Artigo Original
Dossiê: Edição comemorativa dos 10 anos da Rede Cariniana

RBPD Revista Brasileira de Preservação Digital
RBPD Brazilian Journal of Digital Preservation

DOI: 10.20396/rebpred.v3i00.16017

MODELO DE ARQUIVAMENTO DE PÁGINAS WEB PARA PORTAIS DE PERIÓDICOS

um relato de pesquisa no Portal de Periódicos da UNICAMP

WEBPAGE ARCHIVING MODEL FOR JOURNAL PORTALS

a research report on UNICAMP's Periodical Portal

Resumo

Introdução: O crescimento dos documentos em formato digital tem provocado diversas mudanças no cenário da pesquisa científica, e diversos *sites* institucionais e de pesquisa se proliferaram de forma acelerada e necessitam de tratamento. Os arquivos da *Web* têm um papel importante a desempenhar como infraestruturas sociais que permitem a preservação da memória. Objetivo: Preservar digitalmente, por meio da ferramenta Conifer da empresa Rhizome, as páginas *Web* do Portal de Periódicos Eletrônicos Científicos da Universidade Estadual de Campinas (UNICAMP). Metodologia: Um ensaio de cunho exploratório, por meio da abordagem no *"microarchiving"* com enfoque no levantamento bibliográfico para a elaboração da revisão da literatura sobre a temática para avaliação e análise do Portal de Periódicos Eletrônicos da UNICAMP por meio do arquivamento da *Web* pela ferramenta Conifer. O formato trabalhado nesta metodologia é o WARC. O WARC é um formato de arquivo para conteúdo da *Web* que armazena conteúdo da página *Web*, cabeçalhos de respostas e metadados para um grupo de páginas da *Web*. Resultados: Preservadas atualmente 2 revistas do Portal de Periódicos da UNICAMP. Conclusão: Espera-se que com essa metodologia possamos ampliar a preservação e arquivamento da *Web* de páginas institucionais da universidade, bem como oferecer a metodologia para as demais instituições interessadas em arquivar as páginas *Web* de seus Portais.

Palavras-chave: Arquivamento da *Web*. Periódicos eletrônicos. Preservação digital. Portais de periódicos.

Abstract

Introduction: The growth of documents in digital format has brought about several changes in the scientific research landscape, and several institutional and research sites have proliferated at an accelerated rate and are in need of treatment. Web archives have an important role to play as social infrastructures that enable the preservation of memory. Objective: To digitally preserve, using Rhizome's Conifer tool, the Web pages of the Scientific Electronic Journals Portal of the "Universidade Estadual de Campinas" (UNICAMP). Methodology: An exploratory essay, using the "microarchiving" approach with a focus on bibliographic survey for the elaboration of the literature review on the theme for the evaluation and analysis of the Electronic Periodical Portal of UNICAMP through the web archiving by the Conifer tool. The format used in this methodology is WARC. WARC is an archive format for Web content that stores Web page content, response headers and metadata for a group of Web pages. Results: 2 journals from UNICAMP's Portal de Periódicos are currently preserved. Conclusion: It is hoped that with this methodology we can expand the preservation and archiving of the university's institutional Web pages, as well as offer the methodology to other institutions interested in archiving their portals' Web pages.

Keywords: Web archiving. Electronic journals. Digital preservation. Journal portals.

Gildenir Carolino Santos
Universidade Estadual de Campinas
E-mail: gilldenir@unicamp.br
Campinas – SP / Brasil

Danilo Formenton
Universidade Federal de São Carlos
E-mail: formenton.danilo@gmail.com
São Carlos – SP / Brasil

Gabriela Ayres Ferreira Terrada
Fundação Biblioteca Nacional
E-mail: gaby.uff@gmail.com
Rio de Janeiro – RJ / Brasil

RBPD Revista Brasileira de Preservação Digital
RBPD Brazilian Journal of Digital Preservation

CREDIT[1]

• Conceituação	SANTOS, G.C.
• Curadoria de dados	SANTOS, G.C.
• Investigação	SANTOS, G.C.; FORMENTON, D.
• Metodologia	SANTOS, G.C.
• Administração de projetos	SANTOS, G.C.
• Software	SANTOS, G.C.
• Supervisão	SANTOS, G.C.; FORMENTON, D.
• Validação	SANTOS, G.C.; FORMENTON, D.; TERRADA, G. A. F
• Visualização	SANTOS, G.C.; FORMENTON, D.; TERRADA, G. A. F.
• Redação – rascunho original	SANTOS, G.C.; FORMENTON, D.
• Redação – revisão e edição	FORMENTON, D.; TERRADA, G. A. F.

PUBLISHERS
Universidade Estadual de Campinas – Sistema de Bibliotecas / Instituto Brasileiro de Informação em Ciência e Tecnologia – Rede Brasileira de Serviços de Preservação Digital – Cariniana. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Gildenir Carolino Santos, Miguel Angel Márdero Arellano.

Submetido em: 01/05/2022 – Aceito em: 10/06/2022 – Publicado em: 12/07/2022

[1] Sobre o CRediT, consulte o site e conheça outros papéis: https://casrai.org/credit/

1 Introdução

É cada vez mais evidente que o acesso aberto se estabeleceu, bem como a Ciência Aberta para as publicações científicas no âmbito mundial. Este cenário fica comprovado quando observamos o aumento do número de revistas científicas sendo criadas por instituições de ensino superior e importantes universidades nacionais e estrangeiras, buscando também pela preservação digital dos seus portais.

Para uma maior clareza do conceito de preservação digital, o Tesauro Brasileiro de Ciência da Informação defini a preservação digital como sendo "[...] estratégias de preservação que lidam com a obsolescência tecnológica dos objetos digitais de forma a assegurar, no futuro, o acesso aos mesmos" (PINHEIRO; FERREZ, 2014, p. 176). Definição esta complementada por Hedstrom (1998, p. 190, tradução nossa) que a defini tratando-se do "[...] planejamento, alocação de recursos e aplicação de métodos de preservação e tecnologias necessárias para garantir que informações digitais de valor contínuo permaneçam acessíveis e utilizáveis". Além da migração de dados, da emulação tecnológica e dos metadados de preservação, outra importante estratégia de preservação digital se refere à preservação do conteúdo de *websites*[2], ou o arquivamento da *Web* (FORMENTON; GRACIOSO, 2020).

O periódico científico é um canal importante que possibilita a disseminação e institucionalização do conhecimento científico. O ato de publicar inclui a autenticação e legitimação do conhecimento produzido. A contribuição do periódico científico promove melhorias no tempo (determinado pela velocidade de distribuição) e no espaço (em que as revistas são agrupadas em áreas específicas de conhecimento). Os periódicos de acesso aberto, em sentido amplo, apoiam-se em vários pilares relacionados à eliminação de barreiras para compartilhar resultados de investigação científica. Sob esse modelo, a informação tem o potencial de atingir diferentes públicos e contextos e, assim, contribuir para a democratização do conhecimento, bem como possibilitar condições satisfatórias para que sejam preservados digitalmente, garantindo a memória científica (ANGELO; OLIVEIRA, 2021).

