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

PRESERVAÇÃO DE WEBSITES GOVERNAMENTAIS A PARTIR DO ARQUIVAMENTO DA WEB: ABORDAGENS E METODOLOGIAS

GOVERNMENT WEBSITES PRESERVATION BASED ON WEB ARCHIVING: APPROACHES AND METHODOLOGIES

Jonas Ferrigolo Melo - Programa de Pós-graduação em Comunicação e Informação UFGRS - PPGCOM
Lúcia Andréia Nunes de Oliveira Nunes - Programa de Pós-graduação em Ciência da Informação UFGRS - PPGCIN
Moises Rockembach - PPGCOM e PPGCIN da UFRGS

Modalidade: Resumo Expandido

Resumo: O trabalho versa sobre tecnologia e preservação de websites, com objetivo de identificar e demonstrar a produção científica internacional sobre Arquivamento da Web Governamental, a partir da revisão sistemática da literatura e pesquisa documental como base metodológica. *Web of Science*, *Scopus* e *Google Scholar* foram utilizadas para selecionar artigos produzidos de 2013 a 2018 que versem sobre o tema. Dos 147 artigos encontrados, foram analisadas 12 produções científicas. Conclui-se que a produção sobre arquivamento da web governamental ainda é escassa, mesmo que os dados qualitativos sejam substanciais e abordem todas as fases do Ciclo de Vida do Arquivamento da Web.

Palavras-Chave: Arquivamento da web; Arquivo web; Governo.

Abstract: The work is about technology and websites preservation, aiming to identify and demonstrate the international scientific production on Government Web Archiving, based on the systematic literature review and documentary research as a methodological basis. *Web of Science*, *Scopus* and *Google Scholar* were used to select articles produced from 2013 to 2018 that deal with the theme. Of the 147 articles found, 12 scientific productions were analyzed. It is concluded that government web archiving production is still insufficient, even though the qualitative datas is substantial and attend all phases of the Web Archiving Life Cycle Model.

Keywords: Web archiving; Web archive; Government

1 INTRODUÇÃO

A web é um meio essencial para publicação, gerenciamento e disseminação de informações e sua importância social é confirmada pelo uso massivo e exponencial deste ambiente digital. Em função de seu caráter dinâmico e efêmero, exige que o processo de preservação seja pensado de forma sistêmica desde o princípio, incluindo metodologia de coleta dos dados, estabelecimento de políticas para seleção do conteúdo, técnicas e métodos de armazenamento, preservação digital, acesso e outros procedimentos técnicos necessários quando se trabalha com informação associada à tecnologia. A esse conjunto de atividades relativas à preservação do ambiente web é dado o nome de Arquivamento da Web, o qual disponibiliza a futuros pesquisadores os websites e objetos digitais disponíveis na internet por meio de plataformas digitais de armazenamento e recuperação da informação (ROCKEMBACH, 2018b, p. 241), sendo que "[...] de forma objetiva, podemos definir o arquivamento da web como um processo que compreende coletar, armazenar e disponibilizar a informação retrospectiva da *World Wide Web* para futuros pesquisadores" (ROCKEMBACH, 2018a, p. 09).

Os arquivos de páginas da web são sistemas informacionais que adquirem, armazenam e preservam conteúdos publicados na internet; contribuem para pesquisas e podem se consolidar como espaços fundamentais para a salvaguarda de informações de uma época, afinal "[...] são uma nova forma de instituições de patrimônio cultural obrigadas a preservar artefatos semelhantes" (COSTA; GOMES; SILVA, 2016, p. 2, *tradução nossa*). Dentre os pioneiros da iniciativa do arquivamento da web destacamos o *Internet Archive*[1]; o projeto da Biblioteca Nacional Australiana PANDORA[2] (do inglês, *Preserving and Accessing Networked Documentary Resources of Australia*); e a iniciativa Kulturarw3[3], da Suécia, todos iniciados em 1996.

Tendo em vista a elaboração de pesquisas com foco no arquivamento da web governamental brasileira, o objetivo deste artigo é identificar e demonstrar abordagens e metodologias internacionais sobre Arquivamento da Web Governamental, nos últimos 6 anos (2013 a 2018), nas bases de dados *Web of Science* (WoS), *Scopus* e *Google Scholar*, a partir de uma revisão sistemática da literatura como base metodológica.

