

Dia Internacional dos Arquivos 2022

Arquivar a Web: faça-você-mesmo!

Ricardo Basílio, curador digital do Arquivo.pt

Sines, 9 de junho de 2022



Objetivo

- Utilizar ferramentas do Webrecorder.net para **gravar páginas Web** no **próprio computador** num formato normalizado

Os slides correspondem aos conteúdos apresentados na sessão do dia 9 de junho de 2022. Para maior utilidade foi adicionado conteúdo novo:

- Agenda dos tópicos
- Comentários
- Slides sobre o formato WARC (que omitimos no dia da apresentação)

Sendo o conteúdo principal o das demonstrações feitas em vídeos, descreve-se tanto quanto possível o procedimento para gravar páginas Web com o ArchiveWeb.page (Webrecorder.net)

Estes slides podem ser reutilizados ou adaptados para formação de terceiros. Se necessário ajuda ou esclarecimento de alguma questão, envie mail para contacto@arquivo.pt

Agenda

Parte I - Webrecorder.net – tutorial com o ArchiveWeb.page

Parte II - Porque utilizamos o Webrecorder.net

Parte III - Sobre o formato WARC

Parte IV - *SavePageNow* para gravar no Arquivo.pt

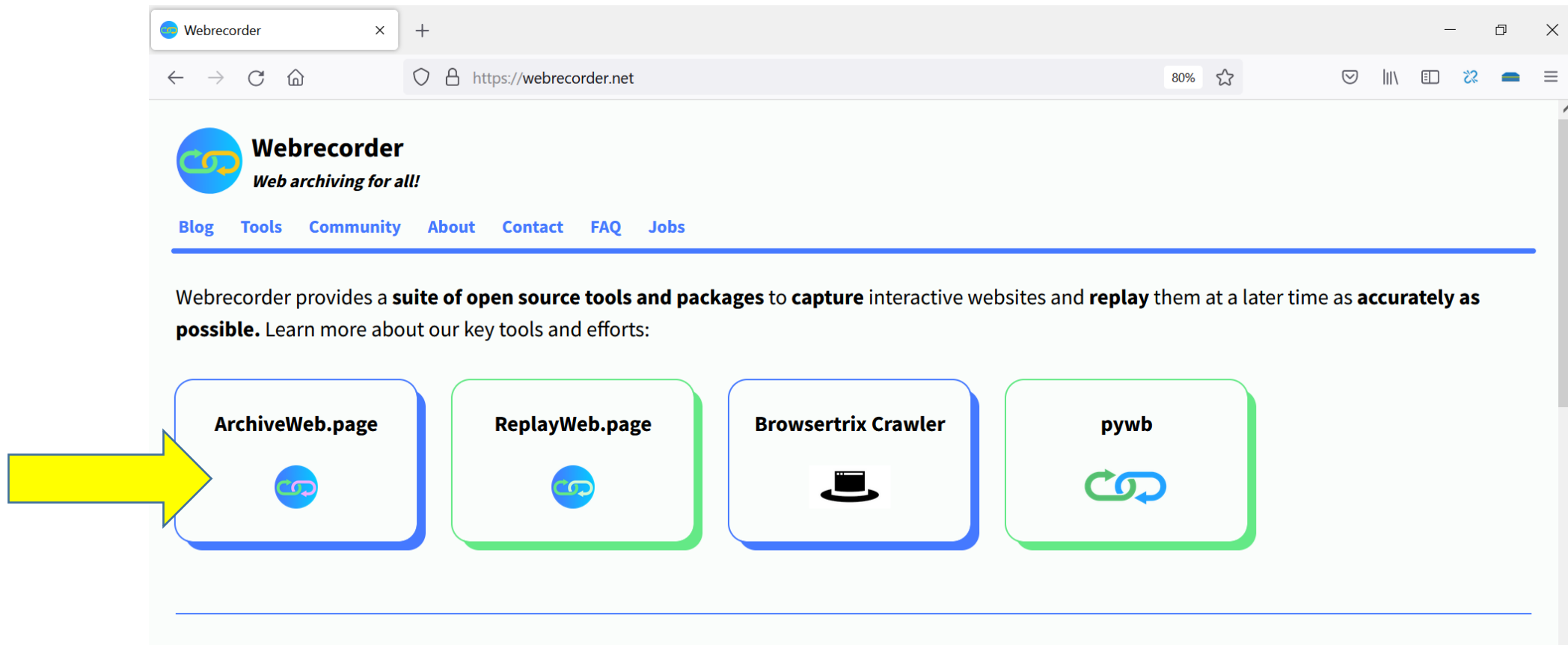
Parte V - Browsertrix para gravar site inteiro

Parte I

Webrecorder.net

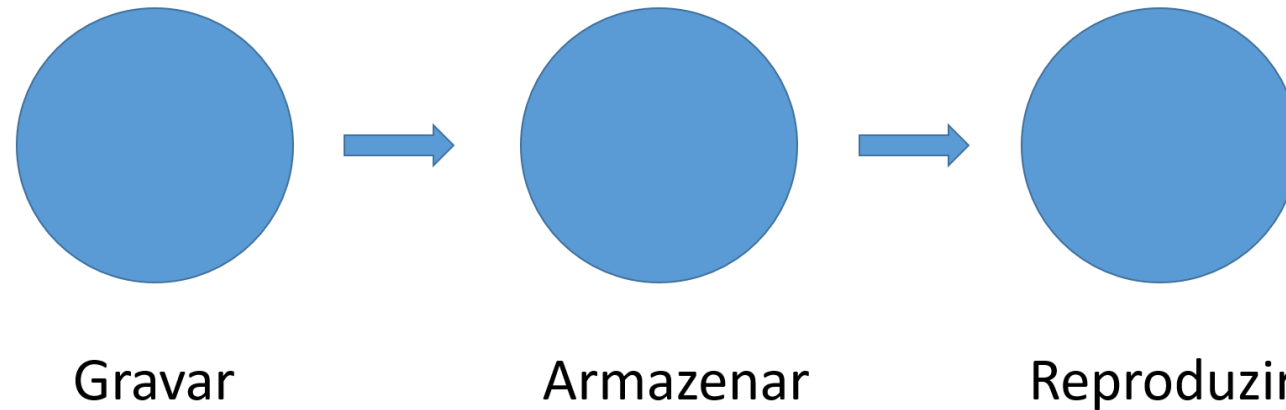
Tutorial – 12 casos de gravação com o ArchiveWeb.page

Ferramentas do Webrecorder.net



[Webrecorder.net](https://webrecorder.net)

3 fases da Preservação da Web



Gravar, Armazenar e Reproduzir são, de forma simplificada, as três fases da preservação (poderíamos falar de outras, tal como a seleção, controle de qualidade).

