XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-8 – Informação e Tecnologia

ARQUIVOS DA WEB UNIVERSITÁRIOS: ANÁLISE DAS PLATAFORMAS DIGITAIS DA UNIVERSIDADE DE HARVARD E UNIVERSIDADE DE COLUMBIA

UNIVERSITY WEB ARCHIVES: ANALYSIS OF THE DIGITAL PLATFORMS OF HARVARD UNIVERSITY AND UNIVERSITY OF COLOMBIA

Marina Rodrigues Martins - Universidade Federal do Rio Grande do Sul e Universidade do Vale do Rio dos Sinos
Moisés Rockembach - Universidade Federal do Rio Grande do Sul

Modalidade: Resumo Expandido

Resumo: expõe parte dos resultados da pesquisa de mestrado, que teve como objetivos averiguar o funcionamento das iniciativas de arquivamento da *web* implantadas pela Universidade de Harvard e pela Universidade de Columbia, concomitante à identificação dos perfis das coleções preservadas. A escolha dos objetos ocorreu pelo cruzamento de dados das primeiras etapas das pesquisas documental e bibliográfica. O estudo concluiu que as coleções são concebidas por várias organizações; o escopo varia conforme os interesses de cada organização coletora; apresentam diferentes classificações: institucional, regional, nacional etc.; preservam a história de suas entidades e promovem o ensino, a pesquisa e a extensão de suas comunidades.

Palavras-Chave: Arquivamento da *web*; *Preservação digital*; Arquivo da *web* universitário.

Abstract: *exposes part of the results of the master's research, which aimed to verify the functioning of web archiving initiatives implemented by Harvard University and Columbia University, concomitantly with the identification of the profiles of the preserved collections. The objects choice occurred by crossing the data from the first stage of documentary and bibliographic research. The study concluded that web archives collections are designed by various organizations; their scopes are according to the interests of each one and them present different classifications: institutional, regional, national, etc.; preserving the history of their entities and promoting the teaching and research of their communities.*

Keywords: *Web archiving; Digital preservation; University web archive.*

1 INTRODUÇÃO

O arquivamento da *web* é o processo que envolve seleção, coleta, armazenamento e recuperação de *websites.* Tem como objetivo preservar parte dos diversos conteúdos digitais presentes na *World Wide Web* (WWW). Os URLs - *Uniform Resource Locator* - capturados e preservados classificam o perfil das coleções. Os perfis podem compreender: acadêmico-científico, institucional, nacional, internacional. Podem servir para constituir uma memória organizacional, de fatos e eventos específicos, conforme os interesses das organizações promotoras e coletoras.

A tecnologia para preservação é desenvolvida e aplicada por diferentes tipos de organização, principalmente, localizadas na Europa e na América do Norte. Descrever duas iniciativas de arquivamento da *web,* promovidas por organizações universitárias nos Estados Unidos da América, possibilitou compreender essas ações no âmbito acadêmico e demonstrar os diferentes perfis de conteúdos preservados. De modo geral, podemos dizer que as coleções observadas preservam informações que beneficiam o ensino e a pesquisa de suas comunidades.

2 O ARQUIVAMENTO DA *WEB*

A necessidade de arquivamento da *web* foi reconhecida no final de 1990. Os primeiros projetos datam dos últimos 20 anos e estão em contínua expansão. Exigem novos estudos, abordagens, ferramentas de preservação, armazenamento e acesso. Entre as organizações que desenvolvem a tecnologia, no exterior, estão entidades não governamentais, públicas e de ensino. Ferreira, Martins e Rockembach (2018) demonstram que algumas iniciativas universitárias têm interesse na disseminação do conhecimento produzido no âmbito acadêmico. As políticas e as tecnologias de arquivamento estão vinculadas aos objetivos e aos interesses das organizações que empregam o processo (INTERNATIONAL STANDARD ORGANIZATION, 2012; BRAGG; HANNA, 2013).

Na maioria dos casos, a motivação para iniciar o arquivamento da *web* tem sido preservar sites institucionais ou governamentais; coleções especiais e/ou baseadas em eventos (jogos olímpicos ou campanhas eleitorais) (FERNANDO; MARENZI; NEJDL, 2018). Os escopos de coleta de conteúdo variam em suas classificações, como por exemplo, institucionais, regionais, nacionais e internacionais (GOMES; MIRANDA e COSTA, 2011).

