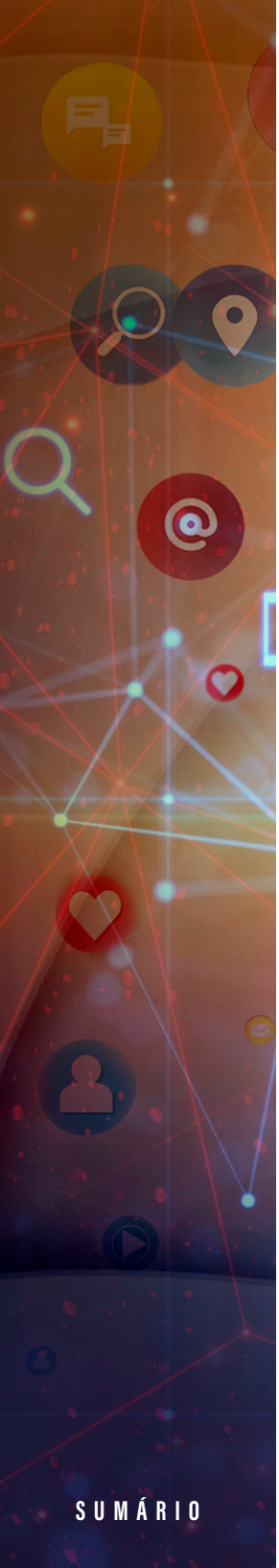
Iniciativas de arquivamento da *web* em nível nacional podem ser empreendimentos complexos e desafiadores, exigindo planejamento e coordenação cuidadosos para serem bem-sucedidos. Algumas das principais considerações para iniciativas nacionais de arquivamento da *web* incluem:

a. escopo e cobertura: o primeiro passo no planejamento de uma iniciativa nacional de arquivamento da *web* é definir o escopo e a cobertura do projeto. Isso envolve decidir qual conteúdo será capturado e preservado e como ele será selecionado e coletado. É importante garantir que o escopo do projeto seja realista e alcançável, dados os recursos e capacidades disponíveis;

b. infraestrutura técnica: outra consideração importante para as iniciativas nacionais de arquivamento da *web* é a infraestrutura técnica que será usada para capturar, armazenar e gerenciar o conteúdo arquivado. Isso pode envolver a construção ou aquisição de *hardware* e *software* especializados, bem como o desenvolvimento e implementação de políticas e procedimentos para acesso e gerenciamento de dados;

c. questões legais e éticas: as iniciativas de arquivamento da *web* em nível nacional devem levantar questões legais e éticas, como questões de privacidade e propriedade intelectual. É importante considerar e abordar cuidadosamente essas questões para garantir que o projeto esteja em conformidade com as leis e regulamentos do país e que respeite os direitos e interesses de indivíduos e organizações e

d. envolvimento das partes interessadas: as iniciativas nacionais de arquivamento da *web* geralmente envolvem uma ampla gama de partes interessadas, incluindo agências governamentais, bibliotecas, arquivos e outras organizações. É importante envolver essas instâncias no planejamento e implementação do projeto, a fim de garantir que suas



necessidades e preocupações sejam atendidas e que possam contribuir para o sucesso da iniciativa.

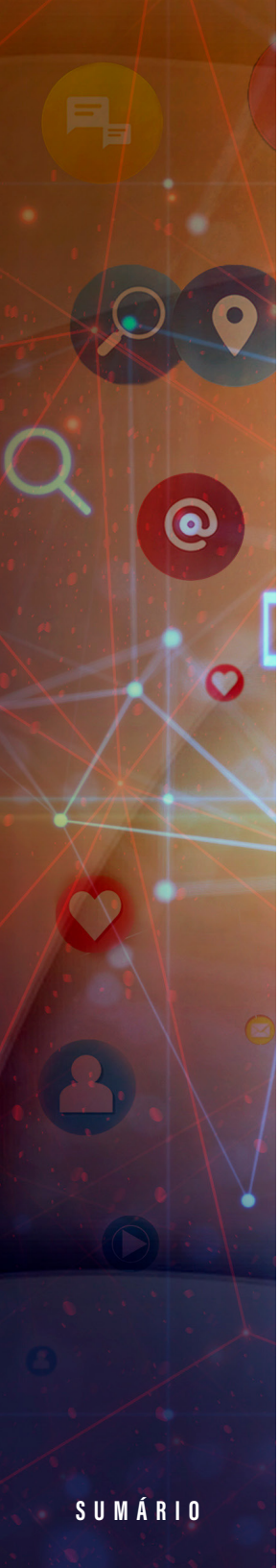
As iniciativas nacionais de arquivamento da *web* exigem planejamento e coordenação cuidadosos para garantir que sejam bem-sucedidas e eficazes. Ao considerar essas e outras questões-chave, é possível desenvolver e implementar iniciativas de arquivamento da *web* que fornecem benefícios de longo prazo para a preservação e acesso ao conteúdo da *web*.

Existem muitas razões pelas quais as iniciativas locais de arquivamento da *web* podem ser benéficas. Alguns dos principais benefícios da criação de iniciativas locais de arquivamento da *web* incluem:

a. preservar a história e a cultura locais: as iniciativas locais de arquivamento da *web* podem ajudar a capturar e preservar a história e a cultura de uma determinada região ou comunidade. Isso envolve a captura e armazenamento de páginas da *web* e conteúdo que documentam eventos, organizações e personalidades locais, bem como a preservação de recursos e materiais *on-line* exclusivos de um determinado local;

b. fornecer acesso a informações locais: as iniciativas locais de arquivamento da *web* também podem fornecer acesso a informações e recursos relevantes para um local específico. Ao coletar e organizar o conteúdo da *web* de fontes locais, é possível criar um repositório de informações pesquisável e acessível que pode ser usado por pesquisadores, educadores e outros interessados em uma determinada região ou comunidade;

c. apoiar instituições locais: iniciativas locais de arquivamento da *web* também podem apoiar instituições como bibliotecas, arquivos e museus, fornecendo-lhes acesso a conteúdo da *web* que seja relevante para suas coleções e missão. Isso vai



ao encontro de ajudar a melhorar os serviços e recursos que essas instituições podem oferecer às suas comunidades e

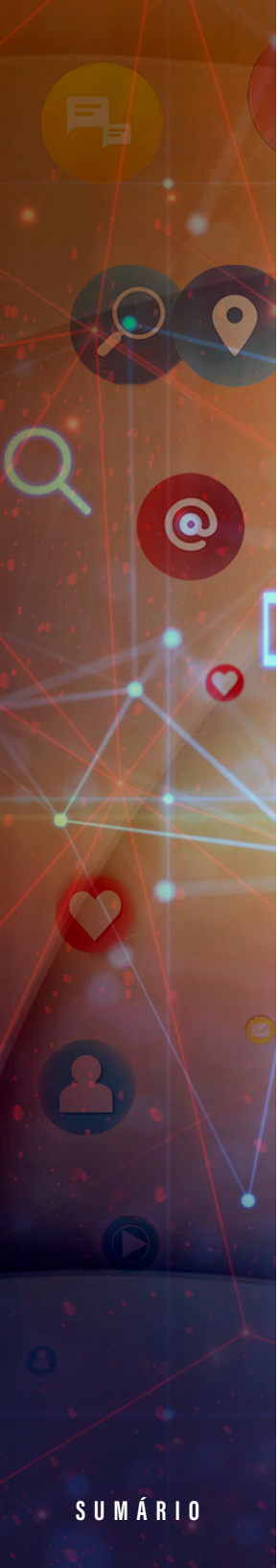
d. melhorar o envolvimento da comunidade: as iniciativas locais de arquivamento da *web* também podem promover o envolvimento da comunidade ao envolver as partes interessadas locais no planejamento e implementação do projeto, para identificar e coletar conteúdo da *web*, bem como oferecer oportunidades para que os membros da comunidade participem da preservação e acesso ao conteúdo arquivado.

As iniciativas locais de arquivamento da *web* podem oferecer muitos benefícios ao capturar e preservar a história e a cultura de uma determinada região ou comunidade e ao fornecer acesso a informações e recursos locais. Ao criar e apoiar essas iniciativas, é possível aumentar a preservação e a acessibilidade do conteúdo da *web* em nível local.

Segundo Dougherty e Meyer (2014), um dos maiores obstáculos ao arquivamento da *web* em instituições é a percepção de que não é relevante o suficiente para investir recursos. Algumas pessoas precisam se esforçar para compreender o ponto de arquivamento da *web*, e há uma necessidade de uma "mudança de mentalidade institucional". Esse obstáculo é maior quando os usuários não compreendem que os conteúdos da *web* ao vivo não estarão sempre disponíveis e não percebem os riscos de não ter mais acesso à informação.

Levando em consideração os desafios de preservar o conteúdo das redes sociais, Hockx-Yu (2014) afirma que a tarefa é muito grande e complexa para instituições individuais realizarem sozinhas e que meios colaborativos com outras organizações e com cidadãos, por meio de *crowdsourcing*, como exemplificado no projeto *Boston Marathon Archive*[62], pode ser benéfico para o arquivamento da *web*. Embora os arquivos e as bibliotecas nacionais devam continuar

62 http://marathon.neu.edu/



desenvolvendo soluções, é importante perceber que os limites geográficos são arbitrários na *web*, e há uma expectativa crescente dos pesquisadores de que as instituições de memória captem e disponibilizem acesso ao universo digital.

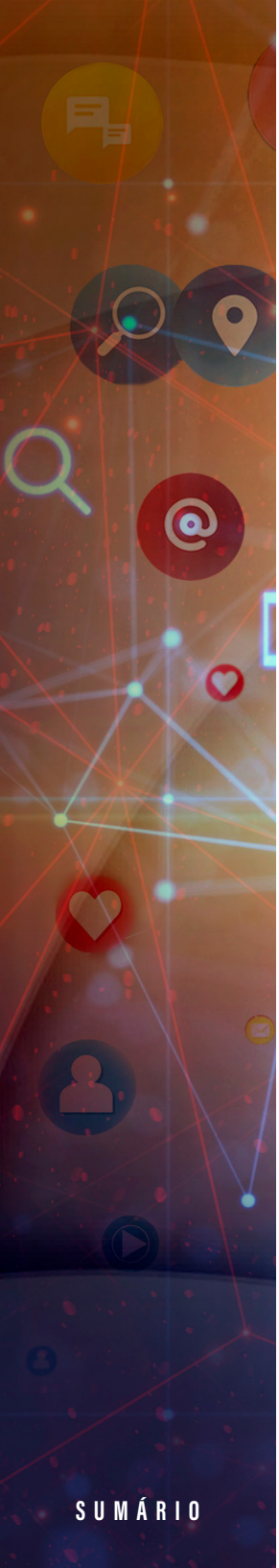
O arquivamento da *web*, ainda, pode ser considerado uma ferramenta de ativismo social e ativismo digital, visando preservar informações e conteúdos *on-line* para fins históricos e de acesso público. Alguns exemplos deste ativismo digital podem ser vistos com o *Archive Team*[63] ou com o projeto Salvando o patrimônio cultural ucraniano *on-line*, *Saving Ukrainian Cultural Heritage On-line* (SUCHO), como já colocado anteriormente.

O ativismo digital, como o desenvolvido pelo *Archive Team*, prioriza a preservação e o acesso, visando garantir que conteúdos em risco não sejam perdidos, o que algumas vezes pode divergir da visão de direitos autorais e consentimento dos produtores. Desta forma, a abordagem do *Archive Team* é vista como um ato político que cria tensões e dilemas éticos (OGDEN, 2020, 2022) e torna-se mais alinhada com a ética *hacker* e seus princípios: compartilhamento, abertura, descentralização, livre acesso aos computadores e melhoria do mundo (LEVY, 1984).

Por outro lado, corporações e plataformas que descontinuam seus serviços *on-line* por não serem mais lucrativos, geram muitas vezes a perda de conteúdo produzido pelo usuário, que pode ter valor tanto para o produtor da informação, quanto para futuros pesquisadores. Isto sem mencionar governos que não priorizam a preservação digital como uma importante ferramenta para garantir acesso à informação ao longo do tempo.

Ainda, como argumenta Webster (2017), tanto o *Archive Team*, quanto outros projetos com a mesma perspectiva, representam uma abordagem mais semelhante à do *Internet Archive* do que

63 https://wiki.archiveteam.org/

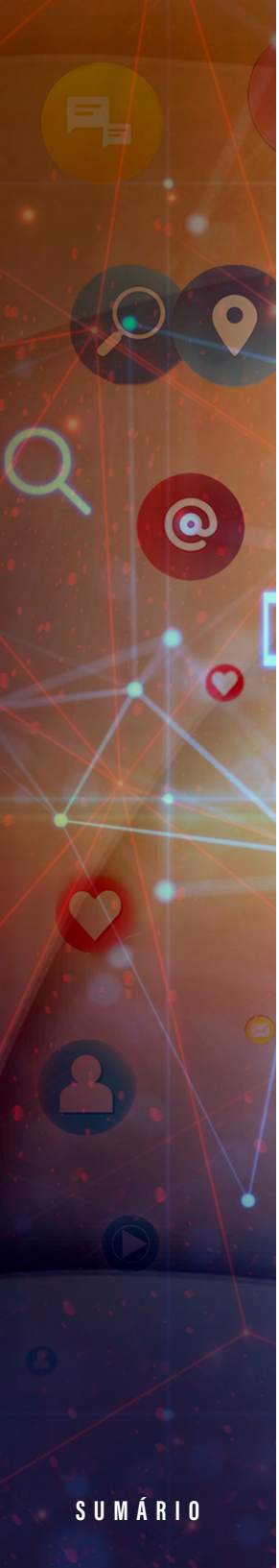


de diversas instituições nacionais, pois torna-se necessário agir rapidamente para preservar um conteúdo que não seria arquivado pelos sistemas institucionais existentes. Essa abordagem foi pragmática e caracterizada pela disposição de preservar e arquivar conteúdo, apesar dos riscos de violação de direitos autorais — riscos que as instituições, por natureza, geralmente evitam. Segundo Webster (2017), estes projetos foram impulsionados por um senso de dever público e por uma visão política e social específica do tipo de espaço que a *web* deve ser e representam uma resposta à nova configuração de interessados após a *web* 2.0: editores, usuários que criam conteúdo e arquivistas que se dedicam a documentar este relacionamento e, em alguns casos, a equilibrar o poder entre eles, dinâmica de interesses que difere da relação binária entre bibliotecas e editores, ou de arquivos e produtores, que historicamente moldou, respectivamente, o desenvolvimento do depósito legal (LARIVIÉRE, 2000) e do recolhimento de documentos.

Com o aumento da importância da Internet na sociedade, muitos ativistas perceberam a necessidade de registrar e preservar o conteúdo que pode ser removido ou censurado pelos governos ou empresas. O arquivamento da *web* é especialmente importante para preservar informações e discursos de grupos marginalizados e minorias, que muitas vezes são invisibilizados ou apagados da história oficial. Por meio do arquivamento da *web*, é possível garantir que as informações sejam preservadas para as futuras gerações e para a construção de uma sociedade mais justa e igualitária.

ARQUIVAMENTO DA *WEB* E COLEÇÕES

As coleções de arquivamento da *web* são uma ferramenta importante para preservar e acessar a história da Internet. Um arquivo da *web* é uma coleção de páginas da *web* que foram salvas

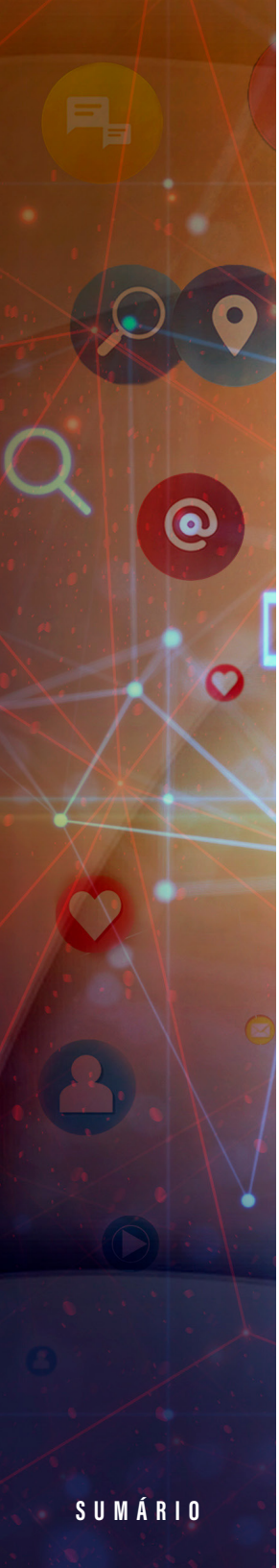


e preservadas, geralmente por uma biblioteca ou outra instituição, para referência futura. Essas coleções são importantes porque nos permitem olhar para o desenvolvimento da Internet e as informações que estavam disponíveis em diferentes momentos.

O arquivamento da *web*, na visão de Ogden (2022), é uma força transformadora que requer atenção das equipes de preservação no que diz respeito às dimensões culturais da prática que fundamentam as decisões cotidianas sobre como a *web* é "salva". Ainda, é necessária uma neutralidade geral ao decidir o valor de preservação de certos *sites* da *web* em detrimento de outros. O arquivamento da *web* deve ser representativo da diversidade da experiência *on-line* para que no futuro tenhamos uma boa visão de como era o mundo. Cada uma dessas observações de Ogden (2022) estão enraizadas nos valores de neutralidade e objetividade do arquivamento da *web* e no sentido de trabalhar para fornecer o contexto para a equipe de arquivamento e qual o seu papel no salvamento da *web*.

Um dos principais benefícios das coleções de arquivamento da *web* é que elas fornecem um registro da Internet como ela existia em um determinado momento. Isso pode ser útil para diversos propósitos, como pesquisa histórica, estudo do desenvolvimento de comunidades *on-line* e rastreamento da disseminação de informações e ideias. Ao preservar as páginas da *web* em sua forma original, as coleções de arquivamento da *web* nos permitem ver como a Internet mudou ao longo do tempo e como ela foi usada.

Outro benefício das coleções de arquivamento da *web* é que elas podem fornecer acesso a informações que podem não estar mais disponíveis na *web* ao vivo. Por exemplo, uma página da *web* que foi retirada do ar ou um *site* que ficou *off-line* ainda pode ser acessado por meio de um arquivo da *web*, podendo ser especialmente útil para pesquisadores e outras pessoas que precisam acessar informações que podem não estar mais disponíveis em outro lugar.

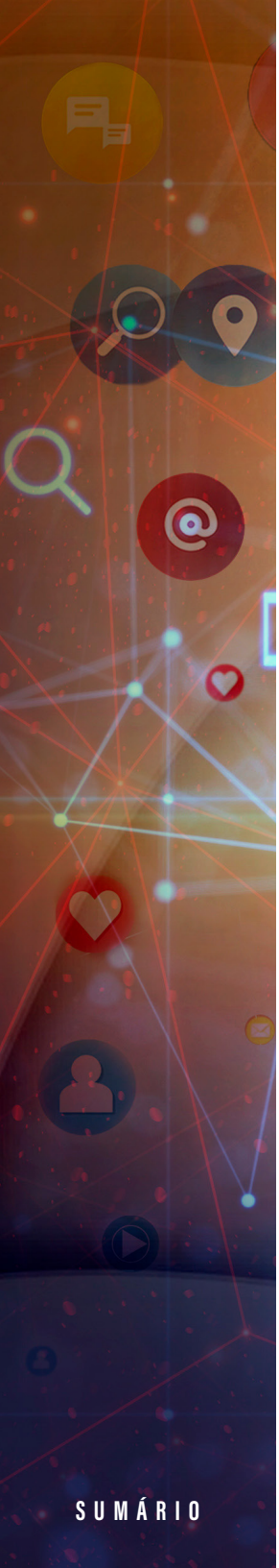


Além desses benefícios, as coleções de arquivamento da *web* também podem ajudar a preservar o patrimônio cultural da Internet. A Internet tornou-se uma parte importante de nossas vidas cotidianas e um repositório significativo de conhecimento e criatividade humanos. Ao preservar as páginas da *web*, as coleções de arquivamento da *web* ajudam a garantir que esse conhecimento e criatividade não sejam perdidos.

As coleções de arquivamento da *web* são uma ferramenta importante para preservar e acessar o histórico da Internet. Elas fornecem um registro da Internet em diferentes pontos no tempo, podem fornecer acesso a informações que não estão mais disponíveis na *web* ao vivo e ajudam a preservar o patrimônio cultural da Internet. Ao apoiar e manter coleções de arquivamento da *web*, podemos ajudar a garantir que a história da Internet seja preservada e acessível futuramente.

As práticas de arquivamento manifestam-se primeiramente por meio das decisões sobre o que arquivar. Apesar da sua posição idealista em relação à seletividade, selecionar significa tomar decisões relativas ao que é arquivado e essas decisões dependem, entre outras coisas, de tempo e infraestruturas disponíveis (largura de banda, IPs), os juízos de valor e prioridades de uma combinação dinâmica de intervenientes e as possibilidades técnicas das plataformas utilizadas.

O arquivamento da *web* e as coleções temáticas são ferramentas importantes para preservar e acessar informações sobre tópicos ou eventos específicos. Uma coleção temática é um conjunto de páginas da *web* com curadoria que se concentra em um determinado assunto ou evento, como a pandemia de Covid-19, mudanças climáticas ou eleições. Essas coleções podem ser criadas por bibliotecas, arquivos ou outras instituições e geralmente são organizadas em torno de um tema ou tópico específico.

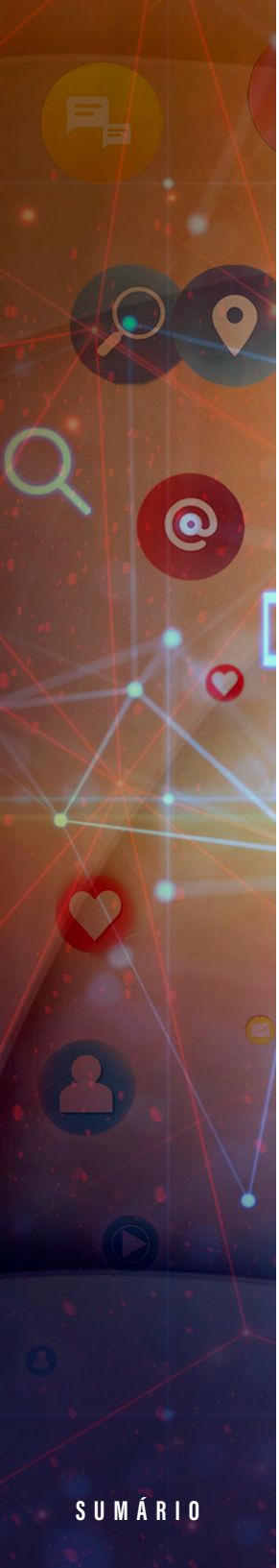


As coleções temáticas podem ser úteis para uma variedade de propósitos, como pesquisa, educação e preservação histórica. Por exemplo, uma coleção temática sobre a pandemia de Covid-19 pode incluir páginas da *web* sobre o próprio vírus, bem como informações sobre seu impacto na sociedade, medidas de saúde pública e outros tópicos relacionados. Esta coleção pode ser útil para pesquisadores que estudam a pandemia, bem como para educadores e estudantes que desejam aprender mais sobre o tema.

No contexto de uma economia global em transformação, cujos líderes estão alterando radicalmente suas posições, a mudança climática tornou-se um tema central e uma questão de sobrevivência tanto para o planeta quanto para o desenvolvimento de nossa sociedade. Seu impacto na política, economia e saúde faz da mudança climática uma das principais preocupações da humanidade – aquela que pode trazer um mundo novo e melhor se decidirmos implementar as transformações necessárias para resolvê-la. Os interesses econômicos e políticos têm frequentemente interferido nos discursos dos cientistas e em outras ações que revelam informações pertinentes para o público em geral. O acesso e a transparência das informações também desempenham um papel importante neste cenário. (ROCKEMBACH; SERRANO, 2021)

A quantidade de informação produzida, armazenada e disponível sobre mudança climática é enorme e não se restringe apenas a recursos governamentais e de pesquisa, é necessário recuperar informações dos arquivos digitais disponíveis na Internet em diversas plataformas. Assim, essas coleções podem ser úteis para pesquisadores que estudam o assunto, bem como para formuladores de políticas e outros que trabalham para lidar com as questões climáticas.

As coleções temáticas também podem ser válidas para acompanhar o desenvolvimento de determinado tópico ou evento ao longo do tempo. Por exemplo, uma coleção sobre eleições pode incluir páginas da *web* de diferentes ciclos eleitorais, fornecendo um registro de como as campanhas e os candidatos evoluíram ao longo



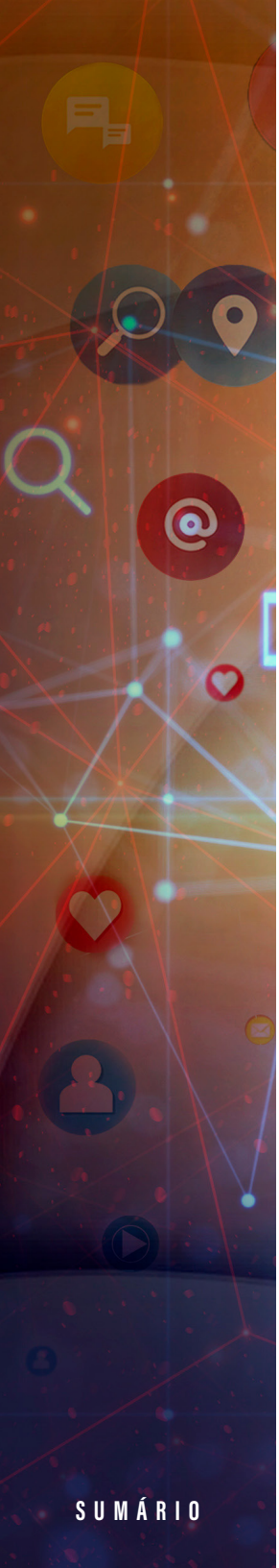
do tempo. Esse tipo de coleção pode ser aproveitável para pesquisadores que estudam o desenvolvimento da política eleitoral, bem como para jornalistas e outros interessados no tema.

De maneira geral, o arquivamento da *web* e as coleções temáticas são ferramentas importantes para preservar e acessar informações sobre tópicos ou eventos específicos. Essas coleções podem fornecer recursos diferenciados para pesquisadores, educadores e outras pessoas interessadas em um determinado assunto ou evento e podem ajudar a garantir que informações importantes não sejam perdidas. Ao apoiar e manter o arquivamento da *web* e as coleções temáticas, podemos ajudar a garantir que essas relevantes informações sejam preservadas para as gerações futuras.

O arquivamento da *web* é uma ferramenta proveitosa para arquivistas, bibliotecários e humanistas documentarem narrativas da pandemia de Covid-19 e protestos como o *Black Lives Matter*. Ao capturar e preservar o conteúdo da *web*, esses profissionais podem garantir que as histórias e experiências dos indivíduos durante esses eventos não sejam perdidas.

O arquivamento da *web* pode ser utilizado para documentar as narrativas da pandemia de Covid-19 coletando páginas da *web* relacionadas à pandemia, como artigos de notícias, *blogs* pessoais e postagens em mídias sociais. Esse conteúdo pode fornecer informações sobre as experiências dos indivíduos durante a pandemia, incluindo seus pensamentos, sentimentos e ações. Ao preservar esse conteúdo, futuros pesquisadores poderão acessar e estudar essas narrativas para entender melhor a pandemia e seu impacto na sociedade.

A coleta de páginas da *web* relacionadas aos protestos *Black Lives Matter*, como artigos de notícias, postagens em mídias sociais, gravações de áudio e vídeo, fotografias e outros conteúdos multimídia, podem fornecer um registro abrangente desses eventos que podem ser utilizados para estudar as experiências dos indivíduos



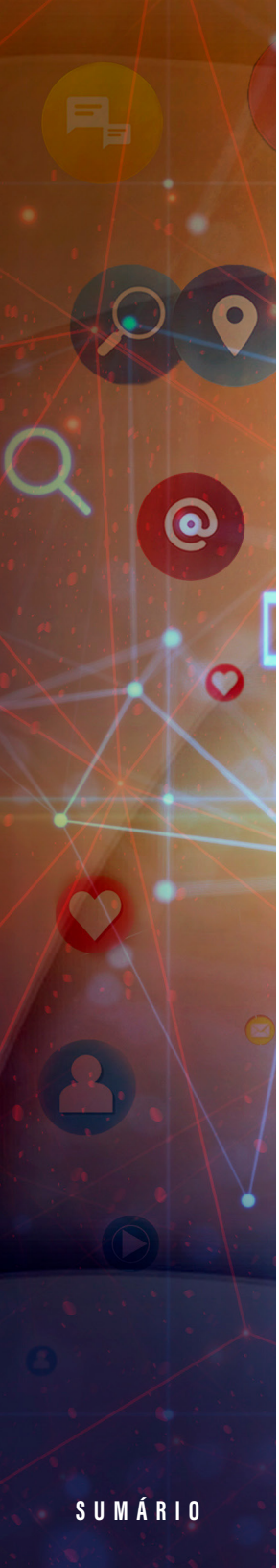
envolvidos nos protestos, bem como o contexto social e político mais amplo em que os protestos ocorreram.

ARQUIVAMENTO DA *WEB* INSTITUCIONAL

O arquivamento institucional da *web* refere-se ao processo de coletar, preservar e fornecer acesso a informações da Internet relacionadas a uma organização específica, como empresa privada ou órgão governamental. O arquivamento institucional da *web* é importante porque permite que as instituições preservem e forneçam acesso a informações valiosas da Internet relacionadas às suas atividades e missão.

Os *site*s da *web* podem ser considerados documentos de arquivo por diversas razões. Em primeiro lugar, eles podem conter informações orgânicas e contextuais que são criadas e publicadas *on-line*, muitas vezes com o objetivo de serem acessadas e consultadas por um público amplo e variado. A organicidade representa as funções, estrutura e atividades da entidade produtora, além das relações, sejam internas ou externas. Essas informações podem ser relevantes para a história, cultura e desenvolvimento de uma determinada sociedade, comunidade ou organização, e, portanto, consideradas como registros que precisam ser preservados. Muitas vezes, a informação produzida em um *site* ou plataforma, incluindo as redes sociais, só existe neste lugar e suporte, portanto, caso haja o apagamento desta informação, não há a possibilidade de recuperação.

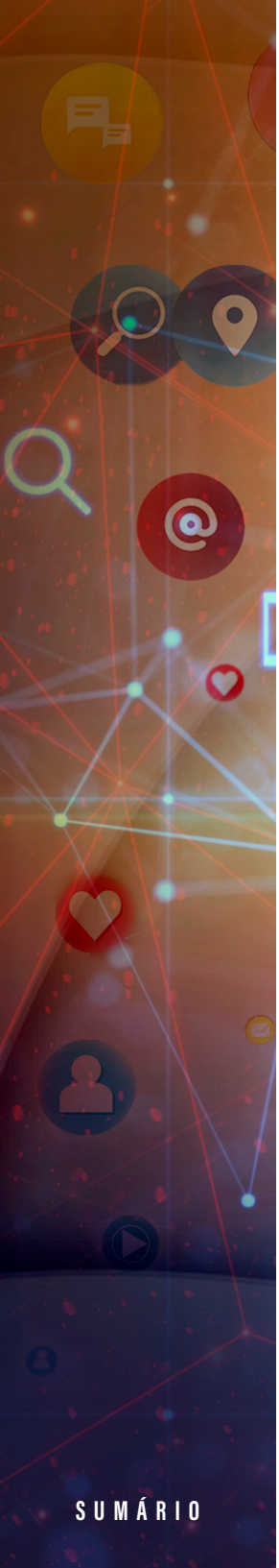
Além disso, os *website*s muitas vezes são criados com a intenção de serem permanentes e acessíveis a longo prazo. Eles são armazenados em servidores e precisam ser conservados por instituições especializadas, como arquivos e bibliotecas, para fins de preservação histórica. Essas instituições têm a responsabilidade de coletar, preservar e disponibilizar esses materiais para pesquisadores



e o público em geral. Eles ainda podem ser utilizados como fontes de evidência em processos legais e administrativos, demonstrando quando uma página *web* ou de rede social esteve *on-line*, a partir do carimbo de tempo (*timestamp*) do momento da captura pelo arquivo da *web*. Portanto, é importante reconhecer a importância dos *websites* como documentos de arquivo e garantir que eles sejam coletados e preservados de forma adequada para que possam ser consultados e utilizados como fontes confiáveis de informação no futuro.

Taylor (2017) argumenta que os tribunais normalmente veem a Internet como um grande catalisador de rumores, insinuações e desinformação e, portanto, provas obtidas na Internet não seriam adequadas. No entanto, os litigantes ainda usam evidências do *Internet Archive* e do *Wayback Machine*, e os tribunais geralmente as aceitam com uma declaração juramentada, testemunho ou por meio de notificação judicial. Os profissionais jurídicos usam evidências dos arquivos da *web* para obter dados históricos exclusivos, ampliar a comunidade de prática e ajudar tribunais e júris a interpretar corretamente as evidências de arquivo da *web*. A autenticação é necessária para satisfazer o requisito de autenticação ou identificação de um item de evidência. Alguns casos que usaram evidências do *Internet Archive* incluem violação de marca registrada, quebra de contrato e *cybersquatting*.

Taylor (2017) ainda cita casos judiciais em que o *Internet Archive* e o *Wayback Machine* foram utilizados para mostrar a presença ou ausência histórica de certas informações em um *site* em um determinado momento. Os casos envolvem questões relacionadas a propaganda enganosa, defesa do consumidor, violação de direitos autorais, violação de marca registrada e difamação. A aceitação do tribunal dos arquivos da *web* como evidência é baseada em fatores como confiabilidade, coerência temporal, proveniência e canonicidade. A precisão do conteúdo da *Wayback Machine* é questionada em alguns casos, quando não foi possível realizar a captura devido a problemas tecnológicos, mas geralmente é considerada um recurso válido como evidência e para a notificação judicial.

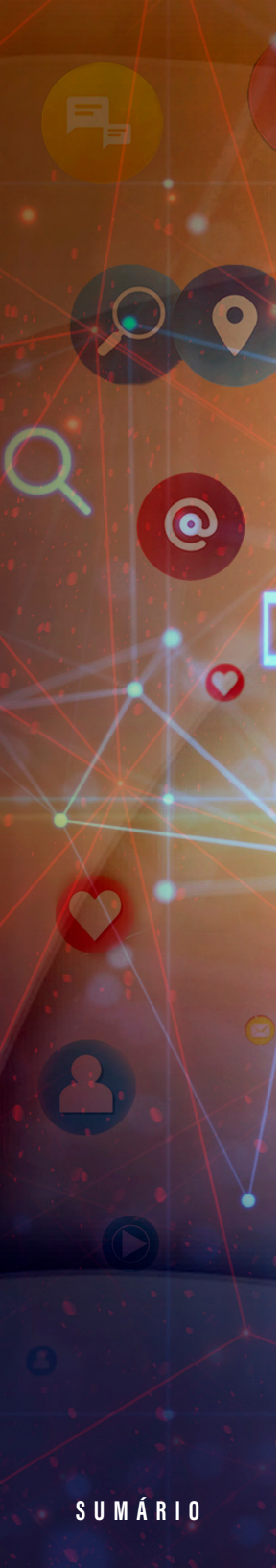


A Carta para a Preservação do Patrimônio Digital da UNESCO (UNESCO, 2003) recomenda que os Estados-Membros, contexto em que se inclui o Brasil, posicionem-se quanto a políticas e ações no sentido da preservação digital, além de reconhecer a efemeridade do digital, representada em textos, imagens, áudio, *software* e páginas da *web*.

Além da Carta da UNESCO para a Preservação do Patrimônio Arquivístico Digital (CONARQ, 2005) e da Política de Preservação Digital do Arquivo Nacional (ARQUIVO NACIONAL, 2016), diversos autores também defendem o entendimento dos *website*s institucionais como documentos de arquivo (MELO, 2020; SANTOS, 2020; MELO; ROCKEMBACH, 2021a) e, portanto, precisam ser preservados e disponibilizados ao público. Além disso, de acordo com Ferreira, Martins e Rockembach (2018), os arquivos da *web* constituem tanto um potencial informacional, com a constituição de uma memória virtual, quanto um potencial probatório, como validação como fonte de evidência ao longo do tempo.

O processo de arquivamento institucional da *web* geralmente envolve a identificação de conteúdo relevante, coleta de conteúdo da *web*, armazenamento de maneira segura e acesso ao conteúdo de maneira amigável. Para garantir que o processo de arquivamento da *web* seja eficaz, as instituições podem usar uma variedade de ferramentas e técnicas, como rastreadores da *web* para identificar conteúdo relevante, criptografia de dados para proteger o conteúdo coletado e portais da *web* para fornecer acesso ao conteúdo arquivado.

É importante ressaltar que uma política a nível institucional deve ser adotada levando em consideração diretrizes nacionais, quando houver. Essas diretrizes devem conter disposições relativas a questões relacionadas com o formato, armazenamento e *backup*, refrescagem e emulação, acesso, descrição de arquivo, responsabilidades, planejamento em caso de catástrofes, utilização da tecnologia,

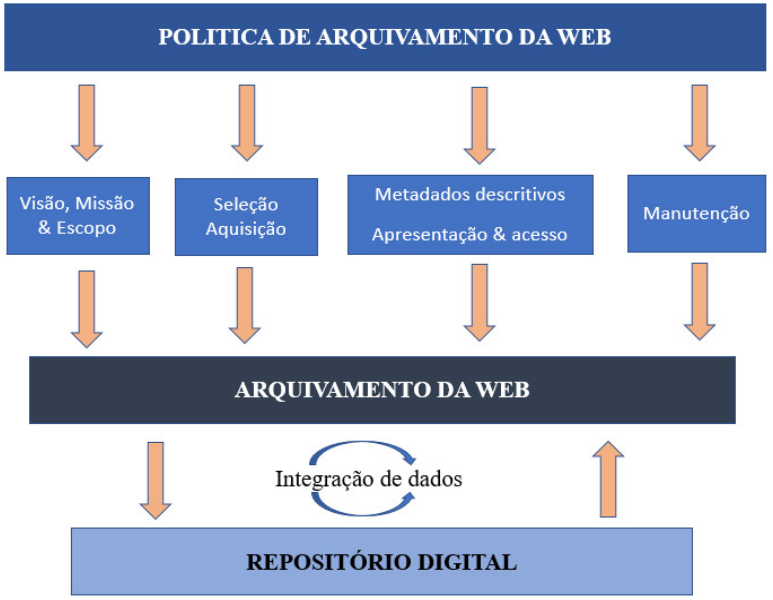


planos de contingência, financiamento, propriedade, direitos de propriedade intelectual e direitos de autor, entre outros temas relevantes.

O arquivamento institucional da *web* é importante por vários motivos. Ele permite que as instituições preservem e forneçam acesso a informações valiosas da Internet relacionadas à sua atuação, funcionamento, compromissos e responsabilidades. Também ajuda a garantir que a herança cultural da instituição na Internet seja preservada ao longo do tempo. Além disso, o arquivamento institucional da *web* pode apoiar a pesquisa, ensino e o público em geral, fornecendo acesso a informações da Internet que podem não estar disponíveis por outros meios.

Geralmente, as instituições preservam, arquivam e dão acesso aos seus *site*s arquivados a partir dos seus repositórios digitais institucionais baseando suas opções e estratégias de arquivamento nas políticas nacionais, internacionais, no *web Archiving Life Cycle Model*, no *Open Archival Information System* (OAIS) e padrões de repositórios digitais confiáveis. A Figura 1 representa um modelo de política de arquivamento da *web* adaptado de Balogun e Kalusopa (2022), cobrindo vários aspectos indispensáveis, como: declaração de missão e âmbito, seleção, aquisição, metadados descritivos, apresentação e acesso, manutenção, pessoal e formação e colaboração.

Figura 16 - Modelo de política de preservação da *web* institucional

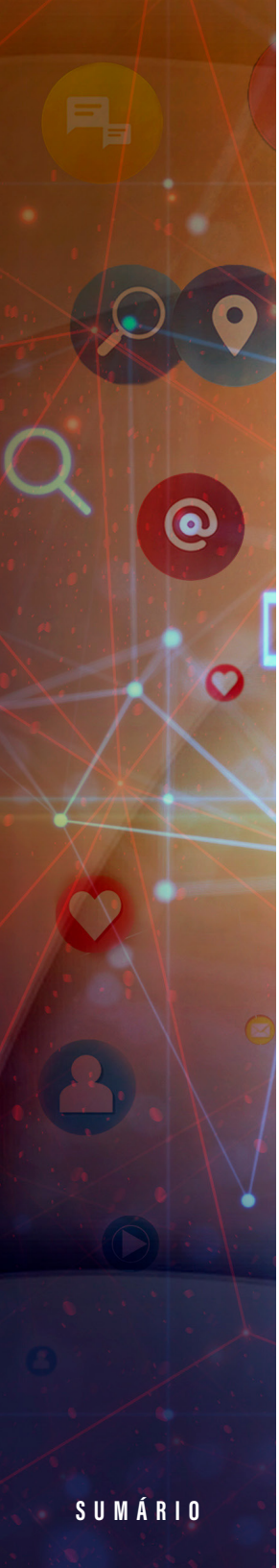


Fonte: Adaptado de Balogun e Kalusopa (2021, p. 8).

A política de arquivamento da *web* deve delinear a missão do projeto de preservação e o seu alcance para orientar a captura, identificar os grupos de utilizadores e os *site*s que serão arquivados.

A política de arquivamento da *web* deve delimitar claramente o que deve ser selecionado para arquivamento. Os critérios de seleção devem ser precisos visto que é um componente chave do arquivamento da *web*. Os *site*s institucionais podem adotar o arquivamento em massa, enquanto os repositórios institucionais com diversos materiais acadêmicos podem adotar o arquivamento seletivo.

No que se refere à aquisição, a política de arquivamento da *web* deve abranger o âmbito de captura, frequência de captura, tipo e formato de material e questões relacionadas com direitos. Nesta fase devemos tomar decisões sobre a captura de dados sensíveis e



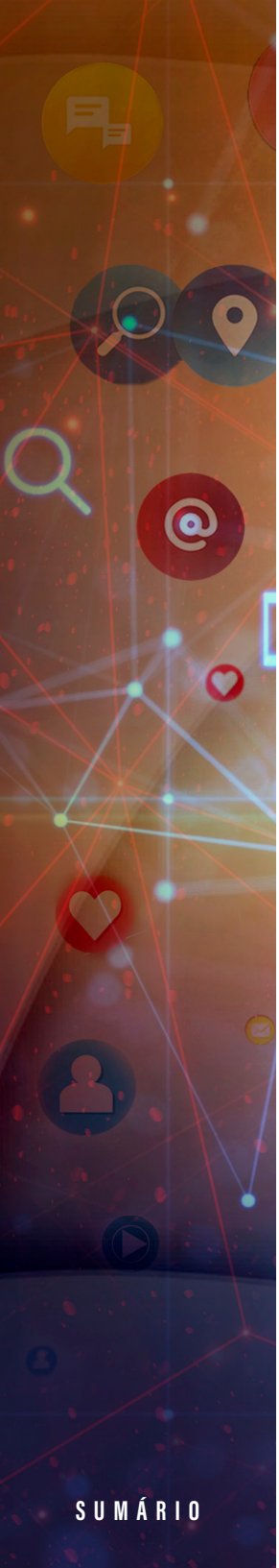
não sensíveis. Quanto à frequência de captura é recomendável que os *website*s sejam capturados com a mesma frequência com que costumam ser atualizados. No que diz respeito ao tipo e formato dos materiais, o formato dos *website*s capturados deve ser preservado tanto quanto possível. A gestão de direitos exige a necessidade de determinar o(s) grupo(s) de pessoas que podem ter permissão para utilizar a *web* arquivada. As questões dos direitos de autor e a propriedade devem ser esclarecidas.

Recomenda-se a utilização de metadados descritivos que atendam padrões internacionais de descrição para os arquivos da *web* como, por exemplo, os metadados descritivos propostos pela *OCLC Research Library Partnership web Archiving Metadata Working Group*[64], que utiliza o padrão *Dublin Core* com seus 15 elementos centrais divididos em três (3) grupos principais: Conteúdo (título, descrição, assunto, tipo, relações, fonte, e cobertura), Propriedade Intelectual (criador, contribuidor, editor, e direitos) e Instanciação (data, formato, identificador, e língua).

A política de arquivamento da *web* deve assegurar, na medida do possível, que os *site*s arquivados sejam idênticos aos do *site* original, exceto quando for necessário excluir conteúdo devido aos direitos autorais envolvidos ou preocupações com a proteção da privacidade. O acesso é uma parte fundamental de qualquer iniciativa de preservação da *web* e envolve a tomada de decisões sobre se e como será facultado o acesso às coleções e o controle da utilização do conteúdo.

A manutenção ou monitoramento periódico oferece uma oportunidade de avaliar o projeto de arquivamento da *web* e melhorá-lo em determinadas áreas, quando necessário. Esta seção da política deve apresentar as questões relativas à manutenção do arquivamento da *web*, é necessário revisitar os *site*s capturados para assegurar a manutenção contínua e periódica do conteúdo a ser arquivado.

64 https://www.oclc.org/research/areas/research-collections/wam.html



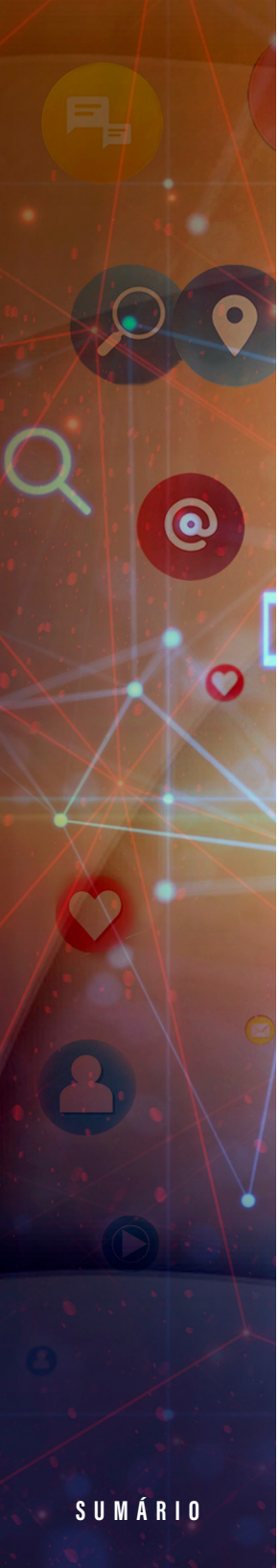
Ainda, a política de arquivamento da *web* deve abranger questões de pessoal e formação. Existe a necessidade de empregar arquivistas e bibliotecários dedicados ao projeto de arquivamento da *web*, que devem receber formação para os capacitar para atuar no arquivamento da *web*, mantendo-os atualizados em termos dos avanços tecnológicos ou tendências atuais no domínio da preservação digital e arquivamento da *web*. Esses profissionais devem trabalhar em estreita colaboração com as TIC selecionadas e com o pessoal de outras instituições para alcançar o objetivo maior de arquivamento da *web* nacional. A colaboração é um elemento chave no arquivamento da *web* como evidenciam as iniciativas colaborativas empreendidas pelo *International Internet Preservation Consortium* (IIPC) e o *National Digital Stewardship Alliance* (NDSA)[65].

O arquivamento institucional da *web* é a prática de preservar e acessar páginas da *web* produzidas por uma organização específica, como uma agência governamental, universidade ou corporação. Essa prática é importante porque permite que as organizações mantenham uma "memória organizacional" de sua presença na *web*, fornecendo um registro das informações e atividades que ocorreram em seus *site*s ao longo do tempo.

Um dos princípios fundamentais do arquivamento institucional da *web* é o conceito de proveniência, que se refere à origem e história de um determinado item ou coleção. No contexto do arquivamento da *web*, a proveniência refere-se ao fato de que uma página da *web* foi preservada em sua forma original, incluindo sua URL original e quaisquer outros metadados relevantes. Isso permite que pesquisadores e outras pessoas interessadas rastreiem as origens das informações na página e entendam seu contexto.

Outro princípio importante do arquivamento institucional da *web* é a ideia de preservação. Para garantir que as páginas da *web* possam ser acessadas e usadas no futuro, elas devem ser

65 https://ndsa.org//

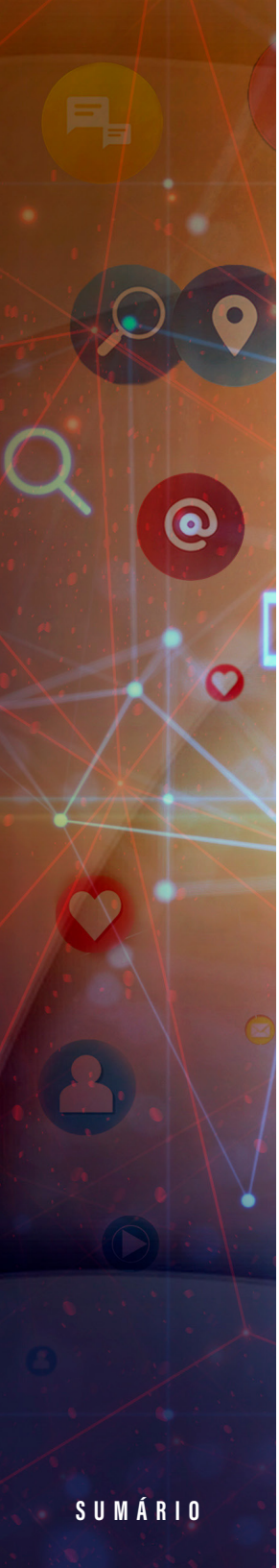


preservadas em um formato estável e acessível. Isso pode envolver uma variedade de atividades, como migrar páginas da *web* para novos formatos, armazená-las em ambientes seguros e fornecer acesso a elas por meio de sistemas especializados.

Além desses princípios, o arquivamento institucional da *web* também cobre uma série de outros, como seleção. Esses princípios envolvem decisões sobre quais páginas da *web* devem ser preservadas, como organizá-las e como fornecer acesso a elas. Essas decisões são normalmente tomadas por arquivistas e/ou bibliotecários que possuem conhecimento especializado sobre a organização e sua presença na *web*, bem como sobre os princípios e práticas de arquivamento da *web*.

O arquivamento institucional da *web* é uma prática importante para preservar, acessar e identificar a presença na *web* de uma organização específica. Ao seguir princípios como proveniência, preservação e outros princípios de arquivamento, as organizações podem manter uma "memória organizacional" de sua presença na *web* e garantir que ela seja preservada para as gerações futuras.

Ainda, segundo Luz (2021), a falta de regulamentação legal no Brasil para preservar arquivos governamentais que não são considerados "legais" no direito à informação pode ser considerado um problema no acesso. O foco da literatura sobre acesso à informação se dá principalmente nos documentos administrativos, e é importante ampliar a compreensão sobre quais informações governamentais devem ser preservadas e tornadas acessíveis. Os *site*s oficiais são atualmente lugares privilegiados para produzir, armazenar e compartilhar conteúdo, sendo urgente um debate sobre a preservação de informações governamentais, já que a Internet facilita tanto a disseminação quanto a exclusão dessas informações.



POLÍTICAS GOVERNAMENTAIS E INSTITUCIONAIS PARA A PRESERVAÇÃO DA *WEB*

O arquivamento da *web* e as políticas institucionais de preservação digital estão intimamente relacionados, pois as instituições geralmente precisam tomar decisões sobre como preservar e acessar as páginas da *web* produzidas por sua organização. Essas decisões são importantes porque podem ter um impacto significativo na memória digital de uma instituição, ou no registro coletivo de suas atividades e informações na *web*.

Uma das principais considerações no arquivamento da *web* e nas políticas institucionais é a questão de quais páginas da *web* devem ser preservadas. Essa decisão envolve uma série de fatores, como a relevância das páginas para a missão e atividades da instituição, o valor potencial das páginas para pesquisa e outros fins, e o custo e viabilidade de preservação das páginas. As instituições podem usar vários critérios para tomar essa decisão, como a idade das páginas, o formato em que são produzidas e seu potencial valor futuro.

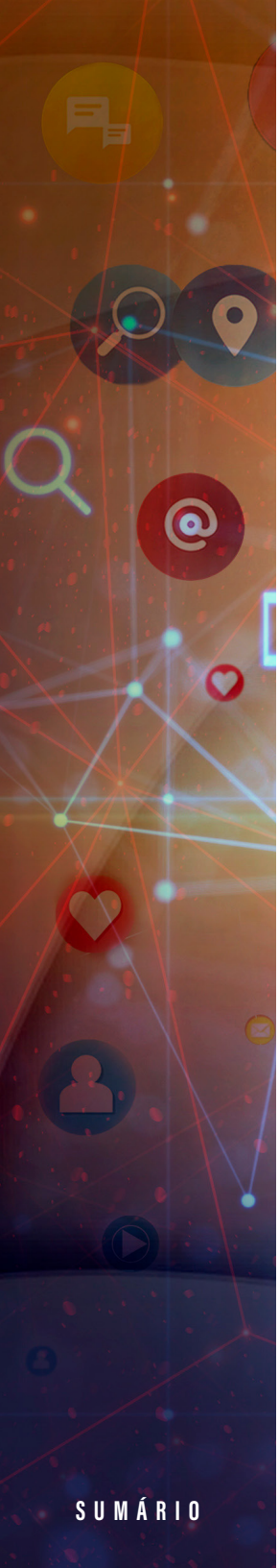
Outra consideração importante no arquivamento da *web* institucional diz respeito ao acesso às páginas da *web* preservadas. Possibilitar o acesso envolve uma variedade de fatores, como as capacidades técnicas da instituição, as necessidades de diferentes grupos de usuários e as questões legais e éticas relacionadas ao acesso a páginas da *web*. As instituições precisam tomar decisões sobre o nível de acesso a fornecer a diferentes grupos, bem como sobre o formato em que as páginas da *web* serão disponibilizadas. Ao tomar decisões ponderadas sobre quais páginas da *web* institucional devem ser preservadas, como e para quem fornece acesso a elas, garantirá a sustentabilidade do serviço oferecido para que a presença institucional na *web* seja preservada para futuros pesquisadores.

Muitas instituições culturais estão desenvolvendo programas para enfrentar o desafio da preservação digital. De acordo com Brayner (2016), o Brasil, com a quantidade de endereços ativos na *web*, precisa implementar políticas de arquivamento da *web* para preservar seu rico patrimônio digital, pois atrasar iniciativas nacionais de arquivamento da *web* poderá resultar na perda irrecuperável de conteúdo cultural valioso.

Algumas instituições brasileiras já estão trabalhando para garantir a preservação dos seus acervos digitais e publicando suas políticas de preservação, nelas incluindo a preservação das páginas *web* da instituição. Na Figura 17 é possível observar algumas características importantes no fluxo do arquivamento de *websites* e mídias sociais.

Figura 17 - Etapas da preservação de *websites* e mídias sociais

Fonte: Conselho Nacional de Arquivos (2023b).



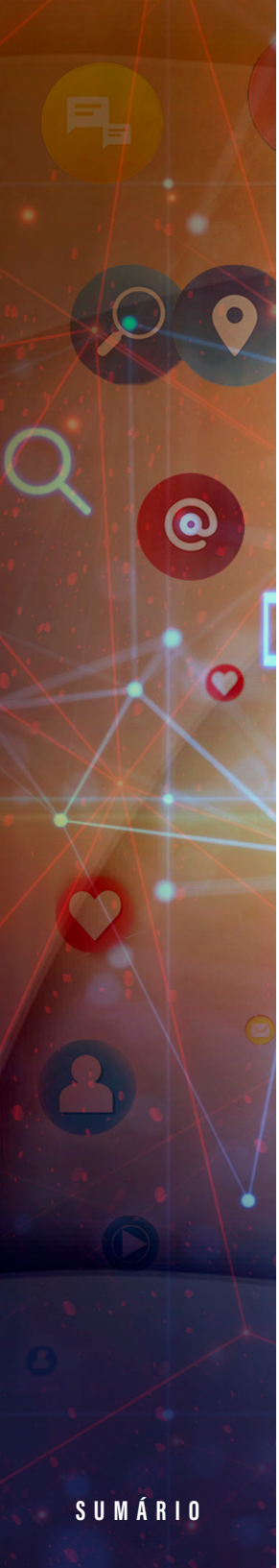
É importante ressaltar que a Universidade Federal do Rio Grande do Sul (UFRGS) aprovou, em março de 2021, a Política de Preservação de Acervos Digitais (PPAD) por meio da Resolução nº 064 (UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2021). Um dos objetivos dessa política é a preservação dos *site*s institucionais da UFRGS, tarefa que foi atribuída ao Núcleo de Pesquisa em Arquivamento da *web* e Preservação Digital da UFRGS (NUA*WEB*), tendo como responsabilidade fornecer suporte técnico e científico para a preservação dos *website*s da universidade.

Em 2020, a Fundação Biblioteca Nacional (FBN), publicou a Política de Preservação Digital da Biblioteca Nacional (PPDBN)[66] estabelecendo princípios, diretrizes, abrangência e responsabilidades para apoiar ações de preservação que possibilitem o acesso a longo prazo ao acervo digital, incluindo os *site*s e portais institucionais próprios ou resultados de acordos de cooperação nacionais ou internacionais.

A necessidade de estudar e promover o arquivamento da *web* para a preservação e acesso da informação a longo prazo é indiscutível. Em um estudo de Redkina (2021), é abordada a experiência bem-sucedida no arquivamento da *web* em várias bibliotecas mundiais, enfoca questões como seleção, acesso, tecnologias, descrição dos recursos, termos de acesso, entre outros. Um dos resultados do estudo demonstra que os arquivos da *web* são selecionados para complementar as coleções digitais das bibliotecas sobre temas emergentes, como Covid-19, ou para atender as exigências de grupos de usuários específicos.

Apesar do crescimento significativo em projetos de arquivamento da *web* e um aumento no número de países que hospedam essas iniciativas, o volume de conteúdos arquivados ainda é insignificante em comparação com a quantidade de informações publicadas na Internet. A fim de preservar o patrimônio cultural, científico

66 https://bndigital.bn.gov.br/wp-content/uploads/2021/01/politica_de_preservacao_digital_FBN_web.pdf

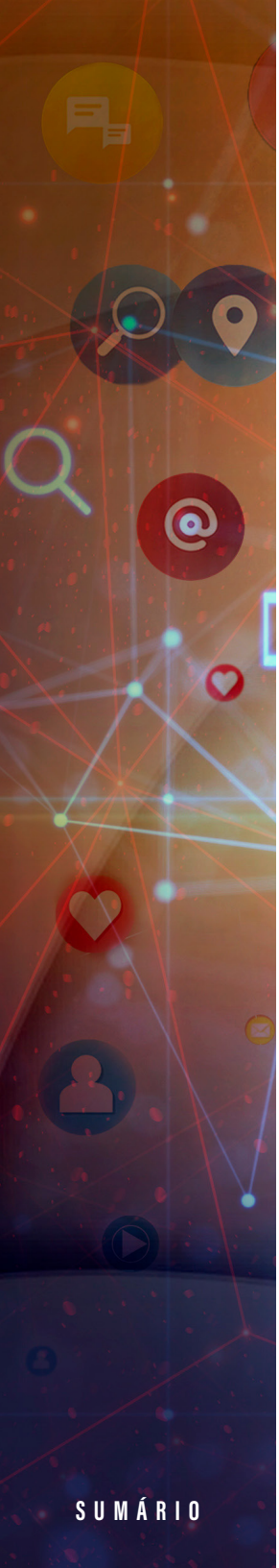


e histórico, uma série de arquivos e bibliotecas em todo o mundo estão implementando projetos para preservar seus próprios recursos e conteúdos da *web*.

Shiozaki e Eisenschitz, em um estudo realizado em 2009, já concluíram que: a) os benefícios da preservação da *web* superam os custos gerais e são justificável na medida em que é possível coletar *site*s públicos amplamente disponíveis a baixo custo, embora seja difícil mensurar de fato essa relação; b) os custos que ora possam ser repassados aos usuários são limitados e quando tais custos existem, são aceitáveis; e c) a necessidade de regulamentação legal, por meio de legislação, contratos e políticas são justificáveis na medida em que os riscos legais são mitigados a um nível aceitável, embora cada solução tenha suas vantagens e desvantagens em termos de negociação, acesso e escopo do arquivamento da *web*. A principal justificativa para o arquivamento da *web*, colocada por Shiozaki e Eisenschitz (2009), baseia-se, principalmente, na suposição de que os interesses públicos, em última análise, superam os direitos individuais, mas a validade de tal ponto de vista, quando baseado no raciocínio econômico, pode apresentar dificuldades para o arquivamento da *web* e o acesso aos *site*s arquivados.

Podemos dizer que com base na justificativa apresentada acima é que as bibliotecas nacionais do mundo todos empreendem iniciativas de preservação da *web*, por exemplo, a *British Library*, por meio de um decreto do governo britânico recebeu a incumbência de coletar publicações digitais e *site*s no Reino Unido para fins de armazenamento legal. Na Dinamarca diversas bibliotecas compartilham sua experiência no arquivamento da *web*, além de fornecer treinamento para especialistas, da mesma forma na Biblioteca Nacional da Hungria são desenvolvidos treinamentos em arquivamento da *web* para profissionais e bibliotecas fora da Europa.

