RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Políticas e tecnologias de preservação digital no arquivamento da *web*

Moises Rockembach
Universidade Federal do Rio Grande do Sul, Departamento de Ciências da Informação, Porto Alegre, RS, Brasil
moises.rockembach@ufrgs.br

Caterina Marta Groposo Pavão
Universidade Federal do Rio Grande do Sul, Centro de Processamento de Dados, Porto Alegre, RS, Brasil
caterina@cpd.ufrgs.br

DOI: https://doi.org/10.26512/rici.v11.n1.2018.8473

Recebido/Recibido/Received: 2017-11-20
Aceitado/Aceptado/Accepted: 2017-12-09

ARTIGOS

Resumo: O objetivo do artigo foi analisar a preservação digital a partir da abordagem de arquivamento da *web*, desde as tecnologias envolvidas no processo de arquivamento, bem como políticas de seleção, preservação e disponibilização destes conteúdos, além do estudo de instituições internacionais que atuam na preservação da *web*. A metodologia utiliza pesquisa bibliográfica e documental sobre iniciativas internacionais de arquivamento da *web* e objetiva fomentar a discussão no Brasil, assim como servir de subsídio para estudos aplicados. Analisa as publicações científicas na base de periódicos Scopus dos últimos cinco anos (2012-2016) que versam sobre o arquivamento da *web*, políticas de seleção dos conteúdos *web* e tecnologias aplicadas à coleta, armazenamento e acesso aos *websites* arquivados. Traz também um panorama das tecnologias utilizadas pela comunidade de iniciativas de arquivamento da *web*, a partir da identificação dos dados disponibilizados no *site* do Consórcio Internacional de Preservação da Internet. Conclui que países que ainda não possuem iniciativas próprias, como o Brasil, com o estabelecimento de políticas de seleção com enfoques específicos (institucionais, temáticas, por domínio, etc.), assim como uma gestão do ciclo de vida do arquivamento da *web* e a adoção de tecnologias no formato código aberto (*open source*) podem não só preservar sua memória digital, mas também contribuir com a comunidade internacional de arquivamento da *web*.
Palavras-chave: Arquivamento da web; Política de preservação; Preservação digital.

Policies and technologies to digital preservation in *web* archiving

Abstract: The objective of this paper was to analyze digital preservation from the web archiving approach, addressing the technologies involved in the archiving process, as well as policies for the selection, preservation and availability of these contents, as well as the study of international institutions that work on preservation of the web. The methodology uses bibliographic and documentary research on international archival web initiatives and aims to foment the discussion in Brazil, as well as to serve as a subsidy for applied studies. It analyzes the scientific publications based on Scopus journals of the last five years (2012-2016) that deal with web archiving, web content selection policies and technologies applied to the harvest, storage and access to archived web*site*s. It also provides an overview of the technologies used by the community of web archiving initiatives, based on the identification of the data available on the web*site* of the International Internet Preservation Consortium. It concludes that countries that do not yet have their own initiatives, such as Brazil, with the establishment of selection policies with specific approaches (institutional, thematic, domain, etc.), as well as web archive adoption of open source technologies can not only preserve your digital memory but also contribute to the international web archiving community.
Keywords: Digital preservation; Preservation policy; Web archiving.

Política y tecnologías de preservación digital en el archivo de la web
Resumen: El objetivo del artículo fue analizar la preservación digital a partir del abordaje del archivamiento de la *web*, desde las tecnologías involucradas en el proceso de archivo, así como políticas de selección, preservación y puesta a disposición de estos contenidos, además del estudio de instituciones internacionales que actúan en la preservación de la información de la *web*. La metodología utiliza fue la investigación bibliográfica y documental sobre iniciativas internacionales de archivado de la *web* y objetiva fomentar la discusión en Brasil, así como servir de subsidio para estudios aplicados. Analiza las publicaciones científicas de la base de datos Scopus en los últimos cinco años (2012-2016) que versan sobre el archivamiento de la *web*, políticas de selección de los contenidos de la *web* y tecnologías aplicadas a la recolección, almacenamiento y acceso a los sitios *web* archivados. También trae un panorama de las tecnologías utilizadas por la comunidad que participa de las iniciativas de archivamiento de la *web*, a partir de la identificación de los datos disponibles en el sitio del Consorcio Internacional de Preservación de Internet. Concluye que países que aún no tienen iniciativas propias, como Brasil, con el establecimiento de políticas de selección con enfoques específicos (institucionales, temáticos, por dominio, etc.), así como una gestión del ciclo de vida del archivo de la *web* y la adopción de tecnologías en el formato de código abierto (*open source*) pueden no sólo preservar su memoria digital, sino también contribuir con la comunidad internacional de archivamiento de la *web*.
Palabras-clave: Archivamiento de la web. Políticas de preservación; Preservación digital.

1 Introdução

A década de 1990 produziu um marco histórico na conexão informacional e comunicacional entre as pessoas de todo mundo a partir da introdução da *web* a um nível global. A partir do conceito proposto por Berners-Lee (1989) e posteriormente aplicado de várias formas, o uso de páginas *web*, de *hiperlinks*, de linguagens de marcação (HTML), além de áudios e vídeos incorporados a estes *websites*, possibilitou um novo *boom* informacional, onde cada vez mais pessoas e Instituições passaram a produzir no ambiente *web*. A navegação pelos conteúdos a partir do pioneiro e já extinto Mosaic, em 1993, e os seus sucessores Netscape, Internet Explorer, Mozilla Firefox Chrome, entre outros, incrementou o acesso a internet por possibilitar a visualização de forma gráfica e multimídia.