O ArchiveWeb.page cumpre as três fases: 1) grava 2) gera um WARC e/ou WACZ para armazenar em qualquer lugar 3) reproduz o conteúdo.

É isso que os arquivos da web também fazem. Por isso é muito adequado utilizar o ArchiveWeb.page

Aceder aos vídeos demonstrativos

Tutorial – 12 casos de gravação com o ArchiveWeb.page

Abaixo há informação sobre 12 casos de gravação que foram apresentados e comentados na sessão. Poderá acompanhar o slide correspondente no vídeo da apresentação.

- [Canal Youtube do Arquivo.pt - Módulo D: Arquivar a Web: faça-você-mesmo](#)
- [Vídeo da apresentação em Sines \(Portugal\) no Dia Internacional dos Arquivo.pt](#)



1

Instalar a extensão ArchiveWeb.page

Tutorial – casos de uso do ArchiveWeb.page



Instalar a extensão ArchiveWeb.page

Demora um minuto.

Pré-requisitos: ter um destes browsers – Chrome, Edge, Opera, Brave. Não funciona no Safari e no Firefox.

Se usa sistema operativo iOS, instale primeiro o Edge e no Edge instale a extensão ArchiveWeb.page

Se usa Linux, instale o Chromium.

Há quem não goste de instalar extensões no browser. Para este tutorial vale a pena abrir uma exceção e instalar o ArchiveWeb.page. Remover é tão fácil como instalar.

Descrição do vídeo:

Neste vídeo demonstra-se como instalar o ArchiveWeb.page no browser Chrome:

- 1- encontrar o site <https://webrecorder.net>
- 2- destaque para a máxima “Archiving for all”; quatro ferramentas “open source”.
- 3- escolher o ArchiveWeb.page
- 4- “Install extension from Chrome Web Store” >> Adicionar
- 5- Verificar ícone do ArchiveWeb.page na barra do browser
- 6- Marcar o alfinete “pin” para manter o ícone sempre visível
- 7- Abrir no símbolo da “Homepage” ou “casinha” e pronto para começar

Para remover a extensão

- 1- ícone semelhante a peça de puzzle canto superior direito >> Clicar >> Manage extensions ou Gerir extensões
- 2- botão remover



2

Gravar uma página única

Tutorial – casos de uso do ArchiveWeb.page



Gravar uma página única.

Esta demo explica o processo de fazer um arquivo da Web local

- 1- cria uma simples pasta no ambiente de trabalho “MeuArquivoDaWeb” para armazenar os ficheiros
- 2- explica que o ArchiveWeb.page vai gravar páginas, vai exportar ficheiros para a pasta que criei, e que os ficheiros vão ser reproduzidos onde e quando se quiser pelo ReplayWeb.page ou pelo próprio ArchiveWeb.page

Descrição do vídeo

Gravar uma página única. Duração 01:55

Esta demo explica o processo de iniciar um arquivo da Web local

- 1- cria uma simples pasta no ambiente de trabalho “MeuArquivoDaWeb” para armazenar os ficheiros
- 2- explica que o ArchiveWeb.page vai gravar páginas, vai exportar ficheiros para a pasta que criei, e que os ficheiros vão ser reproduzidos onde e quando se quiser pelo ReplayWeb.page ou pelo próprio ArchiveWeb.page
- 3- mostra o percurso que vamos adotar em todas as gravações: grava, verifica a gravação, faz download do WARC/WACZ

Tutorial – casos de uso do ArchiveWeb.page



Gravar uma página única.

Gravar:

- 1- entramos no ArchiveWeb.page sempre pela “casinha” “homepage”
- 2- criamos uma pasta para a sessão de gravação, “demo-gravar-uma-pagina”
- 3- clicamos no botão azul para gravar
- 4- indicamos o endereço a gravar (https...) e ok
- 5- basta o conteúdo descarregar no browser; a página está gravada
- 6- no canto superior-direito posso acompanhar a evolução da gravação e parar.

Verificar:

- 1- clico na “casinha”, vou à sessão de gravação
- 2- reproduzo a página e verifico se está bem gravada
- 3- experimento um clic num qualquer botão só para lembrar que só grava os links em que clicamos

Download:

- 1- clico na “casinha” e vou à sessão de gravação (este percurso é bom para não nos perdermos; depois cada pessoa usará o ArchiveWeb.page como entender
- 2- adquiero para o meu “MeuArquivoWeb” no meu ambiente de trabalho os ficheiros em WARC e WACZ

Diferença entre WARC e WACZ:

WARC é o formato standard que serve, por exemplo, para integrar em arquivos da Web como o Arquivo.pt

WACZ (Web Archive Collection Zipped) é um ficheiro comprimido ideal para usar com o ArchiveWeb.page. Dentro de um WACZ vamos encontrar um ficheiro WARC acompanhados de um índice dos URLs que ajudam a reprodução das páginas no ArchiveWeb.page. É só isso.



3

Gravar várias páginas

Tutorial – casos de uso do ArchiveWeb.page



Gravar várias páginas

Descrição do vídeo:

Este vídeo segue os passos do anterior e reforça o procedimento para gravar páginas.

Duração: 01:45

Em vez de uma pagina, vamos gravar três páginas, dando mais alguns cliques e ainda um clique como por engano

Assim, gravar:

- 1- abre o ArchiveWeb.page a partir da “casinha”,
- 2- cria uma pasta “Demo-gravar-várias-páginas”
- 3- inicia a gravação da página de partida <https://www.sines.pt>
- 4- dá mais alguns cliques: “Município >>Sobre sines” >> “História de Sines” >> e por engano um clique a mais por engano em “Contactos”
- 5- Controlar gravação no canto superior direito e parar.