O arquivamento da *web* vem sendo realizado em diversas organizações ao redor do mundo, dentre elas, as universidades, que



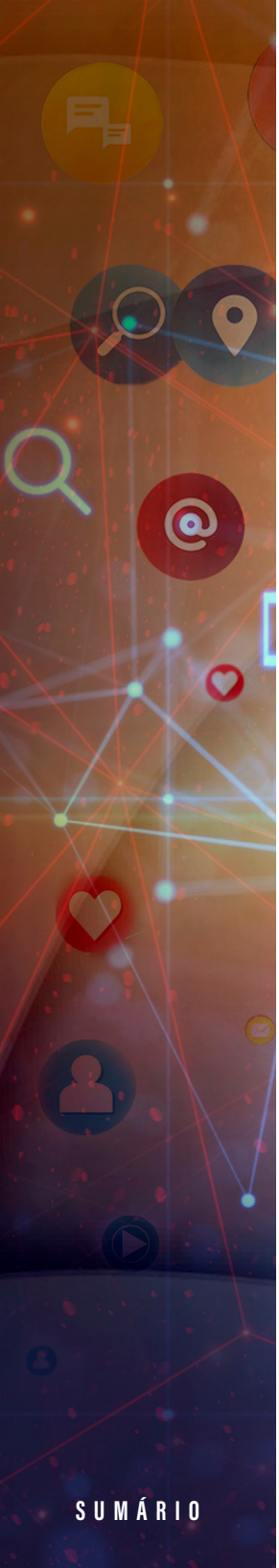
têm desenvolvido projetos para preservar conteúdos que vêm se perdendo com o passar dos anos. A *University of North Texas Libraries* foi uma das primeiras instituições acadêmicas dos Estados Unidos a arquivar *site*s, começando em 1997 com o *CyberCemetery*[67], um arquivo de *site*s governamentais que cessaram sua operação. A iniciativa visa preservar informações do governo e *site*s que apoiam o currículo da Universidade.

Outra iniciativa em destaque é o Projeto *web* em Risco, da *California Digital Library* (CDL), que começou em 2003. A CDL desenvolveu e operou o seu Serviço de Arquivamento da *web*, o qual estabeleceu objetivos como identificar as necessidades de arquivamento da *web* da comunidade de pesquisa em larga escala, identificar como as necessidades da comunidade para o arquivamento da *web* poderiam ter mudado desde o início do trabalho e explorar novas formas de colaboração e parceria.

A proposta de unificar todas as fontes de pesquisa científica em um único repositório proporciona maior eficiência na pesquisa e satisfação dos usuários, além de cruzar fontes e tipos distintos sobre uma mesma temática. A iniciativa da *University of North Texas* é um modelo bem planejado e estruturado que busca atender às distintas demandas do pesquisador de maneira objetiva e prática.

O estudo realizado por Redkina (2021) relaciona mais de 25 iniciativas de arquivamento da *web* de bibliotecas de diferentes tipos, no mundo, mostrando um significativo aumento no número de projetos expandindo os recursos de pesquisa em coleções diversas e condições o acesso para os usuários, se comparado com um estudo realizada 10 anos antes por Shiozaki e Eisenschitz (2009). As principais iniciativas apontadas por Redkina (2021) foram:

67 https://digital.library.unt.edu/explore/collections/GDCC/



a. *National Library of Germany*[68], inclui captura, indexação e arquivamento de *site*s utilizando processo automatizado, criam instantâneos dos *site*s que são indexados no catálogo da biblioteca para acesso;

b. *National and University Library of Iceland*[69], o arquivo da *web* coleta cópias de *site*s islandeses desde 2004 em conformidade com a lei islandesa de depósito legal de 2002. A coleção é limitada para domínios .is e *website*s islandeses selecionado dentro de outros domínios de nível superior;

c. *UK web Archive* (UKWA)[70], a Biblioteca Nacional do Reino Unido coleta milhões de *site*s a cada ano, preservando-os para as próximas gerações. O serviço pode ser utilizado para descobrir versões antigas ou obsoletas de *site*s do Reino Unido, pesquisar o texto dos *site*s e navegar por *site*s com curadoria de diferentes tópicos e temas. O arquivo contém *site*s que refletem vários aspectos da vida em todo o Reino Unido;

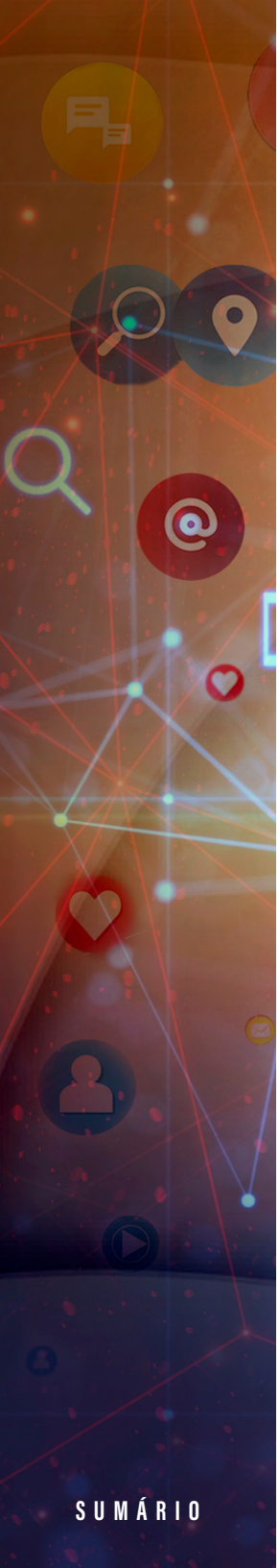
d. *National Library of Wales*[71], o projeto de arquivamento da *web* iniciou em 2003. A Biblioteca Nacional do País de Gales trabalha em parceria com a *British Library*, JISC e a *Wellcome Foundation*, em um projeto de arquivamento da *web* no Reino Unido. O objetivo do projeto é selecionar e capturar *site*s da *web* para preservação e gerar um acervo *on-line* para acesso público.

68 https://www.dnb.de/EN/Professionell/Sammeln/Sammlung_Websites/sammlung_websites_node.html

69 http://beta.vefsafn.is/en/

70 https://www.webarchive.org.uk/en/ukwa/

71 https://www.library.wales/catalogues-searching/about-our-collections/conservation/web-archive-wales



e. *National Library of Ireland* (NLI)[72], está criando um arquivo da *web* irlandês. O NLI reconhece o valor cultural intrínseco das informações publicadas na *web* e a necessidade de preservar esse material para as gerações atuais e futuras. O NLI arquiva a *web* irlandesa de forma seletiva desde 2011. Os *site*s arquivados antes de setembro de 2018 não são exibidos como coleções, eles podem ser vistos na coleção "*National Library of Ireland Collections 2011-2018*".

f. *National Library of Sweden*[73], começou a trabalhar com arquivamento da *web* em 1997. A coleção inclui servidores da *web* no domínio .se e servidores localizados em outros lugares, mas com recursos de identificação para a geolocalização sueca.

g. *National Library of Australia*[74], PANDORA é uma coleção crescente de publicações *on-line* australianas, estabelecida inicialmente pela Biblioteca Nacional da Austrália em 1996, e agora constituída pela colaboração com outras nove bibliotecas australianas e organizações culturais. O nome PANDORA é um acrônimo que resume a missão do arquivamento *web* australiano: *Preserving and Accessing Networked Documentary Resources of Australia*, em português: Preservar e acessar recursos documentais em rede da Austrália.

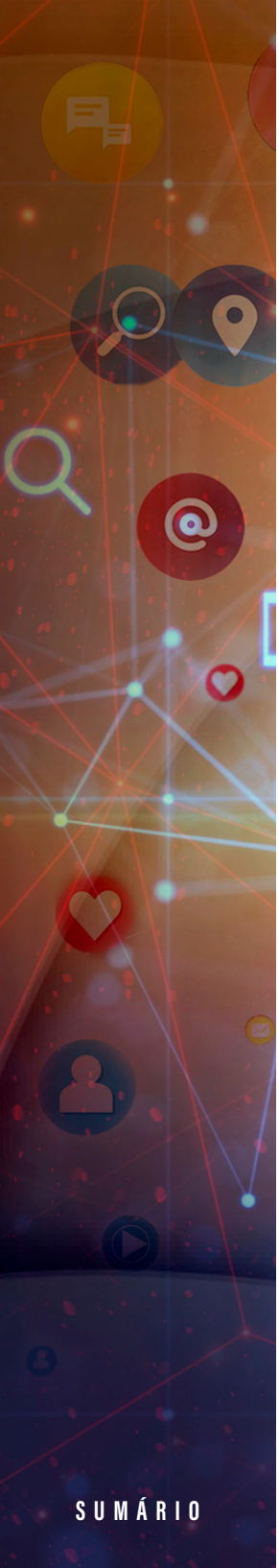
h. *Library of Congress*[75]. Os arquivos estão disponíveis na secção "Colecções Digitais" e são apresentadas como colecções temáticas (Afeganistão, *Elections in India in* 2009, *Manuscript Department*, etc.);

72 https://archive-it.org/home/nli

73 https://www.kb.se/hitta-och-bestall/digitala-kollektioner.html

74 https://pandora.nla.gov.au/

75 https://www.loc.gov/collections/?fa=originalformat:arquivado+web+site



i. *National Library of New Zealand*[76], o Arquivo da *web* da Nova Zelândia é uma coleção de *sites* arquivados da Nova Zelândia e do Pacífico. O arquivo faz parte da coleção *Alexander Turnbull Library*[77]. Inclui *sites* coletados desde 1999 e *sites* que não estão mais *on-line*. A coleção é uma seleção de *sites* sobre a Nova Zelândia e os neozelandeses, incluímos *sites* relacionados ao Pacífico. Os *sites* são arquivados para fins de pesquisa e preservação de longo prazo. A maioria dos *sites* no arquivo da *web* é coletada em intervalos regulares, a coleção de jornais é arquivada diariamente. O acesso é aberto, mas apenas nas dependências da biblioteca;

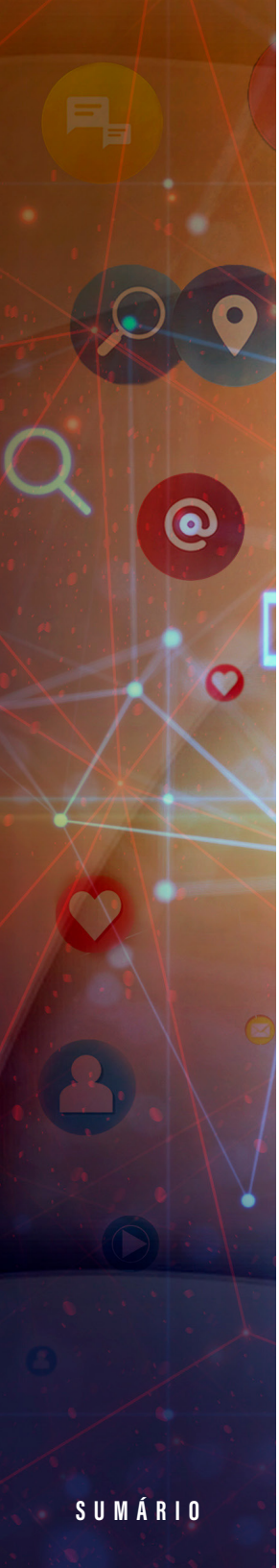
j. *National Digital Archive of Russia*[78], em 2017 foi anunciado o início da primeira iniciativa que criaria um arquivo dos *websites* de organizações e instituições russas a fim de prevenir o irrecuperável desaparecimento dos recursos da rede e assegurar a sua preservação. Periodicamente, são arquivados recursos como o *site* oficial do Presidente e do Governo da Rússia. Novas cópias do arquivo são criadas todos os dias, o que permite ver como o *site* mudou ao longo do tempo. Também, são arquivados canais como o *X.com* (antigo *Twitter*), *Instagram* e *Telegram*. Um critério importante para a escolha de um recurso *web* a ser arquivado foi sua fragilidade, é preciso preservar materiais que, com grande probabilidade, logo desaparecerão do espaço da Internet na forma em que estão no momento.

Alguns dos maiores arquivos e bibliotecas do mundo estão envolvidos no arquivamento da *web*, cobrindo a captura, descrição, indexação e disponibilização da informação. Às iniciativas nacionais juntam-se também arquivos, bibliotecas e instituições públicas

76 https://natlib.govt.nz/collections/az/new-zealand-web-archive

77 https://natlib.govt.nz/collections/a-z/alexander-turnbull-library-collections

78 https://ruarxive.org/



menores, bem como bibliotecas universitárias, interessadas na preservação do patrimônio cultural, científico e histórico local.

Atividades conjuntas são apresentadas no âmbito de projetos como o *End of Term* (EOT)[79] da *California Digital Library*, cuja tarefa é a de manter os *website*s do governo dos EUA no final dos mandatos. O EOT é um projeto colaborativo com instituições parceiras, tais como a *University of California Library*, *Internet Archive*, *Library of Congress* e *Stanford University*. A expansão do arquivamento da *web* tem sido possível, também, graças a alguns programas de financiamento como por exemplo os projetos apoiados pela Fundação *Andrew W. Mellon*[80].

No *site* do IIPC se pode verificar uma lista de instituições membros em todo o mundo, que trabalham em prol do arquivamento da *web*[81]. Também, o IIPC financia projetos técnicos e educacionais com base nos objetivos traçados no Plano Estratégico de Ação. O consórcio também colabora em projetos de pesquisa e desenvolvimento, compartilhando dados e ferramentas de teste. Forças-tarefa são formadas para estudar e fazer recomendações sobre questões ou problemas específicos. (INTERNATIONAL INTERNET PRESERVATION CONSORTIUM, 2023, tradução nossa).

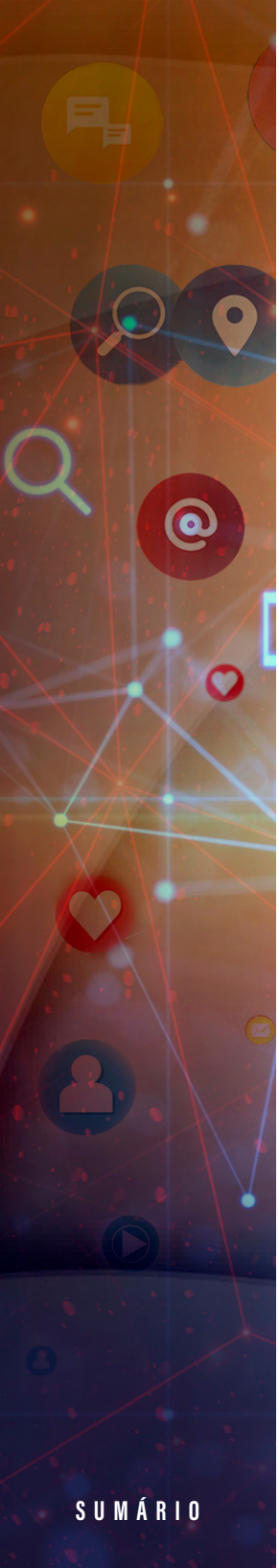
Uma das mais completas fontes dedicada a listar as iniciativas de arquivamento da *web* e em constante atualização com novos dados é a página da *Wikipedia*[82], onde é possível encontrar o ano de criação, tecnologias utilizadas, número de pessoas envolvidas nos projetos de arquivamento, domínios capturados, armazenamento e comentários, dentre outras informações específicas das iniciativas.

79 https://cdlib.org/services/pad/webarchiving/end-of-term-web-archive/

80 https://blog.archive.org/2020/12/08/community-webs-program-receives-1130000-andrew-w-mellonfoundation-award-for-a-national-network-of-public-libraries-building-local-history-web-archives/, https://hyperallergic.com/421336/rhizome-receives-1-million-award-for-webrecorder/

81 https://netpreserve.org/about-us/members/

82 https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives



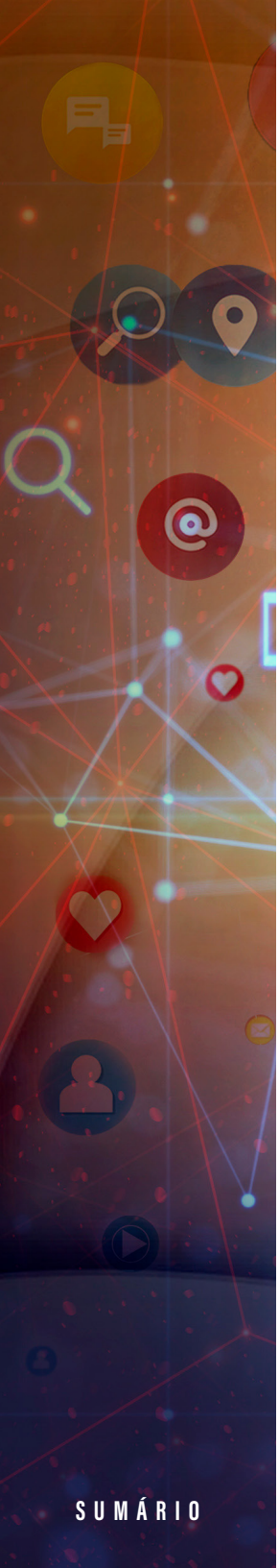
Nos Estados Unidos, de acordo com um *survey* aplicado em 2017[83], uma grande parte dos respondentes, cerca de 61%, representam iniciativas de arquivamento da *web* gerenciadas por faculdades ou universidades.

Segundo Grácio e Madio (2021), a implementação da preservação digital em uma organização depende de uma política de preservação digital, de um plano de ação e dos processos envolvidos em sua implementação. Estes devem incluir tecnologias de informação e comunicação, cultura organizacional e elementos organizacionais, legais e técnicos. A política de preservação digital não define atividades específicas; esse é o papel do plano de ação de preservação digital, que traça procedimentos, operações e responsabilidades para executar parte ou toda a política de preservação digital.

PLANOS DE PRESERVAÇÃO PARA *WEBSITES*

O arquivamento da *web* em contextos institucionais, por exemplo, bibliotecas e arquivos nacionais e ambientes universitários é frequentemente dirigido por normas profissionais de prática de arquivamento e gestão de registos, mandatos e restrições legais, por exemplo, restrições de depósito e direitos de autor, bem como por recursos humanos e técnicos limitados em relação à escala da tarefa. Uma das dificuldades enfrentadas pelas organizações que pretendem realizar o arquivamento da *web* em larga escala é a variedade de problemas com os quais precisam lidar e a pressão constante para acompanhar a evolução da *web*. (HOCKX-YU, 2011).

83 https://doi.org/10.17605/OSF.IO/3QH6N

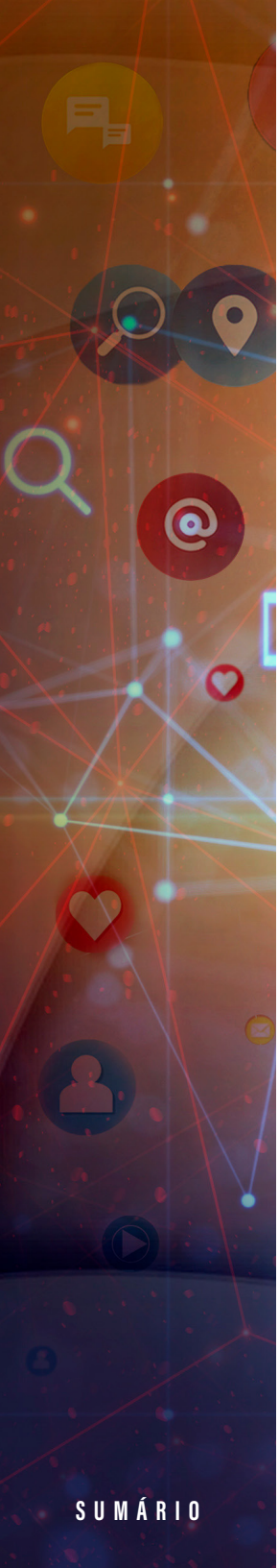


Apesar da longa história de avaliação e práticas de desenvolvimento de coleções e arquivos, o meio digital, ao longo dos anos, tem modificado as formas convencionais em relação à solicitação de consentimento, mesmo em ambientes institucionais. Segundo Summer (2020), os meios de captura impedem frequentemente, ou pelo menos desencorajam os arquivistas da *web* a interagir com os proprietários de conteúdo. Isto criou um panorama onde a maioria dos profissionais envolvidos com o arquivamento da *web* não quer correr riscos no que diz respeito aos limites da seleção do conteúdo, direitos de autor e direitos individuais à privacidade (WINTERS, 2020).

Os projetos de arquivamento da *web* institucional funcionam por meio de várias medidas para conter as perdas, para salvar o que se considera importante antes que desapareça. Ao selecionar um determinado conteúdo, outros podem não ser percebidos, se degradam ou são esquecidos. Esse processo de esquecimento opera por necessidade, porque os profissionais de arquivamento da *web* podem ser incapazes de armazenar todos os registros do *site*. Isso ocorre porque nossa capacidade de gerar dados está superando em muito nossa capacidade de armazená-los a longo prazo.

Para garantir as responsabilidades que a instituição deve assumir por item arquivado é necessário projetar e implementar um plano de preservação digital de longo prazo, reforçando o compromisso com a preservação digital. Projetar e implementar um plano de preservação da *web* irá garantir a preservação e acesso às coleções digitais, o gestor do repositório onde os *site*s serão arquivados, os depositantes e a comunidade designada precisam assumir as responsabilidades inerentes ao arquivamento da *web* institucional. Os procedimentos de arquivamento devem ser documentados e sua conclusão assegurada.

De acordo com Márdero Arellano (2008) um plano de preservação digital deve garantir uma abordagem organizada para a preservação a longo prazo; o acesso contínuo para todos os tipos



de acervo, não importando o formato e suas alterações futuras; a documentação deve ser suficiente para suportar a usabilidade pela comunidade designada; a definição dos níveis de preservação que serão aplicados e quais as mudanças na tecnologia e nos requisitos do usuário que serão tratadas de maneira estável e oportuna.

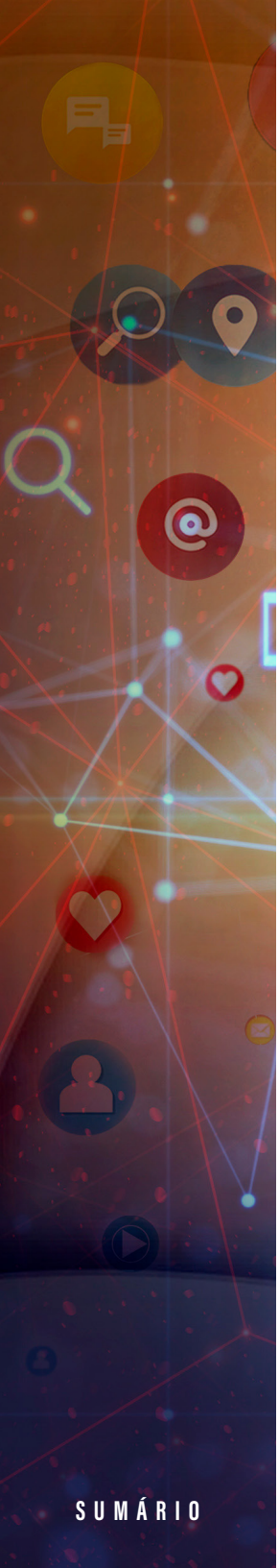
O Plano de Preservação Digital de *website*s da Universidade Federal do Rio Grande do Sul, por exemplo, descreve as políticas e procedimentos necessários para estabelecer uma estrutura técnica, procedimental e organizacional para garantir a preservação contínua da informação por um período de tempo determinado, mantendo os atributos considerados essenciais por meio das ações realizadas nos sistemas de informação e objetos digitais que compõem. Reproduzimos abaixo os principais pontos do plano de preservação digital da UFRGS:

A partir da designação da Política de Preservação de Acervos Digitais da Universidade Federal do Rio Grande do Sul (2021), o Núcleo de Pesquisa em Arquivamento da *web* e Preservação Digital (NUA*WEB*/UFRGS) é responsável por fornecer o suporte necessário para a preservação da *web* institucional da Universidade. Isso inclui a participação de estudantes, bolsistas, professores e pesquisadores.

O diagnóstico das condições de arquivabilidade dos *websites* é realizado a partir de métodos e ferramentas da comunidade de arquivamento da *web*, tais como o *Credible Live Evaluation of Archive Readiness* (CLEAR) e plataformas como o *ArchiveReady*.

As capturas dos *site*s são realizadas conforme a estrutura organizacional da instituição (faculdades e unidades de ensino, pesquisa e extensão). A primeira captura foi realizada como uma única captura de todos os conteúdos, para iniciar o versionamento, utilizando para isso *software* livre e *open source* como ferramentas.

O armazenamento dos *site*s é realizado no formato ISO WARC (*web ARChive*) e o acesso aos *site*s preservados faz-se também por meio de plataformas *open source.* Os *site*s serão preservados por



tempo indeterminado, garantindo o versionamento, histórico, uso e reuso dos dados. Solicitações de remoção de conteúdos (*takedown*) serão analisados caso a caso.

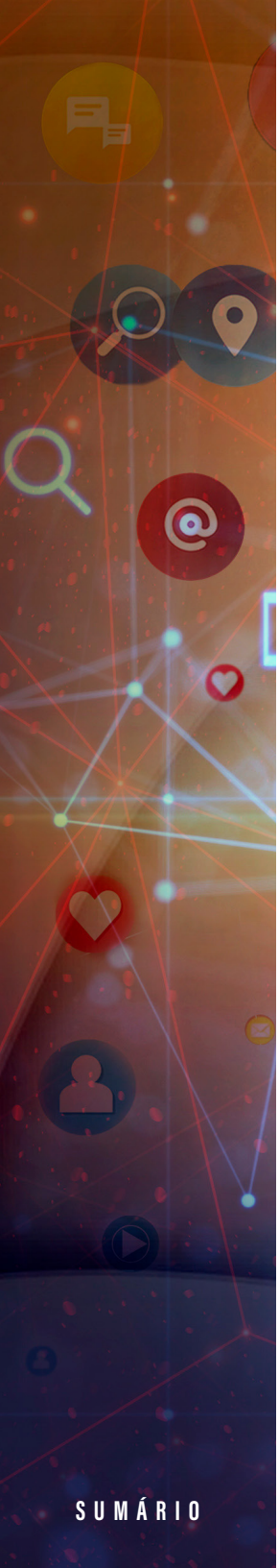
Pretende-se que atualizações, revisões e ampliação do Plano de Preservação Digital de *website*s da Universidade Federal do Rio Grande do Sul sejam realizadas periodicamente a cada dois anos.

QUESTÕES ÉTICAS E LEGAIS

O arquivamento da *web* levanta uma série de questões éticas, pois envolve a preservação e disseminação de informações que podem ser sensíveis, pessoais ou privadas. Essas questões incluem questões sobre consentimento, privacidade, direitos de propriedade e o potencial de dano ou injustiça, que precisam ser cuidadosamente consideradas e abordadas para garantir que o arquivamento da *web* seja conduzido de maneira ética e responsável.

Como argumentado por Formenton e Gracioso (2020), a preservação digital enfrenta desafios em diversas áreas, como gestão, aspectos técnicos, questões jurídicas, políticas, econômicas e sociais. Nos aspectos jurídicos, há desafios relacionados aos direitos de propriedade intelectual e outras obrigações legais que impactam a cópia, o armazenamento, modificação e uso de conteúdo digital para fins de preservação de longo prazo.

Por outro lado, de acordo com Velte (2018), as políticas de permissões variam amplamente e são raras as instituições que preservam a *web* que as implementam. Solicitar autorização aos proprietários de *site*s, considerando também a *web* 2.0, que contém uma grande quantidade de conteúdo gerado por usuários e coletar permissões de todos é uma tarefa quase impossível e demorada, como observado por Pennock (2013).



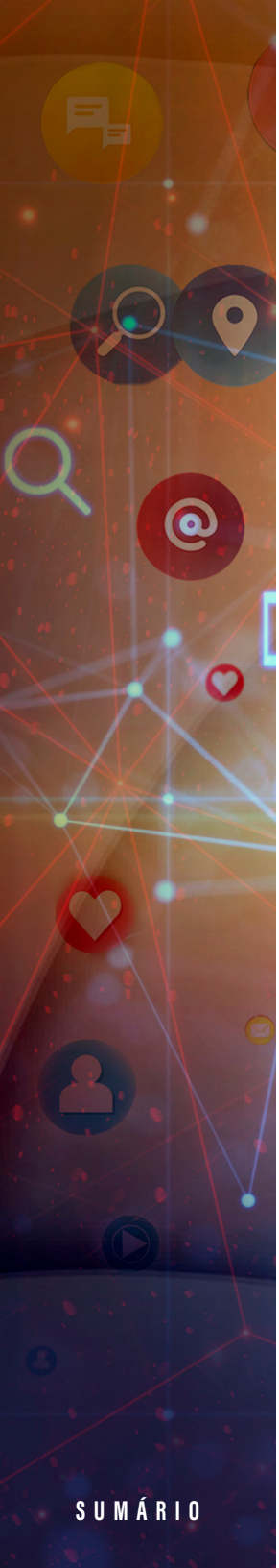
Uma das principais questões éticas no arquivamento da *web* é o consentimento. O arquivamento da *web* envolve a preservação e disseminação de informações que podem ter sido criadas e publicadas por indivíduos ou organizações sem seu conhecimento ou consentimento. Isso levanta questões sobre os direitos desses indivíduos e organizações e se é ético preservar e divulgar suas informações sem sua permissão.

Outro problema ético no arquivamento da *web* é a relação com a privacidade. Os arquivos da *web* podem conter informações pessoais, como nomes, endereços e outros detalhes de identificação, levantando questões sobre se é ético preservar e divulgar essas informações e se medidas devem ser tomadas para proteger a privacidade dos indivíduos cujas informações estão incluídas em arquivos da *web*.

Além disso, o arquivamento da *web* também provoca questões relacionadas aos direitos de propriedade. Os arquivos da *web* geralmente incluem materiais protegidos por direitos autorais, como imagens, texto e outras formas de propriedade intelectual, trazendo a tona questões sobre a ética de preservar e divulgar esses materiais sem a permissão dos detentores dos direitos autorais.

Ainda, segundo estudo realizado por Velte (2018), das organizações que preservam conteúdos *web*, 69% aplicam a regra do embargo, que consiste em um período em que o conteúdo não está disponível e está relacionado com a redução de concorrência de acessos entre o *site* arquivado e o *site* ao vivo. Algumas instituições, como o Arquivo.pt, determinam o embargo de um ano após a captura.

Callister (2021) argumenta que a atividade de arquivamento da *web*, incluindo o uso de ferramentas como o Perma.cc, não está em conformidade com a lei de direitos autorais atual e que a falta de defesas claras de direitos autorais para uso de arquivo digital pode representar um problema legal. A opção nestes casos acaba passando pelo pedido de remoção dos *links* (*takedown*), quando aplicável.



É importante destacar a presença destes problemas legais (DAY, 2003), que envolvem tanto os direitos autorais quanto as responsabilidades pelos conteúdos disponíveis. Além disso, o uso e reúso dessas informações apresentam dificuldades no que se refere às permissões, que nem sempre são claras nas páginas da *web*, principalmente quando se trata das concessões oferecidas pelo *Creative Commons*.

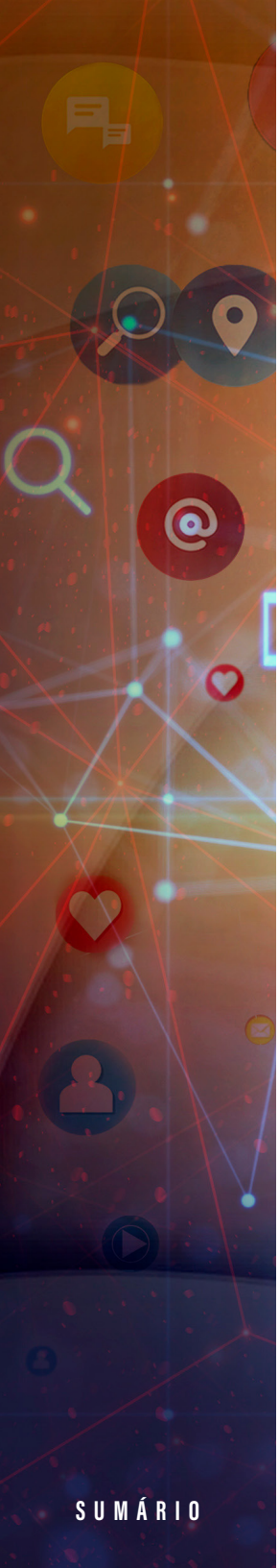
A *Creative Commons (CC)*[84] é uma licença de uso que oferece uma maneira flexível de gerenciar os direitos autorais. É um tipo de acordo que dá aos autores a possibilidade de escolher como querem compartilhar suas obras e também orienta os usuários sobre as permissões de uso. As licenças *Creative Commons* variam em níveis de abertura, desde aquelas mais abertas até as de uso mais limitado. Algumas instituições de arquivamento da *web* mencionam o *Creative Commons* em seus termos, como o *Internet Archive* e o *The National Archives* (UK).

Da mesma forma, é importante a compreensão de ferramentas como o *Creative Commons* para tornar mais transparente o processo de autorização legal por parte de criadores de conteúdos, e materiais como a cartilha *Creative Commons* Brasil (VALENTE; HOUANG, 2020) auxiliam no uso destas licenças.

Essas questões não só acarretam desafios legais, mas também éticos, já que se relacionam com o uso de informações protegidas por direitos autorais em contextos internacionais com diferentes jurisdições, influenciada pela desterritorialização promovida pela Internet e pela *web*. Por isso, é necessário refletir sobre a ética informacional ao utilizar esses conteúdos, incluindo questões concernentes à proteção de dados e privacidade na rede.

Uma das opções legais e éticas para coletar e usar dados arquivados da *web*, mesmo sem possuir os direitos autorais dos *site*s,

84 https://br.creativecommons.net/



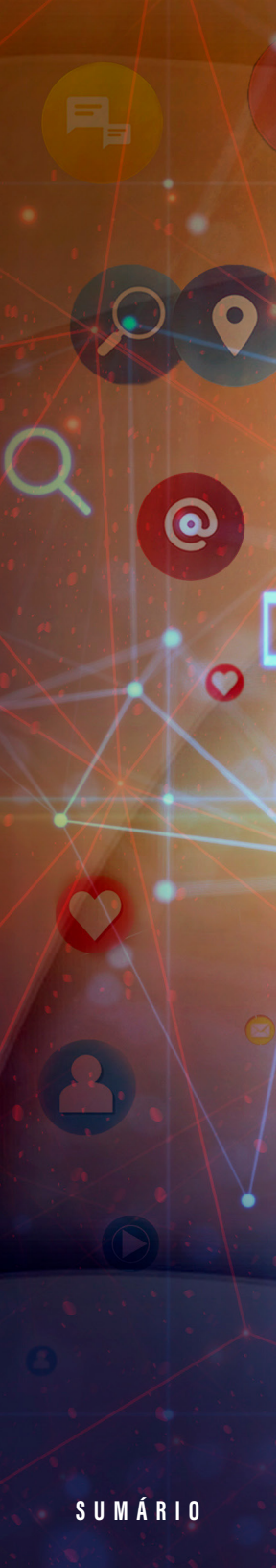
é por meio do conceito de *fair use* ou "uso justo". Esse termo tem origem na legislação americana e está relacionado com a tradição da *Common-law*, que permite o uso de conteúdo protegido por direitos autorais em certas circunstâncias, como fins educacionais, jornalísticos ou de pesquisa, por exemplo. Minow (2003) aborda especificamente essas questões e defende que, devido à crescente importância da *web* como fonte de produção e disseminação de informação, há um interesse em preservar partes de seu conteúdo.

Zittrain, Albert e Lessig (2014) a partir do estudo do uso do Perma.cc, apontam que os acadêmicos devem considerar a lei de direitos autorais antes de arquivar determinados materiais, como *blogs*, artigos de notícias e PDFs que não concedem permissão para cópia, bem como verificar os casos em que pode ser aplicado o "uso justo" (*fair use*).

No entanto, parte dos *site*s na *web* é protegida por direitos autorais, o que cria um dilema ao coletar esses dados. Algumas iniciativas tratam este dilema por meio do uso conjunto do mencionado conceito de "uso justo", sinalizando a intenção de não coletar os dados pelos motores de busca ou controlando as permissões de acesso com o uso de um arquivo *robots.txt*, além de atender a pedidos expressos para remover conteúdo coletado, procedimento a ser adotado nas políticas de *takedown*.

A aplicabilidade dessa legislação e do conceito de "uso justo" varia de país para país, portanto, é necessário avaliar cada caso individualmente. Na Suécia, por exemplo, o acesso aos arquivos da *web* é restrito e só pode ser feito pessoalmente, no local responsável pelo arquivamento, como na iniciativa Kulturarw3[85]. O mesmo ocorre com o acesso às páginas da *web* arquivadas pela Biblioteca Nacional da França, que só podem ser acessadas a partir das salas de leitura da biblioteca.

85 https://dhnb.eu/projects/kulturarw3-the-web-archive-of-the-national-library-of-sweden/

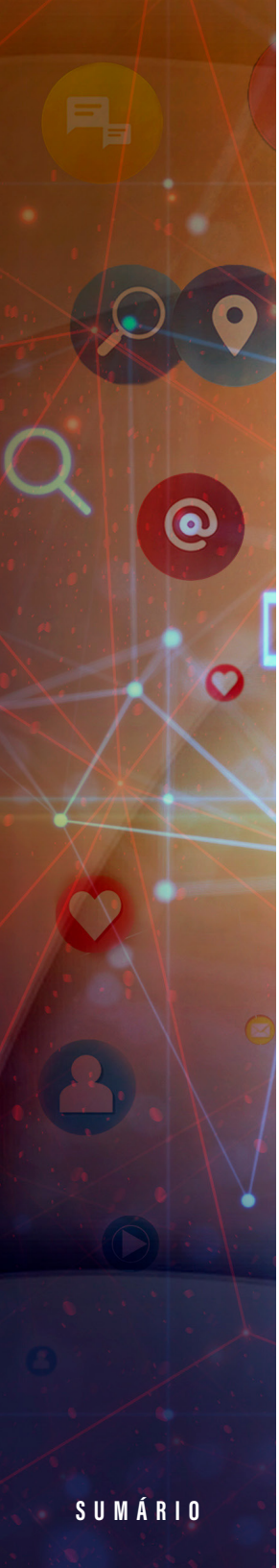


Existem instituições que possuem o mandato de preservação digital expresso em lei ou regulamento e, conforme Pennock (2013), a obrigatoriedade legal de capturar e preservar o conteúdo da *web*, sendo a perda dessas informações uma responsabilidade tanto institucional quanto social. Contudo, existem várias questões éticas no arquivamento da *web* que precisam ser pensadas. É importante que as organizações envolvidas no arquivamento da *web* consultem especialistas jurídicos e tomem as medidas adequadas para garantir a conformidade com as leis e regulamentações relevantes.

Os arquivos da *web* podem refletir vieses e os preconceitos e as perspectivas dos indivíduos ou organizações que os criaram, levando a uma representação incompleta ou distorcida dos materiais que estão sendo preservados. Outra questão é que os arquivos da *web* podem não estar acessíveis a todos os usuários, especialmente usuários com deficiências ou usuários em regiões com acesso limitado à Internet. Isso pode limitar a capacidade desses usuários de acessar e usar os materiais preservados.

O direito ao esquecimento (*right to be forgotten*) é um conceito legal que permite aos indivíduos solicitar a remoção de informações pessoais da Internet. Este direito é reconhecido na União Europeia ao abrigo do Regulamento Geral de Proteção de Dados (RGPD), que permite aos particulares solicitar a eliminação dos seus dados pessoais caso estes deixem de ser necessários para a finalidade para a qual foram recolhidos, ou caso o indivíduo retire o consentimento para o seu processamento. Entretanto, segundo Dougherty (2013), apesar da importância de garantir o direito ao esquecimento, é fundamental ressaltar o valor da preservação da memória e como este propósito é auxiliado pela tecnologia.

As organizações de arquivamento da *web* devem tomar medidas para garantir que estejam cumprindo o direito de ser esquecido, quando aplicável, e outras leis de privacidade. Isso pode envolver o desenvolvimento de políticas e procedimentos para responder a



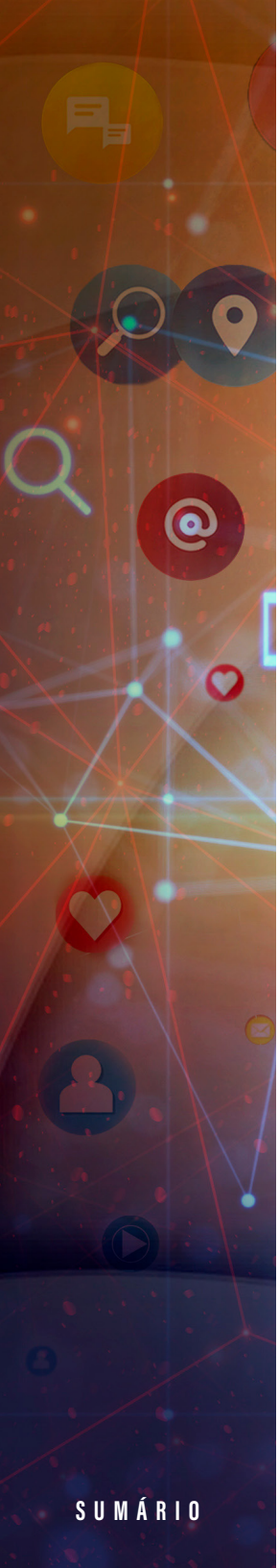
solicitações de remoção e trabalhar com indivíduos para identificar e remover informações pessoais do arquivo conforme necessário.

O *General Data Protection Regulation* (GDPR) é um conjunto de regras que foram implementadas na União Europeia (UE) em 2018 para proteger os dados pessoais de indivíduos na UE. O GDPR se aplica a qualquer organização que processe os dados pessoais de indivíduos na UE, independentemente de a organização estar sediada na UE ou não.

Uma das principais disposições do GDPR é a exigência de que as organizações obtenham consentimento explícito de indivíduos antes de coletar ou processar seus dados pessoais. Isso significa que as organizações devem fornecer informações claras e concisas sobre como os dados pessoais serão usados, e os indivíduos devem dar seu consentimento explícito para a coleta e processamento de seus dados pessoais.

No contexto do arquivamento da *web*, o GDPR e a LGPD podem representar um desafio porque o arquivamento da *web* envolve a coleta e preservação de informações da Internet, que geralmente incluem dados pessoais. Os arquivistas da *web* devem, portanto, garantir que obtenham o consentimento adequado dos indivíduos antes de coletar e preservar seus dados pessoais.

Para cumprir o GDPR e a LGPD, os arquivistas da *web* podem implementar várias medidas, como: a) fornecer informações claras e concisas sobre a finalidade do arquivamento da *web* e os tipos de dados pessoais que serão coletados; b) obter o consentimento explícito dos indivíduos antes de coletar seus dados pessoais; c) implementar medidas técnicas para proteger dados pessoais, como criptografar dados e limitar o acesso apenas a pessoal autorizado e d) revisar e atualizar regularmente suas práticas de arquivamento da *web* para garantir que continuem em conformidade com o GDPR no caso europeu e a LGPD no caso Brasileiro.

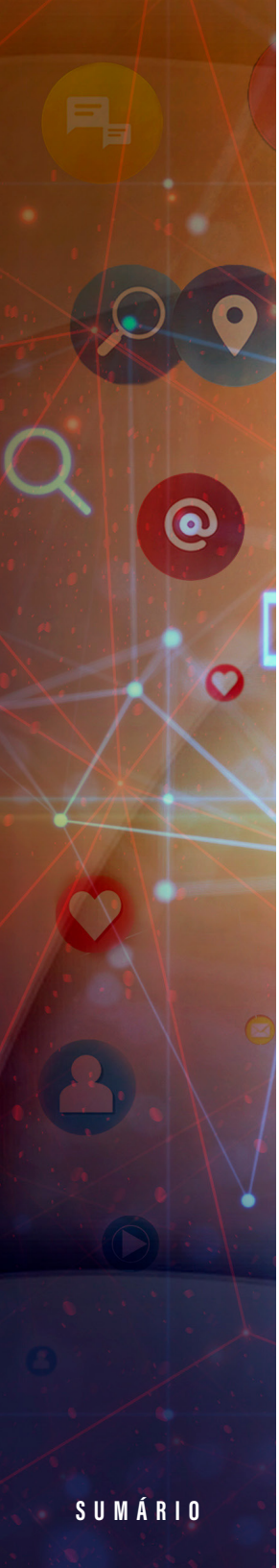


O GDPR e a LGPD apresentam desafios para o arquivamento da *web*, mas ao implementar as medidas apropriadas, os arquivistas da *web* podem garantir que estejam em conformidade com os regulamentos e que os dados pessoais dos indivíduos sejam protegidos.

Outra questão diz respeito ao uso da legislação de depósito legal aplicado aos arquivos da *web*, no Brasil, a lei n. 10.994/2004 rege o depósito legal e não atua sobre os materiais digitais. Atualizações nas legislações, compreendendo o meio digital, dariam maior segurança jurídica e mandato explícito para a guarda destes materiais, contudo, isto não deve ser um impeditivo para realizar ações de preservação digital visando o acesso ao longo do tempo.

Conforme Hockx-Yu (2014), o acesso aos arquivos da *web* é por vezes limitado em alguns países devido a requisitos legais e restrições de direitos autorais. Os *site*s são protegidos por direitos autorais e arquivá-los sem permissão pode vir a violar a legislação. Entretanto, é preciso haver um balanço entre os direitos autorais e o direito ao acesso à informação e preservação do patrimônio digital. Segundo Hockx-Yu (2014), às estruturas de depósito legal podem fornecer isenções à lei de direitos autorais, permitindo que as instituições de memória coletem publicações sistematicamente para o benefício das gerações futuras. A natureza aberta e onipresente da *web* criou uma expectativa de que os *site*s arquivados devem estar acessíveis *on-line* 24 horas por dia, sete dias por semana, no entanto, por vezes, o acesso ao conteúdo é limitado para evitar danos aos interesses dos proprietários dos direitos autorais. Por sua vez, Larivière (2000) já recomendava exceções específicas ao depósito legal, para permitir o acesso aos materiais eletrônicos aos usuários de instituições nacionais.

De acordo com Dougherty e Meyer (2014), as preocupações legais foram identificadas como um obstáculo fundamental à criação de arquivos da *web* no Reino Unido, particularmente o requisito legal de obter permissão prévia explícita para arquivar objetos da *web*.



A Biblioteca Britânica implementou inicialmente uma política para solicitar permissão dos proprietários do *site* antes de arquivar seu conteúdo, mas isso limitou o escopo do que poderia ser coletado, no entanto, os regulamentos que permitem que se preserve todo o conteúdo da *web* do Reino Unido foram implementados desde 2013[86]. Além do depósito legal, as considerações legais podem afetar a direção e os resultados dos projetos, ditando o que pode ser incluído e quem pode acessar os arquivos resultantes.

Uma perspectiva diferente seria sair de uma abordagem estática, centrada no documento, para uma abordagem centrada nos dados e, como argumenta Hockx-Yu (2014), envolver visualização e análise de dados que revelem padrões, tendências e relacionamentos não visíveis ao consultar os *site*s individualmente.

Rockembach (2017) discute questões éticas relacionadas à informação no contexto dos arquivos da *web*, ressaltando aspectos relevantes para a captura, acesso, uso e preservação desses registros. Para refletir sobre estes aspectos, algumas das perguntas que precisam ser realizadas são:

Na captura:

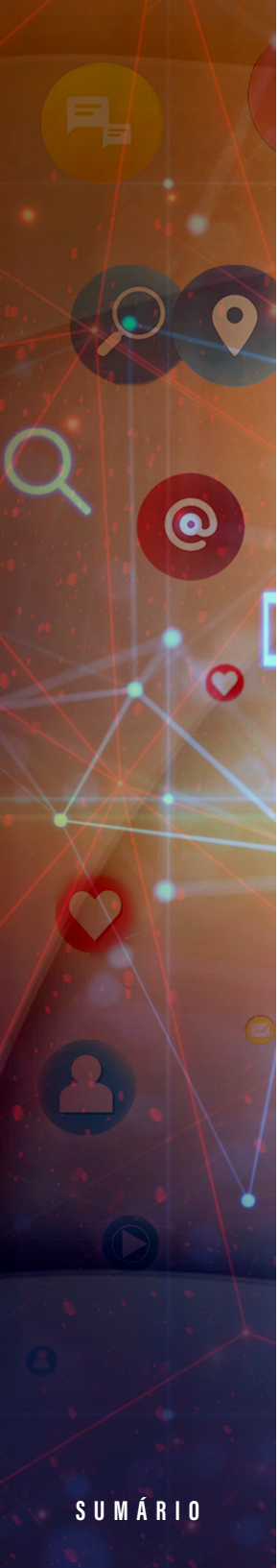
a. Quais parâmetros são usados para avaliar e selecionar as informações a serem arquivadas digitalmente?

b. Qual é a frequência de captura e a influência desta frequência na formação de uma memória digital?

c. Essa captura ocorre igualmente em todas as regiões (caso o escopo seja geográfico)? E se a relação é desigual, por que é assim?

d. Como o direito à privacidade pode ser protegido?

86 https://www.bl.uk/legal-deposit/web-archiving



No acesso e uso:

e. Que perguntas devem ser abordadas em torno de como as informações são usadas e manipuladas?

f. Como os direitos autorais podem ser tratados em termos de uso dessas informações?

Na preservação:

g. Quem é o responsável pela preservação da *web*?

h. Por quanto tempo as informações serão preservadas?

i. O que determinará os parâmetros de tempo de preservação?

Os termos de uso são documentos fundamentais para proteção jurídica de um *site*, permitindo compreender o uso e responsabilidades do usuário, já as políticas de privacidade devem conter informações sobre tratamento de dados pessoais e privacidade. Nunes (2021) realizou estudos sobre os termos de uso e políticas de privacidade de 19 plataformas de arquivos da *web*: *National Library of Australia*[87], *Bibliothèque et Archives Nationales du Québec*[88], *Library and Archives Canada*[89], *National and University Library of Croatia*[90], *Columbia University Libraries*[91], *Cornell University Library*[92], *Harvard Library*[93], *Internet Archive*[94], *Los Alamos National Laboratory Research Library*[95], *Old Dominion University Department of Computer Science*[96],

87 www.nla.gov.au

88 www.banq.qc.ca

89 https://www.collectionscanada.ca/

90 www.nsk.hr

91 https://library.columbia.edu/collections/webarchives.html

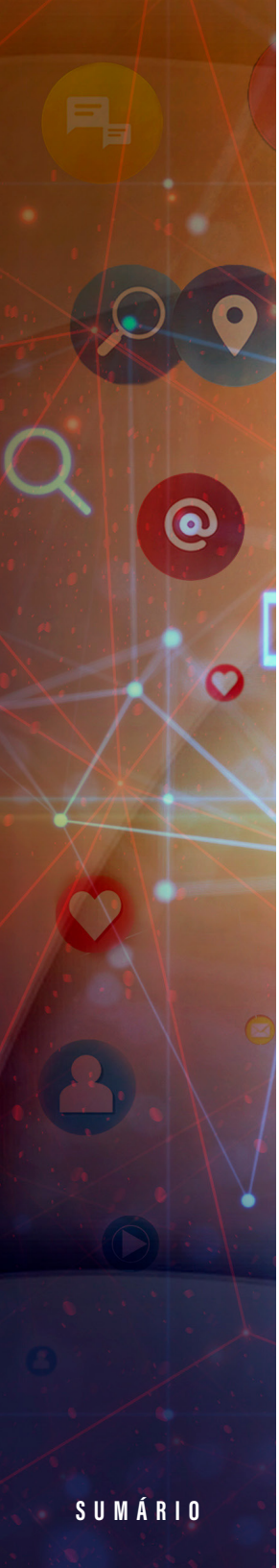
92 https://www.library.cornell.edu/

93 http://library.harvard.edu

94 www.archive.org

95 www.lanl.gov/library

96 www.cs.odu.edu



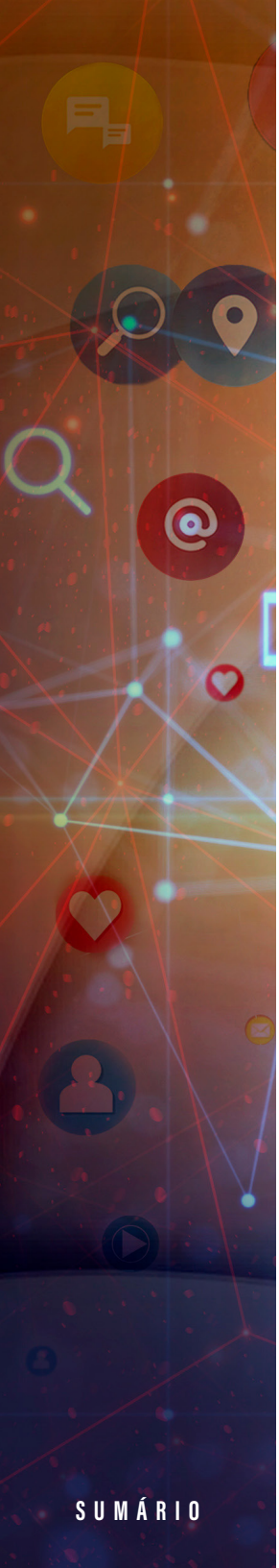
Stanford University Libraries[97], UCLA *Research Library*[98], *University of North Texas Libraries*[99] *Hanzo Archives*[100], *Mirrorweb*[101], *The National Archives UK*[102], *National Library Board Singapore*[103], *National Library of Chile*[104] e Arquivo.pt[105].

Em relação aos termos de uso, os estudos foram baseados em elementos comuns, como a descrição do serviço, proteção legal, uso da plataforma, penalidades por descumprimento, limitações de responsabilidade, garantias para o usuário e declaração de alteração e atualização. A análise das políticas de privacidade foi baseada em elementos como proteção legal, descrição da coleta, dados coletados, compartilhamento e divulgação de dados, monitoramento e controle sobre os dados, e declaração de alteração e atualização do documento.

Em pesquisa sobre proposições em trâmite na Câmara dos Deputados utilizando os termos Internet, páginas *web*, depósito legal e preservação digital, desenvolvida por Santos (2021), foi encontrada a já mencionada PL 2431/2015, que dispõe sobre o patrimônio público digital institucional inserido na rede mundial de computadores e dá outras providências, mas também outros quatro Projetos de Lei, com os seguintes destaques:

> PL 3050/2020 – Esta lei altera o art. 1.788 da Lei n.º 10.406, de 10 de janeiro de 2002, que institui o Código Civil, a fim de dispor sobre a sucessão dos bens e contas digitais do autor da herança de qualidade patrimonial

97 www.library.stanford.edu
98 www.library.ucla.edu/yrl
99 www.library.unt.edu
100 www.hanzoarchives.com
101 www.mirrorweb.com
102 www.nationalarchives.gov.uk
103 www.nlb.gov.sg
104 www.bibliotecanacional.cl
105 www.arquivo.pt



"Serão transmitidos aos herdeiros todos os conteúdos de qualidade patrimonial de contas ou arquivos digitais de titularidade do autor da herança"

PL 3051/2020. Acrescenta o art. 10-A (Marco Civil da Internet), a fim de dispor sobre a destinação das contas de aplicações de Internet após a morte de seu titular.

"Os provedores de aplicações de Internet devem excluir as respectivas contas de usuários brasileiros mortos imediatamente, se for requerido por familiares após a comprovação do óbito"

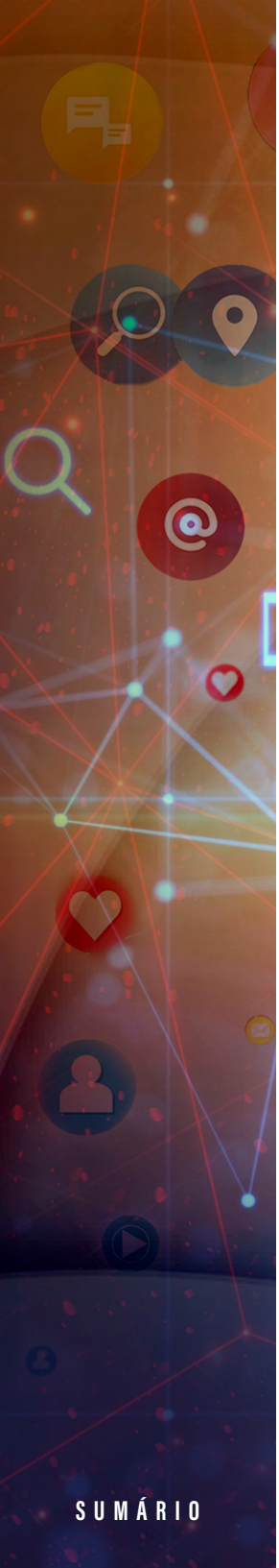
PL 3395/2020. Acrescenta o art. 21-A à Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet), proibindo os provedores de plataformas digitais de remover conteúdos publicados por seus usuários, salvo por força de cumprimento de ordem judicial.

"É vedado às plataformas digitais remover conteúdos publicados por seus usuários, salvo por força de ordem judicial, à exceção da hipótese prevista no art. 21, em que o provedor procederá à indisponibilização do conteúdo independentemente de ordem judicial."

PL 3573/2020. Altera o Marco Civil da Internet para proibir a retirada de conteúdos pelas aplicações de Internet nos casos em que especifica

"Com o intuito de assegurar a liberdade de expressão e impedir a censura, o provedor de aplicações de Internet não poderá retirar conteúdo gerado por terceiro, exceto por ordem judicial ou com a indicação expressa do crime que se está cometendo mediante a divulgação do conteúdo retirado"

Aspectos importantes nas plataformas, os termos de uso e políticas de privacidade não podem ser subestimados na disponibilização de conteúdo. Nunes e Rockembach (2021) elencaram elementos necessários e comuns aos Arquivos da *web*, que expomos abaixo:

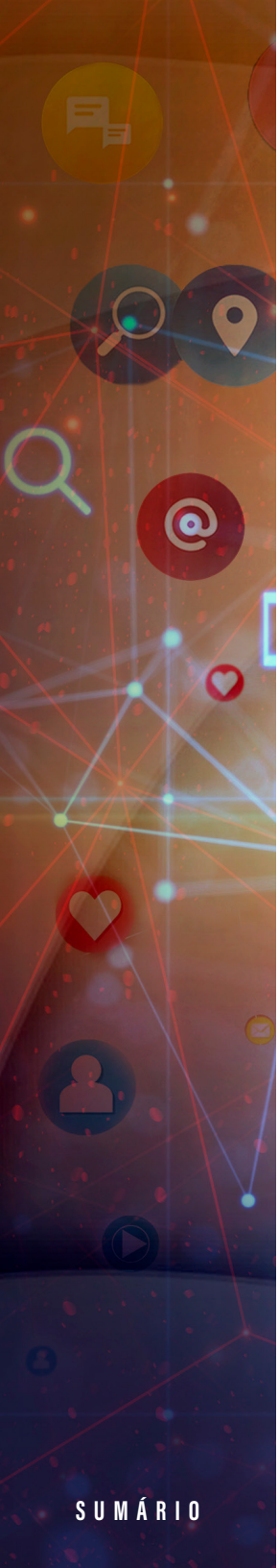


Em relação aos termos de uso:

1. Descrição do serviço
2. Como o serviço é oferecido
3. Se mencionam leis
4. Declaração Direitos Autorais
5. Descrição do uso
6. Limites de responsabilidades
7. Se informam sobre garantias
8. Data atualização termo
9. Documento atualizado
10. Notifica usuário atualização documento

Já em relação às políticas de privacidade:

1. Política em documento único
2. Proteção legal (Leis de proteção de dados pessoais e privacidade)
3. Motivo coleta dados (melhorar os serviços oferecidos; otimizar os *sites*; métricas de uso do *site*)
4. Informa dados coletados (*cookies*, informações pessoais, páginas visualizadas e recursos arquivados)
5. Compartilha/divulga dados (quando exigido por lei/ordem judicial, fornecedores/consultores/prestadores de serviço, equipe/membros da instituição, autoridade/agência governamental)
6. Monitoramento/controle sobre dados (informa onde os dados serão armazenados e processados, e se os mesmos podem ser transferidos para outros países, informam prazos de manutenção dos dados)



7. Data atualização política

8. Documento atualizado

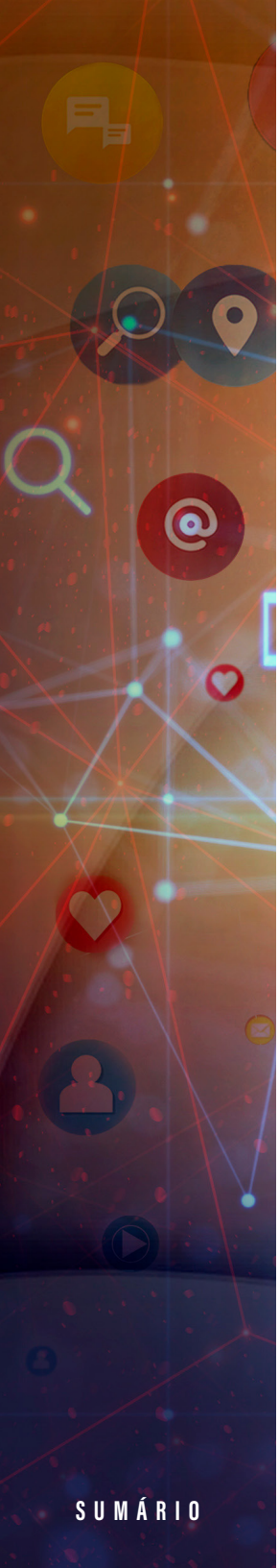
9. Notifica usuário atualização documento

Como argumentado por Valente (2019), é preciso encontrar um equilíbrio entre limitações impostas pelas legislações e a garantia de outros direitos, e no caso do arquivamento da *web*, é necessário ponderar os direitos autorais com o direito ao patrimônio cultural e histórico, beneficiando também o direito à informação e à proteção do patrimônio artístico, histórico e cultural.