Passado mais de duas décadas do surgimento da *web*, é perceptível um fenômeno interessante, com a internet e o uso intensivo das tecnologias o tempo passa a ter outro significado, pois há a sensação na sociedade tecnológica de que a velocidade que as informações são produzidas, publicadas e passam ao esquecimento tornou-se muito maior.

O objetivo deste trabalho é analisar as possibilidades de arquivamento da *web* como formação de uma memória pessoal, organizacional e de fatos e eventos, muitas vezes publicados unicamente neste ambiente digital. Caso não haja uma preservação digital dos conteúdos produzidos na *web*, muito do que foi desenvolvido neste meio se perderá para sempre.

2. Metodologia

Adotamos uma pesquisa quanti-qualitativa, de caráter exploratório-descritivo, onde consideramos as publicações científicas sobre a temática e os dados disponibilizados quanto às políticas e tecnologias consideradas *standards* internacionais sobre arquivamento da *web*.

Como procedimentos metodológicos, procedeu-se com o método bibliográfico e documental, no uso da pesquisa bibliográfica coletou-se informações de artigos de periódicos científicos que tratavam sobre os temas relacionados 'arquivamento da *web*', 'políticas' e 'tecnologias', na base de dados Scopus.

Como delimitadores na pesquisa bibliográfica, utilizamos a busca avançada com o termo de busca '*web archiving*', limitamos o filtro por artigos de periódicos e atas/anais de conferências, produzidos em língua inglesa e restringindo o que foi publicado nos últimos cinco anos (2012-2016). Na exploração dos resultados, identificamos a devida recuperação dos títulos, palavras-chave e resumos dos textos, bem como o acesso aos artigos na íntegra. Por fim, foram verificados todos os resultados para a inclusão ou exclusão de artigos que tinham ou não relação com a abordagem deste trabalho. Na análise dos dados, ilustramos com quantitativos da produção científica sobre arquivamento da *web*.

Na pesquisa documental, procurou-se coletar e analisar os dados de relatórios disponibilizados pelo Consórcio Internacional de Preservação da Internet (http://netpreserve.org), que concentra informações das iniciativas de arquivamento da *web* ao redor do mundo.

Os resultados advindos da coleta e análise de dados procuram responder quais seriam os padrões tecnológicos e políticos vigentes na comunidade de arquivamento da *web*, trazendo uma visibilidade de como as iniciativas vem conduzindo este trabalho e o que podemos apreender destas definições e do que foi desenvolvido até o momento.

3. Pesquisas e publicações em arquivamento da *web*

A primeira página *web* do mundo foi criada em 20 de dezembro de 1990, a partir da concepção de Berners-Lee para um sistema distribuído de informação para a Organização Europeia para a Pesquisa Nuclear, também conhecida como CERN, localizada na Suíça (BERNERS-LEE, 1989). Esta página foi recuperada e ainda encontra-se disponível em (http://info.cern.ch/hypertext/WWW/TheProject.html), graças à ação da entidade em manter as informações do nascimento da *web* acessíveis a todos.

Entretanto, os padrões, políticas e tecnologias da *web*, tanto para seu desenvolvimento quanto para a sua preservação, são discutidos e tratados por,

nomeadamente, dois consórcios internacionais. Em 1994 é criado o World Wide Web Consortium (https://www.w3.org/), uma comunidade internacional onde as organizações membros, os funcionários e o público em geral trabalham para desenvolver padrões na w*eb*. A missão do W3C é aproveitar ao máximo o potencial da *Web* através do desenvolvimento de protocolos e diretrizes que garantam o crescimento a longo prazo da *web*.

O International Internet Preservation Consortium (IIPC) foi lançado oficialmente em julho de 2003 pela Biblioteca Nacional da França, com 12 instituições participante, que assumiram o compromisso financiar e participar de projetos e grupos de trabalho que estivessem alinhados com os objetivos do IIPC. Os membros do IIPC devem coletar, preservar e tornar acessíveis o conhecimento da *web* global.

Para evitar a perda dos *websites* e realizar a preservação digital dos conteúdos, muitas iniciativas de arquivamento da *web* vêm surgindo pelo mundo, algumas com o intuito de arquivar toda a *web*, a um nível mundial, caso do *Internet Archive* (https://archive.org/), outras nacionalmente, como Arquivos e Bibliotecas Nacionais (Estados Unidos, Reino Unido, Alemanha, Portugal, etc.), outras ainda a um nível local (como, por exemplo, o caso da Biblioteca da Catalunha e o projeto PADICAT). A reunião destas iniciativas tenta dar uma resposta ao rápido desaparecimento que a dinamicidade da internet imprime aos *websites* e possuem distintas formações e objetivos, a saber: organizações não governamentais como a já citada iniciativa *Internet Archive*, instituições memorialísticas (Arquivos e Bibliotecas, em diversas esferas), universidades, com políticas de seleção mais restritas a determinados assuntos ou *websites* de instituições e ainda companhias privadas, que oferecem o serviço de arquivamento da *web* a outras empresas (ROCKEMBACH, 2018).

Na pesquisa bibliográfica, a partir da coleta de dados com os procedimentos descritos na metodologia, foram recuperados 83 documentos. Dos autores que publicaram neste período, os mais produtivos foram Michael L. Nelson, da Old Dominion University, com 17 publicações e Michele C. Weigle, de Los Alamos National Laboratory, com 11 publicações, ambos dos Estados Unidos. Coincidentemente estas instituições são, também, as que possuem maior produção científica e os Estados Unidos o país que mais publicou o que pode ser constatado Figura 1.

