UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMUNICAÇÃO E INFORMAÇÃO
MESTRADO EM COMUNICAÇÃO E INFORMAÇÃO

JONAS FERRIGOLO MELO

ARQUIVAMENTO DOS *WEBSITES* DO GOVERNO FEDERAL BRASILEIRO:

preservação do domínio GOV.BR

Porto Alegre
2020

JONAS FERRIGOLO MELO

ARQUIVAMENTO DOS *WEBSITES* DO GOVERNO FEDERAL BRASILEIRO:

preservação do domínio GOV.BR

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Comunicação e Informação da Universidade Federal do Rio Grande do Sul como requisito para obtenção do título de Mestre em Comunicação e Informação.

Orientador: Prof. Dr. Moisés Rockembach

Porto Alegre
2020

CIP - Catalogação na Publicação

Melo, Jonas Ferrigolo
Arquivamento dos websites do Governo Federal Brasileiro: preservação do domínio GOV.BR / Jonas Ferrigolo Melo. -- 2020.
133 f.
Orientador: Moisés Rockembach.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Programa de Pós-Graduação em Comunicação e Informação, Porto Alegre, BR-RS, 2020.

1. Arquivamento da web. 2. Arquivo web. 3. Governo Federal Brasileiro. 4. Web governamental. 5. gov.br. I. Rockembach, Moisés, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Programa de Pós-Graduação em Comunicação - PPGCOM
Rua Ramiro Barcelos, 2705
Porto Alegre, RS – CEP 90035-007
Telefone: (51) 3308.5116
E-mail: ppgcom@ufrgs.br

JONAS FERRIGOLO MELO

ARQUIVAMENTO DOS *WEBSITES* DO GOVERNO FEDERAL BRASILEIRO:
preservação do domínio GOV.BR

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Comunicação e Informação da Universidade Federal do Rio Grande do Sul como requisito para obtenção do título de Mestre em Comunicação e Informação.

Orientador: Prof. Dr. Moisés Rockembach

COMISSÃO EXAMINADORA

_____
Professor Dr. Moisés Rockembach (Orientador)
Universidade Federal do Rio Grande do Sul

_____
Professora Dra. Ana Maria Mielniczuk de Moura
Universidade Federal do Rio Grande do Sul

_____
Professor Dr. Fabiano Couto Corrêa da Silva
Universidade Federal do Rio Grande do Sul

_____
Professora Dra. Caterina Marta Groposo Pavão
Universidade Federal do Rio Grande do Sul

_____
Professor Dr. Armando Malheiro da Silva (Suplente)
Universidade do Porto

Porto Alegre, 20 de fevereiro de 2020.

AGRADECIMENTOS

Agradeço a Universidade Federal do Rio Grande do Sul, a Faculdade de Biblioteconomia e Comunicação, e ao Programa de Pós-Graduação em Comunicação e Informação por me receberem de braços abertos e por serem resistência em um momento de ataques à educação. Ao meu orientador, professor e amigo Moisés Rockembach por estes anos de acompanhamento e parceria acadêmica, de viagens e por ter me apresentando Porto/PT. As professoras e professores membros da banca de defesa que me honraram com uma revisão tão qualifica que só trouxe qualidade a essa pesquisa. Aos amigos e amigas que me auxiliarem neste período, pelos conselhos, pelas ajudas científicas, pelas cervejas, festas e festivais que foram escapes fundamentais neste processo de imersão científica. Aos meus colegas do Arquivo Público do Estado do Rio Grande do Sul por entenderem e me apoiarem durante os períodos em que estive afastado. Aos meus estagiários por entenderam e terem sido autônomos durante minhas ausências. Ao meu amor, Wagner, que foi meu porto seguro, ao me incentivar e me fazendo companhia em momentos de sufoco. A minha família por acompanhar, mesmo que a distância, me dar força e entender os momentos em que tive que abdicar de junções familiares para estudar. Um agradecimento especial à minha mãe, Miriam, a quem dedico esta dissertação, que sempre me apoiou, acreditou em mim, nos meus sonhos e teve coragem de confiar em sua criação. Obrigado!

– Quarenta e dois! – berrou Loonquawl. – É tudo que você tem a nos dizer depois de sete milhões e quinhentos mil anos de trabalho?
– Eu verifiquei cuidadosamente – disse o computador –, e não há dúvida de que a resposta é essa. Para ser franco, acho que o problema é que vocês jamais souberam qual é a pergunta.
– Mas era a Grande Pergunta! A Questão Fundamental da Vida, o Universo e Tudo Mais – gritou Loonquawl.
– É – disse Pensador Profundo, com um tom de voz de quem tem enorme paciência para aturar pessoas estúpidas –, mas qual é exatamente a pergunta?
– Bem, você sabe, é simplesmente tudo... tudo... – começou Phouchg, vacilante.
– Pois é! – disse Pensador Profundo. – Assim quando vocês souberem qual é exatamente a pergunta, vocês saberão o que significa a resposta.

(O Guia do Mochileiro das Galáxias, de Douglas Adams,1979)
Epígrafe à minha amiga, Camila Duarte Ritter.

RESUMO

A presente pesquisa investiga quais são as possibilidades de arquivamento de *websites* do Governo Federal Brasileiro. O objetivo geral é demonstrar as possibilidades de arquivamento de *websites* do governo federal brasileiro a partir de um estudo de caso do domínio gov.br. Para o estudo aplicado, foram selecionados 23 *websites* governamentais, sendo 22 *websites* de ministérios e um *website* do governo central, o Portal Único www.gov.br. A pesquisa consistiu em verificar os recursos oferecidos por estes *websites*; arquivar os *websites* selecionados, com o uso de rastreador de páginas *web* automatizado *Heritrix*; reconstruir os *websites* arquivados com o uso de *software* automatizado WABAC; e comparar os recursos disponibilizados nas versões ao vivo e arquivadas dos *websites* selecionados. A pesquisa foi amparada pelas teorias do arquivamento da *web*, com ênfase nas abordagens de Bragg; Hanna (2013) em seu *The Web Archiving Life Cycle Model*, em Khan; Rahman (2019), considerando *A Systematic Approach Towards Web Preservation*; e no levantamento de informações acerca do arquivamento da *web* governamental. Como procedimentos metodológicos a pesquisa é classificada como de natureza aplicada, exploratória-descritiva, com pesquisa documental e aplicação do estudo de caso, sendo que a abordagem do problema se classifica como mista, considerando que os dados foram analisados quali e quantitativamente. Como considerações finais a pesquisa apresenta os resultados encontrados, permitindo visualizar as possibilidades de arquivamento dos *websites* do Governo Federal Brasileiro. Alguns recursos presentes nos *websites* não foram recuperados, especialmente, quando o formato do arquivo era diferente do textual e da imagem estática. Áudios, vídeos e recursos hospedados em servidores externos tenderam a não recuperação. Entende-se que se faz necessário o uso de ferramenta auxiliar para recuperação destes documentos não textuais, além de outros estudos empíricos para compreender as necessidades e melhores ferramentas para a recuperação destes documentos. A recuperação foi considerada satisfatória quando os resultados mostram que a maioria dos *websites* arquivados apresentam seus conteúdos de forma integral, ainda que alguns não estejam formatados visualmente, tal como o *website* ao vivo. Medidas para garantia de qualidade foram atribuídas e a permanência dos recursos dos *websites* após seu arquivamento são balizadores para definir a qualidade de uma coleta. É apresentado um mapa mental com informações que poderão ser úteis para quem deseja realizar o arquivamento da web no Brasil. O arquivamento da *web* é uma forma de preservar e manter as evidências dos serviços e fazeres do Governo Federal Brasileiro, para tornar acessível para futuros fins de pesquisa e, também, como registros da evolução das ações governamentais. Conclui-se que os *websites* do Governo Federal Brasileiro são arquiváveis sem perda de informações relevantes e que o país carece de uma política pública para sistematizar o arquivamento dos *websites* governamentais.

Palavras-chave: Arquivamento da *web*. *Web* governamental. Governo Federal Brasileiro. gov.br

ASTRACT

This research investigates what are the possibilities for archiving websites of the Brazilian Federal Government. The general objective is to demonstrate the possibilities of archiving Brazilian Federal Government websites from a case study of the domain gov.br. For the applied study, 23 government websites were selected, being that 22 ministry websites and one central government website, Portal Único www.gov.br. The research consisted of checking the resources offered by these websites; archive the selected websites, using the Heritrix web page crawler; rebuild archived websites using the software WABAC; and compare the resources available in the live and archived versions of selected websites. The research was supported by theories of web archiving, with an emphasis on approaches of Bragg; Hanna (2013) The Web Archiving Life Cycle Model, and Khan; Rahman (2019), considering A Systematic Approach Towards Web Preservation; and gathering information about archiving the government web. As methodological procedures the research is classified as applied, exploratory-descriptive, with documentary research and application of the case study, and the problem approach is classified as mixed, considering that the data were analyzed qualitatively and quantitatively. As final considerations the research presents the results found, allowing to visualize the possibilities of archiving the websites of the Brazilian Federal Government. Some resources present on the websites were not recovered, especially when the file format was different from the textual and static image. Audios, videos and resources hosted on external servers tended not to recover. It is understood that it is necessary to use an auxiliary tool to recover these non-text documents, in addition to other empirical studies to understand the needs and better tools for the recovery of these documents. The recovery was considered satisfactory when the results show that most of the archived websites present their contents in full, although some are not visually formatted, such as the live version. Quality assurance measures have been assigned and the permanence of the resources of the websites after their archiving are guidelines to define the quality of a collection. A mind map is presented with information that may be useful for those who wish to archive the web in Brazil. Archiving the web is a way to preserve and maintain evidence of the services and actions of the Brazilian Federal Government, to make it accessible for future research purposes and also as records of the evolution of government actions. It is concluded that the websites of the Brazilian Federal Government are archivable without loss of relevant information and that the country lacks a public policy to systematize the archiving of government websites.

Keywords: *Web* archiving. Governmental *web*. Digital preservation. gov.br

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ANSP	Academic Network at São Paulo
BnF	Bibliotheque Nationale de France
Cetic.br	Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação
CGI.br	Comitê Gestor da Internet no Brasil
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CCTCI	Comissão de Ciência e Tecnologia, Comunicação e Informática
CCULT	Comissão de Cultura
DARPA	Agência de Projetos de Pesquisa Avançada dos Estados Unidos
DCMI	Dublin Core Metadata Iniciative
DSDM	Dynamic Systems Development Method
FAPERGS	Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
FAPERJ	Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
Fermilab	Fermi National Accelerator Laboratory
FINEP	Financiadora de Estudos e Projetos
HTML	Hypertext Markup Language
Ibase	Instituto Brasileiro de Análises Sociais e Econômicas
IIPC	International Internet Preservation Consortium
IPT	Instituto de Pesquisas Tecnológicas do Estado de São Paulo
ISO	International Organization for Standardization
KB	Kungl.biblioteket - Biblioteca Real da Suíça
LARC	Laboratório Nacional de Redes de Computadores
LNCC	Laboratório Nacional de Computação Científica
METS	Metadata Encoding and Transmission Standard
MIT	Instituto Tecnológico de Massachusetts
MODS	Metadata Object Description Schema
NARA	National Archives and Records Administration
Nic.br	Núcleo de Informação e Coordenação do Ponto BR
NISO	National Information Standards Organization
NUAWEB	Núcleo de Pesquisa em Arquivamento da *Web* e Preservação Digital
OAIS	Open Archival Information System

OCLC	*Online* Computer Library Center
ONGs	Organizações não Governamentais
PANDORA	Preserving and Accessing Networked Documentary Resources of Australia
PL	Projeto de Lei
PREMIS	Preservation Metadata Implementation Strategies
RLG	Research Libraries Group
RNP	Rede Nacional de Pesquisa
SIORG	Sistema de Informações Organizacionais
TCP/IP	Transmission Control Protocol/ Internet Protocol
UCLA	University of California Los Angeles
UFMG	Universidade Federal de Minas Gerais
UFRGS	Universidade Federal do Rio Grande do Sul
UFRJ	Universidade Federal do Rio de Janeiro
UKGWA	UK Government *Web* Archive
UKWAC	UK *Web* Archiving Consortium
Unesp	Universidade Estadual Paulista
Unicamp	Universidade de Campinas
URL	Uniform Resource Locator
USP	Universidade de São Paulo
VRA Core	Visual Resources Association Core Strategies
W3C	World Wide *Web* Consortium
WABAC	*Web* Archive Browsing Advanced Client
WAIS	Wide-Area Information Server
WARC	WebARChive
WWW	World Wide *Web*

SUMÁRIO

1 INTRODUÇÃO

O conteúdo publicado na Internet representa uma parte significativa da herança cultural e científica contemporânea de uma sociedade (HOLUB; RUDOMINO, 2014), e muitos recursos que tradicionalmente as instituições de memória coletam, tais como obras de arte, documentos de governo, livros, revistas e notícias, estão agora disponíveis apenas na *web*. Este é um fenômeno já previsto por Manuel Castells, nos anos 90, quando da publicação da trilogia *"A Era da informação: economia, sociedade e cultura"*, em que o autor afirma que a Sociedade em Rede que vivenciamos é o resultado da apropriação social de um conjunto de tecnologias de informação e comunicação surgidas nos últimos 50 anos, como resultado das profundas mudanças nos setores da microeletrônica, computação e telecomunicações e que alteraram, e ainda alteram, a forma como nos comunicamos em sociedade e como nos relacionamos uns com os outros (CASTELLS, 2000, p. 60).

As alterações no paradigma de informação e comunicação na sociedade em rede e as correspondentes mudanças no modelo de funcionamento das mídias de massa provocaram uma erosão nos modelos de negócio tradicionais (MORENO, 2015), uma vez que fomos testemunhas do aumento exponencial da convergência tecnológica entre a Internet e a comunicação sem fio nos anos 2000. Com o advento destas tecnologias de comunicação e a forma predominante em que se estabeleceram no mercado comunicacional fez com que a forma como a sociedade passou a interagir mudasse substancialmente (CASTELLS, 2000, p. XXVI), uma vez em que as relações interpessoais passaram a se estabelecer digitalmente, promovendo uma mudança paradigmática na comunicação e na produção e fluxo da informação.

A arquitetura em rede que sustenta a Internet e a distribuição social de informação não seria possível sem a mediação dos computadores, fazendo com que tenhamos que olhar a migração do analógico para o digital como um elemento central nas transformações tecnológicas, uma vez que o computador torna tudo digital. E, sendo assim, a revolução informacional em curso tenderá a ser mais

impactante do que àquela que resultou da impressão de tipos móveis de Gutenberg, considerando que a primeira foi limitada a um conjunto restrito do corpo social enquanto a comunicação e informação digitais estão presentes em todas as estruturas da sociedade (CASTELLS, 2000, p. 30). Ou seja, para Castells, a Internet, entendida como a rede de ligação entre computadores, é “[...] talvez o meio tecnológico mais revolucionário da era da informação” (CASTELLS, 2000, p. 45).

Passado mais de duas décadas da expansão da Internet, já é perceptível que de fato trata-se de uma revolução tecnológica na forma de agir da sociedade, uma vez que a crescente utilização de meios de comunicação com alto grau de mobilidade e o uso cada vez maior da Internet definem outros espaços e demarcam novas fronteiras para a sociedade contemporânea (RIBEIRO, 2014, p. 97). Por permitir disseminar informações e conhecimento para uma grande quantidade de pessoas que se encontram distantes geograficamente, a Internet é, potencialmente, um meio de comunicação em massa, inserindo na sociedade a possibilidade de pluralizar a emissão de informações, fazendo com que seja produzido conteúdo diverso e de livre expressão, permitindo que a Internet exerça um papel fundamental à existência, manutenção e exercício do direito fundamental de livre manifestação.

A rede mundial de computadores é a maior ferramenta de conexão entre as pessoas, ao mesmo tempo em que possibilita o compartilhamento de informações à sociedade, permitindo oportunidades de desenvolvimento econômico, propagação de ideias e obtenção de reputação. Não obstante, a partir da rede mundial de computadores, pode-se ver pessoas e empresas tendo sua imagem exposta, seja positivamente ou não; presidentes sendo eleitos e democracias sendo postas em risco ou em advento. Mais que a popularização de uma ferramenta de comunicação, assistimos ao incremento de um importante instrumento democrático.

Ao mesmo tempo, este poder de intervir socialmente concede uma capacidade ao cidadão em se posicionar como um parceiro exigente às responsabilidades governamentais, alinhando o cidadão como fonte de informação para o Estado e como coprodutor e consumidor de informação governamental. Tal concepção nos remete à mediação por meio das tecnologias de informação e comunicação, no sentido de que o usuário não é apenas um receptor, mas pode ser um ator ativo na construção da interatividade, questionando e modificando o

conteúdo de que faz uso. Para a junção destas funções de produtor e consumidor da informação, foi cunhado o termo *Prosumer*, em inglês, que significa que o consumidor é ao mesmo tempo produtor de conteúdo. O termo *prosumer* foi utilizado pela primeira vez pelo futurólogo Alvin Tofler, em seu livro *A Terceira Onda*, de 1980 (TROYE; XIE, 2007; FONSECA et al., 2008; ISLAS-CARMONA, 2008).

Com o usuário que consome e, ao mesmo tempo, produz conteúdo, é perceptível que a velocidade com que se produz informação é tão rápida quanto a velocidade com que se perdem e se apagam informações na rede, sendo este um dos fatores que preocupa os pesquisadores da área. Páginas da *web* são documentos dinâmicos, considerando que "[...] ao mesmo tempo em que milhares de informações são criadas, outras são sobrepostas, tornando difícil a recuperação destes dados" (ROCKEMBACH; PAVÃO, 2018). Lawrence et al. (2001, p. 30) citaram uma estimativa da *Alexa Internet* (https://www.alexa.com/) em que as páginas da *web* desaparecem após 75 dias, em média. Brewster Kahle, do *Internet Archive*, ampliou esse número para 100 dias (Kornblum, 2001[1] *apud* DAY, 2003b, p. 7). Mais recentemente, Costa, Gomes e Silva (2016) revelam estudos que apontam que 80% das páginas *web* não estão disponíveis em sua forma original após um ano, 13% das referências da *web* em artigos acadêmicos desaparecem após 27 meses e 11% das publicações em sites de rede social são perdidas após um ano.

Além de perder informações científicas e históricas, a transitoriedade das informações publicadas na *web* faz com que pessoas, de modo geral, percam suas memórias como indivíduos ao desaparecerem, por exemplo, fotos e descrições de eventos exclusivamente publicados na Internet. Além disso, os *websites* mudam de tempos em tempos e suas versões anteriores ficam indisponíveis. Essa é uma complicação para usuários que desejam procurar informações que podem não estar mais disponíveis nos sites. Já estamos enfrentando falta de informações necessárias devido à ausência de páginas ou formatos e extensões antigas de

[1] KORNBLUM, J. (2001). Web-page database goes Wayback when. **USA Today**, 30 out. 2001. Disponível em: <http://www.usatoday.com/life/cyber/tech/2001/10/30/ebrief.htm>. Acesso em: 15 abr. 2020.

documentos. Vint Cerf, um dos pioneiros da Internet, em 2015 já alertou sobre o perigo das futuras gerações que terão pouco ou nenhum registro do século XXI[2].

Por meio dos serviços de informação, acredita-se, então, que o arquivista tem potencial para atuar na mediação da informação[3]. E, neste sentido, surge uma nova preocupação na área da informação e tecnologia: a preservação do conteúdo publicado na *web*. Na verdade, não é algo propriamente novo. Organizações internacionais já manifestaram preocupação com a efemeridade da *web*.

As dezenas de iniciativas de arquivamento da *web* ao redor do mundo demonstram uma variedade de métodos e abordagens para selecionar, adquirir, organizar, armazenar, descrever e fornecer acesso ao conteúdo *web* (NIU, 2012). Estas variedades de métodos e abordagens são consequências de fatores externos, como o ambiente jurídico e os relacionamentos entre os produtores de conteúdo e o próprio arquivo da *web*; além de fatores internos, como a natureza do conteúdo arquivado, a natureza da organização de arquivamento, a escala do arquivo da *web* e a capacidade técnica e financeira da organização de arquivamento (MASANÈS, 2006). Muitas das iniciativas de arquivamento da *web* têm em seus escopos de coleções capturas regulares de *websites* governamentais, que para além de preservar informações valiosas aos cidadãos, pode ajudar o governo na prestação de contas ou na transparência da própria administração pública. A partir deste cenário, elegemos o tema desta pesquisa: arquivamento dos *websites* com domínio gov.br da esfera federal brasileira.

Cumprir essa clara necessidade em abordar o arquivamento da *web* como uma forma de manter o acesso ao registro digital, bem como entender este mecanismo como um instrumento de preservação do conteúdo publicado na *web*, foram os fatores que influenciaram no processo de decisão do problema da pesquisa: quais são as possibilidades de arquivamento da *web* na esfera federal, a partir da análise do domínio gov.br? Sendo que o objetivo geral desta investigação

[2] https://www.bbc.com/news/science-environment-31450389

[3] Cabe ressaltar que autores e estudiosos da literatura da área distinguem os tipos de mediação da informação como, por exemplo, mediação custodial (profissional como custodiador da informação), mediação passiva (usuário apenas como consumidor) e mediação pós-custodial (usuário participante e ativo na construção de sentido) (SILVA, 2010).

é demonstrar as possibilidades de arquivamento de *websites* do Governo Federal Brasileiro a partir de um estudo de caso do domínio gov.br.

Considerando que os propósitos do estudo de caso são os de proporcionar o conhecimento a respeito de determinadas características de um cenário ou população, e o de proporcionar uma visão global do problema (GIL, 2002, p. 55) e, considerando que a análise de um único ou de poucos casos de fato fornece uma base muito frágil para que seja possível generalizar, escolhemos um significativo corpus para amostragem: foram utilizados como objetos de análise 23 *websites* governamentais, sendo 22 *websites* de ministérios, órgãos e secretarias com status de ministérios e um *websites* do governo central, o Portal Único www.gov.br, que incorporou os *websites* www.planalto.gov.br e www.brasil.gov.br.

Como referencial teórico sobre a história da Internet, origem e evolução da *web*, o estudo usou como referência os seguintes autores: Leiner (2009); Abbate (1999); Afonso (2000); Guizzo (1999); Berners-Lee et al. (1994); Vieira (2003); Carvalho (2006); e Castells (2000). Sobre arquivamento da *web*, o estudo tem como base os autores Masanès (2006); Gomes; Miranda e Costa (2011); Costa; Gomes e Silva (2017); Day (2003, 2006); Holub; Rudomiro (2014); Brügger (2005, 2016, 2017); Kelly et al. (2013); Khan; Rahman (2019); Lyman (2002); Bragg; Hanna (2013); Rockembach (2018); Maemura et al. (2018); Niu (2012); Xie (2013); entre outros.

Esta dissertação está estruturada em Introdução; Objetivos; Justificativa; Referencial Teórico com levantamento de informações sobre a história da Internet no contexto do Brasil; arquivamento da *web*, com ênfase nas abordagens de Bragg; Hanna (2013) em seu *The Web Archiving Life Cycle Model*, em Khan; Rahman (2019), considerando *A Systematic Approach Towards Web Preservation*; e no levantamento de informações acerca do arquivamento da *web* governamental; Procedimentos Metodológicos; Análise dos Resultados; Considerações Finais; e Referências Bibliográficas.

1.1 OBJETIVOS

Esta pesquisa foi desenvolvida a partir da prospecção dos objetivos geral e específicos apresentados abaixo:

1.1.1 Objetivo Geral

Demonstrar as possibilidades de arquivamento de *websites* do Governo Federal brasileiro a partir de um estudo de caso do domínio gov.br.

1.1.2 Objetivos Específicos

a) Mapear os *websites* que utilizam o domínio gov.br, para seleção da amostra;
b) analisar os recursos disponibilizados nas versões ao vivo dos *websites* selecionados;
c) arquivar os *websites* selecionados, com o uso de rastreador de páginas *web* automatizado;
d) reconstruir os *websites* arquivados com o uso de *software* automatizado;
e) comparar os recursos disponibilizados na versão arquivada e ao vivo dos *websites* selecionados;
f) apresentar as possibilidades de arquivamento da *web* para a esfera federal brasileira.

1.2 JUSTIFICATIVA

A sociedade contemporânea está imersa na tecnologia desde as ações mais complexas com uso de inteligência artificial, até procedimentos cotidianos e corriqueiros que passaram a ser informatizados. A tecnologia está inserida na sociedade e, como agentes sociais, também acabamos nos inserindo neste cenário. Não obstante, nossa memória social também está sendo produzida em função destes aparatos digitais. Vera Dodebei (2011, p. 41) diz que há de se reconhecer que as instituições mudaram ao longo da vida social e que “As redes sociais são tão reais quanto as relações não intermediadas pela máquina” (DODEBEI, 2011, p. 41). Neste sentido, ao passo que parte de nossa memória social migrou para o ambiente digital e esteja impressa em banco de dados, temos que garantir mecanismos para sua preservação.

Na Internet, temos a construção de uma narrativa visual de acesso aos bancos de dados, contrapondo com o estabelecido ao mundo oral em que “[...] não há possibilidade de formação de memórias auxiliares fixadas em outros suportes porque não existem tais registros fora do pensamento humano [...]” (DODEBEI, 2011, p. 39). Podemos dizer que o suporte da informação, na sociedade digital, oferece maior possibilidade de arquivamento se considerarmos as limitações de espaço físico para o armazenamento de documentos em papel e maior possibilidade de acesso, se considerarmos sua disseminação ainda por meio do ambiente digital. Portanto, precisamos considerar a dinâmica da produção desta memória coletiva constituída por meio das informações digitais, sob a perspectiva do equilíbrio entre a proteção e o acesso às informações para preservar a memória coletiva.

> Reorganizamos deste modo nossa inserção na memória coletiva e, como pode parecer a princípio que anulamos nossa individualidade ao nos mesclarmos no coletivo, pode-se pensar também que não necessariamente haja dissolução de identidades na sociedade digital, mas que, certamente, ocorra uma mudança no processo de subjetivação; surgem novas configurações entre o público e o privado, entre velocidades de interações e produção de informações, entre apropriação e uso de narrativas e, neste caso, podemos supor que o espaço virtual pode ser um ‘lugar’ de memória. (DODEBEI, 2011, p. 40).

Um dos desafios da Sociedade em Rede ligada por tecnologias digitais, é a digitalização de conteúdo das nossas comunicações em sociedade que são feitos cada vez mais por meio dos interesses mercantis e não necessariamente coincidentes com o interesse público (CASTELLS, 2001). Neste sentido, como forma de garantir a memória social, os governos precisam estabelecer rotinas para preservação deste patrimônio que está sendo produzido exclusivamente através da Internet. A revolução informacional sustentada pelo conceito da Sociedade em Rede, reposiciona o Estado como a força que poderá dinamizar a potência que as comunicações e relações digitais estão promovendo com os avanços tecnológicos, tal como Castells (2000) preconiza:

> [...] o que deve ser salientado para entender a relação entre a tecnologia e a sociedade é que o papel do Estado, interrompendo, promovendo ou liderando a inovação tecnológica, é um fator decisivo no processo geral, na medida em que expressa e organiza as forças sociais dominantes num determinado espaço e época. Em grande parte, a tecnologia expressa a capacidade de uma dada sociedade em impulsionar o domínio tecnológico por intermédio das suas instituições sociais, inclusive o Estado. (CASTELLS, 2000, p. 15)

Ainda, de acordo com o *The National Archive*[4], do Reino Unido, uma parte substancial dos registros atuais do governo são produzidos apenas em formato digital e, a falta de uma estratégia de arquivamento e preservação desse conteúdo, inevitavelmente levará ao desaparecimento de informações importantes para o futuro (JAMAIN et al., 2018). Lala e Joe (2006) mencionaram que no século XXI as informações estão sendo produzidas em grandes quantidades, não apenas por meio de formatos tradicionais, mas também cada vez mais em formatos eletrônicos ou digitais. *The National Archives, UK* (2011, *tradução nossa*) declarou que “[...] o arquivamento da *web* é um processo vital para garantir que pessoas e organizações possam acessar e reutilizar conhecimento a longo prazo e atender às necessidades de recuperação de suas informações”.

Isso posto, não restam dúvidas que a produção e uso da *web* transformou a maneira como nos comunicamos nos dias de hoje. Porém, cabe acrescentar que

[4] https://www.nationalarchives.gov.uk/

como forma de garantia de direitos fundamentais dos cidadãos é dever do Estado fornecer o acesso a conteúdo publicado na *web*, especialmente no âmbito de sites governamentais que dispõem de informações críticas e estão sob a ótica e jurisprudência da Lei de Acesso à Informação que regula, dentre outros, o inciso XXXIII do artigo 5º da Constituição Federal do Brasil: “Todos têm direito a receber dos órgãos públicos informações de seu interesse particular, ou de interesse coletivo ou geral [...]” (BRASIL, 1988). Entendemos que além da disponibilização de conteúdo nos portais governamentais, também é dever do Estado garantir o acesso às informações a longo prazo, combatendo possíveis problemas que ocorrem quando não há preservação digital de páginas *web* (conteúdo modificado, deletado, *link* quebrado, etc.).

Para garantir esta ação a nível governamental, o Estado deveria definir sua política de preservação de documentos digitais, incluindo os produzidos em ambiente *web*, nos moldes de projetos semelhantes realizados ao redor do mundo. Selecionar as técnicas de preservação e arquivamento, as tecnologias apropriadas, os conteúdos que serão prioritariamente preservados são alguns dos caminhos que o Brasil poderia seguir para efetivar a implantação de uma iniciativa de arquivamento da *web*. Neste sentido, e como forma de ir em direção à concretude deste cenário, esta pesquisa busca contribuir apresentando possibilidades de arquivamento da *web* na esfera federal brasileira, a partir da análise do domínio gov.br.

O arquivamento da *web* é uma forma de preservar e manter as evidências dos serviços e fazeres do Governo Federal Brasileiro, para torna-los acessíveis para futuros fins de pesquisa e, também, como registros da evolução das ações governamentais. Holub e Rudomiro (2014, p. 1, *tradução nossa*) afirmam que “Devido à natureza dinâmica da *web*, seu crescimento explosivo, vida útil curta, instabilidade e características semelhantes, a importância de seu arquivamento tornou-se inestimável para as gerações futuras”. Rockembach e Pavão (2018) dizem que “caso não haja uma preservação digital dos conteúdos produzidos na *web*, muito do que foi desenvolvido neste meio se perderá para sempre”. Isso exige uma solução de arquivamento da *web* que se apoie em políticas e procedimentos técnicos estruturados. Sendo assim, chega-se ao objetivo desta pesquisa:

demonstrar as possibilidades de arquivamento de *websites* do Governo Federal Brasileiro a partir de um estudo de caso do domínio gov.br.

Xie et al. (2013) destacam a importância histórica, cultural e intelectual do arquivamento da *web* como amplamente reconhecida, em que países com alta taxa de inserção da 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 (2018a) o Brasil ainda não realiza o arquivamento da *web* de forma sistemática, isto é, de maneira contínua e periódica, cobrindo domínios nacionais e, em toda América Latina, somente o Chile[5] constituiu uma iniciativa própria de arquivo da *web*, apresentando quatro (04) coleções.

A sociedade a qual estamos inseridos, frente às tecnologias da informação, a variedade de temáticas de pesquisa, o incremento do uso de dados, representam um conjunto de questões que “[...] colaboram para o crescimento contínuo das publicações científicas na *web* fazendo a comunicação da ciência atingir um novo patamar em sua abrangência e disseminação” (FERREIRA, MARTINS, ROCKEMBACH, 2018, p. 95). Na perspectiva dos usos dos arquivos da *web* na comunicação científica, 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" (FERREIRA, MARTINS, ROCKEMBACH, 2018, p. 95).

A produção de pesquisas e sua disseminação, tanto no âmbito teórico, técnico ou prático, sobre Arquivamento da *Web*, abre um leque de possibilidades à academia e centros de investigação científica, e a necessidade de estudos sobre a temática, especialmente no Brasil, apresenta um amplo potencial à ciência. Este é um dos aspectos pessoais que motivaram a escolha por essa temática.

Estima-se que com essa pesquisa possamos apresentar possibilidades para a realização do arquivamento dos *websites* governamentais no Brasil e criar, quem sabe, uma cultura de preservação sistemática destas fontes de informação e

[5] http://archivoweb.bibliotecanacionaldigital.cl/

memória social. Espera-se também que o arquivamento da *web* ajude a alcançar um senso de comunidade, identidade nacional e enraizamento entre os cidadãos brasileiros, no sentido em que se estará preservando informações que, de certa forma, moldam a identidade nacional, através do seu desenvolvimento nos meios políticos. A *web* é cada vez mais usada como uma ferramenta para comunicação social e interações entre o poder público e a sociedade civil e, com o tempo, o arquivo da *web* poderá formar um registro de eventos que capturam o ambiente da nação e que acompanhe o desenvolvimento da identidade nacional brasileira. Arquivar esse registro fornece uma fonte inestimável de herança documentada às atuais e futuras gerações, criando um senso de comunidade e pertencimento.

2 DA ORIGEM DA WEB ÀS FERRAMENTAS PARA SUA PRESERVAÇÃO: UMA REVISÃO DE LITERATURA

É evidente a relevância da preservação de páginas da *web* para a Ciência da Informação, para a memória e demais disciplinas que perpassam por esses nichos de pesquisa. Da mesma forma, a preservação de *websites* governamentais mostra o quanto já perdemos de informações memorialísticas, de prova e evidência ao longo do tempo, em razão da inexistência de uma política de preservação deste conteúdo informacional. São diversas as abordagens teóricas que se poderia apresentar neste capítulo, perfazendo a literatura da Ciência da Informação, abordagens da Arquivologia, da Memória Social, conformidades frente às teorias da governança de dados, tecnologias de preservação digital, filosofia, por meio da ética, por exemplo. Conforme Rockembach (2017) alguns aspectos éticos necessitam de reflexão no arquivamento da *web*, desde a coleta (transparência na seleção), acesso/uso (como a informação pode ser utilizada e manipulada) e preservação (de quem é a responsabilidade de preservar e por quanto tempo). Isso tudo evidencia o quanto ainda há para ser estudado com a temática do Arquivamento na *Web*.

No entanto, para este capítulo de Revisão de Literatura optamos por apresentar uma abordagem teórica em função do nosso objeto de análise: as páginas *web*. Motivo pela qual apresentamos a evolução da *web*, desde a criação do embrião da Internet, passando pela criação e estabelecimento da *web* como a mais importante aplicação da Internet, as circunstâncias em que ela chegou e se estabeleceu no Brasil, como está hoje e o que, enquanto Cientistas da Informação, podemos fazer para preservar este patrimônio informacional.

Para isso, dividimos o Referencial Teórico em três subcapítulos: Origem da Internet; a Internet no Brasil; e Arquivamento da *Web*, em que além dos conceitos gerais, são abordadas as teorias do *Modelo de Ciclo de Vida do Arquivamento da Web*, proposto por Bragg e Hanna (2013); a *Abordagem Sistemática para a Preservação da Web*, proposto por Khan e Rahman (2019); e uma sessão direcionada ao arquivamento da *web* governamental.

2.1 ORIGEM DA INTERNET

Atualmente, é improvável imaginar a vida sem redes sociais, e-mail, sites de busca e outros recursos provenientes da Internet. As páginas da *web* são um meio essencial para publicação, gerenciamento e disseminação de informações, e sua importância continua aumentando rapidamente dia após dia. A história da Internet, assim como seu atual incremento, também foi rápida. Mesmo que seu *boom* tenha sido nos anos 90, o princípio básico do que viria a ser uma das grandes invenções do homem contemporâneo foi em 1969, no entanto, seu objetivo não era para fins de negócios e entretenimento.