POLÍTICAS DE *TAKEDOWN*

As políticas de *takedown* ou remoção no arquivamento da *web* referem-se aos processos e procedimentos que os arquivos da *web* usam para atender às solicitações de remoção ou restrição de acesso a um conteúdo específico da *web*. As políticas de remoção fornecem uma maneira de indivíduos e organizações removerem ou restringirem seu conteúdo da *web* se acreditarem que isso infringe seus direitos ou é inapropriado.

As políticas de remoção geralmente envolvem um processo de revisão e resposta a solicitações de remoção. Quando um indivíduo ou organização envia uma solicitação de remoção, o arquivo da *web* normalmente analisa a solicitação para determinar se ela atende aos critérios de remoção ou restrição, podendo envolver o exame do conteúdo em questão, bem como a consulta a especialistas jurídicos e outras partes interessadas. Se o arquivo da *web* determinar que o conteúdo atende aos critérios de remoção ou restrição, ele normalmente tomará medidas para remover ou restringir o acesso ao conteúdo, o que significa a remoção do conteúdo do arquivo da *web* ou



a disponibilização do conteúdo apenas para determinados usuários ou sob determinadas condições.

As políticas de remoção são diretrizes ou regras que determinam quando e como os arquivos da *web* devem ser removidos do acesso público, bem como procedimentos para solicitar a remoção de arquivos da *web*. As políticas de remoção são um aspecto importante do arquivamento da *web*, pois ajudam a equilibrar a necessidade de preservar materiais de patrimônio cultural baseados na *web* com a necessidade de respeitar os direitos e interesses de indivíduos e organizações que podem ser apresentados nos arquivos da *web*.

Na elaboração de uma política de *takedown* para um arquivo da *web*, alguns pontos podem ser considerados:

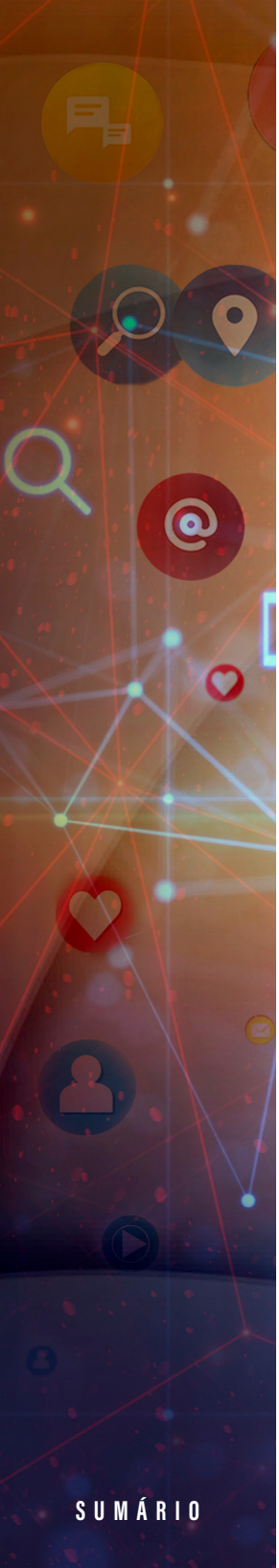
1. Identificação: incluir informações sobre a organização responsável pelo arquivo da *web*, bem como informações de contato para solicitações de *takedown*.

2. Motivos para *takedown*: listar as razões pelas quais um usuário pode solicitar a remoção de um arquivo da *web*, como violação de direitos autorais, difamação, assédio ou outros tipos de conteúdo ilegal.

3. Processo de solicitação: explicar o processo de como os usuários podem solicitar um *takedown*, incluindo informações sobre o que precisa ser fornecido na solicitação e o tempo de resposta estimado para a solicitação.

4. Verificação: incluir informações sobre como a empresa, organização ou indivíduo responsável pelo arquivo da *web* irá verificar se a solicitação de *takedown* é legítima e se o conteúdo a ser removido é de fato ilegal.

5. Ação: descrever quais ações serão tomadas após a verificação da solicitação, incluindo o prazo para a remoção do conteúdo e o aviso ao usuário afetado.



6. Recurso: incluir informações sobre como os usuários podem recorrer da decisão contrária de *takedown*, bem como obter mais informações ou esclarecimentos.

As políticas podem variar dependendo da organização responsável pelos arquivos da *web* e do contexto legal e cultural em que os arquivos estão sendo preservados. Por exemplo, algumas organizações podem ter políticas de remoção que permitem a retirada de arquivos da *web* se eles contiverem informações confidenciais ou pessoais ou violarem os direitos de indivíduos ou organizações. Outras organizações podem ter políticas de remoção mais rígidas que permitem apenas a retirada de arquivos da *web* em circunstâncias específicas e limitadas. Em um dos exemplos de arquivo da web, do *Internet Archive*, a solicitação de remoção deve ser realizada fornecendo a URL dos materiais a serem removidos, o período de tempo que se deseja excluir, o período de tempo durante o qual o requerente teve controle do site ou conta de usuário relevante (se aplicável) e qualquer outra informação que se considere útil para compreender a solicitação[106].

Na captura dos conteúdos e páginas da *web*, normalmente são estabelecidos alguns critérios, como obter o mandato de preservação de determinados *site*s por resoluções ou legislações, no caso de instituições de patrimônio e pesquisa (arquivos, bibliotecas, museus, institutos de pesquisa, universidades). Outra possibilidade inclui obedecer aos padrões de captura constantes no arquivo *robots.txt* de determinado *site*, que contém informações estabelecidas pelo dono do *site* e se ele pode ser rastreado ou não. O arquivo *robots.txt* constitui-se de um arquivo de texto simples e o seu uso adequado pode ajudar a garantir que apenas o conteúdo que o proprietário do *site* deseja arquivar seja capturado. O uso do arquivo *robots.txt*, ou protocolo de exclusão de robôs, foi proposto em 1994[107] como uma forma de impedir ou restringir o acesso de robôs de indexação de *site*s na *web*[108].

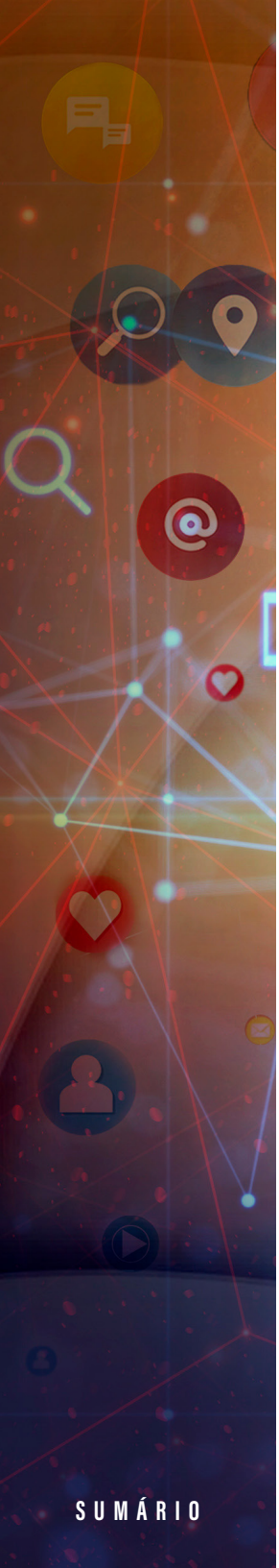
106 https://help.archive.org/help/using-the-wayback-machine/

107 https://web.archive.org/web/20131029200350/http://inkdroid.org/tmp/www-talk/4113.html

108 http://www.robotstxt.org/

3

MODELOS E PROCEDIMENTOS TÉCNICOS NO ARQUIVAMENTO DA *WEB*



Modelos de arquivamento da *web* referem-se às diferentes abordagens e métodos que podem ser usados para preservar e fornecer acesso a materiais baseados na *web*. Existem vários modelos diferentes de arquivamento da *web*, cada um com seus pontos fortes e limitações.

Alguns autores (Bragg; Hanna, 2013; Niu, 2012; Maemura *et al*. 2018) trazem alguns pontos relevantes a serem considerados no estabelecimento de modelos e fluxos de arquivamento da *web*:

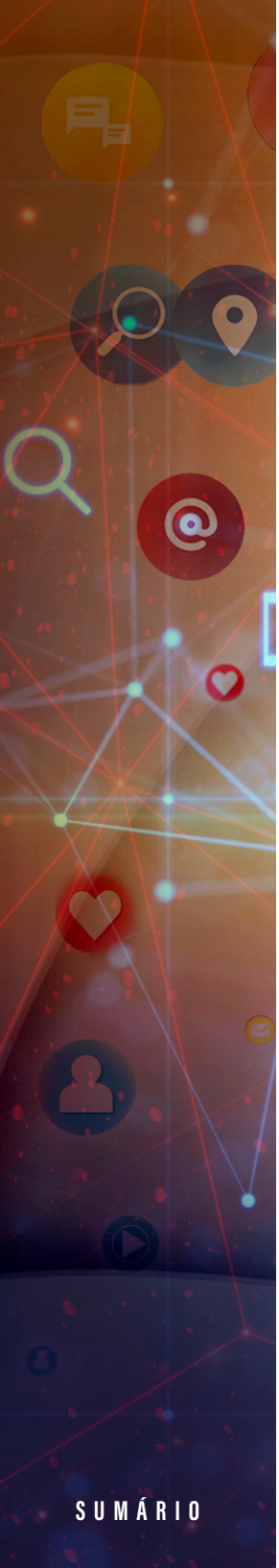
1. a avaliação e a seleção envolvem decidir quais materiais capturar e determinar a lista inicial de URLs para o rastreador da *web* começar as capturas;

2. as decisões de escopo determinam quais recursos serão capturados, com base em fatores como os domínios e tipos de mídia a serem incluídos, a duração e a frequência do rastreamento;

3. a aquisição ou captura de dados começa quando o rastreador da *web* acessa cada recurso em servidores da *web* ao vivo. Para usuários do serviço *Archive-It*, esse processo é automatizado, mas o rastreador gera relatórios e *logs*;

4. a organização e o armazenamento dos dados coletados são padronizados por meio do uso do formato de arquivo WARC e dos arquivos de índice que o acompanham. Para usuários do *Archive-It*, o armazenamento é limitado pelos orçamentos de dados incluídos em sua assinatura anual, e o gerenciamento de arquivos e servidores faz parte do serviço;

5. as revisões de garantia de qualidade verificam a qualidade e integridade do material arquivado, que pode envolver a revisão de *logs* e relatórios de rastreamento para identificar o número e o tamanho dos recursos capturados, bem como quaisquer problemas com as regras de escopo ou restrições de tempo do rastreamento. As páginas arquivadas individuais



também podem ser visualizadas em um navegador da *web* para identificar recursos ausentes ou instâncias em que os recursos da *web* ao vivo não foram registrados na coleção de arquivos da *web* e

6. Descrição e metadados ajudam os usuários a descobrir e acessar recursos arquivados. Os metadados podem ser aplicados em diferentes níveis de detalhe, desde informações gerais sobre a coleção até detalhes específicos sobre *site*s ou páginas individuais. No entanto, metadados mais granulares requerem tempo e esforço adicionais para descrição.

Um modelo de arquivamento da *web* é o modelo centralizado, no qual uma única organização ou grupo é responsável por coletar, armazenar e fornecer acesso a materiais baseados na *web*. Nesse modelo, a organização ou grupo atua como um repositório central de materiais baseados na *web* e é responsável por garantir que os materiais sejam preservados e disponibilizados aos usuários.

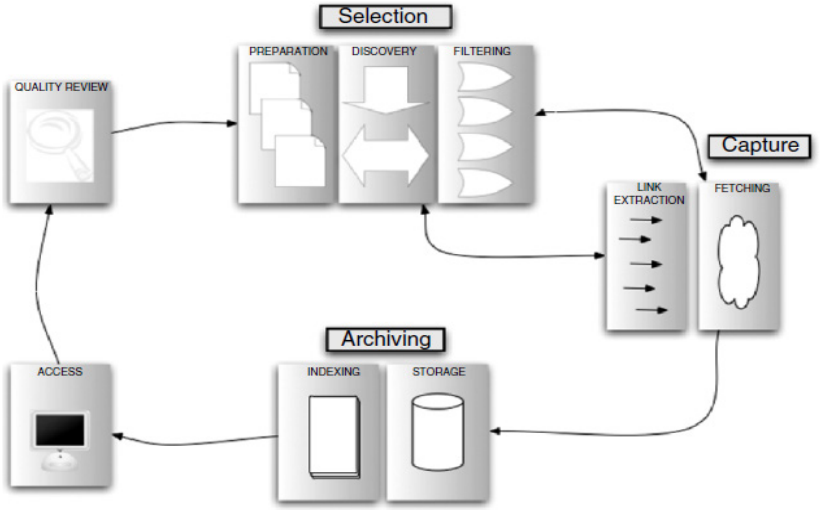
Outra forma de arquivamento da *web* é o modelo distribuído, no qual várias organizações ou grupos são responsáveis por coletar, armazenar e fornecer acesso a materiais baseados na *web*. Neste modelo, cada organização ou grupo atua como um repositório separado para materiais baseados na *web*, e os materiais são compartilhados e distribuídos entre os diferentes repositórios.

Uma outra via de arquivamento da *web* seria o modelo híbrido, que combina elementos dos modelos centralizado e distribuído. Nesse modelo, uma organização ou grupo central coordena a coleta, o armazenamento e o fornecimento de acesso a materiais baseados na *web*, mas os materiais são coletados e armazenados por várias organizações ou grupos.

O modelo usado dependerá das metas e circunstâncias específicas do arquivamento da *web*. Entretanto, podemos observar pontos em comuns em muitos dos modelos apresentados. O ciclo de

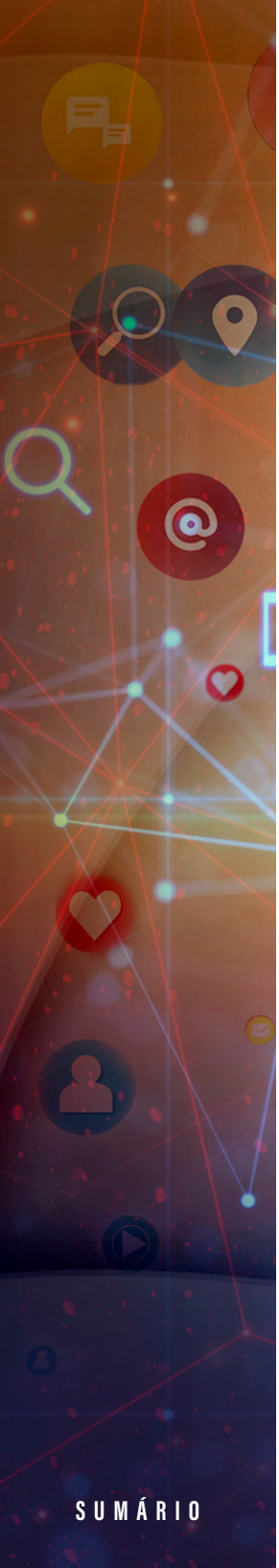
seleção, com as suas três fases: preparação, descoberta e filtragem, tem lugar antes da captura, do arquivamento e da revisão da qualidade, como mostra a Figura abaixo.

Figura 18 - Ciclo de seleção, captura e arquivamento



Fonte: Masanès (2006, p. 71).

A seleção é marcada pelas políticas de cada instituição, as escolhas determinarão o tipo, extensão e qualidade do conteúdo arquivado. Aplicar métodos e práticas desenvolvidas para a seleção de material impresso não é adequado, de acordo com Masanès (2006), a publicação da *web* é diferente o suficiente da publicação tradicional para exigir uma ampla revisão das práticas neste domínio. Porém, Biblarz e colaboradores (2001) colocam que construir coleções de material da *web* requer um documento orientador geral, que defina que o desenvolvimento da coleção trará os mesmos benefícios da política de desenvolvimento de coleções para material impresso, com o intuito de:



g. reduzir o viés pessoal ao tomar as decisões de seleção individual no contexto dos objetivos da prática de construção das coleções;

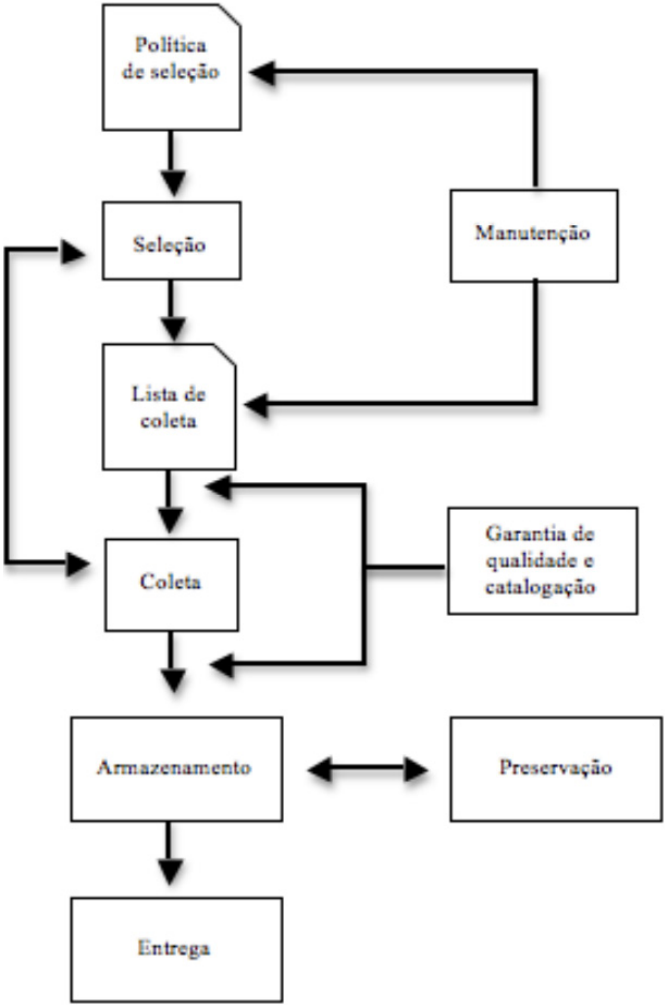
h. permitir planejar e identificar lacunas no desenvolvimento de coleções e garantir continuidade e consistência na seleção e revisão;

i. auxiliar na determinação de prioridades e esclarecer o objetivo e o escopo de cada coleção individual, além de permitir que as decisões de seleção sejam avaliadas, por exemplo, identificando qual proporção de material publicado dentro do escopo foi adquirido e

j. servir como base para uma cooperação mais ampla e compartilhamento de recursos.

Outro processo de arquivamento da *web* (fig. 16) foi descrito por Brown (2006):

Figura 19 - Processo de arquivamento da *web*

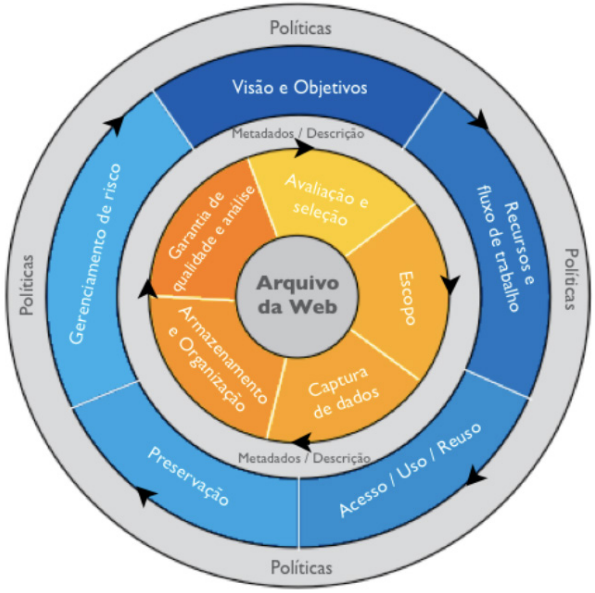


Fonte: Brown (2006, tradução de Rockembach, 2018).

Neste modelo, é possível perceber um fluxo que começa na política de seleção, a seleção propriamente dita, a lista de coleta, a coleta, garantia de qualidade, armazenamento, preservação e entrega (acesso).

Outro exemplo de modelo é o trazido por Bragg e Hanna (2013). Este modelo descreve as decisões de alto nível que uma organização enfrenta ao configurar e gerenciar seu programa de arquivamento da *web*, representado por um círculo azul. As etapas azuis, da Figura 20, incluem definir objetivos, revisar os recursos disponíveis, determinar políticas de acesso/uso/reutilização, tomar decisões de preservação e gerenciar riscos relacionados a direitos autorais e permissões.

Figura 20 – Etapas do ciclo de arquivamento da *web*

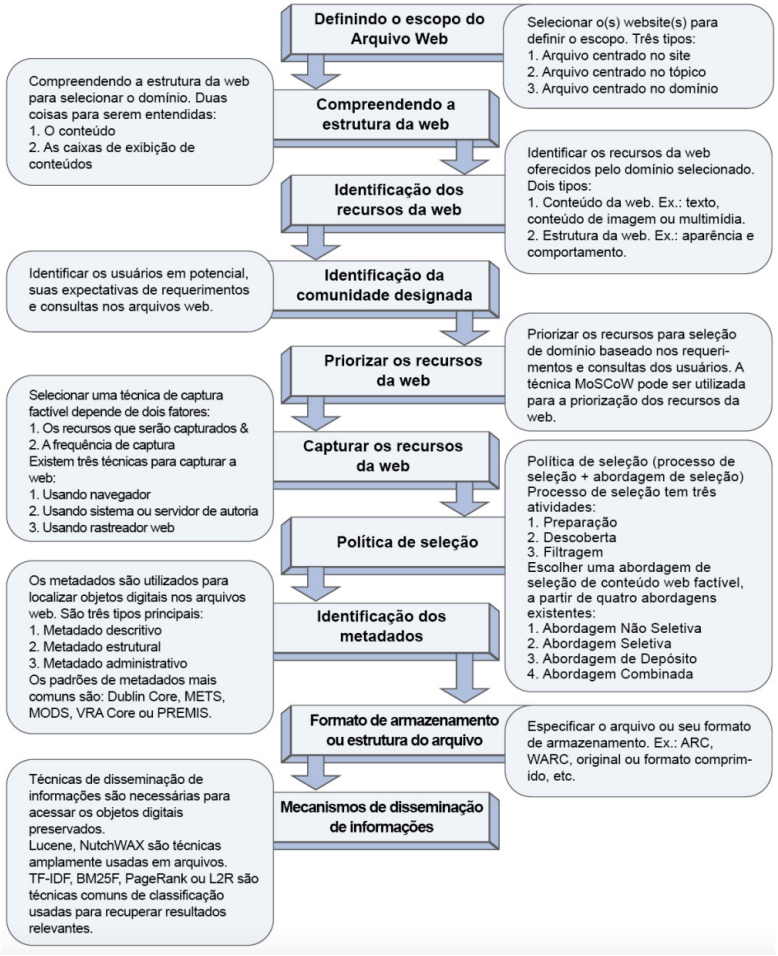


Fonte: Bragg e Hanna (2013, tradução de Rockembach, 2018).

Khan e Rahman (2019) estabeleceram um modelo de abordagem sistemática para a preservação da *web*. A abordagem sistemática de preservação é uma série de etapas que visa garantir a preservação a longo prazo de recursos na *web*. A primeira etapa é definir o escopo do arquivo, ou seja, determinar o tipo de *site*, tópico ou domínio que será preservado, a seguir, é importante entender

a estrutura da *web* e identificar os recursos, priorizando aqueles que serão mais relevantes para a comunidade designada e, em seguida, devemos escolher a técnica de captura mais adequada e criar uma política de seleção para determinar quais conteúdos da *web* serão capturados e preservados (fig. 21).

Figura 21 - Abordagem sistemática de preservação da *web*

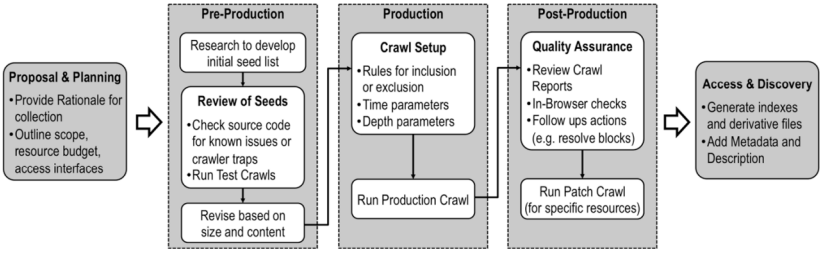


Fonte: Kran; Rahmam (2019, p. 74, tradução de Melo. 2020).

A etapa de preparação e seleção do conteúdo envolve a identificação de metadados e formato de arquivo e mecanismos de disseminação da informação. É importante ter em mente que a preservação a longo prazo de todos os tipos de *site*s não é viável, por isso é necessário priorizar os recursos mais relevantes e ter uma política clara de seleção. A abordagem sistemática de preservação ajuda a garantir a disponibilidade de informações antigas e importantes na *web* para usuários em potencial.

O processo utilizado por Maemura e colaboradores (2018) é composto por três etapas de seleção e captura: pré-produção, rastreamentos de produção e garantia de qualidade de pós-produção. Na fase de pré-produção, são realizados rastreamentos de teste para definir o tamanho e a abrangência da coleção, enquanto os rastreamentos de produção são responsáveis por capturar todos os recursos da *web* da coleção. Para garantir a qualidade na pós-produção, podem ser necessários rastreamentos adicionais e geralmente é comum que haja diversas iterações para cada uma dessas etapas, como podemos verificar no *workflow* da Figura 22.

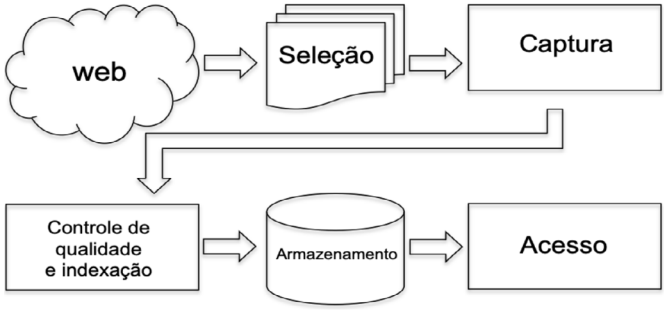
Figura 22 - Workflow para criação de arquivos da *web*



Fonte: Maemura et al. (2018, p. 1228).

Um modelo simplificado e didático do fluxo de arquivamento da *web*, proposto por Rockembach (2021), pode ser visto na Figura 23.

Figura 23 – Fluxo de arquivamento da *web*



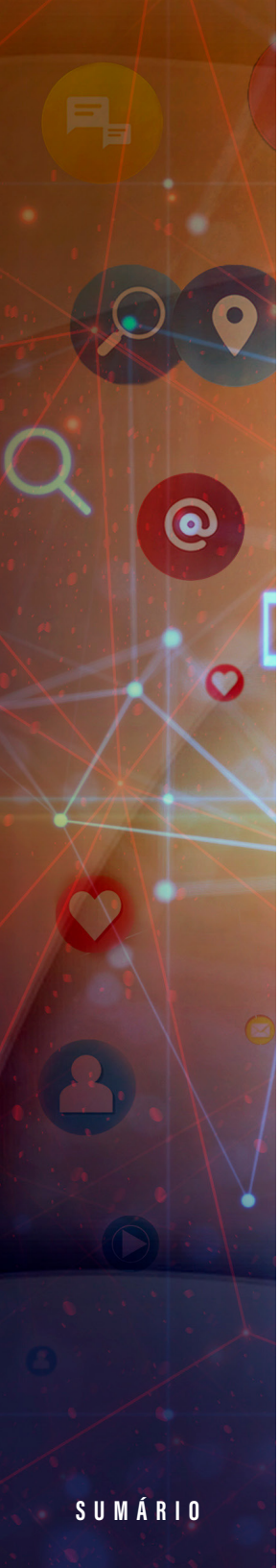
Fonte: Rockembach (2021).

O fluxo de arquivamento da *web* refere-se ao processo de coleta, armazenamento e fornecimento de acesso a materiais baseados na *web*. Esse processo, também conhecido como ciclo de vida do arquivamento da *web*, envolve várias etapas e atividades diferentes, que podem variar dependendo das metas e circunstâncias específicas do esforço de arquivamento da *web*.

Uma das primeiras etapas no fluxo de arquivamento da *web* é a identificação e seleção de materiais baseados na *web* a serem preservados. Isso envolve a identificação dos materiais específicos que devem ser preservados, bem como os critérios que serão usados para determinar quais materiais devem ser incluídos no arquivo da *web*.

Uma vez identificados e selecionados os materiais a serem preservados, o próximo passo no fluxo de arquivamento da *web* é a captura dos materiais, o que envolve o uso de ferramentas e *software* de arquivamento da *web* para capturar e armazenar cópias completas dos materiais, incluindo HTML, CSS e outros arquivos que compõem o conteúdo.

A partir da coleta dos materiais, a próxima etapa no fluxo de arquivamento da *web* é a preservação dos materiais. Isso envolve garantir que os materiais sejam armazenados de maneira a evitar que sejam perdidos ou degradados com o tempo. Isso pode envolver



o armazenamento de materiais em vários locais, usando técnicas de preservação digital ou outras medidas.

Uma vez que os materiais tenham sido coletados e preservados, o próximo passo no fluxo de arquivamento da *web* é o fornecimento de acesso aos materiais. Isso geralmente envolve disponibilizar os materiais aos usuários por meio de pesquisa baseada na *web* e interfaces de navegação, APIs ou outras ferramentas.

O fluxo de arquivamento da *web* envolve várias etapas e atividades diferentes, incluindo a identificação e seleção de materiais, a coleta de materiais, a preservação de materiais e o fornecimento de acesso a materiais. Essas atividades são essenciais para o processo de preservação e acesso a materiais baseados na *web*:

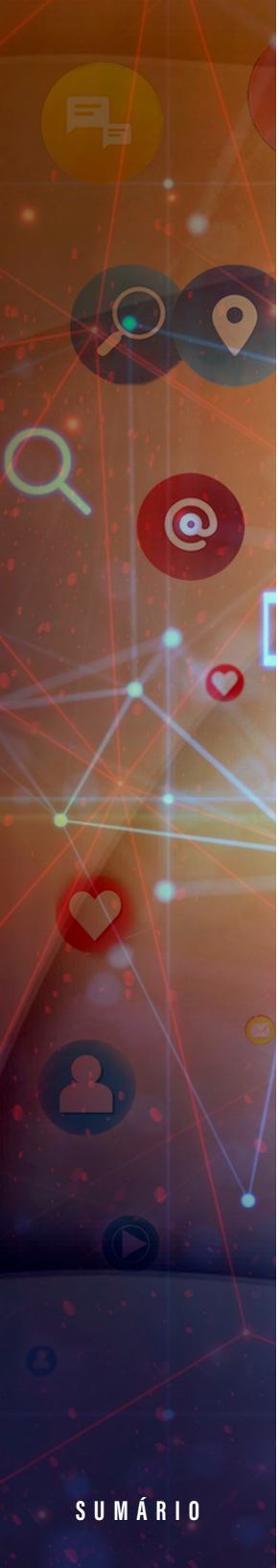
1. Identificar o *site* ou páginas da *web* a serem arquivados, incluindo a seleção do conteúdo específico que será preservado, como um *site* específico ou um grupo de páginas da *web* sobre um tópico específico.

2. Capturar o conteúdo selecionado, processo conhecido como *harvesting*. Isso envolve o uso de *software* especializado para "rastrear" o *site* ou as páginas da *web* selecionadas e capturar o HTML, as imagens e outros arquivos que compõem o conteúdo.

3. Armazenar o conteúdo capturado, salvando o material em um arquivo digital, como um arquivo da *web* em formato WARC.

4. Preservar o conteúdo armazenado. Isso envolve a manutenção da integridade do conteúdo armazenado ao longo do tempo, incluindo a verificação e o reparo de quaisquer erros ou corrupções.

5. Fornecer acesso ao conteúdo arquivado, disponibilizando o conteúdo arquivado aos usuários, seja por meio de uma interface baseada na *web* ou fornecendo acesso aos próprios arquivos digitais.



Por sua vez, a implementação de um projeto de arquivamento da *web* geralmente envolve as seguintes etapas:

1. Definir o escopo e os objetivos do projeto, identificando o conteúdo específico que será arquivado, bem como as razões para fazê-lo (por exemplo, para preservar informações históricas, fornecer acesso a materiais de pesquisa etc.).

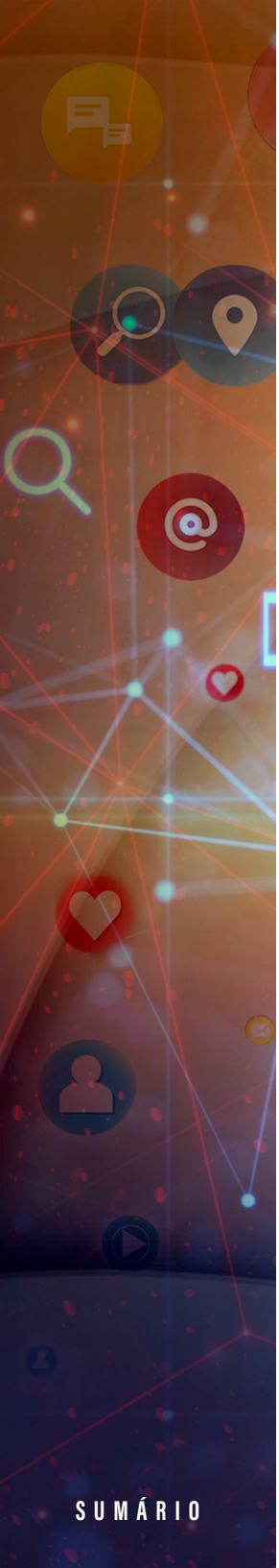
2. Selecionar as ferramentas e *software* a serem usados, parte que envolve a escolha das ferramentas e *software* de arquivamento da *web* que serão usados para capturar, preservar e armazenar o conteúdo. Esta etapa pode incluir rastreadores da *web*, ferramentas de preservação digital e plataformas de arquivamento da *web*.

3. Configurar a infraestrutura para armazenar o conteúdo arquivado, incluindo a configuração de servidores, sistemas de armazenamento e outros *hardwares* e *softwares* necessários para dar suporte aos arquivos da web.

4. A escolha e uso de ferramentas de arquivamento da *web* para capturar o conteúdo selecionado da *web* e salvá-lo no arquivo digital, envolvendo o uso de rastreadores da *web* para capturar sistematicamente grandes quantidades de conteúdo ou selecionar manualmente e salvar páginas da *web* específicas.

5. Preservar o conteúdo, garantindo a manutenção da integridade e acessibilidade do conteúdo arquivado ao longo do tempo, incluindo verificação de erros, reparo de corrupções e migração do conteúdo para novos sistemas de armazenamento conforme necessário.

6. Adotar estratégias de acesso ao conteúdo arquivado, disponibilizando aos usuários por meio de uma interface baseada na *web* ou outros meios, podendo envolver a criação de contas de usuário, o desenvolvimento de ferramentas de pesquisa e o fornecimento de suporte e treinamento aos usuários.



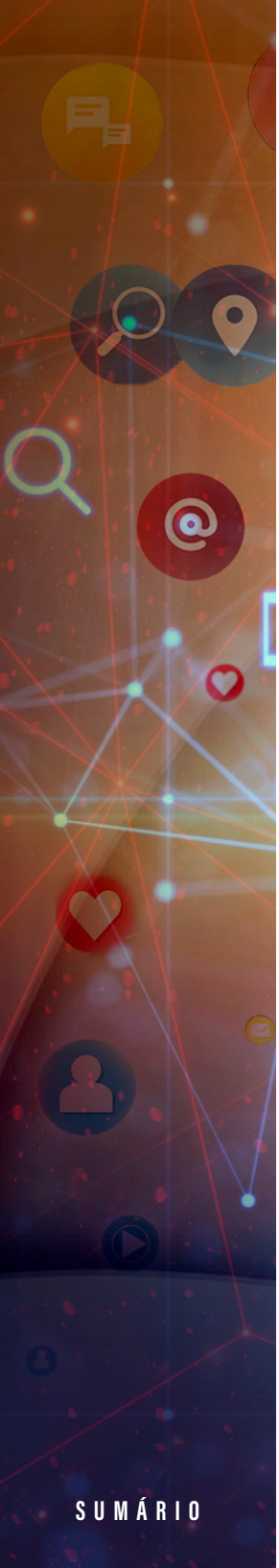
No Quadro 2 destacamos alguns dos pontos importantes que devem ser estabelecidos para o arquivamento da *web*.

Quadro 2 - Definições e conteúdo necessário para a captura e armazenamento da *web*

	Definições	Conteúdo
1	Período de captura	informar período de início e fim do rastreamento
2	Frequência de rastreamento	informar se periódico (quantificar) ou pontual (única vez)
3	Duração do rastreamento	informar horas:minutos:segundos
4	Número de sementes ativas	informar quantidade de sementes
5	Dados das sementes	www.exemplo.com (endereços a serem capturados)
6	Dados totais arquivados	em GB
7	Limites de rastreamento e regras especificadas	por exemplo, ignorar *robots.txt*
8	Detalhes de armazenamento	informar localização, se há restrição de acesso.

Fonte: os autores.

Em relação aos custos e investimentos na preservação digital, Corujo, Revez e Silva (2020) abordam a curadoria digital e seus custos em diferentes fases do processo. A partir da revisão de literatura, os autores identificaram questões a serem observadas em diversas etapas. Na produção de dados científicos em larga escala, há problemas de gestão e estimativas de custos, já na avaliação e seleção de materiais, devem ser considerados o tipo, estado, quantidade, acessibilidade e singularidade, assim como possibilidades de uso futuro. Na ingestão, o controle de qualidade é importante e os custos desta atividade devem ser analisados, na preservação digital, há custos recorrentes que dependem da gama de serviços oferecidos por uma instituição, e sua boa gestão é essencial para evitar problemas práticos. Na armazenagem, os custos de energia, espaço, refrigeração e gestão são crescentes,



e o armazenamento deve cumprir normas abertas e permitir a utilização de *hardware* de vários fornecedores, com mecanismos de acesso de dados flexíveis. O uso de serviços de computação em nuvem pode reduzir custos de armazenamento e aumentar a interoperabilidade com outros sistemas e serviços.

Shiozaki e Eisenschitz (2009) afirmam que os cálculos e a previsão de custos variam de instituição para instituição, mas podem ser amplamente categorizados em três aspectos: equipe, operação e sistema. Algumas instituições enfatizam o desenvolvimento do sistema, enquanto outras os custos com pessoal. Ainda se destacam aspectos particulares das operações diárias, como seleção, obtenção de permissões, processos de garantia de qualidade, indexação e preservação de longo prazo. Os benefícios do arquivamento da *web*, como a preservação do patrimônio cultural, são intangíveis e difíceis de medir, no entanto, é importante explicar o valor intrínseco do arquivamento da *web* quando se trata de gastos públicos. A automação do processo de arquivamento da *web* é uma forma de reduzir alguns dos custos atrelados à preservação digital.

No Quadro 3 sugerimos itens que devem ser considerados no levantamento de custos para a elaboração de um projeto de arquivamento da *web*, podendo ser personalizado conforme o contexto organizacional e os recursos disponíveis. Os valores a serem preenchidos irão variar conforme a extensão do projeto a ser realizado.

Quadro 3 – Itens e de descrição de custos a considerar no arquivamento da *web*

Item	Descrição	Custos / observações
Hardware	Servidores, computadores e armazenamento, bem como a conexão à Internet de alta velocidade.	Pode ser variável, caso os equipamentos sejam adquiridos ou seja contratado um serviço de nuvem (*cloud computing*).
Software	Licenças e treinamento inicial.	Pode variar conforme o uso de *software* livre ou proprietário, bem como contratação de serviço de nuvem (*cloud computing*).

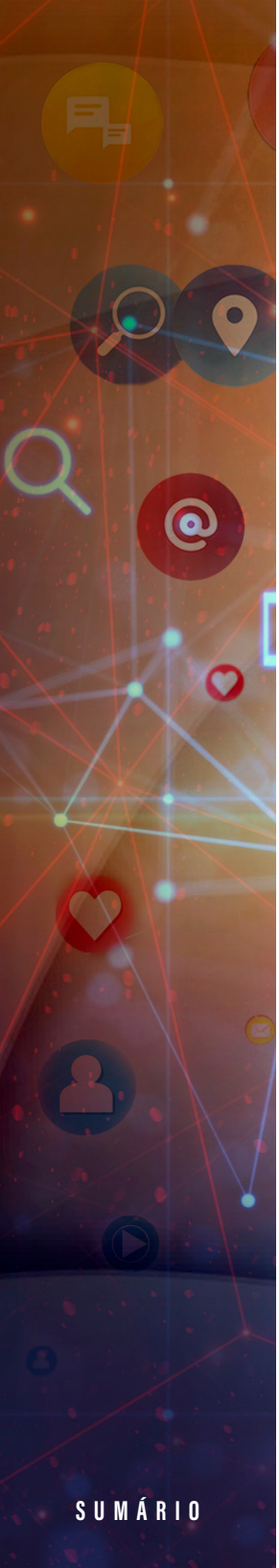
Pessoal	Gestão e execução das atividades envolvidas no processo de arquivamento da *web.*	A automatização de certas atividades pode auxiliar as equipes, entretanto, isso não pode sacrificar a qualidade do processo.
Divulgação e promoção	Promoção do arquivo da *web* para usuários e pesquisadores em potencial.	Pode envolver campanhas internas e externas à organização.

Fonte: os autores.

AVALIAÇÃO E CURADORIA DIGITAL

A avaliação, como uma das funções arquivísticas primordiais, consiste em analisar o que será mantido em caráter permanente ou eliminado. Com a *web*, temos uma quantidade significativa de informação produzida e que serve para muitos fins. O arquivo da *web*, com a funcionalidade essencial do que é ser um arquivo, mas com características distintas do entendimento de um arquivo tradicional, dada às especificidades da *web*, procurará ainda atender às mesmas utilidades. Neste sentido, Delmas (2010, p. 21) traz um importante conceito para levarmos em consideração, "Os arquivos servem para provar, lembrar-se, compreender e identificar-se. Provar seus direitos é uma utilidade jurídica e judiciária. Lembrar-se é uma utilidade de gestão. Compreender é uma utilidade científica de conhecimento. Identificar-se pela transmissão da memória é uma utilidade social".

A curadoria digital e o arquivamento da *web* são práticas intimamente relacionadas que envolvem a preservação e manutenção de conteúdo digital a longo prazo. A curadoria digital é o processo de gerenciar, organizar e preservar ativamente informações digitais para garantir sua autenticidade, confiabilidade e usabilidade ao longo do tempo. Isso envolve a seleção e aquisição de conteúdo digital, organizá-lo, indexar e fornecer acesso de maneira sustentável e fácil para os usuários navegarem. O arquivamento da *web*,



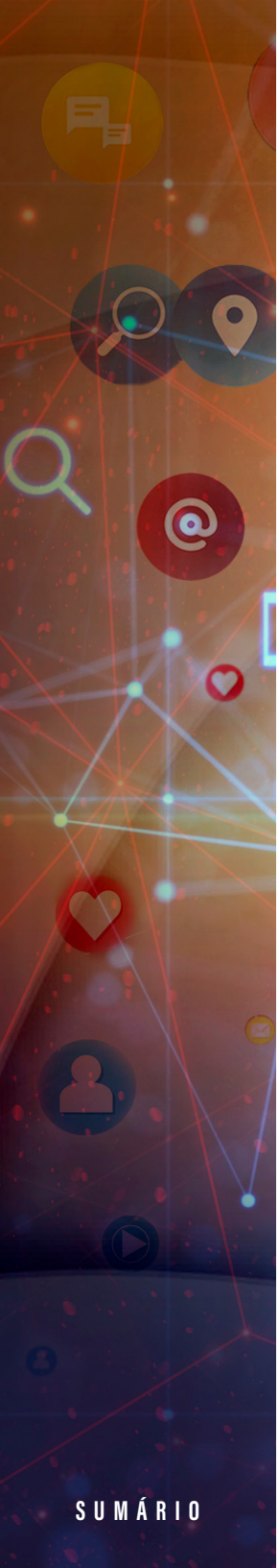
por sua vez, é a prática de preservar *site*s e outras informações baseadas na *web* para a posteridade. Esta atividade envolve a captura de todo o conteúdo de um *site*, incluindo todas as suas páginas, imagens, vídeos e outros elementos multimídia, e armazená-lo de forma que permita seu acesso e uso no futuro.

Outro aspecto relativo à curadoria é a possibilidade de participação cidadã ou da comunidade de usuários na seleção dos conteúdos a serem preservados. Ferramentas como a *Wayback Machine*, do *Internet Archive*[109], possuem a opção "Salve a página agora" (*Save page now*), o que possibilita a inclusão de novos conteúdos no arquivo da *web*. A plataforma Arquivo.pt[110], por sua vez, possibilita a sugestão de *site*s a serem arquivados por meio de formulário.

Segundo Recio, Zaldua e Virgil (2019), a seleção de conteúdos digitais levanta várias questões. Quem é responsável por decidir o que tem valor e deve ser preservado? Alguns conteúdos digitais são mais valiosos do que outros? Como determinamos o valor do conteúdo digital e quem deve ser responsável por preservá-lo? Além disso, o surgimento de plataformas de mídia social como *Instagram* e *YouTube* aumentou o fluxo de fotos e vídeos, mas quem será responsável por salvar todo esse conteúdo? A "Teoria do Equilíbrio" sugere que devemos nos concentrar na preservação de documentos e assuntos que foram historicamente importantes para fins criativos e de pesquisa. Isso significa considerar cuidadosamente o que deve e o que não deve ser preservado em formatos digitais e nas mídias sociais. Também levanta a questão de quem deve determinar o que é essencial e o que não é. Por fim, é importante encontrar um equilíbrio entre ter uma grande quantidade de informações que podem não ser úteis e ter uma seleção com curadoria de conteúdo valiosa e relevante.

109 https://archive.org/web/

110 https://arquivo.pt/sugerir

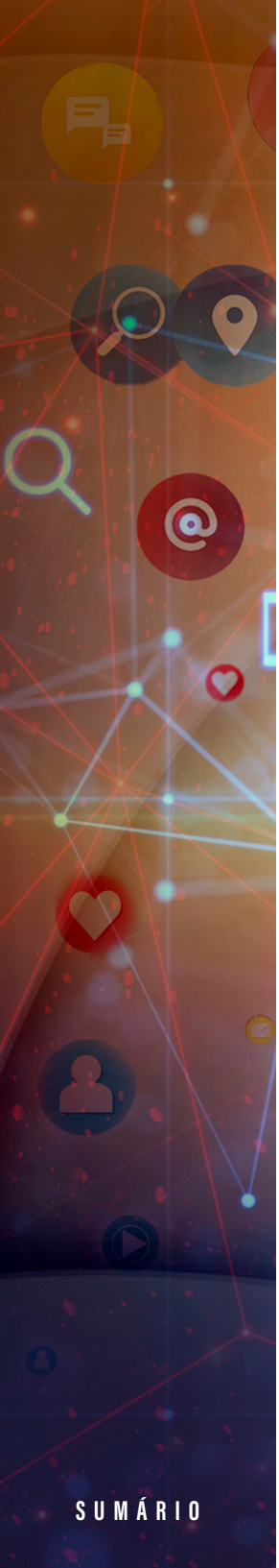


Tanto a curadoria digital quanto o arquivamento da *web* são importantes para garantir que a informação digital seja preservada e acessível com o passar do tempo. Com o rápido crescimento da Internet e a explosão do conteúdo digital, torna-se cada vez mais importante ter sistemas e processos para gerenciar e preservar essas informações. A curadoria digital e o arquivamento da *web* nos permitem garantir que o conteúdo digital valioso não seja perdido ou esquecido e que permaneça acessível e utilizável para pesquisadores, historiadores e outras partes interessadas.

A curadoria digital e o arquivamento da *web* estão voltados para a preservação do conteúdo digital, elas nos permitem manter um registro do mundo digital e sua evolução e usar esse registro para entender melhor nosso passado e informar no futuro. A curadoria digital refere-se ao gerenciamento ativo da informação digital ao longo do tempo, garantindo sua acessibilidade, usabilidade e integridade. Isso inclui selecionar, avaliar e organizar materiais digitais, bem como fornecer metadados e outras informações que ajudem os usuários a entender e contextualizar o conteúdo.

O arquivamento da *web*, por outro lado, concentra-se na preservação de conteúdo baseado na *web*, como *site*s, publicações em mídias sociais e outros materiais *on-line*. Isso geralmente envolve capturar e armazenar instantâneos de páginas da *web* em intervalos regulares, para que o conteúdo possa ser preservado e acessado no futuro. O arquivamento da *web* é importante porque a *web* está em constante mudança e, sem esforços ativos de preservação, informações valiosas e artefatos culturais podem ser perdidos.

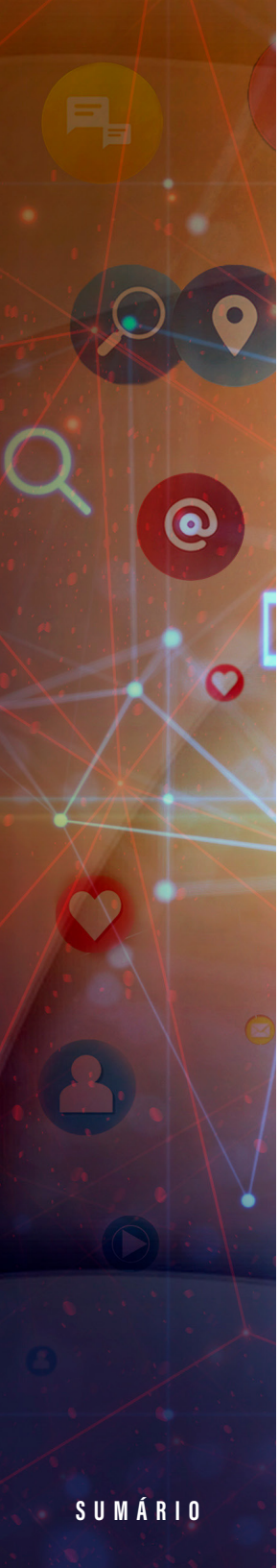
Tanto a curadoria digital quanto o arquivamento da *web* requerem uma abordagem sistemática de preservação, bem como o uso de ferramentas e tecnologias especializadas. Isso pode incluir *software* para organizar e gerenciar coleções digitais, bem como *hardware* para armazenar e fazer *backup* do conteúdo digital.



A curadoria digital e o arquivamento da *web* também desempenham um papel crítico no apoio à preservação do patrimônio cultural. Muitos arquivos, museus, bibliotecas e outras instituições culturais estão usando curadoria digital e arquivamento da *web* para preservar e fornecer acesso a documentos históricos, artefatos e outros tesouros culturais, permitindo que essas instituições tornem suas coleções mais amplamente disponíveis e acessíveis e as preservem para as gerações futuras. Alguns dos recursos que podem ser priorizados em determinados projetos de arquivamento da web são os *sites* com riscos de desaparecimento; originalidade, ou seja, páginas que contém recursos nato-digitais únicos; materiais raros e conteúdos com alta frequência de atualização.

Uma grande parte do patrimônio digital é gerado diretamente por indivíduos, quer trabalhem para instituições ou interajam com recursos patrimoniais de forma independente. Na *UK Web Archive* (UKWA)[111] utilizam as definições de patrimônio propostas por Bonacchi e Krzyzanska (2019) que oferecem uma caracterização do patrimônio digital como interações possíveis na Internet e os resultados de tais processos (as pegadas — incluindo dados — que são produzidos) e consideram a produção patrimonial como qualquer atividade, ocorrendo *on-line* ou *off-line*, com a qual humanos ou não humanos se envolvem. Baseada nesse conceito a UKWA, além de capturar o domínio da *web* que lhe permite um "quadro geral" do universo da *web* do Reino Unido, emprega uma abordagem seletiva incluindo coleções curadas por tópicos e temas. Tópicos e temas são grupos de *site*s reunidos sobre um tema específico por bibliotecários, curadores e outros especialistas, muitas vezes trabalhando em colaboração com organizações-chave no campo. Além disso, os curadores e outros especialistas se concentram regularmente na captura de *site*s sobre eventos, assuntos ou áreas de interesse específicos e os agrupam em tópicos e temas. Os curadores têm um peso epistemológico

111 https://www.webarchive.org.uk/en/ukwa/



substancial, porque decidem que material deve ser capturado e de que forma. (BONACCHI; e KRZYZANSKA, 2019).

Os parceiros acadêmicos do UKWA mostram preferência por selecionar e capturar conteúdo mais erudito e versátil enquanto outros curadores se concentram na seleção de conteúdo de pessoas comuns que estão envolvidas, por exemplo, em campanhas ou *hobbies on-line*. (BINGHAM; BYRNE, 2021). Com vista a aumentar a transparência no processo de arquivamento da *web* vários documentos de orientação foram produzidos por organizações individuais, tais como a *Documenting the Now*[112] projeto que desenvolve ferramentas de código aberto e práticas centradas na comunidade que apoiam a coleta ética, o uso e a preservação de conteúdo publicamente disponível compartilhado na *web* e nas mídias sociais e assim, integrar e documentar as escolhas feitas pelos arquivistas, bibliotecários e pesquisadores no processo de arquivamento da *web*.

Um dos principais desafios na análise de arquivos da *web* é documentar e comunicar suas limitações, o que é especialmente importante porque a natureza das ausências em uma captura pode não ser clara para os usuários. De acordo com diversos autores (BEN-DAVID; HUURDEMAN, 2014; MAEMUR; BECKER; MILLIGAN, 2016), pesquisadores, que normalmente não trabalham com fontes não baseadas na *web*, precisam fazer uma grande mudança em sua abordagem para trabalhar com arquivos da *web*. Isso exigirá que eles reflitam sobre seus métodos, avaliem criticamente as ferramentas e interfaces que usam para acessar arquivos da *web* e documentem sistematicamente essas ferramentas e interfaces.

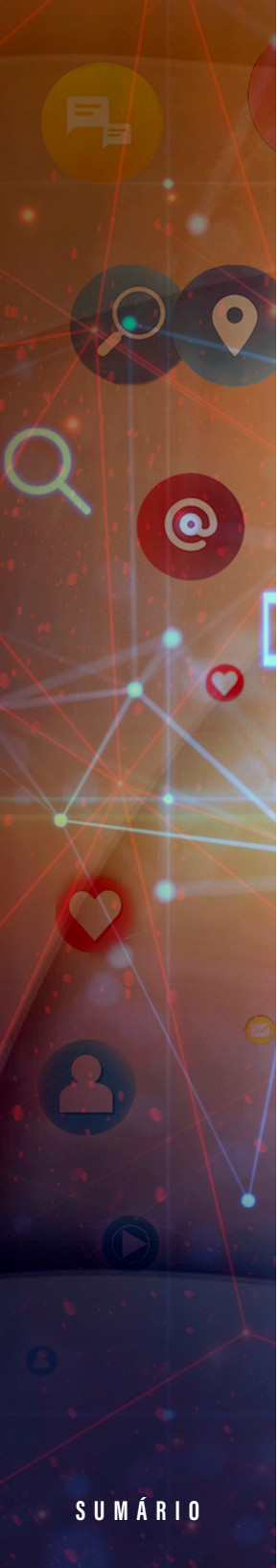
Maemura e colaboradores (2018) sugerem documentar a proveniência a partir dos seguintes elementos de escopo, processo e contexto (Quadro 4, 5 e 6):

112 http://www.docnow.io/

Quadro 4 - Elementos de escopo

Elemento	Principais questões e informações a documentar
Motivação	- Qual é o propósito da coleção? - Seu mandato mudou ao longo do tempo?
Foco	- Quais limites geográficos, temporais, técnicos, políticos, temáticos e/ou sociais são definidos para delimitar a coleção?
Acesso e descoberta	- Qual é o público-alvo? Eles têm características ou necessidades conhecidas? - Quais acordos contratuais, organizacionais, legais ou outros restringem o acesso? - Quais campos de metadados e índices apoiam a descoberta? - Em que grau de granularidade (por coleção, *site* ou recurso individual)? - Quais formatos de dados ou conjuntos de dados derivados estão disponíveis?
Lista de sementes	- Quais sementes foram usadas no escopo da coleção? - Qual foi o processo de descoberta e seleção de sementes?
Frequência de rastreamento	- Qual é a frequência dos rastreamentos? - Quanto tempo os rastreamentos duram ou qual é o limite de tempo definido?
Configuração de rastreamento	- Quais configurações controlam a profundidade do rastreamento? Por exemplo, configurações para capturar a distância a partir da semente original. - O objetivo é ter uma coleção mais abrangente ou mais focada em abrangência?
Inclusões e exclusões	- Certos *sites* ou tipos de mídia são incluídos ou excluídos? Por exemplo, expressões regulares são usadas para direcionar determinados arquivos ou diretórios em uma estrutura de URL.
Permissões dos Administradores do *site*	- Como foram tratadas as restrições dos administradores do site, como robots.txt e bloqueios?

Fonte: Maemura et al. (2018).



Segundo Maemura e colaboradores (2018), o escopo determina quais recursos da *web* serão capturados e em que período de tempo, e as decisões curatoriais para esse escopo variam de posições estratégicas de nível superior a escolhas operacionais de nível inferior. A motivação para iniciar a preservação de *website*s pode envolver colaboração ou coordenação entre diferentes atores ou instituições, e as decisões podem envolver configurações específicas do rastreador. Essas decisões se sobrepõem e as frequências de rastreamento e as listas iniciais podem depender dos *site*s de destino e serem geradas de outras fontes.

Quadro 5 - Elementos do processo

Elemento	Principais Questões e Informações a Documentar
Eventos agendados	- Como são tratados os eventos agendados? Por exemplo, a conclusão de um rastreamento e geração de relatórios de rastreamento provoca novas decisões, como a reapreciação do conteúdo capturado?
Eventos não agendados ou anomalias de processo	- Quais ações ou decisões são tomadas em resposta a eventos não agendados ou imprevistos? Por exemplo, quando os administradores do *site* são contatados diretamente? - Qual é o processo para identificar e capturar recursos que estão faltando em uma coleção?

Fonte: Maemura et al. (2018)

Ainda conforme Maemura, e colaboradores (2018), vários eventos programados e não programados podem ocorrer durante o processo de arquivamento da *web*, como erros de HTTP, anomalias de rastreamento e restrições de *site*, que podem afetar a coleta resultante. O contexto em que as decisões curatoriais são tomadas também é discutido, o que inclui fatores organizacionais e ambientais que orientam e moldam as atividades de arquivamento da *web* e seus resultados. As decisões tomadas devem estar alinhadas com a missão e os mandatos legais da organização, bem como cumprir os regulamentos e

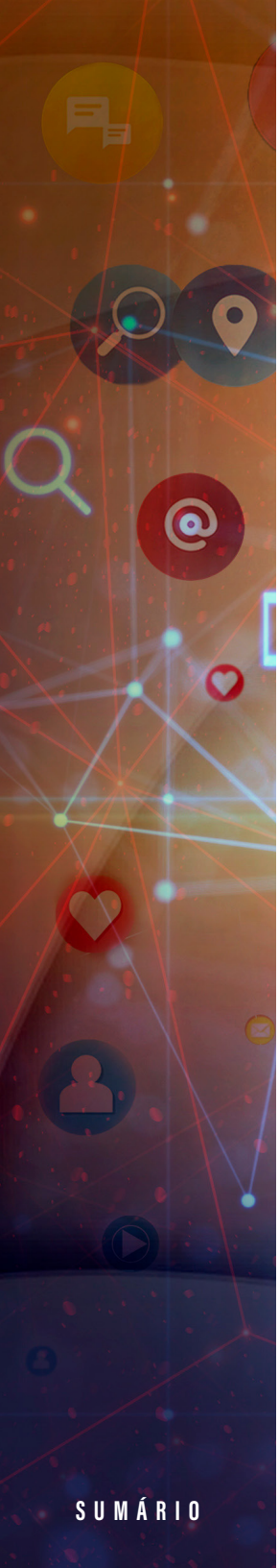
restrições. Essas decisões podem ser documentadas quando as instituições desenvolvem programas de arquivamento da *web*, incluindo a determinação de metas, alocação de recursos, princípios de provisionamento de acesso e abordagem de gerenciamento de risco.