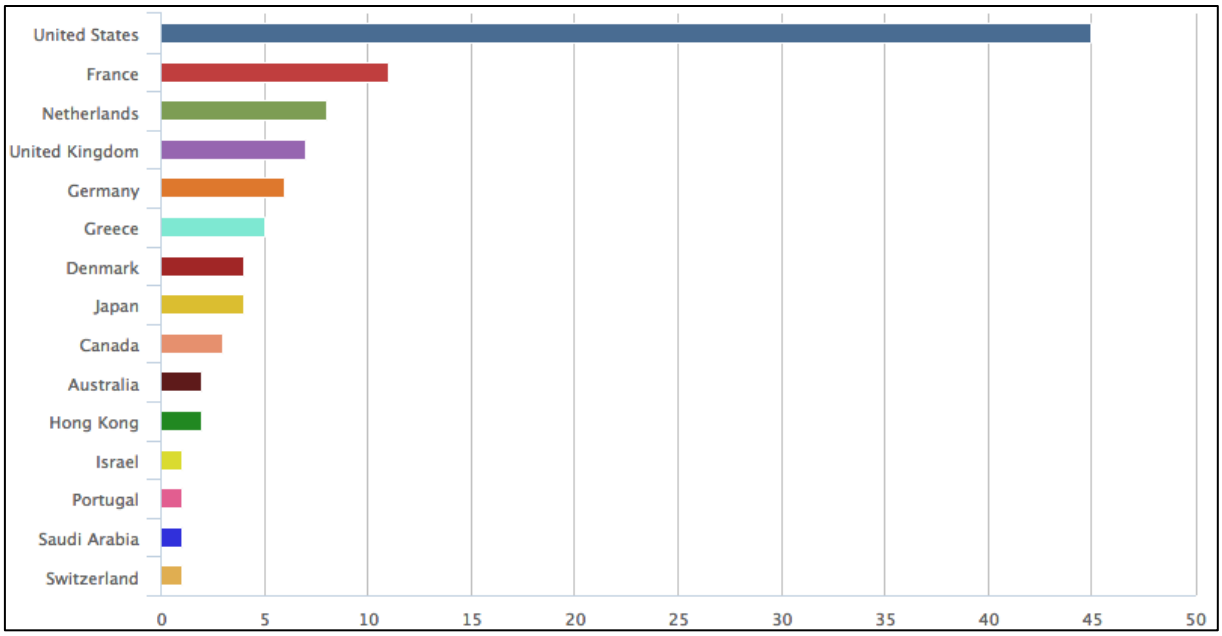

Figura 1: Produção científica por país pela pesquisa "*Web* Archiving" (2012-2016) na base SCOPUS
Fonte: os autores

A Old Dominion University também dirige, junto com o Laboratório Nacional de Los Alamos, o projeto Memento (http://timetravel.memento*web*.org), que permite buscar *websites* armazenados em diversos arquivos da *web* por meio da ferramenta *Time Travel*.

A multi e interdisciplinaridade da área também aparecem nas informações coletadas na pesquisa bibliográfica, como ilustrado na figura 2.

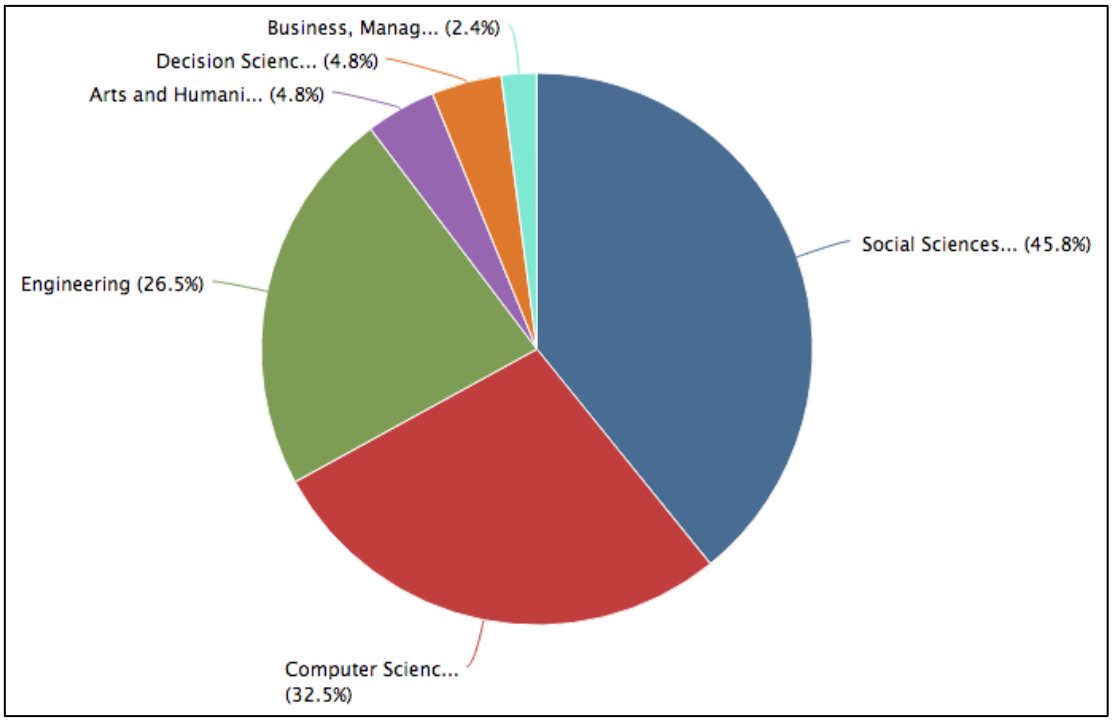

Figura 2 - Divisão das publicações por área de pesquisa
Fonte: os autores

Vemos, portanto, que as três grandes áreas de pesquisa que investigam e publicam sobre o tema de arquivamento da *web* concentram-se nas Ciências Sociais, Ciências da Computação e Engenharia, seguido de Artes e Humanidades, *Decision Sciences* e Negócios, Administração e Contabilidade. Ainda, dos 83 documentos recuperados, 22 utilizam especificamente o termo "*Digital Preservation*" nas suas palavras-chave.