A história da Internet iniciou em 1962, na área militar dos Estados Unidos, em que o engenheiro Joseph Licklider, do Instituto Tecnológico de Massachusetts (MIT), já falava na criação de uma Rede Intergalática de Computadores com a intenção de ajudar a proteger o país durante as guerras (LEINER et al., 2009). Ele imaginou um conjunto globalmente interconectado de computadores através do qual todos pudessem acessar rapidamente dados e programas de qualquer site. Em geral, o conceito era muito parecido com a Internet de hoje, segundo o que Leiner et al. (2009) descrevem no livro "*A Brief History of the Internet*". Licklider foi o primeiro chefe do programa de pesquisa de computadores da DARPA, a Agência de Projetos de Pesquisa Avançada dos Estados Unidos, a partir de outubro de 1962. A rede que conectava os computadores da DARPA ficou estabelecida como o marco do "nascimento da Internet": o projeto da ARPANET inspirou a criação de uma rede que conectava várias redes, um conceito chamado de "*Internetworking*" (ABBATE, 1999) que, por sua vez, deu origem ao termo Internet, utilizado pela primeira vez em 1974 (GUIZZO, 1999).

Os anos 80 marcaram a história da Internet quando um conjunto de redes universitárias interconectadas puderam transmitir dados e trocar mensagens em rede (ABBATE, 1999). Em 1988, a rede foi aberta para interesses comerciais com serviços de correio eletrônico e provedores que faziam a conexão à rede pelo antigo método *dial-up* e, então, começou a "popularização" da grande rede (GUIZZO, 1999). Ainda nos anos 80 a *National Science Foundation* assumiu a

responsabilidade pela Internet e longe dos centros americanos de *networking*, no laboratório de física do CERN, em Genebra, Tim Berners-Lee aproveitou as capacidades únicas da Internet para inventar uma aplicação que ele chamou de *World Wide Web* (ABBATE, 1999; SCHATZ; HARDIN, 1994, p. 896). A *web* mudaria fundamentalmente a Internet, não apenas expandindo sua infraestrutura, mas fornecendo um conjunto de funcionalidades que atrairia milhões de novos usuários. A *web* também mudaria a percepção das pessoas sobre a Internet: em vez de ser vista como uma ferramenta de pesquisa ou até mesmo um canal para mensagens entre pessoas, a rede assumiu novos papéis como meio de entretenimento, vitrine e uma forma de apresentar-se ao mundo. (ABBATE, 1999, p. 213).

Foi no início dos anos 90 que a aplicação se consolidou e, desde então, começou a se utilizar o "www" antes do nome de qualquer site. A primeira página *web* do mundo ainda se encontra disponível na Internet[6], sendo essa considerada a gênese da conexão à Internet atual. Com a associação ao uso de linguagem HTML, áudios e vídeos, incorporados a estes *websites*, a rede passou a ter uma rápida expansão ainda na década de 90, quando as pessoas começaram a ter computadores pessoais e acesso, ainda discado, à grande rede (ABBATE, 1999). A crescente popularidade deste meio de comunicação fez com que se tornasse numa ferramenta usada pelas massas para publicação de informações: cada vez mais pessoas e instituições passaram a produzir conteúdo no ambiente *web* (ABBATE, 1999).

Berners-Lee apreciava o valor do *networking*; no entanto, ele viu uma limitação severa no fato de que, embora os computadores pessoais estivessem se tornando cada vez mais orientados para a imagem, a maioria dos usos da Internet ainda estava limitada ao texto (COMERFORD, 1995, p. 71). Ele imaginou um sistema que ajudasse os cientistas a colaborar, facilitando a criação e o compartilhamento de dados multimídia (BERNERS-LEE et al., 1994, p. 82). O CERN adotou o TCP/IP (*Transmission Control Protocol/ Internet Protocol*), no início dos anos 80, para fornecer um protocolo comum para seus vários sistemas, de modo que Berners-Lee projetou o novo serviço para rodar os protocolos da Internet (BERNERS-LEE et al., 1994, p. 82). Berners-Lee também criou um sistema de

[6] A primeira página *web* do mundo pode ser acessada através do endereço eletrônico http://info.cern.ch/hypertext/WWW/TheProject.html.

hipertexto que ligasse as linhas localizadas em computadores de todo o mundo, formando uma “rede mundial” de informações (ABBATE, 1999). A ideia de *hipertexto* surgiu na contracultura *hacker* dos anos 1960 e 1970, e Ted Nelson, um defensor vocal dessa contracultura, escreveu um manifesto, *Computer Lib*, no qual incitou as pessoas comuns a aprenderem a usar computadores, em vez de deixá-las nas mãos do “sacerdócio dos computadores” (NELSON, 1974 p. 601). Nelson, anos antes, já havia proposto um sistema de organização de informações que ele chamava de *hipertexto*, que possibilitaria vincular informações, em vez de apresentá-las de maneira linear (ABBATE, 1999, p. 214). Isso mudaria a aparência da Internet ao possibilitar o uso de documentos multimídia, como imagens, áudio e vídeos (SCHATZ; HARDIN, 1994, p. 897), ainda que para atingir este objetivo, Berners-Lee e seus colaboradores tivessem que enfrentar alguns desafios técnicos como, por exemplo, criar um formato único para documentos em *hipertexto*, o que eles chamavam de HTML - *Hypertext Markup Language* (ABBATE, 1999).

O HTML foi criado com a intenção de orientar a troca de informações entre navegadores e servidores da *web*, permitindo que ambos localizassem informações na rede. Além disso, também deveria haver alguma maneira uniforme de identificar informações para atender a necessidade de busca de usuário. Para isso foi criado um formato padrão de endereço que especifica tanto o tipo de protocolo de aplicativo que está sendo usado quanto o endereço do computador que possui os dados desejados, o *Uniform Resource Locator* (*URL*). (ABBATE, 1999, p. 215). A *URL* permitiria, também, o uso para outros protocolos anteriores, possibilitando acessar serviços de Internet mais antigos, como *FTP*, *Gopher*, *WAIS* e notícias da *Usenet* (SCHATZ; HARDIN, 1994).

A medida que a popularidade da rede se espalhou, um novo conjunto de grupos de interesse, basicamente formado por operadoras de telecomunicações, vendedores de produtos de rede e órgãos de normalização, exerceram influência sobre a evolução da Internet (WINSTON, 2003). A atual Internet comercialmente orientada e voltada para a comunicação surgiu apenas após um longo processo de reestruturação técnica, organizacional e política, tendo seu *boom* na década de 90, quando da decisão dos Estados Unidos em comercializá-la (WINSTON, 2003). A partir de 1992, a oferta de conteúdo era enorme e aumentava a cada ano com o surgimento dos grandes portais como AOL e Yahoo; ICQ e mIRC para bate-papo e

mensagens instantâneas; os serviços de e-mail gratuitos, como o Hotmail; e, claro, os sites de busca, como Google e o Cadê que passaram a revolucionar a sociedade (RYAN, 2010).

A Internet mudou muito nas duas décadas após o seu surgimento. Sua consolidação para o público em geral aconteceu com a facilidade de aquisição de computadores e também os avanços tecnológicos para as conexões que deixaram de ser discadas e passaram para a Banda Larga; equipamentos de comunicação exclusiva por voz deram lugar às conexões em 3G e 4G; apareceram os canais *streeming* e plataformas multimídia. A produção e uso de conteúdos pela *web* transformou a forma como nos comunicamos atualmente, e a Internet se tornou uma necessidade diária em nossa sociedade.

2.2 A INTERNET NO BRASIL

No Brasil, a implantação da Internet teve origem, inicialmente, no setor acadêmico, sem interferência do governo, e somente anos depois foi destinada a usuários domésticos e empresas. A Internet expandiu e passou a acompanhar a evolução mundial quando do interesse do governo federal na exploração da rede e após uma sucessão de mudanças por ele promovidas.

O Laboratório Nacional de Redes de Computadores (LARC)[7] elaborou, em junho de 1988, um anteprojeto para criação de uma Rede Nacional de Pesquisa (RNP) que passou a reunir as universidades brasileiras com a intenção de desenvolver pesquisas e abrir diálogo sobre o trabalho em rede (CARVALHO, 2006).

> Com o objetivo de integrar esses esforços e coordenar uma iniciativa nacional em redes no âmbito acadêmico, o Ministério da Ciência e Tecnologia formou um grupo composto por representantes do CNPq, da FINEP - Financiadora de Estudos e Projetos, da FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo, da FAPERJ - Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro e da

[7] Laboratório Nacional de Redes de Computadores (LARC) uma entidade “virtual” que visava integrar os esforços institucionais na área de redes de computadores, gerar um know-how de âmbito nacional nesta área, promover o intercâmbio de software e informação científica, através da integração de laboratórios de computação das instituições participantes. (CARVALHO, 2006, p. 74).

> FAPERGS - Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, para discutir o tema. Como resultado, surge em setembro de 1989 o projeto da Rede Nacional de Pesquisa (RNP), uma iniciativa da comunidade científica brasileira sob a égide institucional original da Secretaria de Ciência e Tecnologia da Presidência da República e posteriormente do Ministério da Ciência e Tecnologia, inspirada em iniciativas similares nos Estados Unidos (especialmente a NSFNet). A atuação da RNP é limitada aos âmbitos federal e internacional – nos estados iniciativas de redes estaduais integradas ao projeto nacional seriam estimuladas para a ampliação da capilaridade da rede. (AFONSO, 2000, p. 7).

A RNP veio a consolidar uma espinha dorsal nacional de ensino e pesquisa apesar dos recursos limitados disponíveis, não deixando dúvidas a respeito de seu papel pioneiro na alavancagem da Internet brasileira. Ainda em 1988 já se podia ver a formação de alguns embriões independentes de redes, interligando grandes universidades brasileiras e centros de pesquisa do Rio de Janeiro, São Paulo e Porto Alegre, aos Estados Unidos. (AFONSO, 2000).

No mesmo ano, o Brasil teve o primeiro contato com a Internet por meio da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), à época ligada à Secretaria Estadual de Ciência e Tecnologia. No mesmo período o Laboratório Nacional de Computação Científica (LNCC), localizado no Rio de Janeiro, conseguiu acesso à BITNET, através de uma conexão de 9600 bits por segundo estabelecida com a Universidade de Maryland. (VIEIRA, 2003, p. 9).

> O acesso à BITNET em setembro de 1988 foi uma vitória para o LNCC e para a comunidade acadêmica como um todo, ainda que não fosse possível a implementação do tão esperado gateway internacional no Brasil. A mesma reunião que liberou o acesso à BITNET também concluiu que a Embratel e o LARC envidariam esforços no sentido de uma solução que atendesse à necessidade de comunicação da comunidade acadêmica com as redes no exterior de forma otimizada. O fato é que esta decisão acabou reforçando os interesses de outras instituições que buscavam suas próprias conexões internacionais. (CARVALHO, 2006, p. 84).

Em 1989 a Fapesp também se conectou à BITNET, por meio de uma ligação com o *Fermi National Accelerator Laboratory* (Fermilab), em Chicago. Algum tempo depois, a fundação paulista criou a rede ANSP (*Academic Network at São Paulo*), interligando a Universidade de São Paulo (USP), a Universidade de Campinas (Unicamp), a Universidade Estadual Paulista (Unesp) e o Instituto de Pesquisas

Tecnológicas do Estado de São Paulo (IPT); meses mais tarde ligaram-se à rede paulista a Universidade Federal de Minas Gerais (UFMG) e a Universidade Federal do Rio Grande do Sul (UFRGS) (VIEIRA, 2003, p. 8). Também em 1989, a Universidade Federal do Rio de Janeiro (UFRJ) se ligou à rede BITNET, por meio da *University of California Los Angeles* (UCLA), constituindo-se no terceiro ponto de acesso à rede fora do país (AFONSO, 2000).

Paralelamente a implantação da RNP, em 1987 o Instituto Brasileiro de Análises Sociais e Econômicas (Ibase), no Rio de Janeiro, desenvolvia um projeto para implantação da Internet no âmbito das organizações não governamentais (ONGs): em julho de 1989 o *AlterNex*, criado pelo Ibase, passava a ser o primeiro provedor de serviços de Internet do país fora da comunidade acadêmica (AFONSO, 2000, p. 7). A iniciativa para implantação da Internet no Brasil foi possível graças a junção da RNP e do Ibase, em um esforço comum, para viabilizar o projeto Internet da "*UNCED Information Strategy Project in Rio*", no final de 1990, para a realização de um projeto para a ECO-92 que previa a montagem e operação de uma rede de microcomputadores interligando todos os espaços do evento, entre si e à Internet, o que deu um impulso definitivo à viabilização dos primeiros circuitos IP da nascente espinha dorsal acadêmica brasileira com a NSFNet nos EUA (AFONSO, 1996, p. 7). Eduardo Vieira, autor do livro "Os bastidores da Internet no Brasil", diz que o Ibase foi "[...] a primeira instituição brasileira fora do ambiente acadêmico a utilizar a Internet através do *AlterNex*, um serviço de correio eletrônico e grupos de discussão conectado à rede, em 18 de julho de 1989" (VIEIRA, 2003, p. 9). O teste do *AlterNex* ocorreu em 1992, durante a conferência internacional ECO-92, no qual foi montado um sistema de veiculação de informações eletrônicas para acompanhar o andamento dos debates (VIEIRA, 2003).

Estava consolidada a implantação da Internet no Brasil, ainda que com pequenos avanços e ampliações da rede ao longo dos anos, a situação permaneceu a mesma até meados 1994, quando a Internet ultrapassou as fronteiras acadêmicas e institucionais e começou a chegar à população brasileira (VIEIRA, 2003).

Quase no final de 94, sob o comando de Itamar Franco, o governo brasileiro divulgava, por meio do Ministério de Ciência e Tecnologia e do Ministério das Comunicações, a intenção de investir na nova tecnologia. A criação da estrutura

necessária para a exploração comercial da Internet ficou a cargo da Embratel e da RNP. A Embratel iniciou seu serviço de acesso à Internet comercial em caráter experimental (Figura 1). (VIEIRA, 2003).

Figura 1 - EMBRATEL lança o serviço de Internet comercial

COMUNICADO

No dia 20 de dezembro a EMBRATEL lançou o SERVIÇO INTERNET COMERCIAL. Em sua primeira fase, que vai até abril de 1995, estaremos prestando o serviço em caráter experimental.
A partir de hoje, você já pode se cadastrar no serviço de acesso à INTERNET, a maior rede de computadores do mundo, agora a seu alcance VIA EMBRATEL.
Como se cadastrar:
1 - Configure o seu software de comunicação com os seguintes parâmetros:
— Emulação de terminal - TTY
— 7 BITS, paridade par, 1 Bit de Stop (7 E 1)
2 - Através de seu modem acesse a RENPAC por um dos seguintes números:
078 78228 - (9.600BPS)
078 78224 - (2.400BPS)
3 - Aguarde a resposta "RENPAC" em seu vídeo e logo após digite 0724021050400 e tecle ENTER.
4 - O sistema pedirá seu "USER NAME" e sua "PASS-WORD". Responda INTERNET e EMBRATEL respectivamente.
5 - A partir daí, preencha os campos da tela.
Feliz festas e votos de um bom Ano Novo, navegando na INTERNET VIA EMBRATEL

EMBRATEL

Fonte: Vieira, 2003, p. 20.

Alguns meses depois, em maio de 1995, o acesso à Internet via Embratel começou a funcionar de modo definitivo. Mas a exclusividade da estatal no serviço de acesso a usuários finais desagradou à iniciativa privada, que temia que a Embratel dominasse o mercado, criando um monopólio estatal da Internet no Brasil (VIEIRA, 2003, p. 10). Porém, a eleição presidencial de 1994 trouxe consigo uma agenda política que previa um amplo programa de privatizações, incluindo a desestatização do setor de telecomunicações e diante deste cenário, já sob o governo de Fernando Henrique Cardoso, o Ministério das Comunicações tornou pública a posição do governo de que não haveria monopólio e que o mercado de serviços da Internet no Brasil seria o mais aberto possível (VIEIRA, 2003).

Vieira (2003, p. 10) diz que “Bastou Fernando Henrique Cardoso ascender ao Palácio do Planalto, em 1º de janeiro de 1995, para os planos da Embratel de se estabelecer sozinha no mercado de Internet serem freados bruscamente”. A mídia vinha acompanhando com atenção o andar da nova tecnologia no Brasil: a Revista Veja, edição número 1390, de 3 maio 1995, trouxe a mudança de estratégia do governo federal afirmando que o governo pretendia inaugurar o acesso público à

Internet ainda em maio, mas preferiu reformular as regras ao publicar uma portaria autorizando qualquer empresa, seja ela pública ou privada, a oferecer acesso à Internet (VEJA, 1995).

Em 1995 surgiram nos Estados Unidos alguns dos mais importantes nomes da Internet, como o site de busca *Yahoo!* e a livraria virtual *Amazon.com*, além dos primeiros protagonistas da *web* comercial brasileira e, por isso, pode ser considerado o marco-zero da Internet comercial no Brasil e no mundo (VIEIRA, 2003, p. 11). Foi certamente a partir desse momento que o grande público brasileiro tomou conhecimento desta rede (VIEIRA, 2003). A partir daí apareceram diversos provedores de acesso à Internet no Brasil, assim como grandes portais de conteúdo e comércio eletrônico (CARVALHO; CUKIERMAN, 200-, p. 14).

Ainda sob interferência das reformas governamentais nas políticas para as telecomunicações, em 31 de maio de 1995 é publicada a Portaria Interministerial número 147 que cria o Comitê Gestor da Internet no Brasil (CGI.br) (BRASIL, 1995). O objetivo do CGI.br era "[...] assegurar qualidade e eficiência dos serviços ofertados, justa e livre competição entre provedores, e manutenção de padrões de conduta de usuários e provedores, e considerando a necessidade de coordenar e integrar todas as iniciativas de serviços Internet no país [...]" (BRASIL, 1995), com atribuições tais como: fomentar o desenvolvimento de serviços da Internet no Brasil, recomendar padrões e procedimentos técnicos e operacionais, além de coletar, organizar e disseminar informações sobre os serviços da Internet (BRASIL, 1995).

Com os serviços de Internet já melhorados, em 1996, o Brasil pôde entrar no cenário mundial da Internet e passou a acompanhar a evolução da rede, muito em função da melhoria nos serviços prestados pela Embratel, mas principalmente pelo crescimento natural do mercado (CARVALHO, 2006). A *web* brasileira vivenciava o surgimento de centenas de pequenos provedores de acesso à rede, enquanto acompanhava as regras ditadas pelos pioneiros da Internet comercial, a *Yahoo!* e a *Amazon.com* (VIEIRA, 2003). "A venda de computadores pela primeira vez havia ultrapassado a de aparelhos de TV e o consumo de linhas telefônicas aumentou gradualmente [...]" (VIEIRA, 2003, p. 16), até a privatização do setor em 1998. O *Internet banking* das instituições financeiras do Brasil tornou-se referência internacional; jornais, revistas e demais meios de comunicação impressa também

acompanharam o fenômeno de perto, inaugurando suas versões digitais (MIELNICZUK et al., 2015). Vieira (2003, p. 16) diz que “A prova definitiva de que a Internet havia chegado realmente para valer na vida das pessoas foi a criação da declaração *online* do Imposto de Renda [...]”, pela qual a Receita Federal conseguiu praticamente eliminar o uso do papel ao transportar sua principal operação anual à Internet.

2.3 ARQUIVAMENTO DA *WEB*

Masanès (2006, p.1) diz que “A Internet é atualmente o alicerce sobre o qual a informação é obtida [...]”, sendo o *World Wide Web* o recurso mais utilizado: hospeda centenas de milhões de sites, que conectam indivíduos, cidades e o mundo em geral, usando recursos avançados da tecnologia da *web*. A rede está em constante estado de evolução e não há garantia de que seu conteúdo atual permanecerá acessível em um futuro próximo. A mudança do conteúdo é causada por diferentes motivos, que vão desde a decisão pessoal do proprietário do *website*; edição de partes do conteúdo; até alterações acidentais que eventualmente podem ocorrer; ou até mesmo alterações nos domínios estão sujeitos (DAY, 2003). Devido a essa natureza dinâmica, o conteúdo *web* está em constante mudança e um número considerável de sites tem vida útil bastante curta (DAY, 2003).

Em reconhecimento a esse problema, organizações públicas e privadas de todo o mundo têm investido no desenvolvimento e na aplicação de ferramentas que oferecem suporte e soluções para a preservação da *web*. A comunidade internacional de arquivamento da *web* está em constante desenvolvimento e novas ferramentas são criadas para melhorar as técnicas de preservação e para diminuir a perda sistemática de conteúdo *online* causada pela natureza transitória da *web*. Iniciativas que visam entender que o conteúdo presente da Internet também faz parte do patrimônio cultural e social precisam evoluir para combater a efemeridade desses materiais. Deve-se ter certeza de que essas informações, além de serem acessíveis em todo o mundo, perdurem ao longo do tempo para transmitir conhecimento às gerações futuras. Os arquivos da *web* são sistemas inovadores

que adquirem, armazenam e preservam informações publicadas na Internet (BROWN, 2006) e, assim como os espaços habituais de preservação da memória, os arquivos da *web* também se constituem como fonte para pesquisas e podem se consolidar como espaços fundamentais para a salvaguarda de informações de uma época.

Os artefatos culturais do passado sempre tiveram um importante papel na formação da consciência, na autocompreensão de uma sociedade e na construção de seu futuro. A cultura intangível da nossa sociedade é preservada por instituições culturais, como os museus, bibliotecas e arquivos. Questões relacionadas ao folclore, tradições, linguagem, além do legado de artefatos físicos, como monumentos, livros e obras de arte já possuem, ainda que em algumas instâncias de forma falha, tradição de serem preservados, salvaguardados e acessados pela sociedade. Não obstante, “Os arquivos da *web* [também] são uma nova forma de instituições de patrimônio cultural que preservam artefatos semelhantes” (COSTA; GOMES; SILVA, 2016, p. 2), a diferença é que estes artefatos são natos digitais ou digitalizados. A *World Wide Web*, em suma, é uma mídia difundida e efêmera em que a cultura moderna, em grande medida, encontra uma forma natural de expressão (MASANÈS, 2006, p. 1) e, portanto, diversos aspectos da sociedade estão acontecendo ou são refletidos na Internet em geral, problemáticas como debates, criações, trabalho e interação social em um sentido amplo (MASANÈS, 2006). Alguns desses recursos já desapareceram para sempre e os arquivos da *web* criam várias maneiras de capturar a *web* e manter cópias como evidência de um tempo (HOLUB; RUDOMINO, 2014).

À medida que o tempo vem passando, desde o início da *web*, aumenta, também, a quantidade de *websites* e, com isso, surgem questões metodológicas sobre o tratamento do próprio conteúdo como documentação primária. Embora seja evidente que a quantidade de *URLs* inativos está aumentando gradualmente (LAWRENCE et al., 2001; DELLAVALLE et al., 2003; KOEHLER, 2004), existe o risco de que as futuras gerações possam não ter acesso ao patrimônio cultural acumulado por seus predecessores (LYMAN, 2002). O arquivamento da *web* é uma maneira de neutralizar essas características da rede e de garantir acesso a longo prazo às informações da *web* (BRÜGGER, 2005; MASANÈS, 2006).

Por essa razão, a busca por soluções de preservação da *web* é uma necessidade cultural e histórica. Com a migração do registro escrito do papel para o formato digital, arquivistas, historiadores, cientistas e demais profissionais interessados na preservação do patrimônio documental, devem considerar como o conteúdo da *web* deve ser conservado, recuperado e analisado (MASANÈS, 2006), ainda que se exija uma revisão radical das práticas tradicionais de preservação (MASANÈS, 2006, p. 1). Em 2003, a UNESCO, através da *Carta sobre Preservação do Patrimônio Digital*, manifestou que os materiais digitais incluem textos, bancos de dados, imagens estáticas e em movimento, áudios, gráficos, *software* e páginas da *web*, em uma ampla e crescente gama de formatos, considerando que estes documentos são frequentemente efêmeros e exigem que a produção, a manutenção e a gestão sejam mantidos. Alertou ainda, no mesmo documento, que os materiais nos sites da *web* nem sempre são reconhecidos como registros, mas também são necessários os devidos tratamentos destas informações através de procedimentos para identificá-los e gerenciá-los, fazendo com que estes registros *web* permaneçam válidos (UNESCO, 2003).

No passado, partes importantes de nossa herança cultural foram perdidas porque não foram arquivadas - em parte porque as gerações passadas não reconheceram ou não puderam reconhecer seu valor histórico (LOWENTHAL, 2006). Além disso, as gerações passadas não abordaram o problema técnico de preservar a mídia de armazenamento - filme de nitrato, fita de vídeo, gravações de vinil - ou o equipamento para reproduzi-las e, também, não foi resolvido o problema referente às questões legais, como a criação de leis e acordos de proteção material com direitos autorais e, ao mesmo tempo, permitir sua preservação nos arquivos (LOWENTHAL, 2006). Esses problemas também precisam ser enfrentados em relação a *web*.

A *web* conta com pelo menos 6,25 bilhões de páginas[8] e seu conteúdo é estruturado de diferentes maneiras, sendo composto por muitos formatos de arquivo. No entanto, estas páginas estão sendo constantemente atualizadas, realocadas ou removidas e nem sempre se pode voltar a algo que se viu antes. Sítios da *web* desaparecem diariamente conforme seus proprietários os revisam ou

[8] Dados extraídos do website www.worldwidewebsize.com, em 05 de fevereiro de 2020.

os servidores são colocados fora de serviço, os usuários só descobrem quando inserem uma *URL* e recebem uma mensagem de erro "*404 Site Not Found*" (LYMAN, 2002). O conteúdo é perdido a um ritmo alarmante, arriscando não apenas nossa memória cultural digital, mas também a responsabilidade organizacional (PENNOCK, 2013). Para ajudar a preservar o conteúdo da *web*, os sites são capturados e arquivados para acesso de longo prazo por meio do arquivamento da *web*.

Nicolas Delalande e Julien Vincent observaram que a preservação de *websites* é um desafio bem identificado há vários anos, e cuja relevância ainda está crescendo, mesmo que sua apropriação e estudo por parte dos acadêmicos é mais lenta de desenvolver e os arquivos da *web* ainda carecem de visibilidade como fonte de pesquisa (DELALANDE; VINCENT, 2011). Os arquivos da *web* destacam novos desafios para a captura e armazenamento de informações digitais, especialmente no que diz respeito a novos campos de estudo, como patrimônio digital e humanidades digitais: uma nova área de aplicação de técnicas computacionais de pesquisa e comunicação nas humanidades (HAYLES, 2012). Ainda que Richard Rogers afirme que "[...] falamos até mesmo de uma crise em certos campos das Humanidades Digitais, porque os acadêmicos não receberam/reconheceram esses novos documentos de maneira proporcional à corrida para a digitalização [...]", sendo pouco utilizados para fins científicos (2015, *tradução nossa*); outros pesquisadores, tais como Brügger (2016), Gomes; Costa (2014) e Winters (2017), dizem que os arquivos da *web* foram identificados como fontes de informação para pesquisa e, especialmente, para trabalhos emergentes em humanidades digitais. Exemplos destes trabalhos podem ser encontrados em antologias recentes que refletem uma diversidade de abordagens para o estudo do histórico da *web* com o uso de arquivos da *web* (BRÜGGER, 2017).

Os arquivos da *web* têm grande potencial, tanto como fonte (SCHAFER; THIERRY, 2015), quanto como objeto de análise. Hélène Bourdeloie, observa que a verdadeira inovação é o fato de que as mídias e tecnologias digitais não são mais ferramentas para o serviço de pesquisa, mas em vez disso, elas também são objetos de pesquisa em si mesmos; são, ao mesmo tempo, instrumento, método, campo de pesquisa e objeto de estudo (BOURDELOIE, 2013). Gomes (2010) diz que os arquivos da *web* poderão ser úteis para diversos perfis de usuários, tais

como jornalistas, gestores de *websites*, historiadores, usuários comuns em busca de *links* quebrados, juristas em busca de provas, entre outros casos.

Costa, Gomes e Silva (2016) dizem que arquivos da *web* são um tipo de biblioteca digital, e ambos compartilham a responsabilidade de preservar informações para as gerações futuras. Argumentam ainda que a principal diferença é que os arquivos da *web* geralmente crescem numa proporção maior do que as bibliotecas digitais. A *web*, 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 e acesso. A esse conjunto de atividades relativas à preservação do ambiente *web* é dado o nome de Arquivamento da *web*.

Niu (2012, p.1, *tradução nossa*) definiu arquivamento da *web* como “[...] o processo de coleta de dados que foram registrados na *World Wide Web*, armazenando-os, garantindo que os dados sejam preservados em um arquivo e disponibilizando os dados coletados para pesquisas futuras”. Rockembach (2018b, p. 241) diz que o arquivamento da *web* proporciona a futuros pesquisadores os *websites* e objetos digitais disponíveis na Internet por meio de plataformas digitais de armazenamento e recuperação da informação: “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).

Algumas instituições já desenvolvem programas para preservar *websites* em seus países, encarando de forma efetiva o desafio de preservar o conteúdo produzido na Internet. Desafio que perpassa tanto as questões técnicas e tecnológicas que envolvem o arquivamento da *web*, quanto na coleta destes dados que mudam, somem e se alteram rapidamente, considerando a natureza interativa da Internet.

Ao longo de mais de uma década de estudos em relação à preservação da Internet foram criadas algumas iniciativas de arquivamento da *web* que definem os padrões, políticas e tecnologias para efetivação deste arquivamento. Dois consórcios internacionais foram criados com a intenção de fortalecer os estudos e

a coleta de *websites*: em 1994 é criado o *World Wide Web Consortium*[9] (W3C), liderado pelo inventor da *web*, Tim Berners-Lee, e pelo CEO Jeffrey Jaffe. É uma comunidade internacional em que as organizações membros, os funcionários e o público em geral trabalham para desenvolver padrões da *web*. A missão do W3C é aproveitar ao máximo o potencial da *web* por meio do desenvolvimento de protocolos, diretrizes e padrões que garantam seu crescimento a longo prazo; e discutir o seu estado e examinar possíveis melhorias nessa área. A afiliação ao W3C está aberta a todos os tipos de organizações, incluindo entidades comerciais, educacionais, governamentais e indivíduos. Atualmente, o consórcio conta com 451 membros. (WORLD WIDE WEB CONSORTIUM, website).

Em 2003, a Biblioteca Nacional da França criou o *International Internet Preservation Consortium* (IIPC)[10]. A missão do IIPC é adquirir, preservar e tornar acessível o conhecimento e a informação da Internet para as futuras gerações, promovendo o intercâmbio global e as relações internacionais. Para cumprir com sua missão, o IIPC tem por objetivo a coleta de conteúdo da Internet de todo o mundo; promover o desenvolvimento e uso de ferramentas, técnicas e padrões comuns que permitam a criação de arquivos internacionais e incentivar e apoiar bibliotecas nacionais, arquivos e organizações de pesquisa em todos os lugares para tratar do arquivamento e preservação da Internet. Atualmente, o consórcio conta com aproximadamente 60 membros de mais de 45 países, incluindo bibliotecas e arquivos nacionais, universitários e regionais. Dentre eles estão a maior plataforma do mundo para arquivamento da *web*, a *Internet Archive*; arquivos nacionais como o da França, Canadá, Espanha, Portugal, dentre outros; e entre as bibliotecas que fazem parte do IIPC, destacamos a *British Library (*Biblioteca Britânica) e a *Library of Congress* (Biblioteca do Congresso Americano). (INTERNATIONAL INTERNET PRESERVATION CONSORTIUM, website, 2019).

Em pesquisa realizada por Gomes, Miranda e Costa em 2011, foram levantadas e analisadas 42 iniciativas de arquivamento da *web* (GOMES; MIRANDA; COSTA, 2011). A partir deste estudo, os autores produziram uma página na wikipedia.org[11], que apresenta mais de 80 iniciativas de arquivamento da *web*,

[9] https://www.w3.org
[10] http://netpreserve.org/
[11] https://en.wikipedia.org/wiki/List_of_*web*_archiving_initiatives

tornando evidente como vêm surgindo diversas iniciativas ao redor do mundo. Dentre os pioneiros do arquivamento da *web* destacamos o *Internet Archive*[12]; o projeto da Biblioteca Nacional Australiana PANDORA[13] (do inglês, *Preserving and Accessing Networked Documentary Resources of Australia*); e a iniciativa Kulturarw3[14] da Suécia, todos iniciados em 1996. Além destes, cabe destacar a *UK Web Archive*, a plataforma britânica que iniciou suas atividades em 2003, com o projeto *UK Government Web Archive* e se consolidou como plataforma de arquivamento da *web* nacional em 2004 (UK WEB ARCHIVE, website, 2019).

O *Internet Archive*, uma fundação sem fins lucrativos sediada nos EUA, foi um dos primeiros arquivos da *web* e vem arquivando amplamente *websites* desde 1996, se consolidando como a principal plataforma de arquivamento da *web* no mundo (MASANÈS, 2006). Atualmente, o *Internet Archive* mantém arquivado aproximadamente 330 bilhões de *webpages*[15]. É também um dos membros fundadores do Consórcio Internacional de Preservação da Internet (IIPC) (INTERNET ARCHIVE, website, 2019a). Em 2002, o *Internet Archive* lançou o *Heritrix*[16], um rastreador da *web* de código aberto, uma ferramenta de *software* que captura o conteúdo da *World Wide Web* (INTERNET ARCHIVE, website, 2019a), que se tornou a tecnologia de rastreamento de arquivos da *web* mais popular e difundida atualmente. Os resultados dos rastreamentos são armazenados em um arquivo *WebARChive* (WARC), um formato desenvolvido pelo IIPC, que em 2009 foi adotado como extensão padrão para arquivos *web* (MAEMURA et.al, 2018), definido na ISO 28500[17].

Outra iniciativa que se destaca como uma das pioneiras é o projeto PANDORA, criado em 1996, pela *National Library of Australia* com a ideia de construir um arquivo de sites australianos selecionados e significativos (CATHRO; WEBB; WHITING, 2001). Os objetivos do projeto são o desenvolvimento de políticas e procedimentos para a preservação e acesso às publicações *online* australianas; se constituir como um arquivo de publicações selecionadas; estabelecer os recursos

[12] https://archive.org/
[13] https://pandora.nla.gov.au/
[14] https://www.kb.se/
[15] Dados extraídos de https://archive.org/about/, em 17 de fevereiro de 2020.
[16] http://crawler.archive.org/index.html
[17] https://www.iso.org/standard/68004.html

necessários para a implementação de estratégias de preservação de médio e longo prazo; e desenvolver uma proposta de abordagem nacional para a preservação a longo prazo de publicações *online* (PANDORA, website, 2019). Se constitui em uma abordagem seletiva da *web*, sendo uma amostra fortemente representativa promovida por organizações acadêmicas, governamentais, comerciais e comunitárias, demonstrando que o conteúdo do arquivo *web* PANDORA é baseado em um conjunto de diretrizes de seleção, e reflete o fato de que todo site arquivado é submetido a rigoroso critério de curadoria (CATHRO; WEBB; WHITING, 2001).

Na Suíça, em 1661, a Biblioteca Real (Kungl.biblioteket, abreviada como KB) foi encarregada de coletar todas as publicações impressas suecas. Desde então, a KB coletou, preservou e deu acesso a uma parte importante da memória cultural e histórica da Suécia, até que em 1996, em continuação a esta tarefa, a biblioteca inaugurou um projeto intitulado *Kulturarw3*, que tinha a intenção de testar métodos de coleta, preservação e acesso a documentos eletrônicos suecos, acessíveis *online* (ARVIDSON, 2001, p. 101). O escopo do projeto é coletar todo conteúdo *online* sueco, sem seleção de conteúdo ou formato de arquivo: justificam essa abordagem com o argumento que não se sabe quais informações serão importantes para as futuras gerações; que uma seleção exigiria demanda de mão-de-obra e que o armazenamento de computadores também está ficando mais barato (ARVIDSON, 2001).

