https://doi.org/10.29397/reciis.v14i3.2116

ARTIGOS ORIGINAIS

# Arquivabilidade de *websites* para preservação digital: estudo a partir da área da saúde

## Archivability of websites for digital preservation: study from the health area

## Archivabilidad de sitios *web* para preservación digital: estudio del area de salud

*Jonas Ferrigolo Melo*[1,a]
jonasferrigolo@gmail.com | https://orcid.org/0000-0002-7312-3509

*Moisés Rockembach*[1,b]
moises.rockembach@gmail.com| https://orcid.org/0000-0001-9057-0602

[1] Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Programa de Pós-graduação em Comunicação. Porto Alegre, RS, Brasil.

[a] Mestrado em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul.
[b] Doutorado em Informação e Comunicação em Plataformas Digitais pela Universidade do Porto/Universidade de Aveiro.

## RESUMO

Em busca de uma solução para compreender as razões pelas quais alguns recursos presentes em *websites* não são possíveis de serem arquivados pelas ferramentas de captura, surgiu o conceito de arquivabilidade da *web*. Apresentamos este estudo que propõe iniciar uma discussão acerca do tema, a partir do método CLEAR+ e da ferramenta ArchiveReady, e verificar sua aplicabilidade a partir da identificação de *websites* da área da saúde, com testes de preservação digital por meio do arquivamento da *web*. A pesquisa configurou-se como estudo de caso, com procedimentos envolvendo pesquisa bibliográfica e documental, bem como o uso de *software* para identificar arquivabilidade dos *sites*. Conclui-se que tanto os testes de arquivabilidade quanto os de arquivamento da *web* apontam para poucas dificuldades de captura, em pequeno grau, sugerindo-se, portanto, que para atingir uma melhor qualidade de captura sejam adotados padrões de conformidade na produção dos *websites*, de acordo com o estabelecido pelo World Wide Web Consortium.

**Palavras-chave:** Arquivamento da *web*; Arquivabilidade; *Websites*; Preservação digital; Informações da saúde.

## ABSTRACT

In search for a solution to understand the reasons why some resources present on websites are not possible to be archived by capture tools, we approach the concept of web archivability. We present this study that proposes to initiate a discussion about the evaluation of the archivability, using the CLEAR+ method and the ArchiveReady, and to verify their applicability from the identification of websites in the health studies, with digital preservation tests through the web archiving. The research was configured as a case study, with procedures involving bibliographic and documentary research, as well as the use of software to identify the archivability of the sites. It is concluded that both archivability tests and web archiving tests point to little capture difficulties, to a small degree, therefore suggesting that to achieve better capture quality, compliance standards should be adopted in the production of websites, according to what is established by the World Wide Web Consortium.

**Keywords:** Web Archiving; Archivability; Websites; Digital preservation; Health information.

## RESUMEN

En la búsqueda de una solución para comprender las razones por las cuales las herramientas de captura no pueden archivar algunos recursos presentes en sitios *web*, abordamos el concepto de archivabilidad de la *web*. Presentamos este estudio que propone iniciar una discusión sobre la evaluación de la archivabilidad de los sitios *web*, utilizando el método CLEAR+ y la herramienta ArchiveReady, y verificar su aplicabilidad a partir de la identificación de sitios *web* en los estudios de salud, con pruebas de preservación digital a través del archivo *web*. La investigación se configuró como un estudio de caso, con procedimientos que implican investigación bibliográfica y documental, así como el uso de *software* para identificar la capacidad de archivo de los sitios. Se concluye que tanto las pruebas de archivabilidad como las pruebas de archivo *web* apuntan a pequeñas dificultades de captura, en un pequeño grado, lo que sugiere que para lograr una mejor calidad de captura, se deben adoptar estándares de cumplimiento en la producción de sitios *web* de acuerdo con lo establecido por el World Wide Consorcio Web.

**Palabras clave:** Archivo *web*; Archivabilidad; Sitios *web*; Preservación digital; Información de salud.

INFORMAÇÕES DO ARTIGO

**Este texto compõe o Dossiê Preservação Digital.**

**Contribuição dos autores:**
Concepção e desenho do estudo: Jonas Ferrigolo Melo, Moisés Rockembach.
Aquisição, análise ou interpretação dos dados: Jonas Ferrigolo Melo.
Redação do manuscrito: Jonas Ferrigolo Melo, Moisés Rockembach.
Revisão crítica do conteúdo intelectual: Moisés Rockembach.

**Declaração de conflito de interesses:** não há.

**Fontes de financiamento:** não houve.

**Considerações éticas:** não há.

**Agradecimentos/Contribuições adicionais:** não há.

**Histórico do artigo:** submetido: 18 maio 2020 | aceito: 31 jul. 2020 | publicado: 30 set. 2020.

**Apresentação anterior:** não há.

## INTRODUÇÃO

A *web* é um meio essencial para disseminação de informações e sua importância é confirmada pelo uso massivo na produção e uso de conteúdos no ambiente digital, fazendo com que essa tecnologia tenha um caráter dinâmico e efêmero[1]. Nesse sentido, uma série de princípios metodológicos precisam ser observados para que se promova a preservação de seu conteúdo. A apreensão com a efemeridade do conteúdo na *web* se dá em razão da velocidade com que se perde o acesso às informações produzidas e disponibilizadas, sendo este um dos maiores fatores de preocupação de pesquisadores da área.

A partir do reconhecimento deste 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 de *websites*[2]. A comunidade internacional de arquivamento da *web* possui uma dinâmica de constante desenvolvimento e novas ferramentas são criadas para melhorar as técnicas de preservação e para diminuir a perda sistemática de conteúdo *on-line* causada pela natureza transitória da *web*.

Iniciativas que visam entender que o conteúdo presente na internet também faz parte do patrimônio cultural e social precisam evoluir para combater a efemeridade desses materiais. Torna-se necessário ter certeza de que essas informações, além de serem acessíveis em todo o mundo, perdurarão 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[3] e, assim como os espaços habituais de preservação da memória, também se constituem como fonte para pesquisas e podem se consolidar como espaços fundamentais para a salvaguarda de informações de uma época.

Masanès[4] 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 e a sociedade como um todo, usando recursos avançados da tecnologia da *web*. Devido a sua natureza dinâmica, a rede está em constante estado de evolução e não há garantia de que seu conteúdo atual permanecerá acessível em um futuro próximo, fazendo com que um número considerável de *sites* tenha vida útil bastante curta. Outra questão importante é a mudança do conteúdo, que pode ser causada por diferentes motivos – desde a decisão pessoal do proprietário do *website* em editar partes do conteúdo a alterações acidentais que eventualmente podem ocorrer ou até mesmo alterações nos próprios domínios[5].

Um dos principais desafios do Arquivamento da *Web* é compreender que *websites* podem não ser arquivados corretamente, em razão de problemas que surgem a partir do uso de diferentes tecnologias, padrões e práticas de implementação de páginas. Estes problemas continuam surgindo, mesmo que muitos desses portais tenham sido desenvolvidos com o uso de Web Content Management System (WCMS), o que minimiza a diversificação dessas tecnologias, segundo Banos e Manolopoulos[6]. Ao longo do tempo o WCMS ganhou protagonismo junto ao desenvolvimento de páginas *web*, fazendo com que a rede mundial de computadores passasse de pequenos *websites* informais para grandes e complexos sistemas tecnológicos, exigindo programas que promovam o gerenciamento dessas páginas de forma eficiente. O Web Content Management Systems foi criado em diferentes linguagens de programação, passando a operar em muitos *websites*; como, por exemplo, a dependência de 74,6 milhões de *sites* do WordPress[i], um dos mais conhecidos WCMS.