[1] https://archive.org/
[2] https://pandora.nla.gov.au/
[3] https://www.kb.se/

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa se apoiou na metodologia conhecida como Revisão Sistemática de Literatura (RSL), um método de síntese de evidências que avalia criticamente e interpreta as pesquisas disponíveis para uma questão particular, área do conhecimento ou fenômeno de interesse (BRASIL, 2012), com o objetivo de identificar e demonstrar abordagens e metodologias internacionais sobre Arquivamento da Web Governamental.

Para a realização desta pesquisa elegeu-se como fonte as bases de dados *Web of Science* (WoS), *Scopus* e *Google Scholar*. A escolha das duas primeiras deu-se por tratarem-se das maiores bases de referências bibliográficas de literatura científica revisada por pares e por apresentarem um grande número de artigos científicos na área das ciências da informação; o *Google Scholar* foi escolhido em função do seu uso estar em constante crescimento como principal fonte para pesquisas científicas (ORDUNA-MALEA; et al, 2015).

A equação da pesquisa consistiu na utilização combinada dos descritores *web archive* e *web archiving*. Testes de pesquisa foram realizados considerando os descritores nos campos “resumo” e “palavras-chave”. Logo nas primeiras páginas pode-se perceber de que não se tratavam de artigos relacionados ao assunto que se busca para este estudo. Ao realizar o mesmo teste utilizando o campo “título” pode-se perceber que a recuperação de informações foi mais direcionada à pesquisa que se pretende desenvolver, justificando o uso deste campo para a realização da pesquisa.

Durante a leitura dos resumos dos artigos, foram utilizados como critérios de elegibilidade: (a) artigos científicos; (b) que estivessem diretamente relacionados à temática de interesse; e (c) publicados entre 2013 e 2018. Foram selecionados 173 artigos, dos quais foram extraídas informações tais como *título*, *autor(es)*, *ano de publicação*, *link de acesso* ao trabalho e o *resumo*. Os procedimentos metodológicos adotados estão sistematizados no *Quadro 1*.

Aos artigos selecionados inicialmente foi aplicado um novo critério de inclusão: conter em seu título ou resumo as palavras “government” e/ou “national”. Dos 47 artigos selecionados foram atribuídos outros critérios de exclusão: (a) artigos não redigidos em inglês; e (b) artigos em que não se teve acesso ao texto na íntegra. Desta análise restaram 12 artigos, dos quais tiveram seus dados extraídos para revisão de literatura que se propõe.

Quadro 1 - Procedimentos metodológicos para a revisão sistemática da literatura

Objetivo	Identificar e demonstrar a produção científica sobre arquivamento da web em âmbito nacional e internacional referente ao arquivamento de websites governamentais.
Âmbito da pesquisa	*Web of Science*, *Scopus* e *Google Scholar*.
Equações da pesquisa	**Na WoS:** ti=(web AND archive) **Na Scopus:** Article title: web archive **No Google Scholar:** allintitle: "web archiving"
Critérios de inclusão	Artigos de revistas científicas publicados de 2013 a 2018.
Critérios de exclusão	Literatura cinzenta, capítulos de livros, monografias, dissertações, teses, artigos que não se enquadraram no objetivo e artigos duplicados entre as bases de pesquisa.
Critérios de validade metodológica	Dupla checagem, verificação dos critérios de inclusão e exclusão.
Resultados	Registro dos procedimentos metodológicos e descrição da pesquisa e seus resultados.
Tratamento de dados	Sistematização dos dados em planilhas para organização e posterior análise.

Fonte: Elaborado pelos autores.

Do *corpus* selecionado para análise foram extraídas informações tais como *título do artigo*, *autor(es)*, *nacionalidade do(s) autor(es)*, *ano de produção*, *objetivo central do artigo,* e *abordagem da pesquisa*. A coluna referente aos aspectos da *abordagem da pesquisa* foi dividida em *Teórica* e *Empírica*. A coluna *Teórica* apresenta outra coluna chamada *Conteúdo segundo o ciclo de vida do arquivamento da web* (BRAGG, 2013), onde foram especificados quais aspectos da teoria cada artigo analisado aborda. O Ciclo de vida do arquivamento da web é um modelo desenvolvido pela equipe do Internet Archive que apresenta as fases, sejam elas tecnológicas ou programáticas do arquivamento da web, em uma estrutura que pode ser utilizada por qualquer organização que deseje arquivar a web (BRAGG; HANNA, 2013, p. 28, tradução nossa). Já a coluna *Empírica* foi subdividida em *Objeto que trata o artigo*, *Domínio*, e *Link do arquivamento*. A partir de então, se procedeu a leitura dos 12 artigos com a intenção de extrair as informações supracitadas. Por fim, a pesquisa documental foi utilizada como recurso para complementar o quadro de análise.