Outro repositório de nível nacional é a *British Library*, que está disponibilizando cada vez mais sites britânicos com domínio .uk. Tratam-se de sites cujos domínios foram registrados no Reino Unido, e que a *British Library*, portanto, considera dentro de suas políticas de coleta. Suas *URLs* têm sufixos como: co.uk (empresas); ac.uk (instituições acadêmicas); gov.uk (governo nacional e local) e org.uk (organizações da sociedade civil) (DIKWATTA; DIAS, 2017). A iniciativa de arquivamento da *web* no Reino Unido iniciou com o projeto para coletar arquivos da *web* governamental: o *UK Government Web Archive (UKGWA)*[18], idealizado pelo *The National Archives*, teve origem em 2003, com o objetivo de coletar e arquivar sites governamentais de interesse do público britânico, em parceria com outras instituições de arquivamento da *web*, como o *Internet Archive* e o programa

[18] http://www.nationalarchives.gov.uk/webarchive/

European Archive (THE NATIONAL ARCHIVES, *website*, 2019). No mesmo ano, foi criado o *UK Web Archiving Consortium (UKWAC)*, estabelecendo uma plataforma compartilhada para selecionar, coletar e conceder acesso público às páginas arquivadas do Reino Unido; e uma política integrada, considerando o escopo de coleta diferente de cada membro do consórcio, identificando interesses comuns e pontos institucionais específicos para a preservação do conteúdo da *web* (THE NATIONAL ARCHIVES, *website*, 2019).

2.4 ARQUIVAMENTO DA WEB GOVERNAMENTAL

Em 2003, reconhecendo a eminente produção de informações natodigitais, a UNESCO lançou a *Carta sobre a Preservação do Patrimônio Digital*, que promoveu a adoção de medidas efetivas para a preservação destas informações como, por exemplo, a possibilidade de se instituir o depósito legal para materiais da *web*, identificado nesta carta como uma das ações para a preservação do patrimônio digital (UNESCO, 2003, Art. 8). Essa responsabilidade da coleta documental para o domínio público tem se estabelecido como uma importante ferramenta para preservar a memória oficial (TOSH, 2002), ainda que historicamente, antes da década de 1990, pesquisadores confiavam às agências públicas a coleta e conservação de cópias impressas em papel (THE NATIONAL ARCHIVES, 2006).

Essa alta escala de produção de informações digitais pode ser exemplificada a partir da análise dos dados mais atualizados do *Instituto Brasileiro de Geografia e Estatística* (IBGE), que mostrou que em 2016, 116 milhões de pessoas estavam conectadas à Internet, o que equivale a 64,7% da população brasileira com idade acima de 10 anos (IBGE, *website*, 2018)[19]. Da mesma forma, o CGI.br/NIC.br, *Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação* (Cetic.br), divulgou que em 2018, 67% dos domicílios brasileiros possuíam Internet, e a percentagem de usuários da rede aumentou para 76% da população brasileira (CGI.br, *website*, 2018)[20]. O mesmo instituto, em pesquisa no ano anterior, mostrou

[19] As informações são da Pesquisa Nacional por Amostra de Domicílios Contínua (Pnad C), divulgadas em 21 de fevereiro de 2018, pelo IBGE.

[20] Fonte: https://cetic.br/tics/domicilios/2018/individuos/

que 100% dos órgãos públicos Federais e Estaduais utilizam Internet, sendo que 90% destes órgãos possuem *websites* (CGI.br, *website*, 2017)[21].

Os *websites* das agências governamentais desempenham importante papel na disseminação de informações para o público e seu conteúdo pode mudar diariamente, e algumas informações podem ser removidas permanentemente. Isso pode levar à perda de informações valiosas para pesquisas e até mesmo para a prestação de contas das ações do próprio governo. Atualmente, as informações estão sendo produzidas em grandes quantidades tanto em formatos tradicionais, como, cada vez mais, em formatos eletrônicos ou digitais, porém, ainda são poucos os que reconhecem o valor e a importância em preservar os *websites*, mesmo que existam algumas importantes iniciativas neste contexto, corroboram os autores Lala e Joe (2006).

O *The National Archives* é o mantenedor do *UK Government Web Archive* (UKGWA), um dos maiores e mais utilizados arquivos da *web* do mundo, contendo mais de três bilhões de *URLs* e frequentemente recebendo mais de dez milhões de visualizações de página por mês, com a missão de preservar o conteúdo da *web* pertencente ao governo em todos os seus formatos, ainda que seu conteúdo central seja composto por material publicado pelos departamentos do estado (THE NATIONAL ARCHIVES, 2014). Estes *websites* são identificados pelo *The National Archives* e pelas próprias organizações governamentais (ESPLEY et al., 2014).

A abordagem do *The National Archives* em relação ao arquivamento da *web* governamental envolve a coleta remota e automática de *websites* de acordo com um cronograma, com o uso de um rastreador. A iniciativa é parte de um amplo programa que envolve a gestão do patrimônio da *web* do governo. Em 2017 foi elaborado o documento *The UK Government Web Archive: Guidance for digital and records management teams*. Trata-se de uma orientação às equipes digitais do governo para que entendam como gerenciar e manter *websites* e garantir que a presença do governo na *web* possa ser arquivada com êxito e permanentemente acessível no UKGWA. São fornecidas informações sobre o funcionamento do processo de arquivamento dos *websites*, o cronograma das capturas, as limitações do que pode ser capturado e disponibilizado por meio deste sistema de preservação

[21] Fonte: https://cetic.br/pesquisa/governo-eletronico/

e as circunstâncias em que o conteúdo pode ser removido. (THE NATIONAL ARCHIVES, 2017).

Nos Estados Unidos, a partir de 2008, um grupo de instituições criou um arquivo *web* colaborativo chamado *End of Term Web Archive: U.S. Government Websites* (EOT), com coleções constituídas por *websites* do governo federal (.gov, .mil) nos ramos legislativo, executivo e judicial do governo. *Websites* que corriam o risco de mudar (por exemplo *whitehouse.gov*) ou desaparecer completamente durante as transições do governo, foram capturados. Atualmente, o EOT é composto das coleções de *websites* do final da administração Bush (2008) e do fim dos dois mandatos do governo Obama (2012 e 2016). (EOT, website, 2019).

Outra importante ação norte americana em relação a preservação de *websites* governamentais, diz respeito ao movimento promovido por cientistas para salvaguardar informações governamentais sobre mudanças climáticas, se antecipando ao risco de que dados e informações oriundas do .gov, como por exemplo os *websites* da EPA (*Environmental Protection Agency*)[22] e da NOAA (*National Oceanic and Atmospheric Administration*)[23], pudessem ser perdidos ou ficassem indisponíveis com a transição para a nova administração que estava se efetivando com a eleição de Donald Trump (VICE, *website*, 2016).

Alyssa S. Rosen (2017), em seu artigo sobre o trabalho de proteção dos *websites* com dados a respeito das mudanças climáticas, traz trechos de uma entrevista com a professora Sarah Lamdan[24], que ressalta a importância em arquivar informações oriundas do .gov, seja de natureza ambiental ou de qualquer outra natureza dizendo que do ponto de vista arquivístico, toda a informação é importante, porque o conteúdo na maioria desses sites .gov do Poder Executivo vai mudar nos próximos anos; e do ponto de vista da preservação, salvar tudo é uma boa regra geral.

Outra iniciativa de arquivamento da *web* governamental que merece destaque é a iniciativa do estado de Sarawak, na Malásia. Seu objetivo é preservar evidências dos conteúdos da *web* publicados pelos departamentos e agências da

[22] https://www.epa.gov/
[23] https://www.noaa.gov/
[24] Professora Associada da Biblioteca de Direito da CUNY School of Law.

administração pública do Estado de Sarawak; contribuir para facilitar o acesso e a disponibilização de informações à pesquisa; e cumprir com a legislação da Biblioteca Estadual de Sarawak referentes ao depósito legal (JAMAIN, 2018). A captura de *websites* acontece a cada dois meses e, além dos documentos textuais, inclui a salvaguarda das imagens estáticas, gravação de som, filmes e outros formatos multimídia disponibilizados nos portais do governo.

O depósito legal tem sido uma ferramenta normatizadora utilizada para, em alguns países, justificar a necessidade de preservação de *websites*, ainda que até o século XX seu escopo esteva voltado apenas aos livros e publicações impressas. Com o advento e uso massivo de tecnologias de gravação de som, vídeos e informações digital como um todo, o escopo do depósito legal foi estendido para incluir uma variedade de formatos de documentos. A nível internacional, reconhecendo que a produção em massa de informação natodigital está cada vez mais em uso nas rotinas de produção e consumo da informação e vendo a necessidade de garantir a preservação da informação, a UNESCO divulgou sua carta anteriormente mencionada, promovendo, desta forma, a adoção de medidas efetivas para a preservação da informação digital. (PABÓN CADAVID, 2014).

Tradicionalmente, as bibliotecas nacionais têm a prerrogativa de preservação dos livros e materiais impressos produzidos sob jurisprudência do país, regramento este estabelecido pela legislação do depósito legal. A tradição de recolhimento de material impresso pelas bibliotecas, somado ao advento das tecnologias, uso e consumo da informação em formato digital, atualmente, parte da literatura sobre o arquivamento na *web* indica que as bibliotecas nacionais têm um papel significativo na preservação da *web* (SHVEIKY; BAR-ILAN, 2013).

Em função disso, algumas bibliotecas nacionais começaram a construir coleções de *websites* a partir do início dos anos 90 (DAY, 2003b). Essas bibliotecas passaram a considerar as coleções da *web* como um desenvolvimento natural de sua coleção tradicional e como parte de seu dever em preservar a cultura nacional (LAZINGER, 2013). No início dos anos 2000, alguns países acrescentaram às suas leis de depósito legal uma exigência para depositar publicações eletrônicas em formatos *online* (HAKALA, 2004), tais como Tasmânia, Suíça, Islândia e Nova Zelândia (SHVEIKY; BAR-ILAN, 2013). A Finlândia, Islândia, Noruega e Suécia, por

exemplo, decidiram coletar *websites* locais que atendem ao amplo critério de serem "nacionais" (DAY, 2003a). Ainda que o relatório final da "Conferência Internacional de Arquivamento de Recursos da *Web*", realizada em 2004, relata que os participantes geralmente concordam que é impossível coletar e arquivar tudo e apresenta as abordagens para a questão da seleção, mas conclui que não há uma maneira correta de decidir o que coletar (NATIONAL LIBRARY OF AUSTRALIA, 2004).

Na Croácia, a Biblioteca Nacional e Universitária de Zagreb, em colaboração com a *University of Zagreb Computing Centre* (SRCE) estabeleceu, em 2004, a partir de legislações, que todos *websites* registrados no país teriam uma cópia entregue à Biblioteca. Em 2011 o HAW (*Hrvatski arhiv weba*) ou, em inglês, *Croatia Web Archive*, com a intenção de ampliar o escopo da coleção nacional de *websites*, recolheu o domínio nacional (.hr) pela primeira vez, e iniciou uma coleta temática de conteúdo da *web* relacionada a eventos nacionais, como por exemplo, às eleições locais de 2013 para os territórios governamentais. (HOLUB, 2014).

Com estes exemplos de iniciativas pode-se perceber que o arquivamento da *web* é um trabalho interdisciplinar, que envolve conhecimentos e rotinas vindas das ciências da informação, em que arquivos e bibliotecas convergem para a preservação digital, se associando com as disciplinas da tecnologia, pois tratam-se de documentos que demandam comunicação entre os saberes. Para além destes conteúdos que são fundamentais ao arquivamento da *web*, há outras disciplinas e profissionais que têm espaço na formação do conhecimento referente a esta temática, como é o caso das ciências da comunicação, da engenharia, ética, direito e tantos outros conhecimentos que podem contribuir para o desenvolvimento do tema.

No Brasil, o arquivamento da *web* ainda é uma novidade na academia. A Universidade Federal do Rio Grande do Sul, criou em 2017, o *Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital* (NUAWEB) “[...] com o objetivo de investigar características do arquivamento da *web* por meio de iniciativas nacionais e internacionais, lidando tanto com as políticas, quanto as tecnologias envolvidas no processo” (NUAWEB, *website*, 2019). O grupo de pesquisa estuda aspectos da preservação, uso e acesso ao longo do tempo de objetos digitais disponibilizados

na *web*, com contribuições da Arquivologia, Biblioteconomia, Ciência da Informação, Comunicação e Ciência da Computação. O NUAWEB está desenvolvendo dois projetos de pesquisa simultaneamente: o projeto *AWEB – Arquivamento da Web das Eleições Brasileiras de 2018*; e o projeto *Arquivamento da web brasileira: políticas de preservação e modelos tecnológicos*. O núcleo de pesquisa também apresentou em junho de 2019, durante a *International Internet Preservation Consortium Web Archiving Conference 2019*, em Zagreb, Croácia, o embrião da iniciativa brasileira para arquivamento da *web*, que estará disponível a partir de 2020 através do endereço *http://www.arquivo.ong.br/* (ROCKEMBACH; MELO, 2019).

Ainda no Brasil, porém na esfera política, outra ação a respeito do arquivamento da *web* está em desenvolvimento desde julho de 2015, o Projeto de Lei (PL) 2.431/2015, de autoria da Deputada Luizianne Lins (PT/CE), que "Dispõe sobre o patrimônio público digital institucional inserido na rede mundial de computadores e dá outras providências" (BRASIL, 2015). Ainda em 2015, o PL passou pela *Comissão de Ciência e Tecnologia, Comunicação e Informática* (CCTCI), da Câmara dos Deputados, e teve parecer para aprovação com substitutivo proferido pelo Deputado Fábio Sousa (PSDB/GO), que em seu relatório apresenta o projeto considerando que sua aprovação "[...] visa ampliar as proteções dadas à informação pública, mais especificamente àquela armazenada na internet" (SOUSA, 2015, p. 2). O relatório solicita modificações na redação do projeto e acréscimo de um parágrafo que diz que "[...] devem ser estabelecidas diretrizes em cada órgão ou entidade que orientem a realização de *cópias de segurança* periódicas das informações críticas dos ambientes dos sítios oficiais" (SOUSA, 2015, p. 4, *grifo nosso*).

Anos depois, o PL voltou às discussões nas comissões de trabalho da Câmara dos Deputados, desta vez, junto a *Comissão de Cultura* (CCULT): em outubro de 2017, o Projeto de Lei teve parecer pela rejeição proferido pelo Deputado Evandro Roman (PSD/PR), que em seu relatório justifica dizendo que a obrigatoriedade em manter a totalidade do conteúdo hospedado nos *websites* oficiais do governo "[...] traz uma grande dificuldade operacional, implicando em *gastos crescentes em tecnologias de armazenamento*, podendo tornar a preservação inviável" (ROMAN, 2017, p. 3, *grifo nosso*), segue dizendo que "[...] a

preservação de todo o conteúdo ignora o caráter dinâmico da rede mundial de computadores, que justamente *facilita a atualização e a dispersão de informações* com a maior brevidade para os interessados" (ROMAN, 2017, p. 3, *grifo nosso*). Os nossos grifos mostram que a apresentação do Projeto de Lei, assim como as relatorias são simplórias e desprovidas de qualquer estudo aprofundado sobre a relevância do tema, evidenciando, o que não é nenhuma novidade, a falta de conhecimento qualificado a respeito do arquivamento da *web* no Brasil.

Em março de 2019 foi designado novo relator do PL na CCULT, o Deputado David Miranda (PSOL), que em dezembro do mesmo ano apresentou parecer pela aprovação do projeto com substitutivo, em que acrescenta que "[...] são necessárias providências para que o conteúdo digital dos sítios oficiais não seja apagado à mercê de posicionamentos ideológicos de um candidato ou outro que vença as eleições" (MIRANDA, 2019, p. 2). O relator acrescentou em seu substitutivo, além dos *websites* institucionais já previstos, as redes sociais "[...] tais como *Youtube*, *Facebook*, *Twitter*, etc [...]" (2019, p. 5), além de incluir as contas pessoais em redes sociais de chefes dos Poderes Públicos e titulares de órgãos máximo dos Podes da União durante o exercício de seus mandatos, considerando que "[...] esses atores políticos são os principais porta-vozes de tais instituições" (MIRANDA, 2019, p. 5).

Não surpreende que no meio político no Brasil, o arquivamento da *web* não seja de conhecimento, ainda que superficial, por parte dos agentes políticos. Mas nos surpreende o Arquivo Nacional do Brasil, órgão responsável pela elaboração das políticas de gestão documental a nível nacional, ter excluído os documentos produzidos na *web* do escopo dos formatos de documentos aceitos para recolhimento na instituição, quando da apresentação e publicação da sua *Política de Preservação Digital*, com versões em 2012 e 2016: "Em momento futuro, outros tipos mais complexos de documentos em formato digital, como multimídia e páginas *web*, deverão ser também contemplados" (ARQUIVO NACIONAL, 2016, p. 11).

Recentemente, os *websites* governamentais no Brasil foram assunto quando da publicação do Decreto número 9.756/2019, que "Institui o portal único 'gov.br' e dispõe sobre as regras de unificação dos canais digitais do Governo Federal" (BRASIL, 2019b), que estabelece em seu artigo 1º que "[...] por meio do qual informações institucionais, notícias e serviços públicos prestados pelo Governo

Federal serão disponibilizados de maneira centralizada” (BRASIL, 2019b). O Portal Único GOV.BR foi lançado, oficialmente, em 01 de julho de 2019. Antes de seu lançamento, a primeira página do Portal (Figura 2) informava que “[...] os canais digitais do Governo Federal serão unificados”. O portal único “[...] vai reunir, em um só lugar, serviços para o cidadão e informações sobre a atuação de todas as áreas do governo”, e seguia dizendo que o portal também será “[...] a porta de entrada das páginas institucionais da administração federal, como ministérios, agências reguladoras e outros órgãos” (PORTAL ÚNICO GOV.BR, *website*, 2019).

Figura 2 - Print da página inicial do *website* gov.br

Fonte: www.gov.br

Com isso, percebe-se que a implantação de uma política para preservação de páginas *web* governamentais no Brasil é um grande desafio, seja para fins de convencimento da necessidade, quanto questões técnicas associadas a atividade. Cabe-nos, enquanto comunidade científica, apresentar estudos e possibilidades para arquivamento deste conteúdo informacional produzido exclusivamente na *web*.

2.5 CICLO DE VIDA DO ARQUIVAMENTO DA WEB

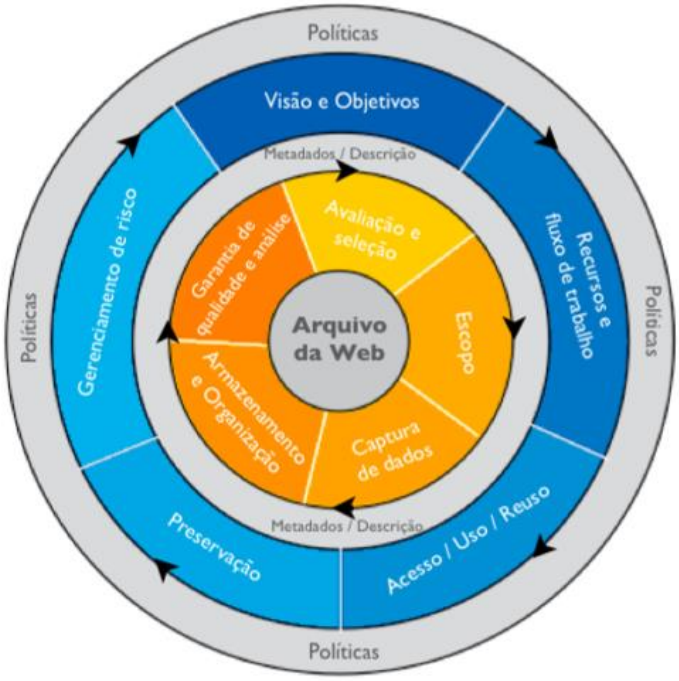
A rotina de preservação de *websites*, segundo Gomes (2010), é dividida em três etapas que envolve a recolha de informação proveniente da *web*, indexação e disponibilização de serviços de pesquisa e acesso, sendo que a primeira etapa se subdivide em coletar o arquivo, armazená-lo, extrair os endereços a partir dos *hiperlinks*, e inserir os endereços para a recolha. Com a intenção de especificar um pouco mais as fases, Niu (2012) diz que com a diversidade de recursos informacionais, o gerenciamento e o fluxo de trabalho do arquivamento da *web* passa a exigir o estabelecimento de passos, rotinas e técnicas que incluem procedimento referentes à avaliação e seleção, aquisição, organização e armazenamento, descrição e acesso, sendo este, basicamente, o fluxo de trabalho do núcleo do arquivamento da *web*. Da mesma forma, em 2013, um grupo de trabalho do *Archive-It*, serviço de arquivamento ligado ao *Internet Archive*, publicou um *White Paper* intitulado "*The web archiving life cycle model*", em que fala sobre o ciclo de vida do arquivamento da *web*. O modelo baseia-se nas experiências da equipe, bem como nas lições aprendidas de inúmeras instituições parceiras, incluindo estudos de caso detalhados de seis dessas instituições e trata-se de um esforço voltado com a intenção de representar fluxos de trabalho comuns e criar um modelo mensurável que possa ser utilizado como referência para organizações que queiram criar ou melhorar seus programas de arquivamento da *web* (BRAGG; HANNA, 2013, p. 2).

> O modelo é uma tentativa de incorporar os braços tecnológicos e programáticos do arquivamento da *web* em uma estrutura que será relevante para qualquer organização que deseje arquivar a *web*, independentemente do tamanho da organização, orçamento ou métodos técnicos de arquivamento da *web*. (BRAGG; HANNA, 2013, p. 28, *tradução nossa*).

O modelo apresenta uma forma de visualizar as diferentes etapas e fases que instituições enfrentam ao desenvolver e gerenciar um programa de arquivamento da *web*, tendo sua forma circular para sugerir a natureza repetitiva dos passos no ciclo de vida (BRAGG; HANNA, 2013, p. 2-3). Rockembach (2018a), ao descrever os casos internacionais e a situação brasileira sobre o arquivamento

da *web*, propôs uma tradução para o português do *Modelo de Ciclo de Vida do Arquivamento da Web*, o qual pode ser visto na Figura 3.

Figura 3 - Modelo de Ciclo de Vida do Arquivamento da *web*



Fonte: Bragg; Hanna (2013, p. 3[25] apud ROCKEMBACH, 2018a, p. 26).

Considerando que quase todas as rotinas do arquivamento da *web* envolvem algum tipo de decisão política, estes aspectos são representados pelo nível mais externo do ciclo de vida, a esfera política. Ao envolver todos os passos do ciclo de vida nesta esfera, o modelo representa a necessidade da constante elaboração de políticas nas rotinas de preservação destes conteúdos. Em uma segunda esfera, estão representados os metadados e a descrição: o serviço *Arquivo-It* escolheu incorporá-los como uma esfera completa, em vez de um segmento da esfera, para enfatizar que criar, importar e exportar metadados é um processo contínuo que ocorre em conjunto com outras atividades que compõem o ciclo de vida (BRAGG; HANNA, 2013).

O círculo azul dentro da banda de políticas representa as decisões de alto

[25] BRAGG, Molly; HANNA, Kristine; DONOVAN, Lori; HUKILL, Graham; PETERSON, Anna. The Web Archiving Life Cycle Model. WhitePaper. 2013. Disponível em: <http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf>. Acesso em: 14 fev. 2019.

nível que uma instituição enfrenta ao configurar e gerenciar seu programa de arquivamento da *web*; por sua vez, o círculo laranja descreve as tarefas do dia-a-dia envolvidas no negócio de arquivamento da *web* (BRAGG; HANNA, 2013, p. 3-4). O centro do modelo representa a própria coleção, o conteúdo da *web* arquivado, considerando que esses dados são o resultado final de todas as etapas anteriores, as quais passamos a descrever:

VISÃO E OBJETIVOS

O desejo em arquivar a *web* e como essa ação se relaciona com a missão da instituição são as questões iniciais que ajudarão a determinar a visão e os objetivos para o arquivamento da *web* proposto por uma instituição. Este passo no ciclo ocorre principalmente quando as instituições planejam seu programa de arquivamento da *web*. No entanto, as instituições tendem a revisitar e redefinir seus objetivos ao longo do programa quando há a mudança de algum recurso ou outras questões políticas que envolvam a preservação da *web*. (BRAGG; HANNA, 2013).

Diversas razões podem ser utilizadas para justificar a preservação de conteúdo *web*: algumas instituições acreditam que o conteúdo corre o risco de desaparecer e, portanto, precisa ser capturado e mantido acessível, como alguns casos de eventos como desastres naturais ou provocados pelo homem, revoltas políticas e memoriais de figuras públicas; já outras instituições optam por arquivar publicações específicas que estão disponíveis apenas em formatos digitais, tais como catálogos de cursos universitários, relatórios e publicações de agências estaduais ou locais; outras, ainda dispõe de instrumentos legais que condicionam o arquivamento de todos os registros oficiais produzidos pela instituição dentro de seu domínio *online*, construindo um registro histórico da presença na *web* daquela instituição. (BRAGG; HANNA, 2013, p. 5-6).

Além do conteúdo institucional de caráter oficial, pesquisadores, acadêmicos e organizações já estão reconhecendo a crescente influência dos sites de redes sociais e a importância de conceber uma forma de preservar este conteúdo, criando arquivo da *web* por tema, assunto ou tópico específico. Independentemente da

visão específica de cada programa de arquivamento da *web*, é importante estabelecer que a visão institucional molda muitas das outras políticas e decisões que terão de ser tomadas nas etapas posteriores do ciclo de vida do arquivamento da *web*. (BRAGG; HANNA, 2013, p. 6).

RECURSOS E FLUXO DE TRABALHO

A fase de recursos e fluxo de trabalho também pode ser abordada de diferentes maneiras, dependendo das escolhas que cada iniciativa de arquivamento da *web* faz. Porém, no contexto da esfera externa do *Modelo de Ciclo de Vida do Arquivamento da Web*, as instituições examinam os recursos e fluxos de trabalho que podem ser aproveitados para criar ou manter o programa de toda uma instituição “dessa forma, os recursos e o fluxo de trabalho podem ser considerados de maneira semelhante à "política", já que podem ser aplicados em várias áreas do modelo de ciclo de vida de arquivamento da *web*” (BRAGG; HANNA, 2013, p. 8, *tradução nossa*). Em termo gerais, os recursos e fluxo de trabalho também podem ser considerados como rotinas de gerenciamento e aplicadas a cada um dos elementos na esfera interna do modelo e, nesse contexto, recursos e fluxo de trabalho se tornam parte das atividades diárias do arquivamento da *web* (BRAGG; HANNA, 2013, p. 8).

ACESSO/USO/REUSO

O acesso aos arquivos da *web* depende de questões legais do país no qual o arquivo está hospedado, sendo esta definição vital ao programa: “Estabelecer políticas de acesso, uso e reutilização é vital para um programa bem-sucedido de arquivamento da *web*” (BRAGG; HANNA, 2013, p. 12, *tradução nossa*). Como exemplo, a legislação de depósito legal da Nova Zelândia, que permite que a Biblioteca Nacional da Nova Zelândia preserve quaisquer páginas disponíveis publicamente de um site do país e forneça acesso à sua cópia arquivada (NEW ZEALAND WEB ARCHIVE, *website*, 2019).

Nos EUA, a *Library of Congress* torna os registros bibliográficos de todos os sites arquivados publicamente acessíveis, porém fornece acesso público às páginas preservada cujos produtores deram permissão (GROTKE; JONES, 2010). Muitos arquivos da *web* são acessíveis apenas no local, como o caso dos arquivos da *web* da *Bibliothèque Nationale de France*[26]; o arquivo *web* finlandês[27]; dinamarquês[28]; *Web* Archive Norway[29], o *Web* Archive da Eslovénia[30], *Web* Archive da Suíça[31] e *Web* Archive da Áustria[32] (NIU, 2012). Alguns arquivos da *web* acessíveis ao público oferecem funcionalidade reduzida e acesso atrasado para evitar a concorrência com os proprietários de sites (MASANÈS, 2006). Por exemplo, há um atraso de pelo menos três meses entre o momento em que um site é coletado e quando ele será exibido no WAX da *Harvard University Library*[33]; e no caso da *IA Wayback Machine*, o atraso é de 6 a 12 meses (ARCHIVE-IT, 2011).

PRESERVAÇÃO

Embora muito se tenha aprendido com os arquivos, bibliotecas e museus, através de seus métodos de preservação, é pertinente salientar que a natureza e as qualidades da *web* exigem que repensemos e adaptemos essas práticas herdadas dessa longa tradição de preservar artefatos culturais físicos para as especificidades dos arquivos *web*. Niu (2012) diz que embora a preservação digital seja definitivamente uma etapa de suma importância, ela não é exclusiva do arquivamento da *web*, ou seja, não é diferente da preservação de outros recursos digitais. Para pensar a manutenção a longo prazo de conteúdo *web*, pesquisadores da área estudam, basicamente, os métodos já alcançados pelas técnicas aplicadas aos artefatos digitais, ainda que muitos delas tenham como base teórica os conceitos e métodos adotados para os artefatos físicos.

[26] https://www.bnf.fr/en
[27] https://digi.kansalliskirjasto.fi
[28] netarchive.dk
[29] https://www.nb.no/en/the-national-library-of-norway/
[30] http://arhiv.nuk.uni-lj.si/
[31] https://www.e-helvetica.nb.admin.ch/
[32] https://webarchiv.onb.ac.at/
[33] https://archive-it.org/organizations/935

O acesso atual aos conteúdos, quando isso for legalmente possível, pode ser fornecido por meio de sites próprios das iniciativas; já o acesso de longo prazo dependerá de iniciativas capazes de preservar o conteúdo da *web* que foi coletado, trazendo-os para o domínio da preservação digital (DAY, 2006). Da mesma forma, as autoras dizem que os dados reunidos para a preparação do modelo sugerem que a preservação ainda é uma questão em evolução para as instituições que arquivam a *web*, pois acompanha a evolução da natureza da preservação digital e o desenvolvimento de repositórios digitais como um todo (BRAGG; HANNA, 2013, p. 17). A equipe do *Archive-It* descobriu que seus parceiros tendem a adotar diferentes estratégias de preservação: algumas utilizam o *Internet Archive* para armazenamento e preservação de seus arquivos WARC e metadados associados; outras iniciativas recebem uma cópia ou baixam seus arquivos WARC diretamente dos servidores do *Internet Archive* para seu disco rígido; algumas instituições parceiras estão trabalhando para incorporar arquivos WARC em seus repositórios digitais locais, embora estes projetos ainda estejam em fase inicial (BRAGG; HANNA, 2013, p. 17).

Para Day (2006, p. 178), a preservação digital pode ser entendida como o conjunto de atividades necessárias que garantam que os objetos digitais permaneçam acessíveis pelo tempo que for necessário. Já Hedstrom (1998, p. 190, *tradução nossa*) diz que a preservação digital envolve “[...] o planejamento, a alocação de recursos e a aplicação de métodos e tecnologias de preservação para garantir que a informação digital de valor contínuo permaneça acessível e utilizável”. A definição sugere que os desafios da preservação digital sejam multifacetados, envolvendo diversas questões técnicas e organizacionais, embora a maioria das dificuldades esteja relacionada com a tecnologia. Smith (2003, p. 2, *tradução nossa*) descreve a preservação digital como uma “[...] série de ações que indivíduos e instituições tomam para garantir que um dado recurso seja acessível para uso em algum momento desconhecido”. Apesar da alta taxa de crescimento da produção de informação digital, a preservação a longo prazo destes documentos ainda está longe de ser uma tarefa simples, pois o problema está na rápida obsolescência das várias tecnologias das quais depende a informação digital (DAY, 2006).

Além das questões técnicas, há uma série de desafios relacionados à preservação a longo prazo de objetos digitais. O primeiro diz respeito às dificuldades

em garantir a autenticidade e a integridade dos objetos ao longo do tempo (DAY, 2006). A informação digital é relativamente fácil de manipular, o que significa que ela pode ser facilmente corrompida, deliberada ou acidentalmente (LYNCH, 1996). Os usuários de recursos digitais precisam ter confiança na autenticidade de objetos preservados, isto é, que eles são o que afirmam ser e que sua integridade não foi comprometida (DAY, 2006). Existem métodos técnicos disponíveis para lidar com esse problema no nível de *bit* (por exemplo, técnicas criptográficas), mas a confiança na autenticidade de um objeto será baseada no nível de segurança que um usuário tem do repositório responsável por manter o objeto digital (DAY, 2006, p. 180).

GERENCIAMENTO DE RISCO

O gerenciamento de risco diz respeito aos direitos autorais sob o conteúdo do site que está sendo arquivado. Ao desenvolver um programa de arquivamento da *web* as instituições devem considerar o nível de risco que estão dispostos a aceitar e como administra-lo (BRAGG; HANNA, 2013). Um dos exemplos mais lúcidos da formulação de políticas de gerenciamento de riscos é SE e COMO as instituições decidem obter permissão dos proprietários do site antes de arquivá-los (BRAGG; HANNA, 2013, p. 18).

As decisões de gerenciamento de risco também podem aparecer na escolha de quais sites serão arquivados; exemplo disso foi a decisão da *State Library of North Carolina* e do *North Carolina State Archives* que optaram por coletar apenas sites de agências estatais. Outras organizações, por serem arquivos ou bibliotecas, entendem que é inerente de sua atuação preservar conteúdo que está disponível publicamente na Internet e, por essa razão, não demandariam autorização prévia. (BRAGG; HANNA, 2013).

Outro exemplo trazido pelas autoras é o caso da *Creighton University*, que retirou o acesso de um *website* preservado em seu arquivo *web* em que eram veiculadas imagens feitas por um fotógrafo que não havia autorizado a publicação. Após o acontecimento, a *Creighton University* decidiu que, se houver risco de

constrangimento ou litígio, o conteúdo do arquivo da *web* será removido. Afinal, “O risco pode ser gerenciado e mitigado preventivamente, e às vezes as instituições necessitarão mediar os possíveis problemas que surgem após o arquivamento do conteúdo” (BRAGG; HANNA, 2013, p. 19, *tradução nossa*).

AVALIAÇÃO E SELEÇÃO

A fase de avaliação e seleção do arquivamento da *web* envolve a escolha de sites para salvaguarda (BRAGG; HANNA, 2013, p. 22). Para o arquivamento da *web* é, essencialmente, um processo de seleção em que são escolhidos os *websites* que serão preservados, com base em um ou mais critérios (NIU, 2012). O desenvolvimento da política de seleção também definirá a forma de coleta dos *websites* (BRAGG; HANNA, 2013).

Day (2003a) classifica três abordagens de coleta de conteúdo para composição de arquivos *web*: coleta automática, coleta seletiva, por depósito ou uma combinação destas abordagens. A coleta automática geralmente utiliza tecnologias de rastreador da *web*, semelhantes aos usados pelos serviços de busca *online*, que acompanham *links* e baixam conteúdos de acordo com regras de coleta específicas. A coleta seletiva é baseada na escolha individual de *websites* para inclusão em um arquivo de acordo com as diretrizes de seleção pré-definidas e com os direitos negociados com seus proprietários, quando houver necessidade. A abordagem de depósito é baseada na decisão de proprietários ou administradores de *websites* que depositam uma cópia de seu site em um repositório. Por fim, a combinação dos três métodos também poderá ser uma opção na formulação de estratégias para coleta de *websites* que comporão o arquivo *web*.

Uma das primeiras questões que as iniciativas de arquivamento da *web* enfrentam ao planejar o arquivamento dos *websites* é o que escolher e coletar da enorme variedade de conteúdos disponíveis (PHILLIPS, 2005). Day (2003b) alega que é impossível arquivar todo o domínio da *web* devido ao seu tamanho e crescimento rápido. Por outro lado, pesquisadores da Universidade de Lisboa, argumentam que o enorme tamanho da *web* impede a implementação de uma

política rigorosa de seleção de conteúdo, e recomendam, portanto, a coleta automática sem intervenção humana, a exemplo do *Internet Archive*, que decidiu preservar sites não seletivamente (GOMES; FREITAS; SILVA, 2006).