Nesse cenário de complexidade dos *websites* – em que o desafio do arquivamento se materializa – e na busca de uma solução para compreender as razões pelas quais alguns recursos presentes em *websites* não são possíveis de arquivamento é que surgiu o conceito de arquivabilidade. Nesse sentido, associado à produção dos autores Banos e Manolopoulos[7,5] e Banos *et al.*[8], apresentamos este estudo que propõe explorar uma discussão acerca da avaliação da arquivabilidade de *websites*, a partir do método CLEAR+ e da

i Disponível em: https://managewp.com/blog/14-surprising-statistics-about-wordpress-usage.

ferramenta ArchiveReady (http://archiveready.com/), bem como analisar a captura de *websites* da área da saúde, como um estudo de caso. E, a partir disso, pretendemos iniciar discussões sobre a temática, visando possibilitar uma articulação entre os saberes dos profissionais que atuam na construção de *websites* e dos que trabalham com a preservação digital. Para o desenvolvimento da pesquisa, procedemos com uma pesquisa bibliográfica e documental, com revisão de literatura e identificação de *websites* relacionados à saúde, bem como o uso de *software* para testes de arquivabilidade e preservação dos *websites*, respectivamente, ArchiveReady e Webrecorder.

## ARQUIVAMENTO DA *WEB* E PRESERVAÇÃO DIGITAL

Após seu surgimento, a internet mudou consideravelmente nas duas décadas. Ela foi concebida com intenções de proteger os Estados Unidos durante a Guerra Fria, mas se expandiu na era dos computadores pessoais e dos computadores de rede. Idealizada inicialmente como suporte a diversas funções, como o compartilhamento de arquivos e *login* remoto, passando pela criação do correio eletrônico e, mais recentemente, a criação da World Wide Web[9].

A consolidação da Internet e da *web* para o público geral teve início com a facilidade de aquisição de computadores e também com os avanços tecnológicos para conexões que deixaram de ser discadas e passaram para a banda larga. Os equipamentos de comunicação exclusiva por voz deram lugar às conexões em 3G e 4G. O crescimento do número de provedores, o estabelecimento do comércio *on-line*, o mercado de jogos, que passou a acontecer também em rede, além do aparecimento dos canais de *streaming* e das plataformas multimídia foram alguns dos fatores determinantes para a expansão e popularização da Internet e da *web*. Nos anos 2000, as plataformas de redes sociais mudaram nossas relações, a produção e o uso de conteúdos pela *web* transformou a forma como nos comunicamos na contemporaneidade, tornando a Internet um recurso de grandes possibilidades.

A partir dessa realidade 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"[10]. Lawrence *et al.*[11] mencionam uma estimativa da Alexa Internet (https://www.alexa.com/) em que as páginas da *web* desaparecem após 75 dias, em média. Recentemente, Costa, Gomes e Silva[12] revelam 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 fotos e descrições de eventos exclusivamente publicados na Internet. Essa é uma dificuldade para usuários que desejam procurar informações que podem não estar mais disponíveis nos *websites*[13]. Já estamos enfrentando falta de informações devido à ausência de páginas ou formatos e extensões antigas de documentos, questão já prevista por Vint Cerf, um dos pioneiros da Internet, que em 2015 alertou sobre o perigo das futuras gerações que terão pouco ou nenhum registro do século XXI[ii]. Segundo Márdero Arellano[14], a preservação digital constitui-se em atividades complexas, a longo prazo, que necessitam de investimentos e, com a criação de redes colaborativas, podem superar os crescentes desafios.

Desta forma, surge uma preocupação na área da informação e tecnologia: a preservação do conteúdo publicado na *web*. Dezenas de iniciativas, ao redor do mundo, já se preocupam com o arquivamento da *web* e demonstram uma variedade de métodos e abordagens para selecionar, adquirir, organizar, armazenar,

ii Disponível em: https://www.bbc.com/news/science-environment-31450389.

descrever e fornecer acesso a esse conteúdo[15]. 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, as políticas e tecnologias para efetivação desse arquivamento.

Dois consórcios internacionais foram criados com a intenção de padronizar e fortalecer os estudos da *web*: em 1994 é criado o World Wide Web Consortium – W3C (https://www.w3.org) –, liderado pelo inventor da *web*, Tim Berners-Lee, e por Jeffrey Jaffe, CEO do consórcio. A missão do W3C é aproveitar ao máximo o potencial desse meio informacional, por meio do desenvolvimento de protocolos, diretrizes e padrões, que garantam seu crescimento a longo prazo. 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.[16]

Em 2003, a Biblioteca Nacional da França criou o International Internet Preservation Consortium – IIPC (http://netpreserve.org/) –, com a missão de 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. Atualmente, o consórcio conta com aproximadamente 60 membros de mais de 45 países, incluindo arquivos e bibliotecas nacionais, universitários e regionais. Entre eles está a maior plataforma do mundo para arquivamento da *web*, a Internet Archive (https://archive.org/), que conta com os arquivos nacionais de países como o Canadá e o Reino Unido. Entre as bibliotecas, destacamos a British Library (Biblioteca Britânica) e a Library of Congress (Biblioteca do Congresso Americano).[17]

A partir das discussões promovidas por essas comunidades internacionais, as organizações trabalham para desenvolver padrões da *web*, além de promover o desenvolvimento e o uso de ferramentas e técnicas que permitem a criação de arquivos digitais, incentivando arquivos, bibliotecas nacionais e organizações de pesquisa que queiram abordar o arquivamento e a preservação da Internet.

Nessa direção, este artigo discorre sobre a arquivabilidade da web, um tópico específico que integra os estudos sobre o arquivamento da web, campo de pesquisa que começa a se desenvolver no Brasil nos últimos anos com o fomento de discussões na área, sobretudo desde 2017[26].

## ARQUIVABILIDADE DE *WEBSITES*

A soma de fatores que tornaram os *websites* documentos complexos fez com que o processo de preservação digital dessas informações fosse um desafio, ao passo que os rastreadores da *web* precisam recuperar o conteúdo com precisão e confiabilidade[18]. Banos *et al.*[8] dizem que o processo automatizado de rastreamento da *web* não estabelece métodos padronizados para acessar os dados completos de um *website*. Os arquivos da *web* estão deixando de apresentar partes significativas de *websites* arquivados, sendo que esse conteúdo pode variar de *website* para *website*, dependendo do tipo de recursos, acessibilidade e tecnologias que foram utilizadas em seu desenvolvimento[8].

Ao concluírem que não existe uma métrica para auxiliar na decisão de um *website* poder ou não ser arquivado com êxito, a partir de 2013, os autores se dedicaram a elaborar um sistema que verificasse as possibilidades de arquivamento de um *website* por *software* de captura, antecipando verificações de garantia de qualidade que devem ser executadas antes do seu arquivamento. Deste modo, permite-se avaliar os resultados e decidir se é possível prosseguir com o arquivamento ou definir estratégias a serem adotadas sobre o processo ideal para a preservação das páginas[7].

Os autores investigaram até que ponto cada Web Content Management System (WCMS) atende às condições para uma transferência segura de seu conteúdo para a preservação em um arquivo da *web*, permitindo identificar seus pontos fortes e fracos e, assim, deduzir recomendações específicas para melhorar o desenvolvimento de páginas *web*, com o objetivo de avançar na prática de rastreamento e

arquivamento dessas informações[8]. O resultado da pesquisa foi o desenvolvimento do método CLEAR (Credible Live Evaluation of Archive Readiness), que propõe uma abordagem para produzir medições *on-line* da arquivabilidade de *websites*[8]. O método foi testado e recebeu retorno de pesquisadores e usuários, em 2015 passou por ajustes, sendo renomeado para CLEAR+[7].

Na pesquisa de Banos e Manolopoulos[7], foi definido um conjunto de métricas para quantificar o nível de arquivabilidade de qualquer *website*. O método CLEAR+ foi desenvolvido para consolidar, estender e complementar práticas empíricas de verificação se um *site* é arquivável ou não, por meio da formulação de um processo padrão automatizado[7]. Os *websites* se beneficiam ao seguir as práticas recomendadas, os padrões internacionais e as tecnologias da *web*, caso queiram ser arquivados. E a soma dos atributos que tornam um *site* passível de arquivamento foi chamado de Web Archivability (WA) ou, em português, arquivabilidade do *site*.