Verificar gravação:

- 1- partir da “casinha” “home”
- 2- encontrar a sessão de gravação
- 3- verifica-se que ficou tudo gravado, incluindo “Município >>Sobre sines” e “História de Sines”

Download:

- 1- voltar à “casinha”, “homepage” (para reforçar o percurso e ninguém se perder)
- 2- encontrar sessão de gravação
- 3- Atenção! Posso escolher as páginas que vou descarregar. Seleciono 3 páginas e deixo para trás a página “Contactos” que não pretendo guardar, por exemplo.
- 3- fazer download do WARC/WACZ – adquirir os conteúdos para o “MeuArquivoWeb”



4

Gravar muitas páginas

Tutorial – casos de uso do ArchiveWeb.page

Gravar muitas páginas

Descrição do vídeo:

Gravar muitas páginas

Duração: 01:50

Desta vez vamos gravar muitas páginas, não três, mas cinco e depois outras cinco, só para exemplificar.

Escolhemos como ponto de partida o endereço da página de notícias do site da Câmara Municipal de Sines.

Quantos links tem esta página? Muitos.

Utilizando o Link Grabber, que é um extrator de links de páginas Web, verificamos que nesta página foram encontrados 128 links.

Significa que para gravar todas estas 128 páginas eu teria de clicar pelo menos 128 vezes.

Como fazer para gravar muitas páginas de forma sistemática, progressiva e sem me perder na navegação

Gravação de cinco notícias:

- 1- Os passos são os mesmos do que nos vídeos anteriores: abre, cria sessão de gravação, grava
- 2- Usa esta página específica como ponto de partida (<https://www.sines.pt/pages/385>)
- 3- Em cada notícia, abrir novo separador, sucessivamente sem fechar nenhum
- 5- Gravar 5 notícias para este exemplo
- 6- O habitual... parar gravação, verificar

Continuar a gravação mais notícias, acrescentando-as a esta sessão:

- 7- Novamente fornecer o endereço de partida (<https://www.sines.pt/pages/385>)
- 8- Em cada notícia, abrir novo separador, sucessivamente sem fechar nenhum
- 9- Quando terminar, botão direito do rato num dos separadores abertos e fechar todas os separadores à direita
- 10- Verificar gravação e fazer download do WARC/WACZ

Este procedimento é bom para fazer gravações sistemáticas e de forma controlada em páginas com muitos links e páginas constantemente alimentadas com novos conteúdos, como por exemplo, as do arquivo de notícias num site.



5

Gravar redes sociais

Twitter

Tutorial – casos de uso do ArchiveWeb.page

Gravar redes sociais – Twitter

Descrição do vídeo:

Processo como nos anteriores
Abre, cria sessão de gravação, gravar,

Neste caso, gravar a página do Twitter institucional, de fora para dentro, ou seja, sem estar “logado”
Neste caso o link é <https://twitter.com/cmsines>

A página do Twitter apresenta os conteúdos verticalmente, à medida que se faz scroll down.
O ArchiveWeb.page tem o botão “Auto pilot”, piloto automático, que rola automaticamente a página e descarrega os conteúdos do post. Não segue links a apontar para fora do Twitter, a não ser que se faça manualmente (abrindo novo separador em cada link que cujo conteúdo queiramos gravar).

Sem login gravamos o conteúdo visível publicamente, evitando recolher outros dados, como por exemplo dados de perfil.
A ter em conta: ao gravar conteúdos sem fazer login podem surgir banner à frente do conteúdo, quando se grava e quando se reproduz.
A gravação sistemática seguindo este método “de fora para dentro”, semanalmente, por exemplo, seria suficiente para ter uma boa representação das publicações.

O resto do processo é como nos exemplos anteriores: verificar a gravação, fazer download.



6

Gravar redes sociais Twitter com login

Gravar redes sociais - Twitter com login

Descrição do vídeo:

Neste vídeo mostra-se o caso da gravação de uma página do Twitter com login.

Há alguns dados de perfil (foto, canais de que é seguidor, etc). Gravamos de dentro para dentro. Não aparecem banners, como acontecia no caso anterior em que estávamos gravar “de fora para dentro”.

No caso de gravar “de dentro para dentro”, deve fazer login antes e começar a gravar depois. Não faça login enquanto está a gravar, porque a sua interação e os seus dados de login podem ficar escritos no ficheiro WARC.

Ao fazer gravação “de dentro para dentro” deve considerar esses conteúdos não públicos. São portanto conteúdos não adequados para acesso público. Podem ser guardados à parte para consulta interna ou pessoal.

O que se refere ao Twitter com login também se aplica a outras redes sociais.

7

Gravar redes sociais Facebook

Tutorial - casos de uso do ArchiveWeb.page

Gravar redes sociais – Facebook

Duração: 01:46

Descrição do vídeo:

O procedimento é semelhante ao Twitter.

O endereço de partida é, por exemplo, <https://www.facebook.com/municipiodesines/>

O Facebook é gravável “de fora para dentro”, seguindo o mesmo procedimento como no caso do Twitter. No entanto, a reprodução dos conteúdos gravados é muito limitada. Essa é uma dificuldade partilhada pela comunidade dos arquivos da Web. Todos apontam da dificuldade de gravar e reproduzir páginas do Facebook.

Ponderação da questão:

- o foco do Facebook não é a preservação digital
- alimenta-se do conteúdos e das interações dos utilizadores
- há alguns uns anos era possível ao dono da conta descarregar um ficheiro HTML com todos os posts na ordem em que foram publicados
- E atualmente? Como ter uma cópia da sequência de posts publicados? Como posso obter de volta uma cópia dos conteúdos no seu contexto (página do Facebook)?

Contudo, muitas instituições ou serviços no interior das instituições têm no Facebook o seu principal canal, por vezes publicando conteúdos em exclusivo no Facebook.

A prática adequada seria as instituições publicarem versões desses conteúdos em Websites ou em plataformas que se possam gravar.

Incentivamos os mais empreendedores a gravarem pequenas porções de conteúdos do Facebook com o ArchiveWeb.page, de modo a avaliarem o resultado e a decidirem o que fazer para assegurarem a sua preservação.

8

Gravar redes sociais e outras plataformas ISSUU

Tutorial – casos de uso do ArchiveWeb.page



Gravar redes sociais e outras plataformas (ISUU)

Duração:01:05

Descrição do vídeo:

Este vídeo mostra que outros canais e plataformas são graváveis com o ArchiveWeb.page

A gravação de conteúdos publicados em plataformas serve principalmente para mostrar o contexto em que a comunicação institucional foi feita.

