RDBCI: Revista Digital Biblioteconomia e Ciência da Informação
RDBCI : Digital Journal of Library and Information Science
DOI 10.20396/rdbci.v16i1.8648747

ARQUIVAMENTO DA WEB: ESTUDOS DE CASO INTERNACIONAIS E O CASO BRASILEIRO

WEB ARCHIVING: INTERNATIONAL CASE STUDIES AND THE BRAZILIAN CASE

ARCHIVO DE LA WEB: ESTUDIOS DE CASO INTERNACIONALES Y EL CASO BRASILEÑO

[1]Moisés Rockembach
[1]Universidade Federal do Rio Grande do Sul

Correspondência

[1]Moisés Rockembach
Universidade Federal do Rio Grande do Sul
Porto Alegre,RS
Email: moises.rockembach@ufrgs.br
ORCID: http://orcid.org/0000-0001-9057-0602

Submetido em: 01/06/2017
Aceito em: 29/08/2017
Publicado em: 12/09/2017

JITA:HC. Archivalmaterials.

RESUMO: O objetivo deste estudo foi delimitar conceitualmente e teoricamente o tema arquivamento da web, além de verificar estudos de caso internacionais e a situação brasileira sobre este tema. Pesquisa de natureza exploratória-descritiva, utiliza abordagem qualitativa, com a aplicação de metodologia de seleção e análise de estudos de caso internacionais para exemplificar o funcionamento do arquivamento da web em vários países, além da análise do contexto brasileiro. São abordados os modelos de processo e ciclo de vida do arquivamento da web, bem como aspectos legais. Dos exemplos encontrados, foram selecionados seis estudos de caso internacionais, contemplando uma variedade de contextos (organizações sem fins lucrativos, arquivos nacionais, bibliotecas regionais, bibliotecas nacionais, universidades, provedores de serviços), não encontrando nenhum sistema específico de arquivamento da web brasileira documentado na literatura, com somente alguns assuntos arquivados de forma esparsa. A investigação diferencia-se por trazer um tópico de pesquisa ainda incipiente no país, mas com muitas possibilidades de estudo. Conclui que é preciso fomentar a discussão no panorama nacional e sugere que iniciativas sejam desenvolvidas para o arquivamento da web brasileira.

PALAVRAS-CHAVE: Arquivamento da web. Tecnologias de arquivamento. Memória digital. Casos internacionais. Caso brasileiro.

ABSTRACT: The objective of this study was to delimit, conceptually and theoretically, the webarchiving, besides, verifying international case studies and the Brazilian situationon this subject. Research with exploratory-descriptive nature, using qualitative approach, with the application of methodology of selection and analysis of international case studies to exemplify the operation of web archiving in several countries, in addition to the analysis of the Brazilian context. It approached the process and life cycle model of web archiving, as well as legal aspects. From the examples found, it has selected six international case studies, covering a variety of contexts (non-profit organizations, national archives, regional libraries, national libraries, universities, service providers), not finding any specific web archiving initiative in Brazil, documented in the literature, with only a few issues archived in a dispersed way. The research differs by bringing a topic of research still incipient in Brazil, but with many possibilities of study. It concludes that it is necessary to foment the discussion in the national panorama and suggests the development of initiatives for the Brazilian web archiving.

KEYWORDS: Web archiving. Archiving technologies. Digital memory. International cases. Brazilian case.

RESUMEN: El objetivo de este estudio fue delimitar, conceptual y teóricamente, el tema archivamiento de la web, además de verificar estudios de caso internacionales y la situación brasilera sobre este tema. Investigación de naturaleza exploratória-descriptiva, utilizando abordaje cualitativa, con la aplicación de metodología de selección y análisis de estudios de caso internacionales para ejemplificar el funcionamiento del archivamiento de la web en varios países, además del análisis del contexto brasilero. Son abordados los modelos de proceso y ciclo de vida del archivamiento de la web, así como aspectos legales. De los ejemplos encontrados, fueron seleccionados seis estudios de caso internacionales, contemplando una variedad de contextos (organizaciones sin fines de lucro, archivos nacionales, bibliotecas regionales, bibliotecas nacionales, universidades, proveedores de servicios), no encontrando cualquier sistema específico de archivamiento de la web brasilera documentado en la literatura, con solamente algunos asuntos archivados de forma dispersa. La investigación se diferencia por traer un tópico de investigación todavía incipiente en el país, pero con muchas posibilidades de estudio. Concluye que es preciso fomentar la discusión en el panorama nacional y sugiere que iniciativas sean desarrolladas para el archivamiento de la web brasilera.

PALAVRAS CLAVE: Archivamiento de la web. Tecnologías de archivamiento. Memoria digital. Casos internacionales. Caso brasileño.

1 INTRODUÇÃO

De forma objetiva, podemos definir o arquivamento da web como 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.

Entretanto, devido à volumosa produção de informação digital no ambiente web, torna-se imprescindível analisar como este arquivamento vem sendo desenvolvido no panorama internacional, bem como identificar estudos que demonstrem a situação nacional. Com base na literatura e estudos publicados, esta investigação converge para os questionamentos: o que é o arquivamento da web e quais são os seus usos? Quais são as pesquisas desenvolvidas sobre o tema arquivamento da web em nível internacional e qual a atual situação brasileira?

Em uma primeira aproximação sobre o tema, identificou-se a dificuldade em encontrar estudos no contexto brasileiro, caracterizando, portanto, a originalidade da pesquisa. Justifica-se a importância e relevância do assunto, por tratar-se de tema recente de pesquisa em âmbito nacional, que merece ser melhor investigado. Além disto, por incidir sobre a memória de tudo o que foi e está sendo produzido e difundido na web brasileira e a perspectiva de acesso futuro a estas informações.