Pesquisadoras israelenses desenvolveram um trabalho que identificou as políticas de seleção de conteúdo que bibliotecas nacionais utilizam. As autoras identificaram quatro abordagens de coleta: *Abordagem extensiva* que coleta automaticamente todo o conteúdo nacional publicado na *web*; *Abordagem seletiva* em que o conteúdo é selecionado por bibliotecários ou por equipes especializadas, considerando critérios pré-definidos de qualidade ou que correspondam à política de aquisição da biblioteca; *Abordagem baseada em assuntos e eventos* que identifica temáticas de interesse da biblioteca, que passa a coletá-los automaticamente ou por curadores humanos; e a *Abordagem combinada* que efetiva a coleta automática de todo o conteúdo nacional publicado na *web*, juntamente com a criação de coleções de assuntos ou eventos de interesse das bibliotecas. (SHVEIKY; BAR-ILAN, 2013, p. 40-41).

ESCOPO

Depois de escolhidos os *websites* que serão preservados, deve-se decidir se o arquivamento será de todo o conteúdo do site ou de partes da página, mesmo que essa definição seja tomada somente depois da captura do conteúdo, como parte da revisão da qualidade da coleta (BRAGG; HANNA, 2013, p. 23-24). Outra maneira útil de classificar os arquivos da *web* é considerando o escopo que eles adotam. Os arquivos da *web* podem ser centrados no *site*, no *tópico* ou no *domínio*, segundo Masanès (2006, p. 42) e Khan e Rahman (2019, p. 73):

Arquivamento centrado no site é aquele que promove o arquivamento de um website específico, usado, especialmente, para arquivo da *web* privados (KHAN; RAHMAN, 2019, p. 73).

O *Arquivamento Centrado em Tópicos,* tem se tornado cada vez mais popular e impulsionado por necessidades diretas de pesquisa, em que pesquisadores têm notado a natureza efêmera de páginas da Internet quando não conseguem efetivar

alguma verificação científica, em razão de algum site estar indisponível, especialmente quando a informação desejada estava presente somente neste endereço, tornado impossível a verificação (KHAN; RAHMAN, 2019). Outros projetos centrados em tópicos foram realizados com escopo para arquivamento de sites eleitorais, como o projeto Minerva da *Library of Congress* (SCHNEIDER et, al., 2003) ou o arquivo *web* das eleições francesas feito pela *Bibliothèque Nationale de France* (MASANÈS, 2005).

O arquivamento centralizado no domínio refere-se a escolha de uma determinada extensão de *URL*. O domínio pode representar uma localidade, tipo de rede ou domínios genéricos. No entanto, é possível distinguir alguns tipos funcionais (como .com e .edu) e tipos geográficos (.ch e .jp). Os domínios de nível superior geográficos geralmente têm subdivisões funcionais (como gov.br e gob.mex). Uma vantagem deste método é que o arquivamento pode acontecer detectando automaticamente sites específicos do domínio escolhido.

CAPTURA DE DADOS

Uma vez que as instituições tenham escolhido quais e quantos sites serão capturados, é que as ações com o *software* de rastreamento são colocadas em funcionamento. Neste processo são definidas a frequência e o tempo em que os rastreamentos ocorrerão. Dada a complexidade de algumas páginas da *web*, a etapa de captura de dados do arquivamento da *web* pode gerar diversas surpresas, como por exemplo, um site pode ser muito maior do que o previsto e, portanto, esgotar os recursos de armazenamento (MASANÈS, 2006); ou uma página pode conter acessos restritos.

Segundo Khan e Rahman (2019, p. 76-77) a escolha de uma técnica de captura viável depende, inicialmente, dos recursos a serem capturados e segundo, a frequência desta tarefa. Para isso existem três técnicas para captura de recursos da *web*: uso de navegadores, uso de um sistema de autoria, e o uso de rastreadores da *web*, detalhadas na sessão *2.6,* tópico *Como capturar os recursos*.

GARANTIA DE QUALIDADE E ANÁLISE

Depois que é realizada a captura dos *websites* desejados, os dados recolhidos são analisados e sua qualidade e integridade são avaliadas por meio de relatórios gerados pelos rastreadores ou nos próprios sites arquivados por meio de uma ferramenta de acesso, como o *software Wayback*, do *Internet Archive* (BRAGG; HANNA, 2013, p. 26). O arquivamento da *web* tem como objetivo capturar e apresentar recursos o mais próximo possível de como eles apareceram nos *websites* ainda acessíveis, porém limitações técnicas e legais na coleta da *web* mostram que uma cópia perfeita de um site raramente é alcançada (SHALLCROSS, 2013; MASANÈS, 2006) e medidas de garantia de qualidade são necessárias para definir o sucesso de uma cópia coletada (BINGHAM, 2014, p. 52). Além disso, a garantia de qualidade também permite que a instituição coletora verifique se sua política de desenvolvimento de coleções esteja sendo atendida, assim como assegura que os recursos estejam adequados para preservação e uso a longo prazo (BROWN, 2006). Os indicadores de qualidade devem ser definidos pelas instituições com base em sua política e objetivos de desenvolvimento de coleções, embora estejam sendo tomadas medidas na comunidade internacional de arquivamento da *web* para abordar essa questão de maneira mais genérica (BINGHAM, 2014).

ARMAZENAMENTO E ORGANIZAÇÃO

Os arquivos da *web* precisam preservar a autenticidade e a integridade do conteúdo arquivado (NIU, 2012), ainda que os requisitos de autenticidade e integridade variem de acordo com o objetivo da coleta. Em alguns cenários, preservar apenas o conteúdo intelectual é suficiente; em outros, como na preservação de evidências legais, a estrutura e o contexto dos recursos também podem precisar ser preservados (NIU, 2012).

Masanès (2006) apresenta três abordagens para organizar e armazenar conteúdo da *web* arquivado: *sistemas de arquivos locais*, *arquivos baseados na web* e *arquivos não baseados na web*. Todas as três abordagens preservam o conteúdo

intelectual das páginas, mas variam no grau de preservação do contexto e da estrutura (NIU, 2012). Em um arquivo da *web* que usa um *sistema de arquivos local*, o navegador pode percorrer o sistema de arquivos da mesma forma que navega na *web* (MASANÈS, 2006). Em um arquivamento *baseado na web*, as páginas e os metadados associados são agrupados e armazenados em pacotes de arquivos e os *links* e *URLs* originais são preservados (MASANÈS, 2006). Esta segunda abordagem preserva a autenticidade ao maior grau (NIU, 2012). A abordagem de arquivamento *não baseado na web* extrai documentos da *web* do contexto de *hipertexto* e reorganiza-os em um modo de acesso baseado em catálogo ou os transforma em arquivos PDF (MASANÈS, 2006). Essa abordagem preserva a autenticidade e a integridade ao menor grau (NIU, 2012).

2.6 ABORDAGEM SISTEMÁTICA PARA A PRESERVAÇÃO DA WEB

Em 2019 foi apresentado à comunidade científica uma pesquisa realizada pelos paquistaneses Muzammil Khan, da *Preston University Islamabad* e Arif Ur Rahman, da *University of Bozen-Bolzano* em que os autores desenvolveram o estudo institulado *A Systematic Approach Towards Web Preservation*, ou, em português, *Uma abordagem sistemática para a preservação da web*, publicado pela revista *Information Technology and Libraries*[34]. O principal objetivo do artigo foi dividir o processo de preservação da *web* em estágios auto-explicativos e projetar um esquema sistematizado deste fluxo, apresentando um modelo que poderá ser utilizado por qualquer iniciativa que pretende promover o arquivamento da *web* (KRAN; RAHMAN, 2019). Nesta sessão, apoiados no estudo referenciado, descrevemos passo a passo deste processo.

Diferentes aspectos precisam ser observados durante o processo de preservação e arquivamento da *web*: gestão dos objetos digitais, formato e armazenamento do objeto digital, gestão do arquivamento, questões administrativas, acesso e segurança ao arquivo, planejamento de preservação (KRAN; RAHMAN, 2019, p. 72), e demais questões que precisam ser entendidas

[34] https://ejournals.bc.edu/index.php/ital/about

para uma eficaz gestão do arquivo *web* e que ajudarão a tratar os desafios que ocorrem durante o processo de preservação. “O modelo de referência *Open Archival Information System* (OAIS) é uma tentativa de fornecer uma estrutura de alto nível para o desenvolvimento e comparação de arquivos digitais” (KRAN; RAHMAN, 2019, p. 72, *tradução nossa*).

Importante destacar que a vida útil da mídia de armazenamento digital pode ser surpreendentemente curta, e a rápida evolução das tecnologias de renderização pode impedir o acesso futuro (LAVOIE, 2000). Neste sentido, o Modelo OAIS se propõe a responder o que é necessário para preservar e manter o acesso às informações digitais a longo prazo. Ainda que seja uma questão de difícil solução, o modelo de referência OAIS apresenta

> [...] uma estrutura conceitual para um sistema de arquivo dedicado a preservar e manter o acesso às informações digitais a longo prazo. O objetivo do modelo de referência é aumentar a conscientização e a compreensão de conceitos relevantes para o arquivamento de objetos digitais, especialmente entre instituições não-arquivísticas; elucidar terminologia e conceitos para descrever e comparar modelos de dados e arquiteturas de arquivo; expandir o consenso sobre os elementos e processos endêmicos da preservação e acesso às informações digitais; e criar uma estrutura para orientar a identificação e o desenvolvimento de padrões. (LAVOIE, 2000, p. 27, *tradução nossa*).

Embora o modelo OAIS seja suficientemente amplo para abranger arquivos de objetos físicos e digitais, é no contexto do documento digital que o modelo foi bem recebido. Publicado inicialmente em 2002, na forma de um padrão recomendado, o OAIS foi aprimorado e, atualmente, está representado na norma *International Organization for Standardization* (ISO) 14721: 2012.

> Talvez a conquista mais importante do modelo de referência da OAIS até hoje seja que ele se tornou quase universalmente aceito como a língua franca da preservação digital, moldando e sustentando conversas sobre preservação digital em domínios diferentes e fornecendo um mapeamento geral da paisagem que administra nossa herança digital deve navegar para garantir a disponibilidade a longo prazo de materiais digitais. [...]. Parece razoável concluir que o OAIS se tornou um recurso fundamental para entender a preservação digital, uma linguagem para falar sobre questões de preservação digital e um ponto de partida para a implementação de soluções de preservação digital. (LAVOIE, 2014, *tradução nossa*).

Considerando o modelo OAIS, os autores dizem que “Este estudo tem como objetivo projetar uma abordagem sistemática passo a passo para a preservação da *web* que ajuda a compreender os desafios das atividades de preservação ou arquivamento [...]” (KRAN; RAHMAN, 2019, p. 72, *tradução nossa*). A abordagem sistemática é uma maneira acessível de analisar, projetar, implementar e avaliar o arquivo com clareza e com diferentes opções para um processo de preservação e desenvolvimento eficaz de arquivos. Um efetivo processo de preservação é aquele que leva a um arquivo da *web* bem organizado, facilmente gerenciado e cumpre os requisitos comunitários designados, que ajudam a resolver os desafios que confrontam arquivistas e analistas durante as atividades de preservação (KRAN; RAHMAN, 2019, p. 72).

Para descrever passo a passo da abordagem sistemática, são necessários dois alinhamentos conceituais: 1) nos apoiaremos no conceito de preservação digital como sendo “[...] o conjunto de processos e atividades que garantem armazenamento sustentado a longo prazo, acesso e interpretação da informação digital” (FARRELL; ASHLEY; DAVIS, 2010, *tradução nossa*); e 2) o *layout* geral da *web* varia de domínio para domínio com base no tipo de informação e sua apresentação, neste sentido, os *websites* podem ser classificados com base no tipo de informação (ou seja, seu conteúdo) e na forma como essas informações são apresentadas (ou seja, o *layout* ou a estrutura da página) (KRAN; RAHMAN, 2019, p. 73).

As seções a seguir explicam as atividades de preservação da *web* e as possibilidades de implementação segundo a abordagem sistemática proposta.

DEFININDO O ESCOPO DO ARQUIVO DA WEB

A *web* oferece a oportunidade de compartilhar informações usando serviços, como blogs, sites de redes sociais, comércio eletrônico, *wikis* e bibliotecas eletrônicas. Esses sites fornecem informações sobre diversos tópicos e abordam

diferentes comunidades com base em seus interesses e necessidades. Da mesma forma, existem diferentes maneiras com as quais são tratadas e apresentadas as informações na *web*. Além disso, seu *layout* geral muda de um domínio para outro, o que torna impraticável desenvolver um sistema único para preservar todos os tipos de sites a longo prazo. Portanto, antes de começar a preservar a *web*, o arquivista deve definir o escopo do que será arquivado. (KRAN; RAHMAN, 2019, p. 73).

O escopo de um arquivo da *web* é determinado pela definição do objeto digital a ser coletado – a página da *web*. Se trata de uma questão complicada, pois do ponto de vista de um usuário, uma página da *web* é a imagem exibida ao colocar um endereço de *URL* em um navegador. Ainda que essa definição operacional seja necessária, não é suficiente, pois um arquivo também deve garantir que o documento seja traduzido de maneira autêntica, ou seja, fazendo com que seja incluído o contexto, evocando a experiência do documento original (a página ao vivo). (LYMAN, 2002).

O arquivo será centrado no *site*, no *tópico* ou no *domínio* (KRAN; RAHMAN, 2019, p. 73; MASANÈS, 2006, p. 42). A sessão *2.5,* tópico *Escopo*, desta dissertação, descreve cada uma das abordagens.

COMPREENDENDO A ESTRUTURA DA WEB

Após definir o escopo do arquivo da *web* pretendido, o arquivista terá uma melhor compreensão do interesse e das consultas esperadas da comunidade pretendida, com base nos recursos disponíveis ou nas informações fornecidas pelo domínio selecionado. O foco nesta etapa é entender o tipo de informação e conteúdo fornecido pelo domínio selecionado e como as informações foram apresentadas. (KRAN; RAHMAN, 2019, p. 73-74).

A *web* pode ser entendida por duas dimensões: a primeira considera a *web* como um meio de comunicação de conteúdo usando protocolos, como HTTP, e a segunda considera a *web* como um *contêiner*, que apresenta o conteúdo aos usuários e não apenas o conteúdo puro, por exemplo, a tecnologia usada para exibir

o conteúdo em códigos binários. A equipe de preservação deve entender estes parâmetros tais como as questões técnicas, as futuras tecnologias e eventuais inclusões esperada de outros conteúdos relacionados. (KRAN; RAHMAN, 2019, p. 75).

IDENTIFICAÇÃO DOS RECURSOS DA WEB

O arquivista deve entender o conteúdo e a representação do domínio selecionado: se blogs, sites de redes sociais, institucionais, educacionais, de notícias ou de entretenimento. Todos esses sites fornecem informações diferentes e abordam diferentes comunidades que possuem necessidades informacionais distintas (KRAN; RAHMAN, 2019, p. 75). Uma página da *web* é a combinação de dois itens: o conteúdo e a estrutura (FARRELL; ASHLEY; DAVIS, 2010). Os recursos que podem ser preservados são os seguintes:

Conteúdo da Web, divididos em três categorias:

1) *Conteúdo textual (texto sem formatação)*: esta categoria descreve as informações textuais que aparecem em uma página da *web* e não inclui *links*, comportamentos e formatação do estilo de apresentação.
2) *Conteúdo visual (imagens)*: esse conteúdo é a forma visual de informação ou é um material complementar às informações fornecidas na forma textual.
3) *Conteúdo multimídia*: o conteúdo multimídia inclui principalmente áudio e vídeo, e também pode incluir animação ou mesmo texto como parte de um vídeo ou uma combinação de texto, áudio e vídeo.

Estrutura da Web, dividida em duas categorias:

1) *Aparência (layout da web ou apresentação)*: esta categoria indica o *layout* geral, a apresentação e a aparência da página da *web*.
2) *Comportamento (navegação de código)*: caracterizado por navegação a partir de *links*, podendo direcionar para o mesmo *website*, diferentes *websites* ou para documentos externos ou recursos dinâmicos e animados, como *feed*

> de uma rede social, comentários, marcações ou favoritos. (KRAN; RAHMAN, 2019, p. 75).

IDENTIFICAR A COMUNIDADE DESIGNADA

Comunidade designada refere-se ao conjunto de usuários em potencial, ou seja, aqueles que podem acessar o conteúdo e as informações arquivadas que poderão não estar mais disponíveis em circunstâncias normais ou no site ao vivo. Os promotores da iniciativa de arquivamento da *web* devem identificar os usuários em potencial do arquivo da *web*, seus requisitos funcionais e consultas esperadas, analisando-os cuidadosamente. (KRAN; RAHMAN, 2019, p. 75).

Segundo Martins (2019, p. 88) “ [...] os públicos exercem poder e influência nas decisões que configuram e gerenciam os programas de arquivamento na *web* [...]”. A autora diz, também, que quando os usuários são categorizados de uma maneira lógica, a partir das funções e responsabilidades da estrutura da organização promotora da iniciativa de arquivamento da *web* “[...] os públicos podem facilitar as tomadas de decisão [...]” (MARTINS, 2019, p. 88). Além da contribuição em relação ao mapeamento e identificação de públicos, a pesquisa deixa evidente a interdisciplinaridade do arquivamento da *web*, quando há a inserção da expertise da área da Comunicação nesta fase do modelo de abordagem sistêmica.

PRIORIZAR OS RECURSOS DA WEB

A complexidade dos recursos da *web* e suas representações causam complicações no processo de preservação digital, pois pode ser indesejável ou inviável preservar todos os recursos utilizados no *website*. Portanto, estabelecer prioridade é uma fase relevante e requer atenção em dois fatores: o *primeiro* diz respeito à reutilização potencial do recurso; e o *segundo* considera a frequência com a qual o recurso será acessado. Recursos que não acrescentam valor relevante

ou aqueles gerenciados em outros locais podem ser excluídos. Os autores sugerem a aplicação do método MoSCoW para estabelecer as prioridades dos recursos. (KRAN; RAHMAN, 2019, p. 75-76).

O Método MoSCoW foi desenvolvido por Dai Clegg, da *Oracle UK*[35] em 1994, e foi popularizado por expoentes do *Dynamic Systems Development Method* (DSDM)[36]. Trata-se de uma técnica de priorização utilizada em gerenciamento e análise de negócios, gerenciamento de projetos e desenvolvimento de *softwares* para alcançar um entendimento comum com as partes interessadas sobre a importância que elas atribuem à entrega de cada requisito; também é conhecido como priorização do MoSCoW ou análise do MoSCoW. (HAUGHEY, 2014). O termo é um acrônimo derivado da primeira letra de cada uma das quatro categorias de priorização: *Must have (Deve ter)*, *Should have (Deveria ter)*, *Could have (Poderia ter)*, e *Won't have (Não terá)*.

COMO CAPTURAR OS RECURSOS

Existem três diferentes técnicas que podem ser utilizadas para capturar páginas da *web*: por navegador, por rastreador da *web* ou por sistema de autoria (FARRELL; ASHLEY; DAVIS, 2010, p. 25). Para a escolha de qual técnica de captura pode ser utilizada é necessário analisar os recursos que serão capturados e a frequência da realização destas capturas (KRAN; RAHMAN, 2019, p. 76). Cada técnica de captura tem vantagens e desvantagens associadas:

[35] https://www.oracle.com/uk/index.html

[36] O DSDM é um método Agile (www.agilebusiness.org) que se concentra no ciclo de vida completo do projeto. O DSDM (formalmente conhecido como *Dynamic System Development Method*) foi criado em 1994, depois que os gerentes de projeto que usam o RAD (*Rapid Application Development*) buscavam mais governança e disciplina para essa nova maneira iterativa de trabalhar. (AGILE BUSINESS, website, 2019). Disponível em: https://www.agilebusiness.org/page/whatisdsdm. Acesso em: 17 dez. 2019.

Captura da web usando navegadores

Capturar conteúdo estático é uma das principais características desta técnica da captura da *web*, o que poderá ser uma vantagem ou não, dependendo do foco da coleta que se pretende; essa abordagem geralmente preservava o conteúdo na forma de imagens, recuperando, basicamente, os recursos que são visíveis para os usuários; é melhor para sites bem organizados; e estão disponíveis ferramentas comerciais para capturar a rede. Além disso, o conteúdo, por ser estático, é tratado como se fossem publicações; perde a estrutura da *web*, como aparência, comportamento e outros atributos característicos da interatividade de páginas da *web*. (KRAN; RAHMAN, 2019).

Captura da web usando rastreadores

Os rastreadores da *web* são talvez a técnica mais usada para captura automatizada de conteúdo da *web*, sendo que sua principal vantagem é que eles extraem conteúdo incorporado no *website*, ainda que não esteja no nível de visualização em tela. Além disso, são amplamente utilizados e, por essa razão, dispõe de comunidades de usuários; captura conteúdo específico ou de todo o *website*; e bloqueia acessos a *links* restritos. As principais desvantagens é que para o uso destas ferramentas é necessário conhecimento e experiência em ferramentas de tecnologias e linguagem de desenvolvimento, e o rastreador pode não capturar tudo que deveria e, às vezes, capturar muito conteúdo. (KRAN; RAHMAN, 2019).

Captura da web usando um sistema/servidor de autoria

Para captura da *web* a partir de um sistema de autoria é necessário ter conhecimento de programação e acesso à infraestrutura de desenvolvimento do *website*. A técnica de captura do sistema de autoria é usada para capturar a *web* diretamente do servidor de hospedagem da página. Todo o conteúdo, seja ele texto, imagens e código fonte são coletados. O sistema de autoria permite preservar diferentes versões do *website*. (KRAN; RAHMAN, 2019).

Algumas vantagens do sistema de autoria é a captura de todo conteúdo disponível e a facilidade em executa-lo, se tiver permissão de acesso adequada ou

ser proprietário do servidor onde está hospedado o *website*. As desvantagens da captura da *web* usando o sistema de autoria é a captura de todas as informações brutas disponíveis (inclusive código fonte); depende da infraestrutura de autoria ou do sistema de gerenciamento de conteúdo e não é viável para acesso de organizações externas. (KRAN; RAHMAN, 2019).

POLÍTICA DE SELEÇÃO DE CONTEÚDO DA WEB

A partir desta fase, a abordagem direciona as demais etapas ao conteúdo a ser coletado, preparando e filtrando-o para uma seleção viável com base naquilo que será arquivado segundo a política de seleção (KRAN, RAHMAN, 2019) que ajudará a determinar e elucidar quais os conteúdos do *website* precisam ser capturados considerando as prioridades, a finalidade e o escopo do conteúdo da *web* previamente definidos (UDAPURE; KALE; DHARMIK, 2014). A decisão da política inclui a descrição do contexto, os usuários pretendidos, o acesso a mecanismos e os usos esperados do arquivo *web* e está dividida em 1) Processo de seleção e, 2) Abordagem de seleção (KRAN, RAHMAN, 2019).

1) O ***processo de seleção*** apresenta uma abordagem qualitativa dos conteúdos selecionados da *web* ao transitar pelas fases de *preparação*, *descoberta* e *filtragem*. Sendo que a fase de **preparação** tem como objetivo determinar os procedimentos de infraestrutura, tais como o espaço de armazenamento, a técnica de captura, ferramentas de captura, categorização de extensão, nível de granularidade e a frequência do arquivamento. O principal objetivo da fase de **descoberta** é determinar a fonte de informação que será arquivada, ou seja, o *website*. Essa definição pode ser alcançada de duas maneiras: primeiro, criando um ponto de entrada manualmente que será usado para determinar a lista dos *links* de entrada, também chamados de *seeds* (ou sementes); ou, segundo, a lista de pontos de entrada é criada automaticamente a partir de uma única semente. A **filtragem** objetiva otimizar e tornar conciso o conteúdo *web* capturado, fazendo com que sejam coletados os conteúdos propriamente ditos, removendo dados indesejados ou duplicados. Normalmente, para preservação, uma filtragem automática é usada

como método, ainda que se tenha a possibilidade de filtragem manual caso os robôs ou ferramentas automáticas não puderem interpretar a *web*. A fase de descoberta e filtragem podem ser combinadas. (KRAN, RAHMAN, 2019).

2) A **abordagem de seleção** pode ser automática ou manual. A seleção manual de conteúdo não é muito utilizada dada sua complexidade e necessidade do uso de ferramentas automáticas para encontrar o conteúdo e para realizar a revisão da coleção capturada (KRAN, RAHMAN, 2019). Já a seleção automática é amplamente utilizada para promover o arquivamento da *web* (DAY, 2003a). Os autores dizem que “a seleção da abordagem de coleta depende da frequência com que o conteúdo *web* será arquivado” (KRAN, RAHMAN, 2019, p. 80, *tradução nossa*) e consideram quatro abordagens de seleção de conteúdo *web*:

Abordagem Não Seletiva

Também chamada de coleta automática (DAY, 2003a), implica em coletar todo o conteúdo possível de um conjunto de *websites* (KRAN, RAHMAN, 2019). A abordagem não seletiva é usada em uma situação em que o escopo define a coleta de todo domínio de determinados *websites*, tornando-se uma técnica mais simplificada de coleta, produzindo uma imagem abrangente da *web*; por outro lado, sua desvantagem é o fato de gerar enormes quantidades de conteúdo não classificados, duplicados, e que consomem mais recursos do que o necessário para armazenamento e recuperação (KRAN, RAHMAN, 2019).

Abordagem Seletiva

A abordagem seletiva foi adotada inicialmente pela Biblioteca Nacional da Austrália, em 1997. Nela o *website* é incluído para arquivamento com base em estratégias, no acesso e nas informações fornecidas pelo arquivo *web*. A decisão de coleta pode ser tomada em um dos seguintes níveis: **Nível do site**, que define quais sites devem ser incluídos considerando um domínio selecionado, por exemplo, arquivar todos sites educacionais do domínio .gov.br*;* **Nível da página da *web*** em que são definidas quais páginas de determinados *websites* devem ser coletados, por exemplo, arquivar as páginas iniciais de todos os sites com extensão .gov.br; e **Nível de conteúdo da *web*** em que o tipo de conteúdo é definido para

arquivamento, por exemplo, arquivar todos as imagens das páginas iniciais dos sites com extensão .gov.br. (KRAN, RAHMAN, 2019).

Abordagem de Depósito

Na abordagem de coleta de depósitos, o pacote de informações é enviado pelo administrador ou proprietário do *website*, que inclui uma cópia da página com arquivos relacionados que podem ser acessados através de *hiperlinks*. O pacote de informações de arquivo é aplicável a pequenas coleções ou o proprietário do *website* pode iniciar o projeto de preservação, por exemplo uma empresa que deseja preservar seu conteúdo *web*. A abordagem de coleta de depósitos foi adotada pela *National Archives and Records Administration* (NARA) para a coleta de *websites* de agências federais americanas em 2001. (KRAN, RAHMAN, 2019).

Abordagem Combinada

Há um contínuo debate sobre qual abordagem é a melhor para cada situação, as vantagens e desvantagens associadas a cada abordagem de coleta precisam ser ponderadas até que se entenda, a partir das necessidades e políticas de cada arquivo *web*, qual melhor atende as necessidades de cada arquivo *web*. Por exemplo, a abordagem de depósito deve ser um acordo econômico aos depositantes, ainda que a combinação de coleta automática e abordagens seletivas sejam mais econômicas em comparação com outras abordagens de seleção. O uso simultâneo de mais de uma abordagem pode ser a opção para o arquivo *web* que pretende colocar, por exemplo, extensões completas de sites governamentais e, ao mesmo tempo, estabelecer coleções de determinados assuntos relevantes ao país ou organização. Esta abordagem combinada foi utilizada pela *Bibliotheque Nationale de France* (BnF) em 2006. (KRAN, RAHMAN, 2019).

IDENTIFICAÇÃO DE METADADOS

Metadados são informações estruturadas que descrevem, localizam, gerenciam, recuperam e acessam recursos de informação digital e são geralmente chamados de “dados sobre dados” ou “informações sobre informações”, mas pode

ser mais útil e informativo descrever esses dados como “documentação descritiva e técnica” (NISO, 2017). Os metadados são entendidos como os dados que permitem a recriação e interpretação da estrutura e conteúdo das informações digitais ao longo do tempo (LUDÄSCHER; MARCIANO; MOORE, 2001). Entendido dessa maneira, fica evidente que esses metadados precisam suportar uma gama de diferentes funções, incluindo acesso, registro de contextos e proveniência de objetos, documentação de ações e políticas de repositório, por exemplo (MASANÈS, 2006, p. 189). Conceitualmente, portanto, os metadados de preservação abrangem a divisão tradicional de metadados em categorias descritivas, estruturais e administrativas (NISO, 2017; MASANÈS, 2006).

Os **metadados descritivos** descrevem um recurso para fins de descoberta e identificação. Pode consistir em elementos para um documento como título, autor(es), resumo e palavras-chave. *Os* **metadados estruturais** descrevem como os objetos compostos são reunidos, por exemplo, como seções são ordenadas para formar capítulos. Já os **metadados administrativos** fornecem informações para facilitar o gerenciamento de recursos: quando e como um arquivo foi criado, quem pode acessar o arquivo, o tipo de arquivo e outras informações técnicas. Os metadados administrativos são classificados em dois tipos: 1) gerenciamento de direitos onde são descritas informações dos direitos de propriedade intelectual; e 2) os metadados de preservação contêm informações necessárias para arquivar e preservar um recurso. (NISO, 2017).

Os metadados desempenham um papel vital na preservação a longo prazo de objetos digitais e são importantes para identificar os metadados que podem ajudar a recuperar um objeto específico do arquivo após a preservação (KRAN, RAHMAN, 2019). Neste sentido, a preservação e o arquivamento de objetos digitais requerem padrões de metadados para rastrear e garantir o acesso a esses objetos (KRAN, RAHMAN, 2019). Exemplos de padrões de metadados: *A Dublin Core Metadata Iniciative* (DCMI)[37]; *Metadata Encoding and Transmission Standard* (METS)[38]; *Metadata Object Description Schema* (MODS)[39]; *Visual Resources*

[37] http://dublincore.org/
[38] http://www.loc.gov/standards/mets/
[39] http://www.loc.gov/standards/mods/

Association Core Strategies (VRA Core)[40] e; *Preservation Metadata Implementation Strategies* (PREMIS)[41].

FORMATO DE ARQUIVO

Considerando todas as estratégias de seleção do conteúdo *web*, os arquivos *web* podem apresentar uma ampla variedade de formatos de arquivos, sejam páginas de conteúdo *online* textual, imagens, vídeos, ou arquivos em PDF e outros formatos. Para preservar esse conteúdo a iniciativa de arquivo da *web* usa diferentes formatos de armazenamento contendo metadados e utiliza técnicas de compactação destes dados (KRAN, RAHMAN, 2019). O *Internet Archive* definiu o formato ARC[42] para salvaguardar os arquivos de *websites*, posteriormente, em 2009, a *Internet Organization for Standardization* (ISO) estabeleceu o formato WARC[43] como um padrão oficial para arquivos *web* (GOMES; MIRANDA; COSTA, 2011). Em levantamento realizado em 2011, foi identificado que, aproximadamente, 54% das iniciativas de arquivamento da *web* utilizavam os formatos ARC e WARC para arquivamento de conteúdo (GOMES; MIRANDA; COSTA, 2011).

MECANISMOS DE DISSEMINAÇÃO DE INFORMAÇÕES

A facilidade de recuperação de conteúdo em um arquivo *web* está implicada em fatores como o processo de seleção, os metadados e, especialmente, a ferramenta de busca utilizada, sendo que o processo de preservação quando bem definido é um fator fundamental para que o arquivo *web* alcance seu objetivo em disseminar informações preservadas em suas bases (KRAN, RAHMAN, 2019).

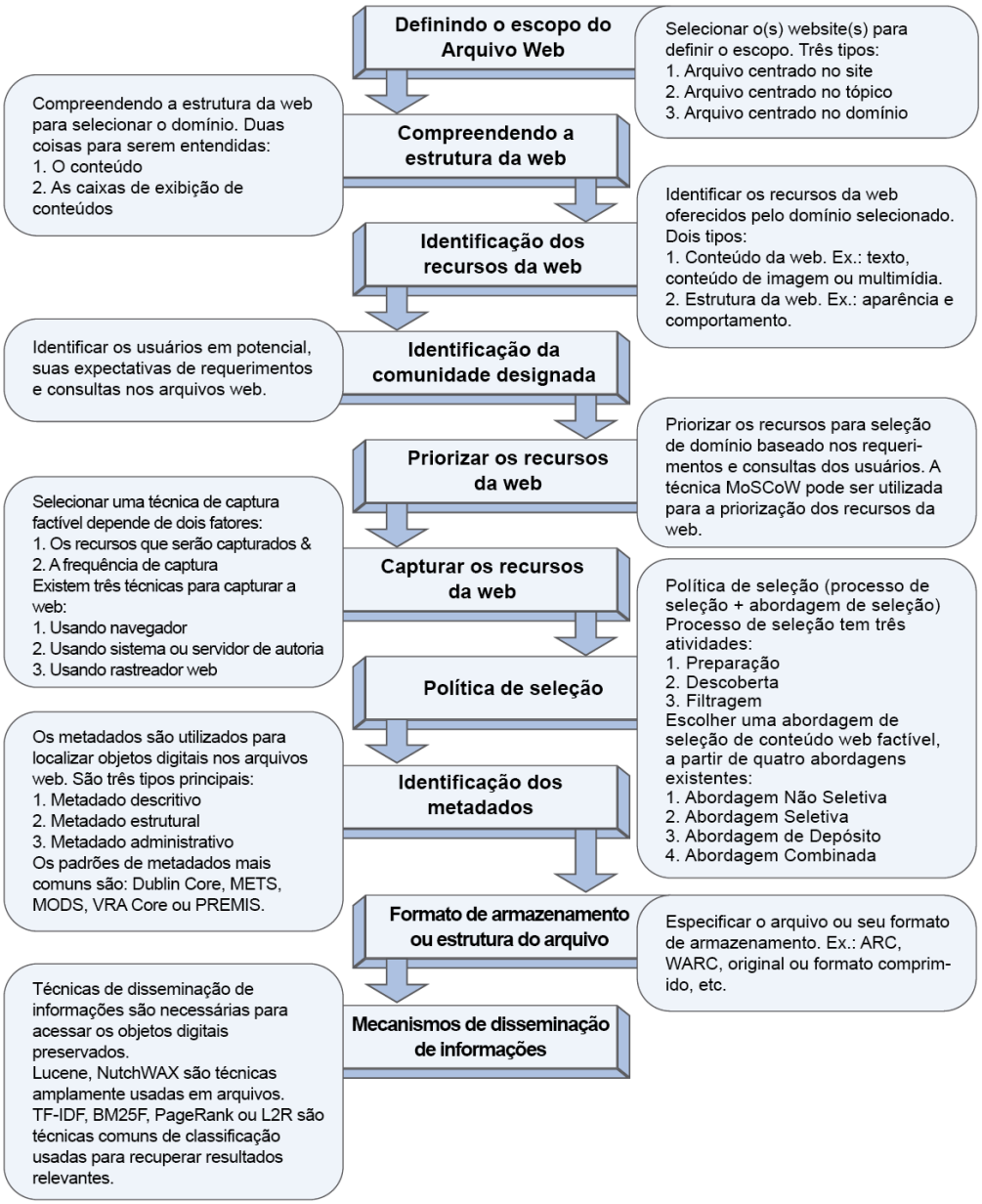
[40] http://www.loc.gov/standards/vracore/
[41] http://www.loc.gov/standards/premis/
[42] http://archive.org/*web*/researcher/ArcFileFormat.php
[43] https://www.iso.org/standard/44717.html

Para usar todo o potencial dos arquivos *web* é necessária uma interface amigável que facilitará a pesquisa ao usuário e, além disso, a utilização de metadados que ajudem a recuperar a partir do texto completo, de palavras-chave ou outras informações que, por ventura, poderão ser relevantes para a recuperação da informação pretendida pelos usuários. Outro fator relevante para otimizar a busca em um arquivo *web* são os sistemas de classificação de conteúdo, que se utilizam de ferramentas para este fim, como por exemplo, as ferramentas *Lucene*[44] e sua extensão *NutchWAX*[45] que são amplamente utilizados no arquivamento da *web*. O modelo de classificação escolhido para o arquivo *web* estima a relevância dos resultados com base nas consultas dos usuários a partir de critérios especificados colocando o resultado mais relevante no topo da lista de busca. Diversos modelos de classificação são apresentados na literatura específica, alguns exemplos de ferramentas comuns de classificação utilizadas para uma recuperação qualificada são o *TF-IDF*[46], *BM25F*[47], *PageRank*[48] e *L2R*[49]. (KRAN, RAHMAN, 2019).

