Usos do Arquivamento da Web na Comunicação Científica

Uses of Web Archiving in Scientific Communication

Lisiane Braga Ferreira

Universidade Federal do Rio Grande do Sul
lisianebf@gmail.com

Marina Rodrigues Martins

Universidade Federal do Rio Grande do Sul
mrodriguesmartins@gmail.com

Moisés Rockembach

Universidade Federal do Rio Grande do Sul
moises.rockembach@ufrgs.br

Resumo

Esta investigação analisa o ambiente web e as informações nele produzidas, procurando enquadrar, como objetos de estudo, o arquivamento da web, como fonte de dados de pesquisa, e a comunicação científica, como prática de disseminação do conhecimento produzido nas universidades. A metodologia delimitou-se pela pesquisa exploratória, a partir de revisão bibliográfica internacional sobre o tema e análise das iniciativas participantes do Consórcio Internacional de Preservação da Internet (IIPC) vinculadas a Universidades. Realizou-se análise qualitativa dos objetivos e projetos desenvolvidos por estas iniciativas. Conclui que o arquivamento da web é um campo ainda pouco explorado internacionalmente, principalmente dentro das universidades. E observa a carência de pesquisas na América Latina, principalmente no Brasil.

Palavras-chave: Arquivamento da web. Comunicação Científica. Ciência da Informação.

Abstract

This research analyzes the web environment and the information produced in this medium, aiming to configure web archiving as an object of study, as a source of research data, along with scientific communication, as a practice of disseminating knowledge produced in universities. The methodology was delimited as exploratory research, based on an international bibliographic review on the subject, and analysis of the Initiatives of the International Consortium for the Preservation of the Internet (IIPC) related with Universities. It uses qualitative analysis of the objectives and projects developed by these initiatives. It concludes that the Web archiving is a field still not explored enough, namely inside the universities, and it observes the lack of research in Latin America context, especially in Brazil.

Keywords: *Web archiving. Scientific Communication. Information Science.*

1. Introdução

Esta investigação procura observar e analisar o ambiente web e as informações nele produzidas, a partir da perspectiva da coleta, preservação e recuperação dos websites e outros objetos digitais também produzidos neste meio, metodologia conhecida como arquivamento da web, e os possíveis usos de seu potencial informacional e probatório.

Procuramos enquadrar, como objetos de estudo, o arquivamento da web, como fonte de dados de pesquisa, e a comunicação científica, como prática para validação e disseminação do conhecimento produzido nas universidades. Como procedimentos metodológicos, a investigação se configura como uma pesquisa exploratória e descritiva, com revisão bibliográfica e análise efetuadas a partir dos estudos de caso levantados sobre o arquivamento da web, realizados por universidades vinculadas ao Consórcio Internacional de Preservação da Internet (*International Internet Preservation Consortium* - IIPC), e sua relação com a comunicação científica.

A comunicação científica não possui uma data específica de origem, estudos apontam que ela é proveniente dos povos gregos de Atenas quando grupos se reuniam para discutir questões filosóficas entre os séculos IV e V a.c. (MEADOWS, 1999). O que de fato se pode compreender é que a ciência só existe se ela for comunicada, e sua comunicação é tão importante quanto a própria pesquisa, portanto, o processo de comunicar vem legitimar a ciência, pois possibilita que a mesma seja analisada pelos pares. (ZIMAN, 1979).

No Brasil, Pinheiro (2012) relata que a comunicação científica é entendida como uma subárea da Ciência da Informação e que, conforme Mueller (2007), tem seu princípio relacionado à necessidade de assegurar o acesso ao crescente volume de publicações científicas. Este tipo de comunicação despontou no país através da inserção de disciplinas e professores estrangeiros nos programas de pós-graduação do Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT, pioneiro na América Latina. Entre os nomes internacionais que colaboraram com esta inclusão através de suas publicações traduzidas em português estão Derek Solla Price, John Ziman e Jack Meadows. Outros professores renomados como Derek Langridge, John Joseph Eyre, Suman Datta e Jack Mills da Polytechnic também contribuíram com seus conhecimentos, ministrando aulas no país na década de 70.

Para um melhor contorno sobre o tema, podemos definir a comunicação científica como um processo que reúne uma série de atividades, incluindo produção, disseminação e uso da informação, desde a concepção do problema de pesquisa até que os resultados sejam aceitos como componentes do conhecimento científico, conforme Pinheiro (2012) baseada em Garvey (1979). Estas atividades geram uma quantidade demasiada de informação como referenciado por Meadows (1999), não apenas pela quantidade de produções e publicações como também pela variedade de temáticas pesquisadas nos diferentes campos científicos. Por causa disso, e da importância de perpetuar o conhecimento para futuras gerações, deve-se resguardar o que se deixou de herança, principalmente quando se fala na era da informação e do ambiente web. Este último se trata de um meio de comunicação formado sobre a rede de computadores (Internet) e que a cada dia se concretiza como o meio mais utilizado para publicações da sociedade moderna (Gomes, 2010).

A ideia de preservar o que é disponibilizado pela web ainda não é consenso entre os pesquisadores que estão estudando o arquivamento da mesma por três motivos principais, como diz Masanès (2006). O primeiro é a qualidade dos dados encontrados que não correspondem aos padrões de preservação. Um dos pontos desta visão acredita que seria adequada uma seleção manual dos conteúdos, mas isso é incompatível considerando a amplitude de dados na web. Podemos também, relacionar este ponto de vista com o crescimento exponencial das publicações científicas. O segundo entende que a web se autopreserva e por isso não há necessidade deste trabalho e o terceiro acredita de modo claro que não é possível arquivá-la.

De modo geral, já existem diversas iniciativas ao redor do mundo que visam o arquivamento da web, basicamente o processo envolve a identificação dos dados de interesse, a captura e o devido armazenamento para possível acesso destas informações por seus diferentes públicos. São, pelo menos, quatro continentes envolvidos com este processo, sendo os primeiros registros de ações do ano de 1996.