3.1 Características do ambiente *web* e recuperação de páginas da internet

As páginas da *web* e outros objetos digitais utilizados nestas páginas possuem a dinamicidade da internet, ou seja, ao mesmo tempo em que milhares de informações são criadas, outras são sobrepostas, dificultando com o tempo a recuperação destes dados. Se observarmos um portal de notícias, facilmente podemos ver este processo acontecer, muitas vezes o conteúdo muda completamente no mesmo dia.

A busca de alguma notícia específica também pode se dar por outros meios, como por exemplo, pelos motores ou serviços de busca, como o Google ou Bing. Contudo, estes buscadores servem como indexadores da informação publicada, remetendo ao endereço do *website* e a responsabilidade de manter a informação disponível fica a cargo do dono do *website*. Conforme Costa, Gomes e Silva (2016), após um ano de publicação, 80% das páginas *web* não estão disponíveis na sua forma original. Outros dados levantados pelos autores também são alarmantes, em relação às referências *online* de artigos acadêmicos 13% desaparecem após 27 meses e 11% dos recursos de mídia social são perdidos após um ano.

É importante salientar que os motores de busca tem a função de armazenamento em *cache*, possibilitando a recuperação de determinada página *web*, a partir da realização de uma captura da página ou instantâneo (*snapshot*) e como ela aparecia em determinada data e hora, identificado no cabeçalho da página recuperada. Este sistema funciona como um *backup* limitado, pois é possível acessar a página *web* que por algum motivo não se encontra disponível, entretanto algumas funções da página pesquisada e mesmo a navegação entre os *hiperlinks* pode ser comprometida no uso do armazenamento em *cache*. Ainda é possível que a página esteja configurada pelo administrador do *website* para não ser arquivada com o uso de *metatags* e, portanto, o *cache* de armazenamento não estará disponível no motor de busca.

Isto significa que o armazenamento em *cache* não é suficiente para a recuperação destes conteúdos e que é preciso que haja uma forma sistematizada de arquivamento e recuperação das informações publicadas na internet. O arquivamento da *web* a partir de tecnologias de coleta, preservação e acesso aos *websites* e o estabelecimento de políticas de arquivamento, que implica a seleção dos conteúdos, os recursos utilizados, os fluxos de

trabalho e tudo o que envolve o ciclo de vida desta atividade são considerados fator primordial na constituição da memória digital *web* e objeto de análise deste estudo.

Uma das dificuldades apontadas para que os arquivistas da *web* executem as atividades de preservação digital diz respeito a algumas tecnologias, linguagens ou plataformas utilizadas na internet. Segundo estudo de Brunelle *et. al* (2016), algumas redes sociais, como o *Twitter*, possuem problemas de arquivamento e em alguns testes realizados chegou a apenas 4,2% de conteúdos perfeitamente arquivados. Outra dificuldade apontada pelos autores diz respeito à linguagem *Javascript*, que por utilizar *scripts* executados do lado do cliente (*client-side*), carregam dados sem alteração do Identificador Uniforme de Recurso (URI) ou exigem interação do usuário, dificultando os métodos de automação de coleta da *web*.

Outro problema apontado por vários especialistas diz respeito ao arquivamento dos conteúdos das redes sociais. Além do volume informacional produzido nestas plataformas, elas possuem uma arquitetura fechada e dependem de parcerias com as empresas desenvolvedoras. Entretanto, mesmo com estas dificuldades, há alguns estudos de caso que importa destacar.

Nas eleições federais canadenses de 2015 utilizou-se a coleta e arquivamento de *tweets* como estratégia para documentar informações do evento, com o uso de uma *hashtag* específica, #elxn42, referindo-se a 42ª eleição federal canadense (RUEST, MILLIGAN, 2016), configurando-se em uma forma colaborativa de registrar as informações e produzir uma memória do respectivo evento.

Arquivos do *Twitter,* de perfis governamentais, são realizados pelo Arquivo Nacional do Reino Unido, como uma forma de preservar registros públicos, assim como os canais oficiais ligados às Olimpíadas e Paralimpíadas de Londres 2012 (THE NATIONAL ARCHIVES, 2017).

3.2. Políticas e tecnologias de preservação digital de páginas *web*

Xie *et. al* (2013) descreve a importância histórica, cultural e intelectual do arquivamento da *web* como amplamente reconhecida, onde todos países com alta taxa de penetração na Internet estabeleceram iniciativas de arquivamento para rastrear e armazenar o conteúdo da *Web*, que desaparece rapidamente e que precisa ser acessada para uso a longo prazo. Entretanto, esta cobertura geográfica ainda é desigual, segundo levantamento de Rockembach (2018) o Brasil ainda não realiza o arquivamento da *web* de forma sistemática e, em toda América Latina, somente o Chile recentemente constituiu uma iniciativa própria de arquivo da *web*.