A arquivabilidade do *site* é definida como a extensão em que ele atende às condições para a transferência segura de seu conteúdo para um arquivo da *web* para fins de preservação[8]. Ou seja, é uma noção estabelecida para capturar os aspectos principais de um *website*, crucial para diagnosticar se ele tem o potencial de ser arquivado com integridade e precisão[7-8]. A ferramenta desenvolvida para calcular a arquivabilidade do *website* fornece uma abordagem para automatizar o controle de qualidade, avaliando a conveniência de ser arquivado antes de qualquer tentativa de fazê-lo, proporcionando ganhos consideráveis, ao reduzir o uso de recursos humanos, computacionais e de rede, e ao não coletar *websites* não colhíveis[8].

Banos e Manolopoulos[7], em 2015, afirmam que "o arquivamento automatizado da *web* em grande escala pode ser bastante aprimorado em relação ao desempenho, recursos e eficácia, usando uma métrica quantitativa para avaliar os *sites* de destino antes do rastreamento e arquivamento. Usando o método CLEAR+, podemos evitar *sites* que não são arquiváveis e fazer melhor uso dos recursos disponíveis. Acreditamos que a escala crescente da *web* e a situação de recursos limitados da maioria dos arquivos da *web* forçarão os arquivistas a considerar o emprego de métricas quantitativas como o CLEAR+ em seus processos de seleção."[7]

Banos *et al.*[8], propõem, com o método CLEAR+, uma abordagem que produz medições *on-line* da arquivabilidade do *site*, sendo que os principais componentes do método são: Facetas de arquivabilidade; Atributos do *site*; e Avaliações. As Facetas de arquivabilidade são medidas ao se realizarem avaliações específicas, considerando-se os Atributos do *site*. Por exemplo: "Qual é a porcentagem de *hiperlinks* válidos *versus* inválidos? [...] Os arquivos CSS usados em um *website* estão em conformidade com os padrões do W3C? [...] Qual é a porcentagem de arquivos de imagem corrompidos, se houver?"[6]. Em outras palavras, podemos dizer que a pontuação para cada Faceta de arquivabilidade é calculada como a média ponderada das pontuações das avaliações associadas à determinada faceta, considerando-se os Atributos do *site*. O significado de cada avaliação define seu peso, ou seja, sua arquivabilidade.

## Facetas de arquivabilidade

As Facetas de arquivabilidade (Figura 1) podem ser resumidas como os fatores que precisam ser considerados para calcular a arquivabilidade total do *website*. Os resultados da avaliação da arquivabilidade dos *websites* são pontuações diferentes na faixa de 0 a 100% para FA, FC, FM e FS (Facet Accessibility, Facet Cohesion, Facet Metadata e Facet Standards Compliance, consecutivamente). A pontuação final da arquivabilidade é a média das pontuações das facetas.

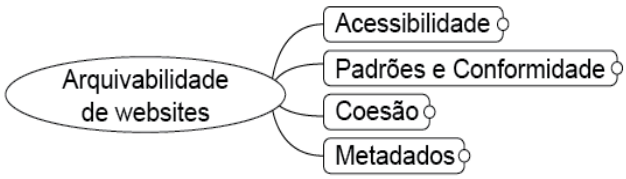


Figura 1 – Facetas de arquivabilidade: visão geral
Fonte: Banos e Manolopoulos (2015).

**FA: Accessibility (Acessibilidade):** "Um *website* é considerado arquivável apenas se os rastreadores da *web* puderem acessar sua página inicial, percorrer seu conteúdo e recuperar via solicitações HTTP padrão"[8]. Caso o rastreador não localize todos os recursos do *website*, não será possível recuperar seu conteúdo. Isso pressupõe que não basta inserir recursos e mecanismos em um *website*, durante seu desenvolvimento, mas é, também, essencial fornecer referências que permitam que os rastreadores localizem e recuperem os conteúdos promovidos por meio desse recurso[8]. É necessário garantir, por exemplo, que um arquivo de imagem em movimento do tipo .gif tenha seus metadados identificados e que seu conteúdo possa ser recuperado pelo rastreador.

Outro fator que contribui para a Acessibilidade é a possibilidade dos *links* permanecerem válidos. Um conjunto de mapas, guias e atualizações de *links* também pode ser fornecido, a partir do mapa do *site* ou em arquivos robots.txt, para ajudar os rastreadores a encontrar todo o conteúdo[8]. Além disso, outro recurso que pode ajudar a determinar a Acessibilidade de um *website* é ter as informações sobre determinada página ou recursos da página preservados em outros locais, como, por exemplo, em um repositório *on-line*. A Figura 2 ilustra a faceta.

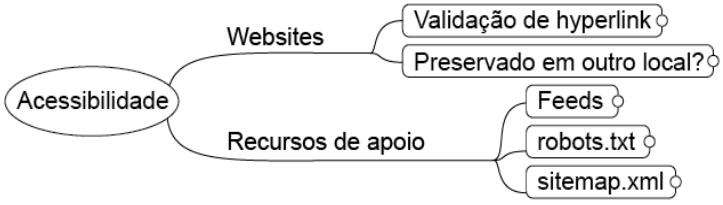


Figura 2 – Facetas de arquivabilidade: Acessibilidade
Fonte: Banos *et al*. (2013).

Importante destacar que, na primeira versão do método CLEAR, em 2013, além das quatro facetas, apresentadas na Figura 1, existia a quinta faceta: Performance. Na revisão do método, em 2015, a faceta Performance foi incorporada à Acessibilidade, considerando que os resultados do seu desempenho, durante os testes, foram de 100% ou próximos a 100%[6-7].

A Universidade Stanford, em seu projeto de arquivamento da *web*, apresenta os critérios de arquivabilidade com a intenção de incentivar desenvolvedores de *websites* a considerar o uso de padrões internacionais. A universidade descreve os critérios de acesso, captura, descrição e recursos. No item "acesso" estão descritos requisitos, tais como: manter os *links* estáveis, conformidade com padrões da *web*, uso de formatos e dados duráveis, relatar tipo de mídia e codificação de caracteres e usar o *design* responsivo para *websites*[19].

**FS: Standards Compliance (Conformidade com os padrões):** O estabelecimento de padrões é um tema frequente, quando o assunto é preservação digital. Acima de tudo, o padrão deve dar suporte à difusão e à transparência com dependências externas mínimas e não deve ter restrição legal em relação

aos processos de preservação que possam ocorrer com o arquivo. Os autores recomendam que "[...] para que os recursos digitais sejam preservados, eles precisam ser representados em padrões conhecidos e transparentes. Os próprios padrões podem ser proprietários, desde que sejam amplamente adotados e bem compreendidos com ferramentas de suporte para validação e acesso"[8].

Se uma página da web não tiver sido criada com padrões aceitos, é improvável que ela seja disponibilizada por navegadores da *web* usando métodos já estabelecidos[7]. Uma das verificações realizadas, a partir dessa faceta, é se o código fonte HTML está em Conformidade com os padrões do W3C, com o uso do validador HTML W3C (https://validator.w3.org/)[6]. Os autores recomendam que a validação seja realizada para três tipos de conteúdo, conforme exemplificado na Figura 3: conteúdo da mídia de referência (áudio, imagem, vídeo, documentos); componentes da página *web* (CSS e HTML, por exemplo); e recursos de apoio (por exemplo, robots.txt, sitemap.xml, JavaScript)[8].

Cabe destacar que o Governo Federal publicou, em 2019, os Padrões Web em Governo Eletrônico, que são recomendações de boas práticas para o desenvolvimento de ferramentas eletrônicas pelos órgãos do Governo Federal. Contemplam informações sobre a padronização de codificação, administração, usabilidade, redação *web*, desenho e arquitetura de conteúdo e modelos e arquivos-base. Trata-se de uma importante iniciativa, considerando que as recomendações se baseiam nos acordos com os padrões estabelecidos pelo W3C, que, entre outras vantagens, são fundamentais para garantir a aquivabilidade da *web*, ainda que o documento não traga informações a esse respeito.