A partir dos estudos de caso delimitados na pesquisa, foi possível identificar os projetos em desenvolvimento de algumas universidades. Em destaque as Universidades de Harvard, Stanford e Norte do Texas que possuem portais independentes do *Archive-It*, fato que acarreta a necessidade de maiores recursos tecnológicos, de pessoal e financeiros. Ainda se

destaca o projeto *Memento - Time Travel for the Web* desenvolvido pela Biblioteca de Pesquisa do Laboratório Nacional de Los Alamos em colaboração com o Departamento de Ciência da Computação da Old Dominion, bem como o projeto *Hiberlink*, também desenvolvido pela Biblioteca de Pesquisa do Laboratório Nacional de Los Alamos.

2. As Comunidades e Sociedades Científicas

As denominadas sociedades ou comunidades tiveram um papel fundamental no desenvolvimento da comunicação científica. Formadas essencialmente por membros sócios tinham como objetivo demonstrar e comprovar suas próprias investigações e, posteriormente, abrir este conhecimento para a sociedade em geral para assim repassá-lo às futuras gerações. Conforme a obra de Meadows (1999) era através de um grupo específico de pessoas que as informações eram difundidas de modo ágil em reuniões que ocorriam regularmente. Mueller (2007) explica as comunidades científicas parafraseando Ziman (1984), definindo-as como grupos de indivíduos ligados a instituições formais, como universidades, institutos de pesquisa, sociedades científicas e também redes informais de colaboração e comunicação, como os colégios invisíveis, formados por pesquisadores que por um determinado momento estão interessados e envolvidos num mesmo problema de pesquisa.

Inicialmente existiu uma diferenciação de nomenclatura entre “academia” e “sociedade” que influenciou na periodicidade da comunicação destes grupos, porém, na contemporaneidade, como afirma Meadows (1999), todas seguem a mesma missão. Entre as primeiras fundadas estão a *Académie Française* (1635), a *Royale de Peinture et de Sculpture* (1648) e a *Royale des Inscriptions et Belles-Lettres* (1663), em Paris; e a *Royal Society* (1662), em Londres. Posteriormente, novos objetivos como controle e fiscalização profissional se agregaram às sociedades e associações, expandindo também à atividade da comunicação.

Meadows (1999, p. 11) afirma que “hoje em dia, a maioria dos sócios tem acesso a acervos adequados, ao alcance da mão, graças às bibliotecas de suas instituições”. Este histórico auxilia a entender como ocorreu todo o processo da comunicação científica desde suas origens, pois estas comunidades e a forma como se organizavam ainda refletem na forma de acesso aos resultados científicos de diferentes áreas.

2.1. Formas, Tipos e Canais de Comunicação Científica

É importante identificar qual a natureza da comunidade científica a qual está se direcionando a informação, pois ela orienta a escolha do tipo de forma e meio adequado para a comunicação. A fala e a escrita são as formas mais antigas e importantes de comunicar a ciência, uma complementa a outra e ambas são utilizadas, desde os primórdios da comunicação científica, pelos gregos. A tecnologia da prensa potencializou a difusão da informação, mantendo o formato manuscrito destinado a um público reduzido ainda nos séculos XVII e XVIII, por meio de cartas, por exemplo. Segundo Meadows (1999) isso acontecia devido à possível censura que a ciência poderia sofrer, deste modo, a circulação de informações manuscritas se tornou estratégica para obtenção de prova e testes dos pares, para que posteriormente estes dados fossem difundidos para um público maior. A partir disto, surgem as revistas científicas, os primeiros registros se deram em 1660 e se tornaram regulares com a formalização da *Royal Society* em 1662.

A bibliografia estudada indica que a comunicação científica se divide em dois tipos, a informal e a formal. O primeiro é caracterizado pela troca de informações entre pesquisadores por canais e meios não oficiais. Entre eles estão “conversas pessoais face a face, por telefone ou carta, aulas e palestras, e circulação de preprints (manuscritos ainda não publicados sobre uma pesquisa), trabalhos apresentados em reuniões” (Mueller, 2007, p. 130). O segundo é caracterizado pela troca via canais reconhecidos como oficiais, onde se incluem capítulos e edição de livros, teses, dissertações, anais de eventos científicos, artigos publicados em revistas científicas, entre outros. A autora observa que o avanço da tecnologia e dos repositórios digitais tornou essa fronteira um pouco turva, porém a divisão ainda permanece válida.

Como afirma Meadows (1999), alguns canais de comunicação são óbvios, como uma conversa face a face em um encontro. e eles também se dividem entre formais e informais, embasando o apresentado por Mueller (2007). As editoras são os principais canais formais se tratando dos impressos em papel, seguidas pelas bibliotecas e unidades de informação, que são os mais importantes compradores destas produções. Só na década de 1990, no período de um ano, as bibliotecas do Reino Unido adquiriram mais de 2 milhões de livros e por volta de 600 mil assinaturas de periódicos.

As bibliotecas são entendidas como depositários de informações passadas e presentes (que ainda estão sendo editadas) e o crescimento desta literatura científica demonstra a necessidade de se pensar o armazenamento destas para acesso futuro. Congressos e conferências se enquadram em canais informais e tem, como principal forma de interação, a fala. Como resultado destes encontros se obtém publicações (anais, livros e periódicos), que são considerados meios formais. A qualidade destes está relacionada ao processo de avaliação e pode variar conforme a orientação dos envolvidos, de acordo com o apresentado por Meadows (1999).

A tecnologia não só ampliou as formas e os canais de comunicação científica como também o modo de processar os dados por parte das editoras e bibliotecas; a consulta, a percepção e a absorção do conhecimento por parte do público, como explicam Meadows (1999) e Mueller (2007). Estes avanços trouxeram alguns benefícios como a agilidade, a amplitude e a comodidade do acesso remoto na disponibilização e na busca das informações.

No entanto, ao passo que o conhecimento científico é reescrito, novos experimentos vão sendo realizados e bibliotecas científicas ou unidades de informação são memórias em constante mutação e crescimento, como diz Ziman (1979). Esta ideia também vai ao encontro do que Meadows (1999) apresenta sobre quantidade exponencial de dados em circulação. Não apenas devido à variedade de produção, mas também a diversidade de temáticas que hoje são pesquisadas, as quais atraem cada vez mais acadêmicos e entusiastas.