Quanto às políticas de seleção dos conteúdos *web* a serem arquivados, assim como existem iniciativas que procuram arquivar grandes quantidades de *websites*, por exemplo, domínios nacionais ou de grandes regiões, como é o caso de arquivos e bibliotecas nacionais de diversos países, ou mesmo o caso da plataforma *Internet Archive*, que procura arquivar toda a *web* mundial, também há casos de arquivamentos em menor escala, que procuram preservar *sites* institucionais ou temáticos e eventos específicos.

Os tipos de arquivamento e quanto à profundidade da coleta, podem ser classificados em seleção extensivas e intensivas, a primeira procura cobrir os domínios em seus primeiros níveis, de uma forma mais abrangente e trazendo um panorama da *web* a partir do seu arquivamento. A segunda seleção, de forma intensiva, procura concentrar-se em alguns *sites* de maneira a arquivar o máximo de níveis, incluindo outros elementos, como banco de dados. Enquanto a seleção extensiva aborda a superfície da *web*, este segundo tipo de arquivamento necessita uma maior acessibilidade aos sistemas e servidores, demanda mais trabalho, mas também possibilita a preservação, não somente dos primeiros níveis, mas de toda a hierarquia, mantendo a navegação entre os *hiperlinks* ativa no arquivo *web*. A figura 3 ilustra a comparação entre os dois tipos de arquivamento extensivo e intensivo.

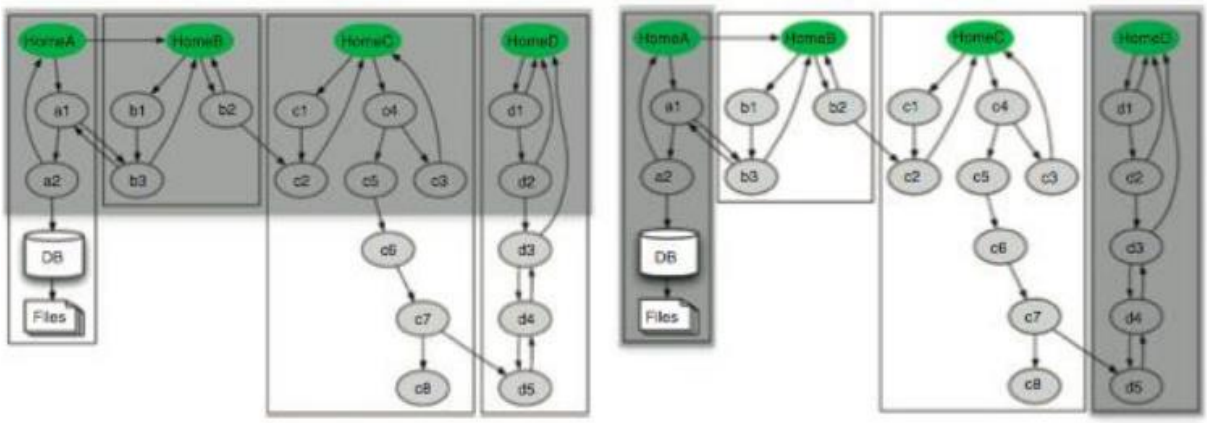

Figura 3 - Comparação entre os tipos de arquivamento extensivo e intensivo
Fonte: Masanès, 2006, p. 39-40

Conforme o planejamento estratégico 2016-2017 do Consórcio Internacional de Preservação da Internet (INTERNATIONAL INTERNET PRESERVATION CONSORTIUM, 2016), o objetivo da Instituição consiste em promover o desenvolvimento e uso de ferramentas, técnicas e padrões comuns, que permitam a criação de arquivos internacionais, estimulando iniciativas em arquivos, bibliotecas e organizações de pesquisa na abordagem do arquivamento da *web*.

Alguns dos pontos chaves, tanto políticos, quanto tecnológicos, no planejamento estratégico do Consórcio dizem respeito à promoção da interoperabilidade entre as ferramentas digitais, com o desenvolvimento modular de software e conjuntos de *Application Programming Interface* (API) para toda cadeia de arquivamento da *web*, o incentivo a estudos de caso e a criação de redes colaborativas e de engajamento social, com o enfoque de permitir a coleta dos conteúdos da internet de todo o mundo para ser arquivado, protegido, acessado e preservado ao longo do tempo.

Dentre as políticas de seleção de conteúdos analisadas no portal do Consórcio Internacional de Preservação da Internet, foram identificados os links de 11 Instituições membros do Consórcio (Bibliothèque Nationale de France, Library of Congress, British Library, The National Archives - UK, National Library of Finland, Portuguese Web Archive, Swiss National Library, Austrian National Library, Columbia University Libraries, Stanford University Libraries, KB National Library of the Netherlands) e sete outras Instituições (Tamiment Library, Bentley Historical Library, North Carolina State Government Website Archives and Access Program, University of Texas San Antonio, University of Alberta Library, University of California Los Angeles e Chesapeake Digital Preservation Group).

Concentramos as análises sobre as Instituições membros do Consórcio, identificando e correlacionando padrões em suas políticas de arquivamento da *web*. Os *links* relativos às políticas de arquivamento das Bibliotecas Nacionais de Suíça e Áustria não foram encontradas *online*.

