



Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação

ISSN: 1518-2924

adilson.pinto@ufsc.br

Universidade Federal de Santa Catarina
Brasil

Pinheiro FARIAS, Juliana; Vieira BOMFIM, Kelen Cândida
A produção científica sobre preservação de websites em língua portuguesa
Encontros Bibli: revista eletrônica de biblioteconomia e
ciência da informação, vol. 24, núm. 55, 2019, Maio-, pp. 1-15
Universidade Federal de Santa Catarina
Brasil

DOI: <https://doi.org/10.5007/1518-2924.2019.e58207>

Disponível em: <https://www.redalyc.org/articulo.oa?id=14763092004>

- Como citar este artigo
- Número completo
- Mais informações do artigo
- Site da revista em redalyc.org

UAEM redalyc.org

Sistema de Informação Científica Redalyc
Rede de Revistas Científicas da América Latina e do Caribe, Espanha e Portugal
Sem fins lucrativos acadêmica projeto, desenvolvido no âmbito da iniciativa
acesso aberto



Encontros Bibli


A PRODUÇÃO CIENTÍFICA SOBRE PRESERVAÇÃO DE WEBSITES EM LÍNGUA PORTUGUESA

The scientific production on preservation of websites in Portuguese language

Juliana Pinheiro **Farias**

Bolsista do Instituto Brasileiro de Informação em Ciência e Tecnologia, graduada em Arquivologia.


Juliana.pf64@gmail.com

<https://orcid.org/0000-0002-9339-7953> 

Kelen Cândida Vieira **Bomfim**

Bolsista do Instituto Brasileiro de Informação em Ciência e Tecnologia, graduada em Biblioteconomia.

kv.bomfim@hotmail.com

<https://orcid.org/0000-0002-1825-0097> 

RESUMO

Objetivo: O presente artigo apresenta uma revisão de literatura a respeito das produções científicas sobre preservação de websites em língua portuguesa, entre o período de 2000 a 2018.

Métodos: Trata-se de um levantamento e análise de dados quantitativos e qualitativos advindos de periódicos científicos que tratam da temática "Preservação de websites" e "Arquivamento da web", com a aplicação da seleção bibliografias voltadas ao tema. Para o desenvolvimento da pesquisa foi feita uma busca na base Oasisbr.

Resultados: Em tabelas foram reunidas os principais métodos e tecnologias que tem sido objeto de estudo para a preservação e arquivamento da Web. Os resultados da pesquisa demonstraram que os quantitativos de produções científicas são escassos.

Conclusões: Com a escassez da produção científica voltada para preservação e arquivamento Web, este trabalho vem auxiliar os pesquisadores e trazer um panorama de como está se comportando a produção científica até o momento.

PALAVRAS-CHAVE: Arquivamento da Web. Preservação Digital de websites. Preservação digital de páginas web. Preservação de websites. Softwares de preservação

ABSTRACT

Objective: This article presents a review of the literature on scientific production on the preservation of websites in Portuguese language, between the period of 2000 and 2018.

Methods: This is a survey and analysis of quantitative and qualitative data from scientific journals dealing with of the theme "Preservation of websites" and "Web archiving", with the application of selection bibliographies focused on the theme. For the development of the research, a search was made at the base Oasisbr.

Results: Tables were gathered the main methods and technologies that have been object of study for the preservation and archiving of the Web. The results of the research demonstrated that the quantitative of scientific productions are scarce.

Conclusions: With the scarcity of scientific production aimed at preservation and archiving Web, this work will help the researchers and bring a panorama of how the scientific production is behaving up to the moment.

KEYWORDS: Web Archiving. Digital Preservation of websites. Digital preservation of web pages. Preservation of websites. Preservation software

1 INTRODUÇÃO

A década de 1990 foi marcada pela expansão das Tecnologias da Informação e Comunicação (TICs), que impactaram a produção de informação, conhecimento e possibilitaram que grandes volumes de informações fossem criados e divulgados em curto espaço de tempo. A construção da rede mundial de computadores, *internet*, impactou os processos de disseminação e acesso à informação, por meio da sua interatividade, velocidade e praticidade.

O grande volume de dados e informações produzidas faz com que a *internet* seja considerada um grande repositório de informação constituído por documento em variados formatos, sendo formada por diversas páginas *web* com conteúdos diversificados para atender e informar diferentes tipos de usuários permitindo o acesso às informações de forma ágil para a geração de conhecimento.