Nesse sentido, as instituições buscam através de suas coleções digitais, implementar os portais de periódicos por meio de parcerias na possibilidade de controlar, guardar e preservar digitalmente essas coleções. Mas, diante dessa busca, procuram também, mais recentemente, implementar projetos para iniciar o processo de arquivamento das páginas dos seus periódicos hospedados no portal em geral, realizando uma sintonia além da preservação digital.

Uma das ações consideradas importantes para esse processo é a utilização de plataformas abertas para gerenciamento e editoração de periódicos científicos e aqui destacamos o *Open Journal System* (OJS), que hoje é considerado uma referência em gestão de fluxo de submissões fazendo com que tanto o editor quanto o autor/leitor discutam e acompanhem o fluxo editorial da revista com transparência e preservando os *logs* do sistema, itens importantes no processo de publicação, criando páginas

[2] *Website* é uma coleção de páginas da *Web* e conteúdo relacionado que é identificado por um nome de domínio comum (*domain name*) ou endereço único que identifica um *site* da *Internet* (por exemplo, "www.unicamp.br") e publicado em pelo menos um servidor *Web*, sendo que todos os *sites* acessíveis ao público constituem coletivamente a *World Wide Web*. Por sua vez, uma página da *Web* (*web page*) é um documento de hipertexto na *Web* acessível por meio do *Hypertext Transfer Protocol* (HTTP) e renderizado por um navegador da *Web* (*browser*) para exibição (WEB PAGE, 2022; WEBSITE, 2022).

dinâmicas de seus conteúdos, e que estão em constante atualizações de versões, e que necessita de preservação e arquivamento das mesmas nas suas instituições.

O OJS foi lançado em 2002 como *software* de código aberto distribuído pelo *Public Knowledge Project* (PKP). No Brasil foi traduzido para o idioma em português pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), oferecendo o ambiente customizado de acordo com a identidade visual da instituição.

É um *software* recomendado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), em que o processo editorial no sistema permite uma melhoria na avaliação da qualidade dos periódicos, e maior rapidez no fluxo das informações. A aceitação do OJS pela comunidade brasileira de editores científicos se deve ao desempenho do sistema e de sua fácil adaptação aos processos de editoração em uso.

Cada vez mais, as universidades desenvolvem e implantam os portais de periódicos para agregar em massa as produções científicas, para disseminá-las e preservá-las em sistemas de preservação digital. De modo geral, um portal é uma página específica na *Internet*, que serve como ponto de acesso direto a outros conjuntos de serviços e informações, contendo subdivisões específicas sobre determinado tema ou área do conhecimento.

Segundo Santos (2017), atualmente, com os sistemas de gerenciamento eletrônico de editoração científica, já são configurados com a estruturação para serem preservados, e possuem facilidade para arquivamento de páginas *Web*. Como citado anteriormente, o OJS é a ferramenta ideal para esses processos, pois cabe a ele o gerenciamento das publicações periódicas *online* no tocante da preservação por possuir *plugins* específicos para isso.

O OJS, em sua terceira versão desde 2018, é um *software* desenvolvido para a construção e a gestão de publicações periódicas eletrônicas, ou de várias publicações gerenciadas por meio de um portal. Essa ferramenta contempla ações essenciais à automação das atividades de editoração de periódicos científicos garantindo a visibilidade das publicações pela *Internet*.

Nesse sentido, as vantagens de um portal de publicações periódicas gerenciado pelo *software* OJS são diversificadas, como ao incluir-se no sistema *Lots of Copies Keep Stuff Safe* (LOCKSS). O LOCKSS trata-se de um sistema de preservação digital de informação eletrônica lançado pela Universidade de *Stanford* em abril de 2004, que permite a preservação de todos os dados da publicação, replicando com segurança os dados de várias publicações por meio de caixas virtuais (*box*) entre uma ou mais instituições. No Brasil, esse serviço de preservação das publicações periódicas estruturadas pelo OJS é gerenciado também pelo IBICT, por meio da Rede Brasileira de Serviços de Preservação Digital (Rede Cariniana)[3]. Assim, os portais de periódicos que utilizam o *software* OJS e que fazem parte da Rede Cariniana garantem a preservação digital dos dados de suas publicações (PEREIRA, 2019).

Dessa forma, esse breve ensaio, tem como objetivo apresentar o processo da preservação digital e arquivamento das páginas *Web*, por meio da ferramenta Conifer da empresa *Rhizome*, as páginas *Web* do Portal de Periódicos Eletrônicos Científicos (PPEC)[4] da Universidade Estadual de Campinas (UNICAMP).

[3] Disponível em: https://cariniana.ibict.br/. Acesso em: 1 jun. 2022.

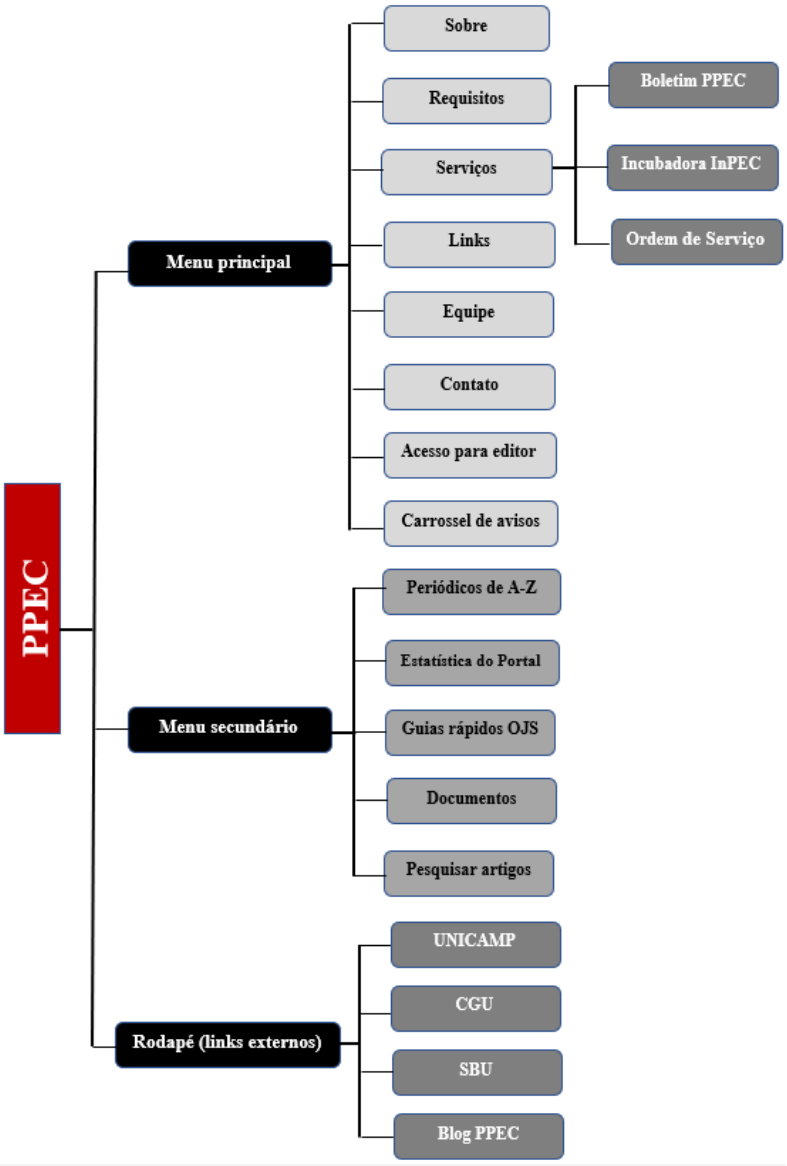
[4] Disponível em: https://periodicos.sbu.unicamp.br/ppec/. Acesso em: 1 jun. 2022.