Para um adequado enquadramentoda pesquisa, sistematizamos com a descrição dos procedimentos metodológicos, seguido da apresentação e discussão dos resultados, que envolvem os resultados recuperados nas bases de dados de periódicos científicos, os conceitos fundamentais relacionados ao arquivamento da web e possíveis usos das informações recuperadas por este processo, bem como as considerações finais sobre a pesquisa realizada.

2 METODOLOGIA

A pesquisa configurou-se numa abordagem qualitativa, exploratória-descritiva, que objetivouidentificarestudos de caso internacionais e nacionaisem arquivamento da web, a partir da busca na literatura científica, artigos publicados em periódicos científicos, e literatura cinzenta, especificamente documentos, relatórios e *whitepapers,* publicados na internet, que estabeleçam conceitos, teorias e políticas de arquivamento da web. Além disso, buscou um enquadramento epistemológico sobre os artigos encontrados, identificando os conceitos e usos atribuídos ao arquivamento da web, bem como a seleção de estudos de casos internacionais e o caso brasileiro.

Iniciamos a pesquisa com a identificação da produção científica sobre o arquivamento da web, pois, por se tratar de um tema novo no contexto nacional e com uma abordagem de pesquisa que tem característica exploratória, torna-se fundamental o levantamento do que de mais recente tem-se produzido sobre a área, bem como a literatura clássica sobre o tópico.

A pesquisa foi realizada na base de dados Scopus, com os termos "web archiving", limitando a pesquisa aos últimos 15 anos (2002-2016).

Como, além da característica exploratória, o trabalho também tem abordagem descritiva, seguimos a análise com o estabelecimento de conceitos e usos do arquivamento da web, delineando o que foi encontrado na pesquisa na base de dados Scopus e nos sites que tratam sobre o arquivamento da web, para um melhor enquadramento teórico.

Ainda foi possível selecionar 6 estudos de caso internacionais sobre arquivamento da web, levantando as principais características, e o contexto brasileiro neste assunto.

3 O QUE É O ARQUIVAMENTO DA WEB E COMO FUNCIONA?

Para um melhor entendimento, traçamos uma resposta a esta pergunta a partir do estabelecimento de conceitos identificados na literatura internacional, bem como a descrição do seu funcionamento.

A partir da pesquisa realizada na base de dados Scopus com as delimitações especificadas na metodologia, recuperamos210 resultados. Dos três maiores tipos documentais encontrados, a maior parte diz respeito a artigos de conferências (*conference papers*), num total de 114 resultados, seguido por artigos de periódicos, 72 resultados e capítulos de livro, 10 resultados. A leitura e análise destes materiais, juntamente com o uso das informações disponibilizadas nos sites que realizam o arquivamento da web, possibilitaram compreender, explicar e exemplificar esta área de estudo.

Convém delimitar o que chamamos de web, além de tudo o que se insere nesta esfera, para também limitar o objeto de estudo da área de arquivamento da web. Também se faz necessário compreender o que significa o arquivamento digital deste material, como funciona, em que fases são divididas e como são recuperadas estas informações, para determinar o processo pelo qual passam, da coleta à recuperação.

Se nos perguntarmos o que é a *world wide web,* ou seu termo abreviado, *web*, teremos que nos reportar imediatamente a Tim Bernes-Lee e seu conceito, que acabou se desenvolvendo até a web que conhecemos hoje. A *world wide web* nasceu em 1989, resultado da proposta de Bernes-Lee para suprir a necessidade de compartilhamento de informações por cientistas, tendo a Organização Europeia para a Pesquisa Nuclear - CERN (*Conseil Européen pour la Recherche Nucléaire*), como contexto inicial de aplicação.

A proposta inicial do surgimento da Web trazia, na sua visão geral, o conceito de organizar, dar acessoa informação e evitar a perda de detalhes importantes dos projetos desenvolvidos no CERN, visto a alta complexidade envolvida e os diversos documentos relacionados. O sistema de *hiperlinks*, termo cunhado por Ted Nelson na década de 1950, tem a função de vincular informações, de formar distintaa estrutura hierárquica em árvores ou com palavras-chave (BERNERS-LEE, 1989). O estabelecimento de uma forma de identificar os objetos dentro da web, por meio do Identificador Uniforme de Recursos (URI, do inglês Uniform Resource Identifier), o uso de linguagem de marcação (HTML, do inglês *Hypertext Markup Language*), de protocolo de transferência de dados (HTTP, do inglês Hypertext Transfer Protocol) e de um navegador – primeiramente o World wide web, posteriormente renomeado como Nexus, e depois o Mosaic, como primeiro navegador gráfico da web – foram elementos fundamentais para a Web tornar-se o ambiente informacional e comunicacional que conhecemos hoje.

A primeira página web do mundo somente foi inserida na rede em 20 de dezembro de 1990 e ainda pode ser acessada online[1]. A manutenção deste endereço web, bem como as primeiras ligações (*hiperlinks*) vem sendo preservadas por um projeto desenvolvido pelo próprio CERN[2]. O World Wide Web Consortium[3] (W3C), fundado em 1994, discute e estabelece padrões e diretrizes para garantir o crescimento da web a longo prazo, fundamentado em uma web aberta e colaborativa.

A velocidade em que perderemoso acesso às informações produzidas e disponibilizadas na web é um dos maiores fatores de preocupação. Emartigo publicado em 2016, revela-se que 80% das páginas web não estão disponíveis na sua forma original após 1 ano, 13% das referências da web em artigos acadêmicos desaparecem após 27 meses e 11% dos recursos de mídia social, como os postados no Twitter, são perdidos após 1 ano (COSTA, GOMES, SILVA, 2016).