A procura por informação passou a ser facilitada pelo desenvolvimento das ferramentas de busca, como os motores de busca na *web* - *Google*, *Mozilla Firefox*, *Internet Explore*, *Opera* entre outras-, possibilitam a indexação e localização de páginas *web* e contribuem para o acesso ao conteúdo informacional.

A criação de páginas *web* voltadas para diversos seguimentos da sociedade e as redes sociais potencializaram a disseminação de informação no meio digital, porém, a grande maioria das páginas *web* e informações são alteradas ou perdidas em um curto espaço de tempo. O arquivamento da *web* e a preservação de páginas *web* passaram a ser discutidos como forma para evitar essas perdas. Segundo Rockembach (2017), entende-se por o arquivamento da *web*, o processo de coletar, armazenar e disponibilizar porções de informações advinda da *World Wide Web*, para que posteriormente possa ser acessado novamente e assim contribuir para a historiografia e construção da memória de instituições, pessoas e serviços.

Pesquisando sobre a temática de arquivamento da *web* no Brasil, esbarra-se no primeiro obstáculo, a falta de documentos/estudos que tratem desta temática. Sabe-se que há uma crescente no que diz respeito à produção de informação digital na *web*, então, é preciso começar a pensar qual a melhor forma de preservar a *World Wide Web*. Para tal, faz-se necessário a análise bibliográfica para avaliação do panorama nacional e internacional.

2 ARQUIVAMENTO E PRESERVAÇÃO DE PÁGINAS WEB/ WEBSITES

As páginas *web* são formadas por diversos formatos documentais com *hiperlinks*, áudios, imagens e vídeos. Manter o acesso futuro da produção documental passou a ser um questionamento para diversas áreas que lidam com a gestão e produção de informação. As perdas de páginas e a falta de acesso é algo que já pode ser observado com frequência um exemplo disso é quando se faz as referências bibliográficas em publicações e as páginas referenciadas com os respectivos *links* trazem resultados inexistentes.

Essa preocupação é levada às páginas *web* por conterem informações relevantes como os outros tipos documentais (BODÊ, 2008):

[...] não há motivos para não considerar páginas da *web* disponíveis na rede *Internet* - principalmente aquelas disponibilizadas através dos protocolos do tipo HTTP - como um documento com o mesmo status que um livro de biblioteca, uma carta histórica ou um relatório financeiro contábil em papel de uma grande empresa (BODÊ, 2008, p. 44).

A preservação digital visa proporcionar métodos para a salvaguarda de informação e documentação buscando resguardar a autenticidade, a integridade do conteúdo, a estrutura e o ambiente, como ressalta Márdero Arellano (2004) “A aplicação de estratégias de preservação para documentos digitais é uma prioridade, pois sem elas não existiria nenhuma garantia de acesso, confiabilidade e integridade dos documentos em longo prazo”.

O arquivo Nacional por meio da Política de Preservação Digital- AN Digital (ARQUIVO NACIONAL, 2016), que é uma portaria feita com o intuito de nortear os procedimentos de preservação digital e acesso aos documentos. A AN Digital traz uma relação de tipos de documentos que serão preservados como [...] texto estruturado com formatação, imagem matricial, imagem vetorial, áudio, audiovisual, mensagem de correio eletrônico, apresentação (*slides*), planilha e base de dados relacional. E esclarece que em momento futuro os tipos documentais mais complexos [...] como multimídia e páginas *web*, deverão ser também contemplados.

As páginas *web* são tipos documentais mais complexos e necessitam de estudos e construção de técnicas para a estruturação de mecanismos que possibilitem a preservação digital desse tipo documental e para que seu acesso seja garantido ao longo

tempo (ARQUIVO NACIONAL, 2016). A falta desses procedimentos fica evidente na Política de Preservação Digital- AN Digital.

O arquivamento web tem contribuído para a guarda e disponibilização futura das páginas web. E conforme conceitua Rockembach (2018) [...] compreende coletar, armazenar e disponibilizar a informação retrospectiva da *World Wide Web* para futuros pesquisadores. O arquivamento web é algo que deve ser feito de forma descentralizada e compartilhada, pois muitas vezes esse processo será realizado com o uso de diferentes servidores de armazenamento que disponibilizaram espaço para a guarda das páginas web, esse processo é feito no projeto Arquivo da Web Portuguesa – AWP (GOMES, 2010).