2 Sobre o Portal de Periódicos Eletrônicos Científicos da UNICAMP

O PPEC da UNICAMP foi criado em 2014 sob a Portaria GR 012/2014[5], para instituir grupo de trabalho e criar o Portal com os periódicos editados pela universidade. Seu desenvolvimento foi fruto do pós-doutorado do bibliotecário Gildenir Carolino Santos junto ao Laboratório de Estudos Avançados em Jornalismo (Labjor/ UNICAMP) e supervisionado pela Professora Vera Regina Toledo Camargo. O PPEC está vinculado ao Sistema de Bibliotecas da UNICAMP (SBU), com o apoio da Coordenadoria Geral da Universidade (CGU). Foi inaugurado em dezembro de 2015, conta com uma estrutura no seu *website* por um acervo digital com 33 (trinta e três) títulos credenciados, e uma incubadora que presta suporte para o melhoramento e adequação aos requisitos solicitados pelo Portal para ingresso futuramente, com 21 (vinte e um) títulos (Figura 1).

[5] Disponível em: https://www.pg.unicamp.br/norma/3575/0. Acesso em: 1 jun. 2022.

Figura 1. Estrutura do PPEC



Fonte: Baseado em Pereira (2019).

Com a criação do portal de periódicos na UNICAMP, a proposta do PPEC é melhorar a disseminação e divulgação das publicações periódicas para a comunidade interna e externa à UNICAMP e permitir o acesso de forma interativa e participativa da produção desenvolvida na UNICAMP em várias áreas do conhecimento. Esse acesso faz parte da iniciativa *open access*, movimento mundial que visa ampliar o acesso aos resultados da produção científica.

A quantidade de publicações periódicas nos espaços acadêmicos vem aumentando e, após o advento da *Internet*, a demanda por periódicos eletrônicos se tornou ainda maior. Na UNICAMP não foi diferente, visto que a produção científica da instituição cresce em todas as áreas do conhecimento (SANTOS, 2012). O portal surge então como uma página centralizadora, que agrega informação de diferentes áreas advindas de diversas unidades de ensino (institutos e faculdades), centros, núcleos e órgãos que contribuem com a ciência através do gerenciamento ou edição

de publicações periódicas eletrônicas. Para o desenvolvimento do portal foram estabelecidos critérios e normas para editoração, indexação, preservação digital e editoração eletrônica para toda a universidade.

De acordo com Santos e Camargo (2013), além de ter a função de agregar informações, aplicações e informações relevantes aos usuários, um portal de periódicos também apresenta a vantagem de permitir o filtro de uma variedade de informações em uma única interface. Os autores também reforçam que a unificação das publicações periódicas em um portal tem a finalidade única de padronização, recuperação e visibilidade da produção da UNICAMP.

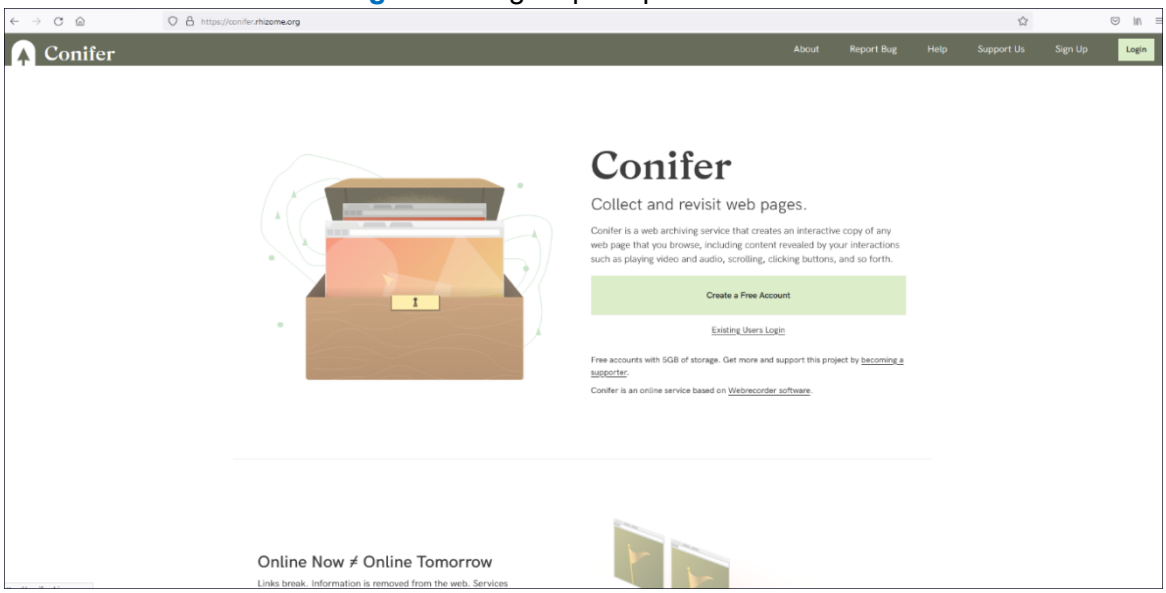
Visando assegurar e preservar as páginas *Web* do Portal, o responsável do PPEC, e participante do Grupo de pesquisa "Estudos e Práticas de Preservação Digital" – Rede de Pesquisa Dríade[6] do IBICT, elaborou uma proposta independente sobre esse assunto, contando com o apoio dos demais coautores neste artigo, e deu início em julho de 2020, com a preservação das páginas *Web* do Portal, com 2 (dois) periódicos inicialmente.

3 Sobre a ferramenta de arquivamento da *Web*: Conifer[7]

O Conifer (Figura 2) é um serviço gratuito que captura a sequência exata de navegação de uma série de páginas *Web* de qualquer *site*, preservando assim a experiência do usuário (SAMOUELIAN; DOOLEY, 2018). De acordo com *International Internet Preservation Consortium* (c2022), *Rhizome.org* ([2022]) e Samouelian e Dooley (c2018) esta ferramenta de *software* de código aberto (chamado anteriormente de *Webrecorder*) cria arquivos de gravações de alta fidelidade, interativas e contextuais de mídia social e outros conteúdos dinâmicos, como *JavaScript* complexo e vídeos integrados, utilizando-se de si mesmo para capturar e reproduzir o *site* (abordagem conhecida como arquivamento simétrico da *Web*).