Quadro 6 - Elementos de contexto

Elemento	Principais Questões e Informações a Documentar
Contexto legal	Quais leis e regulamentos se aplicam à instituição e suas atividades de arquivamento da *web*? Por exemplo, leis de direitos autorais, mandatos de depósito legal, acordos e contratos de usuário.
Ambiente e mandato	Qual é o compromisso organizacional com o arquivamento da *web*? Qual é o papel do arquivamento da *web* dentro da organização?
Políticas e diretrizes	Quais políticas ou regulamentos organizacionais afetam as atividades de arquivamento da *web*? Existem políticas que orientam e limitam especificamente as atividades de arquivamento da *web*? Por exemplo, definindo abordagens para permissões, restrições de acesso ou a divisão de responsabilidades entre departamentos. Essas políticas mudaram ao longo do período de tempo de uma coleção?
Recursos disponíveis	Quais recursos de pessoal dedicados apoiam o arquivamento da *web*? Qual *software* ou plataforma de rastreamento está em uso? Quais limites de armazenamento ou orçamentos de dados limitam a coleta? Quando os recursos são significativamente aumentados ou diminuídos ao longo do período de tempo de uma coleção?

Fonte: Maemura et al. (2018).

O problema da falta de contextualização pode ser resolvido pela descrição da procedência dos arquivos (*web archives provenance*). Maemura e colaboradores (2018) referem-se à proveniência como uma estratégia de documentação que permite prover o usuário de informação sobre como uma coleção *web* foi conformada, o rastreio de um conjunto de dados até à sua origem. Verifica-se a necessidade de novas perspectivas sobre a proveniência dos arquivos da *web*. Os autores identificaram um conjunto inicial de elementos necessários para abordar a proveniência dos arquivos da *web* com o propósito de facilitar a elucidação e documentação de decisões.



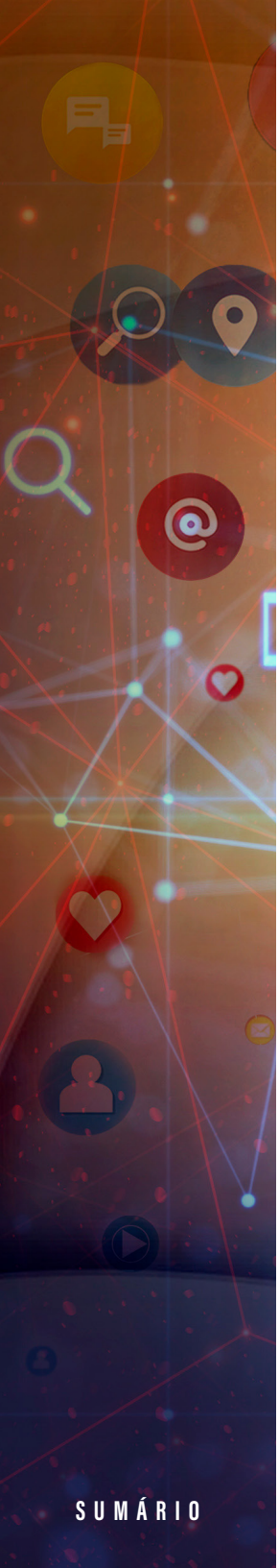
O Escopo é um dos elementos da proveniência dos dados que se refere às decisões de curadoria no âmbito de uma coleção, determinam quais os recursos da *web* que serão capturados, e em que período de tempo. O Processo, outro elemento da proveniência, deve ser monitorado em diferentes graus, e as ações e decisões tomadas em resposta a diferentes eventos de uma captura podem influenciar a presença ou ausência de recursos arquivados e, consequentemente, na coleção resultante. Tanto os eventos programados como não programados podem desencadear uma renegociação de escopo ao longo de todo o processo de captura.

Outro elemento importante a ser documentado na proveniência dos dados arquivados é o contexto no qual as decisões de curadoria são tomadas, incluindo uma série de fatores ambientais que guiam, moldam e condicionam as atividades de arquivamento da *web* e os seus resultados. Quando o arquivamento da *web* tem lugar dentro de uma organização, estas atividades devem alinhar-se com a missão ou mandato legal da organização para a captura, bem como o cumprimento de qualquer regulamento ou mandato legal, restrições organizacionais, técnicas e culturais.

O arquivista da *web* deve fazer-se alguns questionamentos, como os apontado acima sobre escopo, processo e contexto, para não negligenciar nenhum dos aspectos que necessita documentar para informar os usuários sobre todas as questões de proveniência e assim conferir maior transparência às coleções do arquivo da *web*.

Além dos aspectos técnicos da preservação, a curadoria digital e o arquivamento da *web* também exigem planejamento e colaboração cuidadosos, o que pode envolver trabalhar com outras instituições e organizações para compartilhar recursos e conhecimentos, bem como se envolver com comunidades de usuários para entender suas necessidades e garantir que as coleções digitais sejam relevantes e úteis.

No contexto da curadoria digital alguns aspectos são desafiadores conforme Maemura e colaboradores (2018), a curadoria de

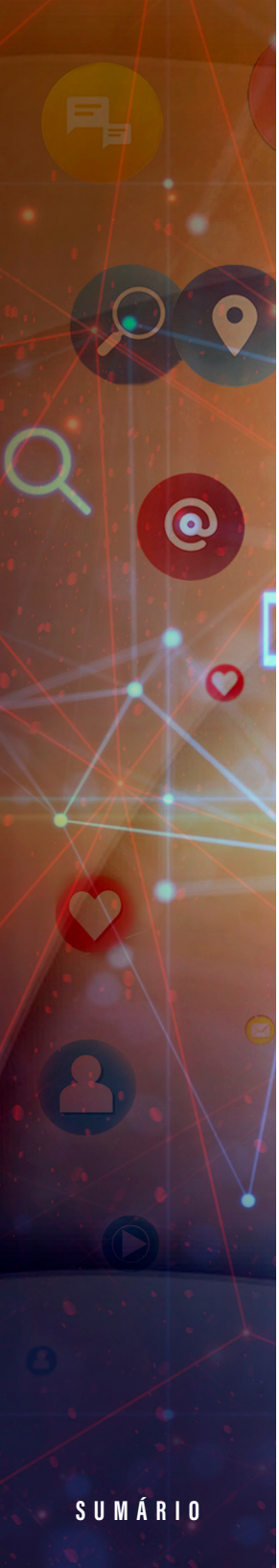


dados envolve rastrear as origens de um conjunto de dados, um processo conhecido como proveniência de dados. Isso é importante para o compartilhamento e reutilização de dados, pois ajuda os pesquisadores a entender as origens de um conjunto de dados e como ele foi criado. A proveniência dos dados é frequentemente associada à *eScience*, mas outras formas de representação do conhecimento são necessárias para entender as origens dos dados na perspectiva das Ciências Sociais e Humanas. As origens dos dados tornaram-se uma grande preocupação na pesquisa em Ciências Sociais envolvendo métodos digitais e "*big data*", bem como nas Humanidades Digitais, o que levou ao desenvolvimento de iniciativas de curadoria de dados com foco em Humanidades.

Vários fatores, incluindo mudanças contínuas nas políticas, novos tipos de arquivos da *web* e variações nos recursos de pessoal e orçamentos de dados, impactam a forma como as atividades de arquivamento da *web* são realizadas. Também, há uma disparidade de critérios entre arquivistas e arquivistas e bibliotecários ao decidir quais os conteúdos digitais que devem ser preservados e por quem. Porém, algumas ponderações importantes devem ser levadas em conta ao selecionar o conteúdo para arquivamento da *web*.

SELEÇÃO NO ARQUIVAMENTO DA *WEB*

As decisões sobre o que deve ser arquivado dependem de muitos fatores, podemos citar o tempo e a infraestrutura disponíveis (largura de banda, IPs, pessoal), o julgamento de valor e as prioridades das partes interessadas. Em geral, o arquivamento da *web* baseia-se numa mistura de APIs (como no caso do arquivamento de redes sociais) e *web* semi automatizada, onde rastreadores indexam e descarregam conteúdos de forma recorrente. Os rastreadores são determinados por "sementes" (*seeds*) que são consideradas instruções que



incluem URLs alvos, a "profundidade" e o alcance da captura. Ainda, para determinar quais serão as sementes devemos levar em conta:

a. Relevância: o conteúdo deve ser relevante para o propósito do arquivo da *web* e para o público a que se destina.

b. Pontualidade: o conteúdo deve estar atualizado e refletir o estado atual da *web*.

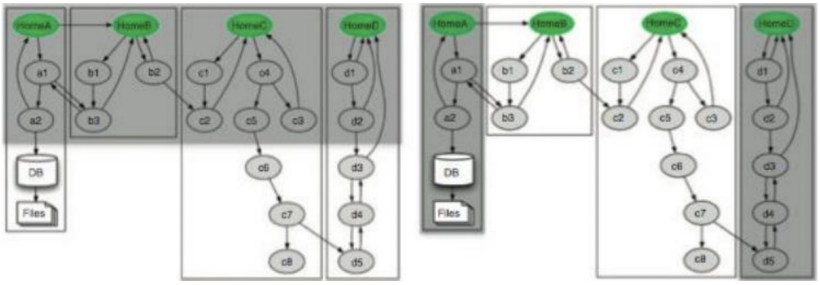
c. Autenticidade: o conteúdo deve ser autêntico e não alterado de forma alguma.

d. Integridade: o conteúdo deve ser completo e não faltar nenhum elemento importante.

e. Conformidade legal: o conteúdo deve ser coletado em conformidade com todas as leis e regulamentos relevantes, tais como o *General Data Protection Regulation* (GDPR) na União Europeia e a Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709/2018, no Brasil.

A profundidade de captura foi abordada por Masanès (2006) e podem ser classificadas de duas formas: seleção extensiva e seleção intensiva. A seleção extensiva visa cobrir os domínios nos primeiros níveis de forma mais ampla, proporcionando uma visão geral da *web* a partir do arquivamento. Já a seleção intensiva concentra-se em alguns *site*s para arquivar o máximo de níveis possíveis, incluindo elementos como bancos de dados. Enquanto a seleção extensiva aborda apenas a superfície da *web*, a seleção intensiva requer maior acessibilidade aos sistemas e servidores, porém permite preservar não apenas os primeiros níveis, mas toda a hierarquia, mantendo a navegação entre os *hiperlinks* ativa no arquivo da *web* (fig. 24).

Figura 24 – Comparação entre os tipos de arquivamento extensivo e intensivo



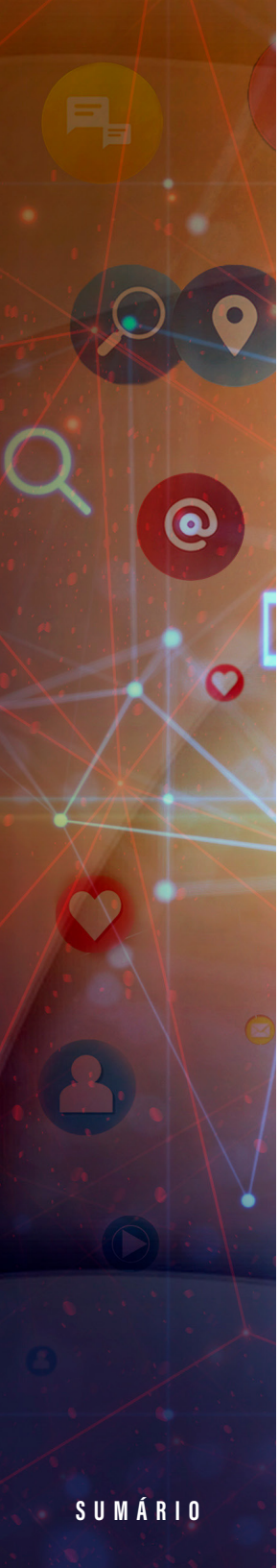
Fonte: Masanès, 2006, p. 39-40

Para garantir que o conteúdo selecionado para arquivamento da *web* atenda a esses critérios, os arquivistas da *web* podem usar uma variedade de ferramentas e técnicas, como rastreadores da *web* para identificar conteúdo relevante, verificações para aferir a qualidade do conteúdo arquivado e revisão manual por equipe treinada para avaliar a integridade e conformidade legal do conteúdo.

Existem várias abordagens diferentes que podem ser utilizadas na seleção de *site*s para arquivamento da *web*, dependendo das metas e objetivos do arquivo da *web*. Algumas abordagens comuns para a seleção de arquivamento da *web* incluem:

1. Amostragem representativa: essa abordagem envolve a seleção de uma amostra representativa de *site*s e outros conteúdos *on-line* que reflitam a diversidade e o alcance das informações disponíveis na *web*. Essa abordagem é útil para criar uma visão ampla da *web* e sua evolução ao longo do tempo.

2. Coleta direcionada: essa abordagem envolve a seleção de *site*s específicos e conteúdo *on-line* que sejam relevantes para as metas e objetivos do arquivo da *web*. Por exemplo, um arquivo da *web* focado em um tópico específico, como mudança climática, pode usar essa abordagem para coletar *site*s e conteúdo *on-line* relacionados à mudança climática.

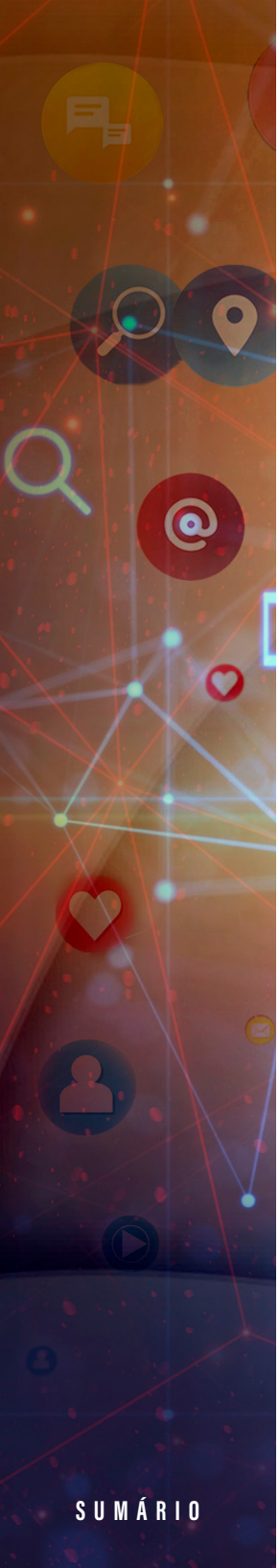


3. Coleta baseada em eventos: essa abordagem envolve a coleta de *sites* e conteúdo *on-line* relacionados a eventos específicos, como desastres naturais ou eleições políticas. Essa abordagem é útil para estudar o impacto dos eventos na *web* e as formas pelas quais a *web* é usada para documentar e discutir esses eventos.

Conforme Summers (2020), avaliar arquivos da *web* envolve um processo complexo de análise e tomada de decisão, isto exige que os arquivistas considerem uma variedade de fatores, incluindo o conteúdo do arquivo da *web*, o contexto em que foi criado e o valor potencial do arquivo para pesquisa e outros propósitos. Os profissionais também devem considerar os aspectos técnicos do arquivamento da *web*, como as ferramentas e técnicas usadas para capturar e preservar o conteúdo, além das implicações éticas. Summers (2020) cita Ketelaar (2001) e o seu conceito de "arquivalização" ou "*archivalization*", que precede o arquivamento e preservação.

A arquivalização de Ketelaar (2001) relaciona-se com a avaliação e seleção e significa estabelecer o que merece, ou é digno, de ser arquivado, uma ação que pode acontecer de forma consciente ou inconsciente. Este processo de julgamento acontece antes da decisão de avaliação e é uma representação de como os profissionais interpretam as informações e o que deve ser preservado, incluindo fatores culturais e sociais.

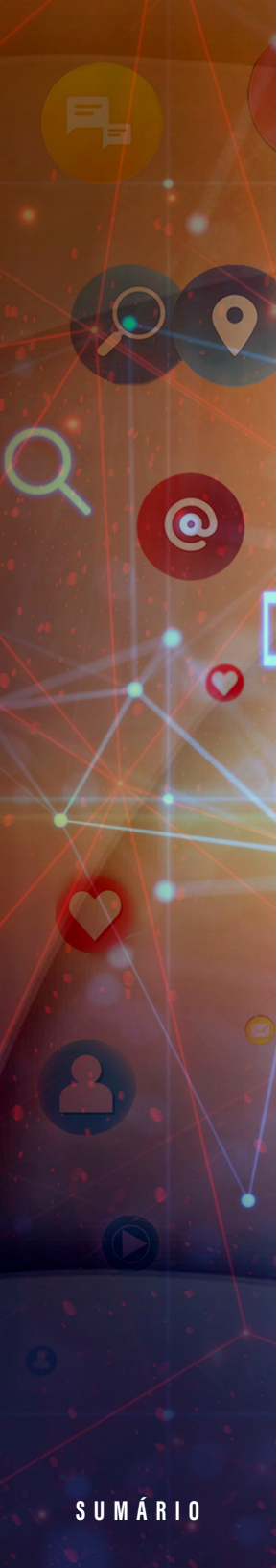
Outro aspecto apontado por Summers (2020) é de que a ideia de transferir ou recolher os arquivos quando eles não são mais usados pelo produtor não se aplica, pois, dada a dinamicidade da *web* e complexidade de formatos e arquivos, é justamente no momento que os *sites* e conteúdo são removidos da *web* é que eles ficam indisponíveis também para o arquivista da *web* e para a preservação digital. As ferramentas atuais também permitem a captura e preservação com menos interação com o proprietário do conteúdo. Neste caso há dois lados, um o de que uma menor interação pode



dificultar o processo de transferência e recolhimento, mas que por outro lado impede que as informações sejam apagadas pelo próprio produtor ou administrador do *site*, pois são capturadas e preservadas nos arquivos da *web* de maneira ativa.

A arquitetura da *web*, portanto, não é desenhada na relação tradicional entre produtor e custodiador, onde o segundo recebe os documentos gerados pelo primeiro, mas em um papel proativo do custodiador em capturar os *site*s produzidos. Entre outras diferenças com os documentos tradicionais, há de se considerar também o versionamento dos *site*s, podendo produzir atualizações muito mais frequentes e a conexão com outros *site*s e documentos externos ao escopo inicial, algo que é típico da arquitetura da *web*, uso de *hiperlinks* e conexão em rede. A falta destes elementos para arquivos da *web* pode proporcionar menos contexto informacional, e o contexto é algo considerado essencial para o entendimento dos arquivos. Além disto, o próprio conceito de territorialidade é transformado na *web*, e em consequência nos arquivos da *web*, pois o conteúdo de uma organização pode estar em mais de um domínio e não circunscrito a um domínio nacional, como é o caso do .br, no Brasil. Um domínio pode ser .com.br, mas também .com ou .net, por exemplo. Ainda, conteúdos relacionados podem estar registrados em domínios de outros países.

Outro ponto fundamental consiste em documentar o processo que foi adotado para avaliar, selecionar, capturar e armazenar os arquivos da *web*. Isto influencia diretamente a compreensão de como os arquivos da *web* são compostos e seus possíveis usos. Uma pesquisa em uma plataforma de arquivamento da *web* que não contém o processo documentado, a forma de seleção e a captura, pode levar à incompreensão da abrangência do arquivo da *web*. Portanto, a transparência quanto aos processos de rastreamento e sobre o que se conseguiu preservar é um fator importante para o uso dos arquivos da *web* como fonte de pesquisa.

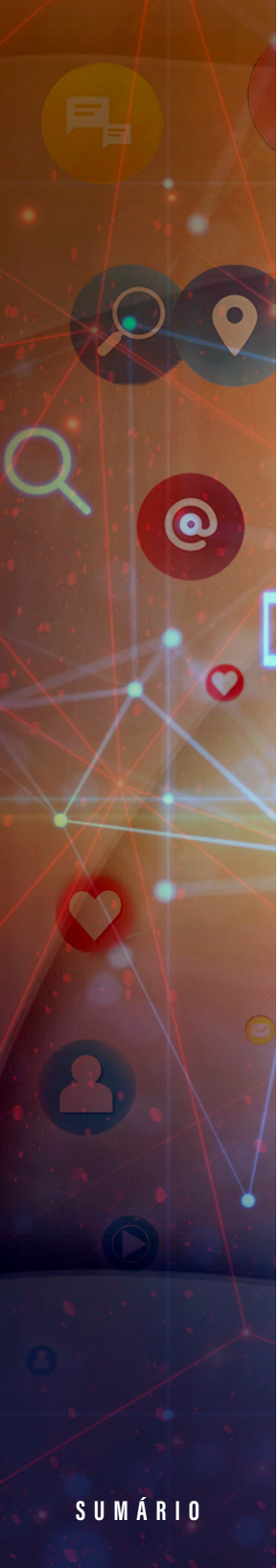


Post (2017) defende que as instituições arquivísticas e demais instituições de patrimônio devem desenvolver métodos para selecionar ativamente as páginas *web* para preservação, com a criação de arquivos da *web* que constituam um registro permanente. Outra questão apontada seria encontrar o equilíbrio entre capturar os materiais da *web* necessários para preservar o contexto, sem, entretanto, lançar-se na ideia utópica de preservar todo o conteúdo dinâmico da *web*. Segundo o mesmo autor, é possível levar em consideração as teorias de avaliação, abordando questões de uso e valor, mas também se faz necessário ampliar a teoria e a prática, procurando construir comunidades mais amplas em torno do arquivamento da *web*.

A preservação da *web* envolve um conjunto de processos e atividades que garantem o armazenamento sustentado de longo prazo, acesso e interpretação de informações digitais. As taxas de crescimento e declínio do conteúdo e a importância das informações apresentadas na *web* tornam um *site* candidato-chave para preservação. Porém, a preservação da *web* enfrenta uma série de desafios devido à sua estrutura complexa, uma variedade de formatos disponíveis e o tipo de informação.

O *layout* também pode variar de domínio para domínio com base no tipo de informação e sua apresentação. De acordo com Khan e Ur Rahman (2019) os *site*s podem ser categorizados com base em dois aspectos: a) o tipo de informação, ou seja, o conteúdo e b) a forma como essas informações são apresentadas, ou seja, o *layout* ou estrutura da página *web*. Portanto, é praticamente inviável desenvolver um único sistema para preservar todos os tipos de *site*s para o longo prazo. Os mesmos autores aconselham que, antes de iniciar o trabalho, o arquivista da *web*, defina o âmbito da *web* a ser arquivado, podendo variar em:

a. Arquivamento centrado em *site*, concentra-se em um *site* específico para preservação. Esses tipos de arquivos são principalmente iniciados pelo criador ou proprietário do *site*. Permite o acesso a versões antigas do *site*.



b. Arquivamento centrado em tópicos, criados para preservar informações sobre um determinado tópico publicado na *web* para uso futuro. Uma das utilizações desse arquivamento é para verificação científica, quando os pesquisadores precisam consultar as informações, embora seja difícil garantir o acesso a esses conteúdos devido à natureza efêmera da rede.

c. Arquivo centrado em domínio, a palavra "domínio" refere-se a um local, rede ou extensão da *web*. Um arquivo centrado no domínio abrange *site*s publicados com um nome de domínio DNS específico. Vários projetos têm um âmbito centrado no domínio, por exemplo, o Arquivo da *web* Portuguesa (PWA), o *Kulturarw3*[113], uma coleção de arquivos da *web* da *National Library of Sweden*[114] do domínio .se e .com e a coleção do *UK Government web Archive*[115] de *site*s do governo britânico, por exemplo, *site*s de domínio .gov.uk.

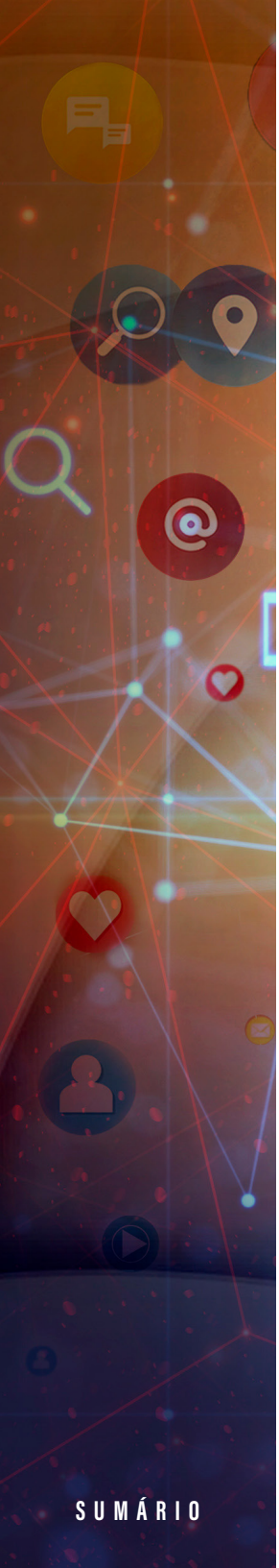
Outro aspecto necessário a ser considerado pelo arquivista da *web* diz respeito aos tipos de recursos que podem ser preservados, conteúdo textual (texto simples), conteúdo visual (imagens) e conteúdo multimídia. Também, precisamos decidir sobre a estrutura da *web* que se deseja preservar, ou seja, se é importante preservar a aparência, o *layout* geral de apresentação da página *web* para garantir a representação do conteúdo e/ou o comportamento (código de navegação) representado por *links* de navegação que podem estar dentro do *site*, *links* de documentos externos ou recursos dinâmicos e animados, como *feed* ao vivo, comentários, marcação ou marcação de favoritos.

Para Khan e Ur Rahman (2019) a complexidade dos recursos da *web* e sua representação podem causar complicações no processo de preservação digital. Os autores acreditam que em alguns

113 https://www.kb.se/hitta-och-bestall/hitta-i-samlingarna/kulturarw3.html

114 https://www.kb.se/in-english.html

115 https://www.nationalarchives.gov.uk/webarchive/



casos, pode ser indesejável ou inviável preservar todos os recursos *web*, portanto, selecionar os recursos da *web* para preservação torna-se prioridade e essa pode ser a primeira decisão, pela potencial reutilização do recurso e segundo, a frequência com que o recurso será acessado. Os recursos sem valor, pouco valor, ou aqueles gerenciados em outro lugar podem ser excluídos da preservação. Na seleção do recurso que será preservado sugerem a utilização do método *MoSCoW*, o acrônimo pode ser entendido como:

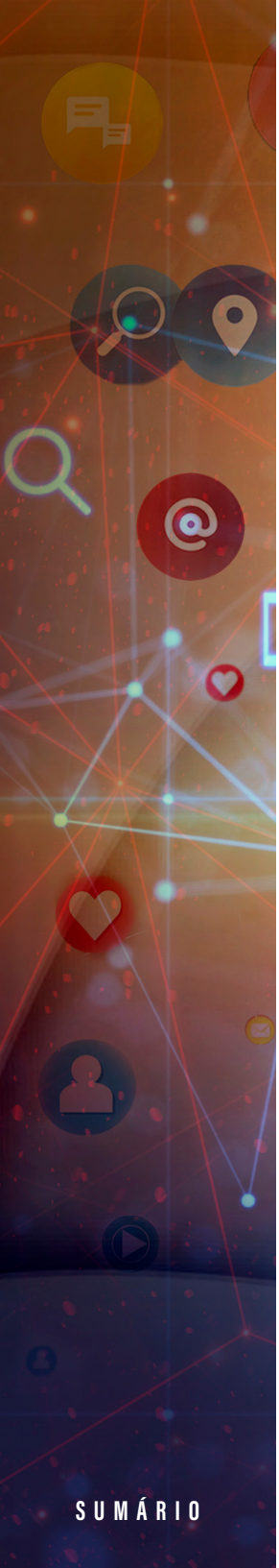
M - MUST have (deve ter), o recurso deve ser preservado ou o recurso que faz parte do arquivo deve ser preservado, por exemplo, uma notícia textual deve ser arquivada porque a ênfase da preservação está numa notícia textual.

S - SHOULD have (deveria ter), o recurso deveria ser preservado, se possível. Quase todas as notícias têm imagens, áudio e vídeo associados que as complementam e devem ser preservados como parte da mesma no arquivamento da *web*.

C - COULD have (poderia ter), o recurso poderia ser preservado se não afetar, por exemplo, a capacidade de armazenamento e a eficiência do sistema.

W - WON'T have (não terá), não seriam incluídas várias versões do *layout*, essa priorização é muito importante no contexto do planejamento de preservação da *web* porque não desperdiça tempo e energia, e é a melhor maneira de atender às necessidades dos usuários.

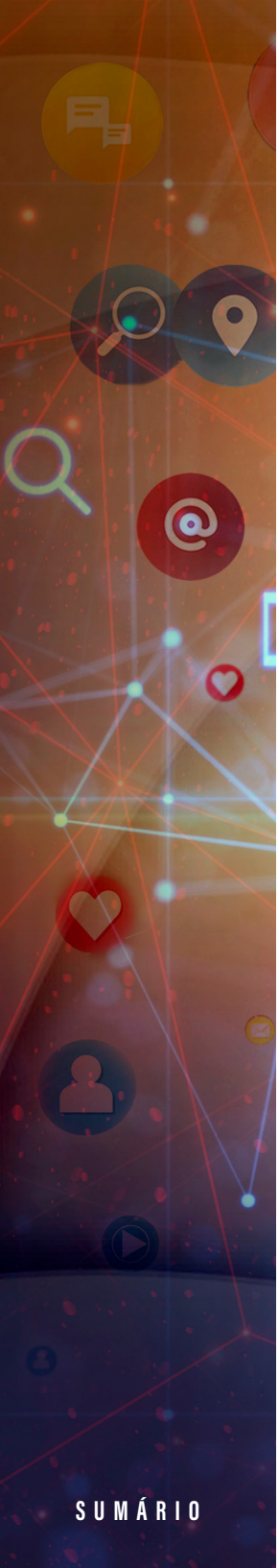
Maemura e colaboradores (2018) afirmam que à medida que os arquivos da *web* se tornam mais proeminentes e amplamente utilizados, é crucial estabelecer mecanismos transparentes para criá-los. Isso facilitará a avaliação da proveniência, escopo e lacunas em um arquivo da *web*. Segundo os autores, a necessidade de transparência na criação de arquivos da *web* é urgente.



Em muitas situações, é possível observar uma falta de transparência na forma como os processos de arquivamento da *web* são comunicados aos usuários dos arquivos da *web*, como mencionado por Maemura e colaboradores (2018). Para resolver esse problema, é necessário documentar corretamente e precisamente como os arquivos da *web* são criados, como essas informações são comunicadas aos usuários e como isso afeta as inferências que podem ser extraídas das análises de todo um arquivo da *web*. Esta informação pode ser compreendida como a proveniência dos arquivos da *web*, e acreditamos que é parte essencial para os usuários entenderem como uma captura foi realizada. Este é um primeiro passo necessário para aumentar a transparência e permitir que os usuários avaliem os arquivos da *web* de forma eficaz.

No contexto de arquivamento da *web*, a política de seleção ajuda a determinar e esclarecer quais conteúdos da *web* devem ser capturados com base nas prioridades, objetivo e âmbito dos conteúdos *web* já definidos. A decisão da política de seleção compreende a descrição do contexto, os usuários pretendidos, os mecanismos de acesso e o uso esperado dos arquivos. O processo de seleção, de acordo com Khan e Ur Rahman (2019), pode ser dividido em subtarefas: preparação, descoberta e filtragem que, combinadas, fornecem uma abordagem qualitativa da seleção de conteúdo da *web*.

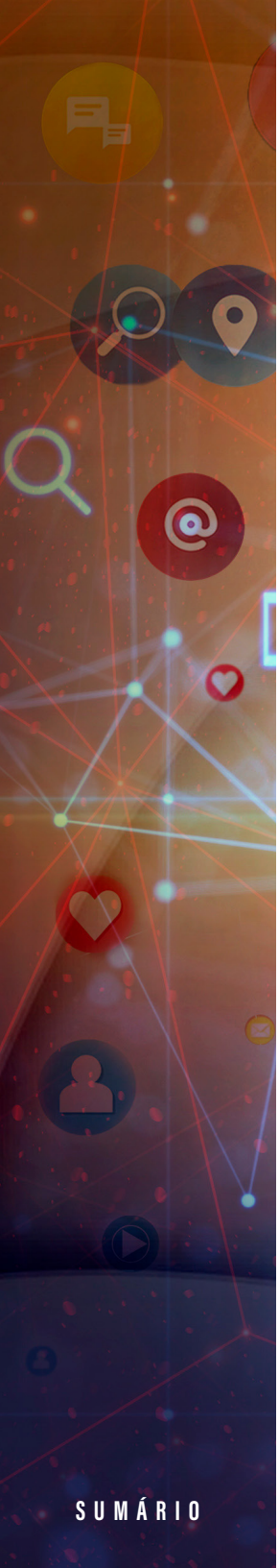
O principal objetivo da fase de **preparação** é determinar as informações direcionadas à técnica de captura, ferramentas de captura, categorização de extensão, nível de granularidade e a frequência da atividade de arquivamento. Esta tarefa deve ser realizada por especialistas de domínio, podem ser os arquivistas, pesquisadores ou bibliotecários. O principal objetivo da fase de **descoberta** é determinar a fonte de informação a ser arquivada. Existem dois métodos de descoberta: exógena, usada na seleção manual e depende principalmente da exploração de uma lista de pontos de entrada (geralmente *links*) para rastrear a coleção manualmente ou endógena usada na seleção automática, depende da extra-



ção de *links* usando rastreadores, quando uma lista de pontos de entrada é criada extraindo *links* automaticamente e obtendo uma lista atualizada todas as vezes durante o rastreamento. O objetivo principal da fase de **filtragem** é otimizar e tornar concisa a descoberta dos conteúdos da *web*, permite coletar um conteúdo da *web* mais específico e remover conteúdo indesejado ou duplicado. Normalmente, utilizam-se métodos automáticos, a filtragem manual é útil se os robôs ou ferramentas automáticas não conseguem interpretar a *web*.

Finalmente, no processo de seleção vale a pena refletir se tudo o que é produzido na *web* tem valor e terá valor no futuro. Por outro lado, é preciso ter a consciência da dificuldade em analisar o que será necessário para futuros pesquisadores. Mas uma coisa é clara: evitar o conteúdo duplicado ou triplicado ajudará sua gestão ao longo do tempo.

Se por um lado, como colocam Marcos Recio; Olivera Zaldua e Sánchez Vigil (2019), as empresas enfrentam o desafio de colocar o usuário no centro da gestão e o sucesso é medido pelo conteúdo que disponibilizam, na premissa de que "quanto mais melhor", as bibliotecas, centros de documentação e arquivos enfrentam o desafio de arquivar apenas o que é imprescindível. Arquivar milhões de documentos apenas pelo fato de ser fácil de armazenar, não engrandece a instituição ou o serviço prestado, pelo contrário, visto que cada vez mais a demanda é por informação mais precisa, pertinente e gerada no menor tempo possível. Os autores ainda questionam se, no momento que as máquinas fizerem a tarefa de seleção de conteúdo, descobriremos se arquivar tudo foi uma boa opção ou se deveríamos ter apostado em guardar apenas o imprescindível.



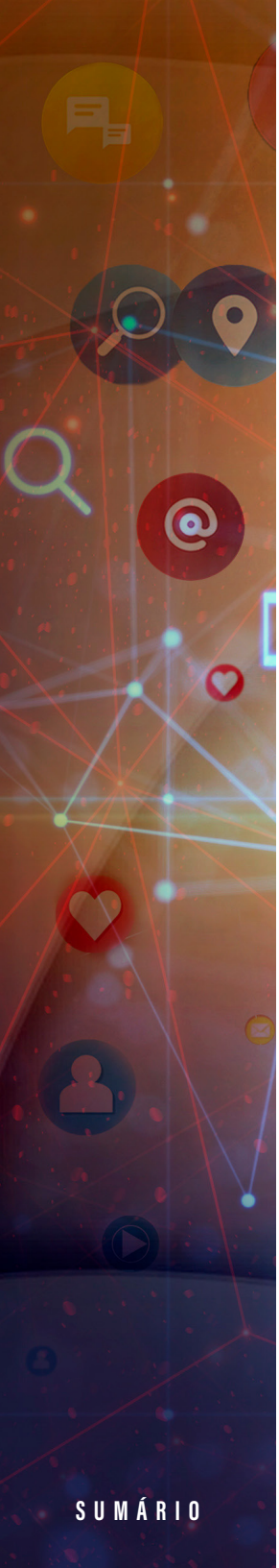
CAPTURA DE *WEBSITES*

O arquivamento da *web* é o processo de coletar, preservar e fornecer acesso ao conteúdo da *web*. Esse conteúdo pode incluir *sites*, *blogs*, postagens de mídia social e outros materiais *on-line*. O arquivamento da *web* é importante porque nos permite preservar e acessar informações que de outra forma podem ser perdidas devido a mudanças na tecnologia ou à natureza efêmera da *web*.

Existem várias técnicas usadas no arquivamento da *web*, incluindo rastreamento da *web* e criação de instantâneos. O rastreamento da *web* envolve o uso de um programa de *software*, conhecido como rastreador da *web*, para percorrer automaticamente a *web* e coletar conteúdo da *web*. Esse conteúdo é então armazenado em um arquivo, onde pode ser acessado e estudado por pesquisadores. Os rastreadores da *web* podem ser configurados para capturar tipos específicos de conteúdo, como páginas HTML, imagens, vídeos e arquivos de áudio.

O arquivamento do lado do cliente por meio de "*crawling*" captura o conteúdo da *web* sem acesso direto aos arquivos armazenados em um servidor. Esta abordagem é comumente usada na comunidade de patrimônio cultural (MAEMURA *et al*. 2018). O rastreamento da *web* foi originalmente usado para indexação de mecanismos de pesquisa, mas foi adaptado para arquivamento, para automatizar a descoberta e a captura de recursos da *web*. Um rastreador começa com uma lista de URLs, chamada de "lista inicial" e, em seguida, visita cada página da lista, baixando o conteúdo e os recursos vinculados ou incorporados a essa página. Ele também identifica todos os *links* nessa página e os adiciona à lista de URLs. O rastreador passa para a próxima URL da lista e repete o processo.

Algumas técnicas permitem salvar e acessar posteriormente o conteúdo da *web*, mas não se configuram em arquivamento da

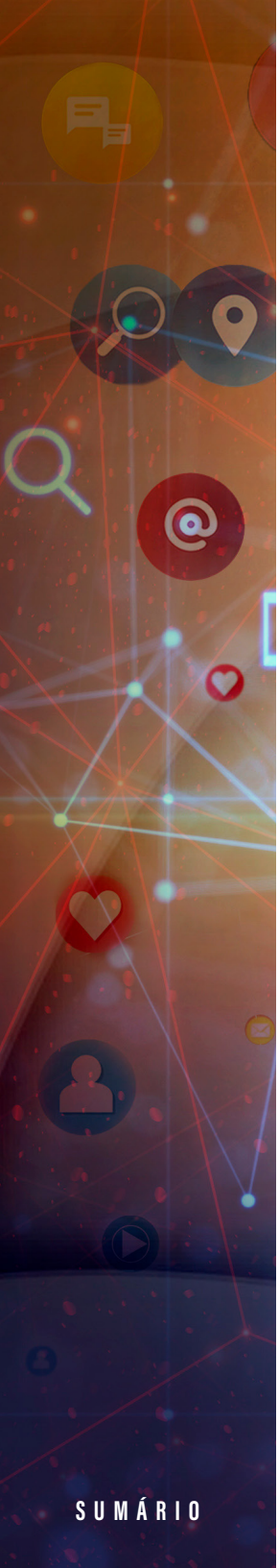


web, como a técnica de *snapshotting* e acesso ao *cache* de motores de busca, como o do Google.

Snapshotting é uma técnica usada para salvar conteúdo da *web*, que envolve tirar uma captura de tela ou "instantâneo" de uma página da *web* em um ponto específico no tempo. Esses instantâneos são armazenados em um arquivo e podem ser acessados para estudar a evolução do conteúdo da *web* ao longo do tempo. Entretanto, é uma técnica limitada, pois não permite a navegação por *links* e a interatividade original da página *web*. Ao contrário das capturas de tela, os arquivos da *web* registram todo o conteúdo de uma página da *web*, incluindo arquivos HTML, CSS, *Javascript* e mídias incorporadas. Isso permite a interação com a página arquivada, incluindo clicar em *links* para explorar a página da *web*.

É relevante destacar que os mecanismos de busca, tais como o Google, possuem uma função de armazenamento em *cache*, que permite a recuperação de uma página da *web* a partir de uma captura da página, mostrando como ela aparecia em uma data e hora específicas, conforme identificado no cabeçalho da página recuperada. Diferente do modelo anteriormente descrito, não se trata de uma imagem capturada do *site*, mas de um *site* com características do original. Esse sistema funciona como um *backup* limitado, pois é possível acessar a página *web* que não está disponível por algum motivo. No entanto, o uso do armazenamento em *cache* pode comprometer algumas funções da página pesquisada e até mesmo a navegação entre os *hiperlinks*. Além disso, é possível que o administrador do *site* tenha configurado a página para não ser arquivada com o uso de *metatags*, e, portanto, o *cache* de armazenamento não estará disponível nos mecanismos de busca.

O arquivamento da *web* propriamente dito envolve diversas práticas e padrões recomendados, além do acompanhamento das ações promovidas pelo *International Internet Preservation Consortium (IIPC)*. A captura, preservação e acesso aos arquivos da



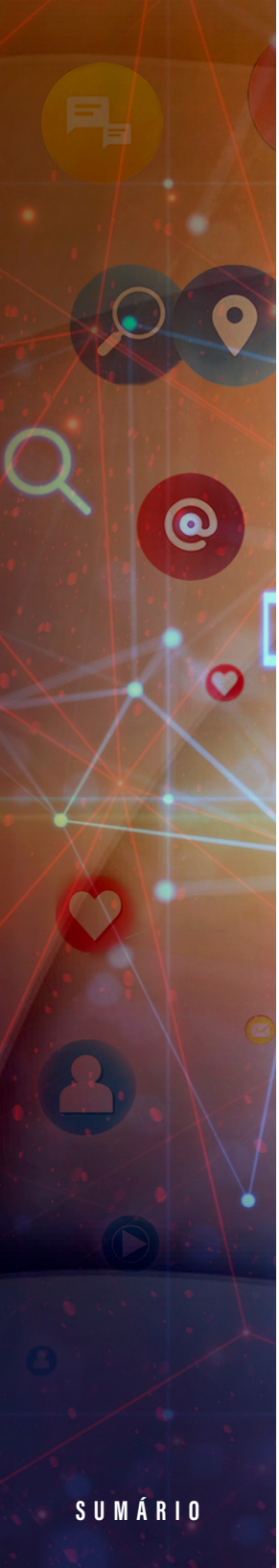
web acontece com o uso de rastreadores, armazenamento em formatos ISO 28500:2017 (WARC) e plataformas de acesso ou *replay* dos *site*s arquivados.

As ações do rastreador da *web* podem causar reações dos administradores dos *site*s que ele está rastreando (MAEMURA *et al*. 2018). A *web* ao vivo pode intencionalmente ou não resistir ao processo de arquivamento, resultando em várias respostas possíveis do arquivista da *web* que conduz o rastreamento.

Os arquivos *robots.txt* podem ser usados para especificar como os robôs de indexação do mecanismo de pesquisa interagem com um *site*. Por exemplo, eles podem ser usados para excluir certas partes do *site* da indexação, definir um limite de tempo entre solicitações ao servidor ou bloquear completamente um rastreador. Esses arquivos fornecem uma forma diferente de resistência ao rastreamento.

De acordo com Maemura e colaboradores (2018), ao agregar dados de vários rastreamentos, há vários desafios e considerações a serem lembrados. Pesquisas anteriores mostraram que o rastreamento da *web* é um método imperfeito de capturar recursos da *web*, mas algumas limitações são mais fáceis de antecipar do que outras. Por exemplo, os parâmetros de rastreamento geralmente são definidos para determinar quais URLs são adicionados à lista ou para limitar o tempo de execução do rastreador. As tecnologias de rastreador também têm limitações conhecidas, como a necessidade de tempo adicional e configuração para capturar materiais da *web* com conteúdo *Javascript*. Essa mudança para a análise de coleções inteiras requer planejamento e execução cuidadosos.

Apesar da utilidade incontestável do arquivamento da *web*, observam-se diversas dificuldades para operacionalizar a prática. Alguns *site*s são fáceis de coletar utilizando rastreadores *Wayback Machine*, robôs da Internet que navegam continuamente na *web* para fins de indexação, criando instantâneos digitais de *site*s provenientes



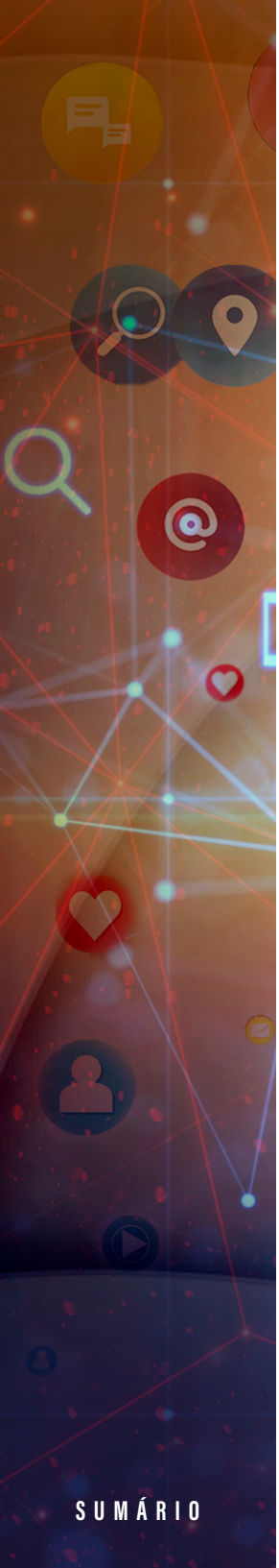
de várias fontes, que mudaram ao longo do tempo. Porém, outros *site*s que são protegidos por senhas, usando certos elementos do *Javascript* ou *site*s bloqueados por extensões *robots.txt* são geralmente mais difíceis de rastrear para uma captura por máquina.

A captura por máquina, ainda que seja um valioso recurso para automatizar a preservação de *site*s, evidencia a necessidade de adotar uma série de medidas para minimizar o problema com *links* quebrados ou inacessíveis. Visto que, mesmo que o conteúdo esteja disponível para os rastreadores *Wayback Machine*, não significa que o *site* esteja preservado e disponível para uso futuro. Um estudo realizado por Balogun e Kalusopa (2022) revelou que os repositórios *Indigenous Knowledge Systems* (IKS) da África do Sul estão disponíveis para a utilização do *Wayback Machine* mas, a maioria dos *links* estão quebrados ou inacessíveis e em outros casos as características do *site* são limitadas devido à *links* quebrados. As instituições precisam fazer um esforço para apropriar-se de informações sobre a importância de arquivamento da *web* e tomar medidas para permitir a captura do conteúdo de seus repositórios para integrá-los às coleções institucionais para uso futuro.

Há diversas ferramentas que coletam e reproduzem o conteúdo da *web*, identificar a mais apropriada é um trabalho que requer planejamento sobre o que se deseja arquivar e com qual objetivo. Dada a natureza de fonte aberta das ferramentas neste domínio, é um desafio fornecer apoio contínuo e desenvolvimento de ferramentas específicas e por esse motivo Dooley e colaboradores (2017) desenvolveram sete questões para avaliar ferramenta dessa natureza:

1. Qual é a finalidade básica do instrumento e as suas funcionalidades principais? (por exemplo, captura, visualização e/ou camada administrativa).

2. Que objetos/arquivos pode absorver e gerar? (ou seja, a unidade principal ou alterações, tais como *Mementos*, WARCs ou PDFs).



3. Em que perfis de metadados faz os registros?

4. Que elementos descritivos são gerados automaticamente?

5. Quais os elementos descritivos podem ser criados ou editados pelo utilizador?

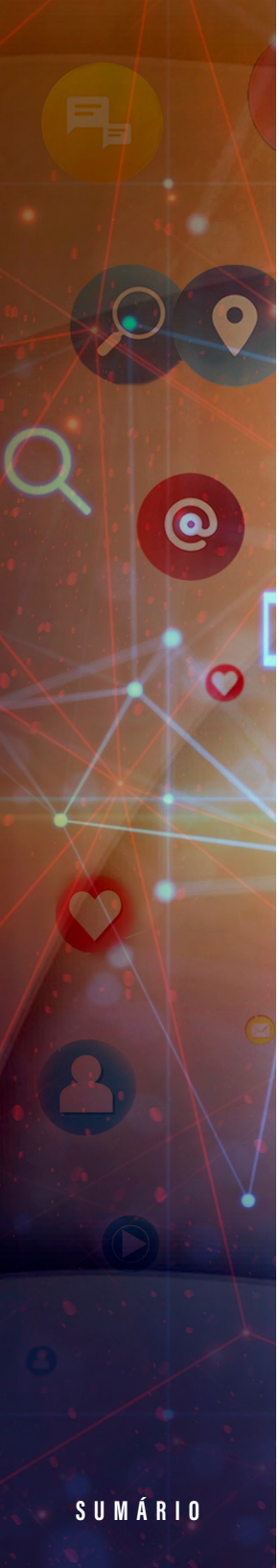
6. Que elementos descritivos de dados podem ser exportados para utilização fora da ferramenta?

7. Que relação tem com outras ferramentas? (por exemplo, *Heritrix* recolhe metadados que estão incorporados em um arquivo WARC, alguns dos quais são utilizados pelo *Archive-It*).

Depois que o conteúdo da *web* é coletado, ele geralmente é armazenado em um arquivo digital. Arquivos digitais são sistemas especializados projetados para armazenar e gerenciar grandes quantidades de dados digitais. Normalmente fornecem recursos como controle de versão, gerenciamento de metadados e planejamento de preservação.

Além dessas técnicas, também existem várias ferramentas e serviços disponíveis para dar suporte ao arquivamento da *web*. Por exemplo, o *Internet Archive* é uma organização sem fins lucrativos que fornece acesso gratuito a uma vasta coleção de conteúdo da *web*, incluindo páginas da *web*, imagens, vídeos e arquivos de áudio. Também, fornece ferramentas e serviços que suportam a criação de arquivos da *web*, como o *Wayback Machine*, que permite aos usuários acessar versões arquivadas de páginas da *web*.

O arquivamento da *web* é uma ferramenta importante para preservar e fornecer acesso à riqueza de informações disponíveis na Internet. Usando rastreadores da *web*, coletores da *web*, arquivos digitais e outras ferramentas e serviços, organizações e indivíduos podem garantir que o conteúdo da *web* seja coletado, preservado e disponibilizado para usos futuros.



PRINCIPAIS FERRAMENTAS DE *WEB ARCHIVING*

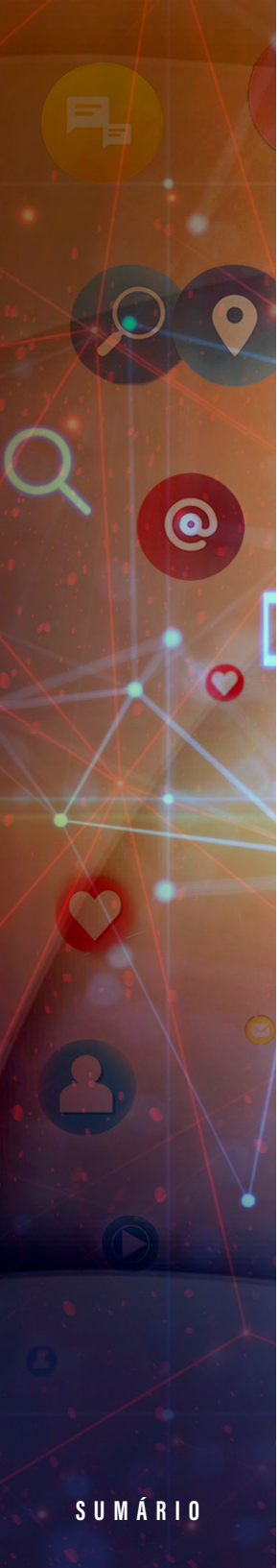
A melhor plataforma para arquivamento da *web* dependerá das necessidades e objetivos específicos do projeto, bem como dos recursos e conhecimentos disponíveis. Pode ser necessário experimentar diferentes ferramentas e plataformas para encontrar a melhor solução para uma determinada situação. Existem várias plataformas e ferramentas que podem ser usadas para arquivamento da *web*, algumas das mais comuns incluem:

1. Arquivos da *web*: existem vários arquivos públicos da *web*, como o *Wayback Machine*, que coletam e preservam o conteúdo da *web* para referência futura. Esses arquivos podem fornecer acesso a uma ampla variedade de páginas e conteúdos da *web*, embora o conteúdo específico disponível possa variar. A preservação de *site*s também pode ser sob demanda do usuário, quando ele insere um *link* específico para salvamento, utilizando serviços como o *Save Page Now*[116] (*Wayback Machine*) e o *Archive.Today*[117].

2. Navegadores da *web*: muitos navegadores da *web*, como *Google Chrome* e *Mozilla Firefox*, possuem ferramentas integradas para salvar páginas da *web* permitindo a visualização *off-line*. Normalmente são disponibilizados *plug-ins* de terceiros para a captura e reprodução desses arquivos da *web*. Essas ferramentas podem ser usadas para capturar e armazenar conteúdo da *web*, embora possam não fornecer tantos recursos e opções quanto às ferramentas dedicadas de arquivamento da *web*.

116 https://web.archive.org/save

117 https://archive.today/



3. Serviços de arquivamento da *web*: plataformas especializadas projetadas especificamente para capturar e armazenar conteúdo da *web*. Esses serviços geralmente fornecem ferramentas para identificar e coletar páginas da *web*, bem como organizar e gerenciar o conteúdo arquivado. Plataformas de assinatura, como o *Archive.it*, criado pelo *Internet Archive* em 2006[118], consiste em um serviço baseado na nuvem e o preço varia de acordo com o número de URLs que são arquivadas e a frequência das coletas.

4. Ferramentas de linha de comando: ferramentas como *Wget* e *HTTrack* podem ser usadas para arquivamento da *web*, capturando e armazenando páginas da *web* a partir da linha de comando, tornando-as úteis para automatizar o processo de arquivamento.

5. Ferramentas de raspagem da *web*: não constituem *softwares* específicos de arquivamento da *web*, mas podem servir como auxílio em projetos, na captura de dados relevantes, sendo usados para extrair automaticamente dados e conteúdo de páginas da *web*. As ferramentas de raspagem da *web* podem ser usadas para coletar grandes volumes de dados da *web* e podem ser personalizadas para extrair tipos específicos de informações. Exemplos de ferramentas de raspagem da *web* incluem *Scrapy*[119], *ParseHub*[120] e *Import.io*[121].

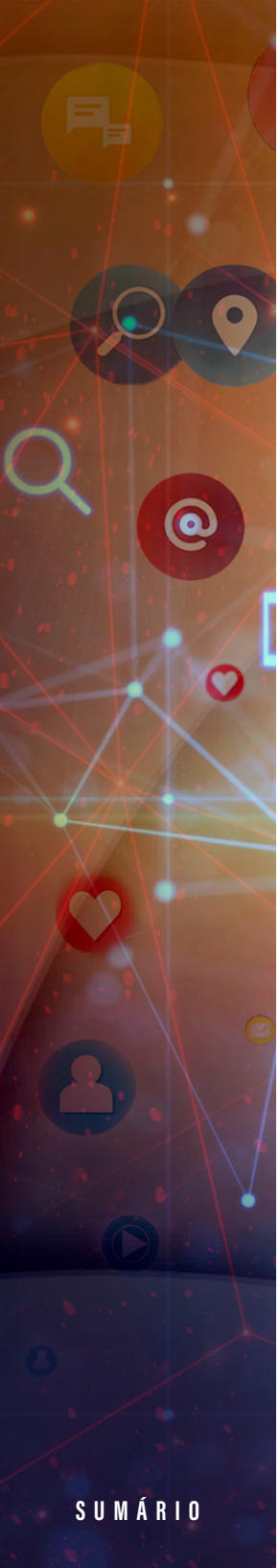
6. *Softwares* de arquivamento da *web*: As ferramentas de arquivamento da *web* são atualizadas constantemente e, portanto, torna-se um desafio manter uma tabela atualizada com todos

118 https://archive-it.org/blog/learn-more/

119 https://scrapy.org/

120 https://www.parsehub.com/

121 https://www.import.io/



os *softwares* disponíveis[122]. Abaixo segue a descrição de alguns dos principais *softwares* utilizados no arquivamento da *web*. Recomenda-se consultar a lista atualizada em português no *site* www.arquivo.ong.br.

Heritrix[123]: é um rastreador de *web* projetado para arquivamento digital. Ele pode ser usado para coletar conteúdo da *web* e criar arquivos WARC (*web ARChive*) que podem ser armazenados para uso futuro. É amplamente utilizado por arquivos, bibliotecas e outras instituições para capturar conteúdo da *web*. Foi desenvolvido pela *Internet Archive* em linguagem Java. É altamente configurável e suporta recursos como autenticação, filtragem e agendamento de tarefas, além de permitir aos usuários definir a profundidade de rastreamento, escolher quais URLs rastrear, filtrar tipos de conteúdo e muito mais. O *Heritrix* é executado em um servidor *web* e é acessado por meio de uma interface gráfica do usuário.

O Núcleo de Pesquisa em Arquivamento da *web* e Preservação Digital da UFRGS (NUA*WEB* UFRGS/CNPq) realizou uma tradução livre para português da documentação de instalação e uso do *Heritrix*, disponível no *GitHub*[124].

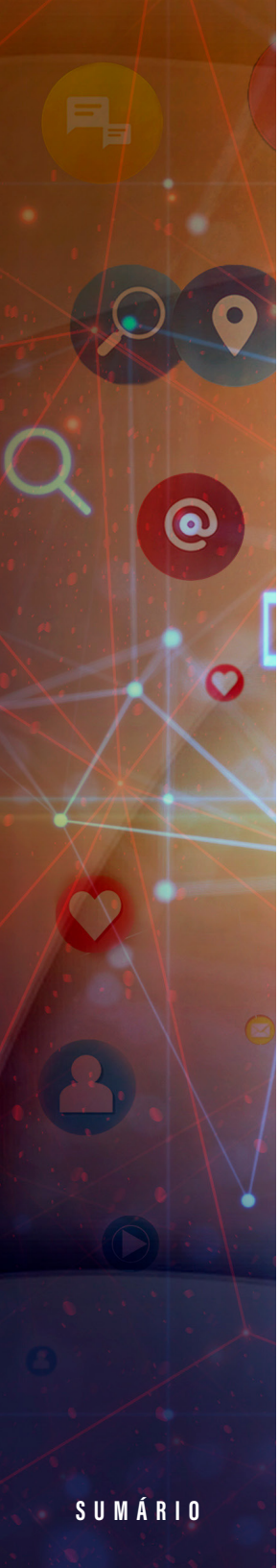
Wget[125]: é um utilitário de linha de comando para baixar conteúdo da *web*. Ele pode baixar arquivos individuais, bem como *site*s inteiros, e também pode criar arquivos WARC. Suporta autenticação, *cookies*, *proxy* e várias outras opções que podem ser configuradas para personalizar o comportamento do *download*. É uma ferramenta popular entre os usuários de sistemas *Unix* e *Linux*. É um *software* de código aberto escrito em C e disponível para sistemas operacionais *Unix, Linux* e *Windows*. Ele suporta vários protocolos de rede,

122 Uma lista no GitHub também pode ser consultada em https://github.com/iipc/awesome-web-archiving

123 https://heritrix.readthedocs.io/en/latest/

124 https://github.com/NUAWEB/heritrix-pt

125 https://www.gnu.org/software/wget/



como HTTP, HTTPS e FTP, e pode ser configurado para baixar arquivos de forma recursiva. O *Wget* é altamente personalizável e suporta diversas opções para personalizar o comportamento do *download*, como autenticação, limitação de largura de banda e resolução de nomes.

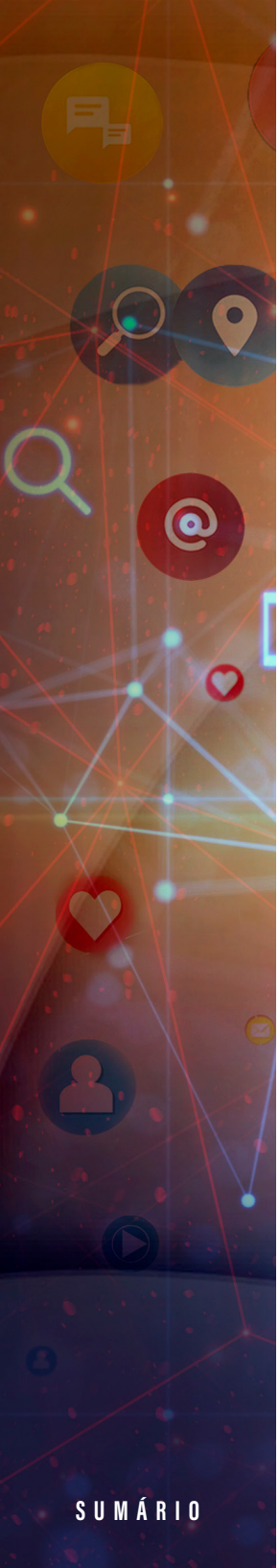
Conifer[126] (anteriormente conhecido como *Webrecorder*): é uma ferramenta de captura da *web* baseada em navegador[127]. Ele permite que os usuários capturem a navegação na *web* em uma sessão de gravação, criando um arquivo WARC com o conteúdo capturado. *Conifer* é fácil de utilizar e é uma opção popular para os usuários que desejam capturar conteúdo da *web* sem precisar de conhecimentos técnicos avançados. Foi desenvolvido pela *Rhizome* em linguagem *Python*. Ele é executado em um servidor *web* e pode ser acessado por meio de uma interface gráfica do usuário. O *Webrecorder* tornou-se um projeto com disponibilização de mais ferramentas relacionadas ao arquivamento da *web*.