Gravamos, por exemplo, o último número da revista “Sines em Agenda” e verificamos que podemos reproduzir esse conteúdo arquivado.

Assim, garantimos mais uma forma de preservação. É bom ter a revista preservada também no formato arquivo da Web. Quanto mais forem os formatos melhor para a preservação dos conteúdos.



9

Gravar vídeos embebidos nos websites

Tutorial – casos de uso do ArchiveWeb.page

Gravar vídeos embebidos nos websites

Duração: 01:15

Descrição do vídeo “demo10-sines-gravação-de-vídeos-embebidos”:

Procedimento semelhante, abreviados alguns passos.

Atenção! Deixar correr o vídeo. Não é necessário esperar até ao fim para descarregar o conteúdo.

O site da Rádio Sines tem vídeos das suas emissões rádio embebidos no website.

Esta prática é bastante frequente em sites de media: rádio, televisão, jornais e até projetos de comunicação que misturam os três media.

Este vídeo mostra que os vídeos embebidos nos Websites são graváveis.

No entanto, devemos ter em conta que nem sempre são reproduzíveis, pois o formato em que foram publicados e gravados nem sempre é compatível com a ferramenta usada mais tarde para reproduzir. Esta dificuldade é partilhada por toda a comunidade dos arquivos da Web.

Assim, devemos considerar a gravação de vídeos como complementar a outras formas de preservação desses recursos. Quando gravamos o vídeo embebido na página Web estamos a preservar o contexto em que foi publicado (se está em destaque na página, se tem legenda, se está associado a uma notícia, etc.

Para gravar bem recomenda-se publicar bem os vídeos, em formatos abertos, bastante utilizados, em plataformas abertas tanto quanto possível, indicando um link para download do ficheiro de vídeo, se possível, etc.

Texto, áudio e vídeo são agora conteúdos de páginas Web. Este tipo de websites são um desafio para a preservação. Como gravá-los? Pode um arquivo de televisão, por exemplo, dispensar gravar também as páginas Web do canal?



10

Gravar vídeos embebidos do Youtube

Gravar vídeos embebidos do Youtube

Descrição do vídeo:

O site do festival Festival Músicas do Mundo, ex-libris de Sines, apresenta um vídeo embebido do Youtube.

Os vídeos embebidos do Youtube são muito frequentes. No entanto, o Youtube é um serviço externo e tende a ser fechado. Em tempos permitiu a gravação dos vídeos através de um software específico (chamado DLL) que depois pediu para ser proibido (em 2018, alegando que este estava a ser usado por outros serviços para difundir conteúdos, violando os direitos. Apesar de não visar os arquivos da Web, estes foram afetados por essa limitação.

O que acontece ao gravar o Youtube com o ArchiveWeb.page?

Recomenda-se remover da lista de extensões pré-instaladas no browser a do Youtube, se ainda lá estiver.

Verifica-se que a gravação corre bem, que o vídeo é gravado rapidamente.

No entanto, a sua reprodução posterior tende a não funcionar. Experimente fazer download e reproduzir com o ArchiveWeb.page, por exemplo.

Não funciona, “playback error”.

Vale a pena gravar vídeos do Youtube?

Na minha opinião sim, com critério, tendo em conta os recursos e o conteúdo em causa.

Se está gravado pode haver forma de reproduzir, ficando pelo menos essa possibilidade.

O contexto em que o vídeo foi embebido e publicado num site é uma informação complementar. Mesmo que o vídeo não esteja reproduzível, fica registado o seu endereço no Youtube. Assim, se ainda estiver online é possível encontrá-lo usando esse metadado.

Tal como para outras redes sociais, é recomendável que os vídeos publicados no Youtube sejam preservados em outros lugares e com outra estratégia.

11

Reproduzir os conteúdos gravados

Tutorial – casos de uso do ArchiveWeb.page



Reproduzir os conteúdos gravados (Novidade!) em contexto totalmente local

Duração: 01:06

Descrição do vídeo:

Como resultado das gravações feitas nos casos que mostrámos neste tutorial, a pasta “MeuArquivoWeb” contém agora vários ficheiros.

Recordamos a diferença entre os ficheiros WARC e os ficheiros WACZ:

- WARC são os ficheiros standard, iguais aos utilizados pelo Arquivo.pt ou o Internet Archive
- WACZ (Web Archive Collection Zipped) são ficheiros que têm dentro o ficheiro WARC e ainda um índice com os endereços gravados, o que é muito útil para serem reproduzidos com o ReplayWeb.page ou com o ArchiveWeb.page.

Já referimos que para reproduzir conteúdos podemos usar o ReplayWeb.page. Importamos ou carregamos para lá o ficheiro que queremos reproduzir e já está. De preferência usar os ficheiros WACZ. Atenção! Não funcionou como esperado quando tentei reproduzir alguns ficheiro. Por isso, optei por instalar o ArchiveWeb.page Desktop App.

Vamos demonstrar como reproduzir os conteúdos gravados em contexto totalmente local:

Para isso é necessário:

- 1- ir a <https://webrecorder.net>
- 2- instalar a versão ArchiveWeb.page Desktop App – é um ficheiro executável, escolher a última versão
- 3- abrir o ArchiveWeb.page Desktop App – verifica que é semelhante à extensão de browser que usámos para gravar
- 4- importar ou carregar o ficheiro que quer reproduzir (de preferência o ficheiro WACZ)
- 5- reproduzir

Tutorial – casos de uso do ArchiveWeb.page



Ao utilizar o ArchiveWeb.page Desktop App podemos ver que estamos em ambiente totalmente local (ver caminho do ficheiro na parte superior do monitor).

É altura ainda de ver o maior ou menor sucesso das nossas gravações. Há conteúdos cuja reprodução não funciona. Essas limitações uma vez identificadas são importantes para decidir o que fazer ou deixar de fazer.