[6] Disponível em: http://dgp.cnpq.br/dgp/espelhogrupo/3997875180380796. Acesso em: 1 jun. 2022.
[7] Disponível em: https://conifer.rhizome.org/. Acesso em: 1 jun. 2022.

Figura 2. Página principal de entrada do Conifer



Fonte: *Site* do *Rhizome.org* ([2022])

Em oposição à maioria das iniciativas de arquivamento da *Web* que criam de forma automática cópias do material encontrado na *Web* pública, o Conifer é uma plataforma orientada ao ser humano e para tornar *sites* capturados acessíveis, permitindo aos usuários (RHIZOME.ORG, [2022]; SAMOUELIAN; DOOLEY, 2018):

- Criar, selecionar e compartilhar as suas próprias coleções de materiais da *Web*, incluindo itens que só seriam revelados após o *login* ou a execução de outras ações complexas em um *site*;
- Capturar a sequência exata de navegação através de uma série de páginas da *Web* ou objetos digitais, preservando a sua experiência única individual em um momento no tempo (abordagem que coloca o controle do arquivo nas mãos do curador);
- Um arquivamento da *Web* de "alta fidelidade", onde itens que dependem de *scripts* complexos, como vídeos integrados, navegação sofisticada ou gráficos 3D, têm uma taxa de sucesso muito maior para captura com a ferramenta;
- Gerar metadados automaticamente durante o arquivamento e incorporá-los em um arquivo em formato *Web ARChive* (WARC)[8] para *download*, sendo que os elementos de metadados incluem nome de usuário (username, chamado "criador"), título (*title*), data/hora da captura (*capture date/time*), formato do arquivo (*archive file format*), título da coleção (*title of collection*, se incluído nos metadados) e todas as *Uniform Resource Locator* (URLs) que o usuário visitou durante uma sessão de gravação; e
- Um nível gratuito limitado com 5 GB de espaço de armazenamento com algumas restrições de cota de rede, porém o acesso às coleções que os usuários tornaram públicas é a todo momento gratuito e ilimitado.

O *Conifer* é o resultado de um projeto de pesquisa e desenvolvimento de vários anos para criar um serviço de arquivamento da *Web* de última geração que foi hospedado do ano de 2015 a 2020 na *Rhizome* (uma organização sem fins lucrativos que comissiona, apresenta e preserva arte digital) sob o nome "*Webrecorder.io*", sendo que os componentes de código aberto criados durante esse período agora

[8] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 28500**: Information and documentation: WARC file format, Geneva: ISO, 2017. Disponível em: https://www.iso.org/standard/68004.html. Acesso em: 1 jun. 2022.

formam a base da ferramenta *Conifer* e são mantidos de forma independente pelo projeto *Webrecorder* (RHIZOME.ORG, [2022]).

Com a renomeação de *Webrecorder.io* para Conifer, a *Rhizome* tornou-se o administrador permanente do serviço; além disto, neste projeto o grande apoio dado advém da Andrew W. Mellon *Foundation*, com suporte adicional para preservação digital pela James S. and John L. Knight *Foundation*, *Google* e *Google Cultural Institute*, *National Endowment for the Arts* e *New York State Council on the Arts* com o apoio do governador Andrew Cuomo e do Legislativo do Estado de Nova York (RHIZOME.ORG, [2022]).

3.1 Metadados de preservação do Conifer */ Rhizome*

Considerados dados ou informações criadas, salvas e compartilhadas para descrever "coisas" ou objetos, os metadados permitem a interação com eles para termos o conhecimento que precisamos, sendo que eles também são difundidos nos sistemas informacionais e se apresentam de diversas formas (RILEY, c2017).

Assim, na definição ampla e clássica de que metadados significa dados sobre dados, Formenton e Gracioso (2022) em um estudo realizado recentemente, buscam uma definição mais profunda e compreensiva sobre metadados. Os autores encontraram em Alves (2010, p. 47), a definição mais completa e aplicável ao domínio da *Web* e em domínios específicos, como o domínio bibliográfico, além de atender aos propósitos da presente investigação e fundamentar-se na construção padronizada e consistente de representações unívocas dos recursos informacionais em diferentes ambientes digitais estruturados. Desta maneira, os metadados podem ser conceituados como:

> [...] elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades [...] com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação (ALVES, 2010, p. 47).

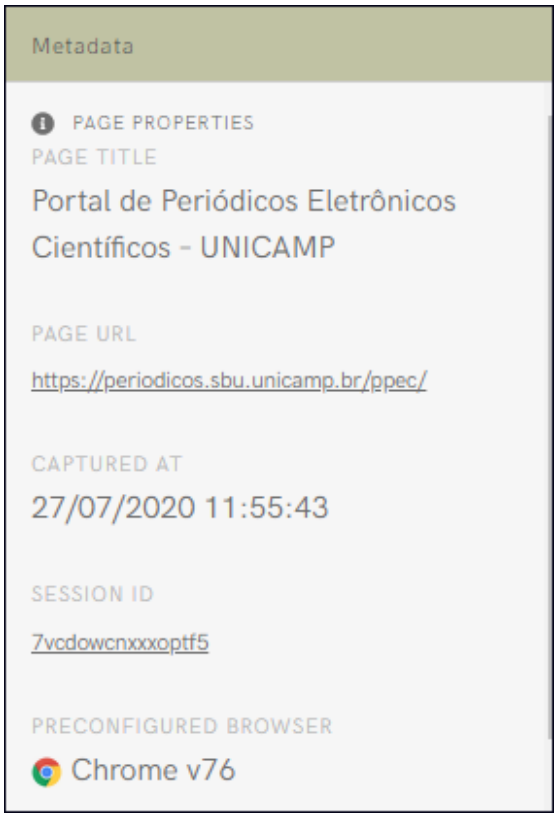
Tal definição pode ser complementada com a de Grácio (2002, p. 23) que entende metadados como o "conjunto de elementos que descrevem as informações contidas em um recurso, com o objetivo de possibilitar sua busca e recuperação", onde recurso se refere a "[...] toda informação que pode ser armazenada em meio eletrônico, podendo estar apresentada como texto, imagem, som, vídeo, página da *Web* etc."

Em relação ao Conifer*,* os metadados de preservação de arquivamento da *Web* são básicos, e não dependem de muita estrutura. Como dito anteriormente, os metadados da estruturação de arquivamento, contém apenas as seguintes informações para a recuperação e utilização na pesquisa das páginas arquivadas:

- nome de usuário (*username*, chamado "criador");
- título (*title*);
- data/hora da captura (*capture date/time*);
- formato do arquivo (*archive file format*);
- título da coleção (*title of collection*);
- URL - *Uniform Resource Locator;*
- ID da sessão (*session ID);* e
- Navegador de captura (*capture browser*).