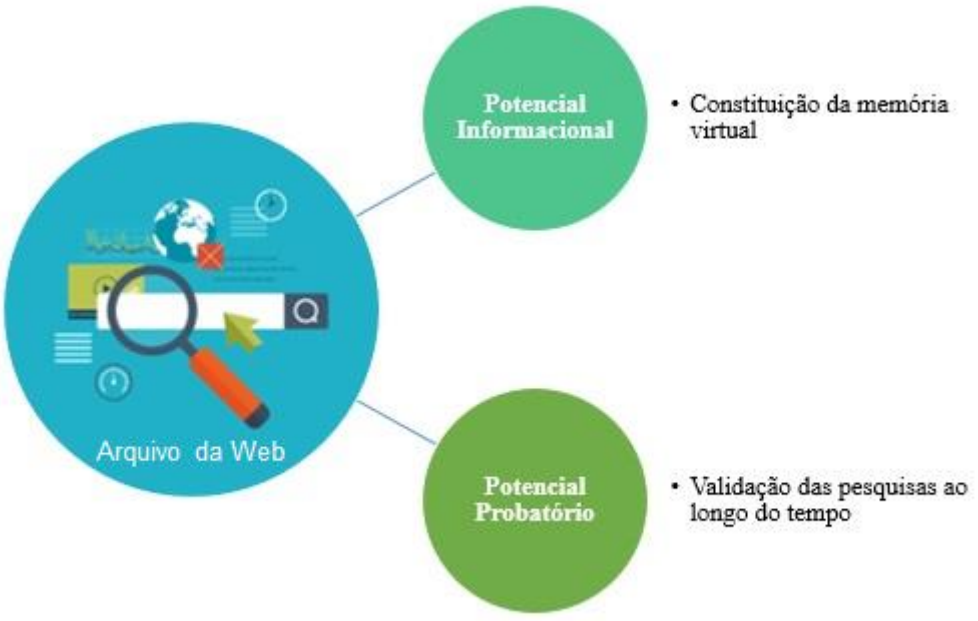
Corroborando com Klein (2014), refletimos sobre os desafios dos usos da Internet que afetam fundamentalmente a maioria dos aspectos da comunicação da informação, incluindo a comunicação científica. E, portanto, o imediatismo que caracteriza a publicação de informações na web, bem como o acesso, permite um aumento dramático na velocidade de disseminação do conhecimento acadêmico. Assim, referenciar fontes - como uma parte fundamental do discurso acadêmico - e a expectativa de que as fontes referenciadas possam e devam ser verificadas por outros, para permitir uma interpretação correta das informações que estão sendo comunicadas e para apoiar a reprodutibilidade dos resultados, passa a ser um problema recorrente no ambiente virtual.

Por este motivo, a discussão sobre o registro e o armazenamento destas informações é importante para acesso futuro e devida comunicação científica, assim, entendemos o

arquivamento da web como uma das práticas que se propõem a garantir o potencial informacional e probatório das referências que dão subsídios e sustentam as pesquisas científicas.

As iniciativas de arquivamento da web dão origem a referências persistentes, como representamos na **Figura 1**, que constituem uma memória virtual, garantindo o seu potencial informacional, bem como possibilitam a validação ao longo do tempo, garantindo o seu potencial probatório.

Figura 1: ***Arquivamento da web e as funções das referências persistentes na Comunicação Científica***



Fonte: autores, 2017

Ao nos referirmos ao potencial probatório de uma referência, estamos incorporando não apenas a possibilidade de acesso para a validação da mesma, mas também sobre a preservação de seu conteúdo, tal qual era no ato da coleta dos dados e informações para uma determinada pesquisa.

Neste sentido, e buscando novas possibilidades para o acesso e a validação da comunicação científica, universidades de diversos países vêm trabalhando no desenvolvimento de projetos de arquivamento da web.

3. Contexto do Arquivamento da Web

Dentro do contexto desta pesquisa o arquivamento da web é significativo e virá a preservar conteúdos pertinentes para o acesso e desenvolvimento de gerações futuras, pois se destaca a grande quantidade de dados disponibilizados na rede, convergindo com a ideia apresentada por Gomes (2010) citada na introdução deste artigo. Neste caso, é importante esclarecer que o arquivamento da web é um processo, uma ação contínua e prolongada que demanda regularidade em todas suas etapas. De modo simplificado, trata-se de um procedimento que identifica a informação, captura e preserva o conteúdo original das páginas, sendo fiel ao que foi postado pelos editores oficiais, que podem ser qualquer pessoa ou instituição que tenha acesso à web.

Compreende-se que para que o arquivamento da web seja executado com eficiência, eficácia e efetividade, deve-se respeitar as etapas que contemplam desde o perfil de páginas/conteúdo a ser arquivado, *hiperlinks* contidos, período e frequência de arquivamento até questões éticas e políticas envolvidas. Isto se constata importante, pois como explanam Costa, Gomes e Silva (2016) 80% do conteúdo disponibilizado na rede é alterado após o período de um ano, não mantendo seu formato original e 13% das referências *on-line* utilizadas por estudantes em suas pesquisas desaparecem em pouco mais de dois anos.

A partir da descrição acima, é perceptível o quão complexo e estratégico é o processo de arquivamento da web, porém não impraticável. Por este motivo é conveniente destacar que não é apenas uma única iniciativa que irá suprir a necessidade de arquivamento da rede, devido ao já mencionado crescimento exponencial da informação, como observam Gomes (2010) e Masanès (2006). Os autores defendem que somente a união das iniciativas existentes permitirá que os usuários usufruam dos reais benefícios do processo como um todo.

Embasados num *survey* realizado em 2010, Gomes, Miranda e Costa (2011) registraram 42 iniciativas ao redor do mundo, sobretudo concentradas na América do Norte, Europa, Oceania e Ásia. Os dados apresentados demonstraram que 24 destas mantinham finalidade nacional para a coleta de conteúdo, e os primeiros registros são datados de 1996 na Austrália, Suécia e nos Estados Unidos. Os demais escopos variam entre interesses regionais, audiovisuais, institucionais e de literatura, por exemplo, caracterizando uma perspectiva heterogênea. Realizando um levantamento a partir dos dados disponibilizados no site do *International*

Internet Preservation Consortium (2017), verifica-se que atualmente existem 52 membros, incluindo uma iniciativa no Chile, a primeira na América Latina. Em Portugal foi identificado o Projeto de Arquivo da Web Portuguesa (AWP), um serviço prestado pela Fundação para a Computação Científica Nacional, que tem como missão capturar, armazenar e preservar a informação que interessa aos portugueses ou a quem se interessar pelo que foi publicado no país.