A captura e a preservação incluem grande variedade de conteúdos disponíveis na rede: textos, imagens, filmes, sons, páginas da *web* interligadas, etc. (INTERNATIONAL STANDARD ORGANIZATION, 2012; TOYODA e KITSUREGAWA, 2012). A captura destes materiais pode ser extensiva, arquivando maior quantidade de *websites*, em um nível superficial; como também intensiva, capturando em menor escala, mas em maior profundidade de navegação (MASANÈS, 2006).

Organizações de grande porte são mais estruturadas para aplicação da tecnologia e formação das coleções. Para organizações menores, arquivar conteúdo *online* ainda é visto como desafio (FERNANDO; MARENZI; NEJDL, 2017/2018). O *Web Archiving Life Cycle Model* apresenta fases importantes para o processo de implantação de iniciativas de arquivamento da *web*, em diferentes estruturas organizacionais. Dentre elas estão: clareza dos objetivos do programa; escopo e tipos de arquivos para coleta; armazenamento das coleções e gerenciamento de riscos, ligado aos direitos autorais dos dados coletados (BRAGG; HANNA, 2013).

A diversificada gama de iniciativas existente varia em seus métodos e abordagens para selecionar, adquirir, organizar, armazenar, descrever e fornecer acesso aos conteúdos arquivados pelas plataformas. Essa variação é causada por fatores externos, como o ambiente legal e as relações entre os produtores dos recursos da *web* e as iniciativas de arquivamento da rede; e também internos, como o escopo a ser arquivado, a natureza da organização que aplica a tecnologia, a escala em que será feita a coleta e a capacidade técnica e financeira de cada organização (NIU, 2012).

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa exploratório-descritiva contou com levantamento bibliográfico nas bases Scopus, *Web of Science*, Google Scholar e SciELO Citation Index; e documental no *website* do Consórcio Internacional de Preservação da Internet (IIPC). A busca pelo termo "*web archiving initiatives*" contemplou publicações de 2010 a 2018 e tinha como propósito compreender o contexto do processo de arquivamento da *web*. A opção pelo termo na língua inglesa ocorreu devido à transnacionalidade e emergência do tema de arquivamento da rede no Brasil. Os dados sobre iniciativas e organizações promotoras foram cruzados, dando origem aos objetos de observação do estudo. As iniciativas da Universidade de Columbia e da Universidade de

Harvard integram o IIPC e também fazem parte da lista das 42 iniciativas citadas por Gomes, Miranda e Costa (2011).

A segunda etapa consistiu na coleta de dados sobre as universidades e suas plataformas de arquivamento da *web*. As fontes - estatutos, regimentos, organogramas, comunicações e mensagens de *websites* - foram escolhidas a partir da regra de pertinência uma vez que corresponderam aos objetivos do estudo, período e procedimentos de análise (BARDIN, 2004; FONSECA JÚNIOR, 2008). O levantamento ocorreu entre 03 e 29 de outubro de 2018, totalizando 26 dias de sondagem. As categorias de dados sobre as organizações universitárias foram: classificação, missão, visão, governança e estrutura. Nas plataformas de arquivamento da *web*: coleções, descrição, direitos autorais, tempo de arquivamento, assuntos, quem coleta os conteúdos, categoria do URL.

Com as informações de cada objeto, elaboraram-se recortes contemplando as categorias escolhidas. A reunião dos dados permitiu a análise documental e o conhecimento sobre os objetos estudados, apresentando os resultados de pesquisa (BARDIN, 2004; MOREIRA, 2008).

4 RESULTADOS

Os 219 documentos - obtidos como resultado da pesquisa nas bases - forneceram conhecimento geral sobre o contexto do processo de arquivamento da *web*. Porém, observa-se que as publicações não apresentaram subsídio informativo, específico, sobre plataformas de arquivamento da *web* que preservam conteúdos digitais no âmbito universitário.

4.1 A iniciativa de arquivamento da *web* da Universidade de Harvard

A iniciativa é viabilizada por meio do *Archive-it* - *s*erviço de arquivamento da *web* oferecido pela organização *Internet Archive*. Este, possibilita que a Universidade autogerencie suas coleções, coletando, criando e preservando conteúdos digitais de interesse de seus públicos (ARCHIVE-IT.ORG, 2014).