Por isto, a compreensão de como funciona o arquivamento da web torna-se ponto importante a ser sublinhado. Conforme Gomes (2010), as 3 etapas de arquivamento da web envolvem recolher a informação, indexar e disponibilizar serviços de pesquisa e acesso, sendo que a primeira etapa subdivide-se em coletar o arquivo, armazená-lo, extrair os endereços para outros arquivos a partir dos hiperlinks, e inserir os novos endereços descobertos para a recolha. Também ressalta a possibilidade de pesquisa por períodos ou intervalos de tempo na web arquivada e que a necessidade de preservação do conteúdo digital torna-se essencial, visto a dinamicidade e inacessibilidade da web pregressa, mesmo com poucos anos da produção informacional.

[1] Disponível em <http://info.cern.ch/hypertext/WWW/TheProject.html>. Acessoem 22 mar. 2017
[2]Restoring the first website - A project to restore info.cern.ch - the world's first website.Disponível em <http://first-website.web.cern.ch/>. Acesso em 30 mar. 2017
[3] Disponível em <https://www.w3.org/>. Acesso em 20 mar. 2017

DOI 10.20396/rdbci.v16i1.8648747

A figura abaixo ilustra o funcionamento do processo de arquivamento da web:

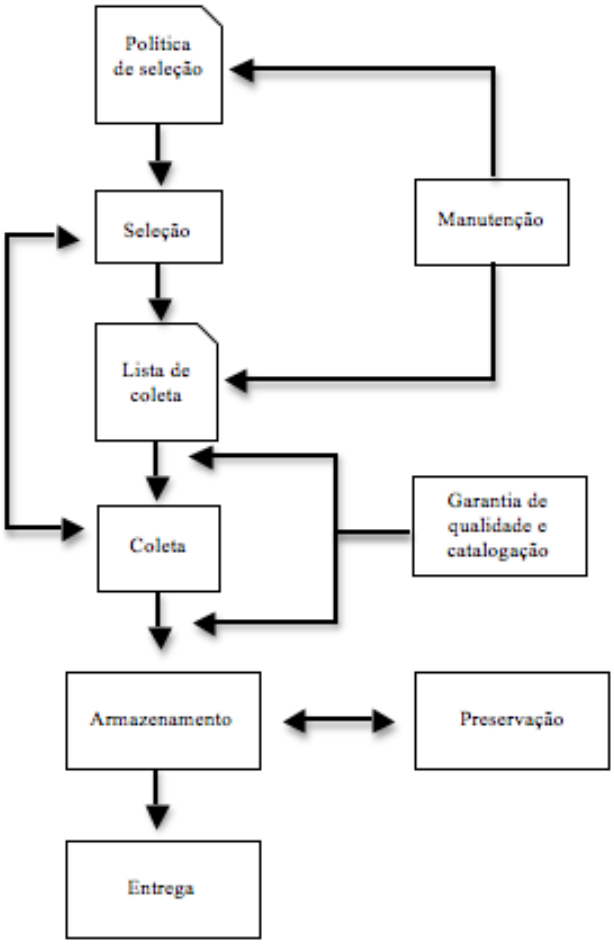

Figura1. Processo de arquivamento da web
Fonte: BROWN, 2006, (*tradução nossa*)

O processo de seleção e coleta é realizado de forma contínua e poderá levar em consideração uma série de fatores, como o conteúdo a ser coletado, se os links externos ao site coletado também serão coletados e até que ponto isto é feito, o equivalente a extensão da coleta, ou ainda a frequência da coleta.

Garantir a qualidade da coleta também é fator fundamental para um melhor arquivamento e recuperação desta informação. Conforme colocado por Hockx-Yu (2012), isto significaria tentar armazenar de forma idêntica ao que vemos ao acessar diretamente um site web, contudo, por uma série de fatores, como *scripts* dinâmicos, streaming de mídia, estruturas das redes sociais e conteúdo baseado em banco de dados, torna-se necessário garantir a qualidade a partir de quatro aspectos também enquadrados por Hockx-Yu (2012), que trazem mais ênfase ao conteúdo do que ao visual gráfico:

DOI 10.20396/rdbci.v16i1.8648747

I. Se o conteúdo pretendido foi coletado integralmente;

II. Se o conteúdo intelectual, em oposição ao estilo e layout, pode ser reproduzido na ferramenta de acesso;

III. Se a cópia coletada pode ser reproduzida, incluindo o comportamento presente no site ao vivo, como a capacidade de navegar interativamente entre links;

IV. Se há a manutenção da aparência de um site.

Ainda cabe salientar que a coleta das informações da web pode ser feita de duas formas, uma onde o autor ou proprietário da informação envia as informações para o arquivamento e outra onde a coleta acontece de forma ativa pela instituição responsável pelo arquivamento. Como abordagens de coleta de web sites, Day (2003) classifica-as como coleta automática, seletiva, por depósito ou uma combinação destas abordagens. Alguns fatores citados por Gomes (2010) apontam para uma vantagem na recolha ativa pelo arquivo, como a automatização da coleta - o que é fundamental visto o crescimento acelerado de dados publicados e que em pouco tempo podem não estar mais acessíveis.

Outro ponto importante diz respeito ao ciclo de vida do arquivamento da web. Um grupo de trabalho do Archive-It, serviço de arquivamento ligado a iniciativa Internet Archive, desenvolveu em 2013 um *White Paper* sobre o ciclo de vida do arquivamento da web, onde podemos observar as diversas fases e dimensões envolvidas neste ciclo.