Numa perspectiva mundial surge, em 2003, pela parceria de 12 instituições com hospedagem oficial na Biblioteca Nacional da França, o Consórcio Internacional de Preservação da Internet (do inglês *International Internet Preservation Consortium* - IIPC). Atualmente, ele conta com a participação de mais de 45 países e tem como membros bibliotecas, museus, arquivos nacionais, universitários, regionais e instituições de patrimônio cultural. Fazem parte do IIPC bibliotecas de países como Áustria, Suíça, Estados Unidos, Finlândia, Los Angeles, Portugal e Reino Unido. O Consórcio tem como missão unir organizações para coletar, preservar e tornar acessível o conhecimento da web com escopo global, nesta busca estão incluídas publicações acadêmicas, obras de arte, como também documentos governamentais. Para ser sócio é necessário demonstrar interesse ou experiência no campo de arquivamento da web. O trabalho é executado de modo colaborativo a fim de compartilhar conhecimento e continuar a desenvolver *softwares* e ferramentas adequadas para a melhor operacionalização da missão do IIPC.

Fundamentado na ideia e na atuação do IIPC se percebe a efetividade da união de forças e como isto torna o arquivamento da web praticável e com mais potencialidade de abrangência. Este ponto de vista se fortalece quando se observa uma das tecnologias utilizadas para o processo de arquivamento da rede pelas instituições que fazem parte do Consórcio, o chamado *WayBack Machine*, disponibilizada pelo *Internet Archive*. Esta última se trata de uma organização sem fins lucrativos, fundada em 1996 que fornece acesso universal ao que é coletado e utiliza o programa *Archive-It* como base para identificar as páginas importantes para captura. Conforme informações disponibilizadas no site do *Internet Archive* (2017) a organização possui em seus arquivos 279 bilhões de páginas da web, 11 milhões de livros e textos, 4 milhões de gravações de áudio (incluindo 160.000 concertos ao vivo), 3 milhões de vídeos (incluindo 1 milhão de programas de televisão), 1 milhão de imagens e 100.000 programas de software. Atualmente o programa é capaz de digitalizar e disponibilizar 1.000

livros acadêmicos diariamente e pode ser acessado gratuitamente por qualquer pessoa no mundo.

A partir desta metodologia de preservação de informações da web, abordaremos a seguir nove iniciativas que a utilizam na academia, visando a comunicação científica e o acesso amplo às informações produzidas em suas pesquisas.

4. O Arquivamento da Web na Comunicação Científica

Alguns projetos de arquivamento da web estão sendo tratados por universidades como extensões de suporte oferecidos por suas bibliotecas universitárias, neste contexto, delimitamos a pesquisa a partir da lista de membros que compõem o IIPC, sendo identificadas nove iniciativas desenvolvidas em universidades – oito localizadas nos Estados Unidos e uma na Eslováquia, as quais trazemos abaixo.

Compreendendo que a Internet tem sido, cada vez mais, o ponto de origem de um grande volume de informações, pesquisas e publicações científicas - muitas universidades voltaram seus recursos tecnológicos e seu conhecimento técnico para a preservação de conteúdos e comunicações, que vêm se perdendo com o passar dos anos. Descreveremos a seguir como algumas dessas universidades têm desenvolvido projetos em cooperação para que o conhecimento científico digital não se perca em sua totalidade.

4.1. Bibliotecas da Universidade do Norte do Texas

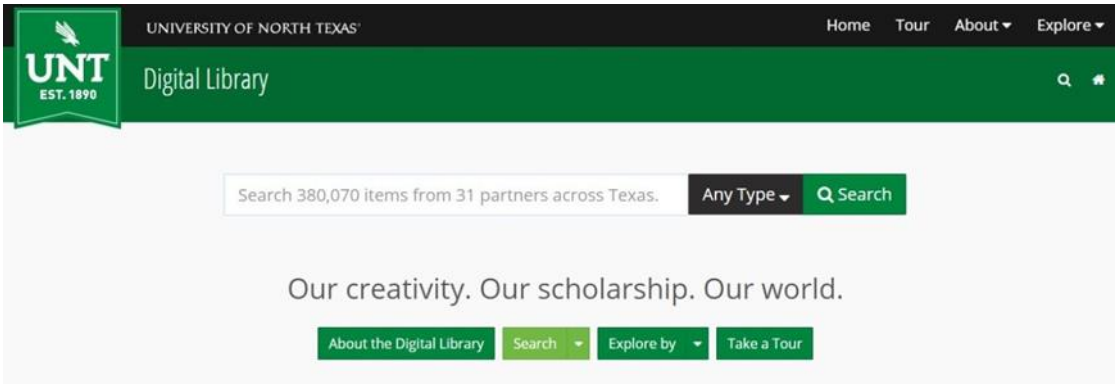
Segundo a *University of North Texas Libraries* (2017), as bibliotecas da instituição assumiram um papel ativo na preservação da Internet, coletando e arquivando domínios da web UNT, sites governamentais e outros conteúdos da web.

A Universidade do Norte do Texas (*International Internet Preservation Consortium*, 2017) foi uma das primeiras instituições acadêmicas dos Estados Unidos a arquivar sites, começando em 1997 com o *CyberCemetery*, um arquivo de sites governamentais que cessaram sua operação. Esta coleção (*University North Texas*, 2017) apresenta uma variedade de tópicos indicativos da natureza ampla das informações do governo e, em particular, sites que abordam temas que apoiam o currículo da universidade e pontos específicos do programa.

Outra coleção em destaque é a *End of Therm Web Archive*, que surgiu da colaboração das Bibliotecas da Universidade do Norte do Texas com a Biblioteca do Congresso, o Arquivo da Internet (*Internet Archive*), a Biblioteca Digital da Califórnia e o Escritório de Impressão do Governo dos EUA visando colher sites que mudariam rapidamente no final dos termos presidenciais em 2008 e 2012.

Apresentamos na **Figura 2**, o portal da Biblioteca Digital da Universidade do Norte do Texas (UNT), onde observamos que o layout da página é bem acessível e oferece a possibilidade de pesquisas, em tipos bem distintos de informações, desde dissertações e artigos até websites.