Essas dez etapas completam a abordagem sistemática para preservação da *web* proposta pelos pesquisadores e estão resumidas na Figura 4, a seguir:

[44] https://lucene.apache.org/
[45] http://archive-access.sourceforge.net/projects/nutchwax/
[46] www.tfidf.com/
[47] https://www.researchgate.net/publication/308991534_A_Tutorial_on_the_BM25F_Model
[48] https://pt.wikipedia.org/wiki/PageRank
[49] http://times.cs.uiuc.edu/course/598f14/l2r.pdf

Figura 4 – Abordagem sistemática para o processo de preservação da *web*



Fonte: KRAN; RAHMAM, 2019, p. 74 (*tradução nossa*).

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa científica é classificada como de **natureza aplicada**, sendo que sua finalidade é gerar conhecimentos de aplicação prática para um problema específico a partir de uma amostragem: o arquivamento dos *websites* do governo federal brasileiro. Considerando os objetivos desta pesquisa, podemos classificá-la como **exploratória-descritiva**, ao proporcionar familiaridade com o problema proposto, com a intenção de torná-lo mais explícito, além disso, descreve as características que suportam o fenômeno do arquivamento da *web*, a partir da aplicação prática deste estudo de caso. Antonio Carlos Gil (2002, p. 42), diz que “As pesquisas descritivas são, juntamente com as exploratórias, as que habitualmente realizam os pesquisadores sociais preocupados com a atuação prática.” Em relação a abordagem do problema classificamos como mista, em razão dos dados terem sido analisados **quali e quantitativamente** com sua visualização facilitada por meio de gráficos, quadros e tabelas. Como procedimentos técnicos, foram adotados o **estudo de caso** e **pesquisa documental**. Conforme Yin (2001, p. 32) “[...] o estudo de caso é uma investigação empírica de um fenômeno contemporâneo dentro de um contexto da vida real, sendo que os limites entre o fenômeno e o contexto não estão claramente definido”. Desta forma, utilizar-se-á duas formas de coleta de dados: os *websites* da esfera federal, como objetos empíricos selecionados, e a coleta e análise das fontes documentais sobre a organização do Poder Executivo Federal. Para organização dos procedimentos, essa pesquisa foi estruturada em quatro etapas.

A **primeira etapa** consistiu na verificação dos *websites* das instituições do Poder Executivo Federal que utilizam o domínio gov.br. O levantamento considerou a estrutura organizacional do Poder Executivo Federal, a partir de informações presentes em plataformas e documentos oficiais do Governo Federal Brasileiro, tais como o *website* do SIORG (Sistema de Informações Organizacionais); o Decreto 9.660/2019, que dispõe “sobre a vinculação das entidades da administração pública federal indireta” (BRASIL, 2019a); e a Medida Provisória 870/2019, que “[...] estabelece a organização básica dos órgãos da Presidência da República e dos Ministérios, definindo suas competências e sua estrutura básica”, que fora

transformada na Lei 13.844, de 18 de junho de 2019 (CONGRESSO NACIONAL, *WEBSITE*, 2019).

Foram escolhidos os *websites* das instituições que compõem o primeiro escalão do Poder Executivo Federal, tais como os ministérios, secretarias e órgãos com status de ministério, por serem os portais que centralizam as informações e tomadas de decisões estruturais do Governo Federal. As instituições foram pesquisadas no portal de busca *Google*, com a intenção de verificar o acesso ao endereço eletrônico de seu *website*. A estrutura administrativa do primeiro escalão do Poder Executivo Federal foi tabelada de forma a permitir sua categorização segundo sua natureza jurídica e função governamental, assim como a identificação do endereço eletrônico de seu sítio *web*. A partir da estrutura administrativa foram selecionados 23 *websites* governamentais que serviram como amostragem, sendo 22 *websites* de ministérios, secretarias e órgão com status de ministério; e um central do Governo, o Portal Único www.gov.br. O Quadro 1 lista os órgãos e os *websites* que compõem o corpus de amostragem.

Quadro 1 - Lista dos *websites* que foram arquivados

Nº	Hierarquia	Nome do órgão	Site que será coletado
01	Órgão central	Poder Executivo Federal	www.gov.br
02	Órgão	Advocacia-Geral da União	www.agu.gov.br
03	Órgão	Banco Central do Brasil	www.bcb.gov.br
04	Órgão	Casa Civil da Presidência da República	https://www.gov.br/casacivil
05	Órgão	Gabinete de Segurança Institucional da Presidência da República	https://www.gov.br/gsi
06	Ministério	Controladoria-Geral da União	www.cgu.gov.br
07	Ministério	Ministério da Agricultura, Pecuária e Abastecimento	www.agricultura.gov.br
08	Ministério	Ministério da Cidadania	www.cidadania.gov.br
09	Ministério	Ministério da Ciência, Tecnologia, Inovações e Comunicações	www.mctic.gov.br
10	Ministério	Ministério da Defesa	www.defesa.gov.br
11	Ministério	Ministério da Economia	www.economia.gov.br
12	Ministério	Ministério da Educação	www.mec.gov.br

Nº	Hierarquia	Nome do órgão	Site que será coletado
13	Ministério	Ministério da Infraestrutura	www.infraestrutura.gov.br
14	Ministério	Ministério da Justiça e Segurança Pública	www.justica.gov.br
15	Ministério	Ministério da Mulher, da Família e dos Direitos Humanos	www.mdh.gov.br
16	Ministério	Ministério da Saúde	www.saude.gov.br
17	Ministério	Ministério das Relações Exteriores	www.itamaraty.gov.br
18	Ministério	Ministério de Minas e Energia	www.mme.gov.br
19	Ministério	Ministério do Desenvolvimento Regional	www.cidades.gov.br
20	Ministério	Ministério do Meio Ambiente	www.mma.gov.br
21	Ministério	Ministério do Turismo	www.turismo.gov.br
22	Secretaria	Secretaria de Governo da Presidência da República	www.gov.br/secretariadegoverno
23	Secretaria	Secretaria-Geral da Presidência da República	www.gov.br/secretariageral

Fonte: Elaborado pelo autor.

Importante destacar que de junho a dezembro de 2019, intervalo de tempo entre o desenvolvimento e execução desta pesquisa, alguns *websites* foram incorporados no Portal Único gov.br, conforme prevê o Decreto 9.756 (BRASIL, 2019b), que institui o Portal Único gov.br, já mencionado nesta pesquisa. Os *websites* da *Casa Civil da Presidência da República*, do *Gabinete de Segurança Institucional da Presidência da República*, da *Secretaria de Governo da Presidência da República* e da *Secretaria-Geral da Presidência da República* compuseram a lista de *websites* que tiveram suas ULRs modificadas no momento da coleta considerando sua incorporação ao Portal Único.

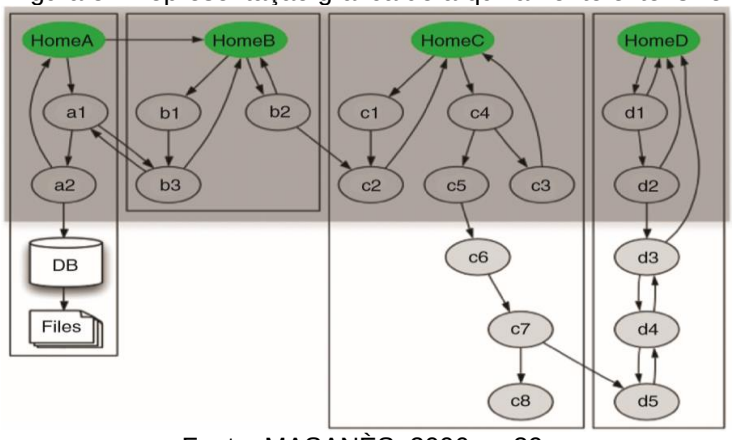
Após a seleção dos *websites*, deu-se início à **segunda etapa** da pesquisa: o arquivamento dos *websites* selecionados. Para o arquivamento dos *websites* utilizamos o *software Heritrix*, um rastreador de *URLs* desenvolvido pelo *Internet Archive*, sob a licença de *Software* Livre e *Open Source*; isso significa que além de ter seu código fonte aberto para consulta, o *Heritrix* pode ser executado, copiado, adaptado e redistribuído pelos usuários gratuitamente. Os usuários possuem livre acesso ao código-fonte do *software* e fazem alterações conforme as suas necessidades. Rockembach e Pavão (2018) destacaram o uso de *softwares* com código aberto e evidenciaram o *Heritrix*:

> 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. (ROCKEMBACH; PAVÃO, 2018, p. 179).

Os 23 pontos de entrada para a coleta foram previamente definidos, conforme apresentado no Quadro 1. Cabe destacar que estes *links* podem direcionar o rastreador para um novo ponto de entrada ou para outros elementos do mesmo site, sendo este o procedimento esperado quando de um arquivamento orientado a sites, conforme Masanès (2006, p. 38-39) discorre. Ainda, “[...] a integralidade [da coleta] pode ser medida horizontalmente pelo número de pontos de entrada relevantes encontrados dentro do perímetro designado e verticalmente pelo número de nós vinculados relevantes encontrados a partir desse ponto de entrada” (MASANÈS, 2006, p. 39, *tradução nossa*). Isso significa dizer que a profundidade da coleta deve ser orientada quando da definição do escopo: a profundidade de coleta é classificada como *extensiva* ou *intensiva* (MASANÈS, 2006).

A profundidade do arquivamento é chamada de “extensiva” quando a integralidade horizontal é preferida à integralidade vertical (MASANÈS, 2006, p. 39), ou seja, “[...] procura cobrir os domínios em seus primeiros níveis, de uma forma mais abrangente e trazendo um panorama da *web* a partir do seu arquivamento [...]” (ROCKEMBACH; PAVÃO, 2018, p. 175). A Figura 5 ilustra uma coleta extensiva: o conteúdo sombreado será arquivado, possibilitando a inclusão de um número maior de *websites*, mas preservando apenas no nível da superfície e fazendo com que páginas profundas na hierarquia (c6, c7, c8, d3, d3, d5), bem como banco de dados ocultos, não sejam capturados.

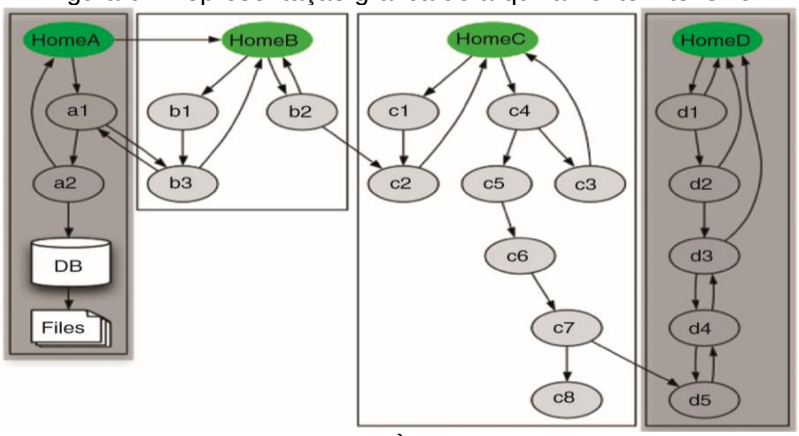
Figura 5 - Representação gráfica do arquivamento extensivo



Fonte: MASANÈS, 2006, p. 39.

Por outro lado, a coleta com profundidade "intensiva" acontece quando a integralidade vertical é preferida à integralidade horizontal (MASANÈS, 2006, p. 40), ou seja, “[...] procura concentrar-se em alguns sites de maneira a arquivar o máximo de níveis, incluindo outros elementos, como banco de dados” (ROCKEMBACH; PAVÃO, 2018, p. 175), necessitando acesso aos sistemas e servidores. Esta abordagem demanda mais trabalho, mas possibilita a preservação, não somente dos primeiros níveis, mas de toda a hierarquia do *website*, permitindo que se mantenha a navegação por *hiperlinks* no arquivo *web* (ROCKEMBACH; PAVÃO, 2018, p. 175). A Figura 6 ilustra uma coleta intensiva: considerando que o conteúdo sombreado será arquivado, nota-se que serão preservados um número menor de *websites*, porém o rastreamento é feito em profundidade, em que apenas o site A e D serão arquivados, incluindo a parte da *web* oculta do site A.

Figura 6 - Representação gráfica do arquivamento intensivo

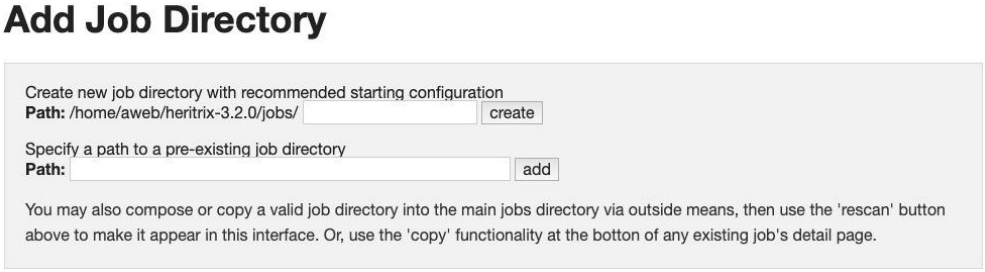


Fonte: MASANÈS, 2006, p. 40.

Para a coleta que foi desenvolvida nesta pesquisa, optou-se pela abordagem extensiva considerando a grande quantidade de *websites* que foram coletados manualmente e o espaço disponível no servidor do núcleo de pesquisa NUAWEB para armazenamento destes arquivos. Em nenhuma das coletas realizadas foi possível navegar pelos *links*, isso significa que apenas os arquivos da página inicial de cada um dos *websites* foram arquivados.

O *Heritrix* utilizado foi a versão 3, último lançamento estável, que está à disposição do Núcleo de Pesquisa NUAWEB e instalado no servidor da UFRGS. O primeiro passo para iniciar uma coleta utilizando o rastreador *web Heritrix* foi adicionar um novo diretório de trabalho (*Add Job Directory*) na tela inicial do rastreador (Figura 7).

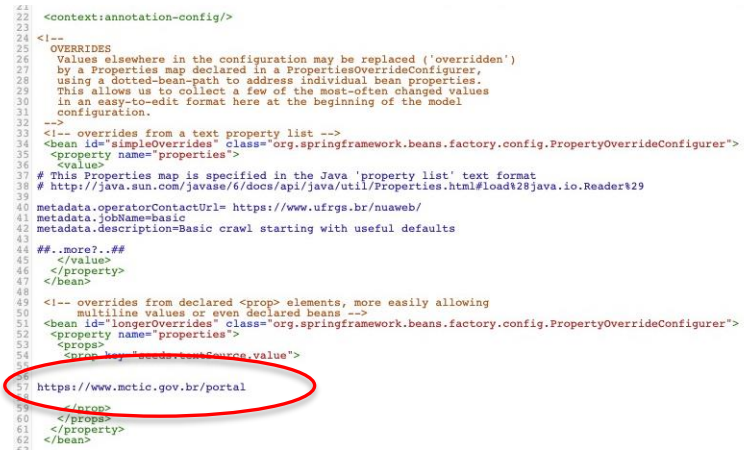
Figura 7 – Tela inicial do Heritrix



Fonte: Software Heritrix, 2019.

Logo após, com o novo diretório aberto, deu-se início às configurações para realização da coleta automatizada. Para isso, foi necessário inserir manualmente as *URLs* que serão coletadas, que na expressão técnica são chamadas de *seeds* (sementes). Para que a coleta acontecesse, algumas linhas da linguagem de programação foram previamente demarcadas para que fossem substituídas, como demonstra a Figura 8, em que é possível ver que o campo *INSERT URL HERE* foi substituído pela *URL* que será coletada. Este procedimento foi realizado para cada um dos 23 *websites* que compuseram o campo amostral.

Figura 8 – Captura de tela da página de configurações do Heritrix



Fonte: Software Heritrix, 2019.

Concluída a configuração, deu-se início a coleta, propriamente dita. A Figura 9 apresenta um exemplo da tela em que a coleta é realizada. Inicialmente, clica-se em ***Build***, habilitando a opção *Info Job instantiated*, este comando irá montar a infraestrutura para que a coleta seja executada; em seguida, clica-se no botão ***Launch*** que iniciará a tarefa no modo ***Pause***, aparecendo o termo PREPARING no campo *Job is Active*; ao atualizar a página será liberado o botão ***Unpause*****,** que ao ser clicado dará início à coleta a partir do *seed* informado no momento das configurações e no campo *Job is Active* aparecerá RUNNING. Neste momento, os arquivos que estão sendo coletados começarão a aparecer no campo *Crawl Log*.

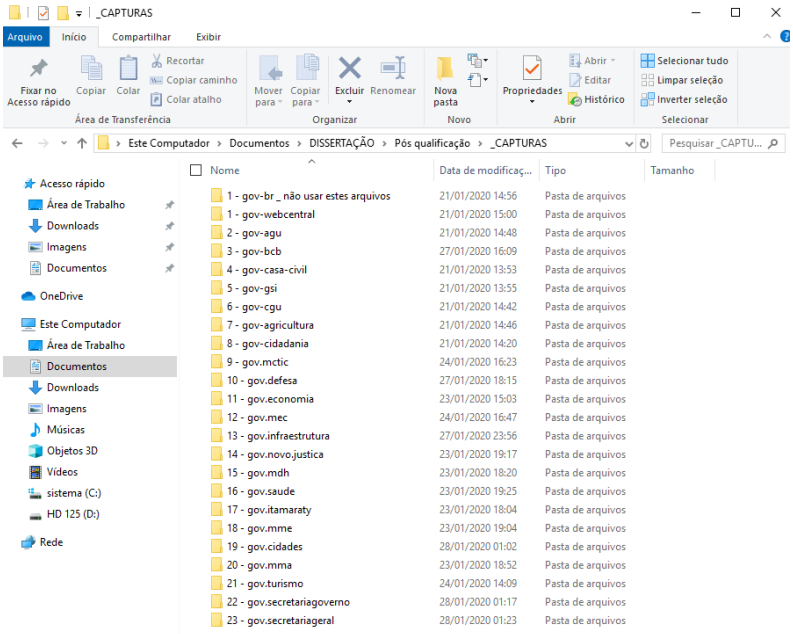
A quantidade de *websites* selecionados como campo amostral desta pesquisa, assim como o espaço disponível no servidor para armazenamento dos arquivos coletados, também foram os balizadores para que se estabelecesse o tempo de 20 minutos de coleta para cada um dos *websites* selecionados. Além disso, a quantidade de arquivos que estavam sendo coletados durante os primeiros experimentos se mostrou suficiente, conforme demonstrado na Figura 9, que em 20 minutos e 1 segundo fez 481 downloads e mais 579 arquivos na fila para download, equivalente ao total de 60 *megabyte*, aproximadamente. Passados os 20 minutos de coleta, clica-se no botão ***Terminate***, para interromper e finalizar a coleta.

Figura 9 – Captura de tela da coleta com Heritrix para o job gov.mctic2

Fonte: Software Heritrix, 2019.

Estes procedimentos foram realizados para as 23 *seeds* previamente selecionadas. Depois de concluída toda coleta, os arquivos foram exportados do servidor da UFRGS para o computador local, descompactados e, as pastas renomeadas com a sequência já estabelecida entre os *websites* para melhor sistematização da análise dos dados (Figura 10) e deu-se início a terceira etapa da pesquisa.

Figura 10 – Captura das pastas com as coletas realizadas

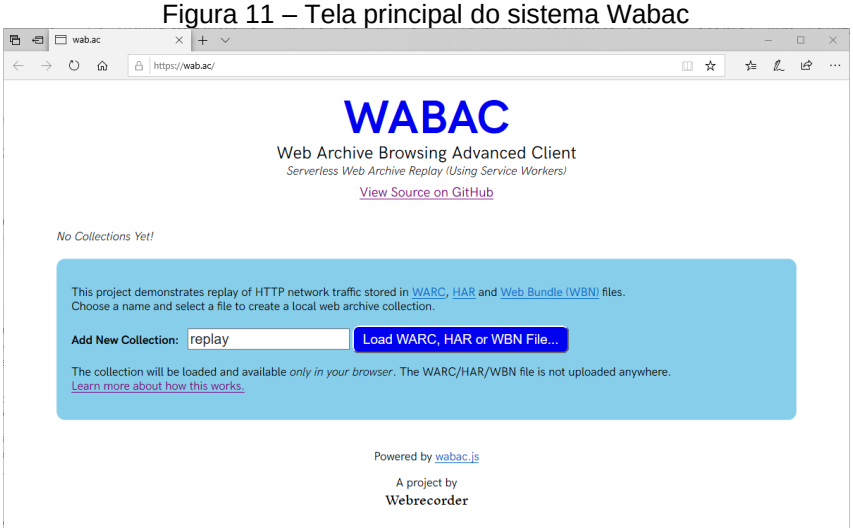


Fonte: Elaborado pelo autor.

A **terceira etapa** da pesquisa foi direcionada à reprodução dos arquivos WARC previamente capturados, com o uso do *Wabac* (*Web Archive Browsing Advanced Client*), disponível no endereço eletrônico https://wab.ac/. Trata-se de um sistema de reprodução de arquivos da *web* implementado por meio de *Service Worker*[50], que permite a navegação em páginas da *web* por meio de arquivos WARC ou HAR com o uso do navegador *Chrome*, sem a necessidade de um componente de servidor (WABAC, *website*, 2020).

O processamento se dá com a inserção do arquivo em formato WARC no sistema, que fará a leitura e reproduzirá o arquivo, montando o *website* a partir daquilo que foi capturado com o uso do rastreador *web*. Na tela inicial do *Wabac*, o arquivo WARC será inserido ao clicar no botão ***Load WARC, HAR or WBN file*** (Figura 11).

Figura 11 – Tela principal do sistema Wabac



Fonte: Software Wabac, 2019.

Quando o arquivo WARC é processado pelo sistema, aparecerá a lista de *links* recuperados pelo rastreador (Figura 12). Ao clicar no *link* principal do *website* que se está arquivando, o sistema irá reconstruir o *website* arquivado a partir das informações e recursos que foram recuperados no momento da coleta feita pelo rastreador *web*. Em ***Source*** pode-se ver o arquivo WARC utilizado e em ***Pages*** é apresentada a lista de *links* que foram recuperados a partir deste arquivo WARC

[50] Um *service worker* é um *script* executado por meio de um navegador em segundo plano.

inserido no sistema. Ao clicar no primeiro *link* da lista apresentada o *website* será reconstruído.

Figura 12 – Captura de tela com a lista de *links* recuperados pelo rastreador

Fonte: Software Wabac, 2019.

A **quarta** etapa da pesquisa foi direcionada à verificação dos recursos que, efetivamente, foram arquivados, promovendo a comparação do *website* ao vivo com o *website* reconstruído na etapa anterior. Inicialmente, a proposta pretendia verificar informações como o tipo de serviços disponibilizados em cada *website* selecionado; a identificação dos recursos oferecidos, tais como áudios ou rádio *web*, vídeos, transmissões *online*, ferramentas de busca no conteúdo do site; formatos de arquivos disponibilizados nos *websites*; o tipo de informações disponibilizadas nos *websites*; outros endereços eletrônicos em que há direcionamento; e a qualificação das abas de direcionamento para conteúdos presentes em cada *website*; e outras informações que forem julgadas necessárias no decorrer do reconhecimento desta amostra selecionada.

Para estabelecer estes itens de verificação foram recolhidas informações prévias na base aberta de microdados do Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br), departamento do Núcleo de Informação e Coordenação do Ponto BR (Nic.br), que implementa as decisões e projetos do Comitê Gestor da Internet do Brasil (Cgi.br). A base aberta de microdados elabora, anualmente, desde 2013, indicadores referentes a diversas

questões relacionadas à tecnologia e ao uso da Internet no Brasil. A pesquisa TIC Governo Federal (CETIC.BR, *websites*, 2019) busca investigar a oferta de e-governo no Brasil e o uso das tecnologias de informação e comunicação no setor público. As pesquisas disponibilizadas no *website* do órgão correspondem a informações genéricas em relação a todos os órgãos do governo federal e seus poderes, além de números referentes a atuação dos governos estaduais nesta área. Pode-se perceber, por exemplo, a partir dos dados coletados na pesquisa TIC Governo Eletrônico de 2017, que 100% dos órgãos federais possuem *website*; no poder executivo federal 96% dos *websites* possuem arquivos em pdf, 19% em odt e 66% em doc ou docx; 57% dos *websites* do poder executivo possuem vídeos em seu conteúdo (CETIC.BR, *website*, 2019).

Para capturar essas informações, foi elaborada uma planilha, em que as colunas representavam cada um dos recursos que estavam sendo verificados e as linhas representam os *websites*, sendo que para cada um dos 23 *websites* analisados haviam dispostas duas linhas: uma representando o *website* ao vivo e outra representando o *website* arquivado. Durante a realização desta análise, se percebeu que muitos recursos estabelecidos para verificação, tais como os tipos de arquivos (pdf, doc, docx...etc), não haviam a possibilidade de serem identificados nos *websites*; informações como, por exemplo, *gerar e pagar boletos*, *preencher formulários*, *realizar agendamento de consultas*, não são recursos tradicionalmente oferecidos pelos órgãos do primeiro escalão do governo, que compõe nosso corpus amostral, mas por órgãos e instituições integrantes à estrutura administrativa destes Ministérios. Neste sentido, optou-se por remover, ajustar e acrescentar alguns recursos para dar início a uma nova análise. A estratégia utilizada foi voltar à literatura já apresentada nesta pesquisa de mestrado (Capítulo 4), com a intenção de extrair recursos que poderiam ser relevantes para estabelecer se o *website* apresenta o potencial de ser arquivado com integridade.

A partir disso, se chegou à nova proposta de itens que seriam analisados, seguidos por sua escala de pontuação que varia de 0 (zero) a 3 (três), com respectiva legenda, aplicada para 14 (quatorze) dos 16 (dezesseis) recursos analisados, conforme Quadro 2. A exceção foi para os itens “*Layout* do *website*”, se antigo ou novo (exemplos no Quadro 3); e se o “*Layout* se manteve” conforme o *website* ao vivo, com variações “não”, “em parte” ou “sim” (exemplos no Quadro 4).

Quadro 2 – Recursos analisados com a respectiva escala de ocorrência

	ESCALA			
RECURSO EM ANÁLISE	**0**	**1**	**2**	**3**
Layout* do *website		antigo	novo	
***Layout* do *website* se manteve?**		não	em partes	sim
Áudio ou rádio *web*	sem vestígio	apenas o cabeçalho ou título	cabeçalho, título e estrutura de ícones	há possibilidade de reprodução
Ferramenta de busca dos conteúdos do *website*	sem vestígio	apenas o cabeçalho ou título	é possível digitar para realizar a busca	a busca é efetivada no conteúdo preservado
Transmissão *online* em tempo real de eventos como sessões, palestras e reuniões	sem vestígio	apenas o cabeçalho ou título	cabeçalho, título e estrutura de ícones	há possibilidade de reprodução
Imagens ilustrativas de notícias	sem vestígio	apenas o cabeçalho ou título	a imagem aparece fora da estrutura original	existe imagem ilustrando a notícia
Imagens ilustrativa para ícones/*links*	sem vestígio	apenas o cabeçalho ou título	a imagem aparece fora da estrutura original	existe imagem ilustrando os ícones
Vídeos	sem vestígio	apenas o cabeçalho ou título	título e estrutura de ícones	há possibilidade de reprodução
Mapa do site	sem vestígio	apenas o cabeçalho ou título	cabeçalho ou título e conteúdo	é possível navegar no mapa do site
Estrutura organizacional do órgão	sem vestígio	apenas o cabeçalho ou título	título e *link*	apresenta título, *link* e estrutura
Interatividade com agenda	sem vestígio	apenas o cabeçalho ou título	título e texto	conteúdo com possibilidade de interação
Acompanhamento em tempo real (horário, previsão do tempo, variação cambial, mapa, etc)	sem vestígio	apenas o cabeçalho ou título	apresenta o conteúdo	conteúdo com possibilidade de interação
Interoperabilidade com redes sociais	sem vestígio	apenas o cabeçalho, título ou ícone	apresenta o conteúdo	conteúdo com possibilidade de interação
Banner rotativo	sem vestígio	apenas o cabeçalho ou título	texto e imagem	texto, imagem e há possibilidade de reprodução
Menu de navegação	sem vestígio	texto e *hiperlinks* desformatados	texto e *hiperlinks* formatados	existe menu de navegação
Apresentação do *feed* de notícias	sem vestígio	texto e *links*	texto, *links* e imagens desformatadas	texto, *link* imagem formatadas

Fonte: Elaborado pelo autor.

A planilha elaborada para a sistematização dos dados teve que ser ajustada considerando os novos critérios de análise. Além das colunas representando cada um dos recursos que seriam verificados, outras colunas compuseram o instrumento: *Nome do arquivo WARC*; *Data da coleta*; *Hora da coleta*; *URL da coleta*; e

Identificação e variação. Essa última refere-se ao número que cada *website* recebeu para sua melhor identificação ao longo da pesquisa e a sinalização se a linha se refere a análise feita no *website* ao vivo ou no arquivado. Os dados da data e hora da coleta foram extraídos do nome do arquivo WARC (*20191121191320454.warc*), que é composto da seguinte forma:

2019	11	21	19	13	20	454
ano	mês	dia	hora	minuto	segundo	código automático

A Figura 13, ilustra a primeira parte da planilha:

Figura 13 – Planilha de sistematização dos dados – Parte 1[51]

Nome arquivo WARC	Data da coleta	Hora de coleta	URL de coleta	Identificação e variação	
gov-webcentral	21/11/2019	17:36'56"	www.gov.br	1	ao vivo
				1	arquivado
gov-agu	21/11/2019	18:04'14"	www.agu.gov.br	2	ao vivo
				2	arquivado
gov-bcb	21/11/2019	18:31'02"	www.bcb.gov.br/	3	ao vivo
				3	arquivado
gov-casa-civil	21/11/2019	18:31'21"	www.casacivil.gov.br/	4	ao vivo
				4	arquivado
gov-gsi	21/11/2019	18:31'35"	www.gsi.gov.br/ *	5	ao vivo
				5	arquivado
gov-cgu	21/11/2019	19:05'16"	www.cgu.gov.br/	6	ao vivo
				6	arquivado
gov-agricultura	21/11/2019	19:13'20"	www.agricultura.gov.br/	7	ao vivo
				7	arquivado
gov-cidadania	21/11/2019	19:05'31"	www.cidadania.gov.br/	8	ao vivo
				8	arquivado

Fonte: Elaborado pelo autor.

A segunda parte da planilha apresenta as colunas dos 16 (dezesseis) recursos analisados, associados à sua escala de ocorrência; e as linhas de cada um dos *websites* do corpo amostral. Foi aplicada a funcionalidade “filtro”, disponível do *Excel*, para facilitar a leitura dos dados no momento de interpretá-los. As Figuras 14, 15, 16 e 17 ilustram os campos da tabela.

[51] Segue na Figura 14 até a Figura 17.

Figura 14 – Planilha de sistematização dos dados – Parte 2.1

RECURSOS OFERECIDOS		Layout	Layout se manteve	Áudio ou rádio web	Ferramenta de busca dos conteúdos do website
LEGENDA DA ESCALA		1 - antigo | 2 - novo	1 - não | 2 - em partes | 3 - sim	0 - sem vestígio | 1 - apenas o cabeçalho ou título | 2 - cabeçalho, título e estrutura de ícones | 3 - há possibilidade de reprodução	0 - sem vestígio | 1 - apenas o cabeçalho ou título | 2 - é possível digitar para realizar a busca | 3 - a busca é efetivada no conteúdo preservado
Identificação e variação					
1	ao vivo				
1	arquivado				
2	ao vivo				
2	arquivado				
3	ao vivo				
3	arquivado				
4	ao vivo				
4	arquivado				
5	ao vivo				
5	arquivado				
6	ao vivo				
6	arquivado				
7	ao vivo				
7	arquivado				

Fonte: Elaborado pelo autor.

Figura 15 – Planilha de sistematização dos dados – Parte 2.2

Ferramenta de busca dos conteúdos do website	Transmissão online em tempo real de eventos como sessões, palestras e reuniões	Imagens ilustrativas de notícias	Imagens ilustrativa para ícones/links	Vídeos
0 - sem vestígio | 1 - apenas o cabeçalho ou título | 2 - é possível digitar para realizar a busca | 3 - a busca é efetivada no conteúdo preservado	0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - cabeçalho, título e estrutura de ícones | 3 - há possibilidade de reprodução	0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - a imagem aparece fora da estrutura original | 3 - existe imagem ilustrando a notícia	0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - a imagem aparece fora da estrutura original | 3 - existe imagem ilustrando os ícones	0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - cabeçalho, título e estrutura de ícones | 3 - há possibilidade de reprodução

Fonte: Elaborado pelo autor.

Figura 16 – Planilha de sistematização dos dados – Parte 2.3

Vídeos	Mapa do site	Estrutura organizacional do órgão	Interatividade com agenda	Acompanhamento em tempo real (horário, previsão do tempo, variação cambial, mapa, etc)
0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - cabeçalho, título e estrutura de ícones | 3 - há possibilidade de reprodução	0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - cabeçalho ou título e conteúdo | 3 - é possível navegar no mapa do site	0 - sem vestígio | 1 - apenas o título | 2 - título e link | 3 - título, link e estrutura	0 - sem vestígio | 1 - apenas o título | 2 - título e texto | 3 - conteúdo com possibilidade de interação	0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - conteúdo | 3 - conteúdo com possibilidade de interação

Fonte: Elaborado pelo autor.

Figura 17 – Planilha de sistematização dos dados – Parte 2.4

Acompanhamento em tempo real (horário, previsão do tempo, variação cambial, mapa, etc)	Interoperabilidade com redes sociais	Banner rotativo	Menu de navegação	Apresentação do feed de notícias
0 - sem vestígio | 1 - apenas cabeçalho ou título | 2 - conteúdo | 3 - conteúdo com possibilidade de interação	0 - sem vestígio | 1 - apenas cabeçalho, título ou ícone | 2 - apresenta o conteúdo | 3 - apresenta conteúdo com possibilidade de interação	0 - sem vestígio | 1 - apenas o texto | 2 - texto e imagem | 3 - texto, imagem e há possibilidade de reprodução	0 - sem vestígio | 1 - texto e hiperlinks desformatados | 2 - texto e hiperlinks formatados | 3 - existe menu de navegação	0 - sem vestígio | 1 - texto e links | 2 - texto, links e imagens, porém desformatadas | 3 - texto, link imagem formatadas

Fonte: Elaborado pelo autor.

Na planilha, foram preenchidos com “x” aqueles recursos que o *website* analisado não possuía em seu formato ao vivo. Para os *websites* que apresentaram o recurso que estava sendo analisado, se estabeleceu uma variação da escala conforme a ocorrência que melhor o representa entre 0 (zero) e 3 (três), conforme a descrição da legenda para cada recurso (Quadro 2). Por exemplo: ao analisar o recurso “Banner rotativo”, o que se pretendia era saber se o *website* arquivado apresentou o recurso: se não, utilizou-se um “x”; se sim, utilizamos as seguintes variações: 0 (zero) caso o *website* não tivesse deixado nenhum vestígio do banner rotativo; 1 (um) para caso o *website* apresentasse apenas o cabeçalho ou título do banner rotativo; 2 (dois) para caso aparecesse apenas o texto ou imagem; ou 3 (três) para caso o recurso estivesse completo, apresentando texto, imagem e a possibilidade de reprodução do banner rotativo.