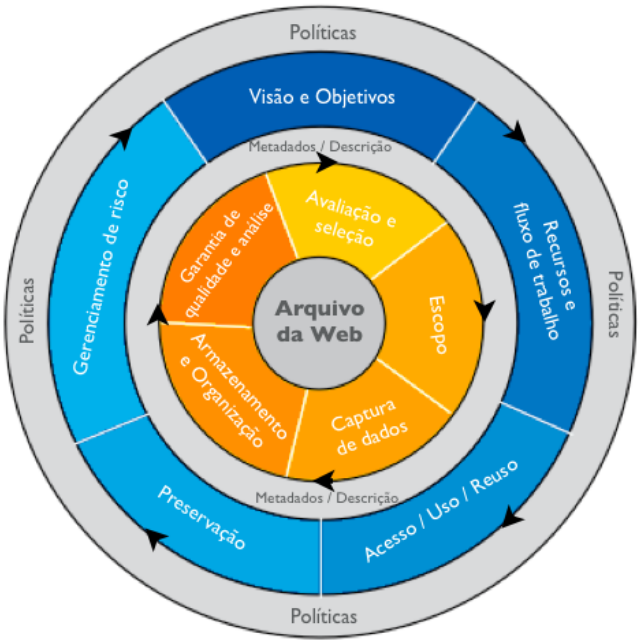

Figura2. Modelo de ciclo de vida do arquivamento da web
Fonte:DONOVAN, HUKILL, PETERSON, 2013 (*tradução nossa*)

A esfera 'Políticas' compreende todo o ciclo de vida do arquivamento da web. No círculo azul são representadas as decisões de alto nível: a definição da visão e objetivos do arquivamento; os recursos e fluxos de trabalho disponíveis; monitorar como acontece o acesso, uso e reuso dos dados; como acontece a preservação; e como procedem ogerenciamento de risco, que envolve questões de direitos autorais, permissões e modalidades de acesso.

A esfera laranja, envolvida pelas definições de metadados e descrição, diz respeito as tarefas operacionais de arquivamento da web, como a avaliar e selecionar quais sites serão coletados; o escopo de coleta; a escolha sobre o tipo e frequência na captura dos dados; a definição de armazenamento de curto ou longo prazo dos arquivos da web e a devida organização; e por fim a garantia de qualidade e análise atrelada aos objetivos definidos para o arquivamento da web.

Outro aspecto encontrado e que se relaciona com a implementação de sistemas de arquivamento da web é a ISO 28500:2009, que especifica o formato *Web ARChive* ou WARC como um padrão de arquivo a ser utilizado, substituindo o antigo formato ARC (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2009).

Sobre os usos do arquivamento da Web, de acordo com Gomes (2010), exemplos que podem ser listados como perfil de usuários vão desde um jornalista buscando informações antigas, um gestor de um site da web recuperando uma versão perdida da página, um historiador estudando documentos digitais, um usuário procurando um link quebrado dos seus 'favoritos', ou mesmo um jurista obtendo provas para um caso.

Resgatar o uso da web retrospectivamente direcionado ao seu lado científico também pode tornar-se uma perspectiva interessante de utilização. O surgimento da web remonta a necessidade de compartilhamento de informações entre cientistas. A aplicação em estudos longitudinais nas mais variadas áreas de conhecimento pode levar em descobertas inovadoras, tanto no processo de investigar e o uso de procedimentos metodológicos diferenciados, quanto nos resultados de pesquisa.

Para além de um paradigma de custódia e preservação (SILVA, et al., 1999) sobre estas informações retrospectivas da web, convém compreender os usos potenciais desta tecnologia e do conteúdo armazenado, para os mais diversos fins, com uma perspectiva sobre os fenômenos informacionais e comunicacionais envolvidos neste processo de arquivamento e recuperação de memórias pessoais e organizacionais.

A forma de recuperação da informação arquivada, a análise sobre o volume ou o tipo de informações recuperadas, levando a abordagens tanto quantitativas, como qualitativas, podem ser considerados campos relevantes de estudo em Ciência da Informação e disciplinas com as quais dialoga, como a Arquivologia, Biblioteconomia e a Museologia, por exemplo.

Sobre questões legais envolvendo o arquivamento da web, Day (2003), destaca alguns problemas legais, como os direitos do autor e demais responsabilidades pelos conteúdos disponibilizados. O uso e reuso destas informações também está em jogo no que diz respeito as permissões, o que nem sempre está explícito na página web, como é o caso da aplicação do *CreativeCommons* e seus formatos de concessões.Isto implica problemas não só legais, como éticos, pois envolve o uso de informações sob direitos autorais, aplicado a contextos internacionais, sob jurisdições distintas e implica também a reflexão sobre ética informacional ao utilizar estes conteúdos, como as questões de proteção de dados e privacidade na rede.

Uma das alternativas legais e éticas a esta coleta e uso do arquivamento da web, mesmo não possuindo os direitos autorais sobre os sites, acontece por meio do *fair use*, ou uso justo, termo usado na legislação americana e que tem relação com a tradição da *Common-law*, aplicada ao uso de conteúdo sob direitos autorais, desde que em certas situações, como o uso pedagógico/educacional, ou como notícia ou ainda como pesquisa, por exemplo. Estas questões são especificamente discutidas por Minow (2003), que defende que, pelo fato da web tornar-se um recurso cada vez mais importante de produção e difusão da informação, surge o interesse em preservar partes do seu conteúdo. E já que a maior parte dos sites são protegidos por direitos autorais, surge um dilema em coletar estes dados, dilema este que poderia ser resolvido com o uso conjunto do fair use, de ferramenta que sinalize a intencionalidade de não-coleta pelos motores de busca ou o controle de permissões de acesso – com o uso de um arquivo robots.txt – e ainda com a expressa solicitação de retirar algum conteúdo capturado.

Esta legislação e o direito ao uso justo não se aplica a todos os países, por isto é preciso ser avaliado caso a caso.O direito Sueco, por exemplo, restringe o acesso ao arquivamento da web de forma online, e só permite que o usuário acesse pessoalmente no local responsável pelo arquivamento da web da Suécia[4], como é o caso da Biblioteca Nacional da Suécia e a iniciativa Kulturarw3. O mesmo acontece com o acesso as páginas web arquivadas pela Biblioteca Nacional da França (BnF), onde só é possível acessar a partir de salas de leitura da biblioteca[5].