Quadro 1: Universidade de Harvard.

Categoria	Dados
Classificação	Privada
Missão	Não tem uma declaração formal de missão, apenas alega que o corpo docente está engajado no ensino e na pesquisa para ampliar os limites do conhecimento humano.
Visão	Cada Escola e Faculdade da estrutura possui suas próprias visões.
Governança	Duas diretorias (sistema de dois blocos) associadas aos seus conselhos administrativos: 1) Corporação de Harvard (o presidente e os membros do *Harvard College*) e 2) Conselho de Superintendentes.
Estrutura	12 Escolas de Graduação

Fonte: Elaborado pelos autores.

Quadro 2: Iniciativa de Arquivamento da *Web* - Universidade de Harvard.

Categoria	Dados
Coleções	69
Descrição	Cada coleção possui uma descrição conforme seu escopo de coleta. Variam entre arquivos pessoais de indivíduos; registros de organizações filiadas à Universidade; registros históricos da Universidade; revistas/periódicos e blogs universitários; arquivos de ensino; registros de todos os departamentos administrativos; jornais estudantis, arquivos institucionais (corpo docente e discente, dados de pesquisa, ensino e extensão), mídias e redes sociais de assuntos gerais que impactam ou são da comunidade de Harvard.
Direitos Autorais	Cada coleção possui uma especificidade que depende do tipo de conteúdo preservado. Algumas possuem direitos compartilhados com a Universidade, outras possuem direitos exclusivos da organização coletora, que não necessariamente a Universidade de Harvard. Algumas coleções são livres de direito autoral, pois os conteúdos são disponíveis *online* gratuitamente.
Tempo de Arquivamento	Desde 2014 até tempo atual
Assuntos	Governo/política/eleições; artes e humanidades; sociedade e cultura; estatísticas socioeconômicas; Universidade; departamentos e livrarias de Harvard; educação; mídia social; administração e negócios; estudos de caso; história nacional; leis financeiras; lideranças sociais; cidades e comunidades nacionais e locais; pesquisa; etc.
Organizações coletoras	9
Categoria dos URLs	Internos e externos ao ambiente digital da Universidade

Fonte: Elaborado pelos autores.

As 69 coleções eram formadas pela captura de 17.083 *links* de *websites* e incluíam textos, imagens e outros formatos multimídia, páginas da *web* interligadas, boletins informativos e *blogs*. A organização coletora mais atuante é a *Harvard Business School (HBS)*, responsável por 42 das coleções identificadas. Estas coleções preservam assuntos como a história institucional da *HBS*; programas acadêmicos; fatos, datas e números sobre a Escola; cultura e lideranças locais.

4.2 A iniciativa de arquivamento da web da Universidade de Columbia

Desde 2010, a iniciativa utiliza o serviço do *Archive-it*. Através da tecnologia as bibliotecas da Universidade capturam os domínios *columbia.edu*, como também URLs

utilizados para publicações da organização e grupos de alunos. Antes disso, a própria Universidade tentou implantar seu serviço de arquivamento da *web* sem sucesso. Desde 2015, grande parte dos domínios é rastreada semestralmente. Os URLs que contêm informações sobre cursos são monitorados trimestralmente para garantir que todas as sessões publicadas sejam preservadas. A alteração da estratégia de tempo de coleta ocorreu quando se descobriu que as coleções de arquivos da *web* eram o único meio de acessar os conteúdos antigos, uma vez que as atualizações ocorriam somente de forma *online*.

Quadro 3: Universidade de Columbia

Categoria	Dados
Classificação	Privada
Missão	Um dos centros de pesquisa mais importantes do mundo e, ao mesmo tempo, um ambiente de aprendizado diferenciado para alunos de graduação e pós-graduação em muitos campos acadêmicos e profissionais. Busca atrair um corpo docente e estudantil diversificado e internacional, apoiar pesquisas e ensinar sobre questões globais. Espera que todas as áreas da Universidade promovam o conhecimento e o aprendizado no mais alto nível e transmitam os produtos de seus esforços para o mundo.
Visão	Não foi encontrada uma visão única para a Columbia em geral, cada escola e faculdade apresenta sua definição.
Governança	A governança geral da Universidade está nas mãos do Senado Universitário. Os curadores (em inglês: *Trustees*) selecionam o presidente da Universidade, supervisionam todos os cargos administrativos, monitoram o orçamento, a doação e protegem a propriedade da universidade. Sendo formado por membros da administração; das faculdades; da representação discente, das instituições filiadas; da equipe da biblioteca; da Pesquisa; dos funcionários administrativos e dos ex-alunos.
Estrutura	16 Faculdades e Escolas

Fonte: Elaborado pelos autores.