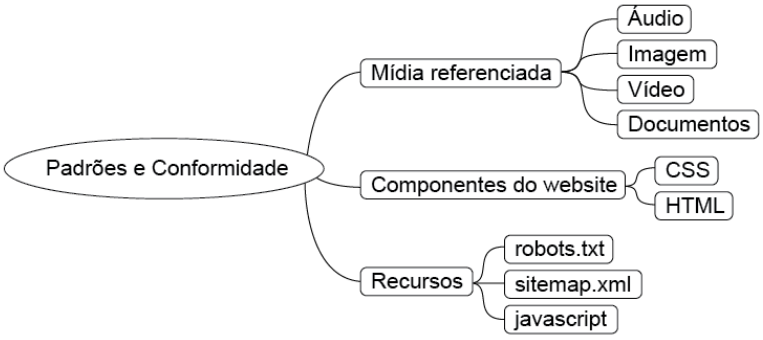


Figura 3 – Facetas de arquivabilidade: Padrões e conformidade
Fonte: Banos *et al*. (2013).

**FC: Cohesion (Coesão):** A Coesão "[...] é a robustez de um *website* contra a falha de diferentes serviços da *web*. Esta faceta refere-se a *websites* dispersos em diferentes serviços [...], em domínios diferentes"[6]. Em outras palavras, a Coesão é relevante para que os rastreadores da *web* operem de forma eficiente e para que seja possível identificar se os arquivos que constituem o *website* estão dispersos por diferentes serviços, como em servidores específicos para imagens – widgets, JavaScript e outros[7]. Para exemplificar: segundo os autores, pode ocorrer de as imagens, que estão sendo usadas em um *website*, estarem hospedadas em outro local, logo elas podem não ser capturadas – o que causaria problemas no arquivamento da página. A premissa é que manter as informações associadas ao mesmo *website* levaria a uma robustez de recursos contra eventuais alterações que possam ocorrer fora do *website* principal[8].

A Coesão é testada em dois níveis[iii]: 1) examinar quantos domínios são empregados em relação à localização do conteúdo da mídia utilizada (imagens, vídeo, áudio, arquivos proprietários); 2) examinar quantos domínios são empregados em relação aos recursos de suporte (por exemplo, robots.txt, sitemap.xml e JavaScript)[7].

iii Na versão de 2013 do método CLEAR, aparecia um terceiro nível para o teste de Coesão que consistia em examinar o número de vezes que *softwares* ou *plugins* proprietários eram referenciados. Em 2015, esse recurso deixou de existir no método CLEAR+.

**FM: Metadata Usage (Uso de metadados):** Como exemplificado por Di Pretoro e Geeraert[27], padrões de metadados de preservação digital são relevantes nos estudos de arquivamento da *web* e podem ser adaptados e aplicados à preservação desses conteúdos, como o MARC21, UNIMARC, MODS, EAD, METS e Dublin Core. "O fornecimento adequado de metadados tem sido uma preocupação constante na curadoria digital. A falta de metadados prejudica a capacidade do arquivo em gerenciar, organizar, recuperar e interagir com o conteúdo de maneira eficaz."[7]. Os autores consideram a verificação dos metadados em Sintaxe, Semântica e Pragmática (Figura 4): "Para evitar os perigos associados ao comprometimento com qualquer modelo de metadados específico, adotamos um ponto de vista geral compartilhado em muitas disciplinas da informação (por exemplo, filosofia, linguística, ciências da computação) com base em sintaxe (por exemplo, como isso é expresso), semântica (por exemplo, sobre o que é isso) e pragmática (por exemplo, o que você pode fazer com isso)"[8].

Uma linguagem entendida pelo usuário final pode ser indicada como parte do atributo do elemento HTML: informações descritivas como nome, palavra-chave, nome de determinado aplicativo e outras informações ajudam a entender como o conteúdo está classificado, podendo ser incluídas no atributo e nos valores do elemento HTML. A existência de elementos nos metadados é entendida como uma maneira de aumentar a probabilidade de implementar a extração e o refinamento automatizados dos metadados, na coleta dos *websites* e, subsequente, no gerenciamento do repositório.[8]

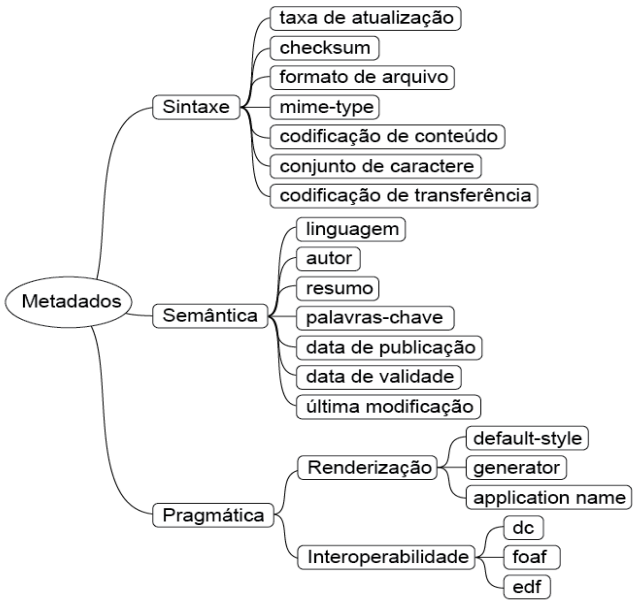


Figura 4 – Facetas de arquivabilidade: Metadados
Fonte: Banos *et al*. (2013).

## Atributos do *site*

Os Atributos do *site* são componentes das páginas *web* analisados com a intenção de avaliar e medir o potencial do *website* em atender os requisitos das Facetas de arquivabilidade[8]. A Figura 5 ilustra os atributos.

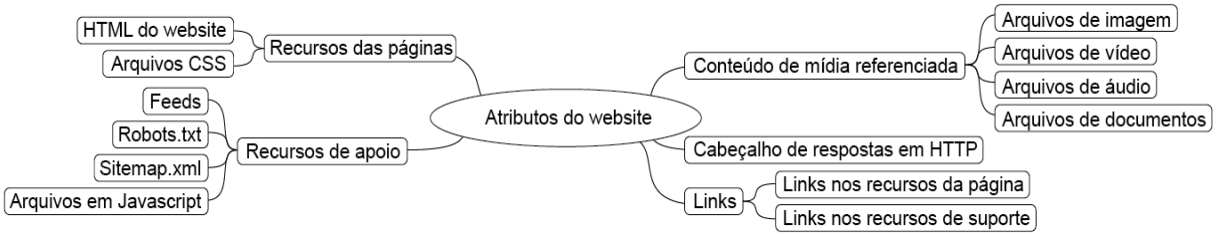


Figura 5 – Atributos do *site*
Fonte: Banos *et al*. (2013); Banos, Manolopoulos (2015).