3 METODOLOGIA

Quanto aos procedimentos metodológicos, optou-se em apresentar uma revisão de literatura com abordagem quantitativa voltada para os elementos descritos nas tabelas acompanhados dos resultados advindos das bases de dados de periódicos científicos.

A pesquisa começou a ser desenvolvida pela delimitação inicial das bases que seriam analisadas, foi assim definido o escopo de realização do trabalho, onde seriam usadas bases que reúnem conteúdo de trabalhos científicos de países ibero-americanos, buscando-se artigos relacionados às temáticas previamente definidas: Arquivamento da web e Preservação digital de páginas web, Preservação digital de Websites.

Esse recorte foi estabelecido com o intuito de poder observar e analisar as produções científicas publicadas nas línguas portuguesa e espanhola usando os termos: Preservação Digital de websites (*Preservación digital de sítios web*), Preservação digital de páginas web (*Preservación digital de páginas web*), Arquivamento da Web (*Archivamiento Web*). Desse modo possibilitando a construção de um cenário relacionado às produções científicas das temáticas definidas.

A base delimitada para uso na pesquisa foi a do “Portal Brasileiro de Acesso Aberto a Publicações Científicas (OASISBR)”, por permitir acesso livre e aberto às produções científicas e por integrar outras bases como a *La Referencia* que tem o intuito de possibilitar visibilidade à produção científica das instituições de educação superior e pesquisa da América Latina e a RCAAP que faz parte do projeto Repositórios Científicos de Acesso Aberto de Portugal. (OASISBR, 2018):

OASISBR é um mecanismo de busca multidisciplinar que permite o acesso gratuito à produção científica de autores vinculados a universidades e institutos de pesquisa brasileiros. Por meio do Oasisbr é possível também realizar buscas em fontes de informação portuguesas. A partir de uma única interface, é possível realizar buscas simultâneas em revistas científicas, repositórios institucionais, repositórios temáticos, bibliotecas digitais de teses e dissertações e outras fontes de informação de natureza científica e tecnológica ou academicamente orientada. (OASISBR, 2018, P).

O período definido para a busca foi entre 2000 e 2018, a primeira data marca o início desse século e a data final delimita o ano de realização da pesquisa. E as buscas foram feitas e refinadas com o objetivo de encontrar teses, dissertações e artigos a respeito da temática.

4 ANÁLISE DOS TRABALHOS E DO CONTEÚDO

A pesquisa na base Oasisbr demonstrou que nos países como Brasil e Portugal os trabalhos voltados para Arquivamento da web e Preservação digital de páginas *Web/ Websites*, são quantitativamente escassos.

Para melhor compreensão dos trabalhos que estão sendo feitos na área, foi realizada uma análise dos conteúdos das teses, dissertações e artigos indexados na base Oasisbr. Essa análise foi feita com a leitura dos trabalhos e pela observação do desenvolvimento dos textos e como foram abordados os pontos relacionados à tecnologia, à metodologia, às iniciativas, às recomendações e ao contexto temporal dos trabalhos. Esses pontos foram selecionados por poderem trazer uma visão da contextualização do trabalho e das principais temáticas que estão sendo abordadas nos trabalhos voltados para preservação e arquivamento *web/ websites*.

O demonstrativo das teses, dissertações e artigos estão agrupados em tabelas, que reúnem a análise do conteúdo dos trabalhos e foram construídas com o objetivo de agrupar os principais temas tratados. As tabelas estão separadas em dois grupos, a primeira tabela reúne os conteúdos e informações referentes aos artigos que foram encontrados e a segunda tabela reúne os conteúdos e informações sobre as teses e dissertações.

4.1 Análise dos artigos

A primeira tabela demonstra as informações relacionadas aos pontos de análise: tecnologia, à metodologia, às iniciativas, às recomendações e ao contexto temporal dos trabalhos. Esses pontos foram usados para estruturar a observação e leitura dos artigos.

Tabela 1: Análise do conteúdo dos artigos.