[4] Restrições encontradas na página da iniciativa de arquivamento da web da Biblioteca Nacional da Suécia, "*Youcanonlysearchthe Web Archives in personattheNational Library, wherethere are specialcomputers for thepurpose. According to Swedish law, the Web Archives may only be displayed inside the library*". Disponível em <http://www.kb.se/english/find/internet/websites/> Acesso em 09 abr. 2017

[5]Restriçõesencontradasnapágina da Biblioteca Nacional da França, "The Web archives are accessible to authorized users of the BnF, in the reading rooms of the Research Library only. This restriction is the same as that which applies to all legal deposit collections".Disponível em <http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html> Acesso em 09 abr. 2017

4 PANORAMA INTERNACIONAL E NACIONAL SOBRE O ARQUIVAMENTO DA WEB

Selecionamos a partir da análise da literatura pesquisada e dos atuais membros do Consórcio Internacional de Preservação da Internet, seis estudos de caso internacionais sobre arquivamento da web e suas principais características. Também analisamos o contexto brasileiro, a partir do que foi encontrado na pesquisa bibliográfica.

4.1Estudos de caso internacionais

Quanto as iniciativas de preservação da web, podemos destacar alguns pioneiros como a iniciativa Internet Archive, o projeto da Biblioteca Nacional Australiana 'PANDORA' (do inglês, *Preserving and Accessing Networked Documentary Resources of Australia*), bem como o Kulturarw3 da Suécia, iniciados em 1996. Ressaltamos o trabalho do Consórcio Internacional de Preservação da Internet (IIPC, do inglês *International Internet Preservation Consortium*), formado em 2003 e que se dedica ao desenvolvimento de padrões e ferramentas que auxiliam no processo de arquivamento da web.

Em 2003, Day publicou uma pesquisa sobre iniciativas de preservação da web, destacando a relevância de algumas Organizações, que seguem listadas em ordem decrescente de informação armazenada: Internet Archive (sediado nos Estados Unidos, mas de abrangência Internacional), Kulturarw3 (Suécia), Bibliothèque Nationale de France (França), AOLA (Áustria), PANDORA (Austrália), Helsinki University Library (Finlândia), Britain on the Web (Reino Unido) e MINERVA (Estados Unidos da América).

Em pesquisa realizada por Gomes, Miranda e Costa em 2011, foram levantadas e analisadas 42 iniciativas de arquivamento da web ao redor do mundo (GOMES, MIRANDA, COSTA, 2011). Este estudo também foi base para as informações que serviram de base para a produção de uma página da Wikipedia[6], da qual trazemos o mapa das iniciativas de arquivamento da web ao redor do mundo.

[6]Página da Wikipedia "*List of Web archiving initiatives*". Disponível em <https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives>. Acesso em 10 fev. 2017

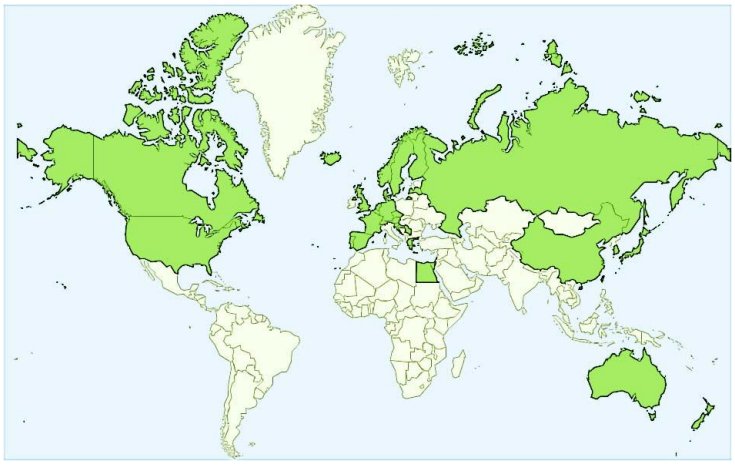

Figura3. Iniciativas de arquivamento da web ao redor do mundo
Fonte: GOMES, MIRANDA, COSTA (2011) e Wikipedia.org

Detodas iniciativas listadas no trabalho de Gomes, Miranda e Costa (2011) selecionamos seis, onde o critério de seleção baseou-se na análise de literatura disponível sobre estes casos, na lista de membros do Consórcio Internacional de Preservação da Internet[7] e na procura por descrever uma variedade de contextos distintos (organizações sem fins lucrativos, arquivos nacionais, bibliotecas regionais, bibliotecas nacionais, universidades, provedores de serviços) ilustrando com informações relativas ao tipo de organização, as formas de coleta web (*harvesting*) e em como fornece acesso aos conteúdos arquivados.

QUADRO 1. Iniciativas internacionais.

Nome/tipo de organização	Tipo de organização/país	Coleta	Acesso
Internet Archive	Organização sem fins lucrativos– Estados Unidos	Exaustiva, por domínios, seletiva, por eventos, temática, coleta em massa, questionário, sites em fim de vida e específicos	Pelo site https://www.archive.org/ e ferramenta *WaybackMachine*
The National Archives	Arquivo Nacional - Reino Unido	Seletiva, por evento, temática.	Pelo site http://nationalarchives.gov.uk/web archive
Biblioteca de Catalunya	Biblioteca regional - Espanha	Domínio '.cat', sites recomendados pelo público do projeto PADICAT e instituições que possuem acordo de cooperação.	Pelos sites http://www.bnc.cat/ e http://www.padicat.cat/
Bibliothèque Nationale de France	Biblioteca Nacional - França	Amostras significativas a partir de duas estratégias: recolha de amostragem em parceria com outras instituições e rastreamento com foco em 20.000 sites	Somente a usuários autorizados em salas de leitura da Biblioteca.