Na Figura 3, a seguir, apresentamos a página de busca no Conifer com os metadados em destaque:

Figura 3. Tela do Conifer apresentando os metadados de arquivamento



Fonte: *Site* do PPEC no *Rhizome.org* ([2022])

4 Revisão da literatura: arquivamento da *Web* de Portais

De acordo com Vlassenroot e colaboradores (2019), o arquivamento da *Web* "nasce entre os anos 1990 do movimento da preservação digital, comandado por instituições de memória, com objetivo de capturar e preservar registros" e as memórias que a sociedade produzia naquele momento (VLASSENROOT *et al*., 2019, tradução nossa).

São apontadas como desafios do arquivamento da *Web* as questões técnicas. Segundo Pennock (2013, p. 4, tradução nossa), "capturar e arquivar um *site* individual pode ser relativamente simples, isto é, baixá-lo no servidor e armazená-lo de maneira *off-line*. No entanto, o arquivamento da *Web* em larga escala e a longo prazo é um assunto mais complicado". As principais abordagens de arquivamento da *Web* são por depósito legal, comumente em bibliotecas nacionais, seja por coleta automática, por seleção ou de forma combinada (DAY, 2003, p. 3).

Na obra de Brügger (2005), intitulada "*Archiving Websites: general considerations and strategies*", o autor menciona dois tipos de abordagem de arquivamento, o micro e o macro. O "*micro archiving* é efetuado em pequena escala tanto no que diz respeito ao espaço (um número limitado de *sites*) como tempo (um período limitado e isolado)", e pode ser realizado "por indivíduos que não têm à sua

disposição um computador consideravelmente potente e capacidade de armazenamento, cujos conhecimentos técnicos de arquivamento ou de o tratamento subsequente ou é inexistente ou a nível amador". Muitas vezes, com "base numa necessidade imediata, aqui e agora, de preservar um objeto de estudo" (BRÜGGER, 2005, p. 10). Essa foi a abordagem de arquivamento adotada no arquivamento das páginas *Web* do PPEC, por meio do *software* Conifer da *Rhizome* na nossa metodologia (TERRADA, 2022).

Hockx-Yu (2011, p. [2]) frisa que "independentemente da abordagem existe um conjunto de processos essenciais ao arquivamento da *Web* que precisam ser executados e geridos, para assegurar a adequação a qualquer sistema de arquivo da *Web*". Os processos que envolvem o arquivamento da *Web* são a seleção, o *harvesting*, o armazenamento, o acesso e a preservação digital. Discorreremos a seguir sobre o armazenamento, acesso e a preservação digital.

Em relação ao **armazenamento,** ele consiste no processo de retenção de *sites* arquivados de forma segura e confiável. Os formatos de arquivo comumente utilizados para *sites* são ARC e WARC. Ambos os formatos foram desenvolvidos especificamente para arquivos da *Web* (HOCKX-YU, 2011; ISO, 2009, 2017).

Conforme explicitado na norma ISO 28500 (2017), o formato WARC,

> [...] oferece uma forma padrão de estruturar, gerir e armazenar os milhares de milhões de recursos recolhidos a partir da *Web* e de outros locais. É utilizado para construir aplicações para a colheita, gestão, acesso, mineração e troca de conteúdos. Embora represente o formato padrão único dos arquivos *Web*, tem sido adotado para além da comunidade de arquivamento da *Web* para armazenar materiais digitais nascidos ou digitalizados (ISO, 2017, p. 1, tradução nossa).

O formato WARC é uma extensão do formato do arquivo ARC que tem sido tradicionalmente utilizado para armazenar '*crawls*' como sequências de blocos de conteúdo coletados na *World Wide Web*. Na norma ISO 28500:2017 é mencionado que "além do conteúdo primário registrado nos ARC, o formato estendido WARC acomoda o conteúdo secundário", tais como "metadados atribuídos", por exemplo (ISO, 2017; TERRADA, 2022).

Em relação ao **acesso**, este refere-se "à reprodução e disponibilização do acesso, aos *sites* arquivados para os usuários nos arquivos *Web"*. E a **preservação digital** "está relacionada com as normas, melhores práticas e tecnologias, que em conjunto são necessárias para assegurar o acesso aos arquivos da *Web* ao longo do tempo" (HOCKX-YU, 2011, p. [3], tradução nossa).

5 Materiais e métodos

O estudo se caracteriza como uma abordagem da pesquisa de cunho exploratório, baseado na documentação da ferramenta Conifer, com o intuito de preservação digital e arquivamento das páginas *Web* do Portal de Periódicos da UNICAMP, permitindo relatar os procedimentos para o desenvolvimento metodológico para o arquivamento da *Web* do PPEC.

Além disso, a abordagem utilizada foi a de "*microarchiving"*, mencionada por Brügger (2005), visto que o arquivamento é realizado por um indivíduo com a intenção de preservar um determinado objeto de estudo.

Como política de arquivamento, os periódicos do PPEC, vinculados ao Sistema de Bibliotecas da UNICAMP, utiliza o sistema LOCKSS, desenvolvido pela Universidade de *Stanford*, para criar um arquivo distribuído entre as bibliotecas participantes, tal sistema permite que essas bibliotecas criem arquivos permanentes do periódico para fins de acesso contínuo e restauração.

Além da preservação no LOCKSS, os periódicos também fazem parte da Rede Cariniana do IBICT, que fornece preservação para qualquer periódico do sistema OJS no Brasil.

5.1 Procedimentos do arquivamento da Web do PPEC

O processo inicial do PPEC com o arquivamento do seu *website* se deu com sua participação na Rede de Pesquisa Dríade da Rede Cariniana, em que foi apresentado um projeto individual para explanação da seleção das páginas *Web* do Portal de Periódicos da UNICAMP, e descrevê-la de forma que se usasse uma ferramenta proprietária, no caso o Conifer da *Rhizome* e apresentar toda a metodologia de arquivamento de forma experimental.

Ao acessar o *site* do *Rhizome*, com a intenção de conhecer a ferramenta Conifer, e ao se cadastrar, consegue-se de imediato o espaço de capacidade de armazenamento de páginas pessoais ou até mesmo institucionais, um espaço em torno de 5 GB.

A seguir, descrevemos as etapas que sucederam o processo de arquivamento das páginas do PPEC.

5.1.1 Etapa 1 - O espaço para o arquivamento

- Então, em julho de 2020, a empresa *Rhizome* lançou uma pesquisa para conhecer seus usuários, e ao responder o formulário para a empresa até o final de julho desse mesmo ano, obteria gratuitamente um espaço de 20 GB para fazer o uso que desejasse para preservação e arquivamento. Dessa forma, o espaço foi cedido para a realização da pesquisa citada acima para a preservação e arquivamento das páginas do PPEC.

5.1.2 Etapa 2 - Material destinado para o arquivamento

- A intenção foi preservar as páginas dos 33 títulos de periódicos, além do *site* de gerenciamento de todo o PPEC, o *software* OJS. Dessa forma, a expectativa foi a inclusão de todas as páginas do Portal de Periódicos, coletando os sites de cada periódico, no entanto, até o momento conseguimos apenas a realização do arquivamento de 02 (dois) títulos.