Título do Trabalho	Tecnologia	Metodologia	Iniciativas	Recomendações	Contexto Temporal
Políticas e tecnologias de preservação digital no arquivamento da web.	. Software Heritrix; . Twarc; . ARC e WARC; . Open Wayback; . Plataformas Wayback Machine e time Travel.	Pesquisa quantitativa de caráter exploratório-descriptivo das publicações científicas a respeito do arquivamento da web. Coleta na base de dados Scopus.	. World Wide Web; . Consortium Internacional Internet Preservation Consortium (IIPC); . Internet Archive; . Arquivos e Bibliotecas Nacionais dos Estados Unidos, Reino Unido, Alemanha, Portugal; . Biblioteca da Catalunha e o projeto PADICAT; . Old Dominion University.	Para garantir a acessibilidade dos conteúdos publicados é necessária a definição de políticas de preservação e delimitação dos conteúdos que serão preservados e a implementação de mais iniciativas de preservação digital na web.	Pesquisa restrita aos anos de (2012-2016). Publicação do artigo 2018.
Arquivamento da web; Estudos de caso internacionais e o caso brasileiro.	. WaybackMachine	Pesquisa qualitativa exploratória-descriptiva. Faz a seleção de estudos de caso internacionais. Feita na base de dados Scopus.	. Grupo de trabalho Archive-It; . ISO 28500:2009; . Biblioteca Nacional Australiana-PANDORA; . Kulturw3; . Consórcio Internacional de Preservação da Internet; . The National Archives; . Biblioteca Catalunya; . Bibliothèque Nationale de France; . Harvard Library; Hanzo Archives Limited; . Projeto Memento (Estados Unidos); . Latin American Web Archiving Project.	Desenvolvimento de iniciativas e estudos voltados para o tema arquivamento na web e preservação da web.	Pesquisa restrita aos anos de (2012-2016). Publicação do artigo 2018.
Preservar a web: um desafio ao alcance de todos	. Heritrix; . NutchWAX e Wayback; . RARC	Pesquisa exploratória. Apresenta um estudo de preservação da web Portuguesa.	. Internet Archive (Estados Unidos); . Arquivo da Web Portuguesa – AWP (Preserva arquivos da Web Portuguesa).	É necessária disponibilização de serviços de preservação da web e que exista integração entre profissionais de informação (bibliotecários, arquivistas e documentalistas) para práticas de preservação.	Início dos trabalhos 2008. Publicação artigo no ano de 2010.
Arquivo e medição da web portuguesa.	. Heritrix; . Hadoop (processamento paralelo); . Wayback Machine; . NutchWAX (Adaptação do motor de busca); . RARC	Pesquisa exploratória. Apresenta um estudo de preservação da Web Portuguesa.	Arquivo da Web Portuguesa (AWP).	Para o sucesso do projeto AWP são necessários conhecimentos de diferentes áreas, com a participação de entidades externas e recursos da comunidade científica.	Início dos trabalhos em 2008.

Fonte: Elaboração Própria

Em relação ao conteúdo da tabela 1 o ponto de análise da parte da observação das tecnologias que foram propostas e estudadas. Nos quatro artigos são apresentadas propostas de uso de recursos tecnológicos para a salvaguarda das páginas web. O

software mais citado é o *Heritrix* e o banco de dados *Wayback Machine*, ambos desenvolvidos pela iniciativa *Internet Archive*.

Fazendo a análise da metodologia usada nos artigos pôde ser constatado que nos dois primeiros artigos a metodologia usada é quantitativa e qualitativa e de caráter exploratório, onde o autor buscou encontrar as referências de trabalhos voltados para arquivamento *web* e estudo de bases e casos internacionais. Os dois últimos artigos são estudos práticos de aplicação da iniciativa de preservação da *web* portuguesa.

As iniciativas mais citadas são a do *Internet Archive*, o *World Wide Web Consortium*, Grupo de trabalho *Archive-It* e o Arquivo da *Web* Portuguesa (AWP). Essas iniciativas estão sendo desenvolvidas por diferentes países como: Estados Unidos, Reino Unido, Alemanha, Portugal, Espanha. No processo de observação das iniciativas apresentadas nos trabalhos foi constatado que nenhuma pertence diretamente ao Brasil e que na América Latina a única iniciativa que foi citada é a *Latin American Web Archiving Project*.