O nível de Acessibilidade é quantificado com base em alguns elementos, tais como: a existência de *feeds* (por exemplo RSS), robots.txt e sitemap.xml; se é mencionado, no robots.txt e no sitemap.xml, a existência de local especificado e/ou o sitemap.xml é encontrado no diretório raiz do servidor; se *hiperlinks* são válidos e acessíveis; e se existem extensões da página da *web* em outros servidores. A existência de um *feed* RSS permite que a publicação de conteúdos em páginas *web* tenha sua distribuição automática, fazendo com que os rastreadores recuperem o conteúdo publicado de forma automatizada, como faz, por exemplo, a BBC, ao utilizar *feeds* que envia aos leitores as notificações de quando há um novo conteúdo adicionado no *website*[iv].[8]

O protocolo Sitemaps, suportado em conjunto por mecanismos de pesquisa mais populares para ajudar criadores de conteúdo, é a maneira cada vez mais utilizada para desbloquear dados ocultos, disponibilizando-os para os mecanismos de pesquisa[20]. Para implementar o protocolo Sitemaps, o arquivo sitemap.xml é usado para listar todas as páginas do *website* e sua localização[8]. Também é importante destacar que à medida que os *websites* se tornam mais sofisticados e complexos aumentam as dificuldades enfrentadas pelos rastreadores na coleta. Alguns rastreadores, por exemplo, têm habilidades limitadas para processar conteúdo dinâmico em JavaScript ou mídias de *streaming*[15]. Para superar esses obstáculos, foram desenvolvidos padrões que ajudam a tornar os *websites* mais acessíveis à coleta por rastreadores da *web*. Dois exemplos são os protocolos sitemap.xml e robots.txt. O protocolo sitemap.xml4 (Simple Website Footprinting) é uma maneira de criar uma imagem detalhada da estrutura e arquitetura de *links* de um *website*, assim como a implementação do protocolo robots.txt fornece aos rastreadores da *web* informações sobre elementos específicos de um *website* e suas permissões de acesso[21].

## Avaliações

As Avaliações são os resultados da análise dos Atributos do *site*, a partir da sua combinação, que são utilizados para calcular a métrica que determinará a percentagem de arquivabilidade do *website*[8].

Os autores definiram fórmulas que calculam o potencial de arquivabilidade de *websites* em relação às facetas, em que uma pergunta binária é feita e o total passa a ser associado à faceta em questão. Essas fórmulas definem a base do produto que foi desenvolvido para estabelecer a arquivabilidade do *site*, o ArchiveReady. O sistema está disponível em http://archiveready.com/ e é descrito como uma ferramenta de avaliação de arquivabilidade de *websites* em tempo real. O ArchiveReady analisa o *website*, a partir de um endereço válido, e realiza avaliações complexas para calcular a arquivabilidade dessas páginas, considerando o conjunto de fatores, anteriormente descritos[22]. A ferramenta é a implementação do método CLEAR+, baseada em *software* de código aberto, e pode ser utilizada gratuitamente.

Banos e Manolopoulos[6] resumem o processo de avaliação da arquivabilidade de *website* da seguinte forma: "1. O ArchiveReady recebe uma URL de destino e executa uma solicitação HTTP para recuperar o *hipertexto* da

iv Disponível em: https://www.bbc.com/news/10628494.

página da *web*. 2. Após analisá-lo, conexões HTTP são iniciadas em paralelo para recuperar todos os recursos da *web* referenciados na página de destino, imitando uma teia de aranha. 3. No estágio 3, os Atributos do *website* são avaliados. Mais detalhadamente: (a) análise e validação de HTML e CSS, (b) análise e validação de cabeçalhos de resposta HTTP, (c) recuperação, análise e validação de arquivos de mídia (imagens, outros objetos), (d) recuperação, análise e validação de sitemap.xml e robots.txt, (e) detecção, recuperação, análise e validação de *feeds* RSS, (f) avaliação de desempenho de transferência de rede. 4. As métricas para as facetas do WA são calculadas de acordo com o método CLEAR+ e a classificação final do WA é produzida"[6].

## PRESERVAÇÃO DE *WEBSITES* DA ÁREA DA SAÚDE

A busca de informações da área da saúde na *web* é um fenômeno ao mesmo tempo recente e crescente. Muitas pessoas pesquisam, a partir dos motores de busca da rede, sobre questões relacionadas à saúde e se deparam com problemas como a falta de credibilidade ou confiabilidade das fontes encontradas na *web*, com possíveis impactos negativos advindos do julgamento de confiança sobre a informação *on-line*[23].

Faz-se necessário o desenvolvimento de uma literacia digital em saúde (eHealth Literacy) como uma medida para que os usuários de Internet avaliem melhor as fontes de informação na *web*[24-25]. Nesse sentido, também é importante selecionar fontes de informações confiáveis, como, por exemplo, os *websites* institucionais e oficiais de organismos que têm a missão de pesquisa e ação em saúde pública.

A partir de um levantamento sobre arquivos da *web*, relativos à área da saúde, destacamos alguns exemplos de instituições governamentais, tal como o U.S. Department of Health & Human Services (HHS)[v]; o National Library of Medicine (NLM), nos Estados Unidos[vi]; o National Health Service (NHS), do Reino Unido, preservado pelos Arquivos Nacionais britânicos[vii]; e *sites* relacionados às políticas públicas em saúde, pela Biblioteca do Congresso americano[viii]. São exemplos de arquivos *web* por apresentarem um grande volume de conteúdo arquivado e, por consequência, aparecem como destaques, quando se busca pelo termo '*health*' nas plataformas de arquivos da *web*.

Para esta pesquisa, optamos por selecionar um *website* de cada tipo de estrutura de negócio do Ministério da Saúde (agência, fundação e empresa), totalizando quatro *websites* de órgãos governamentais brasileiros relativos à área da saúde, que compõem a amostra deste estudo de caso: o *website* central do Ministério da Saúde (http://www.saude.gov.br); o da Agência Nacional de Vigilância Sanitária – Anvisa (http://portal.anvisa.gov.br/); o da Fundação Nacional de Saúde – Funasa (http://www.funasa.gov.br/); e o da Empresa Brasileira de Hemoderivados e Biotecnologia – Hemobrás (www.hemobras.gov.br).

A aplicação do método CLEAR+, na fase relacionada à garantia de qualidade do arquivamento da *web*, é uma forma de prever possíveis erros de arquivamento e, a partir dos resultados, é possível agir previamente em relação à concretização do arquivamento. O resultado ajuda na definição do processo ideal para a preservação dos *websites* e se é possível prosseguir com o arquivamento[6]. Nesse sentido, antes de realizar o arquivamento, examinamos os resultados que a análise do ArchiveReady apresenta para os *websites* que compuseram o corpo amostral.

Importante considerar que a fase de garantia de qualidade deve considerar os indicadores de qualidade predefinidos pelas instituições, com base na política e nos 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[26].

---

v Disponível em: <https://www.hhs.gov/about/archive/index.html>. Acesso em: 30 abr. 2020.

vi Disponível em: National Library of Medicine *web archive*. Disponível em: <https://www.nlm.nih.gov/webcollecting/index.html>. Acesso em: 30 abr. 2020.

vii Disponível em: <http://www.nationalarchives.gov.uk/webarchive/>. Acesso em: 28 abr. 2020.

viii Pesquisa realizada no *web archive* da Biblioteca do Congresso americano. Disponível em: <https://www.loc.gov/collections/public-policy-topics-web-archive/?fa=contributor:dept.+of+public+health>. Acesso em: 30 abr. 2020.

Para os *websites* do Ministério da Saúde e da Anvisa, o resultado da avaliação foi aproximado: 83% e 82%, respectivamente, conforme demonstram as Figuras 6 e 7. Segundo o ArchiveReady, as menores notas foram as das facetas Acessibilidade e Conformidade com os padrões. Em Acessibilidade, foram identificados alguns *links* inválidos e más práticas aplicadas a códigos JavaScript. O desempenho dos *websites* também não foi considerado de alta qualidade e o acesso não foi autorizado para algumas coletas com o uso de *crawler* (robôs). Para a faceta Conformidade com os padrões foram identificados alguns problemas, considerando aspectos relacionados aos CSS e ao HTML.