Brozzler[128] é um *software* de arquivamento da *web* de código aberto criado em 2016 pela equipe do *Internet Archive* que utiliza tecnologias de rastreamento modernas para capturar conteúdo da *web* em grande escala e com alta qualidade. O nome parte da junção dos termos "*browser*" + "*crawler*" = "*brozzler*". Ele usa um navegador *web headless* (sem interface gráfica) para rastrear *site*s da *web*, o que permite uma experiência mais próxima de um usuário real. Isso é importante porque muitos *site*s modernos utilizam tecnologias avançadas, como *Javascript*, que podem não ser facilmente rastreadas por outros *software*s de arquivamento da *web*. Também, é altamente configurável, permitindo aos usuários definir parâmetros como a frequência de rastreamento, profundidade de navegação e filtros de conteúdo. Ele suporta vários formatos de arquivo, incluindo WARC e ARC, e permite que os usuários visualizem o conteúdo capturado de forma

126 https://webrecorder.net/

127 https://conifer.rhizome.org/

128 https://github.com/Internetarchive/brozzler



fácil e eficiente. O *Brozzler* é projetado para ser escalável, é compatível com a plataforma do *Internet Archive*, permitindo que os usuários enviem dados de rastreamento diretamente para a plataforma para preservação a longo prazo. Por ser altamente configurável permite aos usuários definir vários parâmetros, incluindo:

1. Frequência de rastreamento: os usuários podem definir a frequência com que o *Brozzler* deve rastrear um *site* da *web*.

2. Profundidade de navegação: os usuários podem definir a profundidade com que o *Brozzler* deve rastrear um *site* da *web*.

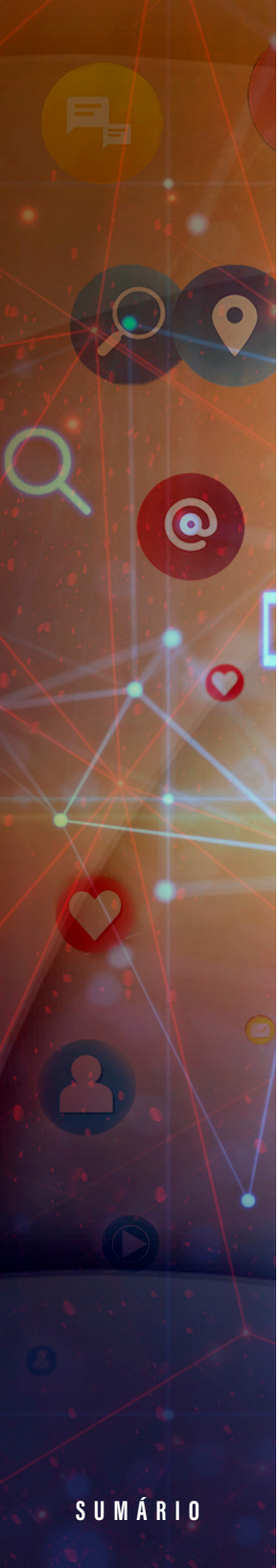
3. Filtros de conteúdo: os usuários podem definir filtros para excluir conteúdo que não é relevante ou que não deve ser arquivado.

4. Formato de arquivo: o *Brozzler* suporta vários formatos de arquivo, incluindo WARC e ARC.

5. Visualização do conteúdo: o *Brozzler* permite que os usuários visualizem o conteúdo capturado de forma fácil e eficiente.

Browsertrix[129]: é um *software* de arquivamento da *web* de código aberto que utiliza a tecnologia de captura de tela baseado no navegador (*Chrome*) para preservar a aparência e o comportamento dos *site*s da *web*. Foi desenvolvido pela equipe do W*ebrecorder*, uma organização sem fins lucrativos que visa preservar a *web*. Projetado para ser fácil de usar, mesmo para usuários que não têm experiência técnica. Ele permite que os usuários capturem *site*s da *web* com aparência e comportamento precisos, mesmo que os *site*s dependam de tecnologias avançadas, como *Javascript* e CSS. Isso é importante porque muitos *site*s modernos têm interfaces complexas que podem ser difíceis de capturar com precisão usando outros *software*s de arquivamento da *web*. Também, é configurável, permitindo aos usuários definir parâmetros como a resolução da tela, o tamanho do nave-

129 https://github.com/webrecorder/browsertrix-crawler



gador e a profundidade de navegação. Ele também suporta vários formatos de arquivo, incluindo WARC, para que os usuários possam armazenar os *site*s capturados para preservação a longo prazo.

Apache Nutch[130]: é um mecanismo de busca baseado em código aberto que inclui um rastreador da *web* capaz de criar arquivos WARC. Altamente configurável, e suporta a filtragem de conteúdo, a autenticação e outras opções avançadas. É frequentemente usado em projetos de arquivamento da *web* e em empresas que desejam criar seus próprios mecanismos de busca personalizados. Foi criado em 2002 por Doug Cutting, o mesmo criador do *Apache Hadoop*. O Nutch é escrito em linguagem de programação *Java* e usa tecnologias como *Apache Lucene*, *Apache Solr* e *Apache Tika* para análise e indexação de dados.

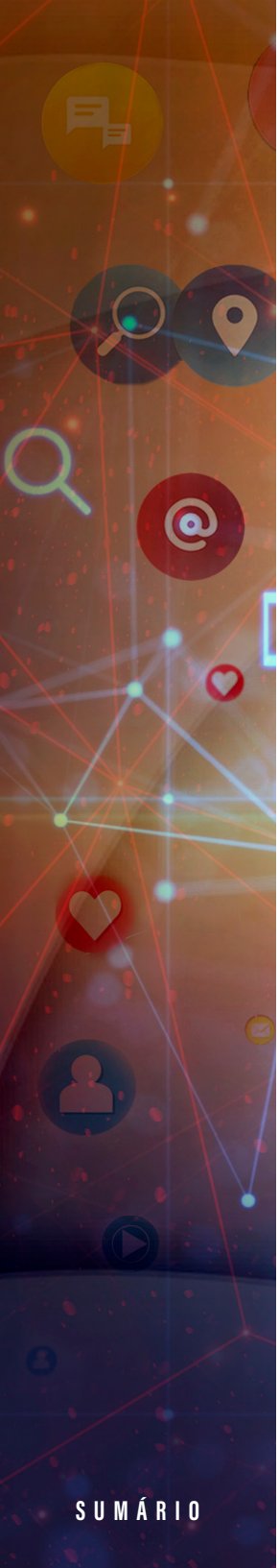
Storm Crawler[131]: é um projeto de código aberto que combina o *Apache Storm* com o *Apache Nutch* para criar um rastreador da *web* altamente escalável capaz de criar arquivos WARC. Altamente configurável, pode ser usado para coletar grandes volumes de conteúdo da *web* em um ambiente de alta velocidade e alta disponibilidade. É frequentemente usado em projetos de arquivamento da *web* em larga escala. Foi criado em 2014 por Julien Nioche e é escrito em linguagem de programação *Java*. Ele combina o poder de processamento distribuído do *Apache Storm* com o *Apache Nutch* para criar um rastreador altamente escalável. Ele também suporta a coleta de dados em tempo real e pode ser usado para alimentar mecanismos de busca personalizados.

Libarchive[132]: é uma biblioteca de código aberto escrita em linguagem de programação C, que suporta a criação, leitura e gravação de arquivos de vários formatos, incluindo AR, ZIP, 7ZIP, RAR, CAB, ISO e WARC. A biblioteca pode ser usada em uma variedade

130 https://nutch.apache.org/

131 http://stormcrawler.net/

132 https://www.libarchive.org/



de projetos que envolvem arquivamento e compactação de dados, e pode ser útil para projetos de arquivamento da *web* que precisam gerenciar grandes volumes de conteúdo.

WARCreate[133] é uma extensão do *Google Chrome* que permite aos usuários criar um arquivo WARC a partir de qualquer página da *web* navegável, que pode ser utilizada com outras ferramentas, como o *Wayback Machine*. O objetivo do *WARCreate* é tornar-se uma solução pessoal de arquivamento da *web* que padroniza os metadados e garante o arquivamento seguro.

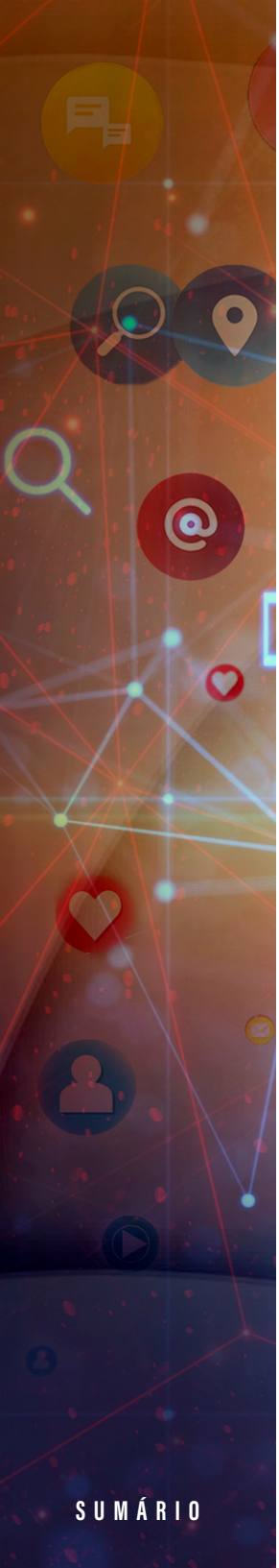
WAIL[134], também conhecido como *web Archiving Integration Layer*, é um aplicativo de *desktop* que combina várias ferramentas de arquivamento da *web* pré-configuradas e oferece uma interface gráfica do usuário (GUI) para permitir fácil preservação e reprodução da página da *web*. Ele vem equipado com várias ferramentas, incluindo *Heritrix* para rastreamento da *web* e *OpenWayback* para reprodução de arquivos da *web*, essas ferramentas são facilmente acessíveis por meio de uma interface de sistema nativa para navegar. Utiliza programação *Python*, compilado em um aplicativo nativo usando *PyInstaller*.

Web Curator Tool[135] (WCT) é um *software* de código aberto desenvolvido pela *National Library of New Zealand* em 2006. O objetivo principal do WCT é permitir que as instituições de patrimônio cultural, como bibliotecas e arquivos, capturem e arquivem o conteúdo da *web* para fins de preservação. Possui uma interface gráfica do usuário e é baseado em *Java*, o que significa que pode ser executado em uma variedade de plataformas, incluindo *Windows, Mac* e *Linux*. Ele suporta vários formatos de arquivo, incluindo WARC e ARC, que são amplamente usados para preservação da *web*. Oferece recursos avançados para gerenciamento de projetos de arquivamento da *web*.

133 https://chrome.google.com/webstore/detail/warcreate/kenncghfghgolcbmckhiljgaabnpcaaa

134 https://machawk1.github.io/wail/

135 https://webcuratortool.org/



Permite que os usuários criem projetos de arquivamento, definam critérios de seleção de conteúdo, gerenciem o processo de coleta e arquivamento, e realizem revisões e avaliações de qualidade.

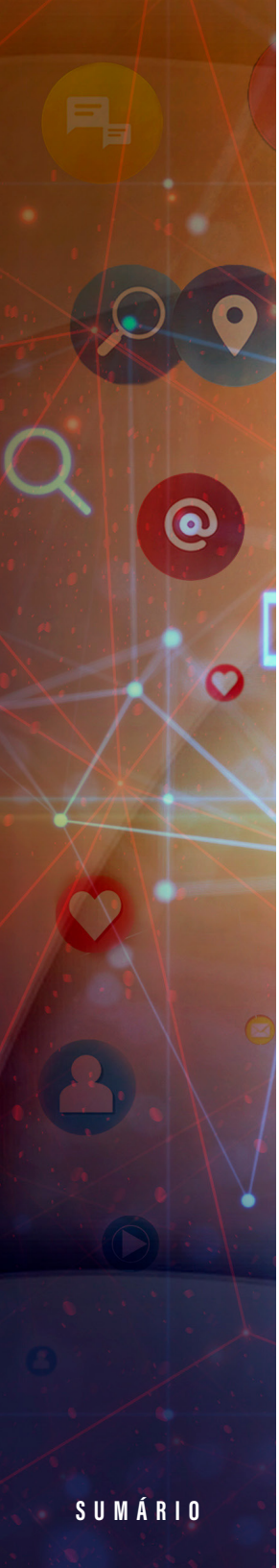
O WCT também possui uma variedade de recursos de personalização, incluindo a capacidade de adicionar *scripts* personalizados para processar o conteúdo da *web* antes do arquivamento e a capacidade de personalizar as regras de exclusão de conteúdo. É frequentemente utilizado por bibliotecas e arquivos em todo o mundo, incluindo a Biblioteca do Congresso dos Estados Unidos e a Biblioteca Nacional da Austrália, para preservar o conteúdo da *web* relacionado a eventos históricos e culturais importantes. Em termos de configurações, o WCT pode ser configurado para executar em servidores locais ou na nuvem. Ele é escalável e pode ser usado para capturar grandes volumes de conteúdo da *web* em um ambiente de alta velocidade e alta disponibilidade. Também, oferece suporte a extensões, permitindo que os usuários personalizem ainda mais o *software* para atender às suas necessidades específicas.

Quanto às ferramentas de acesso ou *replay* dos arquivos da *web*, duas se destacam: a *Open Wayback*[136] e a *Python Wayback* ou PYWB[137].

O *Open Wayback* é uma ferramenta de *replay* de arquivos da *web* amplamente utilizada para acessar e visualizar conteúdo de sites que foram arquivados. Ele permite que os usuários naveguem por versões anteriores de sites e recuperem informações que possam ter sido perdidas ou alteradas ao longo do tempo. Com uma interface amigável, o *Open Wayback* torna possível explorar o passado da *web* e pesquisar instantâneos históricos de páginas da *web*.

136 https://github.com/iipc/openwayback

137 https://github.com/webrecorder/pywb

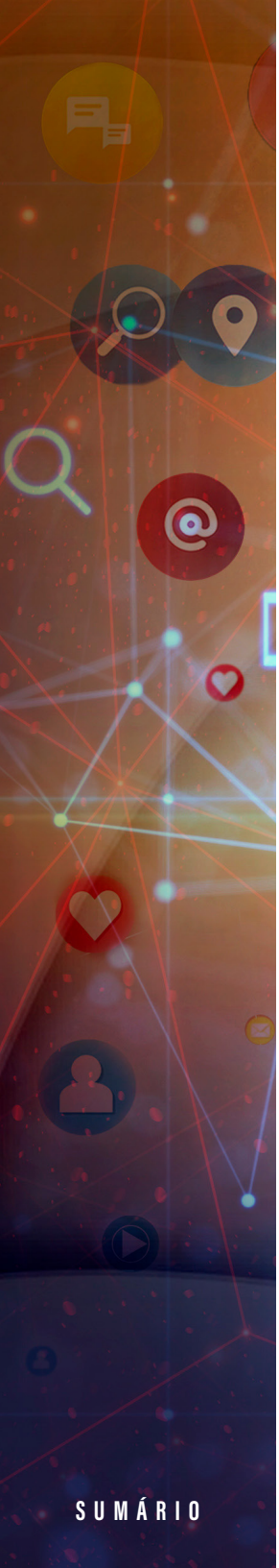


O *Python Wayback*, também conhecido como PYWB, é uma ferramenta de código aberto desenvolvida em *Python* para realizar o *replay* de arquivos da *web*. Com o PYWB, os usuários podem acessar e recuperar versões anteriores de *sites* arquivados. Uma adição importante é o sistema de configuração dinâmica de várias coleções, que permite atualizações sem a necessidade de reiniciar a ferramenta. O PYWB oferece o modo de gravação, permitindo a criação de novos arquivos da *web* a partir de *sites* ativos ou arquivos existentes. Além disso, o PYWB suporta a agregação da API Memento[138], possibilitando consultar várias fontes de arquivos arquivadas, locais e remotos. A ferramenta também possui um modo de Proxy HTTP/S, que inclui uma autoridade de certificação personalizável para gravação e reprodução. O Sistema de Reescrita do Lado do Cliente (wombat.js) também oferece uma compatibilidade aprimorada com *sites* modernos. A interface de consulta permite uma visualização de calendário e agrupamento de resultados por ano e mês.

Existem projetos que têm como objetivo analisar arquivos da *web* e fornecer ferramentas para facilitar essa tarefa. O projeto *Archives Unleashed*[139] tem como objetivo disponibilizar *petabytes* de conteúdo histórico da internet para acadêmicos e pesquisadores interessados em estudar o passado recente. Eles desenvolvem ferramentas de busca e análise de dados de arquivos da *web*, permitindo acesso, compartilhamento e investigação da história desde os primórdios da *World Wide Web*. Com o apoio da Fundação Andrew W. Mellon, o projeto colabora com o *Internet Archive* para integrar a Nuvem ao serviço *Archive-It*, visando a sustentabilidade do projeto e melhorando a usabilidade e acessibilidade dos arquivos da *web*. O *Archives Unleashed Toolkit* (RUEST, LIN, MILLIGAN, FRITZ, 2020) é uma plataforma baseada no *Apache Spark* para análise de arquivos da *web* em

138 https://timetravel.mementoweb.org/guide/api/

139 https://archivesunleashed.org/



grande escala. Além disso, o *Warclight*[140] é um mecanismo de busca que permite a descoberta de arquivos da *web*. Ambas as ferramentas exigem conhecimentos técnicos avançados, mas oferecem recursos interessantes para explorar e analisar dados de arquivos da *web*.

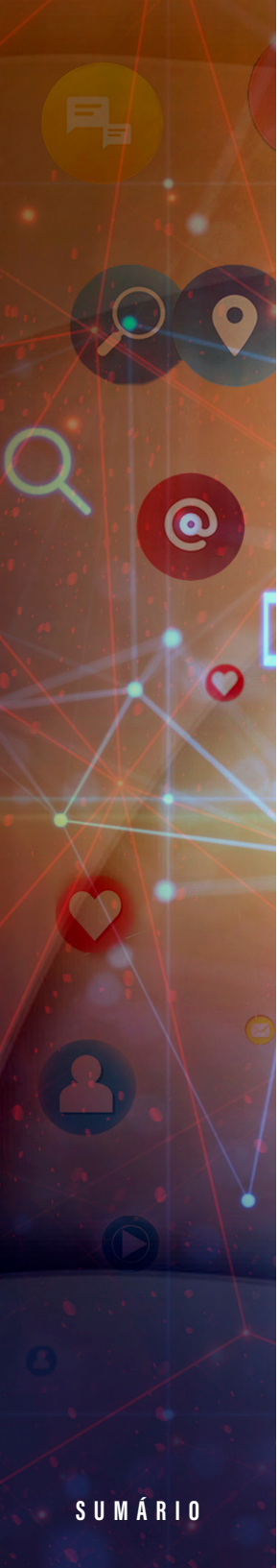
Por fim, conforme Schafer e Winters (2021), a Inteligência Artificial (IA) certamente desempenhará um papel crucial no avanço do campo, mas é importante utilizá-la de forma crítica, fortalecer a "boa governança" dos arquivos da *web* pode ser uma solução para lidar com esses desafios.

ARQUIVABILIDADE DOS *WEBSITES*

A preservação digital de *website*s é um desafio devido à complexidade e diversidade de dados envolvidos no ambiente *web*. O processo automatizado de rastreamento da *web* pode deixar de capturar partes significativas de *website*s arquivados, variando de acordo com o tipo de recursos, acessibilidade e tecnologias utilizadas. Não há uma métrica para determinar se um *website* pode ser arquivado com sucesso, mas na literatura é possível identificar sistemas e métodos para verificar as possibilidades de arquivamento por *software* de captura, antecipando verificações de garantia de qualidade. Isso permite avaliar os resultados e decidir a melhor estratégia para preservar as páginas.

A garantia de qualidade no arquivamento da *web* envolve o exame das características dos *site*s capturados pelo *software* de rastreamento para garantir que sejam arquivos válidos, que cumpram a política de preservação da instituição e estejam adequadamente preparados para uso a longo prazo. As limitações devido às diversas

140 https://archivesunleashed.org/warclight/

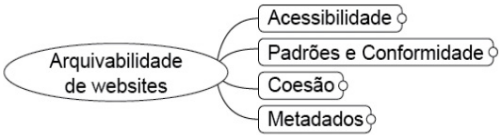


tecnologias utilizadas na *web* significam que uma captura perfeita de todos os *site*s raramente é obtida, tornando necessárias medidas de garantia de qualidade (BINGHAM, 2014). Medir o sucesso de uma captura de arquivos da *web* é difícil, pois conceitos menos tangíveis, como qualidade, amplitude de cobertura e relevância, exigem um certo grau de julgamento de valor. O sucesso pode ser medido pelo tamanho do arquivo, páginas ausentes ou páginas que poderiam ter sido capturadas, mas não foram. A contagem de domínios ou *hosts* não é necessariamente útil para avaliar a qualidade.

Para avaliar a capacidade de um *website* de ser arquivado, são utilizados critérios do W3C e métodos e ferramentas da comunidade de arquivamento da *web* (BANOS *et al.*, 2013; BANOS; MANOLOPOULOS, 2016). Dessa forma, foram elaboradas um conjunto de métricas para quantificar o nível de arquivamento de qualquer *site* e os autores supra citados desenvolveram o método CLEAR+ (*Credible Live Evaluation of Archive Readiness*) para automatizar o processo de determinar se um *site* é arquivável ou não. O conceito de *web Archivability* (WA) identifica os principais aspectos de um *site* necessários para determinar seu potencial de arquivamento com integridade e precisão. Essa ferramenta calcula a arquivabilidade de um *site*, que automatiza o processo de controle de qualidade e evita a coleta de *site*s não arquiváveis, resultando em economia significativa de recursos humanos, computacionais e de rede. O uso de plataformas como o *ArchiveReady*[141] facilitam a análise da arquivabilidade dos *site*s. As facetas de arquivabilidade, baseadas no estudo de Banos e Manolopoulos (2016), podem ser verificadas na Figura 25.

141 https://archiveready.com/

Figura 25 - Facetas de arquivabilidade

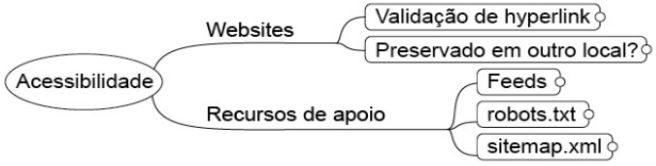


Fonte: Melo; Rockembach (2020, adaptado de Banos; Manolopoulos, 2016).

O funcionamento do *ArchiveReady* compreende executar uma solicitação HTTP para recuperar o *hipertexto* da página da *web*. e avaliar os atributos do *website* em detalhes, incluindo análise e validação de HTML e CSS, cabeçalhos de resposta HTTP, arquivos de mídia, *sitemap.xml* e *robots.txt*, *feeds* RSS e desempenho de transferência de rede.

Nas Figuras 26, 27, 28 e 29 a seguir, serão desdobradas cada uma das facetas de arquivabilidade: acessibilidade, padrões e conformidade, coesão e metadados.

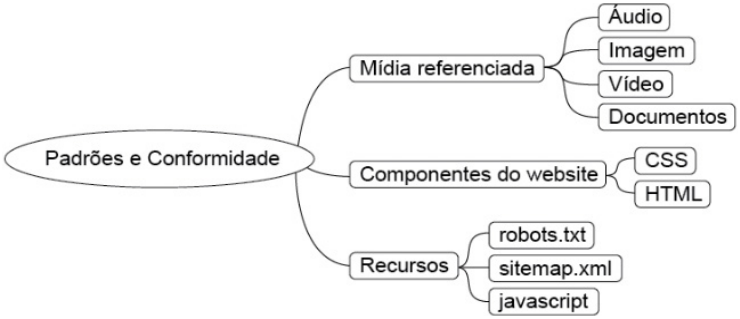
Figura 26 - Acessibilidade



Fonte: Melo; Rockembach (2020, adaptado de Banos; Manolopoulos, 2016).

A acessibilidade é um fator importante para a arquivabilidade de um *website*, já que o rastreador da *web* precisa ser capaz de acessar a página inicial, percorrer seu conteúdo e recuperar todos os recursos do *site* via solicitações HTTP padrão. Para garantir isso, é essencial fornecer referências que permitam que os rastreadores localizem e recuperem os conteúdos promovidos por meio desses recursos. Além disso, a validade dos *links* e a disponibilidade de informações sobre determinadas páginas também contribuem para a acessibilidade e arquivabilidade do *site*.

Figura 27 - Padrões e conformidade

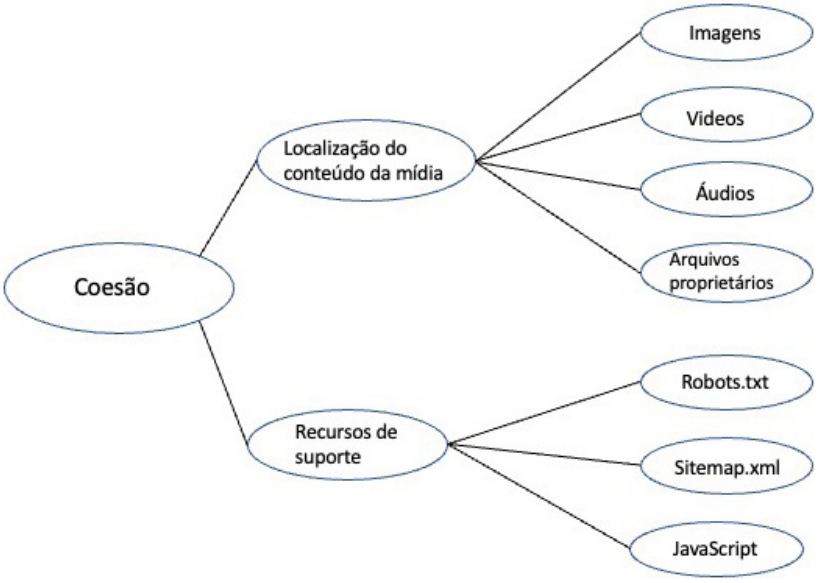


Fonte: Melo; Rockembach (2020, adaptado de Banos; Manolopoulos, 2016).

Neste sentido, recomenda-se que recursos digitais sejam representados em padrões conhecidos e transparentes, incluindo padrões proprietários amplamente adotados com ferramentas de suporte para validação e acesso. Ainda é preciso enfatizar a importância da conformidade com os padrões do W3C e a coesão do *website* para a robustez dos recursos contra alterações externas.

A Coesão se refere à capacidade de um *website* resistir a falhas em diferentes serviços da *web*. Em outras palavras, ela é importante para garantir que os rastreadores da *web* funcionem corretamente e possam identificar se os arquivos que compõem o *website* estão espalhados por diferentes serviços, como servidores específicos para imagens, *widgets*, *JavaScript*, entre outros. Por exemplo, as imagens usadas em um *website* podem estar hospedadas em um local diferente, o que pode resultar na não captura dessas imagens e causar problemas no arquivamento da página. A ideia é que manter todas as informações associadas ao mesmo *website* aumenta a robustez dos recursos, evitando alterações que possam ocorrer fora do *website* principal.

Figura 28 - Coesão

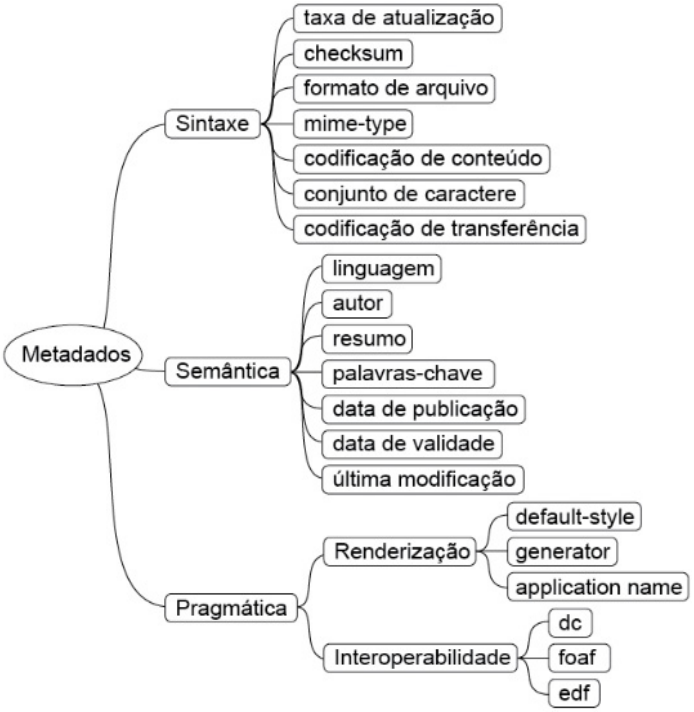


Fonte: Adaptado de Banos; Manolopoulos (2016).

Como pode-se ver na imagem, a Coesão é avaliada em dois níveis distintos: primeiro, é examinado quantos domínios são utilizados em relação à localização do conteúdo da mídia (imagens, vídeos, áudios, arquivos proprietários); segundo, é verificado quantos domínios são empregados em relação aos recursos de suporte, como o arquivo *robots.txt*, *sitemap.xml* e *JavaScript*.

Além disto, aconselha-se a verificação dos metadados em sintaxe, semântica e pragmática. A inclusão de elementos nos metadados é considerada importante para a extração e o gerenciamento dos arquivos da *web*.

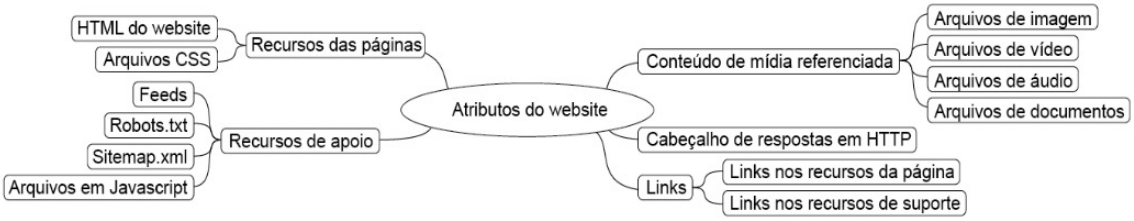
Figura 29 - Metadados



Fonte: Melo; Rockembach (2020, adaptado de Banos; Manolopoulos, 2016).

Seguir as práticas recomendadas e padrões internacionais pode aumentar a capacidade de arquivamento de um *site* e estudos de arquivabilidade de *website*s (MELO; ROCKEMBACH, 2020) podem ajudar no desenvolvimento de programas de preservação digital institucionais. Os atributos têm a intenção de avaliar e medir o potencial do *website* em atender os requisitos das facetas de arquivabilidade, Figura 30.

Figura 30 - Atributos do *site*



Fonte: Melo; Rockembach (2020, adaptado de Banos; Manolopoulos, 2016).

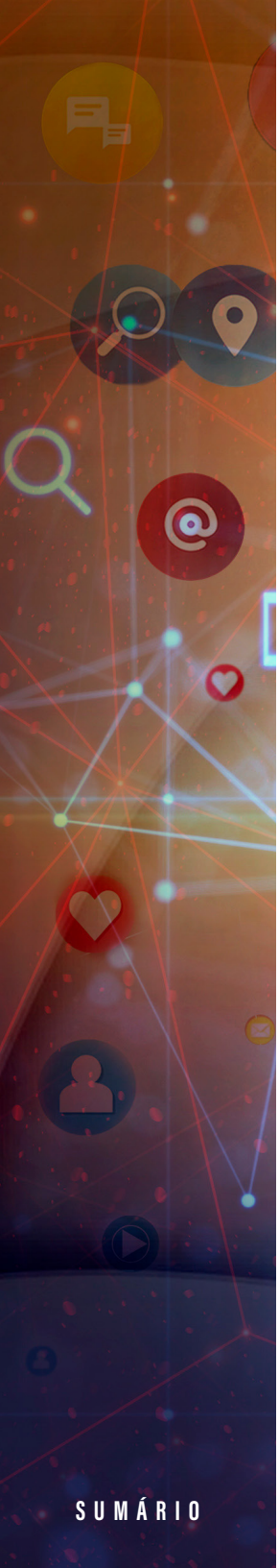
Outras duas ferramentas interessantes para utilização são o validador de padrões *(Markup Validator Service)*[142] e o validador de *links* (W3C *Link Checker*)[143]. Enquanto o primeiro verifica a validade da marcação de documentos da *web* em HTML, XHTML, SMIL, MathML, por exemplo, o segundo validador procura problemas em *links*, âncoras e objetos referenciados em uma página da *web*, folha de estilo CSS ou, recursivamente, em um *site* inteiro. Ambos os validadores fazem parte de uma série de ferramentas disponibilizadas pelo W3C[144].

Ainda, segundo Hockx-Yu (2012), para tornar-se um *site* mais arquivável, é recomendável garantir que o conteúdo importante seja explicitamente referenciado, ter um mapa do *site* e garantir que todos os *links* no *site* funcionem. Outras práticas recomendadas, como as emitidas pelo *Google* para rastreadores de mecanismos de pesquisa, também são aplicáveis a rastreadores de arquivamento da *web*. É importante observar que os rastreadores de arquivamento da *web* pretendem copiar todos os arquivos de todos os formatos pertencentes a um *site*, enquanto os rastreadores de mecanismos de pesquisa estão interessados apenas em arquivos que podem ser indexados.

142 https://validator.w3.org/

143 https://validator.w3.org/checklink

144 https://w3c.github.io/developers/tools/

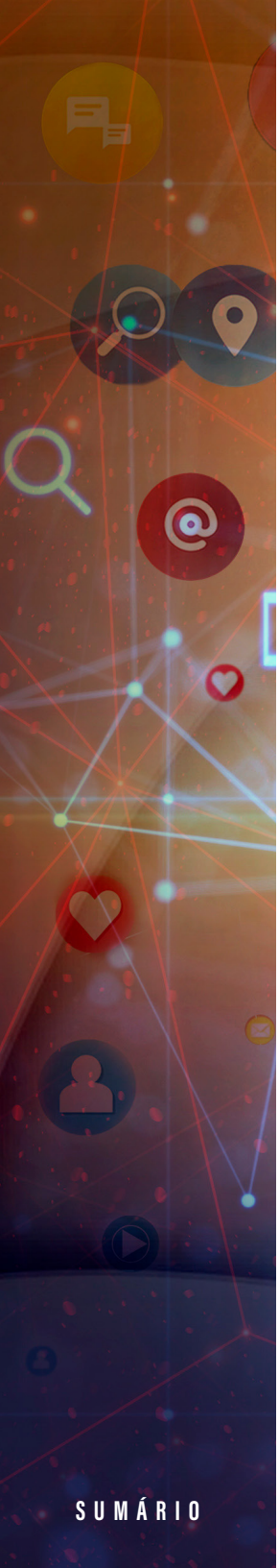


Por fim, a ênfase no arquivamento de *site*s está na captura do conteúdo intelectual, e não na aparência do *site*. Uma cópia imperfeita é melhor do que nenhuma cópia, desde que o conteúdo possa ser reproduzido razoavelmente. Segundo Bingham (2014), após a conclusão da revisão de qualidade, o arquivista pode endossar a cópia, rejeitá-la ou rastreá-la novamente com os parâmetros ajustados. Uma cópia pode ser rejeitada se o conteúdo principal não puder ser acessado, o rastreador sair do escopo, o *site* não renderizar corretamente, os *links* pularem para *site*s ativos ou o rastreador ficar preso em um *loop* indefinido. Também, o *feedback* do usuário durante o acesso ao arquivo é importante para a garantia de qualidade.

CONTROLE DE QUALIDADE

A qualidade em geral pode ser definida em um sentido funcional (adequado a um uso específico) ou em um sentido objetivo (combinando com características mensuráveis). O termo qualidade, como visto por Masanès (2006) é aplicado a acervos culturais em vários contextos e sentidos. Pode ser utilizado para qualificar o estado de conservação, a completude das peças ou de um acervo, o nível de conteúdo intelectual etc. Em cada caso, refere-se a uma escala ideal de perfeição em uma área específica: preservação física, cobertura de domínio, precisão de seleção, entre outros.

Há várias possibilidades de imprimir qualidade ao arquivamento da *web,* mas principalmente pela exaustividade e integridade do material (arquivos vinculados) arquivado dentro de um perímetro designado e a capacidade de renderizar a forma original do *site*, particularmente em relação à navegação e interação com o usuário. Em termos gráficos, a integralidade pode ser medida horizontalmente pelo número de pontos de entrada relevantes encontrados dentro do perímetro designado e verticalmente pelo número



de ligações relevantes encontrados a partir do ponto de entrada. Idealmente, os arquivos da *web* deveriam ser completos tanto na vertical como na horizontal. Mas isto é praticamente impossível de conseguir e por esse motivo, as prioridades têm de ser estabelecidas. (MASANÈS, 2005).

Criar um *site* arquivado que seja o mais próximo possível do *site* original "ao vivo" é um dos desafios mais difíceis na área de arquivamento da *web*. Deixar de capturar um *site* de forma adequada pode significar um registro histórico incompleto ou, pior, não evidenciar que o *site* sequer existiu. Reyes Ayala, Phillips e Ko (2014) colocam que um processo típico de garantia de qualidade nas práticas da comunidade de arquivamento da *web* envolve os seguintes elementos:

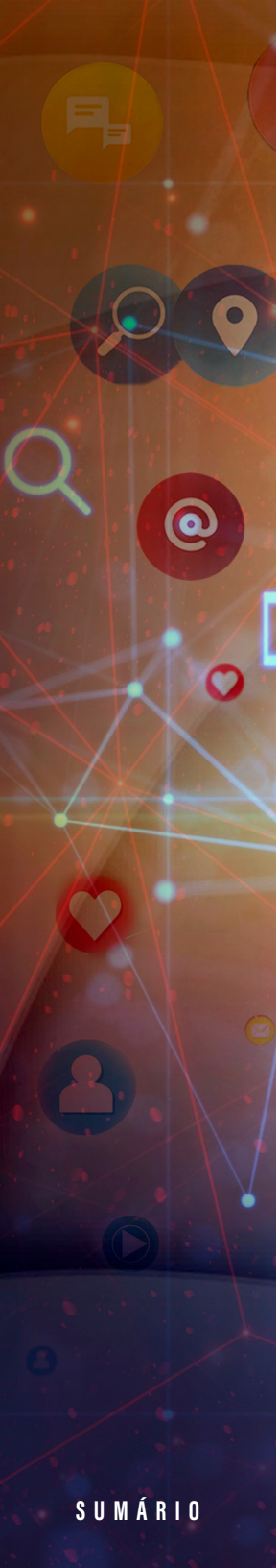
a. controle de qualidade realizado depois dos *site*s serem capturados: neste caso o controle de qualidade não é um processo que se inicia antes da fase de captura. Também, não é realizado durante o processo, mas é feito uma vez e num momento determinado, após o processo de captura;

b. controle de qualidade manual: envolve uma pessoa que olha para a versão arquivada do *site* e avalia a sua qualidade;

c. verificação utilizando *Wayback Machine*: método mais comum de avaliar a qualidade de um *site* arquivado é a sua visualização no *Wayback Machine*;

d. controle de qualidade realizado em todos os *site*s capturados: todo o *site* é verificado por meio do processo de controle de qualidade, e não apenas na página inicial ou de domínios específicos;

e. problemas de qualidade são verificados, quer numa folha de cálculo, quer noutro sistema, como uma base de dados e

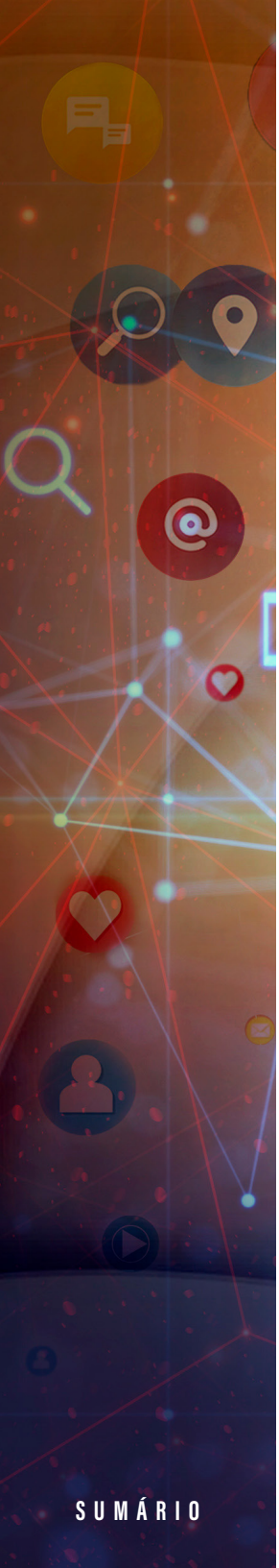


f. o controle de qualidade é realizado pela mesma pessoa que implementou o rastreamento. Este elemento de verificação da qualidade, sugere, de acordo com os resultados da pesquisa de Reyes Ayala, Phillips e Ko (2014), que as equipes de arquivamento da *web* em todo o mundo são pequenas, e uma pessoa pode ser responsável por várias funções diferentes, tais como determinar que *website* deve ser capturados, realizar o processo de captura, e verificar a qualidade de um *crawl*. Evidenciando que poucas instituições têm pessoal dedicado, exclusivamente, ao controle de qualidade.

O controle de qualidade é, portanto, um aspecto importante do processo de arquivamento da *web*, pois garante que o conteúdo da *web* arquivado seja preciso, completo e utilizável. Várias atividades diferentes são necessárias, incluindo a verificação da qualidade técnica dos arquivos da *web*, avaliação de seu conteúdo quanto à relevância e precisão e verificação de que os materiais arquivados são acessíveis e utilizáveis.

Um aspecto fundamental do controle de qualidade no arquivamento da *web* é a qualidade técnica dos próprios arquivos. Isso envolve garantir que as páginas da *web* sejam capturadas de maneira precisa e completa e que o conteúdo arquivado seja organizado e formatado adequadamente. Também envolve verificar se o conteúdo da *web* arquivado pode ser acessado e usado por pesquisadores e outros usuários e se está livre de erros ou defeitos técnicos. Para tal, devemos realizar testes de usabilidade para garantir que os arquivos da *web* possam ser facilmente navegados e pesquisados, e que o conteúdo arquivado seja apresentado de forma clara e compreensível. Também, pode envolver o fornecimento de materiais de apoio e treinamentos, como metadados e informações contextuais, para ajudar os usuários a entender e fazer uso do conteúdo arquivado.

Além da qualidade técnica, o conteúdo dos arquivos da *web* deve ser avaliado quanto à relevância e precisão, o que pode envolver



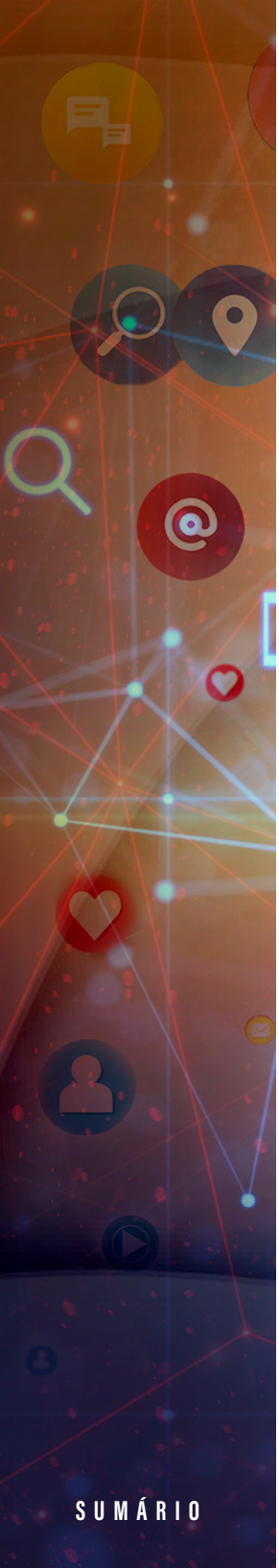
a revisão das páginas da *web* arquivadas para garantir que elas forneçam um registro abrangente e imparcial do conteúdo original e que não contenham erros ou informações enganosas. Também pode envolver a comparação do conteúdo arquivado com outras fontes, como outros arquivos da *web* ou fontes primárias, para verificar sua precisão e integridade.

O controle de qualidade é uma parte crítica do processo de arquivamento da *web*, pois garante que o conteúdo da *web* arquivado seja confiável e útil para pesquisadores e outros usuários. Ao realizar verificações regulares de qualidade e fazer as melhorias necessárias, as organizações de arquivamento da *web* podem ajudar a garantir a preservação e acessibilidade de longo prazo de conteúdo digital valioso.

Existem várias etapas que podem ser realizadas para garantir o controle de qualidade no arquivamento da *web*. Uma das mais importantes é usar ferramentas e tecnologias de arquivamento da *web* projetadas para capturar páginas da *web* de maneira fiel e abrangente. Isso pode incluir o uso de rastreadores da *web* e outras ferramentas que podem seguir *links* e capturar todo o conteúdo de uma determinada página, bem como qualquer mídia associada, como imagens e vídeos.

Outra etapa importante no controle de qualidade é verificar regularmente as páginas da *web* arquivadas para garantir que elas ainda estejam acessíveis e funcionando corretamente. Aqui é preciso realizar o teste das páginas arquivadas para garantir que sejam carregadas corretamente e que todos os *links* e outros recursos estejam funcionando conforme o esperado. Se algum problema for identificado, ele pode ser resolvido por meio do uso de ferramentas e técnicas de preservação.

Além das medidas técnicas de controle de qualidade, também é importante considerar o conteúdo das páginas da *web* arquivadas, isso envolve a revisão do processo de captura e armazenamento. O conteúdo de um *site* é o foco principal do arquivamento da *web*, incluindo todos os textos, imagens, vídeos e outros elementos multimídia que



compõem o *site*. Para manter a integridade do conteúdo arquivado, é importante verificar regularmente erros, omissões e outros problemas que possam afetar a precisão do material arquivado.

Além do próprio conteúdo, os elementos visuais de um *site* também são importantes no arquivamento da *web*, desde o *layout*, o *design* e a estética geral do *site*. Para garantir que o *site* arquivado pareça e funcione o mais próximo possível do original, é importante verificar regularmente inconsistências visuais e outros problemas.

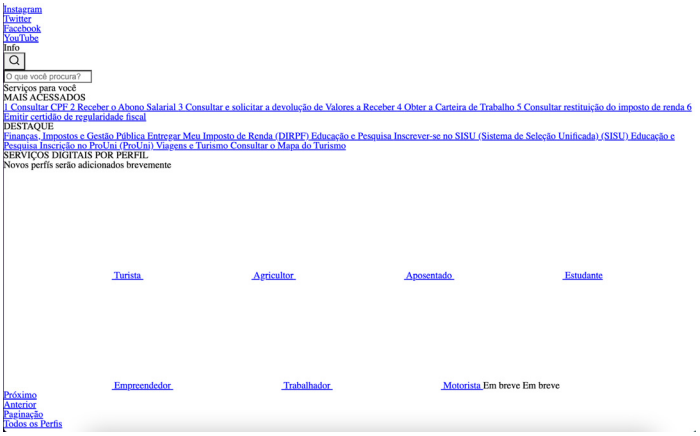
Outro aspecto fundamental do controle de qualidade do arquivamento da *web* é a reprodução do *site* ao longo do tempo. À medida que a tecnologia e os padrões da *web* evoluem, é importante garantir que o *site* arquivado possa ser acessado e visualizado usando os navegadores e dispositivos da *web* mais recentes. Isso envolve testes e atualizações regulares no *site* arquivado para manter sua compatibilidade com as tecnologias em constante mudança.

Finalmente, o acesso ao *site* arquivado também é uma consideração importante no controle de qualidade do arquivamento da *web*, envolvendo não apenas a garantia de que o *site* possa ser acessado pelos usuários, mas também que seja facilmente descoberto e pesquisável. Aqui o uso de metadados e outras técnicas de indexação para tornar o conteúdo arquivado mais acessível aos usuários são fatores a considerar.

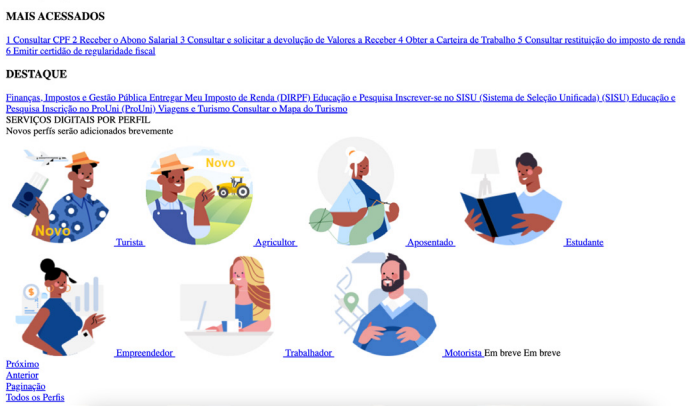
Depois que o rastreamento da *web* for concluído, é essencial executar a garantia de qualidade para que os *links* salvos sejam reproduzidos corretamente e conforme o esperado. É melhor identificar quaisquer problemas nesta fase, quando uma solução ainda pode ser encontrada, em vez de esperar meses até que o conteúdo da *web* "ao vivo" seja alterado. O chamado "rastreamento de *patch*" envolve capturar e corrigir os rastreamentos que não foram bem sucedidos em seu rastreamento original. Na Figura 31 verificamos três exemplos de capturas de *sites* com os respectivos controles de qualidade e os problemas identificados.

Figura 31 - Exemplos de capturas e execução de controle de qualidade

Caso 1: captura com baixa qualidade, podendo apresentar diversos erros de falta de conteúdo textual, multimídia não reproduzível, estilo (css), e mensagens de erro



Caso 2: captura com qualidade melhor do que o caso anterior, com reprodução do conteúdo original e imagens, mas podendo haver problemas de reprodução de alguns arquivos multimídia e de estilo (css)

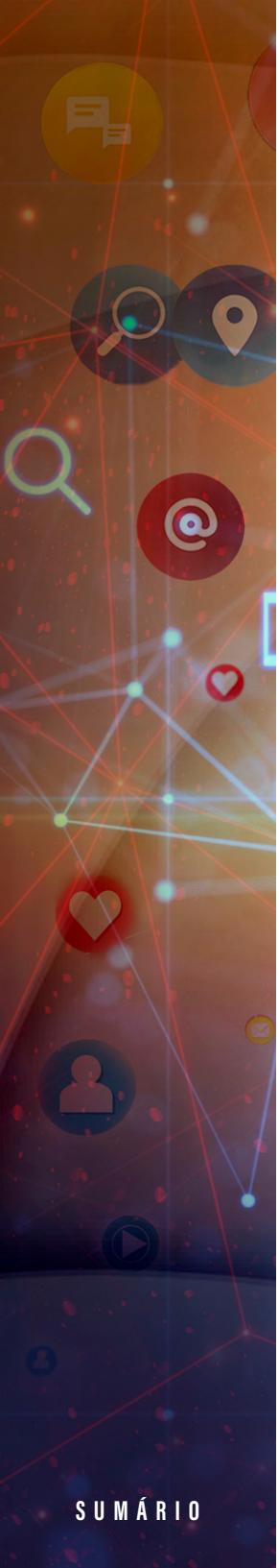


Caso 3: captura de alta qualidade reproduzindo, conforme original, todos os elementos textuais, multimídia e de estilo (css).

Fonte: Os autores, captura do domínio .gov.br

O controle de qualidade do arquivamento da *web* é uma parte essencial da preservação do conteúdo do *site* ao longo do tempo e envolve verificações regulares de precisão, consistência visual, compatibilidade de reprodução e acesso ao material arquivado. Ao manter um alto nível de controle de qualidade, os arquivistas da *web* podem garantir que o conteúdo dos *sites* seja preservado para que as gerações futuras possam acessá-lo e apreciá-lo.

O controle de qualidade do arquivamento da *web* pode não estar relacionado, tão intimamente, apenas com a tecnologia e assumir uma faceta mais geral e abstrata. Reyes Ayala, Phillips e Ko (2014) defendem que a qualidade de um arquivo da *web* consistiria em: a) **correspondência**: dimensão de qualidade mais exclusiva dos arquivos da *web*. Requer equivalência, ou pelo menos uma estreita semelhança, entre o recurso original e o recurso arquivado. Podemos entender esse aspecto com o seguinte exemplo, num arquivo analógico tradicional, o recurso arquivado é ele próprio o recurso original ou uma cópia do mesmo, consequentemente, existe uma correspondência de um para outro, entre o recurso original e o recurso arquivado; b) **completude**: o recurso arquivado conter todos os seus elementos constituintes; c) **coerência**: o recurso arquivado



integra diversos elementos de uma forma lógica e consistente e d) **integridade**: os elementos de dados que constituem o recurso capturado não estão corrompidos e não contêm erros.

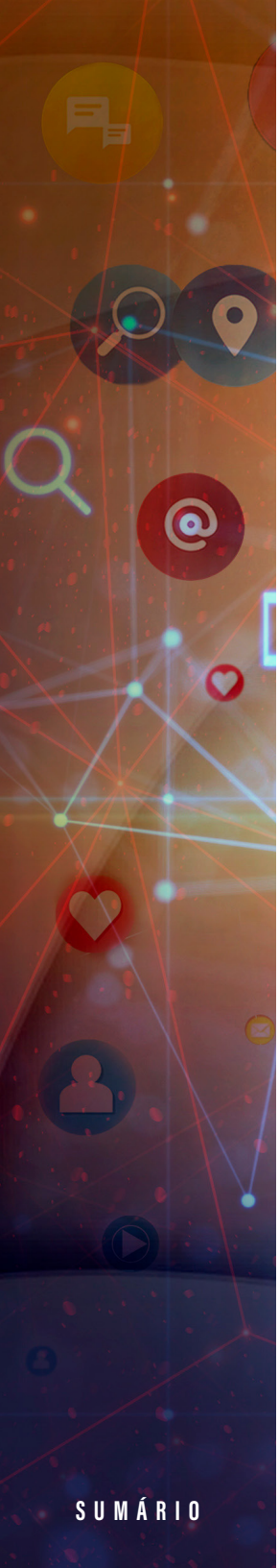
Os mesmos autores citados no parágrafo anterior ao pensar nos erros de acesso ou reprodução como problemas técnicos que ocorrem devido a uma má conexão ou problemas com um servidor, argumentam que ficariam com os dois erros mais importantes: a **representação errada** (qualidade de reprodução) e **falta de conteúdo** (qualidade de captura). A qualidade de captura é uma medida da coerência, completude e integridade de um *site* arquivado, enquanto a qualidade de reprodução seria uma medida que se refere a sua correspondência. A gravidade desses problemas aponta para a necessidade do desenvolvimento de normas e ferramentas para facilitar o processo de controle de qualidade.

METADADOS APLICADOS AOS ARQUIVOS DE *WEBSITES*

Arquivos, bibliotecas, museus e centros de informação utilizam padrões de metadados por várias razões óbvias: para aumentar a qualidade e a estabilidade das informações; para melhorar a compatibilidade das estruturas de dados e para facilitar tanto a recuperação quanto a troca de informações.

O Modelo de Referência para Sistema Aberto de Arquivamento de Informação (SAAI), ABNT NBR 15472:2007, baseado no *Open Archival Information System* (OAIS), norma ISO 14721:2003[145], constitui-se de um modelo abstrato, flexível utilizado por desenvolvedores de sistemas de arquivamento, permitindo que a informação flua sem a perda dos elementos que a compõem (REZENDE, 2022).

145 https://www.iso.org/standard/24683.html

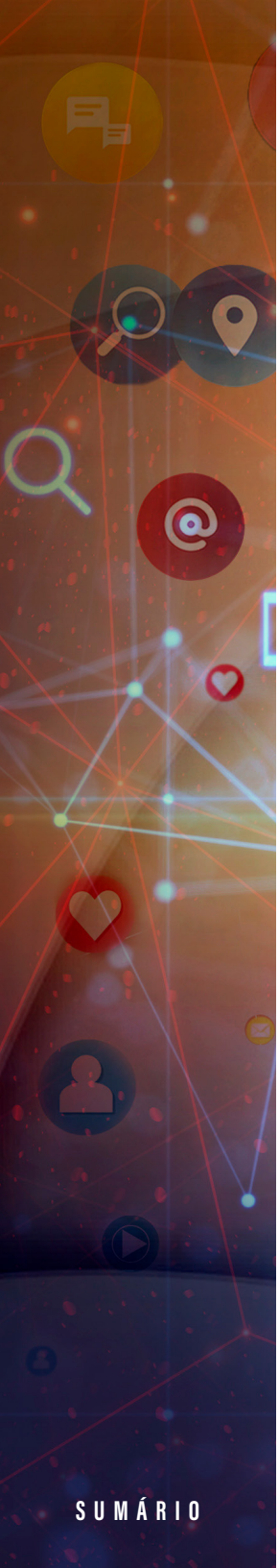


Este modelo de referência destaca especificamente a informação digital, tanto como forma primária para a guarda de informação quanto como informação de apoio a materiais arquivados de forma física e digital. (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2007).

O Modelo de Referência SAAI aborda um conjunto completo de funções arquivísticas para a preservação da informação, incluindo admissão, arquivamento, gerenciamento de dados, acesso e disseminação; a migração de formatos, os modelos de dados usados para representar a informação e a preservação da informação e o intercâmbio de informação digital entre arquivos, dentre outros aspectos. Aplica-se, especificamente, a organizações com responsabilidade de tornar informação disponível por longo prazo.

Alguns dos padrões relacionados ao Modelo SAAI incluem questões sobre metadados: padrão(ões) para submissão de metadados digitais e fontes de dados digitais e padrão(ões) de protocolo para pesquisa e recuperação de metadados de fontes de dados digitais. Os metadados são elementos para apoiar a pesquisa e a recuperação de Pacote de Arquivamento de Informação (PAI), por exemplo: quem, o que, quando, onde, por quê. Estão presentes em todas as etapas do Modelo SAAI: da admissão, que começa com a seleção de todos os documentos eletrônicos produzidos e passíveis de arquivamento a longo prazo, até o acesso. As informações sobre os acervos devem estar disponíveis em múltiplos níveis de detalhamento e por meio de múltiplas fontes, como forma de suprir necessidades de diversos consumidores de informação.

Metadados são informações estruturadas que descrevem e permitem localizar, gerenciar, controlar e preservar informações ao longo do tempo. Os metadados têm a mesma função de um rótulo, e assim como outros rótulos, os metadados fornecem informações sobre um objeto. Metadados descrevem objetos digitais (dados, documentos etc.) e documentam adequadamente os objetos, permitindo que os usuários entendam e rastreiem detalhes importantes de um objeto ou conteúdo e também facilitam a pesquisa e a recuperação.



Além de descrever um objeto digital, os metadados definem permissões, direitos de acesso, compartilhamento, reutilização, redistribuição e políticas, bem como os requisitos técnicos para visualização, acesso e preservação de objetos digitalizados ou concebidos originalmente em formato digital. A norma ISO/IEC 11179[146] apresenta e discute ideias fundamentais de elementos de dados, domínios de valor, conceitos de elementos de dados, domínios conceituais e esquemas de classificação essenciais para a compreensão deste conjunto de padrões.

No contexto do arquivamento da *web*, os metadados são usados para descrever e catalogar as páginas da *web* que estão sendo arquivadas. Esses metadados podem incluir informações como a data em que a página foi arquivada, a URL da página, o título da página e o autor da página, entre outros elementos necessários a sua identificação.

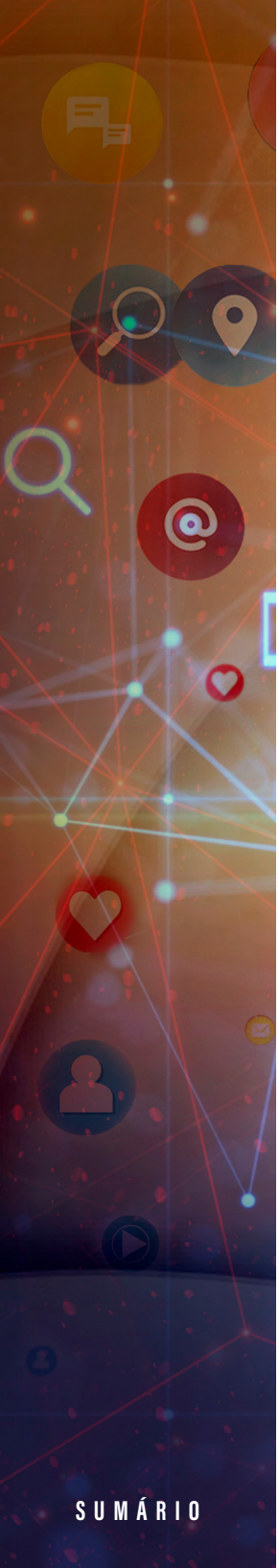
Os metadados são classificados em cinco tipos: descritivos, estruturais, administrativos, técnicos e de preservação (Quadro 6). Porém, nem todos são utilizados para todos os níveis de arquivamento ou tipos de registros.