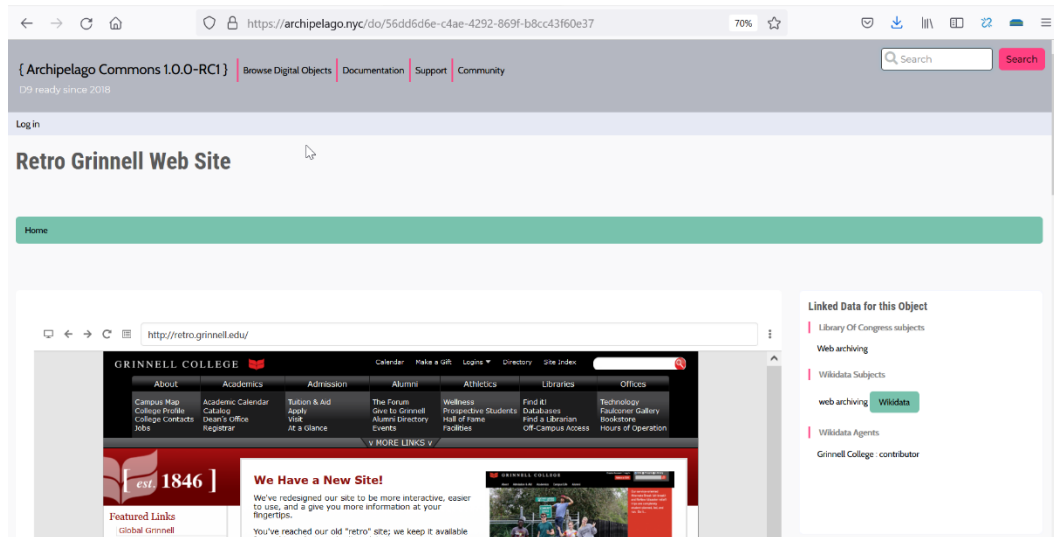
Não podemos esquecer que a preservação digital é o “jogo do gato e do rato” das tecnologias. A preservação deve ser feita colaborativamente, de forma complementar, entre instituições diversas, com estratégias e técnicas diversas.

.

12

Reproduzir sites preservados
num repositório digital
ou no meu site

Integração do ArchiveWeb.page num repositório digital



<https://archipelago.nyc/>

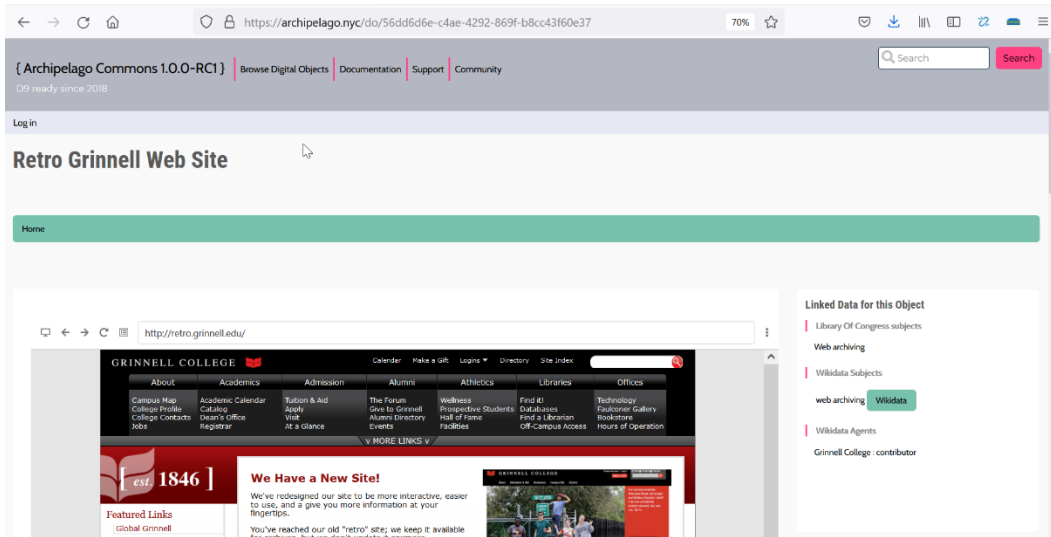
Este slide pretende mostrar que é possível incluir os novos formatos de arquivo da Web (WARC) em repositórios e bibliotecas digitais. Para isso mostramos exemplos de projetos nesse sentido.

Os websites consideram-se novos formatos, nascidos digitais. No entanto, não têm ainda lugar nos repositórios digitais, bibliotecas digitais e outros serviços semelhantes. Estes disponibilizam simultaneamente PDFs, vídeo, áudio, jogos, modelos 3D, etc, O mesmo não acontece com ficheiros WARC.

Para colmatar essa lacuna começam a surgir casos de integração do ReplayWeb.page em repositórios.

O ReplayWeb.page é o software do projeto Webrecorder.net que reproduz os ficheiros do tipo WARC e WACZ, como já referimos.

Integração do ArchiveWeb.page num repositório digital



<https://archipelago.nyc/>

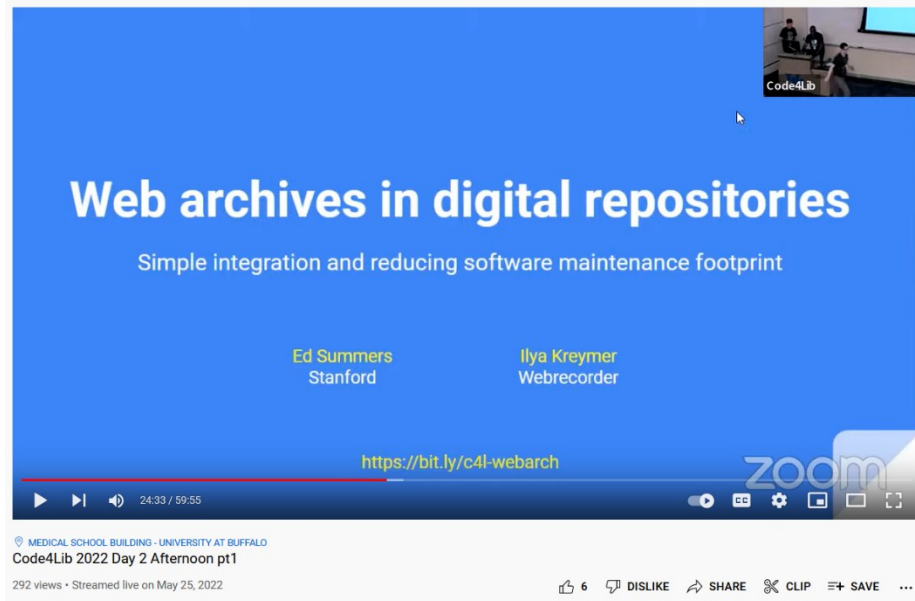
Eis um exemplo de integração: <https://archipelago.nyc/>
Neste repositório digital é possível aceder a sites gravados com o ArchiveWeb.page, do mesmo modo que é possível aceder a outros formatos mais habituais (PDF, JPEG, etc)

Diferentes objetos são descritos com metadados que lhes são próprios (costumizados), o mesmo acontecendo com os websites gravados. A reprodução das páginas Web gravadas é apresentada numa frame no contexto do repositório.