5.1.3 Etapa 3 - Operador da captura das páginas e arquivamento

- O trabalho de captura foi realizado por um bolsista sob a coordenação do bibliotecário responsável pelo PPEC. A intenção era preservar as páginas dos 32 títulos de periódicos, além do *site* de gerenciamento de todo o PPEC, o *software* OJS. O processo de captura das páginas no Conifer é considerado fácil e instrutivo pela própria ferramenta, possibilitando a execução dinâmica do arquivamento da *Web*.

5.1.4 Etapa 4 - Metadados das páginas de arquivamento da *Web*

- Neste trabalho, os metadados utilizados foram os metadados descritivos, que são aqueles que detalham um recurso digital para localização, identificação ou compreensão. A realização desta metodologia pelo Conifer, permitiu a captura das páginas para o arquivamento, com a inclusão básica dos metadados pré-definidos no sistema: nome do criador, data de captura, título da página, URL, sessão ID e tipo de *browser*. Esses dados em que os usos primários permitem ser descobertos, apresentados e interoperáveis entre sistemas.

6 Discussão e Resultados

Como resultados, obtivemos até o momento, o arquivamento de 2 (dois) títulos de periódicos e das páginas principais do PPEC. O formato para *download* das páginas, conforme explicado no início, é o WARC, padrão de arquivo de contêiner para armazenar conteúdo da *Web* em seu contexto original, mantido pelo *International Internet Preservation Consortium* (IIPC)[9] que é um dos principais consórcios internacionais que se dedicam a criação de padrões e ferramentas para o arquivamento da *Web*.

Os resultados de termos o arquivamento da *Web* das páginas do Portal de Periódicos da UNICAMP, levou-nos a entender melhor a significância de se ter uma ferramenta, por mais que pareça simples, com características complexas para a preservação digital e arquivamento da *Web* de páginas. Futuramente pode se tornar obsoleta, todavia os dados históricos podem ser relevantes para a formulação de pesquisas.

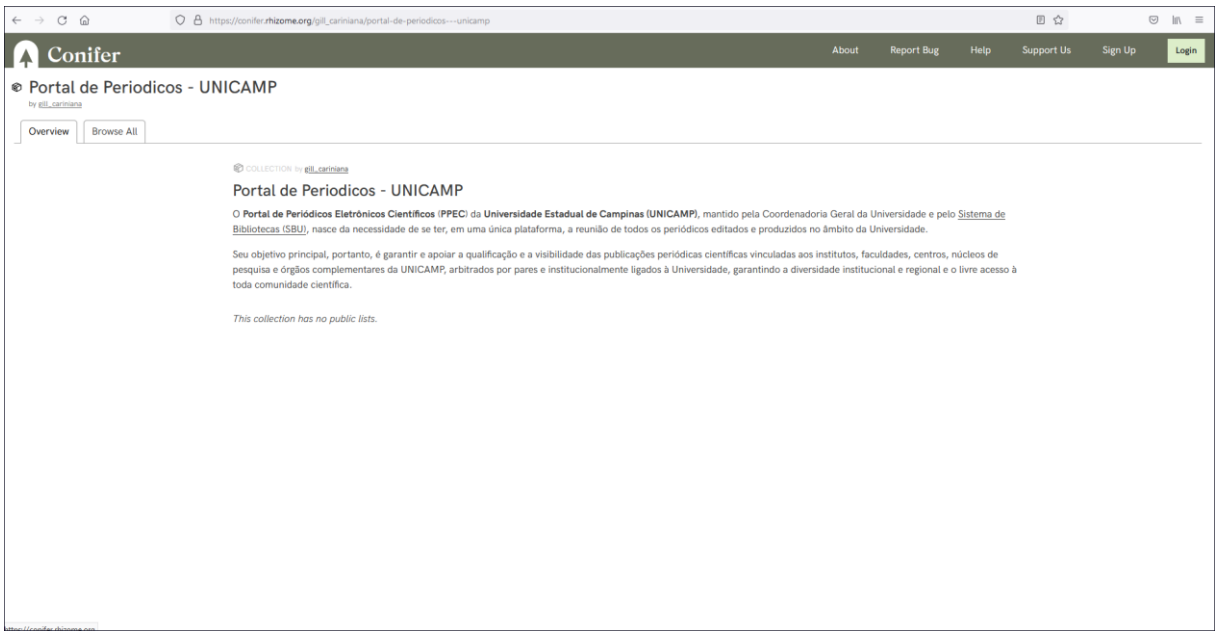
Nesse sentido, obtivemos neste trabalho de arquivamento das páginas do PPEC, etapas que culminaram na realização desse modelo de como realizar o arquivamento da *Web e preservá-las* utilizando o Conifer da *Rhizome*.

Sendo assim, destacamos que podemos obter resultados confiáveis e seguros de um modelo de preservação digital e arquivamento de páginas *Web por meio de* uma ferramenta simples e fácil de usar como o Conifer, *que* é indicado pelo próprio IIPC[10],. Como ilustração desse resultado, apresentamos a seguir as telas do processo de arquivamento da *Web* do PPEC:

[9] Disponível em: http://netpreserve.org/about-us/. Acesso em: 1 jun. 2022.

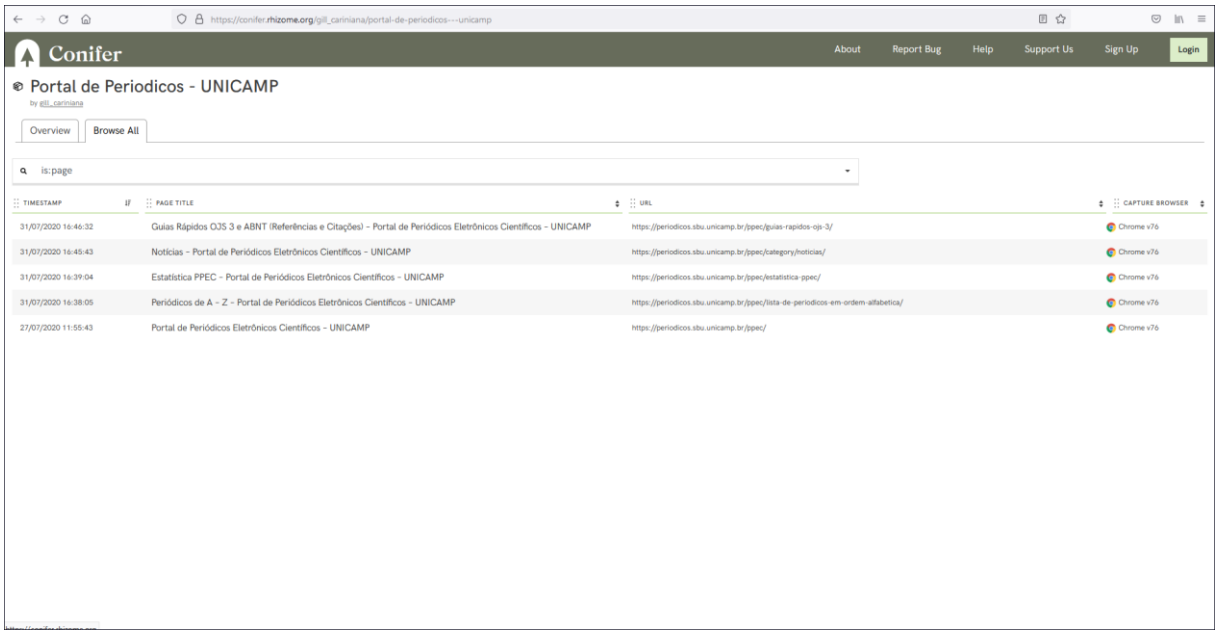
[10] Disponível em: https://netpreserve.org/web-archiving/tools-and-software/. Acesso em: 5 jun. 2022.