Quadro 4: Iniciativa de Arquivamento da *Web* - Universidade de Columbia

Categoria	Dados
Coleções	12
Descrição	Cada coleção possui uma descrição conforme seu escopo de coleta. Variam entre contribuições para o ensino e a pesquisa; o desenvolvimento de escolas, departamentos acadêmicos e programas, institutos e unidades administrativas; vida no campus; serviço público; o papel da Universidade na história das comunidades metropolitanas, nacionais e internacionais; revistas de ex-alunos; atletismo; centros e institutos; boletins de cursos, honras e prêmios; bibliotecas, escolas e departamentos da Universidade; organizações estudantis e publicações científicas de estudantes.
Direitos Autorais	Cada coleção possui uma especificidade que depende do tipo de conteúdo preservado. Algumas possuem direitos compartilhados com a Universidade, outras direitos exclusivos da organização coletora, que não necessariamente a Universidade de Columbia. Outras ainda são livres de direito autoral, pois os conteúdos são disponíveis *online* gratuitamente.
Tempo de Arquivamento	Desde 2009 até tempo atual
Assuntos	Arte e humanidades; sociedade e cultura; governo/política/eleições; história e preservação; espaços públicos; participação pública; direitos humanos; terrorismo; comunismo; mudanças climáticas; etc.
Organizações coletoras	3
Categoria dos URLs	Internos e externos ao ambiente digital da Universidade

Fonte: Elaborado pelos autores.

Formadas por 2.841 *websites* as 12 coleções identificadas preservavam, principalmente, textos, imagens, páginas da *web* interligadas, grupos de notícias, boletins informativos e *blogs*. Grande parte das coleções se caracteriza pelo perfil institucional. A maior delas é a *University Archives* que preserva a memória da Universidade, desde sua fundação em 1754. O objetivo desta coleção é identificar, avaliar, coletar, descrever, preservar e, quando apropriado, disponibilizar aos administradores, pesquisadores e ao público em geral registros da organização. Estes registros documentam a evolução da Universidade de Columbia em toda a sua variedade de atuação. Somente nesta coleção foram encontradas 582 subcoleções, no dia 24 de outubro de 2018. Outras coleções focam em conteúdo internacional/fato (*2015 Nepal Earthquake*) e ainda documentam atividades de resistência política (*Resistence*).

5 CONSIDERAÇÕES FINAIS

O estudo possibilitou entender que ambas as plataformas apresentam coleções que constituem uma espécie de memória *web* de suas entidades, com escopos de conteúdos organizacionais, institucionais, de fatos e eventos específicos. As coleções são concebidas por várias organizações coletoras que não apenas as entidades macropromotoras. Algumas destas organizações fazem parte das estruturas organizacionais das universidades observadas. Elas monitoram e capturam URLs externos – além dos domínios *columbia.edu* e *harvard.edu*.

As duas iniciativas são viabilizadas pelo serviço *Archive.it*, sendo que as próprias universidades selecionam os URLs a serem arquivados. As coleções apresentam diferentes escopos de coleta, conforme políticas e estratégias estabelecidas por suas equipes. Independente do perfil de organização, dos objetivos e das políticas, para garantir o sucesso dos projetos de arquivamento da *web* é fundamental estabelecer estratégias que englobam, desde a seleção do que será preservado até o acesso dos conteúdos pelos usuários.

REFERÊNCIAS

ARCHIVE-IT.ORG. **Website** [*on-line*], 2014. Disponível em: <https://archive-it.org/>. Acesso em: 23 out. 2018.

BARDIN, Laurence. **Análise de conteúdo**. 3. ed. Lisboa: Edições 70, 2004.
BRAGG, Molly; HANNA, Kristine. **The Web Archiving Life Cycle Model**, March 2013. Disponível em: http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. Acesso em: 19 jul. 2019.