O grande benefício de projetos deste género é trazer para a agenda das instituições a preservação dos seus sites e de lhes reconhecer um valor cultural e patrimonial que atualmente não têm.

A integração do ArchiveWeb.page em websites está documentada no site Webrecorder.net e é relativamente simples.

Integração do ArchiveWeb.page num repositório digital



<https://youtu.be/dtd5Os5t0Io>

Este slide é uma referência à apresentação de Ed Summers (Univ. Stanford) e Ilya Kreymer (Webrecorder) sobre a integração do ArchiveWeb.page.

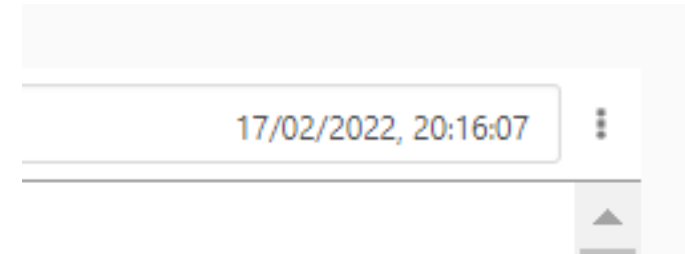
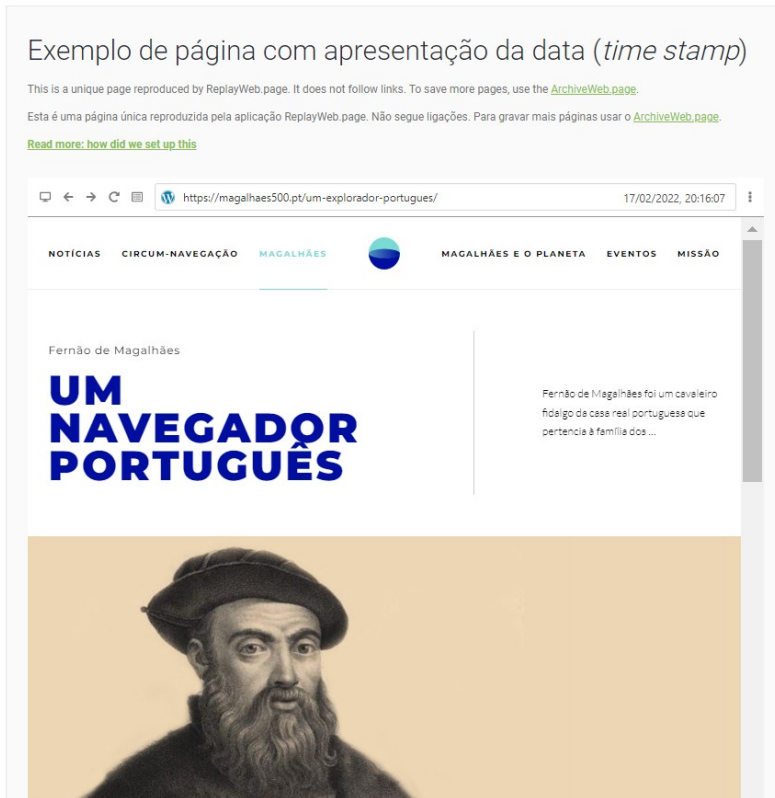
É apresentação útil para quem quiser incluir páginas preservadas num repositório institucional.

Cabe também aos desenvolvedores de produtos de repositórios e biblioteca digitais a tarefa de integrarem estas novas ferramentas.

Se for possível reproduzir um ficheiro WARC ou WACZ num repositório como se reproduz um PDF, muitas instituições passarão a fazer arquivo dos seus websites ou de páginas desses sites.

Quando isso acontecer será um grande avanço para a preservação de conteúdos Web.

Integração do ArchiveWeb.page num repositório digital



Testámos algo mais simples num site wordpress -

<https://webcurator.ddns.net/?p=291>

Uma página gravada pode ser disponibilizada da mesma forma que outros conteúdos embebidos em websites.

Documentação em <https://replayweb.page/docs/embedding>

Página gravada do site das comemorações dos 500 anos da Circum-navegação integrada num site Wordpress:
<https://webcurator.ddns.net/?p=291>

Parte II

Porque utilizamos o Webrecorder.net

O Webrecorder

“Web archiving for all”

- 2016 - Parte do Projeto Rhizome.org (Webrecorder.io)
- 2019 - Independente do Projeto Rhizome, dá suporte ao serviço Conifer
- 2020 - Webrecorder.net

Canal Twitter: [@webrecorder_io](https://twitter.com/webrecorder_io)

O Webrecorder.net

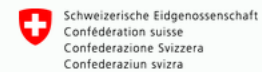
- Open source
- Centrado no utilizador não especializado
- Formatos normalizados
- Compatibilidade com arquivos da Web
- Apoio do International Internet Preservation Consortium (IIPC)



O Webrecorder.net

Partners

Here are some of the partners that we work with or have collaborated with in the past:

Perma.ccLOCKSSRHIZOMETHE NATIONAL ARCHIVEShypothes.isIIPC
INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUMCLOCKSSSchweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizraNATIONAL
LIBRARY
OF AUSTRALIAKIWIXNFB
FNOUKWA
UK WEB ARCHIVEDocNowARQUIVO.PTSTANFORD
UNIVERSITY PRESS

If you're an institution using our tools, let us know and we can add your logo here.