As recomendações ressaltam a importância da construção de políticas de preservação e a implementação de mais iniciativas para a preservação da *web*. Assim como a disponibilização e integração de serviços e métodos que permitam maior integração dos profissionais voltados para a área.

4.2 Análise de conteúdo da dissertação

Em relação ao conteúdo da tabela dois (conteúdo da dissertação), foi realizada a análise da dissertação de Nogueira (2008).

Tabela 2: Conteúdo da dissertação

Título do Trabalho	Tecnologia	Metodologia	Iniciativa	Recomendações	Contexto Temporal
Arquivo da web portuguesa. Recomendações para a criação de conteúdos preserváveis ao longo do tempo. (CONSÓRCIO)	Tem sua versão atual baseada no <i>Heritrix</i> , que é um batedor criado pelo <i>Internet Archive</i> responde aos requisitos de um arquivo da <i>web</i> .	Extração automática de conteúdos da <i>web</i> .	Fundação para a Computação Científica Nacional (FCCN)	Para que se cumpram algumas recomendações é necessário que se siga as instruções de protocolo.	Pesquisa restrita aos anos de (2012-2016). Lista com versão de 2009.

Fonte: Elaboração Própria

Na dissertação, observou-se o uso do modelo rArc (modelo de replicação distribuída), é um projeto de Arquivo da Web Portuguesa que tem por iniciativa a FCCN (Fundação para a Computação Científica Nacional), que almeja replicar os conteúdos contidos no sistema de armazenamento de um arquivo da *Web*. Observou-se também que seu desenvolvimento permite lidar com requisitos que irão suportar alterações ao longo do tempo. E usa a tecnologia *Heritrix*, um batedor especialmente criado pelo *Internet Archive* e tem por metodologia a extração automática de conteúdos da *web*, além de recomendar devida utilização dos mecanismos para o sucesso na utilização do *software* respeitando as políticas de preservação.

4.3 Consórcio

No decorrer do levantamento bibliográfico, foi encontrado uma página na *web* denominada “ARQUIVO.PT.”, onde se encontra recomendações e documentos importantes como o consórcio que leva o título “Arquivo da *web* portuguesa: Recomendações para a criação de conteúdos preserváveis ao longo do tempo”, citado abaixo.

Tabela 3: Consórcio ARQUIVO.PT.

Título do Trabalho	Tecnologia	Metodologia	Iniciativa	Recomendações	Contexto Temporal
Preservação da web através de replicação distribuída em larga escala	O formato Arc é uma especificação de ficheiro desenvolvido pelo Internet Archive que é utilizado pelo batedor Heritrix para armazenar os conteúdos recolhidos da web.	Replicação de conteúdos depositados no armazém de conteúdos do Arquivo da Web por computadores espalhados pela Internet e recuperação dos conteúdos replicados para análise de possíveis perdas.	O trabalho apresentado na tese foi realizado no âmbito do projeto Arquivo da Web Portuguesa (AWP), em curso na Fundação para a Computação Científica Nacional (FCCN).	É necessário alterar alguns mecanismos para permitir este novo modo de funcionamento.	A pesquisa de mestrado consta de 2008.

Fonte: Elaboração própria

Apesar de não fazer parte dos documentos que compõe a análise, trata-se um documento importante, devido sua relevância é interessante que se saiba de sua existência.

Importante lembrar que quando falamos de *web* portuguesa, estamos nos referindo a todo conteúdo que está no domínio.pt e que seu arquivamento está sendo realizado pelo projeto da FCC (fundação que possui um sistema de arquivamento de conteúdos) cujo principal objetivo é a preservação da informação lá contida.

5 FERRAMENTAS E INICIATIVAS DE PRESERVAÇÃO DE PÁGINAS WEB

Algumas iniciativas de arquivamento *web* e preservação digital podem ser destacadas pelo desenvolvimento de projetos e softwares. Dentro dessas iniciativas podem ser citadas a do *Internet Archive*, o consórcio *World Wide Web*, *International Internet Preservation Consortium*, *Archive-It*, *Arquivo da Web Portuguesa (AWP)* e a Fundação para a Computação Científica Nacional (FCCN).

Os *softwares* de arquivamento e preservação são ferramentas importantes nesse processo de salvaguarda das páginas *web*. Alguns *softwares* podem ser destacados como *Heritrix*, *HTTrack* e o *Wayback Machine*.