A Biblioteca Nacional da França (BIBLIOTHÈQUE NATIONALE DE FRANCE, 2017), utiliza como fundamento jurídico o depósito legal da *web*, dando prioridade as coletas automáticas em massa por robôs rastreadores, embasada na Lei do Patrimônio Francês (*Code du patrimoine*) com a lei DADVSI (*Droit d'auteur et droits voisins dans la société de l'information* - Lei nº 2006-961). Também estipula que nenhum obstáculo (*login*, senha ou outra forma de restrição de acesso, pode ser usado pelos produtores para restringir esse processo. As responsabilidades de depósito na *web* são divididas entre o Institut National de l'Audiovisuel (INA) para comunicações audiovisuais, sobretudo rádio e TV, e a Biblioteca Nacional da França para todos os demais *websites*.

Na Biblioteca do Congresso dos Estados Unidos (LIBRARY OF CONGRESS, 2017), o enfoque é a coleta de *site*s selecionados e seus conteúdos em vários formatos para uso do Congresso dos EUA, pesquisadores e público em geral. Os critérios estabelecidos para a coleta são: utilidade no atendimento às necessidades informativas atuais ou futuras do Congresso e pesquisadores, informações exclusivas fornecidas e conteúdo acadêmico, em risco de perda (devido à natureza efêmera de alguns *site*s). Desta forma, a coleta configura-se em conteúdos

muito variados, que vão desde *sites* institucionais, de forma sistemática na esfera federal e não-sistemática na esfera estadual, a *sites* de notícias e *sites* de história em quadrinhos (*cartoons*), estes últimos, por exemplo, por meio de recomendação de *sites* a serem arquivados. *Site*s estrangeiros também são coletados se houver interesse para os cidadãos dos EUA, mas de forma mais seletiva, evitando duplicação de esforços, caso já haja este arquivamento pelos países produtores dos *site*s. Ressalta-se ainda que devem ser considerados o custo do trabalho e dos requisitos de seleção, catalogação, atendimento, armazenamento e preservação na decisão de colecionar *site*s.

A Biblioteca Britânica (BRITISH LIBRARY, 2014) iniciou o programa de arquivamento em 2004, com o *Open UK Web Archive*, com a permissão dos proprietários dos *site*s. Em 6 de abril de 2013 a legislação de depósito legal não imprimível (*Non-print Legal Deposit legislation*), juntamente com o depósito legal (*Legal Deposit Libraries*), permitiu preservar e fornecer acesso a todo domínio da *web* do Reino Unido. O processo compreende a coleta automatizada por meio de rastreador (*crawler)* e um instantâneo (*snapshot*) capturado pelo menos uma vez ao ano, com exceção aos *site*s selecionados por seu valor de pesquisa, que possuem um arquivamento mais frequente, garantindo que conteúdos atualizados rapidamente sejam preservados.

O Arquivo Nacional do Reino Unido (THE NATIONAL ARCHIVES UK, 2012) defende que cada vez mais os registros terão origem no meio digital e isso inclui desde arquivos de computador, a documentos digitalizados e *websites*. Nas suas políticas de coleta de registros consta a interação do Estado com a vida de seus cidadãos, incluindo os *site*s de todos os órgãos do Governo Central do Reino Unido, suas agências e órgãos públicos não departamentais. Ainda ressalta que alguns registros podem ter sido coletados ou duplicados em outros lugares, portanto, não se procurará adquirir coleções duplicadas.

Na Biblioteca Nacional da Finlândia (THE NATIONAL LIBRARY OF FINLAND, 2010) a coleta baseia-se em uma varredura geral da *web* finlandesa, incluindo *site*s de hospedagem com nomes finlandeses, estar localizado no país ou ainda hospedar conteúdos destinados a serem utilizados na Finlândia. A política de preservação foi adotada em 2009 e define princípios gerais relevantes para o arquivamento da *web*, como o aspecto original da publicação como um importante fator informacional, a atenção a preservação desde o momento que um objeto digital é criado e assegurar o armazenamento adequado. Ainda coloca que a preservação ao nível do *bit* (*bit level*) deve ser assegurada por métodos adequados de armazenamento, *backup* e somas de verificação (*checksums*), que os materiais e metadados devem formar pacotes de informações independentes, sendo que os metadados

devem ser tão completos quanto possível e que a migração é o método mais adequado para preservação ao longo do tempo, mantendo o original.

O Arquivo da *Web* Portuguesa (2017) dispõe o acesso aos conteúdos de forma livre e gratuita para usos educativos, científicos ou de investigação, vedado o uso comercial e de distribuição. Na reutilização dos conteúdos, deve ser citado o Arquivo da Web Portuguesa como fonte de informação.

Nas bibliotecas da Universidade de Columbia (COLUMBIA UNIVERSITY LIBRARIES, 2017) a seleção de assuntos a ser coletados é definida pelo coordenador do programa de coleção de recursos *web* em conjunto com especialistas, pesquisadores e proprietários dos *site*s. Os critérios de seleção incluem a relevância do assunto para a pesquisa atual e o ensino, o risco percebido de longevidade do *site* e a complementaridade dos *site*s com as coleções impressas, existentes nas bibliotecas da Columbia University. A coleta é realizada de forma não-intrusiva, notificando todas organizações ou indivíduos cujos *site*s são selecionados para arquivamento, abstendo-se de arquivar *site*s que não desejam ser incluídos e removendo *site*s mediante solicitação do proprietário. Ressalta ainda que, dependendo das diretrizes da coleta, a captura pode acontecer de forma semestral ou trimestral.