3 ARQUIVAMENTO DA WEB DE WEBSITES GOVERNAMENTAIS

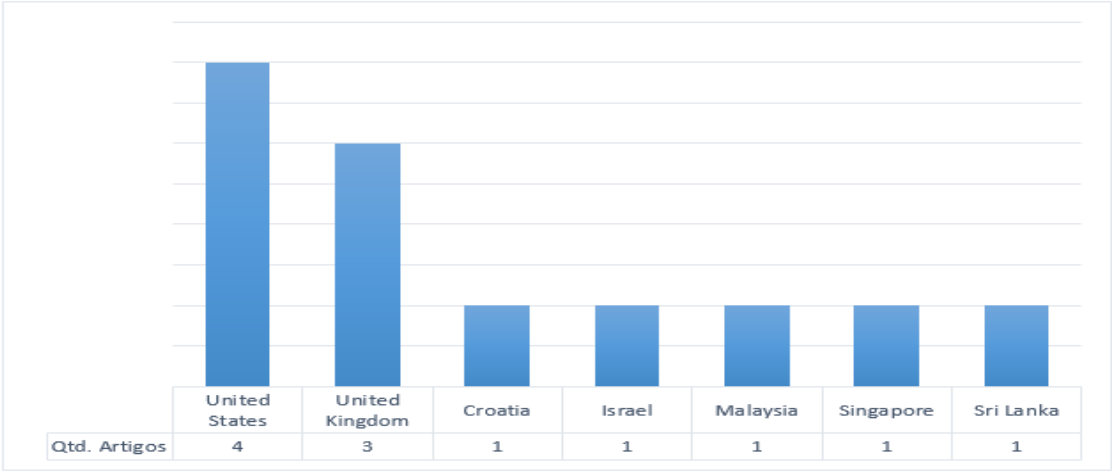
A nuvem de palavra (Imagem 1) produzida a partir dos resumos dos artigos selecionados, com o uso do software online *Wordcloud*, demonstra que a palavra “web” se destaca como a mais utilizada; seguida pelas palavras “national”, “government”, “archive” e “archiving”. A partir desta forma de visualização, notamos que os resultados encontrados, de fato, foram satisfatórios àquilo que se propunha com este estudo.

Imagem 1 - Nuvem de palavras baseada na frequência dos termos dos resumos

Fonte: Elaborado pelos autores.

No *Gráfico 1,* se percebe que sete (7) países estão associados a produção dos 12 artigos selecionados para análise. Estados Unidos com quatro (4) artigos; seguido por Reino Unido com três (3) artigos; Croácia, Israel, Malásia, Singapura e Sri Lanka com um (1) artigos cada.

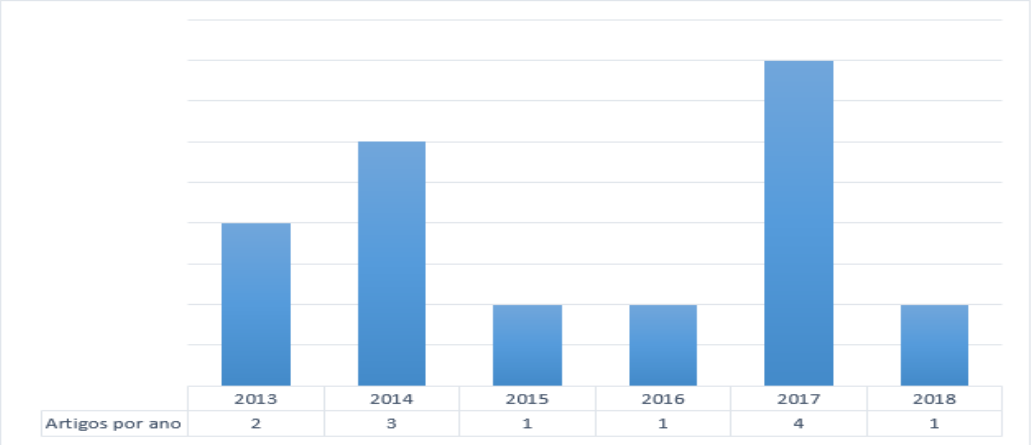
Gráfico 1 - Quantidade de artigos por nacionalidade de autor(es)



Fonte: Elaborado pelos autores.