5.1 Iniciativas

- Internet Archive:

O *Internet Archive* é uma instituição sem fins lucrativos, começou os seus trabalhos em 1996 e tem como objetivo a construção de uma biblioteca digital de sites da internet com a missão de fornecer acesso a todo o tipo de conhecimento. E permite livre acesso aos conteúdos para pesquisadores, historiadores, acadêmicos. (INTERNET ARCHIVE, 2018).

- Consórcio *World Wide Web*:

É um consórcio de nível internacional liderado por Tim Berners-Lee que busca desenvolver padrões para a *web*. Os trabalhos são desenvolvidos por

organizações filiadas, por e por uma equipe interdisciplinar e pelo público. tempo integral e o público trabalham juntos para desenvolver padrões para a *Web*. (CONSÓRCIO WORLD WIDE WEB, 2018).

- Consórcio *International Internet Preservation* (IIPC):

Criado no ano de 2003 foi firmado por 12 instituições participantes. E tem a missão “adquirir, preservar e tornar acessível o conhecimento e as informações da Internet para as gerações futuras em todos os lugares, promovendo o intercâmbio global e relações internacionais.” E tem estudo e desenvolvido padrões para a implementação dos objetivos de preservação da web e disseminação do conhecimento. O consórcio reúne uma rede de bibliotecas e instituições que trabalham em colaboração. (GRUPO DE ESTUDOS DE DIREITO AUTORAL E INDUSTRIAL, 2018).

- Archive-It:

Foi lançado em 2006 é um serviço de arquivamento da *web* por assinatura. O *Archive-It* permite ser feita a coleta, criação e preservação de conteúdo da *web*. Os conteúdos são armazenados nos centros de dados do serviço. E atualmente conta com mais de 400 organizações. (ARCHIVE-IT, 2018).

- Arquivo da *Web* Portuguesa (AWP):

É uma infraestrutura de investigação que permite pesquisar e acessar páginas da *web* arquivadas desde o ano de 1996. E tem como objetivo a preservação da informação publicada na *Web* portuguesa levando em conta seu acesso futuro e como poderíamos evitar a perda de informações existentes em páginas *web* ao longo do tempo. (ARQUIVO DA WEB PORTUGUESA, 2018).

- Fundação para a Computação Científica Nacional (FCCN):

Fundada em janeiro funcionou de 1987 a 2013. FCCN foi uma instituição privada portuguesa sem fins lucrativos de utilidade pública. Tem como atividade primeira o planeamento da Rede Ciência, Tecnologia e Sociedade (RCTS) e o *backbone* de Portugal (WIKIPÉDIA, 2018).

5.2 Softwares

- Heritrix:

Software livre para o rastreamento (*crawler*) de páginas *web*. Software Heritrix é um *software* sob licença livre, que foi desenvolvido pela iniciativa *Internet Archive*, em linguagem *Java*, sendo utilizado por diversas iniciativas de arquivamento da *web* no mundo (HERITRIX, 2018).

- HTTrack:

HTTrack é um utilitário de navegador *offline* e gratuito. Ele permite que sites da *web* sejam baixados para o próprio computador. O utilitário organiza a estrutura de *links* do *site* original, permitindo a visualização conforme o *site* original. (HTTRACK, 2018).

- Wayback Machine:

É o banco de dados digital do *Internet Archive*, lançado em 1996, e que armazena 2 *petabytes* de dados compactados, ou mais de 150 bilhões de páginas da *web*, incluindo conteúdo de todos os domínios de primeiro nível, mais de 200 milhões de sites e mais de 40 idiomas. (WAYBACK MACHINE, 2018).

6 Conclusão

O objetivo do presente artigo foi analisar/revisar a bibliografia da produção científica acerca do tema “Preservação de *websites*” e “arquivamento da *web*” servindo também de norteador para os pesquisadores no entendimento e acompanhamento do tema.

Atualmente observa-se um crescimento ascendente da informação produzida e disseminada na internet. Devido à problemática da curta duração das páginas *web* existe a preocupação de como tais informações serão preservadas de forma *perene* para ser acessada por gerações futuras.