Nas políticas delimitadas pelas bibliotecas da Universidade de Stanford (STANFORD UNIVERSITY LIBRARIES, 2017), coloca-se que, apesar de todo o conteúdo *web* estar sob risco de perda, priorizam-se algumas categorias mais ameaçadas, como as de interesse ou propósito limitado no tempo, conteúdos sujeitos a censura governamental, assim como eventos espontâneos, incluindo desastres, revoluções e tendências de tópicos sociais, que adquirem destaque público durante algum tempo e desaparecem, tornando-se efêmeros. As restrições de recursos devem ser observadas, pois quanto maior o escopo de *site*s designados a serem arquivados, maior também será a sua complexidade e custos. Também é levantada a necessidade de identificar constantemente os arquivos da *web* existentes ao redor do mundo, procurando evitar esforços duplicados, levando em consideração as condições de acesso (*online* ou local), consoante os marcos legais de cada país.

Já na Biblioteca Nacional da Holanda (KONINKLIJKE BIBLIOTHEEK, 2017), a política de coleta dos *websites* segue uma abordagem mais seletiva, ao invés da colheita automatizada por domínio nacional, a partir da escolha de *site*s com conteúdo cultural e acadêmico, mas também *site*s com caráter inovador, que exemplifiquem tendências atuais no domínio holandês. Também são levados em consideração a relevância e popularidade para a sociedade holandesa. Restrições quanto à legislação de direitos autorais ainda não permitem que o acesso seja externo, portanto é preciso que o acesso seja realizado em salas de leitura.

Quanto às ferramentas e tecnologias utilizadas no arquivamento da *web*, analisando o *site* do Consórcio Internacional de Preservação da Internet (IIPC), verificamos que muitas das plataformas são produzidas com base em *software* livre ou em código aberto (*Open Source*), possibilitando um ambiente altamente colaborativo e de implementação de melhorias pelos usuários e desenvolvedores dos sistemas.

Destacamos alguns dos *softwares* que possuem características *Open Source* e que fazem parte dos contributos da comunidade internacional de arquivamento da *web*, a partir de suas funcionalidades e estado atual, estável (*stable*) ou em desenvolvimento *(in development*).

Na funcionalidade aquisição e coleta de *websites*, o processo mais comumente utilizado consiste em coletar os *links* a partir da automatização com o uso de rastreamento (*crawler*), direcionado pelas políticas de seleção definidas pela Instituição. Um dos exemplos de tecnologia que opera nesta funcionalidade é o *software Heritrix*, *open source* e sob licença *software* livre, que foi desenvolvido pela iniciativa *Internet Archive*, em linguagem Java, sendo utilizado por diversas iniciativas de arquivamento da *web* no mundo (HERITRIX, 2017).

Para casos específicos de coleta, como o caso dos *tweets*, identificamos no estudo de caso das eleições canadenses o uso do *Twarc*, como uma ferramenta que auxilia no arquivamento destes dados (RUEST, MILLIGAN, 2016).

Quanto ao armazenamento, destacamos dois formatos, ARC e WARC (Web ARChive). O formato ARC foi tradicionalmente utilizado por muitos anos para compressão de dados e para propósitos de arquivamento da *web*, como padrão para armazenar os dados coletados no rastreamento dos *site*s. Com o desenvolvimento de um novo formato para o arquivamento da *web*, advindo de uma extensão do formato ARC e transformado em um padrão ISO 28500:2009 (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2009), foi disponibilizado o formato WARC para novas coletas da *web*. Entretanto, devido ao legado de anos de arquivamento da *web* em arquivos ARC, é necessária a garantia de acesso e uso ao longo do tempo deste tipo de formato.

Na funcionalidade de acesso e recuperação, destacamos o *software Open Wayback*, plataforma desenvolvida colaborativamente, em Java, pelo Consórcio Internacional de Preservação da Internet e que permite reproduzir no navegador do usuário os *site*s armazenados em arquivos da *web*.

Ainda sobre este quesito, uma das grandes questões consiste na forma de busca que as plataformas de arquivamento da *web* disponibilizam para o usuário. Algumas plataformas permitem somente a busca pelo endereço eletrônico, portanto, a recuperação da informação irá coincidir com os instantâneos passados do endereço pesquisado. Esta experiência de uso pode ser vista nas plataformas *Wayback Machine*, da *Internet Archive*, ou *Time Travel*, da

plataforma Memento, que recuperará a partir da pesquisa de um Identificador Uniforme de Recurso (URI). Uma busca distinta acontece, por exemplo, no portal principal do *Internet Archive* (https://archive.org), que permite pesquisa avançada, incluindo texto e metadados.

4. Considerações finais

Garantir a acessibilidade dos conteúdos publicados na *web* implica em definir uma política de preservação, não apenas para selecionar adequadamente as técnicas de preservação e as tecnologias apropriadas, mas para determinar os conteúdos que serão, prioritariamente, preservados. Os procedimentos adequados, ainda, devem basear-se na permanência, no custo, no tempo, na qualidade e no uso da informação preservada. Além disso, tanto a adoção de políticas de seleção de conteúdos, quanto a preservação dos *sites web* traz à tona questões éticas, sociais e legais. Nestas últimas inserem-se as difíceis questões da propriedade intelectual, da proteção da confidencialidade e da privacidade.

Comparando-se o volume de informações publicadas com a capacidade de arquivamento, o que envolve recursos humanos especializados, infraestrutura tecnológica e custos financeiros, é preciso estabelecer políticas de arquivamento, o que significa, entre outras coisas, delimitar o conteúdo a ser preservado, ou seja, selecionar o que é considerado relevante. A avaliação, seleção e curadoria de conteúdos é uma questão sensível, pois implica a escolha de uma memória e o apagamento do que não permanece.