Figura 2: ***Página inicial da Biblioteca Digital da Universidade do Norte do Texas***



Fonte: UNIVERSISITY OF NORTH TEXAS, 2017

Um website arquivado (*snapshot*) dentro da coleção *CyberCemetery* representa uma fonte de pesquisa e referência com grande potencial informacional, que visa adicionar valor a diversas áreas do conhecimento, como Política, Sociologia, Ciência da Informação, Direito, entre outros. Mas também fornece um potencial probatório, o qual possibilitará a validação de uma determinada pesquisa ao longo do tempo. Ao utilizarmos fontes de referência, nascidas em meio digital, na comunicação científica, principalmente as redes sociais, blogs e websites, devemos ter consciência de que as mesmas dificilmente poderão ser validadas daqui a dez anos a partir de suas fontes originais.

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

4.2. Biblioteca Digital da Califórnia

A *California Digital Library* - CDL também foi uma das pioneiras desta lista que se envolveram no arquivamento da web a partir de 2003, segundo o International Internet Preserving Consortium (2017), com o Projeto Web em Risco (do inglês *Web-at-Risk*) - no qual a CDL desenvolveu e operou o seu Serviço de Arquivamento da Web (WAS).

Em 2010, o Projeto Web em Risco estabeleceu alguns objetivos a serem explorados junto aos pesquisadores e bibliotecários envolvidos no projeto, os quais consistiam em identificar como as necessidades da comunidade para o arquivamento da web poderiam ter mudado desde o início do trabalho; identificar como as necessidades de arquivamento da web da comunidade de pesquisa em larga escala poderiam ser abordadas; identificar/analisar como o WAS, em particular, poderia ser melhorado para melhor ajudar os estudiosos e bibliotecários e mensurar os custos futuros e o crescimento do serviço, além de pensar nas possíveis abordagens para se investir em sustentabilidade (*California Digital Library*, 2017).

A *California Digital Library* (2017), manteve o WAS por oito anos e, em 2015, optou por um trabalho colaborativo com o serviço de arquivamento da web do Archive-It, assim, migrou todos os seus clientes para esta plataforma. A equipe do WAS percebeu que a complexidade e a constante mudança da web representavam desafios significativos para o conjunto de ferramentas de arquivamento da web atual e exigia atualizações frequentes para se manter à frente. Este custo afetaria os objetivos de definir as necessidades técnicas, bem como a estrutura organizacional que poderiam garantir a criação de novas ferramentas e serviços e torná-los amplamente disponíveis em toda a comunidade.

Atualmente, o CDL está explorando oportunidades com Harvard, MIT, Stanford, UCLA e outros para trabalhar em colaboração com o *Archive-It* para criar uma lista expandida de ferramentas e serviços de valor agregado. Esta iniciativa fomenta a convergência dos serviços de arquivamento da web para uma única plataforma, enquanto individualmente cada universidade poderia trabalhar no aprimoramento dos mecanismos de pesquisa e das coleções, entre outros objetivos mais específicos.

4.3. Biblioteca de Harvard

Conforme a *Harvard Library* (2017), foi lançado em 2006 um projeto piloto de seu serviço de coleções de arquivos da web, financiado pela Iniciativa da Biblioteca Digital da Universidade (do inglês *University's Library Digital Initiative* - LDI). Este projeto foi o primeiro do LDI especificamente orientado para preservar o material "nato-digital" e, em 2009, foi lançada a interface pública do WAX.

Os gerentes de coleções, Harvard (2017), que trabalham no ambiente on-line, tem como objetivo adquirir o conteúdo que eles sempre coletaram fisicamente. Como os blogs que substituem diários, o correio eletrônico substituindo a correspondência tradicional e os materiais HTML substituindo muitas formas de garantia impressa, assim, os gerentes de coleções estão cada vez mais preocupados com possíveis lacunas na documentação do patrimônio cultural de Harvard.

Portanto, o WAX foi desenvolvido como uma resposta inicial e parcial a essas e outras preocupações, que vão desde a viabilidade técnica até implicações legais e financeiras. O piloto concentrou-se na colheita de conteúdo da superfície da web - conteúdo que é descoberto para os motores de busca por meio de rastreadores da web, em oposição ao conteúdo escondido dos rastreadores da web em um banco de dados ou restrito por senha ou proteção de *login*.

O serviço de arquivamento WAX possui cinco grandes coleções, (Harvard, 2017), que estão divididas em: *"H-Sites: Harvard life and learning", "SL Sites: Archived Websites from Schlesinger Library Collections", "Capturing Women's Voices on the Web", "Constitutional Revision in Japan Research Project"* e *"A-Sites: Archived Harvard Websites"*.

A *"H-Sites: Harvard life and learning"*, foi criada com a proposta de coletar os materiais, agora criados em sites, que complementam os arquivos pessoais de indivíduos e registros de organizações afiliadas à Universidade. Visa preservar os interesses intelectuais e sociais de um segmento da comunidade de pessoas que vivem, trabalham e aprendem em Harvard. Para a Universidade, as vidas dentro e fora das salas de aula e escritórios são parte integrante da cultura e da história de Harvard.

A *"SL Sites: Archived Websites from Schlesinger Library Collections"* é uma coleção de sites criados e mantidos por organizações e indivíduos cujas coleções são mantidas na Biblioteca Schlesinger. Estes sites completam as coleções baseadas em papel e representam documentação adicional sobre as atividades e contribuições importantes dessas organizações e indivíduos.

A *"Capturing Women's Voices on the Web"* é uma coleção com a missão de capturar as vozes das mulheres cujos pontos de vista não podem ser encontrados em outros lugares, bem como, documentar o uso de blogs e outras formas de publicação na web pelas mulheres americanas no início do século XXI.