Iniciativas importantes foram citadas no decorrer do trabalho como o *Internet Archive*, o consórcio *World Wide Web*, *International Internet Preservation Consortium*, *Archive-It*, Arquivo da *Web* Portuguesa (AWP) e a Fundação para a Computação

Científica Nacional (FCCN). Assim como se observou que os *softwares* mais usados, no que diz respeito à preservação e arquivamento da *web*, foram o *Heritrix*, *HTTrack* e o *Wayback Machine*.

Ao longo do desenvolvimento da pesquisa pôde-se perceber que o quantitativo de artigos é superior ao de teses, dissertações e livros voltados para a área. É importante ressaltar que não foram encontradas teses nas bases de dados pesquisadas.

O Consórcio do ARQUIVO.PT, foi citado no decorrer do trabalho devido sua relevância e por trazer documentos e recomendações importantes para quem precisa de embasamento ou entender mais sobre arquivamento e preservação da *web*.

Após a análise e observação da bibliográfica do panorama nacional e internacional, conclui-se que há uma escassez na produção científica referente ao tema, principalmente na América Latina. A partir desta pesquisa recomenda-se que sejam desenvolvidos estudos para a área e que políticas de preservação sejam implementadas em âmbito nacional e internacional podendo integrar a comunidade da América Latina.

REFERÊNCIAS

ALENCAR BRAYNER, Aquiles. 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, SP, v. 14, n. 2, p. 318-333, maio/ago. 2016. ISSN 1678- 765X. Disponível em: doi:<http://dx.doi.org/10.20396/rdbci.v14i2.8645982>. Acesso em: 24 abr. 2018.

ALVARÉZ, Manuel; et. al. Web Oculta del Lado Cliente: Escala de Crawling. **Departamento de Tecnologías de Información y las Comunicaciones**. Universidade da Coruña. Disponível em: http://www.dl.kuis.kyoto-u.ac.jp/~rafael.lopez/publications/JITEL_2011_2_CR.pdf. Acesso em: 23 mar. 2018.

ARQUIVO NACIONAL. Equipe técnica do programa AN digital. **Política de preservação digital**. 2016. Disponível em: http://arquivonacional.gov.br/siga_antigo/images/an_digital/and_politica_preservacao_digital_v2.pdf . Acesso em: 23 mar. 2018.

ARQUIVO DA WEB PORTUGUESA. **Recomendações para a criação de conteúdos preserváveis ao longo do tempo**. 2009. Disponível em: <http://arquivo.pt/wayback/20090520162044/http://arquivo-web.fccn.pt/>. Acesso em 24 abr. 2018.

BODÊ, Ernesto Carlos. **Preservação de documentos digitais: o papel dos formatos de arquivo**. 2008. 153 f. Dissertação (Mestrado em Ciência da Informação)- Universidade de Brasília, Brasília, 2008. Disponível em:

http://repositorio.unb.br/bitstream/10482/2034/1/2008_ErnestoCarlosBode.pdf. Acesso em: 23 mar. 2018.

DANTAS, Camila Guimarães. **Criptografias da memória: um estudo teórico-prático sobre o arquivamento da web no Brasil**. Disponível em: <http://www.memoriasocial.pro.br/documentos/Teses/Tese42.pdf>. Acesso em 24 abr. 2018.

GOMES, Daniel. **Preservar a Web: um desafio ao alcance de todos**. 2010. Disponível em: <https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/158>. Acesso em: 19 jun. 2018.

FUNDAÇÃO PARA A COMPUTAÇÃO CIENTÍFICA NACIONAL. In: WIKIPÉDIA: a enciclopédia livre. Wikipédia, 2018. Disponível em: https://pt.wikipedia.org/wiki/Funda%C3%A7%C3%A3o_para_a_Computa%C3%A7%C3%A3o_Cient%C3%ADfica_Nacional. Acesso em: 22 jun. 2018.

GOMES, Daniel; MIRANDA, João. **Arquivo e medição da web portuguesa**. 2008. Disponível em: <https://www.rcaap.pt/detail.jsp?id=oai:comum.rcaap.pt:10400.26/471>. Acesso em: 19 jun. 2018.

MÁRDERO ARELLANO, Miguel Ángel. **Crítérios para a preservação digital da informação científica**. Brasília: Universidade de Brasília, 2008. 354 f. Tese (doutorado em Ciência de Informação), Universidade de Brasília, 2008. Disponível em: http://www.repositorio.unb.br/bitstream/10482/1518/1/2008_MiguelAngelMarderoArellano.pdf. Acesso em: 23 mar. 2018.