Após definidos os critérios para preenchimento da planilha de sistematização dos dados coletados, foi dado início a análise, a qual passamos a detalhar no capítulo a seguir.

4 ANÁLISE DOS DADOS: POSSIBILIDADES DE ARQUIVAMENTO DE *WEBSITES*

A sistemática para a análise dos dados e comparação entre o *website* ao vivo e o arquivado foi realizada de forma simultânea ao abrir o *website* ao vivo que está sendo analisado e, ao mesmo tempo, abrindo no sistema WABAC o arquivo WARC capturado. Cada um dos 16 (dezesseis) itens analisados foram confrontados e o escalamento numérico aplicado na planilha de sistematização dos dados, conforme definições apresentadas na metodologia. Os resultados foram interpretados com o auxílio da ferramenta *filtro* do *Excel*, da *Microsoft*, e apresentados em gráficos para melhor visualização.

Importante destacar que dos 23 (vinte e três) *websites* que compõe o corpo amostral, teve-se que se refazer a captura para 03 (três) que, inicialmente, tiveram seus arquivos WARC corrompidos e não recuperaram o conteúdo. São eles: https://www.gov.br/casacivil/pt-br; https://www.gov.br/gsi/pt-br; e http://www.mctic.gov.br/. Na segunda captura destes 03 (três) *websites*, que foi realizada considerando os mesmos procedimentos metodológicos iniciais, os arquivos estavam completos e foi possível analisar os seus recursos.

Foram os recursos analisados, os quais passaremos a apresentar, individualmente:

a) O *website* apresenta o ***Layout*** antigo ou novo
b) O ***Layout* se manteve** íntegro conforme o *website* ao vivo
c) O *website* apresenta **Áudio ou rádio *web***
d) Existe **Ferramenta de busca dos conteúdos do website**
e) O *website* apresenta ferramenta de **Transmissão *online* em tempo real de eventos com sessões, palestras e reuniões**
f) O *website* apresenta **Imagens ilustrativas de notícias**
g) O *website* apresenta **Imagens ilustrativa para ícones/*links***
h) O *website* apresenta **Vídeo**
i) O *website* possui **Mapa do site**

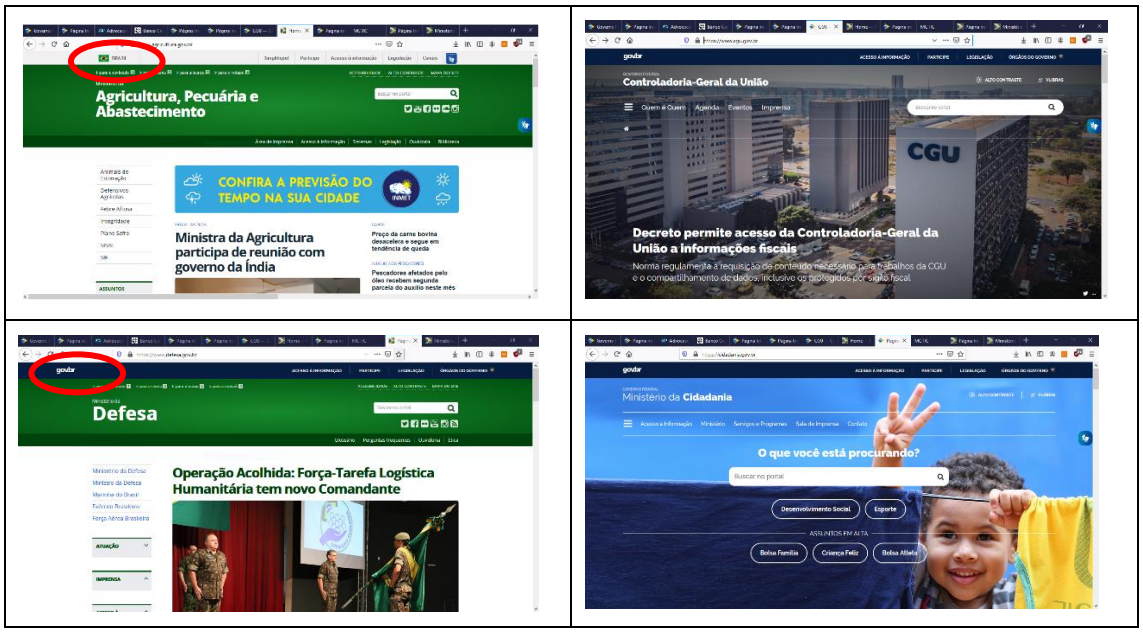
j) O *website* apresenta a **Estrutura organizacional do órgão**
k) É possível **Interatividade com agenda**
l) Existe **Recursos de acompanhamento em tempo real (horário, previsão do tempo, variação cambial, mapa, etc)**
m) O *website* possui **Interoperabilidade com redes sociais**
n) O *website* possui **Banner rotativo**
o) O *website* possui **Menu de navegação**
p) Há permanência da **Apresentação do *feed* de notícias**

LAYOUT

O Gráfico 1 representa o quantitativo de *websites* com o antigo e com o novo *layout*, considerando os parâmetros apresentados no Quadro 3 a seguir. Percebe-se, portanto, que, de fato, trata-se de um momento de transição nas tomadas de decisão relacionadas aos *websites* do Governo Federal Brasileiro. A diferença de *layout* dos *websites* do primeiro escalão do governo é de apenas 03 (três). O Decreto que instituiu o Portal Único do governo de certa forma influenciou neste resultado, considerando que dos *websites* com *layout* novo, 04 (quatro) já estão associados ao Portal Único e os outros 06 (seis) são novos, mas ainda não incorporados no Portal Único gov.br. Outros 13 (treze) *websites* seguem no padrão antigo de *layout*, inclusive, apresentam diferenças entre eles, conforme pode-se notar no Quadro 3.

Quadro 3 – Exemplos de *websites* com *layout* antigo e novo

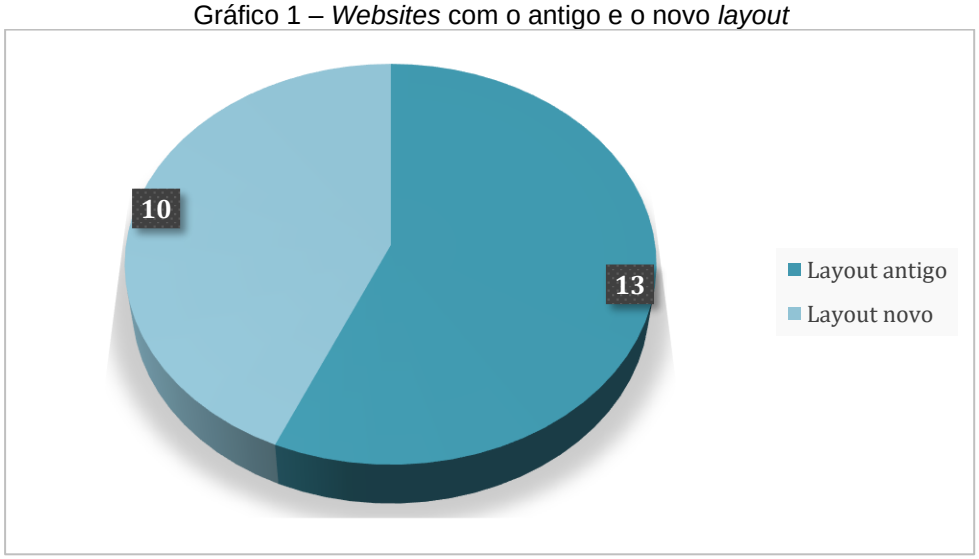
EXEMPLOS DE *LAYOUT* ANTIGO	**EXEMPLOS DE *LAYOUT* NOVO**
	



Fonte: Elaborado pelo autor

Salienta-se que a seleção das *URLs* que seriam coletadas fora realizada no mês de junho de 2019, tendo suas coletas efetivadas nos dias 21 de novembro e em 17 de dezembro de 2019, momento em que se percebeu a incorporação destes 04 (quatro) *websites* no Portal Único gov.br: *Casa Civil da Presidência da República*; *Gabinete de Segurança Institucional da Presidência da República*; *Secretaria de Governo da Presidência da República* e; *Secretaria-Geral da Presidência da República*. Para estes portais teve-se que, no momento da coleta com o *Heritrix*, alterar a *URL*, o que nada interferiu no desenvolvimento da pesquisa, do contrário, vimos como uma oportunidade ter estes dois diferentes parâmetros para estudo e análise científica.

Gráfico 1 – *Websites* com o antigo e o novo *layout*



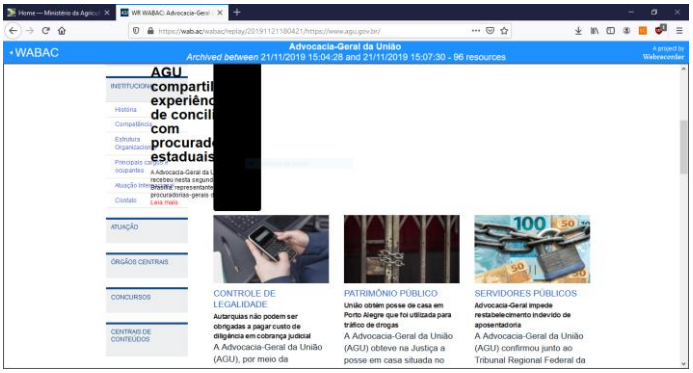
Fonte: Elaborado pelo autor.

LAYOUT SE MANTEVE

Nota-se que o “sim” é maioria em comparação com as outras duas possibilidades, sendo atribuído a 10 (dez) *websites* que se mantiveram com o mesmo *layout* de sua versão ao vivo. Outros 09 (nove) não se mantiveram íntegros em comparação com seu representante ao vivo; e 04 (quatro) *websites* mantiveram seu *layout* em partes, o que significa que pequenas falhas como, por exemplo, não reconhecimento de uma imagem ou deslocamento do título, puderam ter acontecido. O Quadro 4 apresenta exemplos dos *websites* e o Gráfico 2 ilustra o resultado.

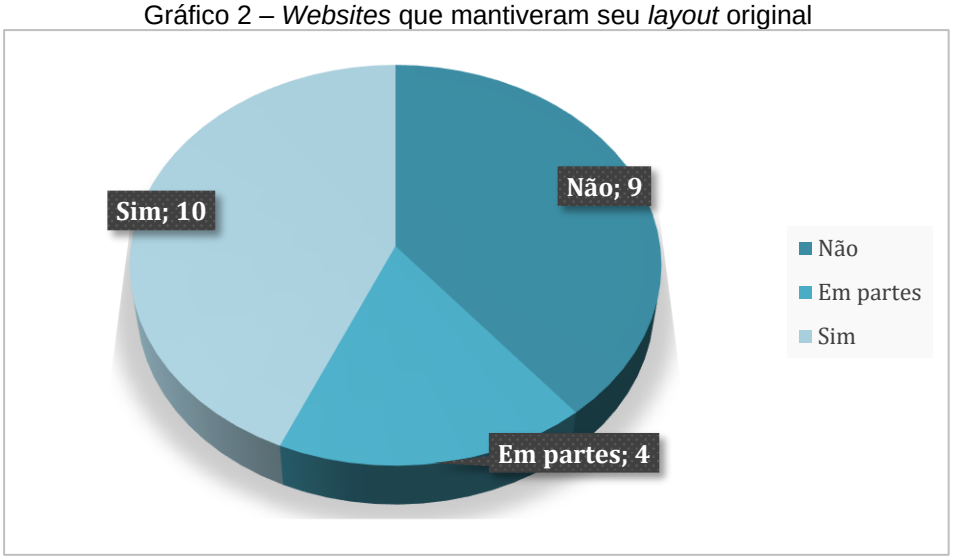
Quadro 4 – Exemplos de *websites* com critério de manutenção do *layout* original

NÃO MANTEVE O *LAYOUT*	

MANTEVE EM PARTES O *LAYOUT*	
SIM, MANTEVE O *LAYOUT*	

Fonte: Elaborado pelo autor.

Gráfico 2 – *Websites* que mantiveram seu *layout* original

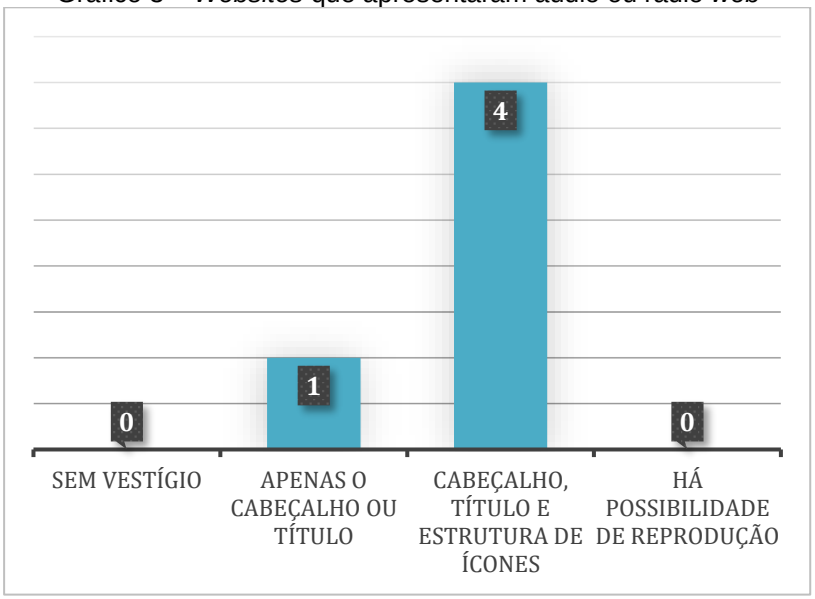


Fonte: Elaborado pelo autor.

ÁUDIO OU RÁDIO WEB

Este é um recurso importante para o arquivamento da *web* por se tratar de um tipo de arquivo específico e com necessidade de tratamento diferenciado dos arquivos textuais, por exemplo. Os arquivos em áudio, para que façam sentido, precisam que seu conteúdo seja reproduzido. Os resultados apresentados no Gráfico 3 mostram que dos 23 (vinte e três) *websites* analisados, apenas 05 (cinco) apresentaram áudios ou rádio *web*. Ao analisar estes *websites* arquivados pode-se perceber que 01 (um) apresentou apenas o cabeçalho ou o título do áudio e os outros 04 (quatro) apresentaram, também, a estrutura de ícones. Em nenhum dos *websites* houve a possibilidade de reprodução dos áudios.

Gráfico 3 – *Websites* que apresentaram áudio ou rádio *web*

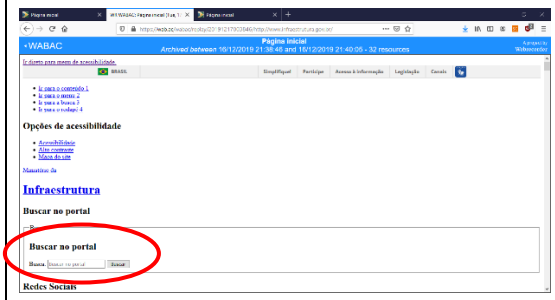
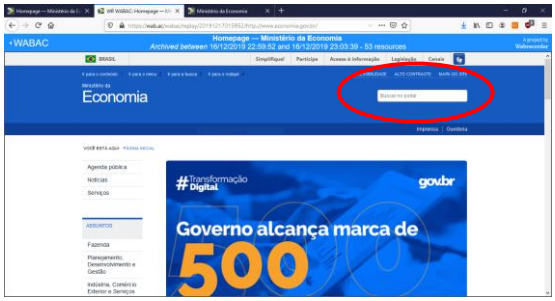


Fonte: Elaborado pelo autor.

FERRAMENTA DE BUSCA DOS CONTEÚDOS DO *WEBSITE*

O recurso de busca de conteúdo é habitualmente utilizado em *websites*, encontra-se presente em 21 (vinte e um) dos analisados, sendo que em 14 (quatorze) *websites* é possível, inclusive, digitar algum termo para busca, porém, em nenhum dos *websites* a busca, de fato, é realizada. Em 06 (seis) *websites* aparece apenas o cabeçalho ou título, em que não é possível digitar o termo para a busca (Quadro 5). Em apenas 01 (um) dos *websites* arquivados não ficou nenhum vestígio dá funcionalidade de busca. Os resultados são apresentados no Gráfico 4.

Quadro 5 – Exemplo de *website* arquivado – ferramenta de busca

APENAS O CABEÇALHO OU TÍTULO	É POSSÍVEL DIGITAR PARA REALIZAR A BUSCA
	

Fonte: Elaborado pelo autor.

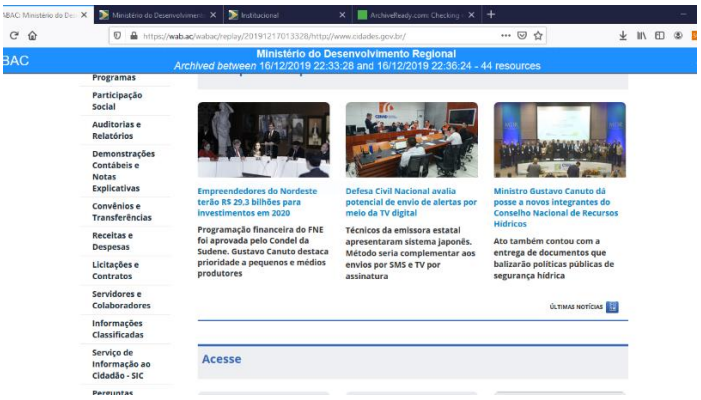
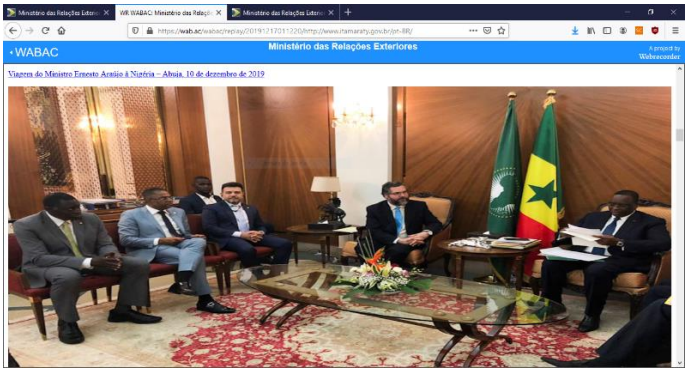
Gráfico 4 – *Websites* que apresentaram a ferramenta de busca

Fonte: Elaborado pelo autor.

IMAGENS ILUSTRATIVAS DE NOTÍCIAS

Ao analisar este recurso, a intenção foi verificar se as notícias eram acompanhadas de imagens ilustrativas e como essas imagens se apresentaram após o arquivamento do *website*. Pode-se observar, incialmente, que dos 23 (vinte e três) *websites* analisados, 19 (dezenove) continham imagem ilustrando suas notícias. Destes, 10 (dez) mostraram as imagens na sua totalidade, conforme o *website* ao vivo; em 02 (dois) *websites* as imagens apareciam fora da estrutura original; em 06 (seis) apareceu apenas o cabeçalho ou o título da imagem; e em 01 (um) dos *websites* arquivados não houve vestígio da existência de imagens nas notícias. O Quadro 6 apresenta um exemplo em cada uma das escalas de ocorrência e o Gráfico 5, ilustra a frequência descrita.

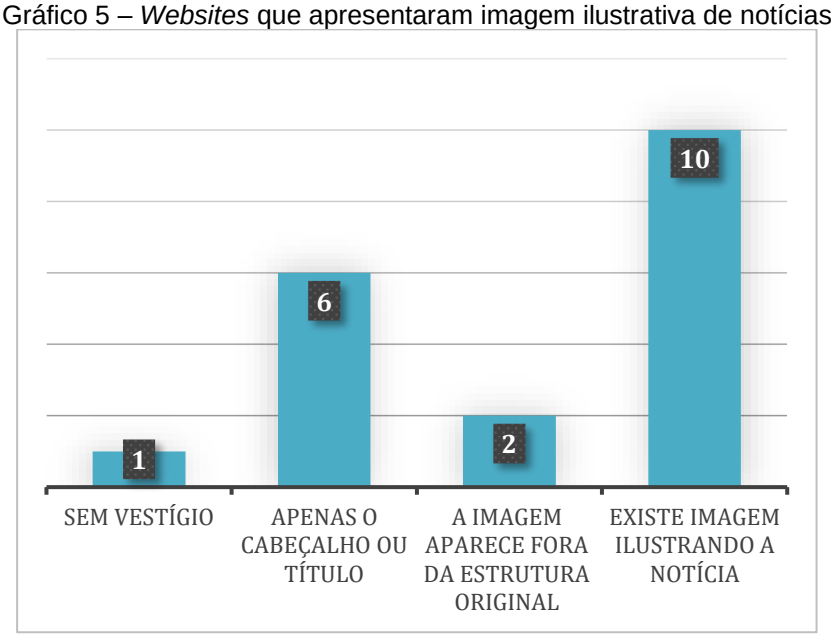
Quadro 6 – Exemplo de *website* arquivado – imagem ilustrativa de notícias

EXISTE IMAGEM ILUSTRANDO A NOTÍCIA	
A IMAGEM APARECE FORA DA ESTRUTURA ORIGINAL	

APENAS CABEÇALHO OU TÍTULO	

Fonte: Elaborado pelo autor.

Gráfico 5 – *Websites* que apresentaram imagem ilustrativa de notícias

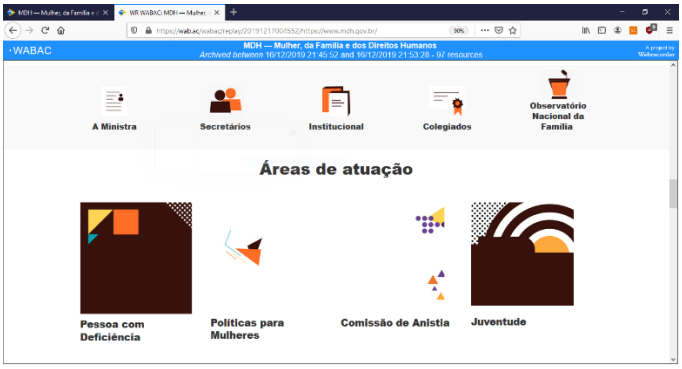
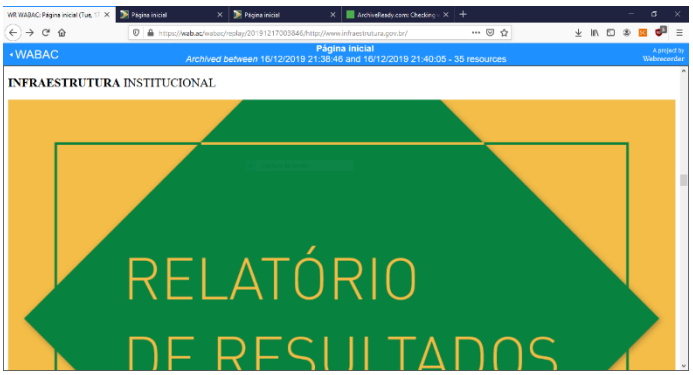


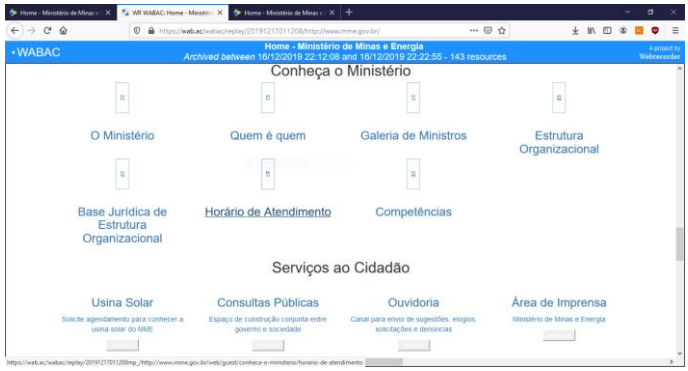
Fonte: Elaborado pelo autor.

IMAGENS ILUSTRATIVAS PARA ÍCONES/*LINKS*

Foi verificado que em 21 (vinte e um) *websites* houve essa ocorrência, sendo que em 12 (doze) a imagem ilustrativa dos ícones foi apresentada intacta em sua versão arquivada; em 02 (dois) a imagem apareceu fora da estrutura original; em 06 (seis) apareceu apenas o cabeçalho ou título no ícone (Quadro 7); e em 01 (um) não houve vestígio das imagens de ícones que, aliás, foi o mesmo que apareceu na mesma escala da categoria acima, o *website* do Banco Central do Brasil (www.bcb.gov.br/), que mostrou somente sua estrutura de cabeçalho na versão arquivada. Os dados são apresentados por meio do Gráfico 6.

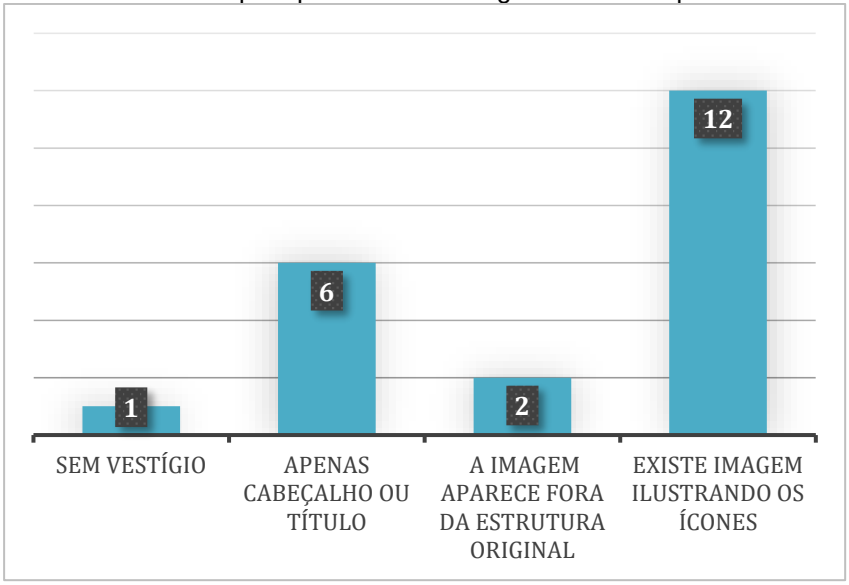
Quadro 7 – Exemplo de *website* arquivado – imagem ilustrativa para ícones/*links*

EXISTE IMAGEM ILUSTRANDO OS ÍCONES	
A IMAGEM APARECE FORA DA ESTRUTURA ORIGINAL	

APENAS CABEÇALHO OU TÍTULO	

Fonte: Elaborado pelo autor.

Gráfico 6 – *Websites* que apresentaram imagem ilustrativa para ícones/*links*

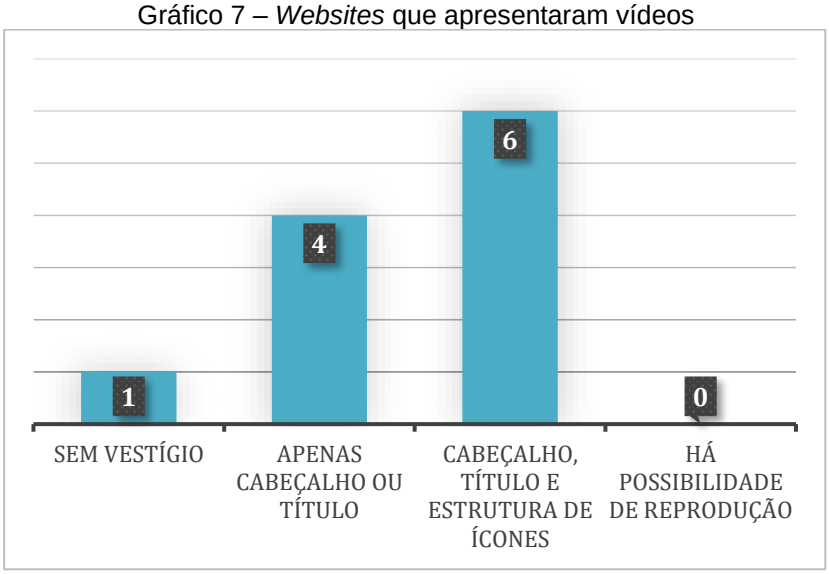


Fonte: Elaborado pelo autor.

VÍDEO

Dos *websites* analisados em suas versões ao vivo, apenas 11 (onze) continham vídeos em seu conteúdo. Ao analisar suas versões arquivadas se observou que destes, apenas 01 (um) não deixou vestígios da presença de vídeo em seu conteúdo; quatro (04) apresentaram apenas cabeçalho ou título; 06 (seis) apresentaram cabeçalho, título e o ícone para o vídeo; e em nenhum dos arquivados houve a possibilidade de reprodução destes vídeos. O Gráfico 7 ilustra os resultados.

Gráfico 7 – *Websites* que apresentaram vídeos

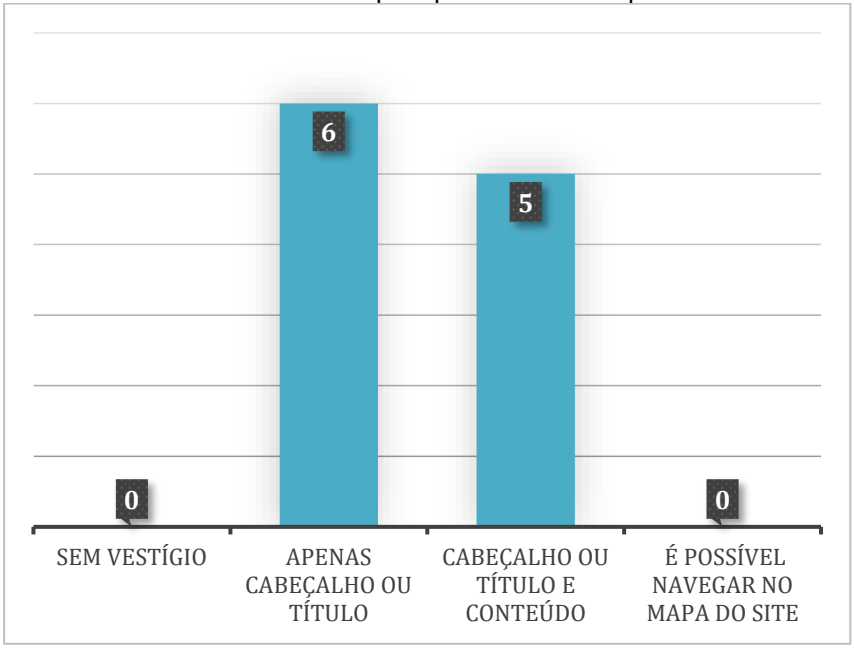


Fonte: Elaborado pelo autor.

MAPA DO SITE

O recurso Mapa do Site, ou em inglês, *Sitemap*, é uma lista com todas as páginas (*URLs*) do site que ajuda o usuário a navegar e encontrar páginas que compõe o portal. Dos 23 (vinte e três) *websites* analisados, apenas 11 (onze) apresentaram o mapa do site. Nas versões arquivadas destes *websites*, 06 (seis) apresentaram apenas o cabeçalho e título; e 05 (cinco) apresentaram cabeçalho, título e conteúdo. Os outros níveis de escalonamento não pontuaram. O Gráfico 8 ilustra os resultados.

Gráfico 8 – *Websites* que apresentaram mapa do site

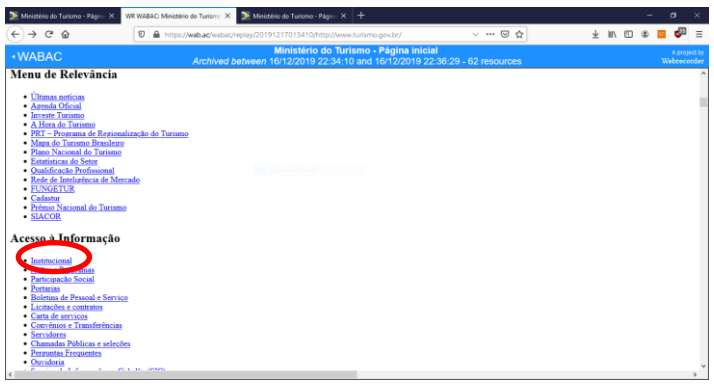


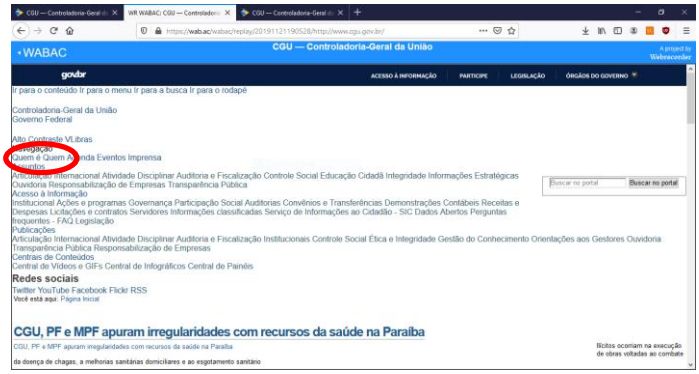
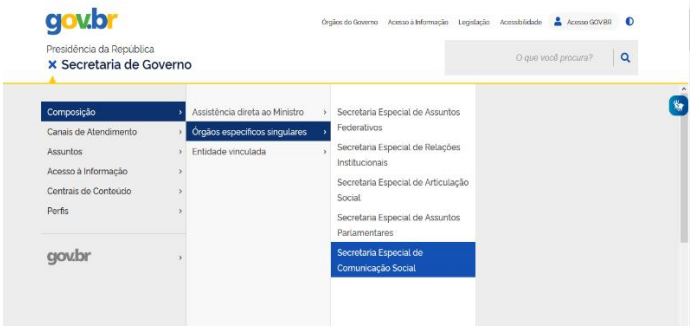
Fonte: Elaborado pelo autor.

ESTRUTURA ORGANIZACIONAL DO ÓRGÃO

Dos *websites* que compuseram o corpo amostral desta pesquisa, 18 (dezoito) apresentaram a estrutura organizacional do órgão. Ao analisar as versões arquivadas destes *websites*, 02 (dois) não deixaram vestígios; 04 (quatro) apresentaram apenas o título do *link*; 08 (oito) apresentaram o título e o *link* para acesso a estrutura administrativa; e 04 (quatro) apresentaram a estrutura completa como no *website* ao vivo. O Quadro 8 mostra alguns exemplos e o Gráfico 9 ilustra os resultados.

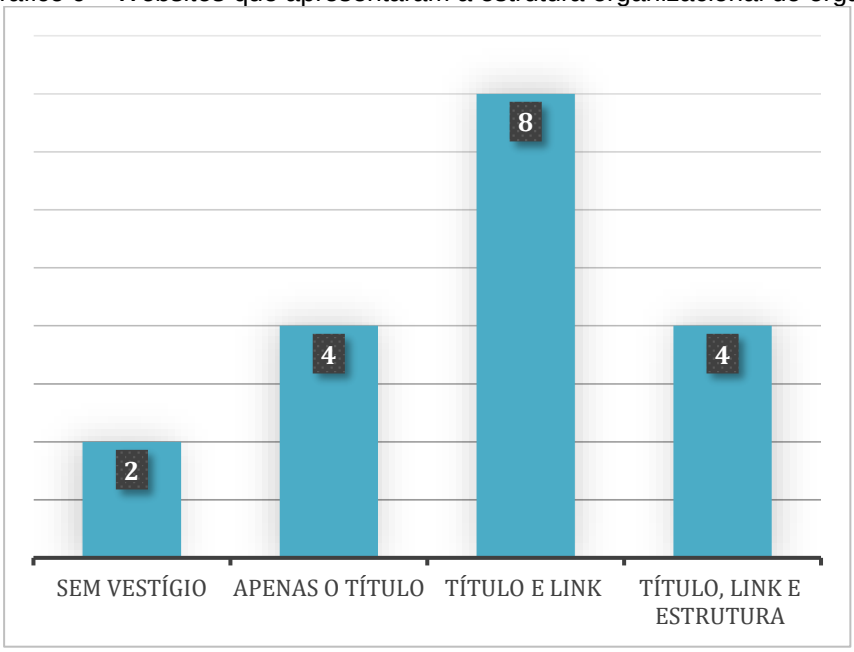
Quadro 8 – Exemplo de *website* arquivado – estrutura organizacional do órgão

APENAS O TÍTULO	

TÍTULO E LINK	
TÍTULO, *LINK* E ESTRUTURA	

Fonte: Elaborado pelo autor.