Na sequência, passamos a analisar o ano de publicação dos artigos demonstrado no *Gráfico 2*. Verificamos que o ano em que mais se publicou artigos sobre a temática analisada foi em 2017, em que veiculam quatro (4) artigos; seguido por três (3) publicações em 2014; duas (2) publicações em 2013; e uma (1) publicação nos anos de 2015, 2016 e 2018.

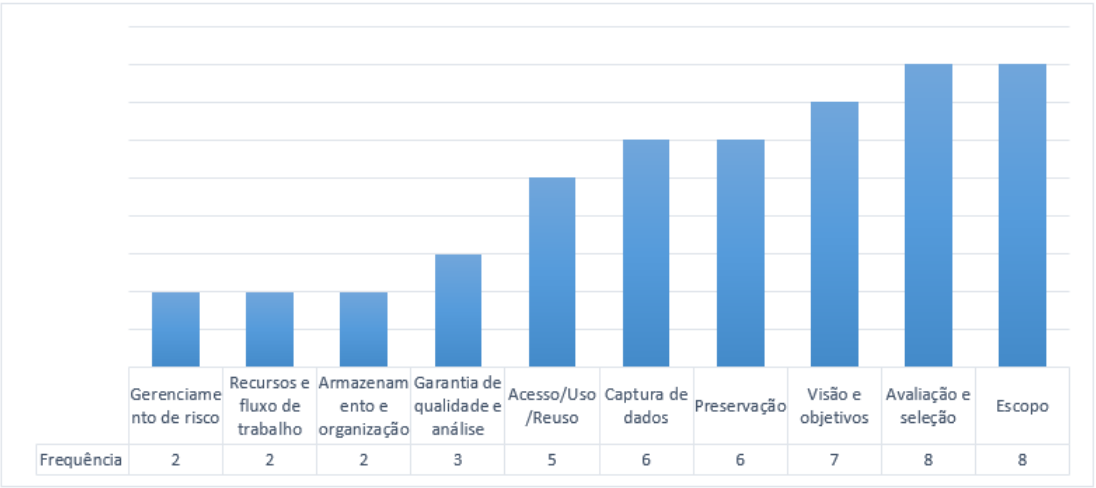
Gráfico 2 - Quantidade de artigos por ano



Fonte: Elaborado pelos autores.

Ainda que os artigos trouxessem análises teóricas sobre o arquivamento da web, todos se utilizaram de estudos empíricos, seja a partir de um domínio ou de website específico. Sendo assim todos os artigos analisados mesclaram a abordagem da pesquisa. Ao analisar os dados em relação ao conteúdo teórico de cada artigo, foi verificada a frequência com que cada fase do Ciclo de Vida do Arquivamento da Web é abordado nos artigos selecionados, demonstrada no *Gráfico 3*. Percebe-se que as fases que mais aparecem nos artigos são o "escopo" e a "avaliação e seleção"; seguido por "visão e objetivos", "preservação", "captura dos dados", "acesso/uso/reuso", "garantia de qualidade e análise", "armazenamento e organização", "recursos e fluxo de trabalho" e "gerenciamento de risco".

Gráfico 3 - Abordagem de pesquisa teórica: conteúdo dos artigos segundo o ciclo de vida do arquivamento da web



Fonte: Elaborado pelos autores.

O *Quadro 2* demonstra o objeto de estudo de cada artigo. Quando a informação não constava nos respectivos artigos, foi utilizada a pesquisa documental, com base na busca dos domínios e dos links dos arquivamentos, possibilitando elaboração do quadro abaixo:

Quadro 2 - Objeto de análise das pesquisas empíricas

OBJETO QUE TRATA O ARTIGO	DOMÍNIO	LINK DO ARQUIVAMENTO
Sites de artistas visuais mulheres	.org	https://wayback.archive-it.org/2973/*/http://womenarts.org | https://wayback.archive-it.org/2973/*/http://wiaf.org
Arquivo web da bibliotecas nacionais da Áustria, Austrália, Grã-Bretanha, Canadá, China, República Tcheca, Dinamarca, Estônia, Finlândia, França, Alemanha, Holanda, Islândia, Japão, Letônia, Lituânia, Nova Zelândia, Noruega , Cingapura, Eslovênia, Espanha (Catalunha), Suécia, Suíça, Tasmânia e EUA.	(domínios de todos os países pesquisados)	(não houve arquivamento neste estudo)
Contas do Twitter e do Youtube do governo central do Reino Unido	twitter.com	https://webarchive.nationalarchives.gov.uk/twitter/
Sites registrados no domínio nacional da Croácia	.hr	http://haw.nsk.hr/
Lei do depósito legal de Singapura	.sg	http://eresources.nlb.gov.sg/webarchives/landing-page
"Dark Domain Archive" preservados pela British Library.	gov.uk	https://www.webarchive.org.uk/wayback/en/archive/20130501145023/http://www.webarchive.org.uk/aadda-discovery/
UK National Archives	gov.uk	(não houve arquivamento neste estudo)
Domínio governamental de Sri Lanka	gov.lk	http://webarchive.loc.gov/all/*/http://www.priu.gov.lk/
Sites com domínio gov.uk	gov.uk	www.nationalarchives.gov.uk/
Coleções EOT2008 e EOT2012 (End of Term)	.gov	http://eotarchive.cdlib.org/
Websites de instituições do governo americano que veiculam informações sobre mudanças climáticas.	.gov	http://eotarchive.cdlib.org/2016.html
Sites dos departamentos e agências do Governo do Estado de Sarawak	.gov.my	(Sarawak State Web Archive não encontrado)

Fonte: Elaborado pelos autores.

A partir das informações extraídas destes artigos, pode-se perceber que países como Estados Unidos e Reino Unido estão à frente quando o assunto é arquivamento da web governamental. Os dois países possuem práticas recorrentes para preservação destas informações, se estabelecendo como líderes no desenvolvimento e uso de tecnologias de informação e comunicação. Cabe destacar, que todas as fases do Ciclo de Vida do Arquivamento da Web são abordadas em pelos menos dois dos artigos analisados. Isso demonstra que as rotinas apresentadas pela teoria do Ciclo de Vida, de fato representam a prática que se espera quando o assunto é arquivamento da web, ao menos pode-se dizer isso a partir da análise que se propõe com essa pesquisa.

4 CONSIDERAÇÕES FINAIS

Foram apresentadas neste artigo possibilidades de abordagens e metodologias internacionais para o arquivamento de websites governamentais a partir da produção científica sobre o arquivamento da web governamental, de modo que pudemos mapear quem, onde, quando e como estão sendo desenvolvidos estudos sobre este tema, de modo

que se conclui que o objetivo deste estudo foi alcançado, especialmente quando se percebe que os procedimentos metodológicos apresentados no capítulo 2 foram satisfatórios e corresponderam às expectativas deste estudo. Os resultados encontrados a respeito dos aspectos teóricas de cada arquivo nos surpreende quando nos deparamos que todos as fases do Ciclo de Vida do Arquivamento da Web são abordadas em pelo menos dois dos artigos selecionais. Mesmo que a produção científica sobre arquivamento da web governamental ainda seja pequena, percebe-se que é um material substancial para o estudo sobre este tema, ainda que seja necessário a realização de outras pesquisas que se preocupem com o desenvolvimento de tecnologias de informação e comunicação desde sua geração, uso, tramitação, preservação, recuperação, difusão, acesso, gestão e as demais rotinas que permeiam os estudos a respeito de informação e tecnologia.

A importância histórica, cultural e intelectual do arquivamento da web tem sido amplamente reconhecida, considerando que diversos países estão estabelecendo iniciativas de arquivamento de conteúdo da web, ficando evidente, desta forma, que a necessidade de o Brasil desenvolver a preservação dos websites governamentais.

REFERÊNCIAS

BRAGG, Molly; HANNA, Kristine; DONOVAN, Lori; HUKILL, Graham; PETERSON, Anna. **The Web Archiving Life Cycle Model.** WhitePaper. 2013.

BRASIL. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Ciência e Tecnologia. **Diretrizes metodológicas:** elaboração de revisão sistemática e metanálise de ensaios clínicos randomizados. Brasília: Editora do Ministério da Saúde. 92p. 2012.

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

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

ORDUNA-MALEA, Enrique; AYLLÓN, J. M.; MARTÍN-MARTÍN, A.; et. al. Methods for estimating the size of Google Scholar. **Scientometrics**, v. 104, n. 3, p. 931-949. 2015.

ROCKEMBACH, Moisés. Arquivamento da Web: estudos de caso internacionais e o caso brasileiro. **RDBCI**: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, SP, v. 16, n. 1, p. 7-24, 2018a.

ROCKEMBACH, Moisés. A web retrospectiva como campo de pesquisa: arquivamento da web e preservação digital. In: BENETTI, Marcia, BALDISSERA, Rudimar (org). **Pesquisa e Perspectivas de Comunicação e Informação.** Porto Alegre: Sulina, 2018b. p. 240-256.