Muitas iniciativas, como o *Internet Archive*, procuram arquivar toda a *web*, porém compreendemos que somente com o esforço conjunto de diversas iniciativas espalhadas pelo mundo é que este objetivo possa se aproximar de algo factível.

O Brasil, apesar de sua grande presença *online* e uso da internet, tanto em acesso quanto na produção de conteúdos e *websites*, ainda não possui um arquivamento da *web* de forma sistematizada. Portanto, a partir da reflexão trazida nesta pesquisa, sugerimos a criação e desenvolvimento de iniciativas que deem conta da preservação digital das informações produzidas na *web*, não somente de forma global, mas também local e regionalmente.

Referências

ARQUIVO DA *WEB* PORTUGUESA. **Termos e condições**. 2017 Disponível em: http://sobre.arquivo.pt/pt/acerca/termos-e-condicoes/ Acesso em: 5 nov. 2017.

BERNERS-LEE, Tim. **Information management**: a proposal. Switzerland: CERN, 1989. Disponível em: http://cds.cern.ch/record/369245/files/dd-89-001.pdf Acesso em: 08 dez. 2017.

BIBLIOTHÈQUE NATIONALE DE FRANCE. **Digital legal deposit**: four questions about *Web* Archiving at the BnF. 2017. Disponível em:

http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html Acesso em: 10 nov. 2017.

BRITISH LIBRARY. **The British Library Collection Development Policy for *websites***. 2014. Disponível em: https://www.bl.uk/aboutus/stratpolprog/digi/webarch/bl_collection_development_policy_v3-0.pdf Acesso em: 3 nov. 2017.

BRUNELLE, J. F., KELLY, M., WEIGLE, M. C., NELSON, M. L. The impact of JavaScript on archivability. **International Journal on Digital Libraries,** v. 17, n. 2, p. 95-117, 2016. Disponível em: https://link.springer.com/article/10.1007/s00799-015-0140-8 Acesso em: 8 dez. 2017.

COLUMBIA UNIVERSITY LIBRARIES. **Web Resource Collection Program - Policies**. 2017. Disponível em: https://library.columbia.edu/bts/*web*_resources_collection/policies.html Acesso em: 6 nov. 2017.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J. The evolution of *web* archiving. **International Journal on Digital Libraries**, v. 18, n. 3, p. 191–205, 2017. Disponível em: https://doi.org/10.1007/s00799-016-0171-9 Acesso em: 8 dez. 2017.

HERITRIX. **Heritrix public wiki**. 2017. Disponível em: https://webarchive.jira.com/wiki/spaces/Heritrix/overview Acesso em: 15 nov. 2017.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. Disponível em: http://netpreserve.org Acesso em: 2 nov. 2017.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. **Strategic Plan (2016-2017)**, 2016. Disponível em: http://netpreserve.org/wp-content/uploads/2017/04/IIPC-Strategic-Plan-2016-2017.pdf Acesso em: 11 nov. 2017.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 28500:2009. Information and documentation - WARC file format**. Geneva: ISO, 2009. Disponível em: https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-1:v1:en Acesso em: 8 dez.2017

KONINKLIJKE BIBLIOTHEEK. **Web Archiving.** 2017. Disponível em: https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/*web*-archiving Acesso em: 12 nov. 2017.

LIBRARY OF CONGRESS **Collections Policy Statements Supplementary Guidelines.** 2017. Disponível em: http://www.loc.gov/acq/devpol/webarchive.pdf Acesso em: 3 nov. 2017.

MASANÈS, Julien. **Web Archiving**. Berlin, Heidelberg: Springer, 2006.

NATIONAL ARCHIVES UK. **Records collection policy**. 2012. Disponível em: http://www.nationalarchives.gov.uk/documents/records-collection-policy-2012.pdf Acesso em: 10 nov. 2017.

NATIONAL ARCHIVES UK. **Twitter Archives**. 2017. Disponível em: http://webarchive.nationalarchives.gov.uk/twitter/ Acesso em: 14 nov. 2017.

NATIONAL LIBRARY OF FINLAND. **Web Archiving in Finland:** Memorandum for the members of the CDNL, 2010. Disponível em:

http://www.doria.fi/bitstream/handle/10024/67051/*web*archivingfinland_cdnl.pdf?sequence=1&isAllowed=y Acesso em: 8 nov. 2017.

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: 8 dez. 2017.

RUEST, N., MILLIGAN, Ian. An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter. **Code4Lib Journal**, n. 32, 2016. Disponível em: http://journal.code4lib.org/articles/11358 Acesso em: 8 dez. 2017.

STANFORD UNIVERSITY LIBRARIES. **Collection Development.** 2017 Disponível em: http://library.stanford.edu/projects/*web*-archiving/collection-development Acesso em: 3 nov. 2017.

XIE, Z., VAN DE SOMPEL, H., LIU, J., VAN REENEN, J., JORDAN, R. Archiving the relaxed consistency *web*. In: ACM INTERNATIONAL CONFERENCE ON CONFERENCE ON INFORMATION & KNOWLEDGE MANAGEMENT, 22., 2013**. Proceedings.** San Francisco: ACM**,** p. 2119-2128, 2013. Disponível em: https://dl.acm.org/citation.cfm?id=2505551 Acesso em: 8 dez. 2017.

Recebido/Recibido/Received: 2017-11-20
Aceitado/Aceptado/Accepted: 2017-12-09