A *"Constitutional Revision in Japan Research Project",* coleção que originou-se a partir do Projeto de Pesquisa de Revisão Constitucional do Japão, criado em 2005, onde são realizadas reuniões para discutir, analisar e documentar o processo de revisão constitucional no Japão. Além das reuniões, são arquivados materiais digitais relevantes de vastas fontes relacionadas à revisão constitucional e importantes para o projeto. Uma vez que a informação sobre as atividades atuais de indivíduos e grupos envolvidos na questão é principalmente gerada em meio digital, uma seleção de cerca de oitenta sites relacionados é coletada periodicamente para garantir que o debate e o processo de revisão constitucional sejam preservados e disponibilizados para os estudiosos. Este projeto, considerando a história entre os países, serve como exemplo de como é possível desenvolver o diálogo entre culturas distintas para se chegar a um ponto de equilíbrio e consenso e é um exemplo de arquivamento que contribui para a pesquisa em diversas áreas da Ciência.

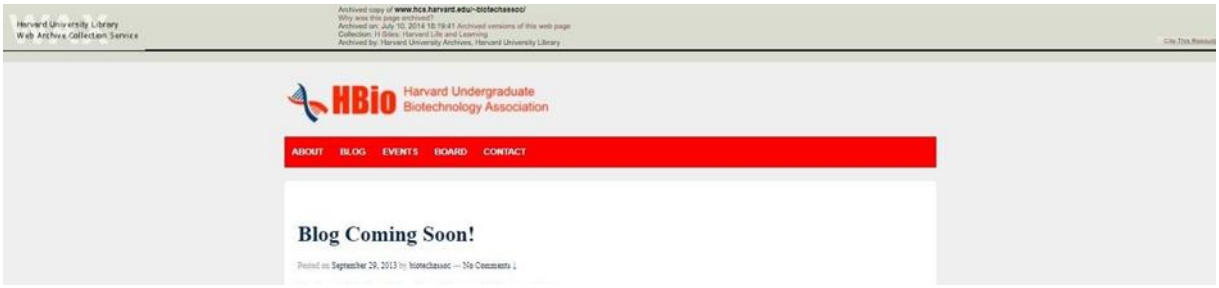
A coleção *"A-Sites: Archived Harvard Websites"*, criada a partir de 2007, consiste na preservação de sites de concessão de diplomas de departamentos e comitês da Faculdade de Artes e Ciências. Estas informações, antes preservadas apenas em formato físico pelos Arquivos da Universidade, agora coletam sites de toda a Instituição, com periodicidades mensais ou anuais. Assim, a lista de sites coletados continua a crescer, proporcionando uma imagem mais completa do espaço web de Harvard.

O serviço de arquivamento da Biblioteca de Harvard (SWAP) é disponibilizado para toda a comunidade acadêmica e não acadêmica de forma aberta e gratuita. As coleções descritas são formadas por listas de sites arquivados, podendo ter mais de uma versão arquivada para cada

site. Estes sites são fontes de informação e referências que se não fossem preservados poderiam prejudicar a validação das pesquisas que as utilizaram, uma vez que a prova delas tende a se perder com o tempo.

E, como mostra a **Figura 3**, ao selecionar uma *snapshot* específica o site arquivado é carregado com as funcionalidades originais da época em que a captura foi realizada. No cabeçalho do site estão descritos alguns dados sobre a captura e a qual coleção pertence.

Figura 3: ***Lista de sites arquivados da coleção H-Sites: Harvard life and learning***



Fonte: *HARVARD LIBRARY*/WAX, 2017.

A coleção possibilita a pesquisa por palavras-chave e termos dentro dos sites antes de acessar qualquer versão e ainda disponibiliza o link atual da página para acesso.

4.4. Bibliotecas da Universidade de Stanford

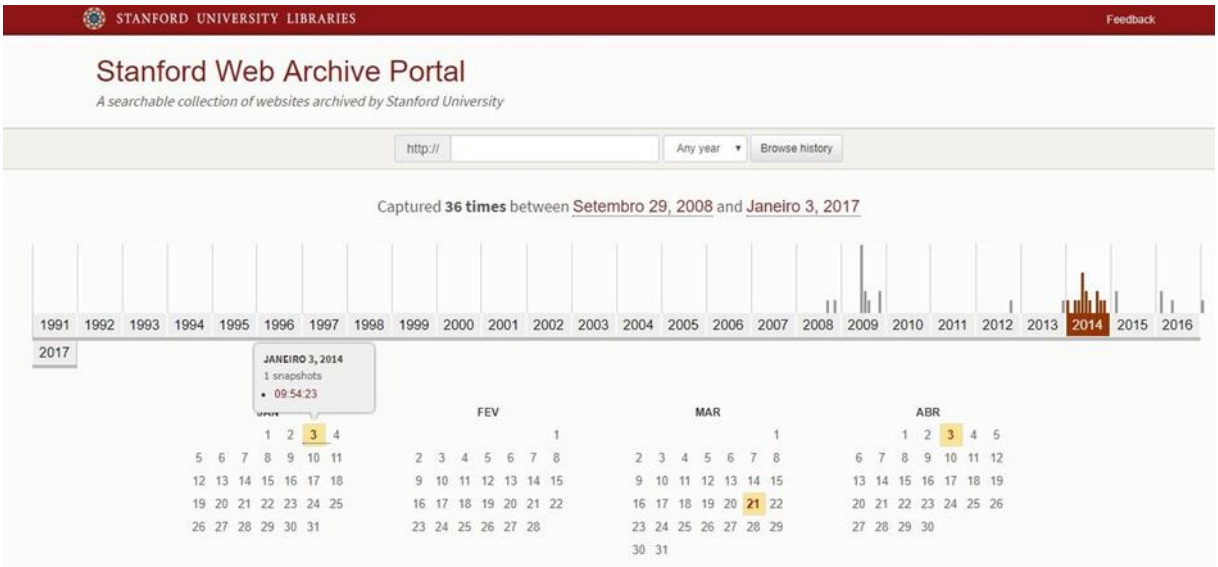
A *Stanford University Libraries* - SUL (2017), em colaboração com bibliotecários, professores, pesquisadores e outros funcionários da Universidade de Stanford, visa identificar os conteúdos selecionados da web para realizar o seu arquivamento. Os objetivos da SUL consistem em armazenar os arquivos da web no Repositório Digital de Stanford, fornecer descobertas através do *SearchWorks* e habilitar a navegação através de uma instância local da plataforma do *Wayback Machine*.