Figura 4. Página principal de pesquisa do PPEC no Conifer



Fonte: *Rhizome.org* ([2022])

Figura 5. Página capturas do arquivamento da *Web* PPEC no Conifer

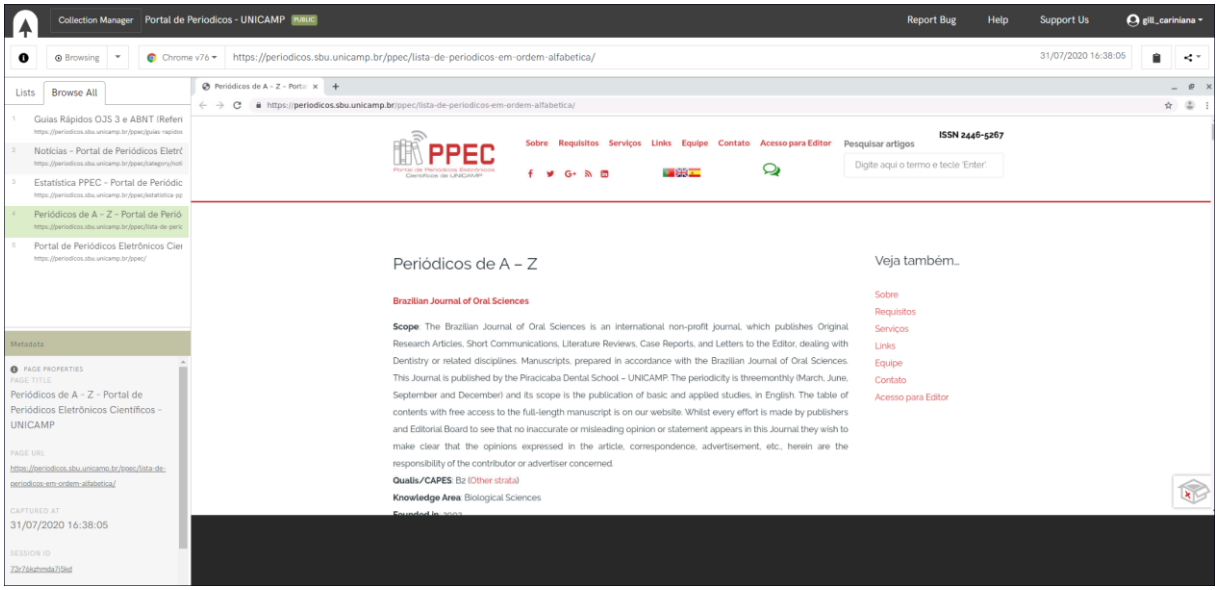


Fonte: *Rhizome.org* ([2022])

Figura 6. Demonstração da página principal do PPEC capturada no Conifer

Fonte: *Rhizome.org* ([2022])

Figura 7. Demonstração da página principal e secundárias dos metadados no Conifer



Fonte: *Rhizome.org* ([2022])

7 Conclusão

Considera-se que este trabalho demandará tempo para sua finalização, mas já é perceptível a sua relevância para a preservação digital e o arquivamento das páginas *Web* do PPEC.

Diante de todo esse relato sobre o processo que gerou um modelo de arquivamento da *Web* das páginas do Portal de Periódicos da UNICAMP, isto nos possibilitou sabermos e entendermos mais sobre os arquivos da *Web*. Mas, afinal, o que são arquivos da *Web* (*web archives*)? Arquivo da *Web* é um registro de recursos da *Web*, e pode incluir *HyperText Markup Language* (HTML) e imagens, *scripts*, folhas de estilo além de vídeo, áudio e outros elementos que compõem as páginas da *Web* e os aplicativos da *Web*, tudo em um arquivo (RHIZOME.ORG, [2022]).

Para mais, os arquivos da *Web* podem, por exemplo, fornecer um contexto melhor do que as capturas de tela isoladas; capturar não apenas o conteúdo, mas os comportamentos dos usuários e seus relacionamentos entre si; e oferecer janelas vívidas em um determinado momento no tempo; contudo, os arquivos da *Web* devem ser criados com muito cuidado e consideração, pois o que nós arquivamos pode se tornar público e, ainda, o que publicamos pode ser usado de forma inesperada (RHIZOME.ORG, [2022]).

Deste modo, *o arquivamento das páginas do PPEC* torna-se importantes por constituírem uma forma de preservar a memória institucional da universidade, ou manter um registro hoje e no futuro da presença da UNICAMP na *Web*; e assim documentar a evolução da universidade, incluindo as suas contribuições para o ensino e a pesquisa, o desenvolvimento de departamentos acadêmicos, unidades administrativas etc.

Também esse modelo pode-se colocar à disposição de administradores, pesquisadores e do público geral, registros universitários. Estes registros podem igualmente estar disponíveis para pesquisadores atuais e futuros de todos os tipos, desde acadêmicos e profissionais até historiadores, jornalistas e aqueles com interesse geral na pesquisa de páginas *Web* contextualizadas na universidade.

Todo esse arquivamento possibilitará que tais pesquisadores futuramente, possam analisar e pesquisar o uso dos arquivos *Web* para ver um histórico visual de como os *sites* mudaram ao longo do tempo.

Enfatizamos por ser uma pesquisa exploratória, com a abordagem no "*microarchiving"*, o arquivamento é realizado por um indivíduo (neste caso sendo uma instituição, referenciando um setor específico), com a intenção de preservar um determinado objeto de estudo (no caso de um Portal de Periódicos dessa instituição).

Por fim, acreditamos que os *sites* constituem um testemunho significativo dos eventos, descobertas etc. do século 21. Todavia, é verossímil que eles sejam criados rapidamente, bem como alterados ou atualizados regularmente em pouco tempo e, muitas vezes, possam ser perdidos completamente para sempre. Portanto, usar ferramentas e sistemas de informação que possibilitem a preservação digital de longo prazo e o arquivamento de páginas *Web*, torna-se primordial e relevante neste contexto. Dessa forma, deixamos aqui esse modelo que poderá ser utilizado por outros portais de periódicos, e como dito acima por toda universidade.