Quadro 6 - Tipo de metadados e sua descrição

TIPO DE METADADOS	DESCRIÇÃO
Descritivos	Visam a pesquisa, recuperação e identificação do objeto digital. Exemplo: título, autor, assunto e palavras-chave.
Estruturais	Vinculam de forma hierárquica distintos objetos digitais (textos, imagens, áudios, dados etc.).
Administrativos	Apresentam informações que suportam a gerência dos recursos eletrônicos. Contêm a forma e a ocasião em que os recursos foram gerados, informações técnicas, direitos ou permissões de acesso e uso.
Técnicos	Especificam os aspectos técnicos dos arquivos e dos seus formatos.
Preservação	Incluem informações necessárias ao arquivamento e salvaguarda dos objetos digitais ao longo do tempo, como o formato de preservação e somas de verificação (*checksum*) para verificações de integridade.

Fonte: os autores.

146 https://www.iso.org/standard/78914.html



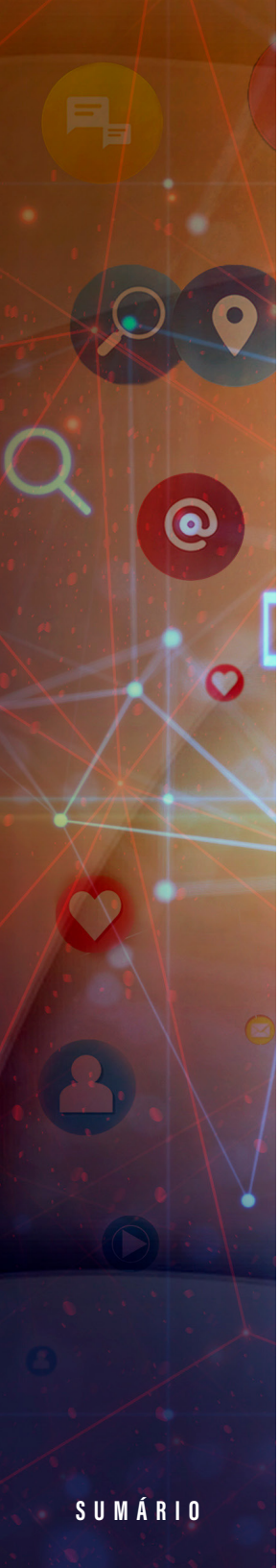
Foementon e Gracioso (2022) identificaram outros tipos de metadados, as linguagens de marcação, que incluem metadados e sinalizadores para outros recursos estruturais ou semânticos no conteúdo. Os mesmos autores subdividem os metadados administrativos em três categorias: metadados técnicos, aqueles que indicam os aspectos e as dependências técnicas de um arquivo digital para decodificá-lo e renderizá-lo; metadados de preservação, aqueles que incluem informações sobre as dependências de *hardware* e de *software* exigidas para a gerência de um arquivo digital a longo prazo e metadados de direitos, que contêm as informações para apoiar à gestão dos direitos de propriedade intelectual associados ao conteúdo.

Desenvolver práticas para a criação de metadados consistentes que abordem características únicas dos *site*s e coleções, mais especificamente os relacionados com os metadados descritivos para tornar possível um recurso *web* ser descoberto pode ser bastante desafiador, principalmente pela falta uma abordagem comum para a criação de metadados.

Na tentativa de desenvolver recomendações para metadados descritivos e facilitar a descoberta de conteúdos arquivados da *web* a *On-line Computer Library Center* (OCLC) criou um W*eb Archiving Metadata Working Group* (WAM)[147] com o objetivo de estabelecer alguns padrões, principalmente para os metadados descritivos, levando em conta as características únicas dos *websites* arquivados e com vistas a auxiliar as instituições a melhorar a consistência e eficiência das suas práticas de metadados nesta área emergente. O estudo desenvolvido por Dooley e Bowers (2018) pretendia:

a. desenvolver práticas comuns de metadados, padronizadas em termos de utilização de normas para metadados descritivos para conteúdo arquivado da *web*, tendo em conta as necessidades dos usuários finais e dos profissionais de metadados;

147 https://www.oclc.org/research/areas/research-collections/wam.html



b. definir um conjunto enxuto de elementos de dados com notas de utilização para orientar a preparação do conteúdo dos dados;

c. assegurar que os elementos de dados pudessem ser utilizados em consonância com outras normas que tenham muito mais conjuntos de elementos de dados granulares;

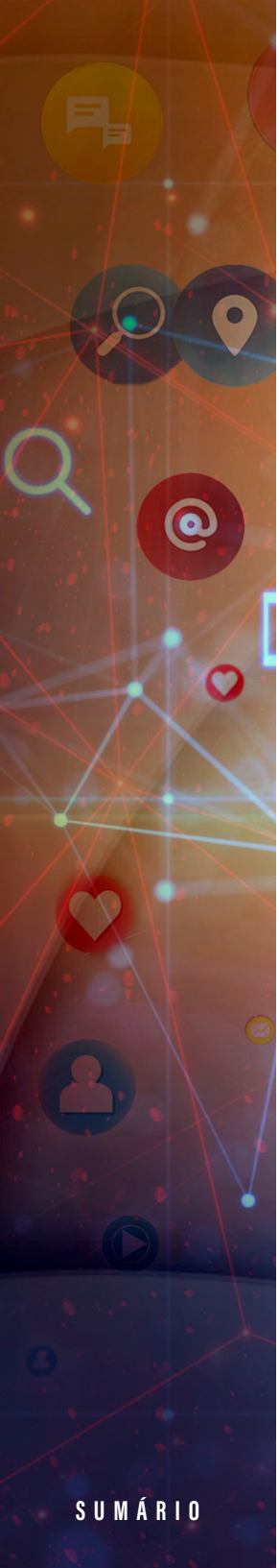
d. fornecer uma ponte entre as abordagens bibliográfica e arquivística para a descrição;

e. utilizar uma abordagem escalável que não exigisse uma descrição em profundidade nem mudanças extensivas para registos ao longo do tempo e

f. permitir que os profissionais tivessem confiança de que estavam aplicando práticas consistentes nesta área emergente.

As práticas recomendadas pela WAM, de acordo com o relatório das autoras citadas anteriormente, podem ser utilizadas por qualquer instituição ou pessoa com necessidade de descrever conteúdo da *web*. As práticas descritivas nas bibliotecas e arquivos diferem em muitos aspectos, basicamente cada comunidade tem o seu próprio conjunto de normas tanto para a descrição como para estruturar os dados. Os metadados para descrever o conteúdo da *web* são criados dependendo de onde se situa a responsabilidade pelos metadados e/ou custódia de materiais, incluindo catalogação em bibliotecas, arquivos e repositórios digitais.

A descrição dos metadados é aplicável tanto à descrição no nível do *site*, como no nível da coleção, o que vai determinar o seu conteúdo. Dooley e Bowers (2018) exemplificam: o título de um *site* geralmente é uma transcrição do texto proeminente do próprio *site*, enquanto o título de uma coleção da *web* é elaborado pela instituição que coletou o conteúdo; o criador de um único *site*, como uma página inicial institucional, *blog* ou *feed* de redes sociais, geralmente é facilmente identificado, enquanto uma coleção de *site*s de



um evento específico ou tópico raramente tem um criador; a data registrada para a descrição de uma coleção temática pode incluir um intervalo de datas entre as quais a captura foi realizada.

O profissional de metadados precisa adotar uma abordagem arquivística e/ou bibliográfica, entendendo os diversos esquemas, padrões e normas de descrição de metadados (DACS, EAD, MARC, *Dublin Core*, RDA, MODS); entender que os elementos de dados variam amplamente e o mesmo elemento pode ter múltiplos significados; decidir o nível de descrição que será adotado, *website* único, coleção de *websites*, URLs iniciais e pensar na escalabilidade e como ultrapassar as barreiras dos recursos limitados.

Estes profissionais também devem agir analisando padrões de metadados e diretrizes institucionais: RDA (bibliotecas), DACS (arquivos), *Dublin Core*; avaliando os tipos de registros de metadados já existentes (*ArchiveGrid*[148], *Archive-It, WorldCat*[149]); identificando dilemas específicos ao arquivamento da *web;* incorporando resultados de estudos e casos de sucesso e preparando documentação e relatórios.

A descrição dos metadados de um *website* arquivado requer tomar decisões em conjunto com e para a comunidade sem esquecer que as soluções encontradas e decisões devem ser documentadas. Alguns dilemas enfrentados pelos profissionais de metadados foram resumidos no Quadro 07.

Quadro 07 - Dilemas enfrentados pelos profissionais de metadados

METADADO	DILEMA
Criador/proprietário do *website*	é o editor? é o criador? é assunto? ou podemos considerar os três?

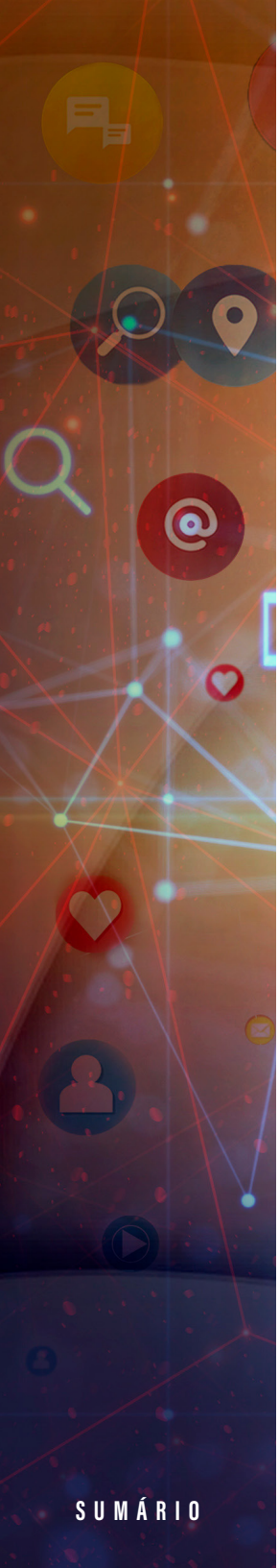
148 https://researchworks.oclc.org/archivegrid/

149 https://www.worldcat.org/

Instituição hospedeira	é a que colhe os *websites?* é a que hospeda os *websites*? é o repositório? é o criador? é o editor? é o selecionador?
Editor	um *website* tem um editor?
Datas: quais são importantes e passíveis de registro?	início/fim da existência do *website*? data(s) de captura pelo repositório? data do conteúdo?
Extensão: como registrar?	cinco *website* arquivados? cinco recurso *on-line*? 6,25 GB? aproximadamente 300 *websites*?
Proveniência: refere-se?	ao proprietário do *website*? ao repositório que coleta e hospeda o *website*? as maneiras pelas quais o *website* evoluiu? a frequência e datas da captura?
Avaliação	qual o motivo para o *website* merecer ser arquivado? é uma coleção de um conjunto de *websites*? quais as partes que foram ou não arquivadas?
Formato: é importante que a descrição indique claramente que o recurso é um *website*? onde?	no título? na extensão? na descrição?
URL: quais URLs devem ser incluídas?	de acesso? da página de destino?
Registro MARC: como um *website* deve ser escrito?	recurso contínuo? recurso eletrônico? publicação textual? material misto? manuscrito? nenhum desses?

Fonte: Adaptado de Dooley et al. (2017, tradução nossa).

Existem diversos padrões de metadados, alguns profissionais seguem tradições bibliográficas, outros adotam uma abordagem arquivística ou ainda, abordagens híbridas. O principal padrão



de descrição em bibliotecas, o *Resource Description and Access* (RDA)[150] adota uma abordagem totalmente bibliográfica, aplicando-se melhor para *sites* "ao vivo" e não recomendável para *sites* arquivados onde o conteúdo arquivado, que muda com frequência, prioriza a ligação entre recursos. Algumas regras não se adaptam a descrição de alguns elementos dos *sites* arquivados e estão intimamente relacionados com o padrão *Machine-Readable Cataloging* (MARC)[151]. Este modelo é utilizado em vários repositórios da *Columbia University*, mas cada repositório da Universidade tem autonomia para adicionar elementos próprios.

A *Library of Congress* também utiliza metadados criados em MARC, exportados no formato *Metadata Object Description Schema* (MODS)[152] para acesso público. A descrição resumida é derivada do cabeçalho HTML da página inicial do *website*. A *University of Texas, Human Rights Documentation Initiative*[153] também utiliza o padrão MODS, registrando uma nota de proveniência indicando por quem o *website* foi coletado e neste caso a Universidade do Texas é registrada como "criadora".

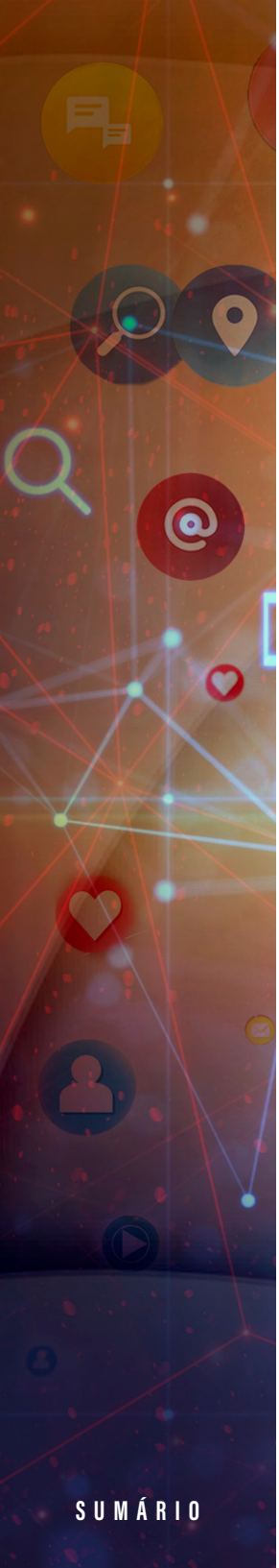
O formato MARC fornece o mecanismo pelo qual os computadores trocam, usam e interpretam informações bibliográficas, e seus elementos de dados constituem a base da maioria dos catálogos de bibliotecas usados atualmente. O MODS é constituído por um conjunto de elementos que podem ser usados para uma variedade de propósitos, e particularmente para aplicativos de biblioteca. Como esquema XML, destina-se a transportar dados selecionados de registros MARC21, bem como permitir a criação de registros originais de descrição de recursos.

150 https://www.loc.gov/aba/rda/

151 https://www.loc.gov/marc/

152 https://www.loc.gov/standards/mods/

153 https://repositories.lib.utexas.edu/handle/2152/4022



O *Describing Archives: A Content Standard* (DACS)[154] é um conjunto de regras de descrição de arquivos, documentos pessoais e colecções de manuscritos e pode ser aplicado a todos os tipos de material, independentemente da forma ou meio. Podem também ser aplicadas à descrição de colecções criadas intencionalmente e a itens individuais. O Modelo é utilizado no *Harvard University Archives*[155] e a descrição é em nível de coleção, ou seja, vários *websites* em um único registro. A abordagem é totalmente arquivística.

O *Dublin Core*[156] (DC) é uma norma de estrutura de dados que contém 15 elementos utilizados na descrição de recursos, com o propósito de encorajar uma abordagem padrão de metadados simplificados para utilização por qualquer comunidade de prática. O *Dublin Core* é amplamente utilizado para descrever objetos digitais, inclusive por utilizadores do *Archive-It*. A *Dublin Core™ Metadata Initiative* (DCMI) apoia a inovação partilhada na concepção de metadados e melhores práticas numa vasta gama de objetivos e modelos de negócios.

Já o *New York Art Resources Consortium* (NYARC)[157], que é composto pela colaboração entre três museus da cidade de Nova York para coletar *websites* que documentam vários segmentos do mundo da arte, os metadados locais são criados no *Archive-It* e, em seguida, reaproveitados em MARC para registro no *WorldCat* da OCLC e são exportados para o catálogo do NYARC. Os dados são registrados em MARC21, *Dublin Core*, MODS e vários outros padrões.

Em vez de simplesmente recomendar o uso do *Dublin Core* para descrição do conteúdo dos arquivos da *web*, Dooley e Bowers (2018) consideraram cuidadosamente quais elementos DC eram aplicáveis e quais poderiam ser adaptáveis. Assim, chegaram à reco-

154 https://saa-ts-dacs.github.io/

155 https://library.harvard.edu/libraries/harvard-university-archives

156 https://www.dublincore.org/

157 https://nyarc.org/

mendação dos seguintes elementos: Colaborador, Criador, Data, Descrição, Idioma, Relação, Assunto e Título, ainda incluíram, por se diferenciarem em nome e/ou significado: Coletor, Extensão, Gênero/ Forma, Direitos, Fonte de Descrição e URL, conforme Quadro 8.

Quadro 8 - Elementos de dados para descrição de *sites* arquivados, definição e uso

ELEMENTO DE DADOS	DEFINIÇÃO E USO
COLLECTOR (Coletor)	Organização responsável pela curadoria e administração de um *site* ou coleção arquivada. **Usado para:** organização que seleciona o conteúdo da *web* para arquivamento, cria metadados e executa outras atividades associadas à "propriedade" de um recurso.
CONTRIBUTOR (Colaborador)	Organização ou pessoa corresponsável pelo conteúdo de um *site* ou coleção arquivada. **Usado para:** entidades que fizeram contribuições significativas, mas secundárias, ao conteúdo de um *site* ou coleção e que não estão especificadas no elemento Criador.
CREATOR (Criador)	Organização ou pessoa responsável principalmente pela criação do conteúdo intelectual de um *site* ou coleção arquivada. **Usado para:** uma organização quando ela tiver claramente a responsabilidade principal pela criação do conteúdo intelectual. Em caso de dúvida, use o *Contributor*. **Não usar:** para descrever uma coleção de *site*s relacionados entre si apenas por assunto.
DATE (Data)	Única data ou intervalo de datas associadas a um evento no ciclo de vida de um *site* ou coleção arquivada. **Usado para:** registrar qualquer data conhecida ou intervalo de datas associado a um *site* ou coleção arquivado, isso ajudará os usuários a entender o conteúdo. Sempre esclareça o significado deste elemento adicionando palavras apropriadas para permitir a compreensão do usuário.
DESCRIPTION (Descrição)	Uma ou mais notas que explicam o conteúdo, contexto e outros aspectos de um *site* ou coleção arquivada. **Usado para:** informação textual sobre múltiplos aspectos do conteúdo. Informação biográfica ou histórica sobre a organização ou pessoa responsável pela criação do conteúdo e um resumo são essenciais para a compreensão dos metadados do *website* pelos usuários. Ainda, o histórico de custódia, avaliação e acréscimos. Como consiste em texto não estruturado, é aqui que o conteúdo e o contexto do *site* ou captura podem ser articulados de forma mais clara.

EXTENT (Extensão)	Indicação do tamanho de um *site* ou coleção arquivada. **Usado para:** informar o número de *site*s, a quantidade de dados armazenados (em *megabytes, gigabytes* ou outra medida) ou o número e/ou tipo de arquivos coletados. Use uma forma que dê aos usuários uma indicação útil de quanto conteúdo foi arquivado.
GENRE/FORM (Gênero/forma)	Termo que especifica o tipo de conteúdo em um *site* ou coleção arquivada. O uso de um vocabulário controlado é fortemente recomendado para evitar inconsistência. Selecione um vocabulário que seja usado pelas comunidades que provavelmente se beneficiarão do conteúdo descrito.
LANGUAGE (Língua/ idioma)	Idioma(s) do conteúdo arquivado, incluindo recursos visuais e de áudio com componentes de idioma.. **Usado para:** qualquer *site* ou coleção em que o idioma seja essencial para a compreensão do conteúdo. Se estiver em mais de um idioma, acrescente todos os que forem significativos.
RIGHTS (Direitos)	Declarações de direitos legais e permissões concedidas pela lei de propriedade intelectual ou outros acordos legais. **Usado para**: para dois tipos distintos de informações: condições que restringem o acesso do usuário ao conteúdo arquivado e se a permissão dos detentores de direitos autorais para reutilização após o acesso foi obtida. Conforme apropriado, ambos os tipos podem ser especificados em uma única ocorrência.
RELATION (Relação)	Apresenta o relacionamento mais comum, entre uma coleção e os subgrupos ou itens dentro dela. Ao descrever um único *site* que faz parte de uma coleção de *site*s arquivados, incluir o título da coleção no elemento *Relation* para fornecer o contexto no qual o *site* foi coletado. **Usado para:** expressar relações parte/todo entre um único *site* arquivado e qualquer coleção à qual ele pertença.
SOURCE OF DESCRIPTION (Fonte da descrição)	Informações sobre a coleta ou criação dos próprios metadados, como fontes de dados ou a data em que os dados de origem foram obtidos. **Usado para:** identificação da fonte de todos ou alguns dos metadados, principalmente para descrições de *sites* únicos. Aspectos básicos de um *site* (nome do criador, título etc.) podem mudar significativamente ao longo do tempo, mas é improvável que a instituição responsável tenha condições de identificar essas mudanças, muito menos atualizar os metadados. Por esse motivo, deve ser incluída a data em que o *site* foi capturado e o local de onde as informações foram retiradas.

SUBJECT (Assunto)	Tópicos que descrevem o conteúdo de um *site* ou coleção arquivada. **Usado para:** identificar o nome pelo qual um *site* ou coleção arquivada é conhecida. Para assuntos tópicos, nomes de lugares geográficos e nomes de pessoas ou organizações que pertencem ao elemento Assunto. Podem ser utilizados quantos termos forem necessários para fornecer acesso ao conteúdo. O uso de um vocabulário controlado é fortemente recomendado para melhorar a consistência. Selecione um vocabulário que seja usado pelas comunidades que provavelmente se beneficiarão do conteúdo descrito. Pode ser apropriado duplicar o nome do Criador no elemento *Subject*, especialmente para *site*s organizacionais nos quais o conteúdo é sobre a organização.
TITLE (Título)	O nome pelo qual um *site* ou coleção arquivada é conhecida Incluir pelo menos um elemento de título.
URL	Endereço de Internet para um *site* ou coleção arquivada. **Usado para:** registrar URLs, URNs ou URIs que são úteis para os usuários, particularmente URLs de divulgação e acesso. Inclua texto explicando sua função. Repita o elemento quantas vezes forem necessárias.

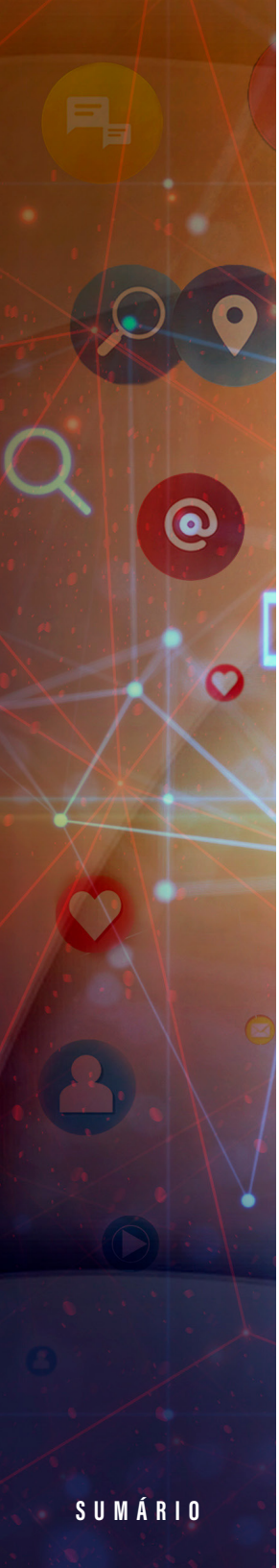
Fonte: Compilado de Dooley e Bowers (2018, tradução nossa).

Outros autores como Kim e Lee (2007) também sugerem metadados descritivos e administrativos para o arquivamento da *web*. Baseados em estudos de metadados de instituições, como a *National Library of Australia* e do *Smithsonian Institution Archives*, concluem que o arquivamento da *web* exige alguns metadados mais detalhados. Assim, recomendam além dos 15 elementos simples do DC, outros elementos como:

a. disponibilidade (*availability*), como obter o conteúdo arquivado da *web*;

b. público (*audience*), usuários a quem se destina;

c. data da captura (*date captured*), a data associada com a captura do *site*;



d. data de validação (*date validated*), a data em que a página *web* foi validada, como sendo de fato codificada, usando o *W3C Markup Validation Service* ou outros serviços;

e. mandato (*mandate*), políticas, termos ou regulamento;

f. arquivo de captura (*harvest file*), informação de que a descrição de metadados do arquivo capturado foi fornecido por uma instituição;

g. título alternativo (*alternative title*), pessoa ou local que pode ser usada como título alternativo;

h. condição de acesso (*access condition*), declaração sobre o escopo do uso dos recursos;

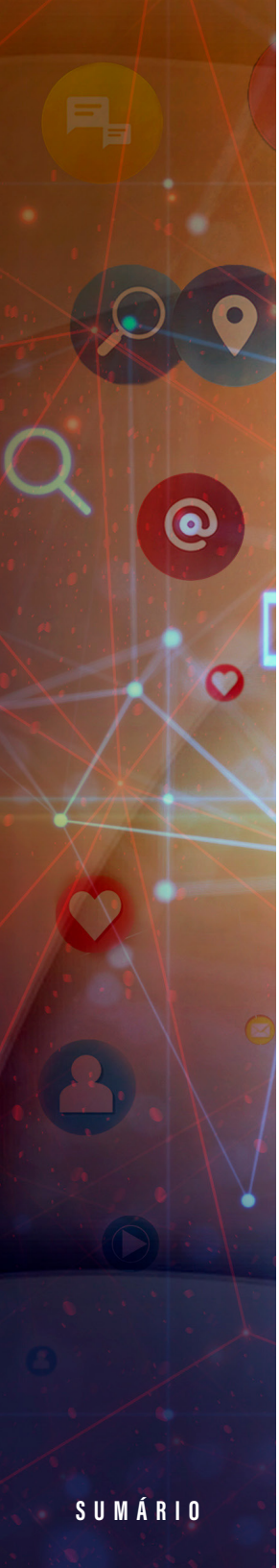
i. título de coleção (*collection title*), nome de um domínio ou projeto especial, ou o nome da coleção de informações especialmente organizadas;

j. data de modificação de metadados (*date metadata modified*), data na qual os metadados sofreram mudanças;

k. método de captura (*collection method*), padronizar o método de captura de recursos em, por exemplo: automático, manual ou transferido e

l. ferramenta de captura (*collection tool*), *softwares* necessários no processo de captura dos recursos.

Desenvolver um conjunto por demais enxuto de elementos de metadados destinados à utilização por uma variedade de padrões de *websites* e não estender os elementos utilizando técnicas tais como modificadores ou subcampos, pode dificultar ações automáticas para elementos que possam ter vários significados, tais como Data, Extensão e URL. Portanto, para definir o perfil de metadados e a organização da informação devemos conhecer o tipo de *website* a



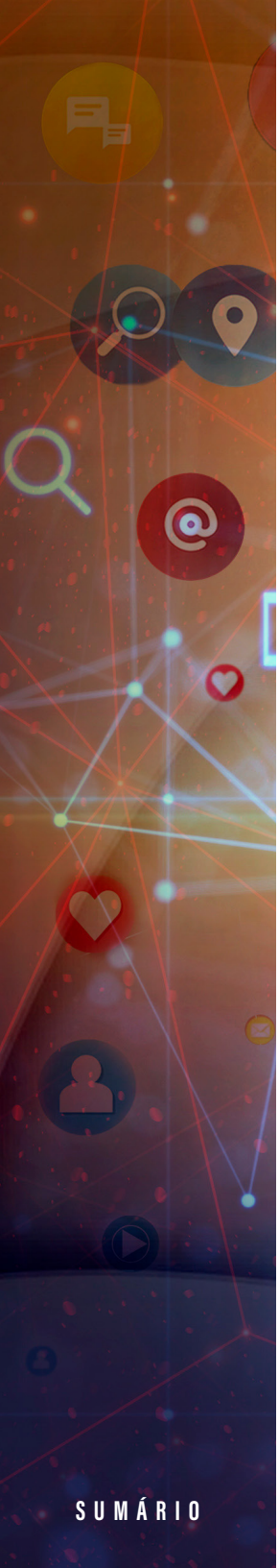
ser arquivado e seu conteúdo. O resultado será a organização, exibição e descoberta da informação mais qualificada.

Os metadados são uma parte importante do arquivamento da *web* porque ajudam a tornar as páginas da *web* arquivadas mais facilmente detectáveis e pesquisáveis. Ao fornecer essas informações, os usuários podem encontrar rápida e facilmente as páginas da *web* que estão procurando e também podem usar os metadados para aprender mais sobre o conteúdo da página.

Além de ajudar na capacidade de descoberta e pesquisa, os metadados também podem colaborar para preservar o contexto e o significado da página da *web*, garantindo que as páginas da *web* arquivadas sejam acessíveis e utilizáveis no futuro, mesmo com a evolução da tecnologia empregada para criá-las. Por exemplo, os metadados podem ser utilizados para registrar informações sobre a hora e o local em que a página foi criada, bem como quaisquer eventos ou desenvolvimentos ocorridos no momento. Isso pode ser útil para historiadores e outros pesquisadores que estudam a história da *web*.

Os metadados de arquivamento da *web* geralmente são criados e gerenciados usando ferramentas de *software* especializadas projetadas para essa finalidade. Essas ferramentas permitem que os arquivistas da *web* criem e gerenciem metadados de acordo com padrões e diretrizes específicas. Há inúmeras formas para a criação dos metadados, assim como diversas ferramentas que realizam essa descrição, porém, a maioria delas concentram-se na captura, que é uma das atividades do processo. Por esse motivo, muitas vezes é preciso a integração de outras ferramentas para fornecer metadados descritivos e de acesso aos recursos.

O estudo realizado em 2018 por Venlet e colaboradores aponta a necessidade de enriquecer a descrição dos recursos arquivados, principalmente no que diz respeito à contextualização da conformação e os conteúdos dos arquivos da *web*. Ainda, as autoras

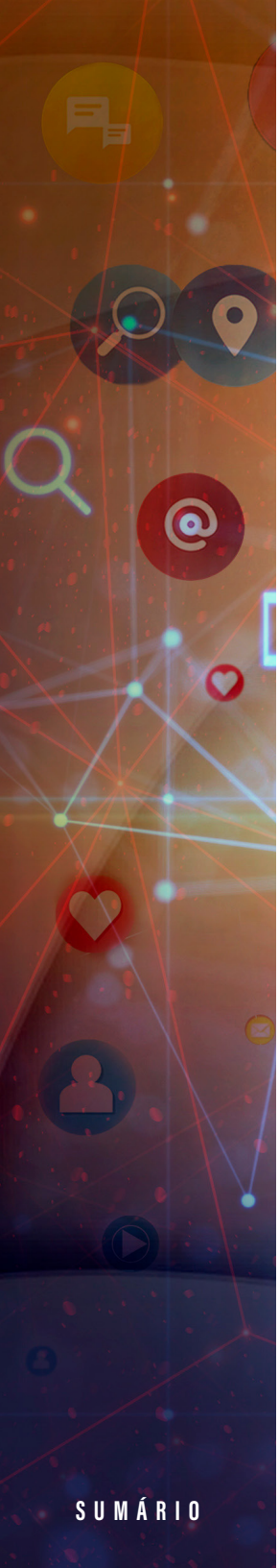


identificaram as necessidades que os usuários manifestam sobre o desejo de ter mais informações sobre a proveniência dos recursos, o que permite fornecer mais contexto sobre a forma como foram gerados. Refletindo o desejo de mais transparência sobre as decisões relativas à seleção de *websites* escolhidos para a conformação dos arquivos, o quão abrangente é o arquivo, bem como a completude das capturas individuais e as alterações e as mudanças que ocorrem ao longo do tempo.

Outro aspecto que deve ser ressaltado quanto à importância da utilização de padrões de metadados é o seu valor para a interoperabilidade, sejam metadados de descrição, gestão ou preservação de objetos digitais. A interoperabilidade somente é assegurada por práticas e por padrões de descrição que se traduzem nas sintaxes para codificação de dados. (FORMENTON; GRACIOSO, 2022).

Tradicionalmente, de acordo com Di Pretoro, Geeraert e Soyez (2019), as bibliotecas e os arquivos descrevem recursos de maneira diferente. Geralmente, nas bibliotecas é feita a catalogação no nível do título e os títulos são copiados literalmente e nos arquivos, por outro lado, trabalham com descrições multiníveis para as quais os títulos são geralmente criados. Isso dificulta a interoperabilidade dos metadados, mas, alguns países e instituições vêm envidando esforços para vincular metadados provenientes de diferentes coleções de arquivos da *web*. É o caso do *Nationaal Register Webarchieven*[158] da Holanda, onde 11 instituições contribuíram para um registro nacional de arquivamento da *web* disponibilizando seus metadados em uma plataforma compartilhada na qual o público pode usar filtros de pesquisa como instituição de arquivamento, nome, *site* arquivado, ferramenta de arquivamento, motivo e ano ou período de arquivamento.

158 https://www.registerwebarchieven.nl/



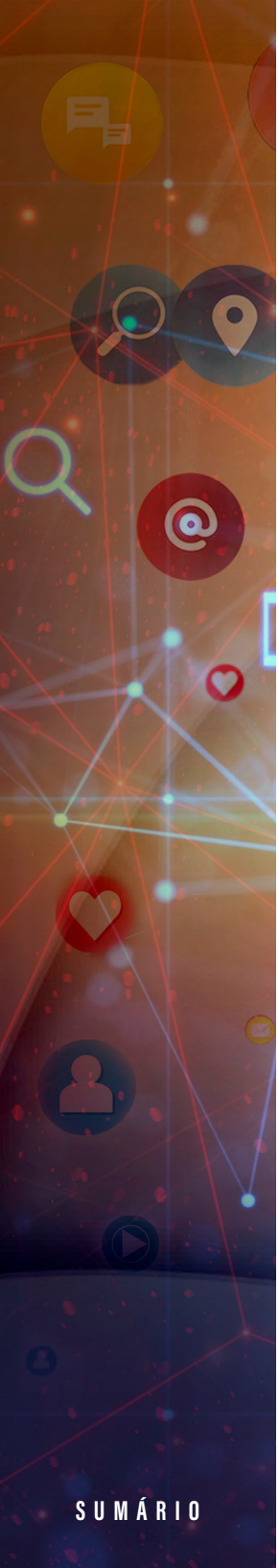
Também, o objetivo do projeto *Preserving On-line Multiple Information: towards a Belgian strategy* (PROMISE)[159] foi garantir a interoperabilidade entre os metadados descritivos provenientes dos *State Archives and the Royal Library for Belgian*[160] para o arquivo da *web* belga desde o início, em 2017. Para tal, adotaram as recomendações da *OCLC Research Library Partnership web Archiving Metadata Working Group* (DOOLEY e BOWERS, 2018). Uma vantagem apontada em relação ao conjunto de metadados da OCLC é que os mapeamentos para EAD e MARC21 foram incluídos.

Se encontramos na literatura estudos que relatam o esforço pela padronização dos metadados descritivos, o mesmo não se verifica sobre os metadados técnicos. Fato que pode ser explicado, pois, ao contrário dos metadados descritivos, que geralmente são gerados manualmente, os metadados técnicos são criados automaticamente no momento do rastreamento. Portanto, existe uma forte ligação entre os metadados técnicos e o formato de arquivo WARC como campos opcionais ou obrigatórios: formato do arquivo, data e hora da captura, tamanho do arquivo, URI de destino, ID do registro WARC, tipo WARC, *software* usado, política de *robots.txt* escolhida, cabeçalhos de resposta HTTP e endereço IP do servidor, podendo haver outros elementos de metadados técnicos adicionais que devem ser preservados para facilitar a pesquisa nos arquivos da *web*.

Para o registro de metadados é importante desenvolver práticas destinadas a uma descrição mais ampla dos recursos, o que, por sua vez, maximiza as possibilidades de acesso e uso. Segundo Blanco-Rivera (2021) os processos de arquivo também abrem portas para mais iniciativas de colaboração, tanto entre arquivistas e bibliotecários e pesquisadores, bem como com outras comunidades interessadas em arquivamento da *web*. Também, abre possibilidades de

159 https://www.ugent.be/mict/en/research/projects/2017/promise-preserving-online-multiple-information-towards-a-belgian-strategy

160 https://arch.arch.be/index.php?l=en



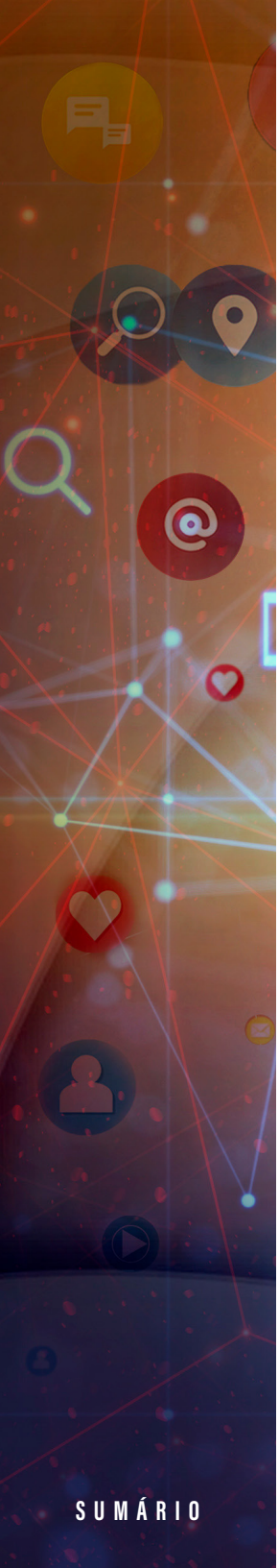
construir pontes entre a biblioteconomia e as práticas arquivísticas na *web*, entre a biblioteconomia e as práticas de arquivo no campo da descrição. Este tipo de colaboração pode ser concretizado a partir de um processo de catalogação colaborativa, o que atribui ainda mais valor ao desenvolvimento e utilização de diretrizes de metadados.

Metadados são atributos indispensáveis para garantir a descrição e identificação de um *website* arquivado. Quando projetados, de maneira cuidadosa, resultarão na boa gestão dos *websites* arquivados a curto e longo prazos e se forem criados, de forma completa e consistente, é possível usá-los de modo a atender às necessidades de recuperação dos usuários.

Os sistemas de descoberta dos arquivos da *web* devem permitir a inclusão de uma ampla gama de informações descritivas e contextuais necessárias para aprofundar o conteúdo dos arquivos da *web* e as ferramentas de análise/mineração de dados são o próximo passo na construção da descoberta para apoiar as necessidades do usuário. Completeza, consistência e qualidade dos metadados e organização dos *websites* arquivados, numa interface simples e amigável, são fatores determinantes para obtenção de bons resultados na descrição e melhoria da satisfação dos usuários.

ARMAZENAMENTO E ACESSO *ON-LINE*

Existem diversas formas de dar acesso aos arquivos da *web*, dependendo das necessidades e requisitos específicos da organização responsável pelos arquivos. Um método comum é disponibilizar uma plataforma de arquivamento da *web*, que é uma ferramenta ou serviço que permite aos usuários acessar e interagir com arquivos da *web*. Os arquivos da *web* também podem ser acessados usando ferramentas especializadas, como leitores de arquivos da *web* ou exten-



sões no navegador da *web*. Essas ferramentas podem ser usadas para visualizar e interagir com arquivos da *web* no formato WARC.

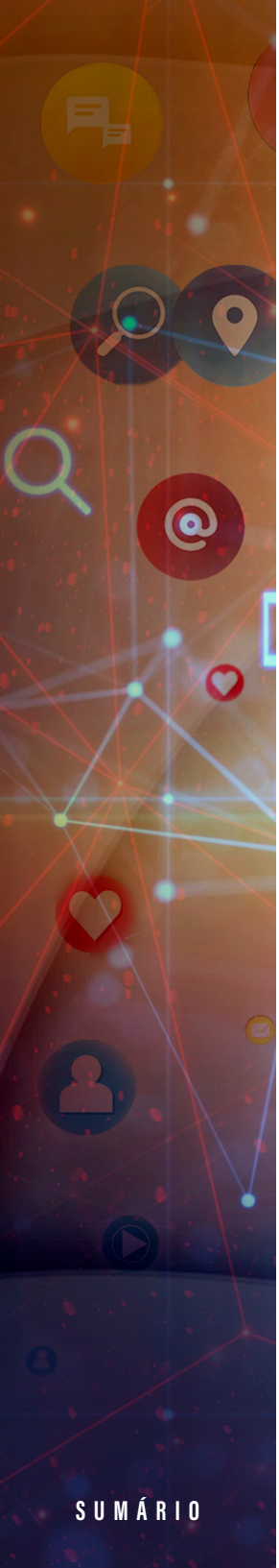
As plataformas de arquivamento da *web* geralmente oferecem vários recursos e ferramentas que facilitam o acesso e o uso de arquivos da *web*. Isso pode incluir recursos de pesquisa, ferramentas de navegação e outras soluções que permitem aos usuários encontrar rápida e facilmente o conteúdo que estão procurando.

Existem iniciativas, como o protocolo *Memento*[161], que têm como objetivo conectar múltiplos arquivos da *web*, com o propósito de oferecer recursos para buscar e acessar informações que foram preservadas em diferentes arquivos, a partir de uma única plataforma. *Memento* é um protocolo que adiciona uma dimensão de tempo ao HTTP, permitindo que os usuários acessem uma versão de um recurso da *web* como existia em uma data específica no passado. Isso é obtido especificando a data desejada em um *plug-in* do navegador e, se existirem versões antigas e forem hospedadas por servidores que suportam o protocolo *Memento*, os usuários podem acessá-las. O protocolo é especificado no RFC 7089[162] e está alinhado com os recursos da *web* de controle de versão.

Como mencionado por Maemura e colaboradores (2018), os rastreamentos da *web* produzem resultados que são armazenados em um arquivo *WebARChive* (WARC), que é um formato padrão ISO criado pelo *International Internet Preservation Consortium* (IIPC). Esses arquivos WARC podem ser usados por exemplo com o *software Open Wayback*, que se constitui em um projeto de código aberto apoiado por membros do IIPC e permite aos usuários visualizar páginas da *web* arquivadas em seu navegador. Este é semelhante ao serviço comumente conhecido como *Wayback Machine* oferecido pelo *Internet Archive*. Embora o *software* de rastreamento da *web Heritrix* seja

161 www.mementoweb.org

162 https://www.rfc-editor.org/rfc/rfc7089



gratuito e de código aberto, ele requer infraestrutura de computação e experiência que podem não estar prontamente disponíveis. Como alternativa, algumas organizações usam provedores de "arquivamento como serviço" para terceirizar parte dessa infraestrutura.

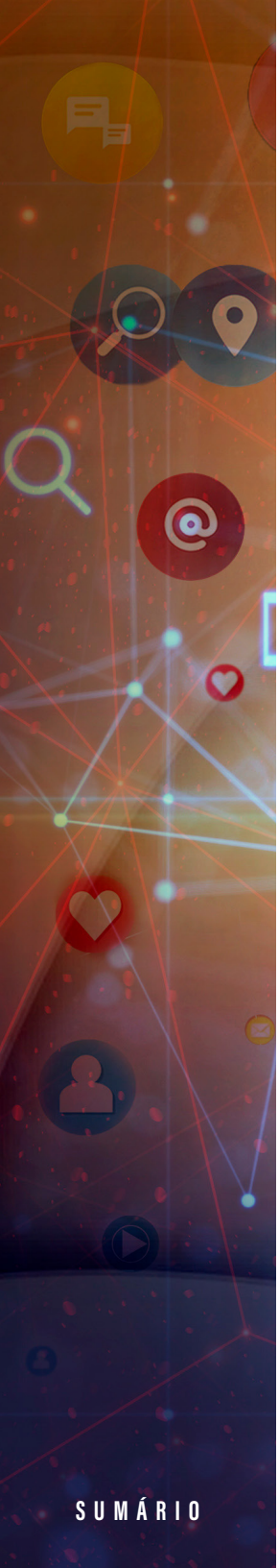
Serviços como o *Conifer*[163], inicialmente chamado de *webrecorder*, e parte do projeto *Rhizome*, permitem o arquivamento de *website*s em uma pequena escala de forma *on-line*. Segundo Santos (2021) e Santos, Formenton e Terrada (2022), a partir de testes e usos de tais ferramentas, projetos mais complexos podem surgir, impulsionando o uso frequente do arquivamento da *web* de instituições.

Dar acesso a arquivos da *web* também envolve a criação de uma interface baseada na *web* que permite aos usuários pesquisar, navegar e acessar o conteúdo arquivado, o que pode ser feito usando *software* de arquivamento da *web* ou sistemas de preservação digital, que normalmente incluem ferramentas integradas para criar uma interface baseada na *web* para o conteúdo arquivado.

A interface baseada na *web* pode ser personalizada para atender às necessidades específicas da organização e de seus usuários, incluindo vários recursos e funções, como a capacidade de pesquisar palavras-chave ou frases específicas, navegar pelos arquivos por data ou outros metadados e acessar o conteúdo arquivado em vários formatos.

Além de criar uma interface baseada na *web*, as organizações também podem considerar outras formas de fornecer acesso a arquivos da *web*, incluindo o acesso por meio de ferramentas ou portais de pesquisa especializados ou trabalhar com arquivos, bibliotecas e outras instituições para tornar os arquivos disponíveis para um público mais amplo.

163 https://conifer.rhizome.org/



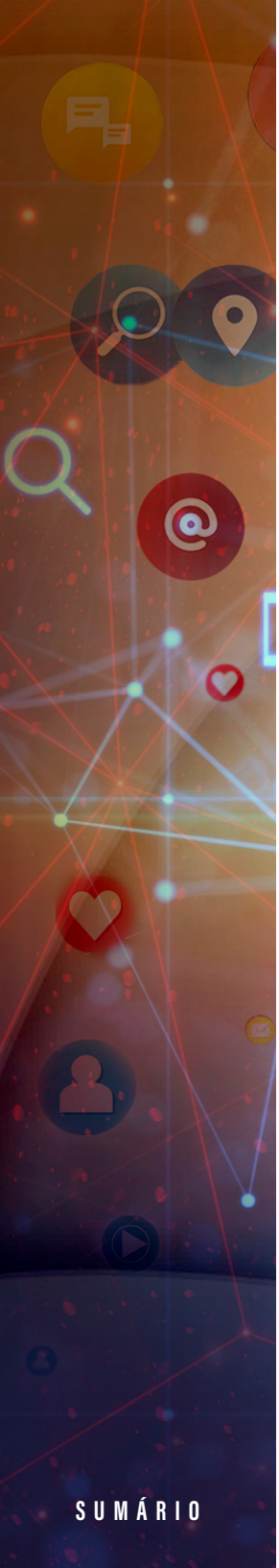
Uma vez armazenados os arquivos da *web*, o próximo passo é fornecer acesso ao conteúdo para que possa ser utilizado por pesquisadores, educadores e demais interessados. Existem várias maneiras de fornecer acesso a arquivos da *web*, dependendo do tipo e formato do conteúdo, bem como das necessidades e recursos da organização responsável pelo acesso.

Algumas plataformas podem fornecer acesso a arquivos da *web* por meio de uma variedade de interfaces, como pesquisa baseada na *web* e interfaces de navegação ou APIs[164] que permitem que os pesquisadores acessem programaticamente o conteúdo dos arquivos. Além das plataformas de arquivamento da *web* e sistemas de bibliotecas digitais, os arquivos da *web* também podem ser acessados por meio de serviços especializados de arquivamento da *web*. Esses serviços geralmente são fornecidos por organizações com experiência em arquivamento da *web* e podem ser usados para hospedar e fornecer arquivos da *web* aos usuários. Para buscar arquivos da *web* hospedados por uma organização ou instituição específica, também pode tentar pesquisar diretamente a plataforma de arquivamento da *web* ou o sistema de biblioteca digital da organização, isso pode ser feito pesquisando o nome da organização, junto com palavras-chave como "arquivo da *web*" ou "*web archive*".

O acesso a arquivos da *web* geralmente envolve o uso de uma plataforma de arquivamento da *web*, sistema de biblioteca digital ou serviço especializado de arquivamento da *web* para pesquisar e acessar o conteúdo de seu interesse. Essas plataformas e serviços são geralmente acessados por meio de uma interface baseada na *web*, permitindo que você possa pesquisar e acessar facilmente arquivos da *web* usando seu navegador da *web*.

Se o arquivo da *web* estiver disponível por meio de uma interface baseada na *web*, normalmente podemos acessar o arquivo

164 https://archive.org/help/wayback_api.php



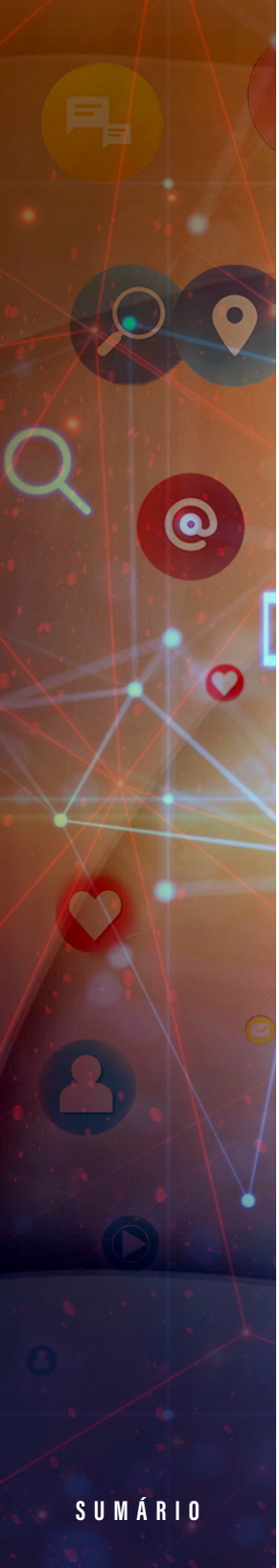
visitando o *site* associado ao arquivo da *web* e usando as ferramentas de pesquisa e navegação fornecidas no *site*. Se o arquivo da *web* estiver disponível por meio de uma API ou outra ferramenta especializada, talvez seja necessário usar um *software* específico ou habilidades de programação para acessar o conteúdo do arquivo.

Para encontrar arquivos da *web*, é possível usar uma variedade de métodos diferentes, dependendo do tipo e formato dos arquivos da *web* que estamos procurando, bem como dos recursos e ferramentas que estão disponíveis. Alguns métodos comuns para encontrar arquivos da *web* incluem o seguinte:

a. Plataformas de arquivamento da *web*: Algumas organizações que preservam arquivos da *web* disponibilizam suas coleções por meio de plataformas de arquivamento da *web*. Essas plataformas geralmente fornecem ferramentas de pesquisa e navegação que permitem aos usuários encontrar e acessar arquivos da *web*. Por exemplo, o *Internet Archive* mantém uma grande coleção de arquivos da *web* e fornece acesso a esses arquivos por meio de seu serviço *Wayback Machine*. Você pode usar o *Wayback Machine* para pesquisar arquivos da *web* por URL, data ou outros critérios.

b. Serviços de arquivamento da *web*: algumas organizações oferecem serviços de arquivamento da *web*, que podem ser usados para localizar e acessar arquivos da *web*. Esses serviços podem fornecer ferramentas de pesquisa e navegação ou podem oferecer outros métodos para encontrar arquivos da *web*, alguns acessos podem ser restritos dependendo da plataforma.

c. Catálogos de bibliotecas, centros de documentação e arquivos: muitas organizações que preservam arquivos da *web* disponibilizam suas coleções por meio de catálogos *on-line*. Podemos pesquisar esses catálogos para encontrar arquivos da *web* mantidos pelo arquivo, biblioteca ou organização.

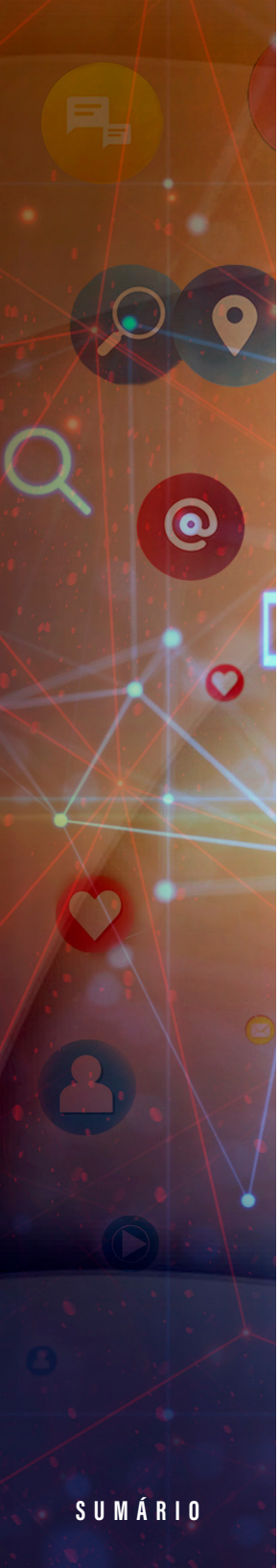


Existem várias maneiras de armazenar arquivos da *web*, e o método usado dependerá das necessidades e requisitos específicos do arquivo. Os arquivos da *web* podem ser armazenados em mídia digital, como discos rígidos ou dispositivos de armazenamento conectados à rede, com o uso de servidores próprios ou compartilhados, também podem ser armazenados em plataformas baseadas em nuvem, como *Amazon web Services* ou *Microsoft Azure*, permitindo que os arquivos sejam acessados de qualquer lugar e fornecendo um alto nível de escalabilidade e flexibilidade. Independentemente do método específico usado, é importante garantir que os arquivos da *web* sejam armazenados de maneira segura e confiável, ajudando a garantir que as informações contidas nos arquivos permaneçam acessíveis e utilizáveis ao longo do tempo.

Os arquivos da *web* são formatos de arquivos *contêiner* usados para preservar o conteúdo da *web*. Existem dois formatos principais: *ARchive Container* (ARC) e o *web ARchive Container* (WARC). O formato ARC foi usado pela primeira vez em 1996 e é utilizado especificamente pelo *Internet Archive*. O formato WARC tornou-se um padrão ISO em 2009 e é o formato predominante empregado pela comunidade internacional de arquivamento da *web*.

Os arquivos ARC e WARC são compostos de diferentes tipos de registros representando diversos aspectos do conteúdo da *web* arquivado, como recursos recuperados da *web*, características do rastreador, respostas e solicitações HTTP. Os metadados técnicos incluídos nesses registros fornecem proveniência e permitem a navegação temporal, bem como oportunidades para extração e análise de metadados.

Recursos arquivados de diferentes *site*s ou partes do mesmo *site* podem ser colocados aleatoriamente em um ou mais arquivos WARC, facilitando o armazenamento. No entanto, isso também explica os tempos de carregamento relativamente lentos dos *site*s no



Wayback Machine, pois cada objeto que compõe uma página pode vir a ser descompactado de arquivos diferentes.

ARC é um formato de arquivo usado para armazenar conteúdo da *web* em arquivos da *web* e a plataforma *Internet Archive* utilizou como padrão no seu rastreador *Heritrix* como um formato simples e compacto para armazenar páginas da *web* e outros objetos digitais, como imagens e vídeos[165]. O formato de arquivo ARC armazena dados em blocos de tamanho fixo e cada bloco contém uma ou mais solicitações e respostas HTTP, juntamente com metadados associados, como carimbo de data/hora, tipo de conteúdo e URL do recurso solicitado, facilitando a identificação e recuperação de recursos específicos do arquivo.

O formato de arquivo ARC, surgido em 1996[166], foi amplamente substituído pelo formato de arquivo WARC (*web ARChive*), que é mais flexível e oferece suporte a recursos adicionais, como segmentação de registro, compactação de conteúdo e metadados. Esse formato, que completou 10 anos em 2019 (Figura 32), é projetado para armazenar conteúdo da *web* de maneira acessível e fácil de preservar, permitindo que o conteúdo seja facilmente acessado e usado por pesquisadores e outros usuários.

165 https://archive.org/web/researcher/ArcFileFormat.php

166 https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml

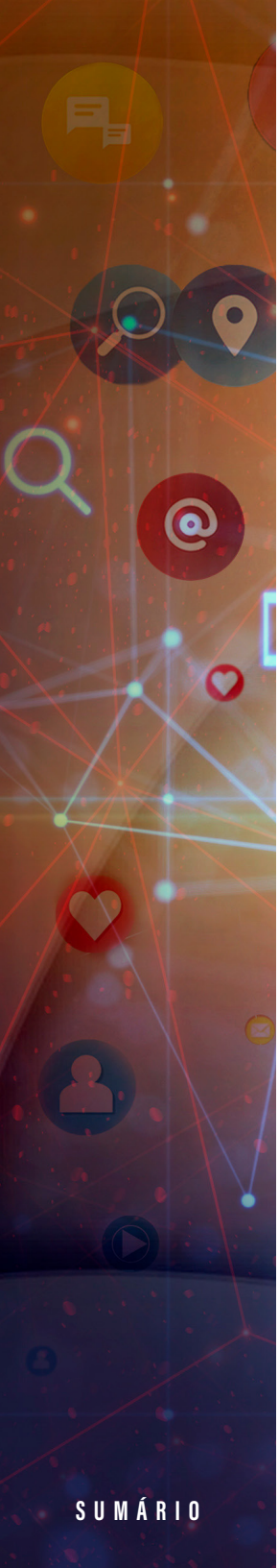
Figura 32 - Logo comemorativo dos 10 anos

Fonte: https://netpreserveblog.wordpress.com/2019/05/29/warc-10th-anniversary/

Outro formato mais recente de armazenamento para arquivamento da *web* é o *Web ARCHive Collection Zipped* (WACZ)[167]. É um formato de arquivo compactado para armazenar arquivos da *web* e derivado do formato WARC, o qual é amplamente utilizado para armazenar arquivos da *web*.

O objetivo da especificação WACZ é fornecer um formato para arquivos da *web* que podem ser facilmente compartilhados e usados para fins sociais (interoperável e contextual) e técnicos (carregando dinamicamente pequenas quantidades de dados, sem exigir o *download* de todo o arquivo), incluindo o conteúdo da *web* arquivado e informações contextuais sobre os *site*s capturados, tais como quando e como ela foi criada. Este formato foi projetado para ser facilmente hospedado e permitir acesso eficiente a dados WARC empacotados, o que permite ao navegador renderizar uma página buscando apenas o que é necessário para aquela página específica. O WACZ não pretende substituir outros formatos de arquivamento da *web*, mas sim estabelecer uma convenção de empacotamento de arquivos para

167 https://specs.webrecorder.net/wacz/latest/



todos os dados necessários a um navegador para a renderização eficiente de uma coleção de arquivos da *web* e sua contextualização.

Além do *software* de arquivamento da *web*, os arquivos da *web* também podem ser armazenados usando sistemas de preservação digital. Esses sistemas são projetados para gerenciar e preservar conteúdo digital a longo prazo, incluindo arquivos da *web*.

Existem várias maneiras de armazenar arquivos da *web*, incluindo o uso de *software* de arquivamento da *web* e sistemas de preservação digital. O método específico usado dependerá do tipo e formato do conteúdo da *web* que está sendo preservado, bem como das necessidades e recursos da organização responsável pela preservação do conteúdo.

A localização e acesso às páginas em um arquivo da *web* pode acontecer diretamente pela URL arquivada, pela pesquisa a partir da URL original ou pesquisa por texto, da mesma forma como pesquisamos no *Google*. Como resultados da pesquisa, temos acesso aos *site*s arquivados ou ainda imagens, como é o caso do Arquivo.pt[168]

Outra estrutura importante é a chamada *timestamp* (ou carimbo do tempo). Normalmente nos arquivos da *web*, uma *timestamp* refere-se a um ponto específico no tempo em que um *site* foi rastreado e arquivado. Quando um usuário visita uma página arquivada, ele pode ver uma exibição de calendário mostrando quando o *site* foi rastreado e arquivado em datas diferentes, bem como uma exibição de linha do tempo mostrando como o *site* mudou ao longo do tempo, caso estes recursos sejam implementados.

Cada captura do *site* é carimbada com a data e hora em que foi arquivado, o que permite aos usuários navegar em diferentes versões do *site* conforme elas apareceram em diferentes pontos no tempo.

168 https://arquivo.pt/image/search?

O *timestamp* é normalmente exibido no URL da página arquivada, seguindo um determinado formato. No caso da plataforma *Wayback Machine*, do *Internet Archive*, o formato é o seguinte: http://*web*.archive.org/*web*/[timestamp]/[URL original] (Figura 33).