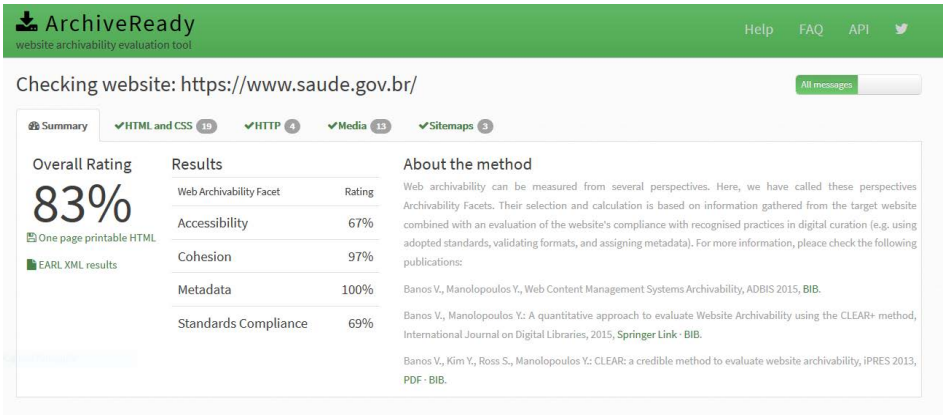


Figura 6 – Resultado do ArchiveReady para o *website* do Ministério da Saúde
Fonte: Os autores (2020).

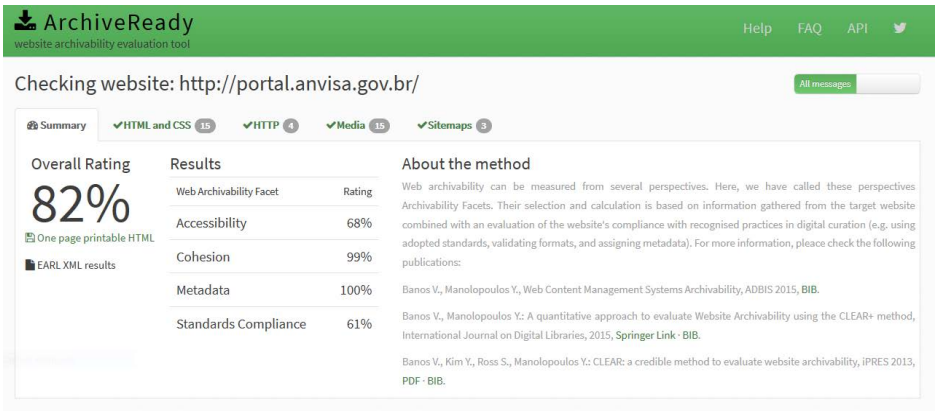


Figura 7 – Resultado do ArchiveReady para o *website* da Anvisa
Fonte: Os autores (2020).

No caso dos *websites* da Funasa e da Hemobrás, o resultado da avaliação da arquivabilidade foi de 70% e 75%, respectivamente, conforme apresentado nas Figuras 8 e 9. As facetas Acessibilidade, Metadados e Conformidade com os padrões apresentaram resultados abaixo do esperado, considerando, além dos mesmos problemas verificados nos dois *websites* anteriores em maior escala, problemas no cabeçalho de cache do HTTP.

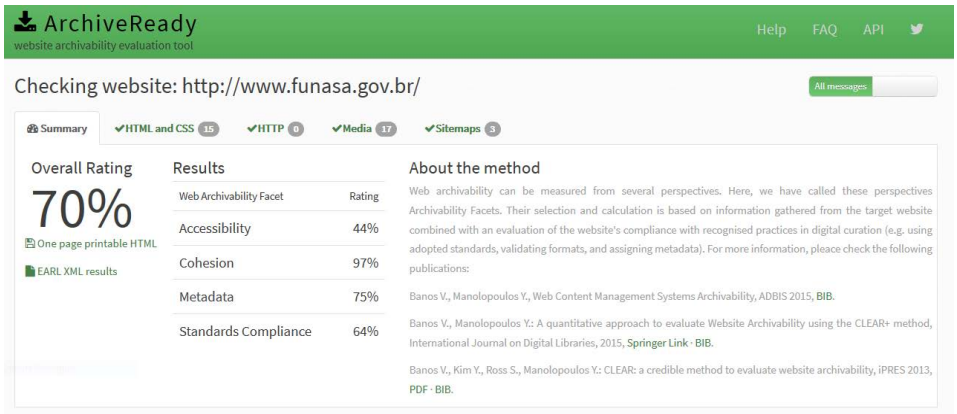


Figura 8 – Resultado do ArchiveReady para o *website* da Funasa
Fonte: Os autores (2020).

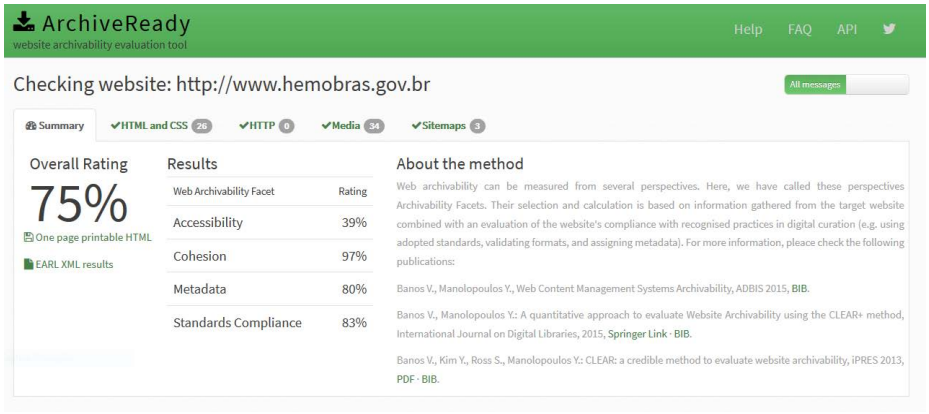


Figura 9 – Resultado do ArchiveReady para o *website* da Hemobrás
Fonte: Os autores (2020).

Depois de aplicar a avaliação, foram iniciadas as coletas para arquivamento com o uso da ferramenta Webrecorder. Foram coletadas 27 URLs, no dia 02 de maio de 2020, agrupadas na coleção 'Saúde GOV.BR'. Foram arquivados a página inicial, o Mapa do *Site*, e, pelo menos, outras três URLs de cada um dos *websites* que compuseram o corpo amostral. As Figuras 10, 11, 12 e 13 ilustram os resultados dos arquivamentos.

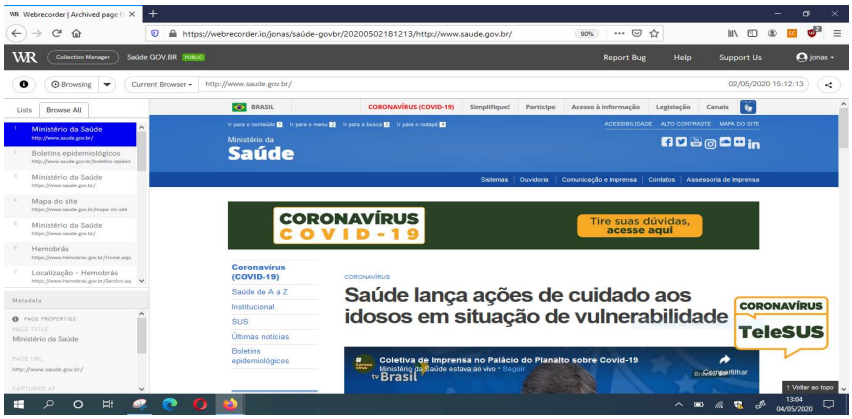


Figura 10 – Arquivamento do *website* do Ministério da Saúde
Fonte: Os autores (2020).

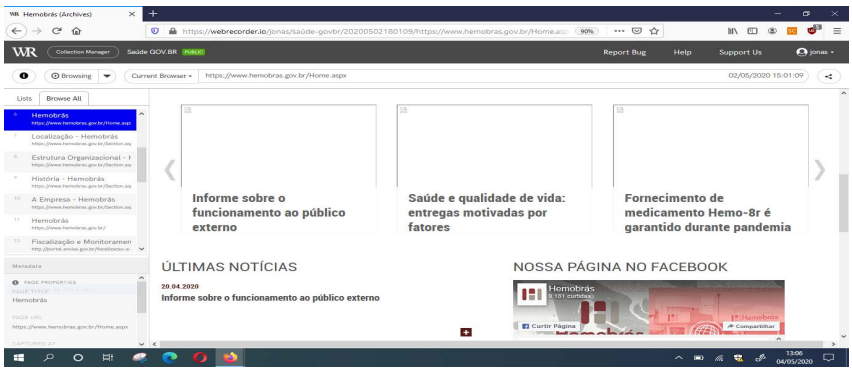


Figura 11 – Arquivamento do *website* da Hemobrás
Fonte: Os autores (2020).