COLUMBIA UNIVERSITY, in the City of New York. **Charters and Statutes**. The Charters and Statutes are maintained by The Office of the Secretary. Edition of April 6, 1959. Disponível em: https://secretary.columbia.edu/files/secretary/university_charters_ and_ statutes/UniversityStatues_December2017.pdf. Acesso em: 25 out. 2018.

CONSÓRCIO INTERNACIONAL DE PRESERVAÇÃO DA INTERNET. **Website** [*on-line*] Disponível em: http://netpreserve.org/. Acesso em: 24 abr. 2018.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J. The evolution of web archiving. **International Journal on Digital Libraries**, [s.l.], v. 18, n.3, p.191-205, set. 2017. https://doi.org/10.1007/s00799-016-0171-9

FERNANDO, Zeon Trevor; MARENZI, Ivana; NEJDL, Wolfgang. **Archive Web**: Collaboratively Extending and Exploring Web Archive Collections. **Int J Digit Libr,** [s.l.], v.19, n.1, p. 39-55, 2018. https://doi.org/10.1007/s00799-016-0206-2

FERREIRA, Lisiane Braga; MARTINS, Marina Rodrigues; ROCKEMBACH, Moisés. Usos do arquivamento da web na comunicação científica. **Prisma**, Porto, n.36, p. 78-98, 2018.

FONSECA JÚNIOR, Wilson C. da. Análise de conteúdo. In: DUARTE, Jorge; BARROS, Antonio (Org.) **Métodos e técnicas de pesquisa em comunicação**. 2. ed. São Paulo: Atlas, 2008, p. 280-303.

GOMES, Daniel. Preservar a web: um desafio ao alcance de todos. **Actas:** Congresso Nacional de bibliotecários, arquivistas e documentalistas, Lisboa, n. 10, 2010.

GOMES, Daniel; MIRANDA, João; COSTA, Miguel. A survey on web archiving initiatives. International Journal on Digital Libraries, Springer, v.18, n.3, p. 408-420, 2011.

GOMES, Daniel; MIRANDA, João; COSTA, Miguel. A Survey on Web Archiving Initiatives. *In:* GRADMANN, S.; BORRI, F.; MEGHINI, C.; SCHULDT, H. (eds.). **Research and Advanced Technology for Digital Libraries**. Berlin: Springer, 2011.

HARVARD UNIVERSITY. **Laws and statutes of Harvard**. 2019. Disponível em: https://hollisarchives.lib.harvard.edu/repositories/4/resources/4369. Acesso em: 29 jul. 2019.

INTERNET ARCHIVE. **Website.** 2018. Disponível em: https://archive.org/. Acesso em: 24 abr. 2018.

INTERNATIONAL STANDARD ORGANIZATION. **Relatório ISO**: Statistics and Quality Indicators for Web Archiving. Technical Report, 2012.

MASANÈS, Julien. **Web Archiving**. Paris, FRA: Springer-Verlag Berlin Heidelberg, 2006.

MOREIRA, Sônia. Análise documental como método e como técnica. *In:* DUARTE, Jorge; BARROS, Antonio (org.). **Métodos e técnicas de pesquisa em comunicação**. 2. ed. São Paulo: Atlas, 2008, p. 269-279.

NIU, Jinfang. An Overview of Web Archiving. **D-LIB Magazine**, [s.l.], v. 18, n. 3-4, mar./abr., 2012. doi:10.1045/march2012-niu1.

ROCKEMBACH, Moisés. Arquivamento da *web*: Estudos de caso internacionais e o caso Brasileiro. **Revista Digital Biblioteconomia e Ciência da Informação**, [s.l.], v. 16, n. 1, jan./abr. 2018.

STUMPF, Ida R. C. Pesquisa bibliográfica. *In:* DUARTE, Jorge; BARROS, Antonio (org.). **Métodos e técnicas de pesquisa em comunicação**. 2. ed. São Paulo: Atlas, 2008, p. 51-61.

TOYODA, Masashi; KITSUREGAWA, Masaru. **The history of web archiving**. Proceedings of the IEEE, 2012. Disponível em: https://ieeexplore.ieee.org/ stamp /stamp.jsp? arnumber=6182575. Acesso em: 2 jun. 2018.