Gráfico 9 – *Websites* que apresentaram a estrutura organizacional do órgão

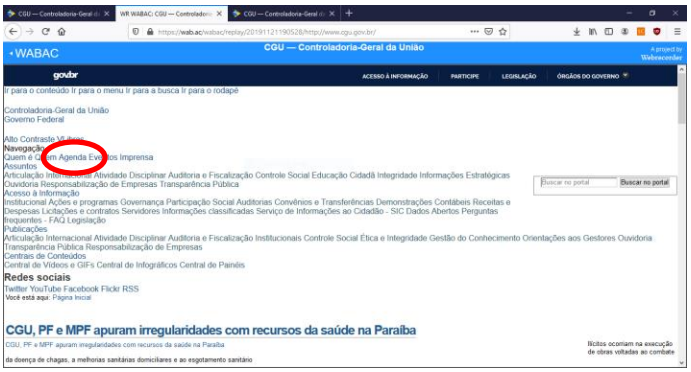
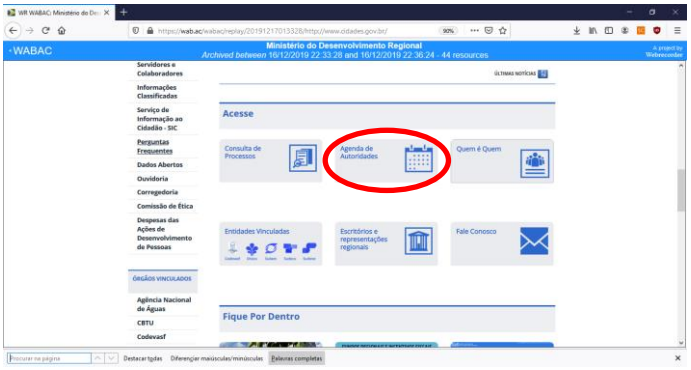
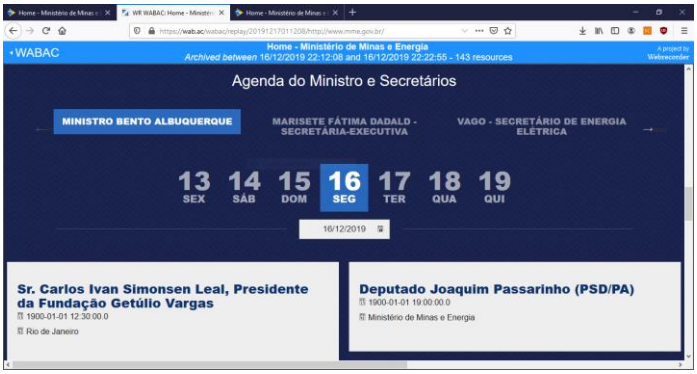


Fonte: Elaborado pelo autor.

INTERATIVIDADE COM AGENDA

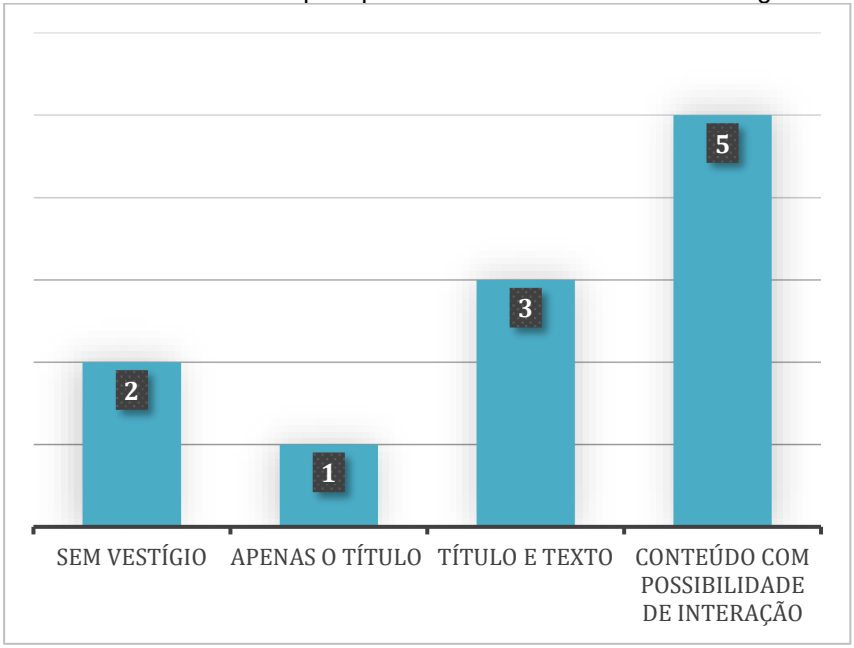
Ao conhecer os *websites*, verificamos que alguns deles apresentavam a agenda dos ministros das respectivas pastas, em tempo real e com possibilidade de interação. Foi verificado que 11 (onze) *websites* apresentavam este recurso, sendo que em 02 (dois) não houve vestígio da existência da agenda após o arquivamento; em 01 (um) apareceu apenas o título da agenda; em 03 (três) foi possível ver o título e o texto da agenda; e em 05 (cinco) houve a possibilidade de interação com a agenda após o arquivamento. O Quadro 9 mostra alguns exemplos e o Gráfico 10 ilustra os resultados.

Quadro 9 – Exemplo de *website* arquivado – interatividade com agenda

APENAS O TÍTULO	
TÍTULO E *LINK*	
TÍTULO, *LINK* E ESTRUTURA	

Fonte: Elaborado pelo autor.

Gráfico 10 – *Websites* que apresentaram interatividade com agenda

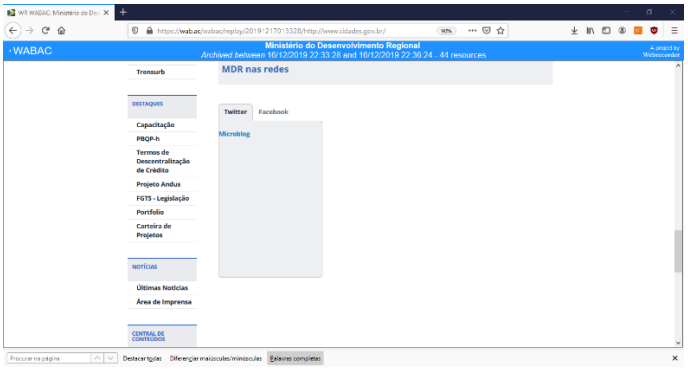
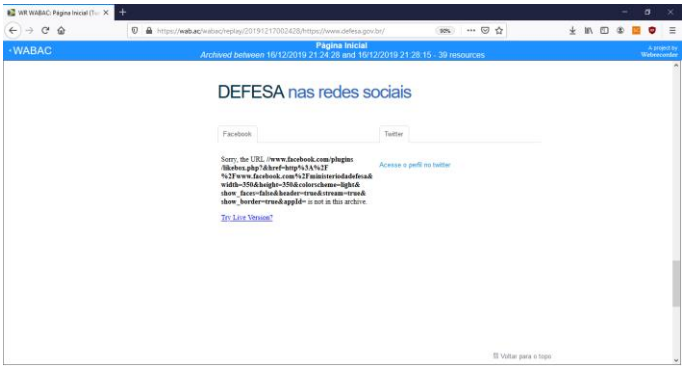


Fonte: Elaborado pelo autor.

INTEROPERABILIDADE COM REDES SOCIAIS

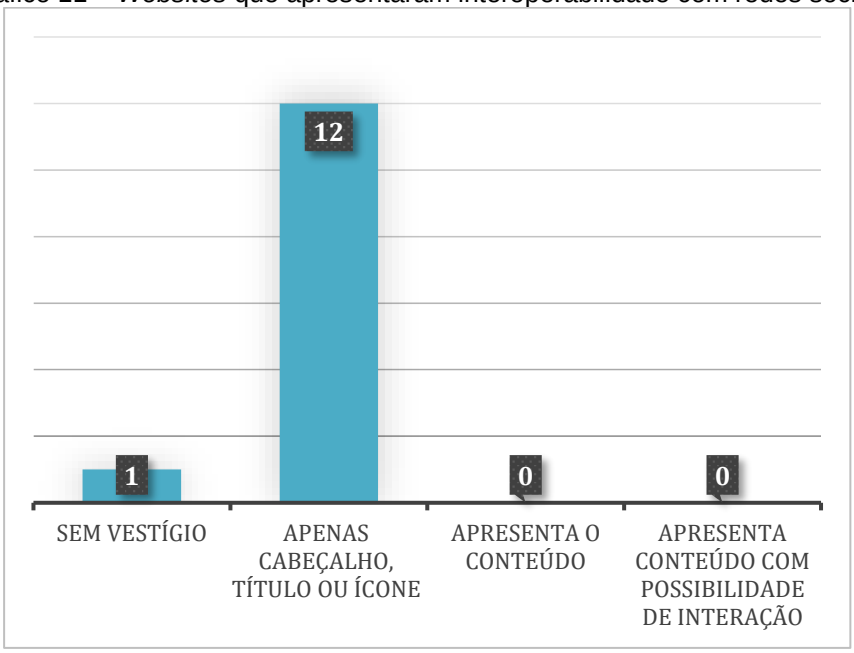
Inicialmente, cabe destacar que o termo *Interoperabilidade* utilizado neste item de análise, é definido pela NISO como a capacidade de sistemas de diferentes plataformas de *hardware* e *software*, estrutura de dados e interfaces, em trocar dados com perda mínima de conteúdo e funcionalidade (NISO, 2017). Considerando este conceito, foi verificado que dos 23 (vinte e três) *websites* que compuseram o corpo amostral desta pesquisa, 13 (treze) apresentaram o recurso de interoperabilidade com redes sociais (especialmente *Facebook* e *Twitter*). Após o arquivamento, observou-se que em 01 (um) não houve vestígio deste recurso; em 12 (dose) apareceu apenas o cabeçalho, título ou ícone da rede social; e em nenhum apareceu o conteúdo ou houve possibilidade de interação, como curtir, comentar ou compartilhar. O Quadro 10 mostra alguns exemplos e o Gráfico 11 ilustra os resultados.

Quadro 10 – Exemplo de *website* arquivado – interoperabilidade redes sociais

APENAS O CABEÇALHO, TÍTULO OU ÍCONE	 

Fonte: Elaborado pelo autor.

Gráfico 11 – *Websites* que apresentaram interoperabilidade com redes sociais

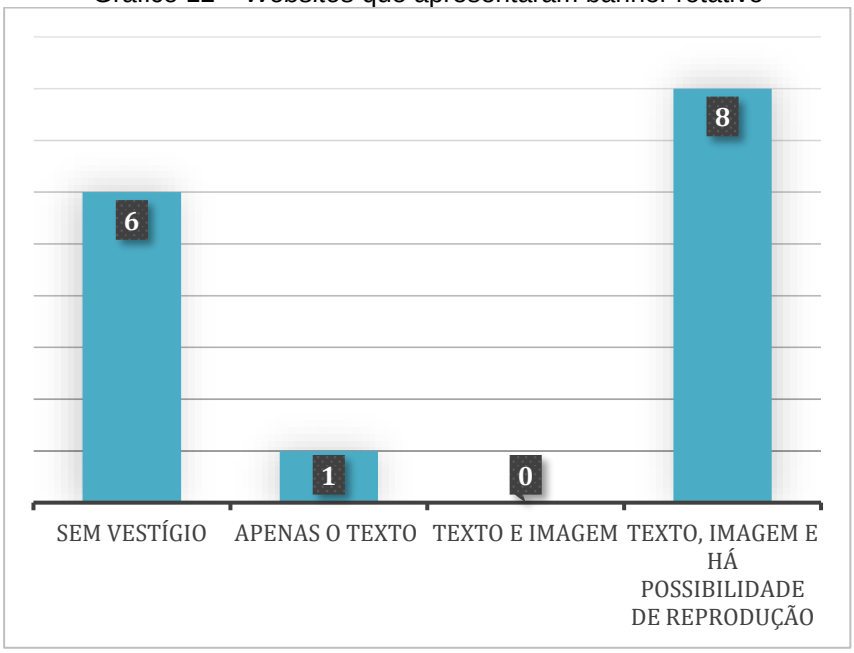


Fonte: Elaborado pelo autor.

BANNER ROTATIVO

Em 15 (quinze) *websites* se verificou o uso de banner rotativo, que é uma ferramenta em que são veiculadas diferentes imagens que se alternam ininterruptamente. Após o arquivamento destes *websites* pode-se perceber que em 06 (seis) não houve registros da existência destes banners; em 01 (um) aparece apenas o texto; e em 08 (oito) foi possível verificar que o banner rotativo permaneceu exatamente como no *website* ao vivo, veiculando o texto e imagem com possibilidade de reprodução. O Gráfico 12 ilustra os resultados.

Gráfico 12 – *Websites* que apresentaram banner rotativo

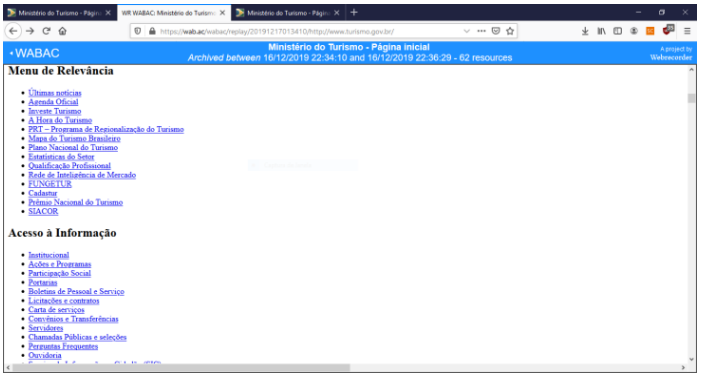
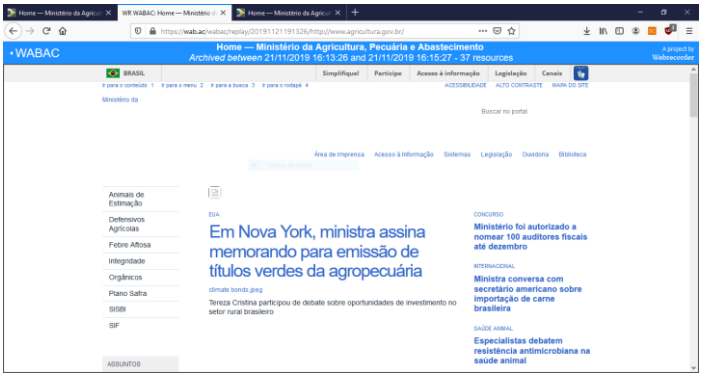
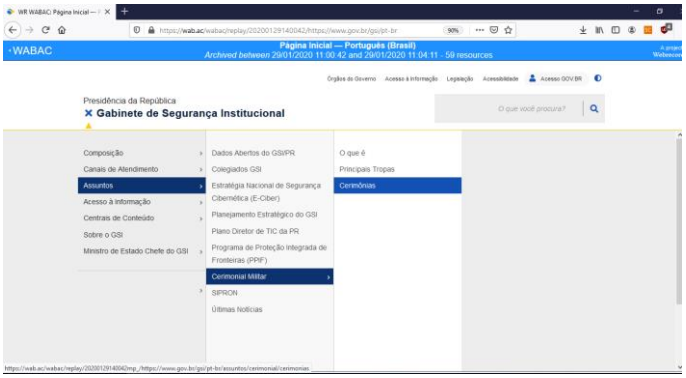


Fonte: Elaborado pelo autor.

MENU DE NAVEGAÇÃO

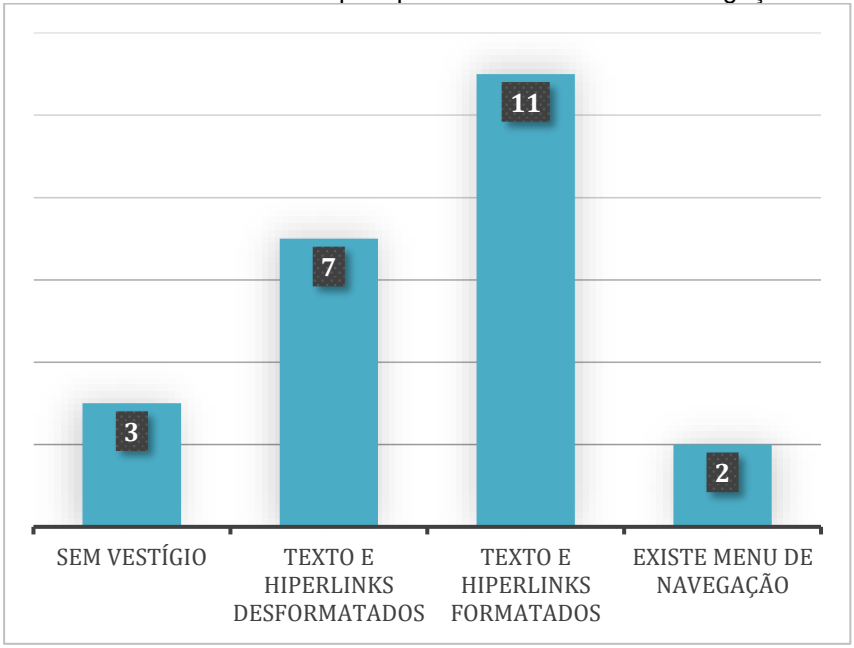
Inicialmente, em todos os 23 (vinte e três) *websites* foi verificado a existência de menu de navegação, sendo que após o arquivamento, em 03 (três) não houve registro da existência destes menus de navegação; em 07 (sete) apareceu apenas os textos e *hiperlinks* desformatados; em 11 (onze) foi possível verificar o texto e *hiperlinks* formatados; e em 02 (dois) foi possível, inclusive, a navegação pelo menu. O Quadro 11 apresenta alguns exemplos e Gráfico 13 ilustra os resultados.

Quadro 11 – Exemplo de *website* arquivado – menu de navegação

TEXTOS E *HIPERLINKS* DESFORMATADOS	
TEXTOS E *HIPERLINKS* FORMATADOS	
EXISTE MENU DE NAVEGAÇÃO	

Fonte: Elaborado pelo autor.

Gráfico 13 – *Websites* que apresentaram menu de navegação

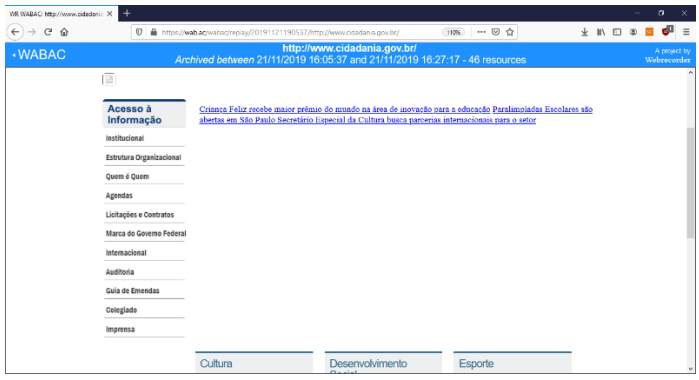


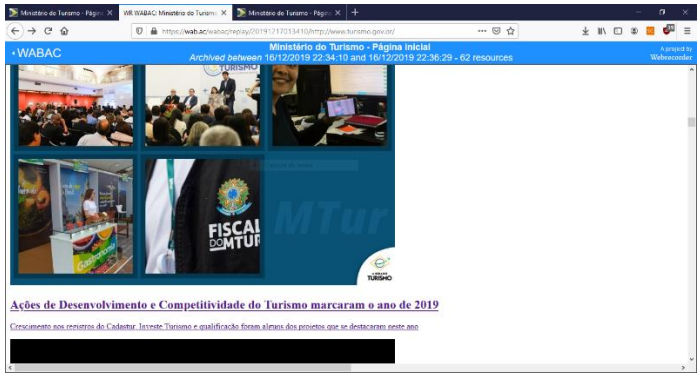
Fonte: Elaborado pelo autor.

APRESENTAÇÃO DO *FEED* DE NOTÍCIAS

Dos 21 (vinte e um) *websites* que apresentavam *feed* de notícias, foi verificado que 01 (um) não apresentou vestígios após seu arquivamento; 05 (cinco) apareceu apenas texto e *links*; em outros 05 (cinco) apareceu texto, *links* e imagens, porém desformatadas; e em 10 (dez) texto, *link* e imagens formatadas. O Quadro 12 apresenta alguns exemplos e Gráfico 14 ilustra os resultados.

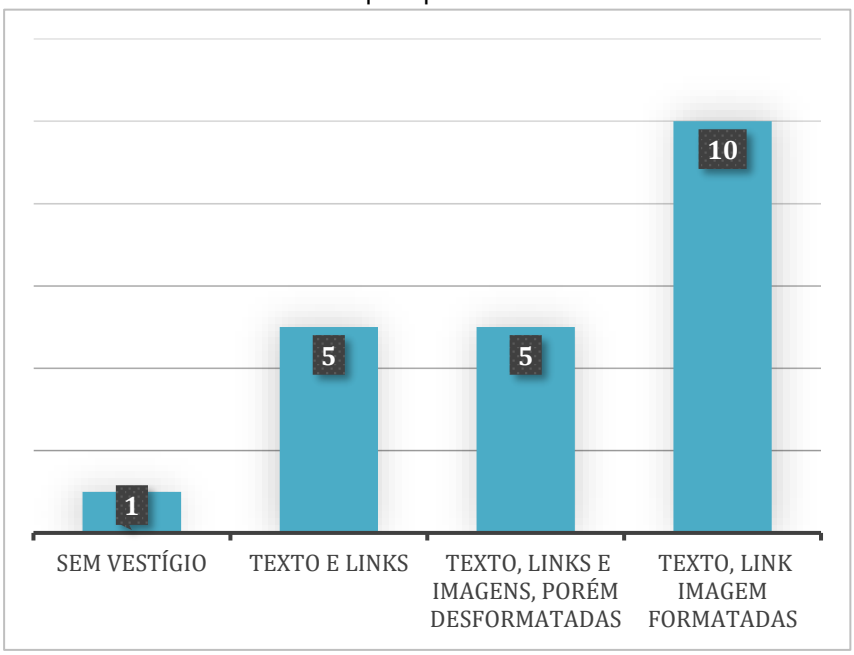
Quadro 12 – Exemplo de *website* arquivado – *feed* de notícias

TEXTO E *LINKS*	

TEXTO, *LINKS* E IMAGENS DESFORMATADAS	
TEXTO, *LINKS* E IMAGENS FORMATADAS	

Fonte: Elaborado pelo autor.

Gráfico 14 – *Websites* que apresentaram *feed* de notícias



Fonte: Elaborado pelo autor.

Para o recurso **Acompanhamento em tempo real (horário, previsão do tempo, variação cambial, mapa, etc)** apenas 01 (um) *website* o apresentou em sua versão ao vivo, porém este recurso não foi recuperado no momento do

arquivamento. Já o recurso **Transmissão *online* em tempo real de eventos como sessões, palestras e reuniões** não apresentou resultados, considerando que em nenhum dos *websites* analisados apresentou este recurso.

Depois de analisar individualmente os dados recolhidos em cada um dos recursos, foi elaborado um quadro com a intenção de verificar a pontuação total que cada *website* apresentou em sua versão ao vivo e na sua versão arquivada, considerando a soma das pontuações atribuídas para cada recurso. Deste total, foi verificada a diferença de pontuações entre as versões dos *websites*, em que intervalos numéricos maiores significam que possuem mais distinções entre si; por outro lado, para intervalos menores entre as versões, significa que os *websites* apresentam menos distinções entre sua versão ao vivo e arquivada. Considerando que o *website* X somou 23 pontos em sua versão ao vivo e 22 pontos em sua versão arquivada, a diferença entre os dois *websites* seria de 1 ponto e, portanto, os *websites* se assemelham entre si; se caso a pontuação da versão arquivada deste mesmo *websites* tivesse somado 7 pontos, por exemplo, significaria que a diferença entre suas versões seria de 16 pontos (23-7=16) e, portanto, suas diferenças seriam evidentes.

Após a constatação desta lógica atribuída à diferença de pontuação entre as versões, se notou que havia uma coerência com a pontuação atribuída ao recurso "*Layout* se manteve", presente na relação de análises previamente apresentadas. Essa coerência pode ser observada se compararmos os dados dispostos no Quadro 13, a seguir:

Quadro 13 – Sistematização da pontuação dos *websites*

***Website* coletado**	**Versão**	**Pontuação total em cada versão**	**Diferença de pontuação entre as versões**	**Comparação com a análise "*Layout* se manteve"**
www.gov.br	ao vivo	21	9	2
	arquivado	12		
www.agu.gov.br	ao vivo	30	9	2
	arquivado	21		
www.bcb.gov.br	ao vivo	25	25	1
	arquivado	0		

Website coletado	Versão	Pontuação total em cada versão	Diferença de pontuação entre as versões	Comparação com a análise "*Layout* se manteve"
https://www.gov.br/casacivil	ao vivo	21	2	3
	arquivado	19		
https://www.gov.br/gsi	ao vivo	27	2	3
	arquivado	25		
www.cgu.gov.br	ao vivo	24	15	1
	arquivado	9		
www.agricultura.gov.br	ao vivo	33	18	1
	arquivado	15		
www.cidadania.gov.br	ao vivo	15	11	1
	arquivado	4		
www.mctic.gov.br	ao vivo	30	7	3
	arquivado	23		
www.defesa.gov.br	ao vivo	33	8	3
	arquivado	25		
www.economia.gov.br	ao vivo	24	5	3
	arquivado	19		
www.mec.gov.br	ao vivo	6	1	3
	arquivado	5		
www.infraestrutura.gov.br	ao vivo	24	15	1
	arquivado	9		
www.justica.gov.br	ao vivo	18	13	1
	arquivado	5		
www.mdh.gov.br	ao vivo	19	3	3
	arquivado	16		
www.saude.gov.br	ao vivo	27	10	2
	arquivado	17		
www.itamaraty.gov.br	ao vivo	24	16	1
	arquivado	8		
www.mme.gov.br	ao vivo	25	10	2
	arquivado	15		

Website coletado	Versão	Pontuação total em cada versão	Diferença de pontuação entre as versões	Comparação com a análise "*Layout* se manteve"
www.cidades.gov.br	ao vivo	30	7	3
	arquivado	23		
www.mma.gov.br	ao vivo	30	21	1
	arquivado	9		
www.turismo.gov.br	ao vivo	33	21	1
	arquivado	12		
www.gov.br/secretariadegoverno	ao vivo	21	2	3
	arquivado	19		
www.gov.br/secretariageral	ao vivo	24	2	3
	arquivado	22		

Fonte: Elaborado pelo autor.

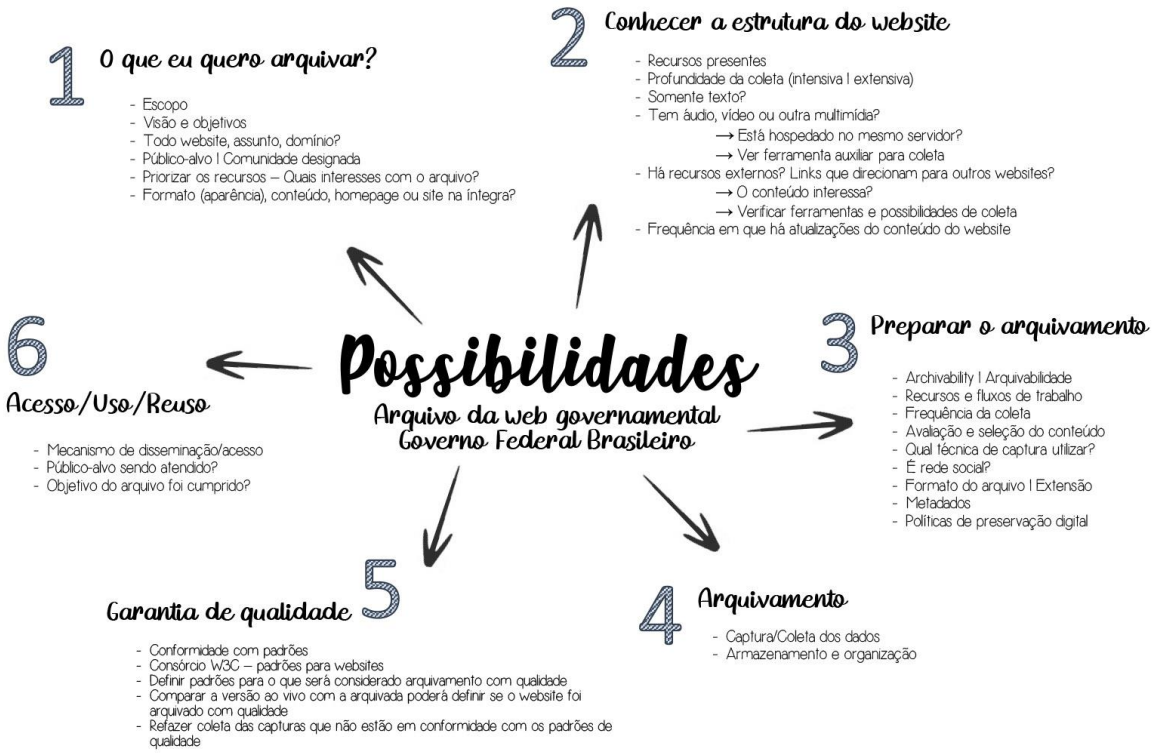
Pode-se perceber que há uma lógica recorrente associada a essa comparação: as diferenças de pontuação entre as versões do mesmo *website*, quando de 1 a 8 pontos, representam os *websites* que mantiveram seu *layout* após o arquivamento (sendo atribuídos 3 pontos); diferenças de 9 a 10 pontos, compreendem os *websites* que mantiveram seu *layout* em partes (sendo atribuídos 2 pontos); e pontuações de 11 a 25, correspondem os *websites* que não mantiveram seu *layout* após o arquivamento (com atribuição de 1 ponto). Portanto, podemos afirmar que quanto maior a diferença entre a pontuação das versões ao vivo e arquivada de um mesmo *website*, menos recursos foram arquivados e, neste caso, a qualidade do arquivamento foi baixa. Ao mesmo tempo, quando menor a diferença da pontuação das duas versões, melhor qualidade teve o arquivamento.

Para finalizar a análise dos dados, elaboramos um Mapa mental[52] que sistematiza o processo que poderá ser seguido para realizar o arquivamento da *web* governamental no Brasil, demonstrando suas possibilidades. O mapa mental apresenta uma combinação das teorias do *Modelo de Ciclo de Vida do*

[52] Mapa mental (ou mapa da mente) é um tipo de diagrama voltado para a sistematização de dados, informações, conhecimentos; para a compreensão e solução de problemas; na memorização e aprendizado; na criação de manuais, livros e palestras; como ferramenta de brainstorming; e no auxílio da gestão estratégica de uma empresa ou negócio. Foi idealizado pelo psicólogo inglês Tony Buzan.

Arquivamento da Web (BRAGG; HANNA, 2013) e da A*bordagem sistemática para a preservação da web* (KRAN; RAHMAN, 2019), associado as anotações feitas pelo autor durante o desenvolvimento da pesquisa. O mapa mental é apresentado na Figura 18, a seguir.

Figura 18 – Mapa mental das possibilidades de arquivamento dos websites governamentais



Fonte: Elaborado pelo autor.

Por fim, passamos a discorrer a respeito das considerações finais desta pesquisa.

5 CONSIDERAÇÕES FINAIS

A partir das informações levantadas com o desenvolvimento desta pesquisa, somados aos aprendizados trazidos pelo referencial teórico, passamos a apresentar as considerações finais desta investigação. Inicialmente, cabe destacar que o problema de pesquisa foi originado a partir da necessidade em abordar o arquivamento da *web* como um recurso capaz de manter o acesso aos registros digitais, e entende-lo como um instrumento de preservação do conteúdo publicado na *web*. Associado a isso, foram utilizados os *websites* das instituições que compõem o primeiro escalão do Poder Executivo do Governo Federal Brasileiro, tais como os ministérios, secretarias e órgãos com status de ministério, como objeto empírico, que nos auxiliaram a demonstrar as possibilidades de arquivamento da *web* na esfera federal, a partir de um estudo de caso do domínio gov.br.

Foram utilizados 23 (vinte e três) *websites* que consideramos suficientes para cumprir com o objetivo desta pesquisa. O escopo amostral foi objeto de análise para os objetivos específicos que eram, além de selecionar e verificar estes *websites*, identificar os recursos disponibilizados nas versões ao vivo dos *websites* selecionados; realizar o arquivamento dos *websites* selecionados, com o uso de rastreador de páginas *web* automatizado; reconstruir os *websites* arquivados com o uso de *software* automatizado; e identificar os recursos disponibilizados nas versões arquivadas dos *websites* selecionados. Com o cumprimento destes objetivos específicos conforme o estabelecido nos procedimentos metodológicos, chegamos ao objetivo principal da pesquisa que era demonstrar as possibilidades de arquivamento de *websites* do Governo Federal Brasileiro a partir de um estudo de caso do domínio gov.br.

Do referencial teórico foram extraídos 16 (dezesseis) recursos, que deram origem ao instrumento de sistematização da análise. Cada um dos 23 (vinte e três) *websites* tiveram suas versões ao vivo comparadas com sua versão arquivada, de modo a identificar quais recursos foram arquivados e em que condições, a partir de uma escala de ocorrência. Consideramos que a metodologia utilizada foi satisfatória para a realização dos procedimentos e auxiliou para a resolução do problema de pesquisa.

Ainda que definir medidas para avaliar o sucesso de uma coleção de arquivos da *web* seja difícil de implementar, considerando que os indicadores quantitativos nem sempre são apropriados, assim como conceitos menos tangíveis, como qualidade e valor, não são fáceis de definir e denotam um grau de julgamento, algumas tentativas de medição são necessárias para auxiliar na visualização das possibilidades de arquivamento dos *websites* do Governo Federal. Para isso, consideramos que o comportamento e as funcionalidades dos *websites* coletados foram avaliados como parte do processo de revisão de qualidade, verificando até que ponto os elementos interativos (*links*, menus, etc) funcionaram como no original e como os *websites* arquivados condizem com o *website* ao vivo em relação as imagens, *links* ou botões, por exemplo. Essas questões foram analisadas para cada um dos 16 (dezesseis) recursos previamente estipulados. Da mesma forma, entendemos que ao determinar se um *website* é recuperado com sucesso, deve-se dar ênfase na captura do conteúdo intelectual e não necessariamente na aparência do *website*. Contanto que o conteúdo de um site seja capturado e possa ser reproduzido de maneira razoável na ferramenta de acesso, a versão coletada será entendida como de qualidade.

Em relação a verificação do *layout* dos *websites*, pode-se perceber que do corpus amostral, 13 (treze) ainda estavam com seu *layout* antigo e não passaram por nenhuma transformação em relação a sua estética; por outro lado, 10 (dez) *websites* estão com o novo *layout* dos portais do Governo Federal, sendo que 04 (quatro) já haviam sido incorporados no Portal Único gov.br, que pretende reunir todos *websites* do Poder Executivo do Governo Federal em um único portal. Estes resultados mostram o potencial que essa investigação nos proporcionou: o de verificar a potencialidade de arquivamento destes *websites* em diferentes parâmetros e em um momento de transição. O resultado foi positivo quando percebemos que mais da metade dos *websites* utilizados como objetos desta pesquisa tiveram seus recursos arquivados, ainda que uma quantidade relevante não tenha recuperado alguns dos recursos disponibilizados, especialmente, quando o formato do arquivo era diferente do textual e da imagem estática. Para áudios e vídeos, por exemplo, percebemos que o tempo de coleta ou a ferramenta WABAC, utilizada para reconstrução dos arquivos WARC capturados, não foram suficientes para recuperar estes arquivos. Recursos hospedados em páginas diferentes das

arquivadas, também tenderam a não recuperação, tais como instrumentos de medição de câmbios monetários, de previsão do tempo e a interoperabilidade com as redes sociais, por exemplo. Essas falhas podem ter ocorrido por diversos fatores como, por exemplo, o fato destes recursos estarem hospedados em servidores diferentes do que está o *website*, originalmente. Entendemos, portanto, que se faz necessário o uso de alguma ferramenta auxiliar para recuperação destes documentos não textuais, tais como áudios e vídeos, além de outros estudos empíricos para compreender as necessidades e melhores ferramentas para a recuperação destes arquivos.