Figura 33 – Exemplo de formato de URL contendo *timestamp*

Fonte: autores, baseado na plataforma Wayback Machine.

Por exemplo, se um *site* foi arquivado em 19 de setembro de 2013, às 04h 46min e 12 segundos, o carimbo de data/hora será "20130919044612" no URL. Ao especificar diferentes carimbos de data/hora no URL, os usuários podem visualizar diferentes versões da página arquivada e ver como ela mudou ao longo do tempo.

PLATAFORMAS E REDES SOCIAIS

O arquivamento de redes sociais na *web* refere-se à prática de preservar e fornecer acesso ao conteúdo de rede social (também conhecido como mídia social), como *tweets*, postagens e outros materiais publicados em plataformas da web 2.0. O arquivamento da *web* nas redes sociais é um aspecto importante do arquivamento da *web*, pois permite a preservação e o acesso a uma ampla gama de materiais publicados nestes meios, que não se encontram em outras plataformas.

As redes sociais representam um elemento fundamental da *web* atual, caracterizada pelo crescente engajamento e interação dos usuários *on-line*, que passou de *site*s estáticos para dinâmicos, onde os usuários não só consomem, organizam e comunicam conteúdo,

mas criam e compartilham (VLASSENROOT *et al.*, 2021), tornando uma experiência altamente personalizável e interativa.

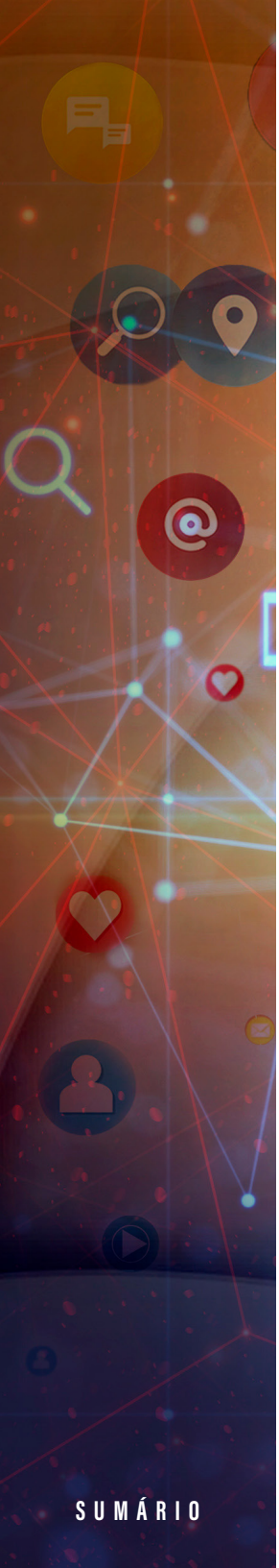
O arquivamento de redes sociais na *web* geralmente envolve a captura e o armazenamento de cópias o mais completas possíveis do material disponibilizado, incluindo HTML, CSS e outros arquivos que compõem o conteúdo, permitindo a preservação da formatação e aparência originais, bem como quaisquer metadados, como data e hora da publicação, autor do conteúdo e qualquer outra informação associada ao conteúdo.

Depois que o conteúdo da rede social é capturado e armazenado, ele pode ser disponibilizado aos usuários por vários meios, como pesquisa baseada na *web* e interfaces de navegação, APIs ou outras ferramentas, de forma que os usuários acessem e usem o conteúdo de rede social preservado, seja para fins de pesquisa ou para outros fins. A Figura 34 mostra o exemplo de uma página arquivada da Rede Social *Orkut.*

Figura 34 – Página arquivada do *Orkut*

Fonte: Rede Social Orkut em 01 de setembro de 2004[169].

169 https://web.archive.org/web/20040901081541/http://www.orkut.com/



Algumas comunidades do *Orkut* foram preservadas e ainda podem ser acessadas *on-line* por meio do *Wayback Machine*[170]. Vlassenroot e colaboradores (2021) fizeram um levantamento com iniciativas de arquivamento da *web* que trabalham com preservação de mídias sociais e quais plataformas preservam (Quadro 9).

Quadro 9 - Visão geral de mídias sociais arquivadas por arquivos da *web*

País	Sigla da iniciativa	*Facebook*	*Twitter (X.com)*	*You Tube*	*Instagram*	*Flickr*
Canadá	LAC[171]	Sim	Sim	Sim	Sim	Sim
Canadá	BAnQ[172]	Sim	Sim	Não	Não	Não
Dinamarca	Netarkivet[173]	Sim	Sim	Sim	Sim	Não
Estonia	Eesti Veebiarhiiv[174]	Uma pagina	Não	Não	Não	Não
França	BnF[175]	Sim	Sim	Sim	Sim	Não
França[176]	INA[177]	Não	Sim	Sim	Não	Não
Hungria	NSL[178]	Não	Não	Não	Sim	Não
Irlanda	NLI[179]	Não	Sim	Sim	Não	Não
Luxemburgo	BnL[180]	Sim	Sim	Sim	Não	Não

170 https://web.archive.org/web/20141001005309/http://orkut.google.com/

171 https://library-archives.canada.ca/eng

172 http://bndigital.bn.gov.br/bnds/canada-biblioteca-nacional-do-quebec-banq/

173 https://www.kb.dk/find-materiale/samlinger/netarkivet

174 https://veebiarhiiv.digar.ee/

175 https://www.bnf.fr/fr

176 Também: *Dailymotion*, Vimeo SoundCloud

177 https://www.ina.fr/

178 https://www.civic-epistemologies.eu/partners/national-szechenyi-library/

179 https://www.nli.ie/

180 https://pt.wikipedia.org/wiki/Biblioteca_Nacional_do_Luxemburgo

Nova Zelandia	NLNZ[181]	Sim	Sim	Não	Sim	Não
Suíça	Webarchiv Schwiez[182]	Não	Não	Não	Não	Não
Holanda[183]	KB[184]	Não	Não	Não	Não	Não
Holanda	NA[185]	Não	Não	Não	Não	Não
UK	UKWA[186]	Sim	Sim	Não	Não	Não
USA	GWUL[187]	Não	Sim	Não	Não	Não

Fonte: Vlasseroot et al. (2021, tradução nossa).

O arquivamento de redes sociais é um aspecto importante do arquivamento da *web*, pois permite a preservação a uma ampla variedade de conteúdos, isso pode ajudar a garantir que os materiais sejam acessíveis para as próximas gerações. Existem algumas questões envolvidas no arquivamento de redes sociais que precisam de atenção, incluindo:

1. Privacidade: o conteúdo da rede social podem conter informações pessoais, fornecer acesso a essas informações sem o consentimento dos indivíduos envolvidos pode ser uma violação de seus direitos de privacidade.

2. Propriedade intelectual: o conteúdo da rede social pode conter materiais protegidos por direitos autorais, como imagens, vídeos

181 https://natlib.govt.nz/

182 https://www.nb.admin.ch/snl/de/home/fachinformationen/e-helvetica/webarchiv-schweiz.html

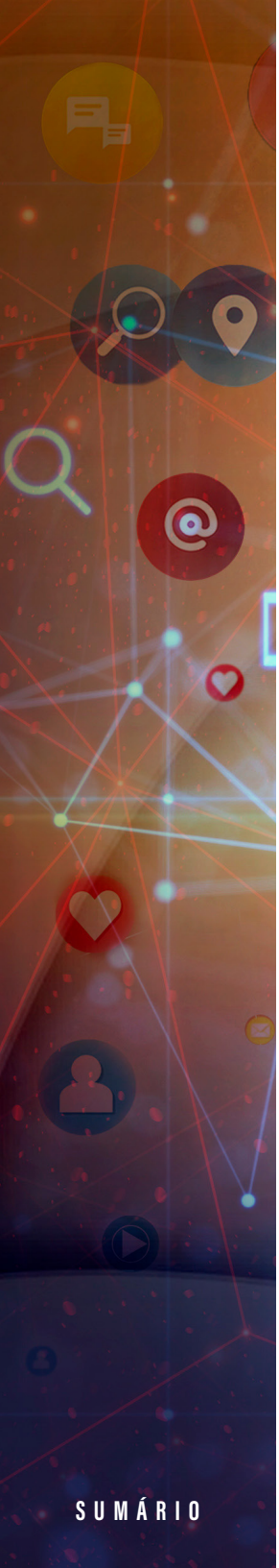
183 Também: *WhatsApp*

184 https://www.kb.nl/en

185 https://www.nationaalarchief.nl/en/about-na

186 https://www.webarchive.org.uk/ukwa/

187 https://library.gwu.edu/



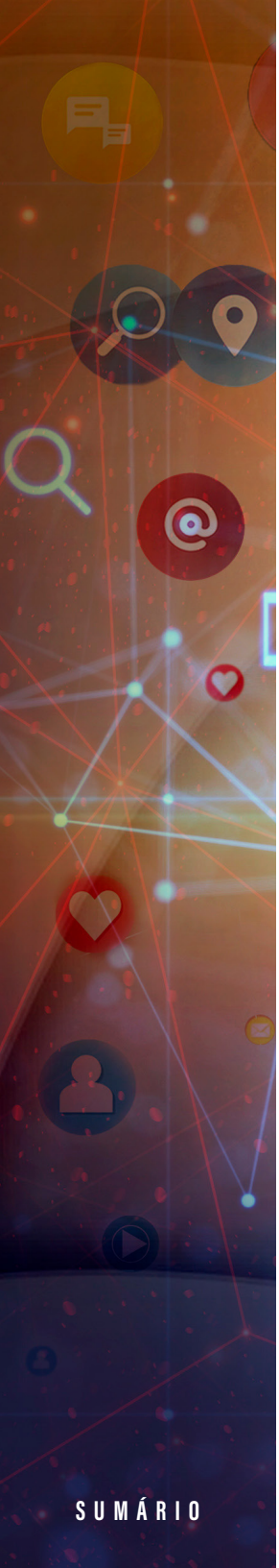
e texto, e fornecer acesso a esses materiais sem a permissão dos detentores dos direitos autorais pode incorrer em violação de seus direitos de propriedade intelectual em certos casos.

3. Natureza dinâmica da rede social: o conteúdo da rede social geralmente é dinâmico e mutável, com usuários adicionando, modificando ou excluindo conteúdo regularmente, isso dificulta a preservação e o acesso de uma forma que reflita o estado original do conteúdo.

4. Acesso limitado a conteúdo da rede social: algumas plataformas de rede social podem limitar o acesso ao seu conteúdo, seja por meio de medidas técnicas ou por meio de acordos de termos de serviço, podendo tornar difícil para os arquivistas da *web* capturar e preservar estes materiais.

5. Viés e perspectiva: o conteúdo da rede social pode refletir os preconceitos e perspectivas dos indivíduos ou organizações que o criaram, levando a uma representação incompleta ou distorcida dos materiais que estão sendo preservados.

De acordo com Hockx-Yu (2014), quando se trata de arquivar conteúdo de rede social, as instituições de memória geralmente usam abordagens seletivas com ferramentas de arquivamento da *web* existentes, capturando apenas um pequeno número de páginas para coleções de tópicos específicos ou indivíduos/organizações. Este método tem limitações e requer a personalização do processo de rastreamento e embora existam alguns arquivos públicos do *Twitter*, *Facebook* e *YouTube*, os arquivos da *web* nacionais geralmente têm acesso restrito ao conteúdo da rede social. As páginas de redes sociais em arquivos da *web* geralmente estão incompletas devido às limitações do rastreador com conteúdo dinâmico. Recuperar conteúdo de redes sociais em arquivos da *web* pode ser um desafio sem as funções de pesquisa adequadas.



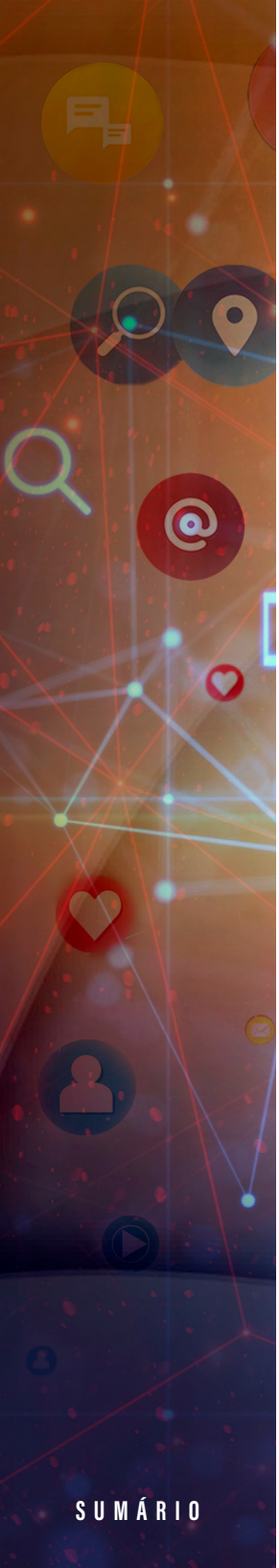
Outro desafio é que os dados da rede social geralmente são gerados pelo usuário, o que significa que podem não ser confiáveis ou ser tendenciosos. O conteúdo nas redes sociais nem sempre é factualmente preciso, e os usuários podem ter diferentes motivações para compartilhar informações, o que pode dificultar a interpretação dos dados.

Os dados nas redes sociais estão sujeitos a alterações e atualizações frequentes, o que pode dificultar a manutenção de um conjunto de dados consistente e preciso ao longo do tempo. Os dados podem ser excluídos ou alterados pelos usuários, e as próprias plataformas podem alterar suas políticas ou algoritmos, o que pode afetar os dados que estão disponíveis.

Os desafios de recuperar dados das mídias sociais destacam a importância de um planejamento cuidadoso e práticas robustas de gerenciamento de dados para garantir a precisão e a confiabilidade das informações coletadas. Segundo Ferreira e Rockembach (2019), algumas dificuldades podem surgir especialmente na captura dos conteúdos das redes sociais.

Um dos principais desafios é o acesso restrito aos dados, que podem exigir permissões especiais ou autenticação para acessar. Isso pode dificultar a coleta e a análise dos dados, especialmente se as permissões ou os requisitos de autenticação forem complexos ou mudarem constantemente. Outro fator é o grande volume de dados gerados nas plataformas de rede social, que pode dificultar a identificação e coleta de informações relevantes.

Isto de certa forma está relacionado ao fenômeno da plataformização (VAN DJICK; POELL; DE WAAL, 2018). Um dos desafios que isto impõe é que as plataformas privadas podem ter seus próprios formatos ou estruturas de dados proprietárias, o que dificulta a integração dos dados com outros sistemas ou ferramentas, e que também podem dificultar o acesso e a análise. Além disso,



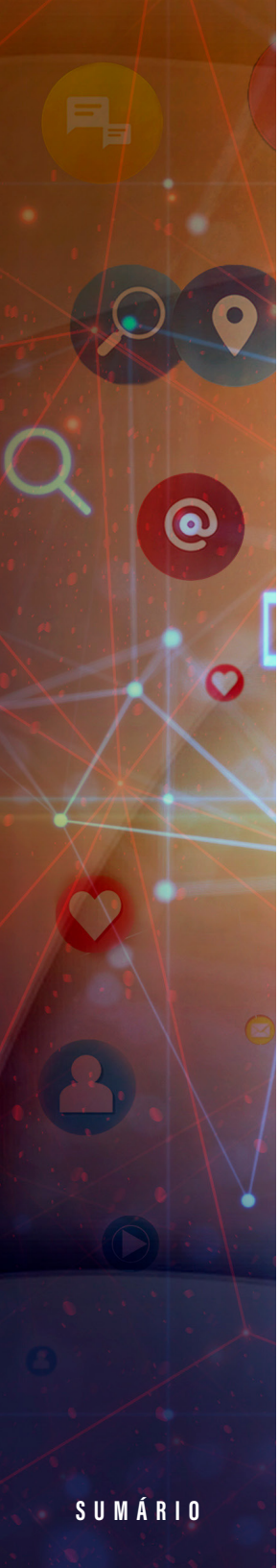
as plataformas privadas muitas vezes possuem suas próprias políticas e práticas exclusivas de gerenciamento de dados, o que pode influenciar no entendimento e no cumprimento das regras de acesso e uso dos dados.

As plataformas de rede social costumam ser chamadas de "arquiváveis por *design*" (BEN-DAVID, 2019), o que significa que não são projetadas para preservar o conteúdo para acesso e uso de longo prazo. Isso contrasta com as práticas arquivísticas tradicionais, que normalmente envolvem a preservação e organização cuidadosa de registros e documentos para referência futura.

Uma das principais razões pelas quais as plataformas de rede social são inarquiváveis por *design* é que elas se concentram em gerar e compartilhar conteúdo em tempo real, em vez de preservá-lo para o futuro. As plataformas de rede social são projetadas para serem dinâmicas e interativas, com os usuários constantemente postando e comentando o conteúdo, o que pode dificultar a captura e o armazenamento das informações em um formato estável e acessível.

Além disso, as plataformas de rede social geralmente possuem políticas e algoritmos que excluem ou ocultam automaticamente o conteúdo mais antigo, a fim de priorizar informações mais recentes e relevantes. Esta medida pode dificultar o acesso e a preservação de conteúdos mais antigos, mesmo que tenham sido amplamente compartilhados e discutidos quando foram postados.

A natureza não arquivística das plataformas de rede social destaca a necessidade de planejamento e gerenciamento cuidadosos para preservar as informações e o conhecimento gerados nessas plataformas para referência futura, envolvendo o uso de ferramentas e técnicas especializadas para capturar e armazenar conteúdo, bem como desenvolver estratégias para organizar e acessar o conteúdo arquivado ao longo do tempo.

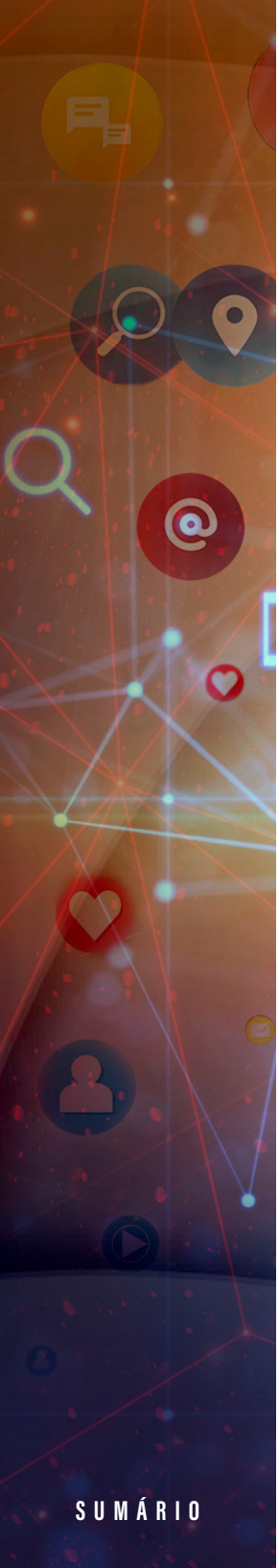


Ben-David (2019) propõe o pensamento arquivístico como uma estrutura analítica para estudar plataformas como o *Facebook*, relacionando colonialismo de dados e o papel que o *Facebook* assume como um novo *archeion* de documentos públicos, embora a plataforma seja inarquivável por *design*, ou seja, não é projetado para ser arquivado.

O mesmo conceito de inarquivável por *design* pode ser aplicado a muitas redes sociais, onde é produzido o documento, mas não é projetado para arquivamento, preservação e diversidade de métodos de recuperação da informação. Muitas vezes a recuperação é restringida por filtros e algoritmos pré determinados pela plataforma, não um acesso completo e irrestrito a todos os dados.

Para combater esta ideia, Ben-David (2019) apresenta o contra-arquivo - uma prática desenvolvida para resistir à hegemonia epistêmica dos arquivos coloniais - como um método que permite o estudo crítico da plataforma de rede social depois que ela bloqueou o acesso do pesquisador aos dados públicos por meio de sua Interface de Programação de Aplicativo ou API (*Application Programming Interface*). A autora conclui discutindo a mudança de fronteiras entre o arquivista, o ativista e o acadêmico, como imperativo dos métodos de pesquisa após o fenômeno da dataficação. Outro argumento trazido por Ben-David (2019) afirma que, após o fenômeno da dataficação, o acesso a informações que até recentemente eram de domínio público e digno de preservação por instituições de memória, agora depende muito da benevolência das plataformas e redes sociais.

Atualmente, o conteúdo público de rede social é arquivado por uma combinação de pesquisadores, empresas, governos e indivíduos que usam uma ampla variedade de ferramentas. O *X.com* (antigo *Twitter*), em particular, implementou uma gama de serviços para garantir a longevidade e a disponibilidade de dados históricos, fornecendo acesso a *tweets* antigos, como o primeiro *tweet* publicado



em março de 2006[188], no entanto, arquivar as mídias sociais não deve ser a única missão das instituições de memória, mas sim um esforço colaborativo envolvendo arquivistas, bibliotecários, tecnólogos, pesquisadores, editores, representantes do governo e empresas.

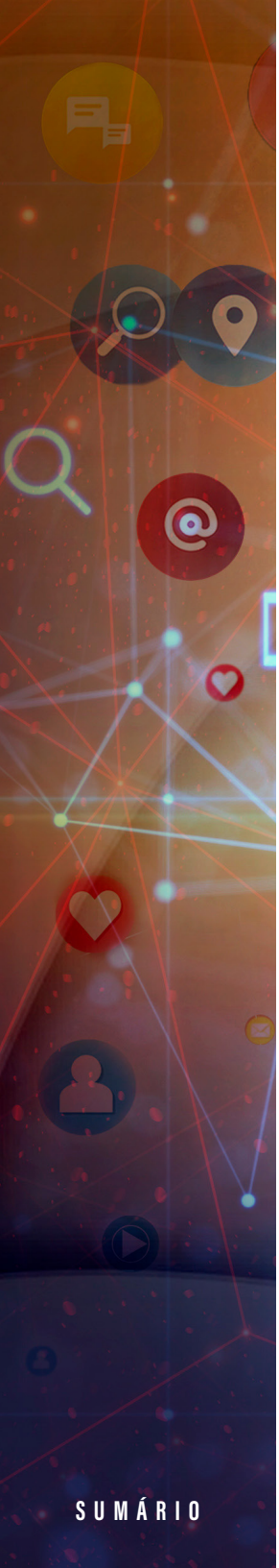
Umas das primeiras ações a tomar é decidir os dados a serem arquivados, definindo os perfis, palavras-chave ou *hashtags*. Pehlivan, Thièvre e Drugeon (2021) trazem algumas possibilidades, principalmente focadas na preservação do *X.com* (antigo *Twitter*):

a. Rastreadores: o uso de rastreadores da *web* (*crawler*) pode ser usado para páginas da rede social, realizando o *download* em massa, mas com certas limitações, pois apenas os dados exibidos podem ser extraídos, enquanto outros tipos de metadados podem não ser capturados, além disso, pode não ser possível navegar em todos os *tweets* exibidos de forma dinâmica. Sem os mecanismos de segurança adequados, os rastreadores também podem realizar acidentalmente um ataque de negação de serviço.

b. Raspagem da *web*: esta abordagem consiste na raspagem da *web* (*web scraping*) envolvendo o uso de programas automatizados para extrair informações específicas de páginas da *web*, no entanto, os dados extraídos da *web* podem vir a incluir apenas as informações disponíveis para o navegador, que podem diferir dos dados obtidos diretamente das APIs oficiais. As plataformas de rede social geralmente possuem políticas contra o uso da raspagem da *web*, mas toleram rastreadores "educados", ou seja, que não realizam um processo intensivo e de sobrecarga sobre a raspagem de dados das plataformas.

c. O uso de APIs: o *X.com* fornece APIs que permitem que desenvolvedores terceirizados acessem determinados recursos ou

188 https://twitter.com/jack/status/20



dados no *site*, contudo, as APIs também possuem limitações, como os dados que disponibilizam, além de estarem sujeitas a alterações e restrições frequentes. Pehlivan, Thièvre e Drugeon (2021) mencionam diferentes aplicativos e bibliotecas como *Twarc*[189], *Social Feed Manager*[190], *TAGS*[191] e *Digital Methods Initiatives*[192], que foram desenvolvidos para capturar *tweets* por meio de APIs. Para garantir a completude de um *tweet*, todos os objetos como imagens, vídeos ou URLs devem ser arquivados e disponibilizados.

Helmond e Van der Vlist (2019) sugerem que, ao enxergar as redes sociais a partir dos seus diversos atores, que incluem pesquisadores, desenvolvedores e usuários finais, além do ecossistema, interface, arquitetura e diversos níveis que compõem uma plataforma, torna-se possível aos pesquisadores reconstruir histórias de plataformas como redes sociais, artefatos técnicos e organizações empresariais. Eles ainda argumentam que o arquivamento de redes sociais é frequentemente entendido como uma extensão do arquivamento tradicional da *web*, mas apresenta desafios específicos devido à natureza dinâmica e personalizada do conteúdo da rede social e à natureza distribuída das plataformas.

189 https://github.com/DocNow/twarc

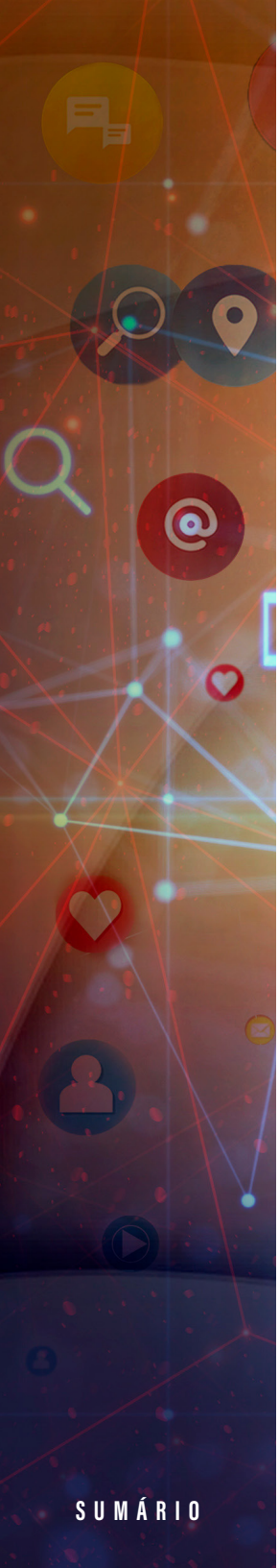
190 https://gwu-libraries.github.io/sfm-ui/

191 https://tags.hawksey.info/

192 https://github.com/digitalmethodsinitiative/dmi-tcat

4

PERSPECTIVAS DE ESTUDOS E APLICAÇÕES PROFISSIONAIS DO ARQUIVAMENTO DA *WEB*



Levando em consideração a bibliografia, as pesquisas e as práticas da área, compreendemos que algumas das perspectivas de estudo, aplicações práticas e profissionais passam pelos diversos capítulos abordados neste livro. Estes campos de atuação podem ir dos fluxos de trabalho, os usos da *web* como objeto de pesquisa e os arquivos da *web* como fonte documental, às reflexões ético-legais e de disseminação de conhecimento. Alguns dos tópicos podem incluir:

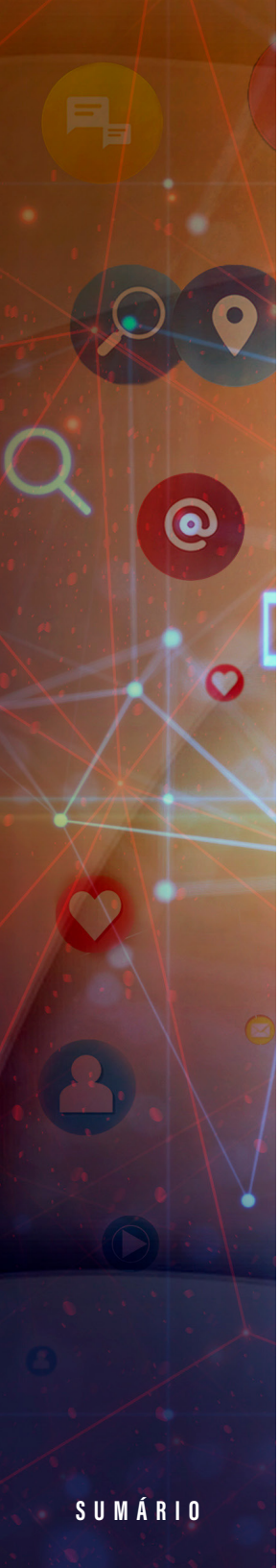
1. Elaboração de políticas de seleção e curadoria digital que incluam conteúdos que ainda não foram arquivados de forma sistemática e que podem tornar-se tópicos de pesquisa interessantes, além de formar novas fontes para usuários e pesquisadores.

2. Estudo das características de arquivabilidade dos *site*s a serem coletados, bem como dos metadados atribuídos, fornecendo autenticidade e proveniência.

3. Estudos quantitativos e qualitativos sobre os arquivos da *web*, incluindo análise de *Big Data*, abordagem estatística, análise de conteúdo, análise do discurso, análise de redes, netnografia, entre outras perspectivas.

4. Análise de contextos específicos do arquivamento da *web*, formação da memória digital como, por exemplo, em eventos como eleições, grandes acontecimentos sociais, tendências de caráter efêmero, temáticos, institucionais, por localização geográfica ou por domínio, bem como a observação da atuação das iniciativas de arquivamento da *web* em diversos países e composição dos acervos digitais.

5. Pesquisa aplicada sobre tecnologias, *softwares*, interfaces, preservação digital e plataformas de arquivamento da *web*, atuando sobre princípios de *software* livre, *open source* e interoperabilidade entre sistemas, favorecendo o modelo colaborativo de desenvolvimento de soluções, além de soluções



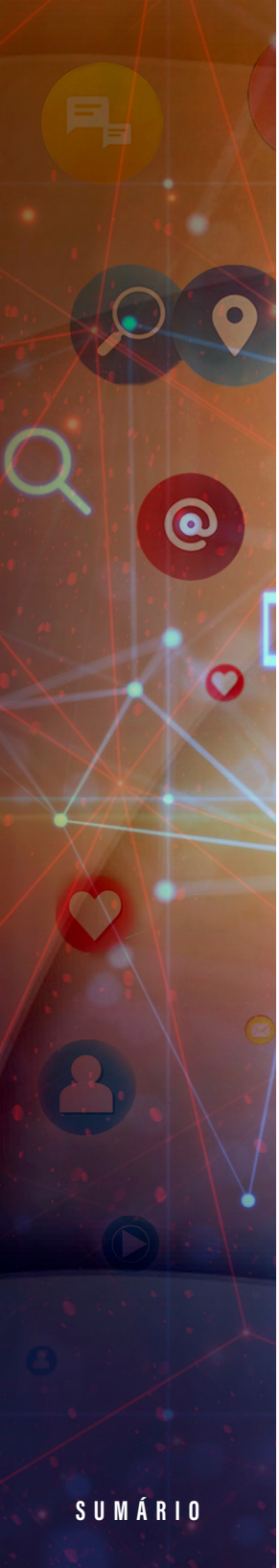
que incluam o uso de Inteligência Artificial. Enfoque especial nos estudos de rastreadores usados no arquivamento da *web*, projetadas para capturar conteúdo de *site*s específicos ou rastreamento mais amplo (micro e macroarquivamento).

6. Questões éticas e legais a respeito do arquivamento da *web*, investigando riscos e possibilidades relacionadas a direitos autorais, patrimoniais, financeiros, *copyright* e licenças de uso, coleta e armazenamento de informações pessoais, depósito legal e responsabilidade ética na preservação e no acesso a informações públicas.

7. Investigar possibilidades de desenvolvimento e disseminação de expertise entre equipes interdisciplinares que trabalhem com o desenvolvimento de arquivos da *web* e capacitações direcionadas aos produtores dos *site*s (*webmaster*s, *webdesigners*, profissionais de tecnologia da informação), para que trabalhem em formato colaborativo

8. Estudo de perfis de usuários e de públicos, bem como o engajamento de comunidades que auxiliem o arquivamento da *web* e a difusão das informações.

Nos últimos anos, observamos uma mudança significativa na maneira como concebemos instituições de memória, tais como arquivos, bibliotecas e museus. Antes vistos como locais de armazenamento de documentos, esses espaços passaram a ser entendidos como fontes de dados para uma ampla variedade de finalidades. Esse novo paradigma de dados vem transformando a forma como a informação é utilizada em diferentes áreas, permitindo que dados antes isolados sejam acessados e analisados de maneira mais eficiente e eficaz. Em um mundo cada vez mais conectado e orientado por dados, essa mudança representa uma oportunidade para que instituições de memória ampliem seu papel, atuando como fontes de informação confiáveis e contribuindo com o seu papel na



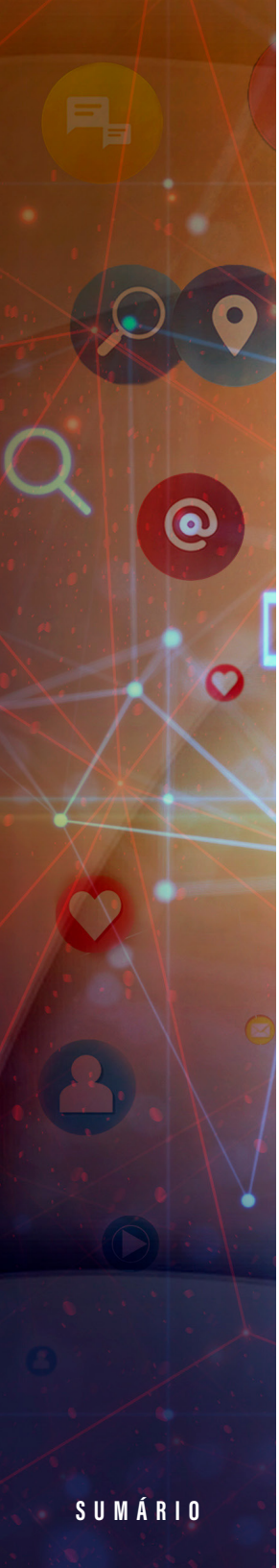
sociedade. Desta forma, o arquivamento da *web* vem ao encontro da necessidade da preservação de uma memória digital do nosso tempo e de seu acesso futuro.

Ainda, nos objetivos do projeto de arquivamento da *web* devemos definir se faremos um arquivamento em grande ou pequena escala. Devem ser analisadas e escolhidas as ferramentas de *software* que serão utilizadas durante os processos, do rastreamento até a descoberta e acesso pelo público, assim como os metadados e mecanismos de busca que serão disponibilizados.

Para auxiliar na elaboração de um projeto de arquivamento da *web* que aborde os aspectos elencados acima, como ponto de partida para a estruturação de um projeto de arquivamento da *web*, levando em conta os objetivos da organização interessada em preservar e manter seu patrimônio digital, elencamos algumas boas práticas, modelos e fluxos utilizados em diversas instituições e na literatura da área.

Qualquer organização interessada em realizar um projeto de arquivamento da *web* deve avaliar e analisar suas funções, plano estratégico, missão e visão, o que lhe permitirá definir o alcance e a precisão dos objetivos do projeto. A correta definição dos objetivos, garantirá o sucesso e a sustentabilidade do arquivamento da *web*, dado que o projeto se enquadra nos propósitos da organização, permitirá selecionar especificamente quais *site*s da *web* serão capturados, dimensionará a complexidade do processo de arquivamento, identificará a extensão da captura, se escolherá o tipo de arquivamento e as estratégias certas para superar desafios e riscos associados ao projeto.

Manter um projeto de arquivamento *web* implica comprometimento a longo prazo. Para superar este desafio administrativo a organização deve identificar aliados que se articulam com a iniciativa e estão interessados em aderir ao projeto. Mas devem ser analisados os pontos fortes e recursos disponíveis de cada parceiro para garantir a definição e padronização de processos e fluxos de trabalho e atribuir responsabilidades e níveis de participação.

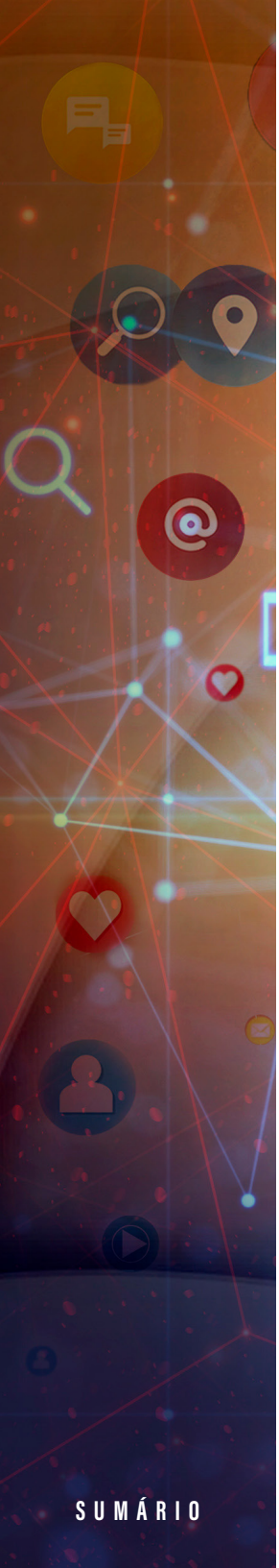


Definir a política de arquivamento da *web* orienta e facilita a tomada de decisões na execução do projeto, orienta na escolha das ferramentas de *software*, na definição e padronização de processos e no desenho dos fluxos de trabalho, na atribuição de responsabilidades e na administração, no uso, reuso e acesso de seus arquivos pela comunidade de usuários interessados.

O trabalho também envolve escolher estratégias de preservação apropriadas, adequadas ao projeto e de acordo com a complexidade dos arquivos da *web* que se deseja preservar. Estas estratégias devem garantir a disponibilidade e acesso aos recursos a longo prazo. Adotar as melhores práticas e padrões internacionais é fundamental para enfrentar os principais desafios de arquivamento da *web*. No entanto, a preservação digital é um assunto em constante evolução, que exige uma contínua atualização e treinamento pelos líderes do projeto.

Monitorar e analisar a conformidade do processo estabelecido e as responsabilidades atribuídas, as ferramentas utilizadas, o desempenho, a assertividade das estratégias adotadas para superar os desafios e riscos, tanto técnico como administrativo, é uma atividade que deve ser gerenciada transversalmente e continuamente durante a toda a execução do projeto para identificar oportunidades de melhorias e evitar desvios na abordagem do método de trabalho. Como resultado desta etapa deve-se gerar estratégias ou alternativas para garantir qualidade no arquivamento da *web*.

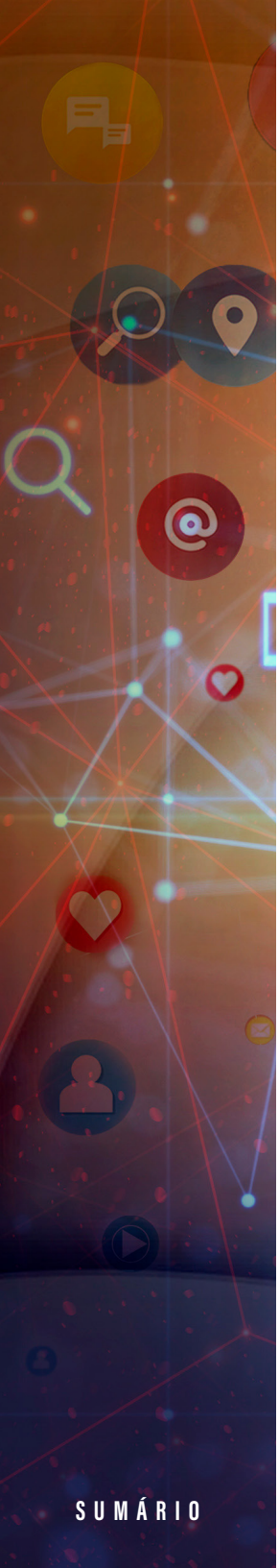
Apesar da crença comum de que o conteúdo *on-line* é permanente, pesquisas mostram que uma parte significativa dele desaparece com o tempo. A preservação da *web* é importante não apenas para a pesquisa acadêmica e responsabilidade institucional, mas também porque cada vez mais nossas vidas profissionais e pessoais acontecem no ambiente *on-line*.



Existem várias razões pelas quais um *site* pode ficar inacessível ou ser perdido para sempre. Um dos motivos mais comuns é uma falha do servidor, que pode ocorrer durante as operações regulares de uma organização, levando à perda de dados em pequena ou grande escala. Outro fator é a quebra de *link* (*link rot*), onde os *hiperlinks* não levam mais ao seu destino original, resultando em conexões perdidas. Em alguns casos, empresas ou instituições públicas podem encerrar suas atividades, passar por fusões ou transformações políticas, resultando na perda de dados quando não há preocupação com a preservação digital. O não pagamento ou manutenção dos serviços de hospedagem também pode levar à perda do conteúdo, que pode ser ocasionado por abandono de projetos ou falências de empresas. Às vezes, mudanças na liderança podem resultar na exclusão ou alteração de informações *on-line*. Além disso, o apagamento de conteúdo pode ocorrer quando empresas privadas ou governos suprimem palavras, imagens e ideias contrárias, instalando um ambiente de censura. O arquivamento da *web* procura suprimir esses apagamentos e manter viva a história digital e os recursos da *web* enquanto fonte de informação para pesquisa e o público em geral.

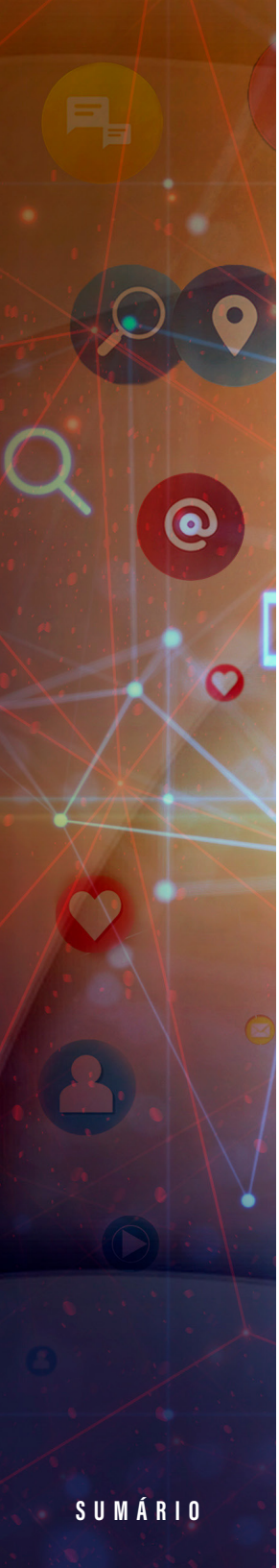
Por esses motivos, o arquivamento da *web* nos apresenta diversos desafios, alguns de natureza administrativa como os jurídicos, de seleção e alcance e mesmo de atribuição de responsabilidades entre os membros da equipe executora ou os desafios técnicos como, por exemplo, os decorrentes das constantes atualizações das páginas *web*, das limitações dos rastreadores, dos vírus e *malware*, da duplicação de conteúdos e da preservação a longo prazo.

Procuramos ilustrar aqui as oportunidades e desafios envolvidos na gestão e preservação, o patrimônio digital salvo pelos processos de arquivamento da *web* ou o patrimônio perdido pela falta destes processos. É notório que os projetos de arquivamento da *web* pretendem ser tão abrangentes quanto possível, por mais que a gama de conteúdo da *web* algumas vezes possa ser limitada pela legislação ou por questões técnicas ou administrativas, mas apesar



das barreiras oferece excitantes oportunidades de explorar um universo de oportunidades de estudo e pesquisas. Embora exista uma literatura em crescimento nesta área, os arquivos da *web* continuam sendo uma área de estudo a ser explorada. Também, devemos pensar em desenvolver trabalhos com o envolvimento da comunidade de pesquisadores que utilizam dos ricos materiais arquivados, permitir aos pesquisadores explorar e analisar os arquivos da *web* sem a necessidade de aprender recursos avançados de codificação, que muitas vezes provam ser uma barreira ao trabalho com os arquivos da *web*, propor um processo de curadoria participativa que auxilie a dar forma a um ambiente mais democrático, aberto e inclusivo ao arquivamento da *web* nas organizações.

Finalmente, a preservação da *web* deveria ser um programa obrigatório para os arquivos, bibliotecas e sistemas de informação em um futuro próximo, bem como nos cursos que formam estes profissionais. Os *websites* contém recursos especiais de conhecimento humano, mas a fragilidade da rede digital pode prejudicar a acessibilidade a esses conteúdos para as próximas gerações. Somente adotando o compromisso da preservação digital é que os arquivos, bibliotecas e centros de informação serão capazes de manter a memória coletiva a partir dos conteúdos da *web*.



DICIONÁRIO DE TERMINOLOGIA EM ARQUIVAMENTO DA *WEB* E PRESERVAÇÃO DIGITAL

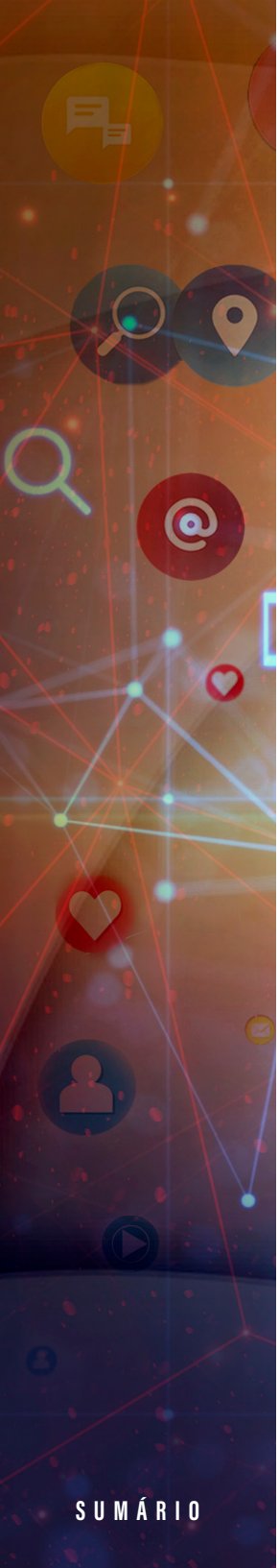
Os termos aqui apresentados servem como auxílio para a compreensão dos principais conceitos utilizados nesta obra e em projetos envolvendo arquivos da *web*[193]. Até o acesso dos arquivos da *web* pelos usuários, uma série de estratégias, ações, tecnologias e métodos são necessários e para isso precisam ser levados em consideração os recursos organizacionais, bem como as características específicas de cada contexto. O conhecimento dos termos mais utilizados neste campo visa a padronização e entendimento mútuo por parte dos diversos profissionais que atuam nesses projetos.

Arquivamento da *web*: processo que compreende capturar, armazenar e disponibilizar a informação retrospectiva da *World Wide Web* para cidadãos, servindo como preservação da memória institucional.

Arquivo da *web*: conteúdos publicados na *web*, sejam *website*s ou mídias sociais, que uma instituição tomou a responsabilidade e providências para a preservação digital e que mantém as mesmas características de informação e navegabilidade das páginas *web* originais.

Autenticidade: credibilidade de um documento, isto é, a qualidade de um documento ser o que diz ser e de estar livre de adulteração ou qualquer outro tipo de corrupção.

193 Alguns dos termos apresentados são síntese da revisão de literatura e dos documentos "Política de Preservação de Acervos Digitais" da Universidade Federal do Rio Grande do Sul e "Requisitos mínimos de preservação para *websites* e mídias sociais" do Conselho Nacional de Arquivos (CONARQ) (2023b).



Captura: etapa do método de preservação dos *website*s e mídias sociais, correspondente ao recolhimento das páginas *web* baseada em uma lista de entradas (URLs pré-definidas). Também chamado em inglês de *harvesting* (colheita).

Conteúdo dinâmico: página ou *site* gerados automaticamente pelo servidor da *web*, em oposição ao conteúdo estático, onde o resultado da requisição é uma página ou *site* previamente construído. O conteúdo dinâmico pode ser personalizado a partir da solicitação direta do usuário ou de preferências do usuário, com base em *cookies* ou *login*.

Controle de qualidade: verificação do processo de captura e rastreamento, de forma a identificar a precisão e integridade do conteúdo da *web* arquivado.

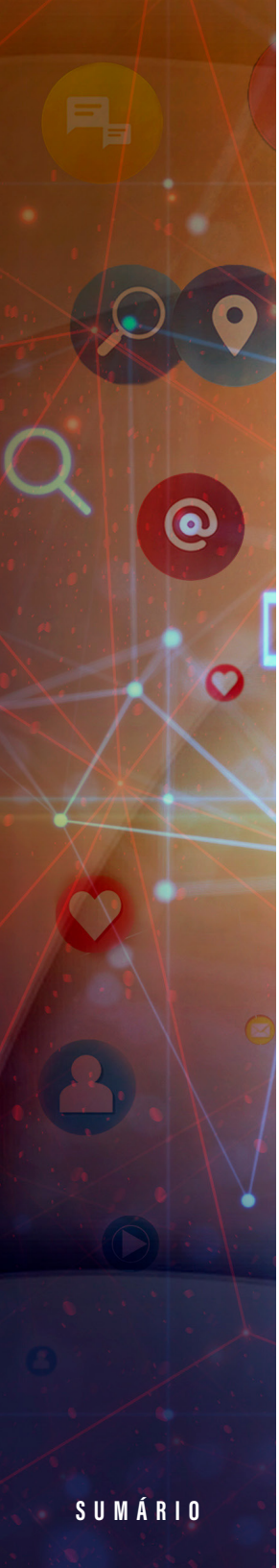
Documento digital complexo: documentos não estáveis que agregam dados, metadados e às vezes serviços em uma única entidade digital lógica.

Elementos da *web*: compreendem o conteúdo (textual, visual, multimídia etc.) e a estrutura (leiaute, apresentação, comportamento de navegação) da *web*.

Embargo: o embargo aplicado ao arquivamento da *web* é uma restrição temporária aplicada a determinados recursos, a fim de limitar o acesso por um período de tempo específico, geralmente aplicado aos conteúdos recém-capturados, de forma a evitar acessos concorrentes com o conteúdo original.

Encurtador de link: técnica utilizada na Internet para transformar um endereço HTTP em um *link* mais curto. A URL original passa a ser acessada pelo novo *link* com menos caracteres.

Fim de vida: *site*s que estão sob risco de desaparecimento, com maior probabilidade de não estarem mais disponíveis na *web* ou não serem mais acessíveis por diversos motivos, por exemplo, organizacionais, políticos e técnicos.



Hyperlink: também abreviado como *link*, é uma referência digital que, por meio de um clique, leva a outro recurso *on-line*, desta forma pode conectar uma página ou documento a diversos outros recursos.

Mídias sociais: conteúdos criados e compartilhados pelos usuários, em plataformas de redes sociais.

Objeto digital: conjunto de uma ou mais cadeias de *bits* que registram o conteúdo do objeto e de seus metadados associados.

Perfil de rede social: ambiente particular em uma rede social utilizado para interação de indivíduos e entidades, bem como para a disseminação de informações. No arquivamento da *web* pode servir como referência para ser preservada.

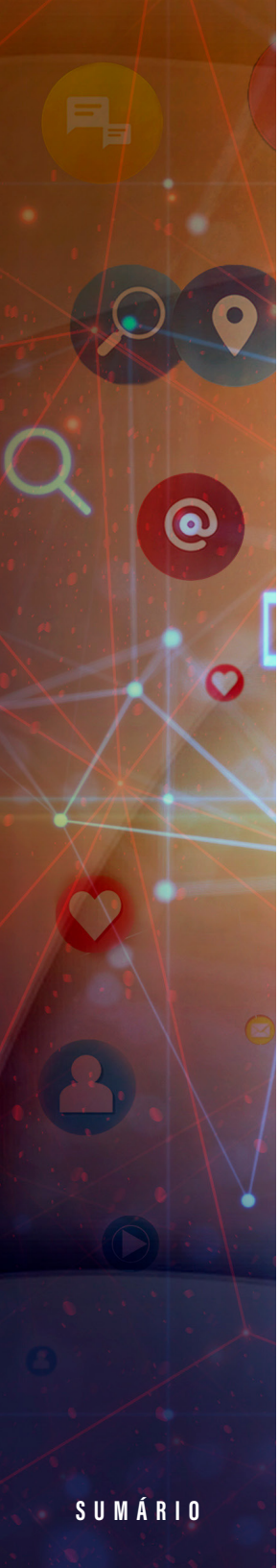
Quebra de link: também conhecido como *link rot*, consiste no fenômeno de *hiperlinks* se tornarem quebrados ou inativos com o passar do tempo.

Rastreador: em inglês *crawler*, ferramenta de *software* que pode capturar o conteúdo da *World Wide Web* de forma automatizada.

Robots.txt: arquivo colocado em um *site* que instrui os rastreadores da *web* sobre quais páginas excluir da indexação.

Semente: em inglês *seed*, constitui um ponto de partida para os rastreadores da *web*, geralmente um URL ou domínio específico.

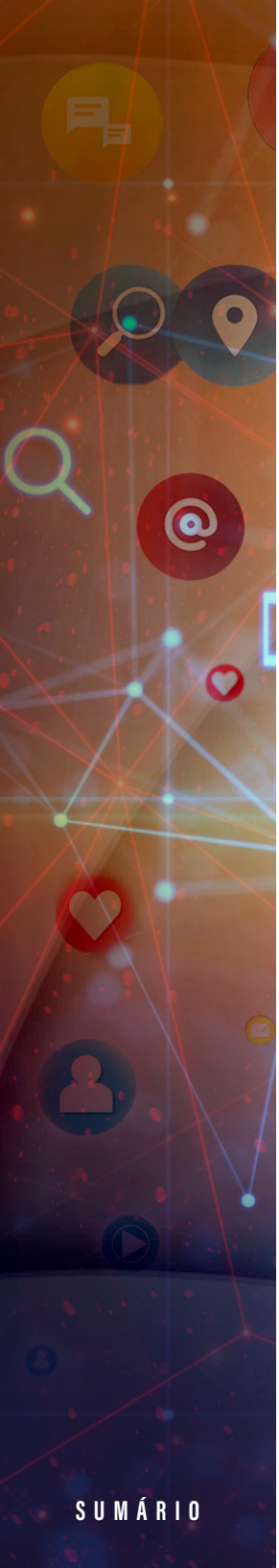
Site: também chamado de *website*, sítio ou página da *web*, geralmente é composto por diversos arquivos, normalmente HTML (*HyperText Markup Language*), CSS (*Cascading Style Sheets*) e *Javascript*, podendo conter outros arquivos e conteúdo multimídia. Estes recursos são agrupados e acessíveis por um endereço comum.



Streaming: tecnologia que transmite dados pela Internet, principalmente áudio e vídeo, sem a necessidade de baixar o conteúdo. O arquivo é acessado pelo usuário de forma *on-line*.

URL: do inglês *Uniform Resource Locator* ou localizador uniforme de recursos, identifica um recurso em uma rede informática.

WARC: formato de arquivo em modelo *open source*, lançado em 2009 pela *International Organization for Standardization* (ISO), com versão 2017, ISO 28500:2017, destinada à preservação de arquivos da *web* em longo prazo.



REFERÊNCIAS

ABBATE, Janet. **Inventing the Internet**. Massachusets: MIT, 1999.

AFONSO, Carlos Alberto. **Internet no Brasil:** o acesso para todos é possível? Policy Paper, n. 26, 2000. Disponível em: https://docplayer.com.br/520918-Internet-no-brasil-o-acesso-para-todos-e-possivel.html. Acesso em: 18 jan. 2023.

AGATA, Teru *et al*. Life span of *web* pages: a survey of 10 million pages collected in 2001. In: IEEE/ACM JOINT CONFERENCE ON DIGITAL LIBRARIES, 14., 2014, London. **Proceedings [...].** London, 2014. p. 463-464.

ALVES, Daniel. As humanidades digitais como uma comunidade de práticas dentro do formalismo acadêmico: dos exemplos internacionais ao caso português. **Ler História**, v. 69, p. 91-103, 2016.

ARONSKY, Dominik, *et al*. The prevalence and inaccessibility of Internet references in the biomedical literature at the time of publication. **Journal of the American Medical Informatics Association**, v. 14, n. 2, p.232-234, 2007.

ARQUIVO NACIONAL DO BRASIL. **Política de preservação digital**. Versão 2. Dezembro de 2016. Disponível em: http://www.siga.arquivonacional.gov.br/images/an_digital/and_politica_preservacao_digital_v2.pdf. Acesso em: 05 fev. 2023.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 15472**: sistemas espaciais de dados e informações - modelo de referência para um sistema aberto de arquivamento de informação (SAAI). Rio de Janeiro: ABNT, 2007.

AUSTRALIA. NATIONAL ARCHIVES OF AUSTRALIA. **Archiving *web* resources**: guidelines for keeping records of *web*-based activity in the commonwealth government. Canberra, 2001. Disponível em: https://www.ltu.se/cms_fs/1.67312!/file/Archiving*web*ResourcesGuidelinesOsv.pdf. Acesso em: 04 abr. 2023.

BAKER, Mary; KEETON, Kimberly; MARTIN, Sean. Why traditional storage systems don't help us save stuff forever. In: IEEE WORKSHOP ON HOT TOPICS IN SYSTEM DEPENDABILITY, 1., 2005, Yokohama. **Proceedings [...].** Yokohama, 2005.

BALOGUN, Tolulope; KALUSOPA, Trywell. Web archiving of indigenous knowledge systems in South Africa. **Information Development**, v. 38, n. 4, p. 658-671, 2022. DOI: https://doi.org/10.1177/02666669211005522.

BANOS, Vangelis; MANOLOPOULOS, Yannis. A quantitative approach to evaluate *website* Archivability using the CLEAR+ method. **International Journal on Digital Libraries**, v. 12, n. 2, p. 119-141, 2016.

BANOS, Vangelis *et al*. CLEAR: a credible method to evaluate *website* archivability. *In*: CONFERENCE ON PRESERVATION OF DIGITAL OBJECTS, 2013. **Proceedings** [...]. v. 2, 2013.

BENBOW, S. Mary P. File not found: The problems of changing URLs for the World Wide Web. **Internet Res**. 1998, 8, 247–250. Disponível em: http://dx.doi.org/10.1108/10662249810217867 Acesso em: 27 nov. 2022.

BEN-DAVID, Anat. (2019): 2014 not found: a cross-platform approach to retrospective web archiving. **Internet Histories**, v. 3, n. 3-4, p. 316-342, 2019. DOI: https://doi.org/10.1080/24701475.2019.1654290.

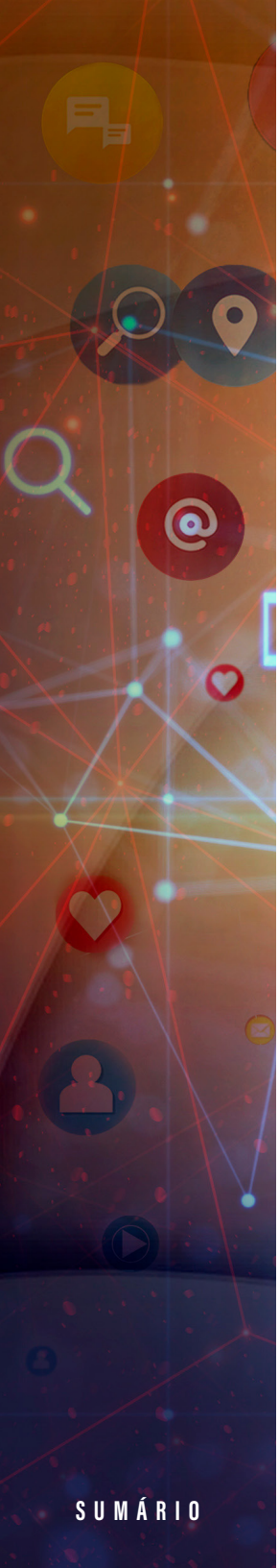
BEN-DAVID, Anat; HUURDEMAN, HUGO. Web Archive Search as Research: Methodological and Theoretical Implications. **Alexandria**, v. 25, n. 1-2, p. 93-111, 2014. DOI: https://doi.org/10.7227/ALX.0022.

BERMÈS, Emmanuelle. **Des identifiants pérennes pour les ressources numériques**. Bibliothèque Nationale de France, 2006a. Disponível em: http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf. Acesso em: 12 mar. 2023.

BERMÈS, Emmanuelle. **Des identifiants pérennes pour les ressources numériques**: L'expérience de la BnF. Bibliothèque Nationale de France, 2006b. Disponível em: http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf. Acesso em: 11 mar. 2023.

BERNERS-LEE, Tim. **Information management**: a proposal. 1998. Disponível em: https://www.w3.org/History/1989/proposal.html. Acesso em: 13 fev. 2023.

BIBLARZ, Dora *et al*. **Guidelines for a Collection Development Policy Using the Conspectus Model**. IFLA, 2001. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/acquisition-collection-development/publications/gcdp-en.pdf. Acesso em: 29 abr. 2023.



BINGHAM, Nicola. Quality Assurance Paradigms in *web* Archiving Pre and Post Legal Deposit. **Alexandria**: The Journal of National and International Library and Information, v. 25, n. 1-2, p. 51-68, 2014. DOI 10.7227/alx.0020.