[7] IIPC members. Disponível em < http://netpreserve.org/about-us/members >. Acesso em 02 abr. 2017

		selecionados.	
Harvard Library	Universidade – Estados Unidos	Seletiva	Web Archive Collection Service (WAX) pelo site http://wax.lib.harvard.edu/collections/home.do
HanzoArchives Limited	Provedor de serviço – Reino Unido	Conforme demanda de empresas e organizações, que vai desde a coleta do site corporativo a redes sociais.	Restrito a clientes

Fonte: o autor

Além dos dados apresentados, cabe descrever a maior das iniciativas, a Internet Archive, uma fundação sem fins lucrativos, considerada uma das mais antigas iniciativas de arquivamento da web, datada de 1996. Nestes 21 anos, a Internet Archive coletou e mantém arquivado aproximadamente 286 bilhões de páginas, de mais de 361 milhões de websites (INTERNET ARCHIVE, 2017). É também um dos membros fundadores do Consórcio Internacional de Preservação da Internet.

A iniciativa Internet Archive dispõe de uma ferramenta bastante popular, chamada *Wayback Machine*[8], que permite acessar como um determinado site era em determinada época (*snapshot*). Para isto, é preciso inserir o domínio a ser pesquisado e, com os resultados da pesquisa, selecionar alguma data específica, a partir dos dados capturados.

Dentre os métodos de coleta utilizados, podemos identificar a coleta de dados da Web de forma exaustiva, coleta por domínios de primeiro nível, nacionais e regionais, seletivo, por eventos, temático, coleta em massa, *survey*, sites que serão encerrados (em fim de vida) e sites específicos. A coleta pode ser feita de forma ativa pelo usuário, por exemplo, na página da ferramenta Wayback Machine, é possível inserir o endereço eletrônico que se pretende capturar e salvar esta página no banco de dados da Internet Archive, podendo com isto fazer uma citação confiável desta referência no futuro. Entretanto, esta ferramenta só está acessível para sites que permitam o uso de rastreadores web (*crawlers*), para a devida indexação destes sites, que é a mesma tecnologia utilizada por motores de busca.

São arquivados pela Internet Archive, não só sites da web, mas também livros, textos, filmes, softwares, música e imagens. Também são arquivadas coleções das mais diversas temáticas, muitas aparecem na página principal da Internet Archive, com o total de itens disponíveis em cada coleção.

Já em iniciativas como o da Biblioteca Regional da Catalunha e a Biblioteca Nacional da França, são arquivados domínios específicos, no primeiro caso os ".cat" e no segundo caso os domínios ".fr" e de territórios franceses, como ".re" (Reunião), ".nc" (Nova Caledônia),

[8]WaybackMachine. Disponível em <https://archive.org/web/> Acesso em 05 abr. 2017

".gf" (Guiana Francesa) entre outros, bem como domínios relacionados a França ou a cultura francesa (.bzh, .alsace, .paris) e outros domínios comuns (.com, .org, etc.), criados no território francês ou com conteúdo produzido na França.

As parcerias entre instituições também são frequentes nos casos estudados. O Reino Unido possui um arquivamento da Web composto por contribuições de várias instituições, liderada pela Biblioteca Britânica (*British Library*), a *UK Web Archive* ou *UKWA*[9], que tem o objetivo de preservar os sites do Reino Unido e conta com a parceria da Biblioteca Nacional de Wales (*National Library of Wales)*ea Biblioteca Nacional da Escócia(*National Library of Scotland*).A UKWA também já teve participação do Arquivo Nacional do Reino Unido, *JISC (Joint Information Systems Committee) e The Wellcome Library.*

Cabe também ressaltar que é desenvolvida em Portugal uma estrutura específica para tratar o arquivamento da web, o Arquivo da Web Portuguesa ou AWP[10]. Consideramos importantes alguns pontos levantados pelos pesquisadores portugueses e os quais descrevemos abaixo, um ligado a recomendações de como os sites podem ser melhor arquivados e outro ligado a um projeto de preservação colaborativa da Web.

Oarquivo da Web Portuguesa (2015) publicou recomendações para que as páginas web possam ser melhor arquivadas, divididas entre 'organização do site web', 'conteúdo de cada página web' e 'Geral', cada um com recomendações fundamentais e/ aconselháveis. Quanto à organização do site web, as recomendações fundamentais são links para o endereço de cada conteúdo epágina principal ou de entrada em formato amigável para os rastreadores web (*crawlers*); e as recomendações aconselháveis são manter o mesmo endereço para um conteúdo ao longo do tempo e uso do Robots Exclusion Protocol (REP), por meio do arquivo robots.txt, permitindo ao autor da página distinguir o que deve ou não ser coletado. Quanto ao conteúdo de cada página web, como recomendações fundamentais surgem links nas páginas publicadas utilizando HTML, textos publicados em formatos textuais e tipo de conteúdo (*MIME Type*) e codificação de caracteres identificados corretamente; e como aconselháveis, a existência de metadados acerca dos conteúdos, o uso de normas de formato de arquivos (validados pela W3C), a data de publicação identificada e uso de formatos adequados para preservação. Por fim, como recomendação aconselhável no âmbito geral, o respeito pelas recomendações de usabilidade e acessibilidade para pessoas com deficiência, onde é valido buscar as diretrizes da W3C sobre acessibilidade.