Segundo o *International Internet Preservation Consortium* (2017), a SUL iniciou seu envolvimento com o arquivamento da web em 2007, focando em informações governamentais e institucionais da web. E após uma pesquisa realizada na Universidade, entre 2011 e 2012, sobre o potencial uso deste recurso, em 2013 começou a ser construído um programa de arquivamento da web mais abrangente e a iniciativa passou a fazer parte do IIPC.

A *Stanford University Libraries* (2017) lista alguns dos fatores que motivaram o seu envolvimento no arquivamento da web, entre eles: sites de campanha de candidatos políticos, que estão disponíveis apenas durante a temporada eleitoral; sites de projetos financiados, que ao deixarem de receber os subsídios, são retirados da web, mesmo com o projeto em andamento; discurso político dissidente sujeito à censura governamental; notícias on-line relacionadas a eventos rápidos, que da mesma forma são rapidamente alterados e submersos; presença na web de personagens importantes falecidos; entre outros.

Na página inicial do Portal de Arquivamento da Web de Stanford (*Stanford Web Archive Portal* - SWAP), a pesquisa é realizada a partir da URL de interesse. E, na **Figura 4**, podemos visualizar como são apresentados os resultados da pesquisa por URL, onde, a partir de um calendário, é possível selecionar por ano, mês e dia a *snapshot* desejada.

Figura 4: ***Página inicial da Stanford Web Archive Portal - SWAP***



Fonte: STANFORD UNIVERSITY LIBRARIES/SWAP, 2017.

Assim como no WAX de Harvard, o site arquivado é apresentado com um cabeçalho contendo a data de arquivamento. No caso do SWAP, é possível navegar pelos demais *snapshots* nesta mesma barra/cabeçalho.

Segundo a SUL (2017), assegurar a continuidade da capacidade de acesso ao conteúdo da web que desapareceu ou foi substituído, está diretamente relacionado com objetivos diversos, de pesquisa, ensino, construção de coleções de bibliotecas, legado institucional, conformidade legal e administração de informações governamentais.

4.5. Biblioteca de Pesquisa do Laboratório Nacional de Los Alamos e Departamento de Ciência da Computação da Old Dominion

A equipe de prototipagem da *Los Alamos National Laboratory Research Library* - LANL (2017), explora vários aspectos da comunicação acadêmica na era digital, com foco principal na infraestrutura e interoperabilidade da informação e persistência a longo prazo do registro acadêmico. Dois projetos associados com o arquivamento da web ganharam destaque dentro da iniciativa: o *Hiberlink* e o *Memento - Time Travel for the Web*.

O Projeto *Hiberlink*, Klein (2014), utiliza o termo "*reference rot*" para denotar a combinação de dois problemas envolvidos no uso de referências URI, que se relacionam com a natureza dinâmica e efêmera da web: o ***link rot***, onde o recurso identificado por um URI pode deixar de existir e, portanto, uma referência URI para esse recurso deixará de fornecer acesso ao conteúdo referenciado; e o ***content drift*,** onde o recurso identificado por um URI pode mudar ao longo do tempo e, portanto, o conteúdo no final do URI pode evoluir, até mesmo deixando de ser representativo do conteúdo originalmente referenciado.

O Projeto *Memento - Time Travel for the Web* (2017) tem sido desenvolvido em colaboração com o *Old Dominion Department of Computer Science*. Este projeto consiste em um protocolo, que adiciona uma dimensão de tempo ao protocolo HTTP. Inspirado na negociação de conteúdo HTTP, o protocolo introduz a noção de negociação de data e hora que permite que um cliente solicite a versão de um recurso tal como existia em um tempo específico no passado.

4.6. Bibliotecas da Universidade da Columbia e Biblioteca de Pesquisa da UCLA

As Iniciativas da *Columbia University Libraries* e UCLA Research Library são parceiras na utilização da plataforma e do serviço do Archive-It para coletar e arquivar suas coleções. Contudo, nos sites institucionais existem poucas informações a respeito de suas coleções - sendo necessária a pesquisa direta no *Archive-It* pelo nome da instituição.

4.7. Biblioteca da Universidade de Bratislava

A Iniciativa mais recente a incorporar o IIPC (2017), a *Univerzitná knižnica v Bratislave*, a partir de 2015, começou a executar o arquivamento de domínios nacionais da web e de documentos nascidos no ambiente digital, mas ainda não se encontra nenhuma plataforma disponível para acesso aberto, bem como a localização de nenhuma coleção no *Archive-It*.

Em uma análise geral, percebemos que as iniciativas de arquivamento da web preocupam-se tanto com os aspectos informacionais, os quais são definidos como a formação da memória virtual - este muito mais evidente nos projetos, quanto com os aspectos probatórios desta informação que se origina, tramita e é comunicada na web e pode vir a ter o seu acesso interrompido ou alterado.

5. Considerações Finais

A sociedade da informação, as tecnologias, a variedade de produção e de temáticas de pesquisa colaboram para o crescimento contínuo das publicações científicas na web fazendo a comunicação da ciência atingir um novo patamar em sua abrangência e disseminação. Estes são alguns dos motivos que despertaram o interesse de estudiosos da Ciência da Informação acerca de novas fontes de referência, como é o caso do arquivamento da web para a comunicação científica.

Iniciativas nesta perspectiva são realidade há pouco mais de duas décadas e demonstram a importância de se pensar o armazenamento dos dados publicados neste meio de comunicação. Estas ações propõem - pouco a pouco - coletar partes criteriosamente selecionadas de informações formando uma memória coletiva, o mais fiel e completa possível, de uma determinada comunidade, a fim de preservar e garantir o conhecimento para gerações futuras.

No entanto, observa-se que é necessário rever e ampliar os métodos do processo de arquivamento da web para a ciência, a fim de garantir a validação das pesquisas atuais, onde as fontes por vezes se limitam ao ambiente virtual. Para tanto, a aferição das referências deve ser assegurada pelos mecanismos já existentes de preservação da web para que, no futuro, tais investigações possam seguir válidas ou postas a prova em função da evolução das fontes e dados e não por falta de acesso ou por inconsistência dos links.

Assim, se constata que o arquivamento da web é um campo ainda pouco explorado internacionalmente, principalmente dentro das universidades e do campo científico. E ainda se observa a carência de pesquisas na América Latina, principalmente no Brasil sobre esta temática.