NOGUEIRA, André Ricardo Lopes. **Preservação da web através de replicação distribuída em larga escala**. 2008. Dissertação (Mestrado) - Universidade Nova de Lisboa, 2008. Disponível em: <http://repositorio.ul.pt/handle/10455/3189>. Acesso em 24 abr. 2018

ROCKEMBACH, Moisés. Arquivamento da Web: estudos de caso internacionais e o caso brasileiro. **Revista Digital de Biblioteconomia e Ciência da Informação**. Campinas, v. 16, n. 1, 2018. Disponível em: <http://hdl.handle.net/10183/169433>. Acesso em 24 abr. 2018.

ROCKEMBACH, Moisés; PAVÃO, Caterina Marta Groposo. Políticas e tecnologias de preservação digital no arquivamento da web. **RICI: Revista Ibero-americana de Ciência da Informação**. ISSN 1983-5213, Brasília, v. 11, n. 1, p. 168-182, jan. /abr. 2018. Disponível em: <http://periodicos.unb.br/index.php/RICI/article/view/27950/0>. Acesso em: 23 mar. 2018.

SITES CONSULTADOS

ARCHIVE-IT. Disponível em: <https://archive-it.org/>. Acesso em: 2018.

ARQUIVO DA WEB PORTUGUESA. Disponível em: <http://arquivo.pt/>. Acesso em: 2018.

CONSÓRCIO WORLD WIDE WEB. Disponível em: <http://www.w3c.br/Sobre/>. Acesso em: 2018.

GRUPO DE ESTUDOS DE DIREITO AUTORAL E INDUSTRIAL. Disponível em: <https://gedai.com.br/?q=pt-br/content/o-cons%C3%B3rcio-internacional-de-preserva%C3%A7%C3%A3o-da-internet-international-internet-preservation>. Acesso em: 2018.

HERITRIX. Disponível em: <http://crawler.archive.org/index.html>. Acesso em: 2018.

HTTRACK. Disponível em: <https://www.httrack.com/>. Acesso em: 2018.

INTERNET ARCHIVE. Disponível em: <https://archive.org/>. Acesso em: 2018.

OASISBR. Disponível em: <http://oasisbr.ibict.br/vufind/>. Acesso em: 2018.

WIKIPÉDIA. Disponível em: https://pt.wikipedia.org/wiki/Funda%C3%A7%C3%A3o_para_a_Computa%C3%A7%C3%A3o_Cient%C3%ADfica_Nacional. Acesso em: 2018.

NOTAS

AGRADECIMENTOS

Agradecemos ao Doutor Miguel Ángel Márdero Arellano pela orientação e todos os ensinamentos.

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: J.P. Farias, K. C. V. Bomfim

Coleta de dados: J.P. Farias, K. C. V. Bomfim

Análise de dados: J.P. Farias, K. C. V. Bomfim

Discussão dos resultados: J.P. Farias, K. C. V. Bomfim

Revisão e aprovação: J.P. Farias, K. C. V. Bomfim

CONJUNTO DE DADOS DE PESQUISA

- 1) Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

FINANCIAMENTO

Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)/ Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

CONSENTIMENTO DE USO DE IMAGEM

Não se aplica.

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA

Não se aplica.

CONFLITO DE INTERESSES

Não se aplica.

LICENÇA DE USO –

Os autores cedem à **Encontros Bibli** os direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob a [Licença Creative Commons Attribution](#) (CC BY) 4.0 International. Esta licença permite que **terceiros** remixem, adaptem e criem a partir do trabalho publicado, atribuindo o devido crédito de autoria e publicação inicial neste periódico. Os **autores** têm autorização para assumir contratos adicionais separadamente, para distribuição não exclusiva da versão do trabalho publicada neste periódico (ex.: publicar em repositório institucional, em site pessoal, publicar uma tradução, ou como capítulo de livro), com reconhecimento de autoria e publicação inicial neste periódico.

PUBLISHER –

Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no [Portal de Periódicos UFSC](#). As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando,



necessariamente, a opinião dos editores ou da universidade.

EDITORES –

Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.

HISTÓRICO –

Recebido em: 13/07/2018 – Aprovado em: 27/02/2019