O Webrecorder.net



Este slide repete o que apresentamos no início para concluir que:

- As ferramentas do Webrecorder.net que utilizámos, mais propriamente o ArchiveWeb.page dão-nos a possibilidade de cumprir as três fases da preservação: gravar, armazenar e reproduzir,
- Tudo isso formato normalizado WARC,
- Podendo ser reproduzido através de software de arquivos da Web geralmente conhecido por Wayback

Parte III

Sobre o formato WARC

Formato WARC

- ISO 28500:2017
- Independente de plataforma
- Reproduzido por software específico (geralmente designado *wayback*)

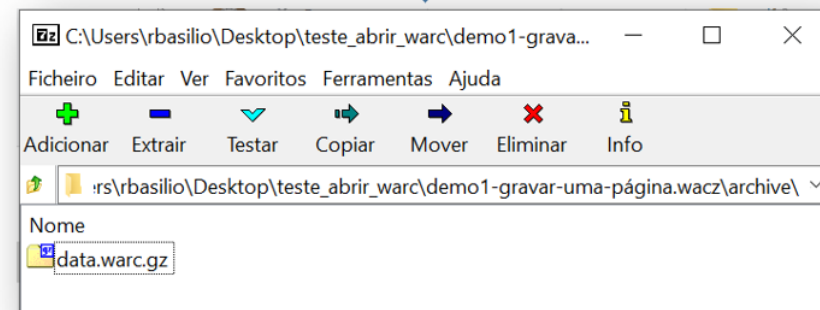
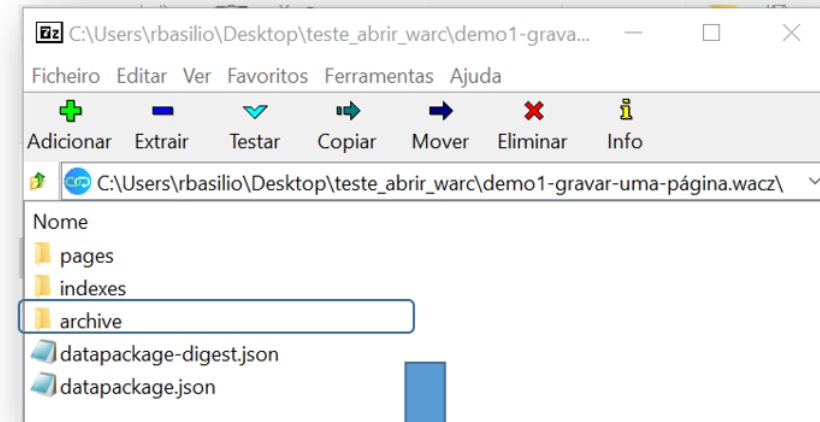
WARC e WACZ

WARC – standard para ser lido por qualquer arquivo da Web

WACZ – criado pelo Webrecorder.net para ser lido por ferramentas específicas como o ArchiveWeb.page ou o ReplayWeb.page

WARC e WACZ

Name	Date modified
 demo1-gravar-uma-página.wacz	08/06/2022 17:23



Nome	Tamanho
 data.warc	1 723

Dentro de um WACZ
há sempre um WARC

Figura que demonstra como pode encontrar o ficheiro WARC dentro de um WACZ que o ArchiveWeb.page exporta

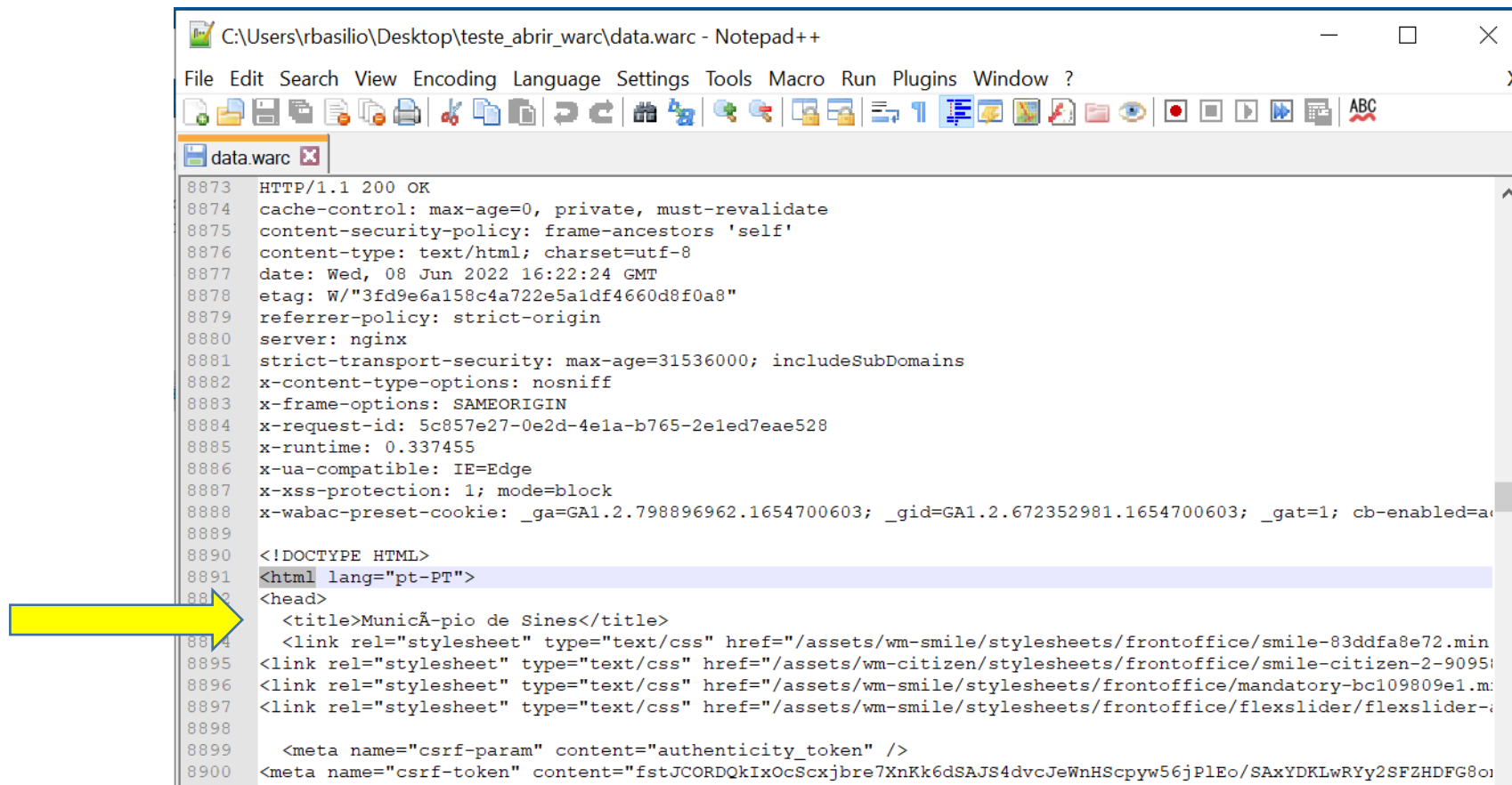
WARC e WACZ



```
C:\Users\rbasilio\Desktop\teste_abrir_warc\data.warc - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
data.warc
1 WARC/1.1
2 WARC-Record-ID: <urn:uuid:dbb3c348-61da-5314-ad2c-c3cae121blaf>
3 WARC-Filename: demo1-gravar-uma-p%C3%AAlgina.wacz#/archive/data.warc.gz
4 WARC-Date: 2022-06-08T16:22:05.389Z
5 WARC-Type: warcinfo
6 Content-Type: application/warc-fields
7 Content-Length: 134
8
9 software: Webrecorder ArchiveWeb.page 0.7.9, using warcio.js 1.5.0
10 format: WARC File Format 1.1
11 isPartOf: demo1-gravar uma pÃgina
12
13
14 WARC/1.1
15 WARC-Record-ID: <urn:uuid:b4ac2271-2aa8-5fdd-80cc-aaf9f86aa69b>
16 WARC-Page-ID: z3fnpjflmjsg83mamacr8
17 WARC-Target-URI: https://m.addthis.com/live/red_lojson/300lo.json?si=62a0ccb9634cec
18 WARC-Date: 2022-06-08T16:22:17.831Z
19 WARC-Type: response
20 Content-Type: application/http; msgtype=response
21 WARC-Payload-Digest: sha256:ebf4f635a17d10d6eb46ba680b70142419aa3220f228001a036d311a
22 WARC-Block-Digest: sha256:d8856aad0ddbca63ff5a8c6457984696ab69e516102583db4700d1f8c
23 Content-Length: 399
```