De 2007 a 2011 foi desenvolvido um projeto que visava a preservação colaborativa da Web, chamado rARC (replicador de arquivos ARC), onde usuários da internet cediam parte do espaço de armazenamento dos seus discos rígidos para o arquivamento, contando com a instalação de um programa para efetuar automaticamente este arquivamento. O projeto teve

[9] UK Web Archive (UKWA). Disponível em <https://www.webarchive.org.uk/ukwa/>Acesso em 10 abr. 2017

[10] Arquivo da Web Portuguesa - Disponível em <http://arquivo.pt/> Acesso em 19 fev. 2017

como contributo uma dissertação de mestrado, intitulada "Preservação da web através de replicação distribuída em larga escala" (NOGUEIRA, 2008). O projeto de preservação colaborativa encontra-se suspenso, mas por tratar-se de código aberto, disponibilizou o código-fonte para o uso em outras iniciativas de arquivamento da web[11].

Já nos Estados Unidos, na busca de integrar a pesquisa em diversos arquivos da web, surge o projeto Memento, financiado pelo Programa Nacional de Infra-estrutura e Preservação de Informação Digital (NDIIPP, do inglês *National Digital Information Infrastructure and Preservation Program*) e a ferramenta *Time Travel*, que permite a pesquisa em arquivos da web que tenham conformidade com o protocolo Memento Time Travel para a web,RFC7089[12].

4.2 O caso brasileiro

Algumas iniciativas globais, como a Internet Archive, procuram coletar e armazenar toda a web mundial. Entretanto, como demonstrado na figura 3, não há iniciativas de arquivamento da web identificadas no Brasil e o mesmo exemplo pode ser visualizado nos demais países da América Latina. O Chile ainda não consta na figura 3, mas em 2014 ingressou como membro do Consórcio Internacional de Preservação da Internet, a partir da Biblioteca Nacional do Chile[13].

O Programa Permanente de Preservação e Acesso a Documentos Arquivísticos Digitais do Arquivo Nacional do Brasil (AN Digital), iniciado em 2010, disponibilizou em sua página web o documento Política de Preservação Digital, com versões de 2012 e outra atualizada em 2016. Em ambos documentos, ressalta a necessidade de preservação dos documentos digitais, contudo, dispõe que os a preservação incidirá sobre os tipos documentais "texto estruturado com formatação, imagem matricial, imagem vetorial, áudio, audiovisual, mensagem de correio eletrônico, apresentação (slides), planilha e base de dados relacional" (ARQUIVO NACIONAL, 2016, p.11), e que "em momento futuro, outros tipos mais complexos de documentos em formato digital, como multimídia e páginas web, deverão ser também contemplados" (ARQUIVO NACIONAL, 2016, p.11).

Há uma iniciativa que possui relação com a web brasileira, chamada Latin American Web Archiving Project, hospedada em endereço eletrônico da Universidade do Texas em Austin[14], com foco sobre documentos governamentais e de expressão política. O início da coleta para arquivamento iniciou em 2005, mas nem todas coleções continuam sendo

[11] Projeto de preservação colaborativa rARC - Disponível em <https://code.google.com/archive/p/rarc/>. Acesso em 16 fev. 2017
[12] Sobre o serviço *Time Travel*. Disponível em <http://timetravel.mementoweb.org/about/>. Acesso em 10 abr. 2017
[13]Disponível em <http://www.netpreserve.org/member-organizations/biblioteca-nacional-de-chile-national-library-chile > Acesso em 10 abr. 2017
[14]Disponível em <http://lanic.utexas.edu/project/archives/>. Acesso em 16 fev. 2017

atualizadas. Pesquisando na página da Latin American Web Archiving Project, foram encontradas 4 coleções: Latin American Government Documents Archive (LAGDA)[15]; México 2010[16]; Archive of Venezuelan Political Discourse (ARVEPODIS)[17]; e Archive of Political Parties and Elections in Latin America (APPELA)[18].

Outra iniciativa pontual foido Grupo de Trabalho de Desenvolvimento de Conteúdo (CDG, do inglês, *Content Development Working Group*)[19] vinculado ao Consórcio Internacional de Preservação da Internet, que coletou sites, artigos, notícias, blogs e mídias sociais (Twitter, Facebook) sobre as Olimpíadas Rio 2016. O grupo realizou formulário público[20] para pessoas colaborarem com a seleção de temas relacionados às Olimpíadas 2016. A hashtag '#RIO2016WA' no Twitter também foi uma forma de marcar e acompanhar informações relativas ao arquivamento das Olimpíadas 2016, além de conectar pessoas dispostas a contribuir com o processo de arquivamento da web do referido evento.

A tese de doutorado de Dantas (2015) abordou questões relativas ao arquivamento da web a partir de abordagem teórico-prática, com uma discussão tanto de aspectos culturais e sobre o que a sociedade considera como memória digital, como aspectos técnicos, o que levou a identificação de iniciativas de arquivamento da web internacionais, mas a constatação de não haver coleções de páginas web em instituições brasileiras. Também trouxe experiências empíricas, de demonstrar como o processo de arquivamento é realizado e a formação de uma coleção relativa a ferramentas de busca no Brasil.

Após a pesquisa no site Internet Archive, foi possível encontrar algumas coleções que tem relação com conteúdo produzido no Brasil, mas de forma muito dispersa, sem uma linha definida ou políticas de seleção e arquivamento estabelecidas. Por exemplo, coleções de revistas antigas como a coleção "Geração Prológica (Brazil)"[21], que traz edições da revista de mesmo nome, dos anos de 1984, 1985 e 1986; ou a coleção "Brazilian Web Engines (1997-2013)"[22], desenvolvida em projeto de pesquisa na Universidade Federal do Estado do Rio de Janeiro (Unirio), que procurou armazenar *snapshots* de motores de busca brasileiros, vinculado a pesquisa de Dantas (2015).