Referências

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, 2010.

ANGELO, E. da S.; OLIVEIRA, M. Estudo altmétrico de repercussão social das revistas científicas brasileiras de acesso aberto. **Bibliotecas**: Anales de Investigación, v. 17, n. 1, p. 14-26, 2021. Disponível em: https://urless.in/iYHLd. Acesso em: 22 maio 2022.

BRÜGGER, N. **Archiving websites**: general considerations and strategies. Århus: The Centre for Internet Research, 2005. Disponível em: https://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf. Acesso em: 22 maio 2022.

DAY, M. Preserving the fabric of our lives: a survey of web preservation initiatives. In: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 7., 2003, Trondheim. **Proceedings** [...]. Trondheim, Norway: Springer Verlag, 2003a. Disponível em: http://www.ukoln.ac.uk/metadata/presentations/ecdl2003-day/day-paper.pdf. Acesso em: 22 maio 2022.

FORMENTON, D.; GRACIOSO, L. de S. Padrões de metadados no arquivamento da Web: recursos tecnológicos para a garantia da preservação digital de websites arquivados. **RDBCI**: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, SP, v. 20, e022001, 2022. DOI: 10.20396/rdbci.v20i00.8666263. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8666263. Acesso em: 5 jun. 2022.

FORMENTON, D.; GRACIOSO, L. de S. Preservação digital: desafios, requisitos, estratégias e produção científica. **RDBCI**: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, SP, v. 18, e020012, 2020. DOI: 10.20396/rdbci.v18i0.8659259. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8659259. Acesso em: 3 jun. 2022.

GRÁCIO, J. C. A. **Metadados para a descrição de recursos da Internet**: o padrão Dublin Core, aplicações e a questão da interoperabilidade. 2002. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, 2002.

HEDSTROM, Margaret. Digital preservation: a time bomb for digital libraries. **Computers and the Humanities**, Netherlands, v. 31, p. 189-202, 1998. Disponível em: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/42573/?sequence=1. Acesso em: 3 jun. 2022.

HOCKX-YU, H. The past issue of the web. *In*: INTERNATIONAL WEB SCIENCE CONFERENCE, 3., 2011, NewYork. **Proceedings** [...]. New York, NY: Association

for ComputingMachinery, 2011. p. 1-8. DOI: https://doi.org/10.1145/2527031.2527050.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. Web archiving. **Tools & software**. [*S. l.*], c2022. Disponível em: https://netpreserve.org/web-archiving/tools-and-software/. Acesso em: 3 jun. 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION - ISO. **BS ISO 28500**: 2009: information and documentation: WARC file format. Switzerland: ISO, 2009.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION - ISO. **BS ISO 28500**: 2017: Information and documentation: WARC file format. Switzerland: ISO, 2017. Disponível em: https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-2:v1:en. Acesso em: 22 maio 2022.

PENNOCK, M. Web archiving. **DPC technology watch report 13-01 March 2013**. Great Britain: DCP, 2013.DOI: http://dx.doi.org/10.7207/twr13-01.Disponível em: https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file. Acesso em: 9 jun. 2021.

PEREIRA, P. C. **Avaliação da usabilidade do Portal de Periódicos Eletrônicos Científicos da UNICAMP**. 2019. 1 recurso online (251 p.) Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas, SP. Disponível em: https://hdl.handle.net/20.500.12733/1636189. Acesso em: 1 jun. 2022.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesauro brasileiro de ciência da informação**. Rio de Janeiro; Brasília, DF: Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 2014. 384 p. Disponível em: http://sitehistorico.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1/copy_of_TESAUROCOMPLETOFINALCOMCAPA24102014.pdf. Acesso em: 3 jun. 2022.

RILEY, J. **Understanding metadata**: what is metadata, and what is it for? Baltimore, Maryland: National Information Standards Organization (NISO), c2017. 45 p. Disponível em: https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf. Acesso em: 1 jun. 2022.

RHIZOME.ORG. **Conifer**. About. [New York, United States], [2022]. Disponível em: https://conifer.rhizome.org/_faq. Acesso em: 3 jun. 2022.

SAMOUELIAN, M.; DOOLEY, J. **Descriptive metadata for web archiving**: review of harvesting tools. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 23 p. Disponível em: https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-harvesting-tools.pdf. Acesso em: 3 jun. 2022.

SANTOS, G. C. Ensaio sobre arquivamento de páginas web: foco na experiência do Portal de Periódicos da UNICAMP, utilizando o Conifer (Rhizome). *In*: SEMINÁRIO

INTERNACIONAL DE PRESERVAÇÃO DIGITAL, 5., 2021, Campinas. **Resumos** [...]. Campinas: UNICAMP: IBICT, 2021. Disponível em: http://eventoscariniana.ibict.br/index.php/sinpred/issue/view/5. Acesso em: 31 maio 2021.

SANTOS, G. C. **Organização, registro e a divulgação do conhecimento científico: metodologia para a criação do Portal de Periódicos Científicos produzido na UNICAMP**. 2012. 82 f. Relatório (Pós-doutorado) - Universidade Estadual de Campinas, Laboratório de Estudos Avançados em Jornalismo, 2012.

SANTOS, G. C. Visibilidade e vantagens na publicação de periódicos em portais institucionais. **Blog PPEC**, Campinas, SP, v.1, n.1, 2017. ISSN 2526-9429. Disponível em: https://periodicos.sbu.unicamp.br/blog/index.php/2017/06/12/portais-2/. Acesso em: 22 maio 2022.

SANTOS, G.C.; CAMARGO, V. R. T. Portal da informação e comunicação: proposta de desenvolvimento do portal de periódicos científicos eletrônicos da Universidade Estadual de Campinas. *In*: CONGRESSO INTERNACIONAL DE CIDADES CRIATIVAS,3., 2013, Campinas, SP. **Actas Icono14.** Campinas, SP: [*S.l.*], 2013. v. 1. p. 1691-1706.

TERRADA, G. A. F. **Preservação digital da web**: uma reflexão sobre política e práticas. 2022. 213 f. Dissertação (Mestrado). Universidade Federal Fluminense, Instituto de Arte e Comunicação Social, 2022. Disponível em: http://dx.doi.org/10.22409/PPGCI.2022.m.11797917706. Acesso em: 22 maio 2022.

VLASSENROOT, E. *et al*. Web archives as a data resource for digital scholars. International Journal of Digital Humanities, London, v. 1, p. 85-111, 2019. DOI: https://doi.org/10.1007/s42803-019-00007-7. Disponível em: https://link.springer.com/article/10.1007/s42803-019-00007-7. Acesso em: 02 fev. 2021.

WEB PAGE. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation, 2022]. Disponível em: https://en.wikipedia.org/wiki/Web_page. Acesso em: 3 jun. 2022.

WEBSITE. *In*: WIKIPEDIA: the free encyclipedia. [San Francisco, CA: Wikimedia Foundation, 2022]. Disponível em: https://en.m.wikipedia.org/wiki/Website. Acesso em: 3 jun. 2022.