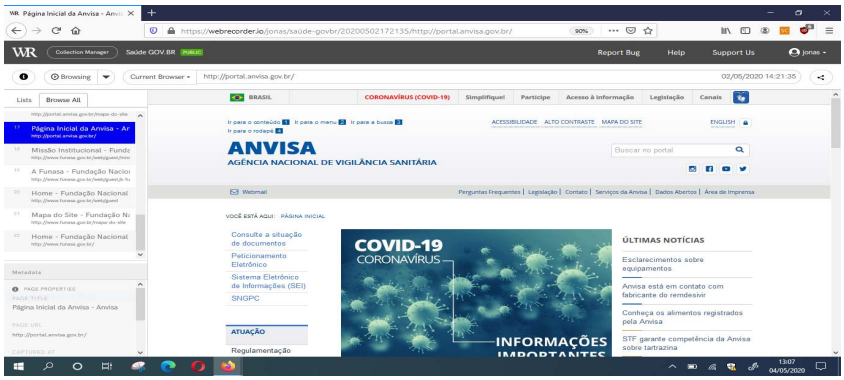


Figura 12 – Arquivamento do *website* da Anvisa
Fonte: Os autores (2020).

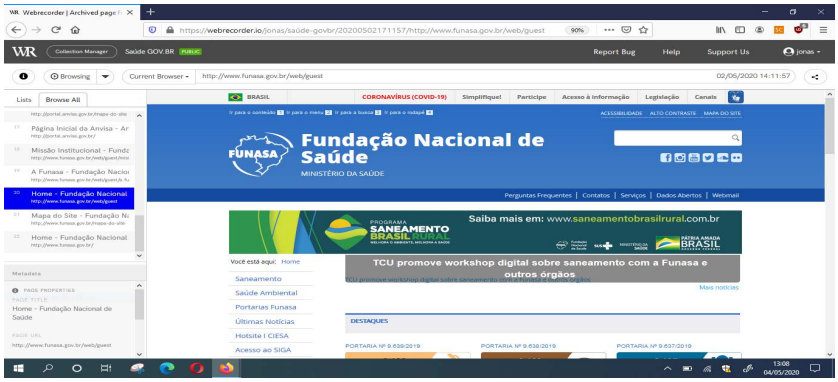


Figura 13 – Arquivamento do *website* da Funasa
Fonte: Os autores (2020).

Depois de arquivados, foram acessadas todas as URLs para comparar com a versão 'ao vivo' dos *websites*. Da totalidade de imagens, textos, vídeos e demais recursos presentes nas páginas arquivadas, somente o *website* da Hemobrás teve problemas com o arquivamento de algumas imagens presentes na página inicial do *site*. Segundo o ArchiveReady, os problemas para arquivamento das imagens têm relação estreita com os padrões de conformidade, que para o *website* em questão pontuou 83% nessa faceta. Mensagens indicando que o formato de imagens não estava em conformidade com o padrão pretendido apareceram nos resultados do ArchiveReady, assim como erros no processamento de alguns arquivos de imagem. Ainda que a faceta Acessibilidade tenha apresentado o menor desempenho indicando 39%, em razão de 117 erros de HTML, não foi verificado nenhum problema com as URLs arquivadas.

As demais URLs e todos os outros *websites* arquivados não tiveram nenhum resultado diferente dos que apresentam suas versões 'ao vivo'. Foi possível, inclusive, reproduzir o vídeo presente na página inicial do *site* do Ministério da Saúde.

## CONSIDERAÇÕES FINAIS

Em busca de informações a respeito da arquivabilidade da *web*, se desenvolveu este estudo, a partir das investigações dos autores Banos e Manolopoulos, responsáveis por introduzir o conceito no cenário científico e nos estudos que versam sobre o arquivamento da *web*. A intenção deste artigo é iniciar uma discussão sobre o tema e encontrar canais de comunicação entre os profissionais de desenvolvimento de *websites* e os profissionais com interesse na preservação desses conteúdos.

Como ponto de partida, foi apresentada uma introdução sobre o arquivamento da *web* e sua contextualização no cenário internacional. Em seguida, procuramos abordar o conceito de arquivabilidade e sua aplicação a partir do método CLEAR+. Dessa forma, tornou-se necessário discorrer sobre as fases que dão sustentação aos cálculos que definem a porcentagem de arquivabilidade de *websites*, quais sejam: Facetas de arquivabilidade; Atributos do *site*; e Avaliações. A aplicação destas fases e de cálculos é operacionalizada pela utilização do ArchiveReady.

A fase de coleta de conteúdo das páginas da Internet é um aspecto complexo e delicado no fluxo de trabalho de arquivamento da *web*, especialmente quando se depende de sistemas externos, tais como os servidores da *web*, e de aplicativos, *proxies* e outras infraestruturas de rede que porventura existam entre o arquivo da *web* que está sendo coletado e o *website* ou servidor de destino. Por outro lado, ao aplicar o método CLEAR+, a fase relacionada à garantia de qualidade no arquivamento da *web* poderá ser melhor explorada, considerando que os resultados já estarão preestabelecidos, uma vez que os testes de arquivabilidade podem ser executados antes do arquivamento, de modo a avaliar os resultados e decidir sobre o processo ideal da preservação dos *websites* e, se é possível, prosseguir com o arquivamento.

Outro aspecto que cabe discutir é a possibilidade de os desenvolvedores da *web* não saberem – ou até mesmo desconhecerem o assunto – se seus *websites* podem ou não ser arquivados corretamente por um arquivo da *web*. Ao incluir o cálculo de arquivabilidade, na fase de desenvolvimento do seu fluxo de trabalho, esses profissonais poderão entender melhor o nível de arquivabilidade de seus sistemas e evitar possíveis problemas com o futuro arquivamento. Falhas de arquivamento de páginas da *web* poderiam ser evitadas em muitos casos, se os desenvolvedores incluíssem o cálculo do arquivamento da *web* em seus testes, pois os problemas poderiam ser identificados com antecedência. Ao mesmo tempo, os padrões estabelecidos pelo consórcio W3C são desconhecidos por parte dos desenvolvedores. A consideração das diretrizes estabelecidas como padrões internacionais, no desenvolvimento dos *websites*, pode levar a um alto grau de arquivabilidade e preservação digital pelos arquivos da *web*.

Outro benefício desse método é que os arquivos da *web* podem calcular a arquivabilidade dos *websites* de destino e informar aos seus proprietários sobre problemas específicos de arquivamento, possibilitando que os rastreamentos subsequentes tenham mais sucesso na coleta. Os arquivos da *web* também podem evitar a captura de *websites* específicos, se suas pontuações no WA forem muito baixas ou se algumas avaliações específicas falharem, economizando recursos. Observando os *sites* capturados, sugere-se que, para atingir uma melhor qualidade de captura, sejam adotados padrões de conformidade na produção dos *websites*, de acordo com o estabelecido pelo World Wide Web Consortium.