[15]Disponível em <http://lanic.utexas.edu/project/archives/lagda/>. Acesso em 17 fev. 2017
[16]Disponível em <http://lanic.utexas.edu/project/archives/mexico2010/>. Acesso em 17 fev. 2017
[17]Disponível em <http://lanic.utexas.edu/project/archives/arvepodis/>. Acesso em 17 fev. 2017
[18]Disponível em <http://lanic.utexas.edu/project/archives/appela/>. Acesso em 17 fev. 2017
[19]Disponível em <http://www.netpreserve.org/working-groups/content-development-working-group>Acesso em 10 abr. 2017
[20]Disponível em <https://netpreserveblog.wordpress.com/2016/06/27/2016-rio-games-collection-how-to-get-involved/>. Acesso em 02 abr. 2017
[21] Coleção Geração Prológica. Disponível em <https://archive.org/details/geracoprologica>. Acesso em 02 abr. 2017
[22] Coleção *Brazilian Web Engines*. Disponível em <https://archive.org/details/ArchiveIt-Collection-4266>. Acesso em 02 abr. 2017

Como levantado por Brayner (2016), países como o Brasil, que ainda não tem uma preocupação em armazenar e preservar a web no nível nacional deveriam ter políticas de arquivamento da web como objetivo para resguardar o patrimônio digital brasileiro.

5 CONSIDERAÇÕES FINAIS

A pesquisa realizada demonstra que a atribuição ou iniciativa de desenvolver o arquivamento da web não está vinculada somente a um determinado tipo de organização. Aparecem como exemplos analisados, tanto instituições públicas, memorialísticas como Arquivos e Bibliotecas Nacionais e Regionais, bem como organizações privadas e também aquelas sem fins lucrativos ou ainda as relacionadas com pesquisa, como é o caso das universidades.

Estas organizações utilizam diferentes tecnologias e estabelecem distintas formas de coleta de dados, não somente baseados na região geográfica ou domínio eletrônico, mas, como verificamos, nas políticas de seleção, vinculadas a eventos, temáticas, uns de forma extensiva, tentando abarcar toda a web, outros mais focados em determinados contextos.

A web hoje configura-se em um ambiente que contém informações sobre diversos campos do conhecimento eque pode possuir diversos formatos de arquivo. A produção e uso de conteúdos pela web transformou a forma como nos comunicamos atualmente. Muitos estudos surgem a partir das interações neste ambiente digital e muitos outros vêm surgindo sobre os usos que se fazem da informação retrospectiva da web. Na medida em que esta tecnologia de arquivamento seja aplicada a vários contextos e países, mais pesquisas poderão ser realizadas, levando em conta as particularidades locais.

Caso não haja responsabilidades atribuídas ao próprio país ou localmente e as suas devidas Instituições governamentais ou de pesquisa em conseguir recuperar o que foi produzido neste ambiente informacional, muito do que já foi e vem sendo gerado digitalmente simplesmente estará perdido. Além disto, todas as possibilidades de pesquisa sobre estes conteúdos não serão aproveitadas por futuros pesquisadores das mais diversas áreas de conhecimento.

REFERÊNCIAS

ARQUIVO DA WEB PORTUGUESA. **Recomendações para a criação de conteúdos preserváveis ao longo do tempo.** Disponível em: <http://arquivo-web.fccn.pt/colaboracoes/recomendacoes-para-autores-de-sitios-web>, 2015.

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>. Acessoem: 05 abr. 2017.

BERNERS-LEE, Tim. “Information management: a proposal.” **Word Journal Of The International Linguistic Association**, 1989.

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: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8645982>. Acesso em: 10fev. 2017. doi:http://dx.doi.org/10.20396/rdbci.v14i2.8645982.

BROWN, Adrian. **Archiving Websites**: a practical guide for information management professionals.Facet publishing, London, 2006.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J.The evolution of web archiving.**International Journal on Digital Libraries**, p. 1-15, 2016.

DANTAS, Camila Guimarães. **Criptografias da memória**: um estudo teórico-prático sobre o arquivamento da web no Brasil. 2015. Tese (Doutorado em Memória Social) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2015.

DAY, Michael. Preserving the fabric of our lives: a survey of web preservation initiatives. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 7.: 2003: Berlin. [**Proceedings of…**]. Berlin: Springer-Verlag, 2003.

DONOVAN, Lori; HUKILL, Graham; PETERSON, Anna.**The web archiving life cycle model**.2013. Disponível em:<http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf>. Acesso em: 10 fev. 2017.

GOMES, Daniel. Preservar a Web: um desafio ao alcance de todos. In: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS. 2010.**Actas do...** [S.l.] : [s.n.], 2010.

GOMES, Daniel; MIRANDA, João; COSTA, Miguel.A survey on web archiving initiatives. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 15.: 2011: Berlin. [**Proceedings of…**]. Berlin: Springer-Verlag, 2011. p. 408-420.

HOCKX-YU, Helen. **How good is good enough?** – qualityassurance of harvested web resources, 2012.Disponívelem:<http://blogs.bl.uk/webarchive/2012/10/how-good-is-good-enough-quality-assurance-of-harvested-web-resources.html>Acessoem: 07 abr. 2017

INTERNET ARCHIVE. Disponível em:<https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/>. Acessoem: 16 fev. 2017

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 28500:2009. Information and documentation - WARC file format, 2009.

MINOW, Mary. **Digital preservation and copyright by Peter Hirtle**,2003.Disponívelem:<http://fairuse.stanford.edu/2003/11/10/digital_preservation_and_copyr/>. Acesso em: 06 abr. 2017.

NOGUEIRA, André Ricardo Lopes. **Preservação da web através de replicação distribuída em larga escala**. 2008. Dissertação (Mestrado) - Universidade Nova de Lisboa, 2008.

SILVA, Armando Malheiro da et al. **Arquivística**: teoria e prática de uma ciência da informação. Porto: Edições Afrontamento, 1999.