BINGHAM, Nicola Jayne; BYRNE, Helena. Archival strategies for contemporary collecting in a world of big data: challenges and opportunities with curating the UK *web* archive. **Big Data & Society**, v. 8, n. 1, p. 1-6, jan./jun. 2021.

BLANCO-RIVERA, Joel Antonio. Tendencias y desafíos en la creación de metadatos en proyectos de archivo de la *web*. **e-Ciencias de la Información**, v. 12, n. 1, p. 79-95, 2022.

BODÊ, Ernesto Carlos. **Preservação de documentos digitais**: o papel dos formatos de arquivo. Brasília, 2008. 153 f. Dissertação (Mestrado em Ciência da Informação) - Universidade de Brasília, Brasília, 2008.

BOERES, Sonia Araújo de Assis; SAAD, Rondineli Gama. Arquivamento da Web: definições, estratégias, fluxos e iniciativas. **Revista Brasileira de Preservação Digital**, Campinas, SP, v. 4, 2023. DOI 10.20396/rebpred.v4i00.17934. Disponível em: https://econtents.bc.unicamp.br/inpec/index.php/rebpred/article/view/17934. Acesso em: 20 jun. 2023

BONACCHI, Chiara; KRZYZANSKA, Marta. Digital heritage research retheorised: ontologies and epistemologies in a world of big data. **International Journal of Heritage Studies**, v. 25, n. 12, feb. 2019.

BORGES, Jorge Luis. La biblioteca de Babel: prólogos. Emecé: Buenos Aires, 2000.

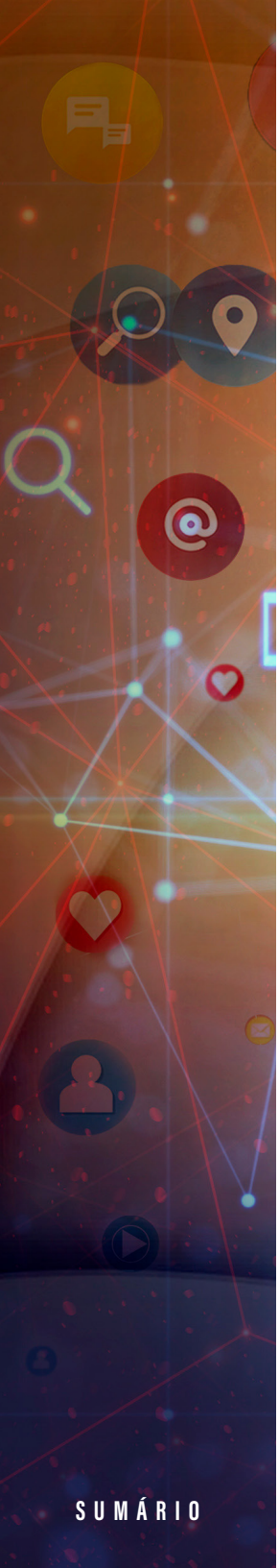
BORGES, Jorge Luis. O jardim dos caminhos que se bifurcam. **Ficções**, v. 4, p. 01-83, 2000.

BOWERS, John. **A million squandered**: the million dollar homepage as a decaying digital artifact. 2017. Disponível em: https://goo.gl/rppeQV. Acesso em: 24 fev. 2023.

BRAGG, Molly, HANNA, Kristine. **The *web* archiving life cycle model**. 2013. Disponível em: https://archive-it.org/static/files/archiveit_life_cycle_model.pdf Acesso em: 04 abr. 2023.

BRAND, Stewart. Escaping the digital dark age. **Library Journal**, v. 124, n. 2, p. 46-48, 1999.

BRAYNER, A. Alencar. Programa de arquivo de páginas *web* no reino unido: uma breve história de oportunidades e desafios**. RDBCI:** Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 14, n. 2, p. 318–333, 2016.



BROWN, Adrian. **Archiving Websites**: a practical guide for information management professionals. Facet publishing: London, 2006.

BRÜGGER, Niels. Digital humanities and *web* archives: possible new paths for combining datasets. **International Journal of Digital Humanities**, v. 2, n. 1/3, p. 145-168, 2021.

BRÜGGER, Niels; MILLIGAN, Ian (Org.) **The SAGE handbook of *web* history**. SAGE Publications Limited, 2019.

BRÜGGER, Niels; FINNEMANN, Niels Ole. The *web* and digital humanities: theoretical and methodological concerns. **Journal of Broadcasting and Electronic Media**, v. 57, n. 1, 2013.

BRÜGGER, Niels. *web* archiving: between Past, Present, and Future. *In*: CONSALVO, Mia; ESS, Charles. **The Handbook of Internet Studies**. Oxford: Blackwell, 2011. p. 24-42.

BRÜGGER, Niels. **Archiving *websites***: general considerations and strategies. Århus: CFI, 2005.

BRUNELLE, Justin F. *et al*. Not all mementos are created equal: measuring the impact of missing resources. **International Journal on Digital Libraries**, v. 16, p. 283-301, 2015. DOI: https://doi.org/10.1007/s00799-015-0150-6.

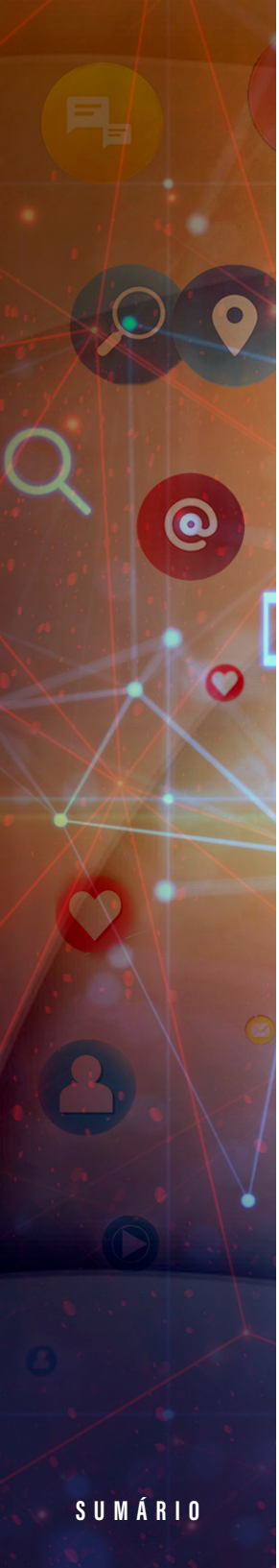
BURNHILL, Peter; MEWISSEN, Muriel; WINCEWICZ, Richard. Reference rot in scholarly statement: threat and remedy. **Insights**, v. 28, n. 2, 2015.

CALLISTER, Paul Douglas. Perma. cc and Web Archival Dissonance with Copyright Law. **Legal Reference Services Quarterly**, v. 40, n. 1, p. 1-57, 2021.

CASTELLS, Manuel *et al*. **Comunicação móvel e sociedade**: uma perspectiva Global. Lisboa: Fundação Calouste Gulbenkian, 2009.

CONSELHO NACIONAL DE ARQUIVOS. **104 a reunião ordinária do CONARQ**, 07 de dezembro de 2022, quarta-feira, das 15h às 17h. Brasil, 7 dez. 2022. Disponível em: https://www.facebook.com/ConselhoNacionaldeArquivos/videos/678607297071001. Acesso em: 13 fev. 2023.

CONSELHO NACIONAL DE ARQUIVOS. **Carta para a preservação do patrimônio arquivístico digital**. Rio de Janeiro: Arquivo Nacional, 2005. Disponível em: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/conarq_carta_preservacao_patrimonio_arquivistico_digital.pdf. Acesso em: 11 jan. 2023.



CONSELHO NACIONAL DE ARQUIVOS. **Portaria CONARQ n. 131**, de 9 de novembro de 2021. Institui a Câmara Técnica Consultiva para definir diretrizes para a elaboração de estudos, proposições e soluções para a preservação de *site*s e mídias sociais. Rio de Janeiro, 2021a. Disponivel em: https://www.gov.br/conarq/pt-br/assuntos/acesso-a-informacao/portarias-conarq-1/PORTARIACONARQN131DE9DENOVEMBRODE2021.pdf. Acesso em: 03 mar. 2023.

CONSELHO NACIONAL DE ARQUIVOS. **Requerimento ao Conselho Nacional de Arquivos, instituição de câmara técnica consultiva para a elaboração de estudos, proposições e soluções para a preservação de *site*s e mídias sociais**. Rio de Janeiro, 2021b.

CONSELHO NACIONAL DE ARQUIVOS. **Resolução CONARQ nº 52, de 25 de agosto de 2023**. Estabelece a Política de Preservação de Websites e Mídias Sociais no âmbito do Sistema Nacional de Arquivos (SINAR), 2023a. Disponível em: https://www.in.gov.br/en/web/dou/-/resolucao-conarq-n-52-de-25-de-agosto-de-2023-530263510. Acesso em: 13 dez. 2023.

CONSELHO NACIONAL DE ARQUIVOS. **Resolução CONARQ n° 53, de 25 de agosto de 2023**. Define requisitos mínimos de preservação para websites e mídias sociais no âmbito do Sistema Nacional de Arquivos (SINAR), 2023b Disponível em: https://www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/resolucao-no-53-de-25-de-agosto-de-2023. Acesso em: 28 dez. 2023.

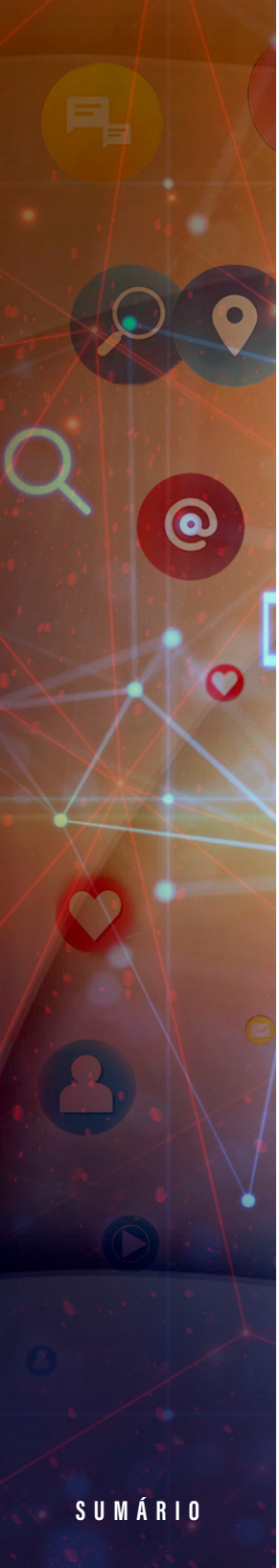
CORUJO, Luis; REVEZ, Jorge; SILVA, Carlos Guardado. Digital curation and its costs: a study of practices and insights. **DHQ: Digital Humanities Quarterly**, v. 14, n. 2, 2020.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário. The evolution of *web* archiving. **International Journal on Digital Libraries**, v. 18, n. 3, p. 1-15, 2016.

DANTAS, Camila Guimarães. **Criptografias da memória**: um estudo teórico-prático sobre o arquivamento da *web* no Brasil. Rio de Janeiro, 2014. 228 f. Tese (Doutorado em Memória Social) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

DAY, Michael. Preserving the fabric of our lives: a survey of *web* preservation initiatives. In: RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES: EUROPEAN CONFERENCE, 7., 2003, Trondheim. **Proceedings [...].** Trondheim: Springer Berlin Heidelberg, 2003.

DELLAVALLE, Robert, *et al*. Going, going, gone: lost Internet references. **Science** v. 302, n. 5646, p. 787-788, 2003.



DELJANIN, Sandra. Digital obsolescence. **INFOtheca**, v. 13, n. 1, p. 43-53, 2012.

DELMAS, Bruno. **Arquivos para quê?** Textos escolhidos. São Paulo: Instituto Fernando Henrique Cardoso, 2010.

DENEMARK, Howard A. The death of law reviews has been predicted: what might be lost when the last law review shuts down? **Seton Hall Law Review**, v. 27 n. 1, p. 1-32, 1996.

DI PRETORO, Emmanuel; GEERAERT, Friedel. Behind the scenes of *web* archiving: metadata of harvested *website*s. **Archives et Bibliothèques de Belgique-Archief-en Bibliotheekwezen in België**, v. 106, p. 63-74, 2019.

DI PRETORO, Emmanuel; GEERAERT, Friedel; SOYEZ, Sébastien. Behind the Scenes of Web Archiving Metadata of Harvested Websites. **Archives et Bibliothèques de Belgique**, n. 106, p. 63-73, 2019. Disponível em: https://hal.science/hal-02124714/document. Acesso em: 27 nov. 2022.

DOOLEY, Jackie; BOWERS, Kate. **Descriptive metadata for *web* archiving**: Recommendations of the OCLC Research Library Partnership *web* Archiving Metadata Working Group. Dublin: OCLC Research, 2018.

DOOLEY, Jackie M. *et al*. Developing *web* Archiving Metadata Best Practices to Meet User Needs. **Journal of Western Archives**, v. 8, n. 2, artigo 5, 2017. DOI: https://doi.org/10.26077/cffd-294a

DOUGHERTY, Meghan; MEYER, Eric T. Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs. **Journal of the Association for Information Science and Technology**, v. 65, n. 11, p. 2195-2209, 2014. DOI: https://doi.org/10.1002/asi.23099.

DOUGHERTY, Meghan. Property or privacy? reconfiguring ethical concerns around *web* archival research methods. 2013. Disponível em: https://journals.uic.edu/ojs/index.php/spir/article/view/8804. Acesso em: 12 jan. 2023.

EVANGELOU, Evangelos; TRIKALINOS, Thomas A.; LOANNIDIS, John P. Unavailability of *online* supplementary scientific information from articles published in major journals. **The FASEB Journal,** v. 19, n. 14, 2005.

FARIAS, Juliana Pinheiro; BOMFIM, Kelen Cândida Vieira. A produção científica sobre preservação de *website*s em língua portuguesa. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 24, n. 55, p. 1-15, 2019.

FARRELL, Susan; (ed). **A guide to *web* preservation**: Practical advice for *web* and records managers based on best practices from the JISC-funded PoWR project, 2010. Disponível em: http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf Acesso em: 10 mar. 2023.

FAZANO, Igor Ferreira. **Arquivamento de páginas da *web* como recurso de preservação digital**: estudo de caso das coleções de *site*s brasileiros da Library of Congress. Rio de Janeiro, 2022. Dissertação (Mestrado Profissional em Bens Culturais e Projetos Sociais) - Fundação Getúlio Vargas, Rio de Janeiro, 2022.

FERREIRA, Lisiane B.; MARTINS, Marina R.; ROCKEMBACH, Moisés. Usos do arquivamento da *web* na comunicação científica. **Prisma.com**, v. 36, p. 78-98, 2018.

FERREIRA, Lisiane. **Arquivamento da *web* e mídias sociais**: preservação digital de vídeos da campanha presidencial brasileira de 2018. Porto Alegre, 2019. 105 f. Dissertação (Mestrado em Comunicação e Informação) - Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

FERREIRA, Lisiane, ROCKEMBACH, Moisés. Preservação de mídias sociais e arquivamento da web: um estudo acerca das eleições presidenciais brasileiras de 2018. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO. 2019. **Anais Eletrônicos** [...]. Florianópolis: ANCIB, 2019.

FONTOURA, Marcelo Carneiro da. **A Documentação de Paul Otlet**: uma proposta para a organização racional da produção intelectual do homem. Brasília, 2012. 219 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade Federal de Brasília, Brasília, 2012. Disponível em: https://repositorio.unb.br/handle/10482/11909. Acesso em: 29 mar. 2023.

FORMENTON, Danilo; GRACIOSO, Luciana de Souza. Padrões de metadados no arquivamento da *Web*: recursos tecnológicos para a garantia da preservação digital de *website*s arquivados. **RDBCI:** Revista Digital de Biblioteconomia e Ciência da Informação, v. 20, p. e022001, 2022. DOI https://doi.org/10.20396/rdbci.v20i00.8666263.

FORMENTON, Danilo; GRACIOSO, Luciana de Souza. Preservação digital: desafios, requisitos, estratégias e produção científica. **RDBCI:** Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 18, p. e020012, 2020.

GHOSH, Pallab. Google's Vint Cerf warns of 'digital Dark Age.' **BBC news**, v. 13, 2015. Disponível em: https://www.bbc.com/news/science-environment-31450389. Acesso em: 10 mar. 2023.

GOH, Dion Hoe Lian; Ng, Peng Kin. Link decay in leading information science journals. **Journal of the American Society for Information Science and Technology**, v. 58, n. 1, 2007.

GOMES, Daniel. Preservar a *web*: um desafio ao alcance de todos. *In*: ATAS DO CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS. BAD Portugal, v. 10, 2010.

GRÁCIO, José Carlos Abbud; MADIO, Telma Campanha de Carvalho. O papel da preservação digital na curadoria digital. *In:* JORENTE, Maria José Vicentini *et al.* **Curadoria digital e gênero na Ciência da Informação**. São Paulo: Oficina Universitária, 2021.

HARRIS, Kayla; BEIS, Christina A.; SHREFFLER Stephanie. Citizen *web* archivists: applying *web* archiving as a pedagogical tool. **Journal of Electronic Resources Librarianship**, v. 33, n. 4, p. 262-272, 2021.

HAYES, Brian. Bit rot. **American Scientist**, v. 86, n. 5, p. 410-415, 1998.

HEDSTROM, Margaret. Digital preservation: a time bomb for digital libraries. **Computers and the Humanities**, v. 31, n. 3, p. 189-202, 1997.

HELMOND, Anne; VAN DER VLIST, Fernando N. Social media and platform historiography: Challenges and opportunities. **TMG–Journal for Media History**, , v. 22, n. 1, 2019.

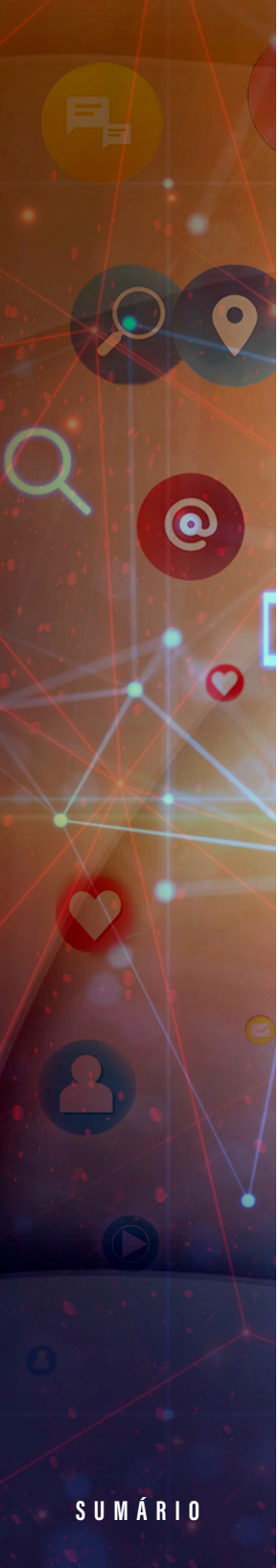
HOCKX-YU, Helen. Access and scholarly use of *web* archives. **Alexandria**, v. 25, n. 1/2, p. 113-127, 2014.

HOCKX-YU, Helen. Archiving social media in the context of non-print legal deposit. *In*: IFLA WORLD LIBRARY AND INFORMATION CONGRESS, 2014, Lyon. Libraries, Citizens, Societies: Confluence for Knowledge, 2014. Disponível em: https://library.ifla.org/id/eprint/999. Acesso em: 30 abr. 2023.

HOCKX-YU, Helen. **How to make *website*s more archivable?** UK *web* Archive blog, 2012. Disponível em: https://britishlibrary.typepad.co.uk/*web*archive/2012/09/how-to-make-*website*s-more-archivable.html. Acesso em: 02 mar. 2023.

HOCKX-YU, Helen. The past issue of the *web. In:* INTERNATIONAL *WEB* SCIENCE CONFERENCE, 3., 2011, New York. **Proceedings [...].** New York: Association for Computing Machinery, 2011.

HUDGINS, Allison. M. Preservation of the video game**. Provenance, Journal of the Society of Georgia Archivists**, v. 29, n. 1, p. 32-48, 2011.



INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. **About IIPC**. 2023. Disponível em: https://netpreserve.org/about-us/. Acesso em: 12 fev. 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 28500**: information and documentation: WARC file format. Geneve, 2017.

JENKINS, Henry. **Cultura da convergência**. São Paulo: Aleph, 2008.

JUTY, Nic *et al*. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. **Data Intelligence**, v. 2, p. 30–39, 2020. DOI 10.1162/dint_a_00025.

KETELAAR, Eric. Tacit Narratives: The Meanings of Archives. **Archival Science,** v. 1, n. 2, p. 31–41, 2001. Disponível em: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/41812/10502_2004_Article_359685.pdf;jsessionid=191C545CFA393364136FEE3AB4CE36D4?sequence=1. Acesso em: 20 mar. 2023.

KHAN, Muzammil; UR RAHMAN, Arif. A systematic approach towards *web* preservation. **Information Technology and Libraries**, v. 38, n. 1, p. 71-90, 2019.

KIM, Heejung; LEE, Hyewon. Development of metadata elements for intensive *web* archiving. **Journal of the Korean Society for Information Management**, Songdo, South Korea, v. 24, n. 2, p. 143-160, June 2007. Disponível em: https://kosim.accesson.kr/assets/pdf/303/journal-24-2-143.pdf. Acesso em: 09 mar. 2023.

KLEIN, Martin, *et al*. Scholarly context not found: one in five articles suffers from reference rot. **PloS one**, , v. 9, n. 12, e115253, 2014.

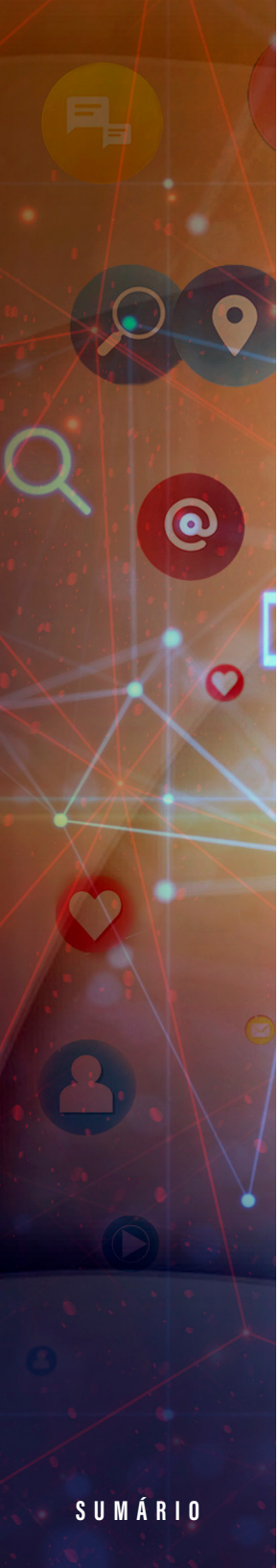
KNIGHT, Steve. Early learnings from the national library of New Zealand's national digital heritage archive project. **Program**, v. 44, n. 2, p. 85-97, 2010.

KOBAYASHI, Mei; TAKEDA, Koichi. Information retrieval on the web. **ACM Computing Surveys** (CSUR), v. 32, n. 2, p. 144-173, 2000.

KOHLER, Wallace. A longitudinal study of *web* pages continued: a consideration of document persistence. **Information Research**, v. 9, n. 2, 2004. Disponível em: https://www.informationr.net/ir/9-2/paper174.html. Acesso em: 30 out. 2022.

KRÓL, Karol; ZDONEK, Dariusz. Peculiarity of the bit rot and link rot phenomena. **Global Knowledge, Memory and Communication**, v. 69, n. 1/2, p. 20-37, 2020.

KSHETRI, Nir. *Web* 3.0 and the metaverse shaping organizations' brand and product strategies. **IT Professional**, v. 24, n. 2, p. 11-15, 2022.



KUNY, Terry. A digital dark ages? Challenges in the preservation of electronic information of electronic information. *In:* IFLA COUNCIL AND GENERAL CONFERENCE, 63., 1997, Copenhagen. **Proceedings** [...]. Copenhagen, 1997.

LAITANO, Bruno Grigoletti. **Digitalizar o arquivo, arquivar o digital**: a história e suas fontes diante das velhas e novas tecnologias. Porto Alegre, 2021. Dissertação (Programa de Pós-Graduação em História.) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021.

LARIVIÈRE, Jules. **Guidelines for legal deposit legislation**. United Nations, 2000. Disponível em: http://unesdoc.unesco.org/images/0012/001214/121413eo.pdf. Acesso em: 20 mar. 2023.

LEETARU, Kalev. How much of the Internet does The Wayback Machine really archive? **Forbes**. Disponível em: https://www.forbes.com/*sites*/kalevleetaru/2015/11/16/how-much-of-the-Internetdoes-the-wayback-machine-really-archive/#6bc039f59446. Acesso em: 20 jan. 2023.

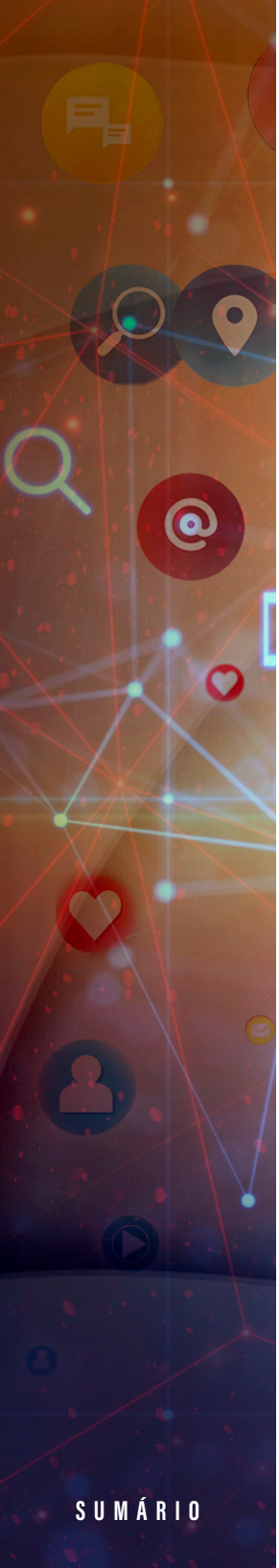
LEUNG, Shun-Tak A.; PERL, Sharon; STATA, Raymie; WIENER, Janet L. Towards *web*-scale *web* archaeology. **SCR Research Report**, v. 174, p. 1-5, 2001.

LEVY, Steven. **Hackers**: heroes of the computer revolution. Garden City: Anchor Press/ Doubleday, 1984.

LINS, Luziane. **Projeto de lei N. 2.431/2015**. Dispõe sobre o patrimônio público digital institucional inserido na rede mundial de computadores e dá outras providências. Disponível em: https://www.camara.leg.br/proposicoes*web*/prop_ mostrarintegra?codteor=1363213. Acesso em: 10 fev. 2023.

LOR, Peter Johan; BRITZ, Johannes J. An ethical perspective on political-economic issues in the long-term preservation of digital heritage. **Journal of the American Society for Information Science and Technology**, v. 63, n. 11, p. 2153-2164, 2012.

LUZ, Ana Javes Andrade da. **Comunicação pública e memória comunicacional**: revelações e apagamentos sobre o governo da presidenta Dilma Rousseff. Porto Alegre, 2021. 253 f. Tese (Doutorado em Comunicação e Informação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021. Disponível em: http://hdl.handle.net/10183/235405. Acesso em: 19 fev. 2023.



LUZ, Ana Javes Andrade da. **Comunicação pública e memória das cidades**: a preservação dos sistemas de comunicação nos *site*s das capitais brasileiras. Porto Alegre, 2016. Dissertação (Mestrado em Comunicação e Informação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

LYMAN, Peter; KAHLE, Brewster. Archiving digital cultural artifacts: organizing an agenda for action. **D-Lib Magazine**, v. 4 , n. 7 , julho 1998.

LYONS, Bertram. There will be no digital dark age. **Issues and Advocacy**, may 2016. Disponível em: https://issuesandadvocacy.wordpress.com/2016/05/11/there-will-be-no-digital-dark-age/. Acesso em: 24 jan. 2019.

MAEMURA, Emily *et al*. If these crawls could talk: Studying and documenting *web* archives provenance. **Journal of the Association for Information Science and Technology**, v. 69, n. 10, p. 1223–1233, 2018.

MAEMURA, Emily; BECKER, Christoph; MILLIGAN, Ian. Understanding computational *web* archives research methods using research objects. *In:* IEEE INTERNATIONAL CONFERENCE ON BIG DATA, Washington, DC, 2016. **Proceedings** [...]. p. 3250-3259, 2016. DOI 10.1109/BigData.2016.7840982.

MAJOR, Daniela. The Problem of *web* Ephemera. *In*: GOMES, Daniel; *et al*. **The Past *web***: Exploring *web* Archives. Cham: Springer International Publishing, p. 5-10, 2021.

MARCOS RECIO, Juan Carlos; SÁNCHEZ VIGIL, Juan Miguel; OLIVERA ZALDÚA, Maria. La conservación de todos los contenidos digitales no es necesaria guardemos solo lo imprescindible de los medios. **Ibersid**: Revista de Sistemas de Información y Documentación, v. 13, n. 2, p. 31-38, 2019. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=7112011. Acesso em: 02 dez. 2022.

MÁRDERO ARELLANO, Miguel Ángel. Critérios para a preservação digital da informação científica. Brasília, 2008. 356 f. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília, Brasília, 2008.

MÁRDERO ARELLANO, Miguel Ángel. Preservação de acervos digitais em repositórios institucionais. *In*: ENCONTRO DA REDE SUDESTE DE REPOSITÓRIOS INSTITUCIONAIS, 1., 2019, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: Fiocruz/Icict/UFRJ, 2019.

MÁRDERO ARELLANO, Miguel Ángel; SANTOS, Gildenir Carolino (org.). **Bibliografia sobre preservação digital:** um levantamento nos diversos suportes informacionais. Campinas: BCCL/UNICAMP, 2021. DOI: https://doi.org/10.20396/ISBN9786588816110.

MARTINS, Marina R. **Mapeamento de públicos estratégicos para iniciativas brasileiras universitárias de arquivamento da *web* no âmbito acadêmico: uma projeção para a Universidade Federal do Rio Grande do Sul e seu Programa de Pós-graduação em Comunicação**. Porto Alegre, 2019. Dissertação (Mestrado em Comunicação e Informação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

MARKWELL, John; BROOKS, David W. Broken links: The ephemeral nature of educational WWW hyperlinks. **Journal of Science Education and Technology**, v. 11, p. 105-108, 2002.

MARTINS, Marina R.; ROCKEMBACH, Moisés. Promoção de iniciativas de arquivamento da web: um estudo a partir da rede de públicos estratégicos da UFRGS. **AtoZ**: novas práticas em informação e conhecimento. Curitiba. v. 8, n. 2, p. 99-105, 2019.

MARTINS, Marina R., ROCKEMBACH, Moisés. Mapeamento de públicos para iniciativas acadêmicas de arquivamento da web. **Intercom**: Revista Brasileira de Ciências da Comunicação, v. 43, n. 1, p. 71-88, 2020. DOI: https://doi.org/10.1590/1809-5844202014.

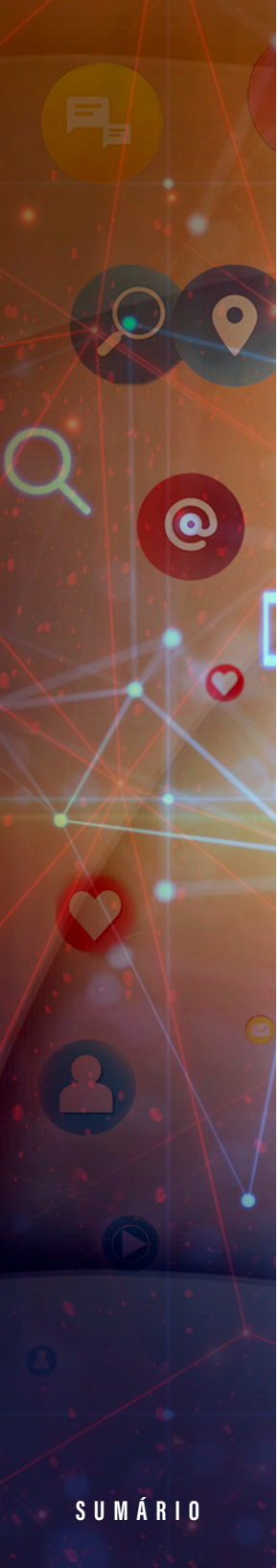
MASANÉS, Julien. Selection for *web* Archives. *In*: MASANÉS, Julien. ***Web* Archiving**. Berlin: Springer, Heidelberg, 2006. p. 71-90.

MASANÉS, Julien. *Web* archiving methods and approaches: a comparative study. **Library Trends**, v. 54, n. 1, p. 72-90, Summer 2005.

MELO, Jonas Ferrigolo. **Arquivamento dos *website*s do governo federal brasileiro**: preservação do domínio GOV.BR. Porto Alegre, 2020. 133 f. Dissertação (Mestrado em Comunicação e Informação) – Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2020. Disponível em: http://hdl.handle.net/10183/210671. Acesso em: 23 abr. 2022.

MELO, Jonas Ferrigolo; ROCKEMBACH, Moises. Arquivabilidade de *website*s para preservação digital: estudo a partir da área da saúde. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, v. 14, n. 3, p. 529-545, jul./set. 2020. Disponível em: http://hdl.handle.net/10183/216041. Acesso em: 03 dez. 2022.

MELO, Jonas Ferrigolo; ROCKEMBACH, Moises. International initiatives and advances in Brazil for government *web* archiving. *In*: INTERNATIONAL CONFERENCE ON DATA AND INFORMATION IN *ON-LINE*. Springer, Cham, 2021. p. 83-95.



MELO, Jonas Ferrigolo; ROCKEMBACH, Moises. Public policies for governmental web archiving in Brazil. **International Internet Preservation Consortium General Assembly and Web Archiving Conference**, Hilversum, The Netherlands, 2023. Disponível em: https://netpreserve.org/ga2023/programme/wac/ Acesso em: 27 abr. 2023.

MELO, Jonas Ferrigolo; OLIVEIRA, Carolina; ROCKEMBACH, Moisés. História do arquivamento da web no Brasil: um percurso entre a academia, o legislativo e executivo brasileiro. *In*: SIMPOSIO DA HISTÓRIA DOS ARQUIVOS E DA ARQUIVOLOGIA, 2023, Niterói, Rio de Janeiro. **Anais** [...]. 2023. DOI 10.31235/osf.io/u4ydh. Disponível em: https://doi.org/10.31235/osf.io/u4ydh Acesso em: 10 jun. 2023.

MIESSLER, Daniel. **The Real Difference Between a URL and URI**. Disponível em: https://danielmiessler.com/study/difference-between-uri-url/. Acesso em: 27 abr. 2023.

MILLIGAN, Ian; RUESST, Nick; LIN, Jimmy. Content selection and curation for *web* archiving: The gatekeepers vs. the masses. *In*: ACM/IEEE-CS ON JOINT CONFERENCE ON DIGITAL LIBRARIES, 16., 2016. **Proceedings** [...]. 2016. p. 107-110. DOI: https://doi.org/10.1145/2910896.2910913.

MILLIGAN, Ian. Exploring *web* archives in the age of abundance: a social history case study of GeoCities *In*: BRÜGGER, Niels; MILLIGAN, Ian (org.). **The SAGE handbook of *web* history**. SAGE Publications Limited, 2019. p. 344-358.

MILLIGAN, Ian. **Welcome to the web**: The online community of GeoCities during the early years of the World Wide Web. 2017. Disponível em: http://hdl.handle.net/10012/11859. Acesso em: 02 dez. 2022.

MINOW, Mary. **Digital preservation and copyright by Peter Hirtle**. 2003. Disponível em: http://fairuse.stanford.edu/2003/11/10/digital_preservation_and_copyr/. Acesso em: 06 fev. 2023.

MOLINARI, William. **Desconstruindo a *web***: as tecnologias por trás de uma requisição. São Paulo: Casa do Código, 2016.

MOTTL, Judy. Internet pioneer Vint Cerf warns of bit rot and digital dark age but don't panic yet. **Tech Times**, 16 fev. 2015. Disponível em: https://goo.gl/Wmo4nu. Acesso em: 24 jan. 2019.

NAZAROVETS, Serhi; KULYK, Yevheniia. Library 4.0: next generation services and technologies. **Bìblìotečnij vìsnik**, v. 5, p. 3-14, jan. 2017.

NELSON, Theodor Holm. Complex information processing: a file structure for the complex, the changing and the indeterminate. *In*: ACM´65 NATIONAL CONFERENCE, 20., 1965. **Proceedings** [...]. 1965. p. 84-100. DOI: https://doi.org/10.1145/800197.806036.

NICODEMO, Thiago Lima; ROTA, Alesson Ramon; MARINO, Ian Kisil. **Caminhos da História digital no Brasil**. Milfontes: Vitória, 2022.

NIELSEN, Jakob. Fighting linkrot, Nielsen Norman Group. 1998 Disponível em: www.nngroup.com/articles/fighting-linkrot/. Acesso em: 02 dez. 2022.

NIU, Jinfang. An Overview of Web Archiving. **D-Lib Magazine**, v. 18, n. 3-4, mar./apr. 2012. DOI 10.1045/march2012-niu1.

NOH, Younghee. Imagining Library 4.0: creating a model for future libraries. **The Journal of Academic Librarianship**, v. 41, n. 6, p. 786-797, 2015.

NTOULAS, Alexandros; CHO, Junghoo; OLSTON, Christopher. What's new on the *web*? The evolution of the *web* from a search engine perspective. *In*: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB. 13., 2004. New York. **Proceedings** [...]. 2004. p. 1-2. DOI: https://doi.org/10.1145/988672.988674

NUNES, Lucia N.; ROCKEMBACH, Moisés. Análise de conteúdo de termos de uso e políticas de privacidade de arquivos da web. *In*: FORUM DE ESTUDOS EM INFORMACAO, SOCIEDADE E CIENCIA, 4., 2021, Porto Alegre. **Resumos** [...]. Porto Alegre: PPGCIN, 2021. p. 188-195. Disponível em: http://hdl.handle.net/10183/232333. Acesso em: 20 abr. 2022.

NUNES, Lucia N. O. **Arquivamento da *web***: aspectos éticos e legais no acesso e uso da informação. Porto Alegre, 2021. 162 f. Dissertação (Mestrado em Comunicação e Informação) - Faculdade de Biblioteconomia e Comunicação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2022.

NYHAN, Julianne; PASSAROTTI; Marco. **One Origin of Digital Humanities**: Fr Roberto Busa in His Own Words. London: Springer, 2019.

ODERSKY, Martin; MOORS, Adriaan. Fighting bit rot with types (experience report: scala collections). *In*: IARCS ANNUAL CONFERENCE ON FOUNDATIONS OF SOFTWARE TECHNOLOGY AND THEORETICAL COMPUTER SCIENCE. **Leibniz International Proceedings in Informatics** [...], v. 4, 2009. p. 427-451. DOI: http://dx.doi.org/10.4230/LIPIcs.FSTTCS.2009.2338.

OGDEN, Jessica; HALFORD, Susan; CARR, Leslie. Observing *web* archives: The case for an ethnographic study of *web* archiving. *In*: ACM ON WEB SCIENCE CONFERENCE. 2017, New York. **Proceedings** [...]. New York: Association for Computing Machinery, 2017. p. 299-308. DOI: https://doi.org/10.1145/3091478.3091506.

OGDEN, Jessica. Everything on the Internet can be saved: archive team, Tumblr and the cultural significance of *web* archiving. **Internet Histories**, v. 6, n. 1/2, 2022.

OGDEN, Jessica. **Saving the *web*: facets of *web* archiving in everyday practice**. Southampton, 2020. Thesis (Doctoral Dissertation) – University of Southampton, Southampton, 2020.

O'REILLY, Tim. What is Web 2.0: Design patterns and business models for the next generation of software. **Communications & strategies**, n. 1, p. 17, 2007.

PANOS, Patrick. Technotes: the Internet archive: an end to the digital dark age. **Journal of Social Work Education**, v. 39, n. 2, p. 343-347, 2003.

PEHLIVAN, Zeynep; THIÉVRE, Jérôme; DRUGEON, Thomas. Archiving social media: the case of Twitter. GOMES, Daniel *et al*. **The Past *web***: Exploring *web* Archives. Cham: Springer International Publishing, 2021. p. 43-56.

PENNOCK, Maureen. ***Web-archiving***: DPC Technology Watch Report 13-01 March 2013. Inglaterra: DPC, 2013. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.5280&rep=rep1&type=pdf. Acesso em: 15 jan. 2023.

PIMENTA, Ricardo M. Os objetos técnicos e seus papéis no horizonte das Humanidades digitais: um caso para a Ciência da Informação. **Revista Conhecimento em Ação**, v. 1, n. 2, p. 20-33, 2016.

PIMENTA, Ricardo M. Fontes demais, tempo de menos: uma breve crítica à escalada tecnoinformacional para a escrita da História do tempo presente *In*: NICODEMO, Thiago Lima; ROTA, Alesson Ramon; MARINO, Ian Kisil (ed.). **Caminhos da História digital no Brasil**. Ed. Milfontes: Vitória, 2022.

POST, Colin. Building a living, breathing archive: a review of appraisal theories and approaches for *web* archives. **Preservation, Digital Technology & Culture**, v. 46, n. 2, p. 69-77, 2017.

RAUBER, Andreas; KAISER, Max; WACHTER, Bernhard. Ethical issues in *web* archive creation and usage–towards a research agenda. 2008. Disponível em: fhttps://www.researchgate.net/publication/228638059_Ethical_Issues_in_Web_Archive_Creation_and_Usage_-_Towards_a_Research_Agenda. Acesso em: 04 abr. 2023.

RECIO, Juan Carlos Marcos; VIGIL, Juan Miguel Sánchez; ZALDÚA, María Olivera. La conservación de todos los contenidos digitales no es necesaria: guardemos solo lo imprescindible de los medios. **Ibersid:** Revista de Sistemas de Información y Documentación, v. 13, n. 2, p. 31-38, 2019.

REDKINA, Natalya. S. Global trends in library *web*-archives. **Scientific and Technical Libraries,** v. 1, n. 1, 99-114, 2021.

REYES AYALA, Brenda; PHILLIPS, Mark Edward; KO, Lauren. **Current quality assurance practices in *web* archiving**. *web*Archiving@UNT, 2014. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc333026/. Acesso em: 22 fev. 2023.

REZENDE, Laura Vilela Rodrigues. Caracterização de repositórios digitais Dataverse conforme o Modelo OAIS. **Revista Brasileira de Preservação Digital**, Campinas, SP, v. 3, e022011, 2012. DOI 10.20396/rebpred.v3i00.16581.

ROCKEMBACH, Moisés. Arquivos da web como fonte de pesquisa em humanidades digitais. *In*: NICODEMO, Thiago Lima; ROTA, Alesson Ramon; MARINO, Ian Kisil (ed.). Caminhos da História digital no Brasil. Milfontes: Vitória, 2022.

ROCKEMBACH, Moisés; SERRANO, Anabela. Climate change and web archives: an Ibero-American study based on the Portuguese and Brazilian contexts. Records Management Journal, Bingley, v. 31, n. 3, p. 222-239, 2021.

ROCKEMBACH, Moisés. A *web* brasileira na Covid-19: arquivamento da *web* e preservação digital. **Liinc em Revista**. v. 17, n. 1, 2021a.

ROCKEMBACH, Moises. Brazilian Election Web Archive. *In*: INTERNATIONAL INTERNET PRESERVATION CONSORTIUM GENERAL ASSEMBLY AND WEB ARCHIVING CONFERENCE. 2021. **Proceedings** [...]. 2021b. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc1827574/ Acesso em: 06 abr. 2023.

ROCKEMBACH, Moisés. Arquivamento da Web no contexto das humanidades digitais: da produção à preservação da informação digital. **LIINC em revista**. Rio de Janeiro, RJ. v. 15, n. 1, p. 131-139, 2019.

ROCKEMBACH, Moisés; MELO, Jonas Ferrigolo. Brazilian Web Archive. *In*: INTERNATIONAL INTERNET PRESERVATION CONSORTIUM GENERAL ASSEMBLY AND WEB ARCHIVING CONFERENCE. 2019, Zagreb, Croatia. **Proceedings** [...]. 2019. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc1609000/ Acesso em: 06 abr. 2023.

ROCKEMBACH, Moisés. Arquivamento da *web*: estudos de caso internacionais e o caso brasileiro. **RDBCI:** Revista Digital de Biblioteconomia e Ciência da Informação, v. 16, n. 1, p. 7-24, 2018.

ROCKEMBACH, Moisés; PAVÃO, Caterina M. Groposo. Políticas e tecnologias de preservação digital no arquivamento da web. Revista Ibero-Americana de Ciência da Informação, v. 11, n. 1, p. 168-182, 2018.

ROCKEMBACH, Moisés. Inequalities in digital memory: ethical and geographical aspects of *web* archiving. **The International Review of Information Ethics**, v. 26, p. 138-149, 2017. Disponível em: https://informationethics.ca/index.php/irie/article/view/286/284. Acesso em: 06 abr. 2023.

RODRIGUES, Vander L. D.; ROCKEMBACH, Moisés. Arquivos da web como fonte historiográfica. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 19 Campinas, 2021.

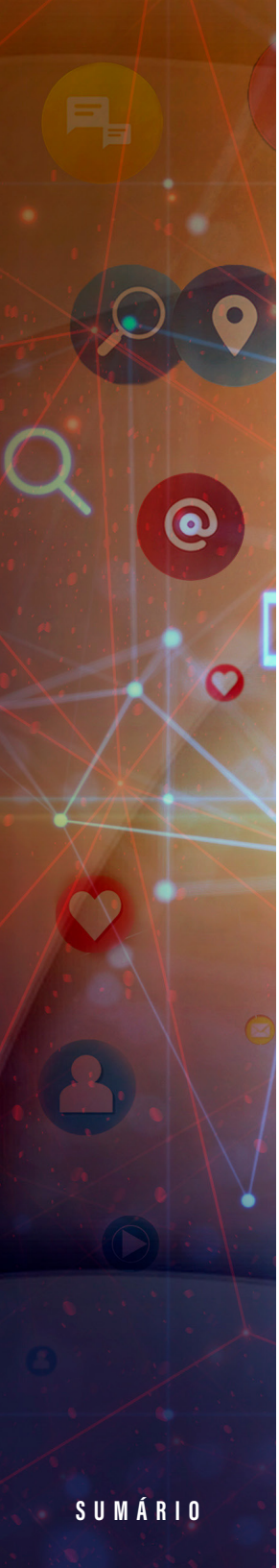
ROGERS, Richard. O fim do virtual: os métodos digitais. **Lumina**. v. 10, n. 3, 2016.

ROUSE, Margaret. **Definition bit rot**. TechTarget, 2019. Disponível em: https://searchstorage.techtarget.com/definition/bit-rot. Acesso em: 25 jan. 2023.

RUEST, Nick, *et al*. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. *In*: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES. 2020, New York. **Proceedings** [...]. Association for Computing Machinery: New York, 2020. p. 157–166. DOI: https://doi.org/10.1145/3383583.3398513

SALAHELDEEN, Hany; NELSON, Michael L. Losing my revolution: how many resources shared on social media have been lost? *In*: THEORY AND PRACTICE OF DIGITAL LIBRARIES: INTERNATIONAL CONFERENCE. 2., Heidelberg, 2012. **Proceedings** [...]. Springer: Berlin, 2012. p. 125–137.

SAMOUELIAN, Mary; DOOLEY, Jackie. **Descriptive metadata for *web* archiving**: review of harvesting tools. Dublin: OCLC Research, 2018.



SANTOS, Gildenir Carolino. Ensaio sobre arquivamento de páginas *web*: foco na experiência do Portal de Periódicos da UNICAMP, utilizando o Conifer (Rhizome). *In*: SIMPÓSIO INTERNACIONAL SOBRE PRESERVAÇÃO DIGITAL. 5.; 2021; Campinas, SP. **Resumos** [...]. Campinas, SP: SBU/UNICAMP; IBICT, 2021.

SANTOS, Gildenir Carolino; FORMENTON, Danilo; TERRADA, Gabriela Ayres Ferreira. Modelo de arquivamento de páginas *web* para Portais de Periódicos: um relato de pesquisa no Portal de Periódicos da UNICAMP. **Revista Brasileira de Preservação Digital**, Campinas, v. 3, p. e022001, 2022.

SANTOS, Vanderlei Batista. Arquivamento *web*: legislação correlata. **Revista Brasileira de Preservação Digital**, v. 1, 2020. Disponível em: http://hdl.handle.net/20.500.11959/brapci/158975. Acesso em: 06 abr. 2023.

SCHAFER, Valérie; WINTERS, Jane. The values of web archives. **International Journal of Digital Humanities,** v. 2, p. 1-3, 2021. Disponível em: https://doi.org/10.1007/s42803-021-00037-0 Acesso em: 07 abr. 2023.

SCHLIEDER, Christoph. Digital heritage: semantic challenges of long-term preservation, **Semantic *web*,** v. 1, n. 1/2, p. 143-147, 2010.

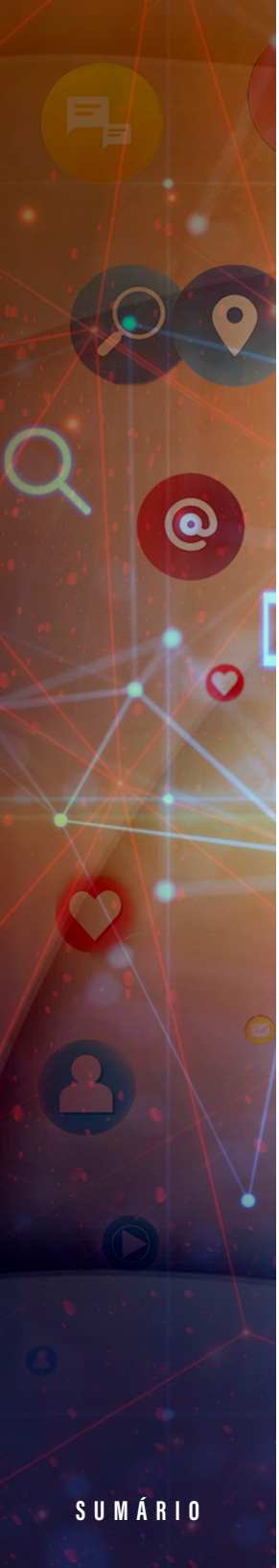
SHIOZAKI, Ryo; EISENSCHITZ, Tamara. Role and justification of *web* archiving by national libraries: a questionnaire survey. **Journal of Librarianship and Information Science**, v. 41, n. 2, p. 90–107, 2009.

SCHNEIDER, Steven M.; FOOT, Kristen A.; WOUTERS, Paul. *Web* archiving as e-research. *In*: **E-Research**, Routledge: Londres, 2010. p. 221-237.

SIFE, Alfred; BERNARD, Ronald. Persistence and decay of *web* citations used in theses and dissertations available at the Sokoine National Agricultural Library, Tanzania. **International Journal of Education and Development Using ICT**, v. 9, n. 2, ago. 2013. p. 85-94. Disponível em: file:///C:/Users/User/Downloads/article_130281.pdf. Acesso em 15 ago. 2022.

SUMMERS, Ed. Appraisal talk in *web* archives. **Archivaria**, v. 89, p. 70-103, May 2020. Disponível em: https://archivaria.ca/index.php/archivaria/article/view/13733. Acesso em: 16 fev. 2023.

TAIT, Elizabeth *et al*. Linking to the past: an analysis of community digital heritage initiatives. **Aslib Proceedings:** New Information Perspectives, v. 65, n. 6, p. 564-580, 2013.



TAYLOR, Mary K.; HUDSON, Diane. Linkrot and the usefulness of web site bibliographies. **Reference and User Services Quarterly**, v. 39, n. 3, p. 273-277, 2000.

TAYLOR, Nicholas. Understanding legal use cases for *web* archives. *In*: IIPC WEB ARCHIVING CONFERENCE. 2017, Lisboa. **Proceedings** [...]. 2017. Disponível em: https://nullhandle.org/pdf/2017-06-16_understanding_legal_use_cases_for_web_archives.pdf. Acesso em: 02 mar. 2023.

TERRADA, Gabriela Ayres Ferreira. **Preservação digital da *web***: uma reflexão sobre políticas e práticas. Niterói, 2022. Dissertação (Mestrado em Ciência da Informação) - Instituto de Arte e Comunicação Social. Universidade Federal Fluminense, Niterói, 2022. Disponível em: http://app.uff.br/riuff/handle/1/26276. Acesso em: 01 nov. 2022.

TERRAS, Melissa. Disciplined: using educational studies to analyse "humanities computing". **Literary and Linguistic Computing**, v. 21, n. 2, p. 229-246, 2006.

UNESCO. **Charter on the preservation of digital heritage**. 2003. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000229034. Acesso em: 23 fev. 2023.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. Conselho Universitário. **Resolução N° 064/2021, de 19 de março de 2021**. Aprovar a Política de Preservação de Acervos Digitais da Universidade Federal do Rio Grande do Sul. Porto Alegre: Conselho Universitário, 2021. Disponível em: https://www.ufrgs.br/consun/legislacao/resolucao-no-064-2021/. Acesso em: 20 fev. 2023.

VALENTE, Mariana. Direito autoral e plataformas de Internet: um assunto em aberto. INTERNETLAB, 18/04/2019. Disponível em: https://Internetlab.org.br/pt/especial/direito-autoral-e-plataformas-de-Internet-um-assunto-em-aberto/. Acesso em: 04 abr. 2023.

VALENTE, Mariana. Preservação do conteúdo *web* brasileiro. *In:* FÓRUM DA INTERNET DO BRASIL, 9., 2019, Manaus. **Anais** [...]. Manaus, 2019. Disponivel em: https://www.youtube.com/watch?v=y_k-k0c254Y. Acesso em: 20 fev. 2023.

VALENTE, Mariana; HOUANG, André. **O que você precisa saber sobre licenças CC.** 2020. Disponível em: https://br.creativecommons.net/wpcontent/uploads/*sites*/30/2021/02/CartilhaCCBrasil.pdf. Acesso em: 16 jan. 2023.

VAN DICK, José; POELL, Thomas; DE WAAL, Martijn. **The platform society:** Public values in a connective world. Oxônia: Oxford University Press, 2018.

VLASSENROOT, Eveline *et al*. *Web*-archiving and social media: an exploratory analysis: Call for papers digital humanities and *web* archives–A special issue of international journal of digital humanities. **International Journal of Digital Humanities**, v. 2, n. 1-3, 107-128, 2021. DOI 10.1007/s42803-021-00036-1

VELTE, Ashlyn. Ethical challenges and current practices in activist social media archives. **American Archivist**, v. 81, n. 1, p. 112-134, 2018.

VENLET, Jessica *et al*. **Descriptive metadata for *web* archiving**: literature review of user needs. Dublin: OCLC, 2018.

VIRGIL, Johnny. A Biblioteca de Babel: uma metáfora para a sociedade da informação. **DataGramaZero**-Revista de Ciência da Informação. Rio de Janeiro, v. 8, n. 4, 2007.

*WEB*STER, Peter. Users, technologies, organisations: towards a cultural history of world *web* archiving. *In*: BRÜGGER, Niels *et al*. ***Web* 25**: histories from 25 years of the *World Wide Web*. New York: Peter Lang, 2017. p. 179-190.

WINTERS, Jane. Giving with one click, taking with the other: e-legal deposit, *web* archives and researcher access. *In*: GOODING, Paul; TERRAS, Melissa. **Electronic Legal Deposit**: shaping the library collections of the future. London: Facet Publishing, 2020. p. 159-178.

ZAFALON, Zaira Regina; NÓBREGA DE SÁ, Mariana. Mundaneum e Biblioteca Digital Mundial: relações possíveis? **EmQuestão**, Porto Alegre, v. 25, p. 216-242, 2019. DOI 10.19132/1808-5245250.216-242

ZHOU, Ke *et al*. No more 404s: predicting referenced link rot in scholarly articles for pro-active archiving. *In*: PROCEEDINGS OF THE 15TH ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES. **Proceeding** [...]. 2015. p. 233-236.

ZITTRAIN, Jonathan; ALBERT, Kendra; LESSIG, Lawrence. Perma: scoping and addressing the problem of link and reference rot in legal citations. **Legal Information Management**, v. 14, n. 2, p. 88-99, 2014.

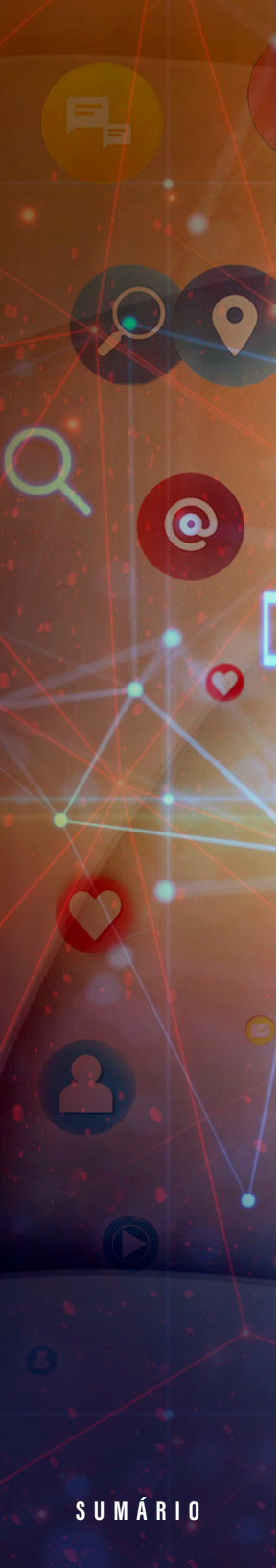
SOBRE O AUTOR E A AUTORA

Moisés Rockembach

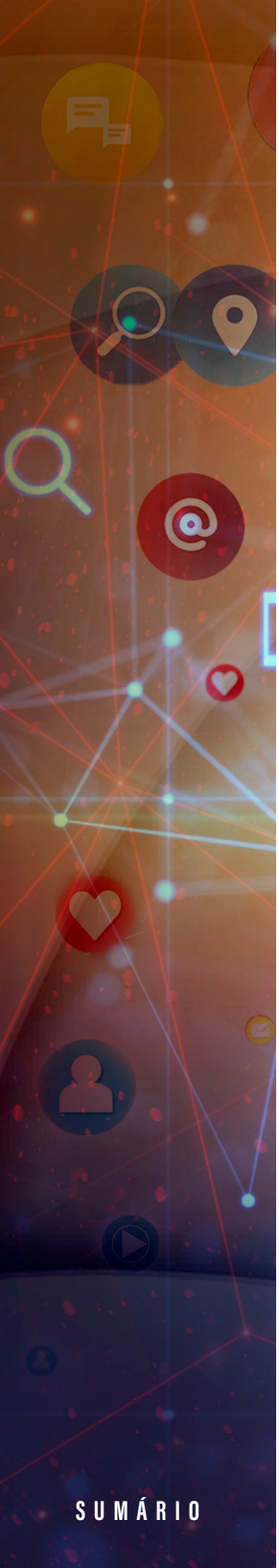
Professor de Ciência da Informação da Universidade de Coimbra (Portugal), e professor dos Programas de Pós-Graduação em Ciência da Informação (PPGCIN) e em Comunicação (PPGCOM) da Universidade Federal do Rio Grande do Sul. Bacharel em Arquivologia, Mestre em Comunicação e Informação (UFRGS) e Doutor em Informação e Comunicação em Plataformas Digitais (Universidade do Porto e Universidade de Aveiro - Portugal), com Pós-Doutorado na Universidade do Porto. É pesquisador do CITCEM – Centro de Investigação Transdisciplinar para Cultura, Espaço e Memória (Universidade do Porto / Portugal) e do InterPARES Trust AI – Inteligência Artificial (University of British Columbia / Canadá). Atuou como professor visitante na KU Leuven (Bélgica), envolvido em atividades do KU Leuven Digital Society Institute e do Meaningful Interactions Lab (Mintlab). É líder do Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital (NUAWEB / UFRGS - CNPq).

Caterina Groposo Pavão

Professora de Biblioteconomia da Faculdade de Biblioteconomia e Comunicação, e professora do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio Grande do Sul. Bacharel em Biblioteconomia (UFRGS), Mestre e Doutora em Comunicação e Informação (UFRGS). É pesquisadora dos Grupos de Pesquisas: Comunicação Científica (UFRGS-CNPq), Estudos e Práticas de Preservação Digital (CARINIANA/IBICT-CNPq) e Laboratório do Ecossistema da Pesquisa Científica Brasileira (LaEPeCBr/IBICT-CNPq). É líder do Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital (NUAWEB/UFRGS - CNPq). É Diretora do Centro de Documentação e Acervo Digital da Pesquisa (CEDAP/FABICO-UFRGS). É membro do grupo RDP Brasil – Acesso Aberto a Dados de Pesquisa (CEDAP/FABICO).



ÍNDICE REMISSIVO



www.PIMENTACULTURAL.com

ARQUIVAMENTO DA WEB E PRESERVAÇÃO DIGITAL

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES
PPGCIN
pimenta cultural