Em relação à recuperação do conteúdo textual, de imagens ilustrativas, sejam das notícias ou dos ícones, e da permanência do *feed* de notícias, percebemos que a recuperação foi satisfatória quando os resultados mostram que a maioria dos *websites* arquivados apresentam estes conteúdos de forma integral, ainda que alguns não estejam formatados visualmente, tal como o *website* ao vivo. Em relação ao mapa do site, a estrutura organizacional, a interatividade com agenda, o banner rotativo e o menu de navegação verificamos que os resultados da recuperação não foram tão satisfatórios. A visualização destes recursos poderia ter sido melhor se a coleta tivesse sido realizada sem a atribuição de tempo específico, considerando que estes recursos exigem o aprofundamento de camadas de *links* no *website*. Mesmo que a pesquisa não tenha definido a quantidade de aprofundamento de *links* na hora da coleta, o tempo de 20 minutos pode não ter sido suficiente para a recuperação destes recursos. Por essa razão, atribuímos o resultado ao fato do pouco tempo em que o rastreador foi programado para realizar a coleta.

Os recursos como a Ferramenta de busca dos conteúdos do *website*, o acompanhamento em tempo real e a transmissão *online* em tempo real não foram recursos que tiveram sucesso nos resultados desta pesquisa. Seja em função dos *websites* ao vivo não apresentarem estas ferramentas ou, como no caso da ferramenta de busca, ser um recurso que não funciona em *websites* arquivados. Para garantir essa informação, fizemos buscas em *websites* governamentais hospedados no *Internet Archive*; garantimos que o website consultado recuperou mais de uma profundidade de coleta e tentamos fazer uma pesquisa no campo

“Busca”. A rápida experiência foi suficiente para mostrar que este é um recurso que, de fato, não funciona em *websites* arquivados.

Ao somar as pontuações atribuídas aos *websites*, verificar a diferença de pontuação entre as versões e comparar os resultados com o recurso “*Layout* se manteve”, nos permite afirmar que a permanência dos recursos dos *websites* após seu arquivamento estabelece uma relação com a garantia de qualidade do arquivamento, ao passo que quanto mais recursos forem arquivados, mais qualidade teve o rastreamento. Portanto, atribuição de medidas para garantia de qualidade são necessárias para definir o sucesso de uma coleta e os indicadores podem ser definidos pelas instituições com base em sua política e objetivos de desenvolvimento de coleções.

Após a conclusão da revisão de qualidade, o arquivista tem três opções: endossar uma coleta, se considerar que seu conteúdo é suficiente para acesso público; rejeitá-la, ponto em que os arquivos serão excluídos; ou rastreá-la novamente com parâmetros de rastreamento ajustados (BINGHAM, 2014). Em geral, uma coleta de *website* será rejeitada se o conteúdo principal não puder ser acessado ou estiver incompleto; o rastreador sair do escopo desejado; ou o *website* não é processado corretamente e pode ser melhorado ajustando as configurações do rastreador. Considerando os resultados desta pesquisa, entendemos que de todo o escopo amostral arquivado, apenas o *websites* do Banco Central do Brasil (www.bcb.gov.br/) não seria disponibilizado em uma plataforma de acesso, considerando que não recuperou, minimamente, seu conteúdo. Acreditamos que o *website* do Banco Central possa ter alguma configuração feita em seu desenvolvimento para que intervenções não possam ser acontecer em razão de ser um portal com dados sensíveis. Os demais *websites* arquivados, ainda que apresentem falhas em alguns recursos, todos tiveram seu conteúdo informacional coletado. Neste sentido, entendemos que o problema da pesquisa foi solucionado ao demonstrarmos, a partir deste estudo de caso, quais são as possibilidades de arquivamento da *web* na esfera federal, mesmo que ajustes como, por exemplo, o tempo de coleta, poderia ser considerado dependendo do que se pretende com a formação do arquivo da *web* governamental no Brasil.

Por fim, desenvolvemos o mapa mental que sistematiza as possibilidades de arquivamento dos *websites* do Governo Federal Brasileiro. O mapa busca reunir as duas principais teorias apresentadas nesta pesquisa: o *Modelo de Ciclo de Vida do Arquivamento da Web* (BRAGG; HANNA, 2013) e a *Abordagem sistemática para a preservação da web* (KRAN; RAHMAN, 2019). Estruturamos as fases apresentadas pelas teorias em seis momentos, associando com anotações realizadas durante o desenvolvimento da pesquisa, que poderá ser útil para quem deseja realizar o arquivamento da *web* governamental no Brasil, assim como para a formação de outros arquivos *web*, resguardando as especificações estabelecidas para arquivos *web* governamentais.

Para além dos requisitos técnicos aprendidos com essa pesquisa, cabe destacar nas considerações finais, o papel do arquivista neste cenário de mudanças paradigmáticas na arquivística, ou ainda, na mudança do fazer profissional, partindo do pressuposto que envolve tornar acessível as informações ao usuário e que, ao desenvolver atividades como identificar, analisar, avaliar e disseminar a informação, o arquivista está, mesmo que implicitamente, estabelecendo uma relação com os usuários e, portanto, se percebe a necessidade de uma mudança na forma de agir do profissional da informação, não em relação ao seu objeto de trabalho, que segue sendo a informação em qualquer de seus formatos, mas em relação a quantidade e a velocidade que se faz necessária para seu gerenciamento, armazenamento e, especialmente, sua recuperação. Essa mudança de paradigma trouxe novas exigências para os profissionais que atuarão com o gerenciamento de dados: “Não se trata apenas de algo a mais na formação do profissional da informação, mas em uma mudança exponencial na forma como esses profissionais lidam com dados, informações, documentos e tecnologia” (MELO; ROCKEMBACH, 2019, p. 15-16).

Outro aspecto aprendido com esta pesquisa foi o estabelecimento de uma relação entre o arquivamento da *web*, com a produção governamental no ambiente *web*. É sabido que coletar, preservar e fornecer ferramentas para uma ampla gama de registros requer investimentos e experimentações consideráveis. Existe um claro compromisso do governo em melhorar os serviços digitais oferecidos e a coleta desses serviços continuará a ser um desafio para a tecnologia de arquivamento na *web*. Recomendamos que estas melhorias considerem os padrões estipulados pelo consórcio W3C, considerando que a partir destes padrões é que serão

desenvolvidas ferramentas e metodologias para arquivamento da *web*, possibilitando melhores resultados nos arquivamentos, à medida que mais documentos também estão sendo produzidos em páginas da *web*.

Deste modo, estudar e analisar a tecnologia e os procedimentos para coleta de páginas *web* pode ajudar a criar uma maneira própria de capturar documentos criados em páginas da *web*, sem negligenciar o patrimônio documental analógico que deve continuar sendo tratado, custodiado e dado acesso. Não se trata de puro investimento tecnológico, mas de reconfigurar os modos de gestão dos documentos digitais. Arquivistas, cientistas da informação e profissionais voltados à tecnologia devem ser responsáveis por liderar e conduzir essas mudanças organizacionais, e por isso que é imperativo que estes profissionais aprimorem suas competências tecnológicas para melhorar os novos sistemas de produção digital de documentos.

REFERÊNCIAS

ABBATE, Janet. **Inventing the Internet.** Massachusets: MIT, 1999.

AFONSO, Carlos Alberto. The Internet and the community in Brazil: background, issues and options. **IEEE Communications Magazine**, v. 34, n. 7, Jul. 1996, p. 62-68.

AFONSO, Carlos Alberto. Internet no Brasil: o acesso para todos é possível? **Policy Paper**, n. 26, 2000. Disponível em: <https://docplayer.com.br/520918-Internet-no-brasil-o-acesso-para-todos-e-possivel.html>. Acesso em: 20 jul. 2019.

ARCHIVE-IT. FAQ. 2011.

ARCHIVE READY. **Website** [*online*], 2014. Disponível em: <http://archiveready.com/>. Acesso em: 24 ago. 2019.

ARQUIVO NACIONAL DO BRASIL. **Política de preservação digital.** Versão 2. 2016. Disponível em: <http://www.siga.arquivonacional.gov.br/images/an_digital/and_politica_preservacao_digital_v2.pdf> . Acesso em: 14 fev. 2019.

ARVIDSON, Allan. Kulturarw3. **Preserving the Present for the future**. In: Conference on strategies for the Internet, proceedings, Copenhagen, jun./2001. p. 101-104.

BERNERS-LEE, Tim; CAILLIAU, Robert; LUOTONEN, Ari; NIELSEN, Henrik Frystyk; SECRET, Arthur. The World-wide *web*. **Communication of the ACM**, v. 37, n. 8, Aug./1994.

BINGHAM, Nicola. Quality Assurance Paradigms in *Web* Archiving Pre and Post Legal Deposit. **Alexandria**, v. 25, n. 1/2, 2014. http://dx.doi.org/10.7227/ALX.0020

BOURDELOIE, H. Ce que le numérique fait aux sciences humaines et sociales. **tic&société**, v. 7, n. 2, 2013. Disponível em: <https://ticetsociete.revues.org/1500#text. Acesso em: 23. jul. 2019.

BRAGG, Molly; HANNA, Kristine; DONOVAN, Lori; HUKILL, Graham; PETERSON, Anna. **The *Web* Archiving Life Cycle Model.** WhitePaper. 2013. Disponível em: http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. Acesso em: 14 fev. 2019.

BRASIL. **Constituição da República Federativa do Brasil.** Brasília, DF: Senado Federal: Centro Gráfico, 1988. 292 p.

BRASIL. Portaria Interministerial nº 147, de 31 de maio de 1995. **Diário Oficial [da] República Federativa do Brasil,** Poder Executivo, Brasília, DF, 31 maio 1995.

BRASIL. Câmara dos Deputados. **Projeto de Lei 2431/2015**. Dispõe sobre o patrimônio público digital institucional inserido na rede mundial de computadores e dá outras providências. 2015. Disponível em: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=1594241>. Acesso em: 02 dez. 2019.

BRASIL. Decreto nº 9.660, de 1º de janeiro de 2019. Dispõe sobre a vinculação das entidades da administração pública federal indireta. **Diário Oficial [da] República Federativa do Brasil,** Poder Executivo, Brasília, DF, 01 Jan. 2019a.

BRASIL. Decreto nº 9.756, de 11 de abril de 2019. Institui o portal único “gov.br” e dispõe sobre as regras de unificação dos canais digitais do Governo federal. **Diário Oficial [da] República Federativa do Brasil,** Poder Executivo, Brasília, DF, 11 Abr. 2019b.

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

BRÜGGER, Niel. **Archiving Websites:** General Considerations and Strategies. Denmark: The Centre for Internet Research, 2005.

BRÜGGER, Niel. Digital Humanities in the 21st Century: Digital Material as a Driving Force. **Digital Humanities Quarterly**, v. 10, n. 2, 2016.

BRÜGGER, Niels (Ed.). ***Web* 25:** histories from the first 25 years of the World Wide *Web*. Switzerland: Peter Lang. 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.

CADAVID, Jhonny Antonio Pabón. Copyright Challenges of Legal Deposit and *Web* Archiving in the National Library of Singapore. **Alexandria**, v. 25, n. 1/2, 2014. http://dx.doi.org/10.7227/ALX.0017.

CARVALHO, Marcelo Sávio Revoredo Menezes de. **A trajetória da Internet no Brasil:** do surgimento das redes de computadores à instituição dos mecanismos de governança. 2006. 259 f. Dissertação (Mestrado em Ciências de Engenharia de Sistemas e Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

CARVALHO, Marcelo Sávio Revoredo Menezes de; CUKIERMAN, Henrique Luiz. **Os primórdios da Internet no Brasil.** 200-, p. 14. Disponível em: < http://www.nethistory.info/Resources/Os%20primordios%20da%20Internet%20no%20Brasil.pdf>. Acesso em: 20 jul. 2019.

CASTELLS, Manuel. **A Sociedade em Rede**. A Era da Informação: economia, sociedade e cultura. Lisboa: Fundação Calouste Gulbenkian, 2000. Disponível em: <https://www.researchgate.net/publication/322155696_Castells_M_2002_A_Era_d

a_Informacao_Economia_Sociedade_e_Cultura_Vol_I_A_Sociedade_em_Rede_Lisboa_Fundacao_Calouste_Gulbenkian_Castells_M_2003_A_Era_da_Informacao_Economia_Sociedade_e_Cultura_Vol_II_O>. Acesso em: 23 jun. 2019.

CASTELLS, Manuel. **A Galáxia internet.** Reflexões sobre Internet, Negócios e Sociedade. Lisboa: Fundação Calouste Gulbenkian, 2001.

CATHRO, Warwick; WEBB, Colin; WHITING, Julie. Archiving the *web*: The PANDORA archive at the National Library of Australia. **Preserving the Present for the future**. In: Conference on strategies for the Internet, proceedings, Copenhagen, jun./2001. p. 105-118.

CETIC.BR. **Website** [*online*], 2019. Disponível em: < https://cetic.br/>. Acesso em: 23 jul. 2019.

CGI.br. **Website** [*online*], 2017. Pesquisa TIC Governo Eletrônico. Disponível em: <https://cetic.br/pesquisa/governo-eletronico/>. Acesso em: 23 jul. 2019.

CGI.br. **Website** [*online*], 2018. Pesquisa TIC Domicílios 2018 - Individuos. Disponível em: <https://cetic.br/tics/domicilios/2018/individuos/>. Acesso em: 23 jul. 2019.

COMERFORD, Richard. The *Web*: A Two-Minute Tutorial. **IEEE Spectrum**, v. 32, n. 71, 1995.

CONGRESSO NACIONAL. **Website** [*online*], 2019. Medida Provisória 870/2019. Estabelece a organização básica dos órgãos da Presidência da República e dos Ministérios, definindo suas competências e sua estrutura básica. Disponível em: < https://www.congressonacional.leg.br/materias/medidas-provisorias/-/mpv/135064>. Acesso em: 8 ago. 2019.

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

DAY, Michael. **Preserving the fabric of our lives:** a survey of *web* preservation initiatives. In: International Conference on Theory and Practice of Digital Libraries, 7. 2003. Berlin. Berlin: Springer-Verlag, 2003a.

DAY, Michael. **Collecting and preserving the World Wide *Web*:** a feasibility study undertaken for the JISC and Wellcome Trust. United Kingdon: UKOLN; University of Bath, 2003b.

DAY, Michael. The long-term preservation of *web* content. In: MANESÈS, Julien (Org.). ***Web* archiving.** Berlin: Springer, 2006. Cap. 8, p. 177-199.

DELALANDE, Nicolas; VINCENT, Julien. Portrait de l'historien-ne en cyborg. **Revue d'histoire moderne et contemporaine**, v. 5, n. 58-4bis, p. 5-29. 2011.

DELLAVALLE, Robert; HESTER, Eric J.; HEILIG, Lauren F.; DRAKE, Amanda L.; KUNTZMAN, Jeff W.; GRABER, Marla; SCHILLING, Lisa M. Information Science

Going, Going, Gone: Lost Internet References. **Science**, New York, n. 302, v. 5646, p. 787-788, 2003. Disponível em: <https://www.researchgate.net/publication/9030760_Going_Going_Gone_Lost_Internet_References>. Acesso em: 02 ago. 2019.

DIKWATTA, Umanda; DIAS, Gihan. *Web* Archiving for Sri Lanka. National Information Technology Conferente (NITC), 13-15 Sep., Colombo, Sri Lanka, 2017. **IEEE**, 2017.

DODEBEI, Vera Doyle. MEMORIA E PATRIMÔNIO perspectivas de acumulação/dissolução no ciberespaço. **Aurora. Revista de Arte, Mídia e Política**, [S.l.], n. 10, p. p. 36, jan. 2011. ISSN 1982-6672. Disponível em: <https://revistas.pucsp.br/aurora/article/view/4614>. Acesso em: 03 dez. 2019.

DODEBEI, Vera. Patrimônio e memória digital. **Revista Morpheus - Estudos Interdisciplinares em Memória Social**, [S.l.], v. 5, n. 8, mar. 2015. Disponível em: <http://www.seer.unirio.br/index.php/morpheus/article/view/4759>. Acesso em: 16 fev. 2019.

EOT - End of Term *Web* Archive. **Website** [*online*], 2019. Disponível em: <http://eotarchive.cdlib.org/>. Acesso em: 20 ago. 2019.

ESPLEY, S. et al. Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government *Web* Archive to Social Media Content. **Alexandria**, v. 25, p. 31-50. 2014. DOI: https://doi.org/10.7227/ALX.0019.

FARRELL, Susan; ASHLEY, K.; DAVIS, R. **A guide to *web* Preservation:** practical advice for *web* and records managers based on best practices from the JISC- Funded PoWR Project. 2010. Disponível em: <https://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf>. Acesso em: 04 jan. 2020.

FERREIRA, Lisiane Braga; MARTINS, Marina Rodrigues; ROCKEMBACH, Moisés. Usos do arquivamento da *web* na comunicação científica. **Prisma.com**, v. 36, 2018. Disponível em <https://ojs.letras.up.pt/index.php/prismacom/article/view/3927>. Acesso em: 11 abr. 2020.

FONSECA, M. A. et al. Tendências sobre as comunidades virtuais na perspectiva dos prosumers. In: Encontro de Marketing da Anpad, 3, 2008, Curitiba. **Anais...** Curitiba: Anpad, 2008.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

GOMES, Daniel. **Preservar a *Web*:** um desafio ao alcance de todos. In: Actas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas. 2010. Lisboa. Disponível em: http://sobre.arquivo.pt/wpcontent/uploads/PreservarAWebBADFormat-v.14.pdf. Acesso em: 14. Fev. 2019.

GOMES, Daniel; COSTA, Miguel. The Importance of *Web* Archives for Humanities. **International Journal of Humanities and Arts Computing**, v. 8, n. 1, p. 106-123, Apr. 2014. https://doi.org/10.3366/ijhac.2014.0122

GOMES, Daniel; FREITAS, Sergio; SILVA, Mário J. 2006. **Design and selection criteria for a national *web* archive**. In European Conference on Research and Advanced Technology for Digital Libraries, 10. Springer-Verlag, Berlin, Heidelberg, 196-207. DOI=http://dx.doi.org/10.1007/11863878_17.

GOMES, Daniel; MIRANDA, João; COSTA, Miguel. **A survey on *web* archiving initiatives**. In: International Conference on Theory and Practice of Digital Libraries. Lisboa: Springer Berlin Heidelberg, 2011, p. 408-420. Disponível em: <https://link.springer.com/ content/pdf/10.1007%2F978-3-642-24469-8_41.pdf>. Acesso em: 06 abr. 2019.

GROTKE, A.; JONES, G. DigiBoard: A tool to streamline complex *web* archiving activities at the Library of Congress. In: **International *Web* Archiving Workshop**, 10. Vienna, 2010.

GUIZZO, Érico. **Internet - o que é, o que oferece, como conectar-se**. São Paulo: Ática, 1999.

HAKALA, Juha. Archiving the *Web*: European experiences. **Program-electronic Library and Information Systems** - PROGRAM-ELECTRON LIBR INFORM, v. 38, p. 176-183. 2004. DOI: 10.1108/00330330410547223.

HARPER, Corey A. The Dublin Core Metadata Initiative: beyond the element set. **Information Standards Quarterly**, v. 22, n. 1, p. 19-28, 2010. Disponível em: <https://www.niso.org/sites/default/files/stories/2019-11/FE_DCMI_Harper_isqv22no1.pdf>. Acesso em: 07 jan. 2020.

HAUGHEY, Duncan. **Moscow Method**. Project Smart website. 2014. Disponível em: <https://www.projectsmart.co.uk/moscow-method.php>. Acesso em: 17 dez. 2019.

HAYLES, Katherine. **How we think:** digital media and contemporary technogenesis. Chicago: The University of Chicago Press, 2012.

HEDSTROM, Margaret. Digital preservation: a time bomb for digital libraries. **Computers and the Humanities**, v. 31, n. 3, p. 189–202, 1998.

HOLUB, Karolina; RUDOMINO, Inge. Croatian *Web* Archive: on overview. **Преглед НЦД**, v. 25, p. 11-16, 2014.

IBGE. **Website** [*online*], 2018. Pesquisa Nacional por amostra de domicílio. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=o-que-e>. Acesso em: 25 nov. 2019.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. **Website** [*online*], 2019. Disponível em: < http://netpreserve.org/>. Acesso em: 23 jul. 2019.

INTERNET ARCHIVE. **Website** [*online*], 2019a. Disponível em: <https://archive.org/>. Acesso em: 23 jul. 2019.

INTERNET ARCHIVE. **Website** [*online*], 2019b. Internet Archive Help Center: The Wayback Machine. Disponível em: < https://help.archive.org/hc/en-us/categories/360000553851-The-Wayback-Machine>. Acesso em: 23 jul. 2019.

ISLAS-CARMONA, J. O. El prosumidor. El actor comunicativo de la sociedad de la ubicuidad. **Palabra Clave**, v. 11, n. 1, p. 29-39, 2008. Disponivel em: <http://www.scielo.org.co/pdf/pacla/ v11n1/v11n01a03.pdf>. Acesso em: 01 dez. 2019.

JAMAIN, Jassalini; YAHYA, Ayu Lestari; MUHAMMAD, Natalia; RAHMAN, Musa Ayob Abdul. ***Web* archiving issues and challenges in State Government of Sarawak (Malaysia):** Do they really need their website to be archived?. IFLA WLIC. 2018.

KELLY, M. **An Extensible Framework For Creating Personal Archives Of *Web* Resources Requiring Authentication.** Master's thesis, Old Dominion University (2012).

KELLY, M.; BRUNELLE, J. F.; WEIGLE, M. C.; NELSON, M. L. On the change in archivability of websites over time, **International Conference on Theory and Practice of Digital Libraries**. Springer, Berlin, Heidelberg. 2013. p. 35-47.

KHAN, Muzammil; RAHMAN, Arif Ur. A systematic approach towards *web* preservation. **Information Technology and Libraries,** v. 38, n. 1, p. 71-90. 2019. Disponível em: <https://doi.org/10.6017/ital.v38i1.10181>. Acesso em: 01 ago. 2019.

KOEHLER, Wallace. A Longitudinal Study of *Web* Pages Continued: Consideration of Document Persistence. **Information Research**, v. 9, n. 2, 2004. Disponível em: <http://informationr.net/ir/9-2/paper174.html>. Acesso em: 02 ago. 2019.

LALA, Vanita; JOE, Susanna. ***Web* Archiving At The National Library of New Zealand**. LIANZA Conference, 2006.

LAVOIE, Brian. Meeting the challenges of digital preservation: The OAIS reference model. **OCLC Newsletter**, v. 243, p. 26-30, 2000. Disponível em: <http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000001747>. Acesso em: 17 dez. 2019.

LAVOIE, Brian. The Open Archival Information System (OAIS): Introductory Guide. 2 ed. **DPC Technology Watch Report**, october, 2014. 37 p. Disponível em: <https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file>. Acesso em: 09 jan. 2020.

LAWRENCE, Steve; COETZEE, Frans M.; GLOVER, Eric; PENNOCK, David M.; FLAKE, Gary W.; NIELSEN, Finn Arup; KROVETZ, Robert; KRUGER, Aandries; GILES, C. Lee. Persistence of *Web* References in Scientific Research. **IEEE Computer**, v. 34, n. 2, p. 26-31, 2001.

LAZINGER, Susan. S. **Digital preservation and metatada:** History, theory, practice. Englewood, Colorado: Libraries Unlimited, 2013.

LEINER, Barry. M; CERF, Vinton G.; CLARK, David D; KAHN, Robert E.; KLEINROCK, Leonard; LYNCH, Daniel C.; POSTEL, Jon; ROBERTS, Larry G.; WOLFF, Stephen. A brief history of the Internet. **ACM SIGCOMM Computer Communication Review,** v. 39, n. 5, p. 22-31. 2009.

LOWENTHAL, David. Archives, heritage, and history. In: BLOUIN, Francis X. (Jr.); ROSENBERG, William G. (Eds.). **Archives, Documentation, and Institutions of Social Memory:** Essays from the Sawyer Seminar. Michigan: The University of Michigan Press, 2006. p. 193-206.

LUDÄSCHER, B.; MARCIANO, R.; MOORE, R. Preservation of digital data with self-validating, self-instantiating knowledge-based archives. **SIGMOD Record**, v. 30, n. 3, p. 54-63, 2001.

LYMAN, Peter. Archiving the World Wide *Web*. In: COUNCIL ON LIBRATY AND INFORMATION RESOURCES; LIBRARY OF CONGRESS. **Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving**. Council on Library and Information Resources; Library of Congress: Washington, DC, 2002. p. 38-51. Disponível em: <https://www.clir.org/pubs/reports/pub106/*web*/>. Acesso em: 02 ago. 2019.

LYNCH, C. Integrity issues in electronic publishing. In: PEEK, R.P., NEWBY, G.B. (Eds.). **Scholarly publishing:** the electronic frontier. Cambridge: MIT Press, 1996. p. 133-145.

MAEMURA, Emily; WORBY, Nicholas; MILLIGAN, Ian; BECKER, Christoph. If these crawls could talk: Studying and documenting *web* archives provenance. **JASIST**, v. 69, p. 1223-1233. 2018.

MASANÈS, Julien. *Web* Archiving Methods and Approaches: A Comparative Study, **Library Trends**, v. 54, n. 1, p. 72-90. 2005.

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

MELO, Jonas Ferrigolo; ROCKEMBACH, Moisés. Arquivologia e Ciência da Informação na Era do Big Data: perspectivas de pesquisa e atuação profissional em arquivos digitais. **Prisma.com**, n. 39, p. 14-28, 2019a. Disponível em: <https://doi.org/10.21747/16463153/39a2>. Acesso em: 02 dez. 2019.

MIELNICZUK, Luciana; BACCIN, Alciane; BRENOL, Marlise; SOUSA, Maíra; DANIEL, Priscila. Vinte anos de Zero Hora na Internet (1995-2015). **Alcar 2015,** 10ª Encontro Nacional de História da Mídia, UFRGS, Porto Alegre, 2015.

MORENO, José. O valor económico da informação na sociedade em rede. **OBS***, Lisboa, v. 9, n. 2, p. 1-28, jun. 2015. Disponível em: <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S1646-59542015000200001&lng=pt&nrm=iso>. Acesso em: 03 dez. 2019.

NATIONAL LIBRARY OF AUSTRALIA. Themes emerging from archiving *web* resources: issues for cultural heritage organizations. Conference… National Library of Australia, Canberra, 9-11 november 2004. Disponível em: <http://www.nla.gov.au/webarchiving/ConferenceReport.rtf>. Acesso em: 11 jul. 2019.

NELSON, Theodor. **Computer Lib/Dream Machines:** New Freedoms Through Computer Screens. 1974. Self-published.

NEW ZEALAND WEB ARCHIVE. **Website** [*online*], 2019. Disponível em: <https://natlib.govt.nz/collections/a-z/new-zealand-*web*-archive>. Acesso em: 23 jul. 2019.

NISO - NATIONAL INFORMATION STANDARDS ORGANIZATION. **Understanding metadata:** what is metadata, and what is it for?. Beltimore: NISO, 2017. Disponível em: < https://www.niso.org/publications/understanding-metadata-2017>. Acesso em:> 07 jan. 2020.

NIU, Jinfang. An Overview of *Web* Archiving. **D-Lib Magazine**, v. 18, n. 3/4. 2012.

NUAWEB – Núcleo de Pesquisa em Arquivamento da *Web* e Preservação Digital. **Website** [*online*], 2019. Disponível em: <https://www.ufrgs.br/nuaweb/>. Acesso em: 20 nov. 2019.

PABÓN CADAVID, Jhonny. (2014). Copyright Challenges of Legal Deposit and *Web* Archiving in the National Library of Singapore. **Alexandria**, v. 25, p. 1-19, 2014. DOI 10.7227/ALX.0017.

PANDORA. **Website** [*online*], 2019. Disponível em: <https://pandora.nla.gov.au/>. Acesso em: 01 jul. 2019.

PENNOCK, Maureen. *Web*-Archiving. **DPC Technology Watch Report**, v. 13, mar. 2013.

PHILLIPS, Margaret E. Selective archiving of *Web* Resources: A study of acquisition costs at the National Library of Australia. **RLG DigiNews**, v. 9, n. 3, 2005.

PORTAL GOV.BR. **Website** [*online*], 2019. Disponível em: <https://www.gov.br>. Acesso em: 01 jul. 2019.

RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia**, v. 1, n. 1, 96-105. 2014.

ROCKEMBACH, Moisés. Inequalities in digital memory: ethical and geographical aspects of *web* archiving. **International Review of Information Ethics**, v. 26, 2017. Disponível em: <http://www.i-r-i-e.net/inhalt/026/IRIE-26-Marx-12-2017.pdf#page=141>. Acesso em: 11 abr. 2020.

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. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8648747>. Acesso em: 16 fev. 2019.

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

ROCKEMBACH, Moisés; PAVÃO, Caterina Marta Groposo. Políticas e tecnologias de preservação digital no arquivamento da *Web*. **Revista Ibero-Americana de Ciência da Informacão (RICI).** Brasília, v. 11, n. 1, 2018. Disponível em: http://periodicos.unb.br/index.php/RICI/article/view/27950. Acesso em: 12 fev. 2019.

ROCKEMBACH, Moisés; MELO, Jonas Ferrigolo. *Web* archiving of Brazilian websites. In: IIPC WEB ARCHIVING CONFERENCE 2019, 2019, Zagreb, Croatia. **Abstracts & presentations....** Drop-in Talks & Drop-in Slides. Disponível em: <http://netpreserve.org/ga2019/programme/abstracts/>. Acesso em: 01 jul. 2019.

ROGERS, Richard. Au-delà de la critique big data. La recherche sociale et politique à l'ère numérique. In: M. Severo; A. Romele (Eds.). **Traces numériques et territoires**, 18. Paris: Presses des Mines, 2015.

ROMAN, Evandro. Relatório de Deputado Federal Evandro Roman sobre o PL 2431/2015. 2017. Disponível em: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=1594241>. Acesso em: 02 dez. 2019.

ROSEN, Alyssa S. Scientists & Librarians Turn to "End of Presidential Term" *Web* Archive to Safeguard Climate Change Data, **Law Lines**, Jan. 2017. Disponível em: http://digitalcommons.pace.edu/lawfaculty/1041. Acesso em: 25 jul. 2019.

RYAN, Johnny. **A History of the Internet and the Digital Future.** London: Reaktion Books, 2010.

SCHAFER, Valérie; THIERRY, Benjamin. L'ogre et la Toile: Le rendez-vous de l'histoire et des archives du *Web*. **Socio**, v. 4, p. 75-96. 2015.

SCHATZ, Bruce R.; HARDIN, Joseph B. 1994. NCSA Mosaic and the World Wide *Web*: Global Hypermedia Protocols for the Internet, **Science**, v. 265, p. 895-901.

SCHNEIDER, Steven. M.; FOOT, Kirsten; Kimpton, Michele; JONES, Gina. **Building thematic *Web* collections: Challenges and experiences from the September 11 *Web* Archive and the Election 2002 *Web* Archive**. Paper presented at the 3rd Workshop on *Web* Archives (IWAW'03), Trondheim, Norway, 2003.

SHALLCROSS, Michael. **Quality assurance for the Bentley Historical Library *Web* Archives**: guidelines and procedures. Version 3.0. Sep. 2013. Disponível em: <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/94162/BHL_WebArchivesQA-v3-20130909.pdf>. Acesso em: 27 jun. 2019.

SHVEIKY, Rivka; BAR-ILAN, Judit. National Libraries' Traditional Collection Policy Facing *Web* Archiving. **Alexandria**, v. 24, n. 3, p. 37-72. 2013.

SILVA, Armando Malheiro da. Mediações e mediadores em Ciência da Informação. **Prisma.Com**, Porto, n. 9, p. 1-36, 2010.

SMITH, Abby. **New-model scholarship:** how will it survive? Washington, D.C.: Council on Library and Information Resources, 2003.

SOUZA, Fábio. Relatório de Deputado Federal Fábio de Souza sobre o PL 2431/2015. 2015. Disponível em: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=1594241>. Acesso em: 02 dez. 2019.

THE NATIONAL ARCHIVES. **General Guidelines for the selection of records.** United Kingdom: The National Archives, 2006.

THE NATIONAL ARCHIVES. **Operational Selection Policy OSP27UK.** United Kingdom: The National Archives, 2014.

THE NATIONAL ARCHIVES. **The UK Government *Web* Archive:** Guidance for digital and records management teams. United Kingdom: The National Archives, 2017.

THE NATIONAL ARCHIVES. **Website** [*online*], 2019. Disponível em: <http://www.nationalarchives.gov.uk/>. Acesso em: 01 jul. 2019.

TOSH, John. **The Pursuit of History**. 3 ed. London: Pearson Education, 2002. p. 76-83. Disponível em: <https://www.academia.edu/22557305/John_Tosh_-_The_Pursuit_of_History?auto=download>. Acesso em: 11 out. 2019.

TROYE, S.; XIE, C. The active consumer: conceptual, methodological, and managerial challenges of prosumption. In: FIBE, 2007, Bergen. **Conference...** Bergen: NHH, 2007.

UDAPURE, Trupti V.; KALE, Ravindra D.; DHARMIK, Rajesh C. Study of *Web* Crawler and its Different Types. **IOSR Journal of Computer Engineering** (IOSR-JCE), v. 16, n. 1, versão VI, Feb. 2014, p. 01-05.

UK WEB ARCHIVE. **Website** [*online*], 2019. Disponível em: <https://www.webarchive.org.uk/>. Acesso em: 01 jul. 2019.

UNESCO. **Charter on the Preservation of Digital Heritage**. 2003.

VEJA. Ed. Abril. São Paulo, 03 maio 1995.

VICE. **Website** [*online*], 2016. Disponível em: <https://www.vice.com/en_us/article/d7yeej/all-references-to-climate-change-have-been-deleted-from-the-white-house-website-5886b75d0b367c453f87dd14>. Acesso em: 10 out. 2019.

VIEIRA, Eduardo. **Os bastidores da Internet no Brasil:** as histórias de sucesso e de fracasso que marcaram a *Web* brasileira. Barueri, SP: Manole, 2003.

XIE, Zhiwu; VAN DE SOMPEL, Herbert; LIU, Jinyang, VAN REENEN, Johann; JORDAN, Ramiro. Archiving the relaxed consistency *web*. **Proceedings of** ACM International Conference On Information & Knowledge Management, 22. 2013. San Francisco: ACM, p. 2119-2128, 2013. Disponível em: https://dl.acm.org/citation.cfm?id=2505551. Acesso em: 8 jun. 2017.

WABAC - *Web* Archive Browsing Advanced Client. **Website** [*online*], 2020. Disponível em: < https://wab.ac/>. Acesso em: 26 jan. 2020.

WINSTON, Brian. **Media, Technology and Society:** a history. London, Routledge, 2003.

WINTERS, Jane. *Web* archives for humanities research: some reflections. In BRÜGGER, Niels; SCHROEDER, Ralph. **The *Web* as History:** Using *Web* Archives to Understand the Past and the Present. London: UCL Press, 2017. p. 238-248.

WORLD WIDE WEB CONSORTIUM. **Website** [*online*], 2019. Disponível em: <https://www.w3.org/>. Acesso em: 01 jul. 2019.

YIN, R. K. Estudo de caso: planejamento e métodos. 2 ed. Porto Alegre: Bookman, 2001.