Além da parte teórica sobre arquivabilidade, este estudo apresentou a análise de quatro *websites* de órgãos governamentais brasileiros que tratam de saúde. São eles: Ministério da Saúde, Anvisa, Funasa e Hemobrás. Foi realizada análise no *software* ArchiveReady e posterior arquivamento com o uso do Webrecorder. Na sequência, os *websites* arquivados e suas versões 'ao vivo' foram comparadas e apenas o *website* da Hemobrás

apresentou falhas no arquivamento de imagens que estavam presentes na página inicial do portal. Essas falhas aconteceram em razão das imagens não estarem em conformidade com os padrões estabelecidos pelo consórcio W3C. Importante destacar que os resultados do ArchiveReady se referem à análise da totalidade do *website* e não apenas a URLs arquivadas. Isso significa que poderiam haver outros erros perceptíveis, se os *websites* tivessem sido arquivados em sua totalidade. De todo modo, para o corpo amostral selecionado, o resultado do arquivamento foi satisfatório, mesmo que o ArchiveReady tenha apresentado algumas inconsistências, que poderão ser consideradas para a melhoria dos *websites*, em relação à preservação de seu conteúdo na integralidade. Isso demonstra que novas pesquisas podem ser desenvolvidas sobre essa temática, a partir das discussões preliminares que apresentamos neste estudo.

Por fim, além de melhorar a usabilidade da *web* e ajudar a garantir a preservação do patrimônio cultural coletivo, o atendimento aos critérios de arquivabilidade também tenderá a otimizar o *website* para acessos de rastreadores, aumentar o desempenho da página, aprimorar a usabilidade e melhorar os aspectos de consulta e recuperação de versões históricas do seu conteúdo. Conclui-se, portanto, que o uso do ArchiveReady configura-se como uma importante ferramenta que está à disposição de forma gratuita e livre.

## REFERÊNCIAS


1. Brügger N, Milligan I, editors. The SAGE handbook of web history. Los Angeles: SAGE; 2018.
2. Melo JF, Nunes L, Rockembach M. Preservação de websites governamentais a partir do arquivamento da web: abordagens e metodologias [Internet]. In: A Ciência da Informação e a era da Ciência de Dados. ENANCIB 2019: Anais do 20º Encontro Nacional de Pesquisa em Ciência da Informação; 2019 out. 21-25; Florianópolis, Brasil. Florianópolis: UFSC; 2019 [citado em 2020 maio 1]. p. 1-8. Disponível em: http://hdl.handle.net/10183/202684.
3. Ferreira LB, Martins MR, Rockembach M. Usos do arquivamento da web na comunicação científica. Prisma.com [Internet]. 2018 [citado em 2020 abr. 27];(36):78-98. Disponível em: https://ojs.letras.up.pt/index.php/prismacom/article/view/3927.
4. Masanès J, editor. Web archiving. Berlin: Springer; 2006.
5. Rockembach M. Arquivamento da web no contexto das Humanidades Digitais: da produção à preservação da informação digital. Liinc em Revista [Internet]. 2019 [citado em 2020 abr. 27];15(1):131-9. Disponível em: http://revista.ibict.br/liinc/article/view/4578.
6. Banos V, Manolopoulos Y. Web content management systems archivability. In: Tadeusz M, Valduriez P, Bellatreche L, editors. Advances in databases and information systems. ADBIS 2015: Proceedings of the 19th East European Conference on Advances in Databases and Information Systems; 2015 Sept. 8-11; Poitiers, France. Berlin: Springer; 2015. p. 198-212.
7. Banos V, Manolopoulos Y. A quantitative approach to evaluate website archivability using the CLEAR+ method. Int J Digit Libr [Internet]. 2015 [cited 2020 Mar. 13];17(2):119-41. Disponível em: https://link.springer.com/article/10.1007/s00799-015-0144-4?sa_campaign=email/event/articleAuthor/onlineFirst.
8. Banos V, Kim Y, Ross S, Manolopoulos Y. CLEAR: a credible method to evaluate website archivability [Internet]. In: Borbinha J, Nelson M, Knight S, editors. iPRES 2013: Proceedings of the 10th International Conference on Preservation of Digital Objects; 2013 Sept. 3-5; Lisbon, Portugal. Lisbon: IST; 2013 [cited 2020 Mar. 13]. p. 9-18. Disponível em: http://purl.pt/24107/1/iPres2013_PDF/CLEAR%20a%20credible%20method%20to%20evaluate%20website%20archivability.pdf.
9. Leiner BM, Cerf VG, Clark DD, Kahn RE, Kleinrock L, Lynch DC, et al. A brief history of the Internet. Comput Commun Rev. 2009;39(5):22-31.
10. Rockembach M, Pavão CMG. Políticas e tecnologias de preservação digital no arquivamento da web. R Ibero-amer Ci Inf [Internet]. 2018 [citado em 2020 fev. 12];11(1):168-82. Disponível em: https://www.lume.ufrgs.br/handle/10183/175153.
11. Lawrence S, Coetzee FM, Glover E, Pennock DM, Flake GW, Nielsen FA, et al. Persistence of web references in scientific research. Computer. 2001;34(2):26-31.
12. Costa M, Gomes D, Silva MJ. The evolution of web archiving. Int J Digit Libr. 2016;18(3):191-205.

13. Rockembach M. Arquivamento da web: estudos de caso internacionais e o caso brasileiro. Rev Digit Bibliotecon Cienc Inf [Internet]. 2018 [citado em 2010 abr. 18];16(1):7-24. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8648747.

14. Márdero Arellano MA. La preservación digital y la red Cariniana. In: Márdero Arellano MA, Araújo LMS., organizadores. Tendências para a gestão e preservação da informação digital. Brasília, DF: IBICT; 2017. p. 200-18.

15. Niu J. An overview of web archiving. D-Lib Magazine. 2012;18(3-4).

16. W3C: World Wide Web Consortium [Internet]. Cambridge: W3C; c2020 [cited 2020 Apr. 1]. Disponível em: https://www.w3.org/.

17. International Internet Preservation Consortium [Internet]. c2020 [cited 2020 Apr. 1]. Disponível em: http://netpreserve.org/.

18. Van Ballegooie M, Duff W. DCC digital curation manual: instalment on archival metadata [Internet]. London: UKOLN; 2006 [cited 2020 May 15]. Disponível em: http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/archival-metadata.

19. Stanford University [Internet]. Stanford: ST University; c2020. [cited 2020 Apr. 1]. Disponível em: https://www.stanford.edu/.

20. Schonfeld U, Shivakumar N. Sitemaps: above and beyond the crawl of duty [Internet]. In: Quemada J, León G, editors. WWW '09: Proceedings of the 18th International Conference on World Wide Web; 2009. Apr. 20-24; Madrid, Spain. New York: ACM; 2009 [cited 2020 Apr. 27]. p. 991-1000. Disponível em: https://dl.acm.org/doi/abs/10.1145/1526709.1526842.

21. Mansfield-Devine, S. Simple website footprinting. Network Security, 2009;4:7-9.

22. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and agenda for future research. J Med Internet Res. 2017;19(6):e218.

23. Paige SR, Stellefson M, Krieger JL, Anderson-Lewis C, Cheong J, Stopka C. Proposing a transactional model of eHealth Literacy: concept analysis. J Med Internet Res. 2018;20(10):e10175.

24. Huhta AM, Hirvonen N, Huotari ML. Health Literacy in web-based health information environments: systematic review of concepts, definitions, and operationalization for measurement. J Med Internet Res. 2018;20(12):e10273.

25. Bingham N. Quality assurance paradigms in web archiving pre and post legal deposit. Alexandria [Internet]. 2014 [cited 2020 May 15];25(1-2):51-68. Disponível em: http://dx.doi.org/10.7227/ALX.0020.

26. Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital [Internet]. Porto Alegre: UFRGS/CNPq; c2020. [citado em 2020 jun. 20]. Disponível em: https://www.ufrgs.br/nuaweb.

27. Di Pretoro E, Geeraert F. Behind the scenes of web archiving: metadata of harvested websites. Archives et Bibliothèques de Belgique – Archief – En Bibliotheekwezen in Belgie; Archief, in press, trust an Undertanding: The value of metadata en a digitally joined-up world [Internet]. 2019 [cited 2020 Jul. 22]. Disponível em: https://hal.archives-ouvertes.fr/HAL-02124714/DOCUMENT.