Referências Bibliográficas

BUENO, W. C. (2010). Comunicação científica e divulgação científica: aproximações e rupturas conceituais. *Informação & Informação*, 15, 1-12. doi: 10.5433/1981-8920

COSTA, M; GOMES, D; SILVA, M.J. (2016). The evolution of web archiving. *International Journal on Digital Libraries*, 1-15. doi: 10.1007/s00799-016-0171-9

GARVEY, W. D. (1979) *Communication: essence of science; facilitating information exchange among librarians, scientists, engineers and students*. Oxford: Pergamon Press.

GOMES, D.(2010.) Preservar a Web: um desafio ao alcance de todos. In: *Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas*. (pp. 1-9). Lisboa, POR: Retirado de http://sobre.arquivo.pt/wp-content/uploads/PreservarAWebBADFormat-v.14.pdf

GOMES, D.; MIRANDA, J. & COSTA, M. (2011). A survey on web archiving initiatives. In: *International Conference on Theory and Practice of Digital Libraries*. (pp. 408-420). Lisboa, POR: Springer Berlin Heidelberg. Retirado de https://link.springer.com/content/pdf/10.1007%2F978-3-642-24469-8_41.pdf

MASANÈS, J. (2006). Web Archiving: Issues and Methods. In Julien Masanès, *Web Archiving* (pp. 1-46). Paris, FRA: Springer-Verlag Berlin Heidelberg.

MEADOWS, A. J. (1999). *A Comunicação Científica.* Brasília, DF: Briquet de Lemos.

MUELLER, S. P. M. (2007). Literatura Científica, Comunicação Científica e Ciência da Informação. In: TOUTAIN, L. B. (Org.). *Para entender a Ciência da Informação.* (pp. 125-144). Salvador, BA: Editora da Universidade Federal da Bahia. Retirado de: https://repositorio.ufba.br/ri/bitstream/ufba/145/1/Para%20entender%20a%20ciencia%20da%20informacao.pdf

PINHEIRO, L. V. R. (2012). Constituição epistemológica e social da comunicação científica no Brasil. In: PINHEIRO, L. V. R.; OLIVEIRA, E. da C. P. de (Orgs.). *Múltiplas facetas da comunicação e divulgação científicas: transformações em cinco séculos*. (pp. 115-148) Brasília, DF: Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT. Retirado de: http://livroaberto.ibict.br/handle/1/711

ZIMAN, J. (1979). *Conhecimento público.* Belo Horizonte, MG: Itatiaia.

ZIMAN, J. (1984). *An introduction to science studies: The philosophical and social aspects of science and technology*. Cambridge, GB: Cambridge University Press.

KLEIN, M.; SOMPEL, H. V. de; SANDERSON, R. et al. (2014). *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot.* Retirado de: https://doi.org/10.1371/journal.pone.0115253

Hiperlinks

ARQUIVO.PT. Retirado de: http://arquivo.pt/ Versão de referência arquivada no Archive.is: http://archive.is/CIHTO

HARVARD LIBRARY. WEB ARCHIVE COLLECTION SERVICE (WAX). Retirado de: http://wax.lib.harvard.edu/collections/about.do?kind=about&lang=eng Versão de referência arquivada no Archive.is: https://archive.is/O6aev

HARVARD LIBRARY. Constitutional Revision in Japan Research Project. Retirado de: http://wax.lib.harvard.edu/collections/collection.do?coll=101&lang=eng Versão de referência arquivada no Archive.is: http://archive.is/uOCGf

IMAGEM REFERÊNCIAS, FIGURA 1. Retirado de: https://comunidade.rockcontent.com/o-que-sao-referencias-na-hora-de-escrever-um-texto/ Versão de referência arquivada no Archive.is: https://archive.is/ZFf84

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC). Membros. Retirado de: http://netpreserve.org/about-us/members/ Versão de referência arquivada no Archive.is: http://archive.is/RLaVv

INTERNET ARCHIVE. About.Retirado de: http://archive.org/about/ Versão de referência arquivada no Archive.is: http://archive.is/xokXX

INTERNET ARCHIVE. Wayback Machine. Retirado de: https://archive.org/web/ Versão de referência arquivada no Archive.is: http://archive.is/xokXX

LOS ALAMOS NATIONAL LABORATORY. Digital Library Research and Prototyping. Retirado de: http://www.lanl.gov/library/about/research-prototyping.php Versão de referência arquivada no Archive.is: http://archive.is/8cdqA

MEMENTO - TIME TRAVEL SERVICE. Retirado de: http://timetravel.mementoweb.org/about/Versão de referência arquivada no Archive.is: http://archive.is/4Lvj9

STANFORD UNIVERSITY LIBRARIES. Stanford Web Archive Portal. Retirado de: https://swap.stanford.edu/ Versão de referência arquivada no Archive.is: http://archive.is/YcObD

STANFORD UNIVERSITY LIBRARIES. Web Archiving. Retirado de: https://library.stanford.edu/projects/web-archiving Versão de referência arquivada no Archive.is: http://archive.is/oVFbZ

UNIVERSITY OF CALIFORNIA. California Digital Library – CDL. The Web-At-Risk: Preserving Our Nations's Digital Cultural Heritage. Retirado de: http://www.cdlib.org/services/uc3/partners/webatrisk.html Versão de referência arquivada no Archive.is: http://archive.is/geMrM

UNIVERSITY OF CALIFORNIA. California Digital Library – CDL. Announcing a New Partnership: California Digital Library, UC Libraries, and Internet Archive's Archive-It Service. Retirado de: http://www.cdlib.org/cdlinfo/2015/01/14/announcing-a-new-partnership-california-digital-library-uc-libraries-and-internet-archives-archive-it-service/ Versão de referência arquivada no Archive.is: http://archive.is/bLpkE

UNIVERSITY OF NORTH TEXAS. Website. Retirado de: https://digital.library.unt.edu/ Versão de referência arquivada no Archive.is: http://archive.is/6fncP

UNIVERSITY OF NORTH TEXAS. CyberCemetery. Retirado de: https://digital.library.unt.edu/explore/collections/GDCC/ Versão de referência arquivada no Archive.is: http://archive.is/vzYxP

WEB ARCHIVING SERVICE. Website. Retirado de: https://archive-it.org/ Versão de referência arquivada no Archive.is: http://archive.is/RMAQK