Ficheiro WARC aberto com NotePad++ onde se pode diversos sobre a gravação
(nome do ficheiro, data, versão do software que gravou, etc)

WARC e WACZ



```

C:\Users\rbasilio\Desktop\teste_abrir_warc\data.warc - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
data.warc
8873 HTTP/1.1 200 OK
8874 cache-control: max-age=0, private, must-revalidate
8875 content-security-policy: frame-ancestors 'self'
8876 content-type: text/html; charset=utf-8
8877 date: Wed, 08 Jun 2022 16:22:24 GMT
8878 etag: W/"3fd9e6a158c4a722e5a1df4660d8f0a8"
8879 referrer-policy: strict-origin
8880 server: nginx
8881 strict-transport-security: max-age=31536000; includeSubDomains
8882 x-content-type-options: nosniff
8883 x-frame-options: SAMEORIGIN
8884 x-request-id: 5c857e27-0e2d-4e1a-b765-2e1ed7eae528
8885 x-runtime: 0.337455
8886 x-ua-compatible: IE=Edge
8887 x-xss-protection: 1; mode=block
8888 x-wabac-preset-cookie: _ga=GA1.2.798896962.1654700603; _gid=GA1.2.672352981.1654700603; _gat=1; cb-enabled=a
8889
8890 <!DOCTYPE HTML>
8891 <html lang="pt-PT">
8892 <head>
8893   <title>Munic pio de Sines</title>
8894   <link rel="stylesheet" type="text/css" href="/assets/wm-smile/stylesheets/frontoffice/smile-83ddfa8e72.min
8895   <link rel="stylesheet" type="text/css" href="/assets/wm-citizen/stylesheets/frontoffice/smile-citizen-2-9095
8896   <link rel="stylesheet" type="text/css" href="/assets/wm-smile/stylesheets/frontoffice/mandatory-bc109809e1.m
8897   <link rel="stylesheet" type="text/css" href="/assets/wm-smile/stylesheets/frontoffice/flexslider/flexslider-
8898
8899   <meta name="csrf-param" content="authenticity_token" />
8900   <meta name="csrf-token" content="fstJCORDQkIxOcScxjb7XnKk6dSAJS4dvcJeWnHScpyw56jPlEo/SAXYDKLwRYy2SFZHDFG8oi

```

Ficheiro WARC aberto com NotePad++, gravação do site da Câmara Municipal de Sines com o ArchiveWeb.page, onde se pode ver por exemplo o código HTML

Parte IV

SavePageNow
para gravar no Arquivo.pt

Parte IV – SavePageNow para gravar no Arquivo.pt

Depois de termos apresentado o ArchiveWeb.page, ferramenta do Webrecorder.net que permite iniciar um arquivo da Web local, vamos agora demonstrar o SavePageNow.

O SavePageNow é um serviço de gravação de páginas Web na hora que o Arquivo.pt criou para qualquer pessoa poder usar, em qualquer local sem ter que instalar nenhuma aplicação. Qual a relação com o ArchiveWeb.page? Usa o mesmo software e grava de modo semelhante.

O SavePageNow, podemos dizer desta forma, é como o ArchiveWeb.page mas do lado Arquivo.pt. O que gravarmos fica imediatamente arquivado e, em 48 horas, está disponível no Arquivo.pt.

Funciona desta forma:

- Encontrou uma página que quer gravar? Copie o link
- Abra o SavePageNow (Grave páginas) do Arquivo.pt – Encontra-o no menu lado direito ou em arquivo.pt/savepagenow
- Cole o link e comece a gravar como fez no ArchiveWeb.page
- E pronto. Os conteúdos são imediatamente gravados no Arquivo.pt e disponibilizados em 48 horas.

Atenção! Nem sempre os conteúdos de uma página Web que estamos a ver ficam gravados. O streaming de vídeos, por exemplo, foi limitado no SavePageNow para evitar carregar o sistema.

Parte IV – SavePageNow para gravar no Arquivo.pt

Descrição do vídeo:

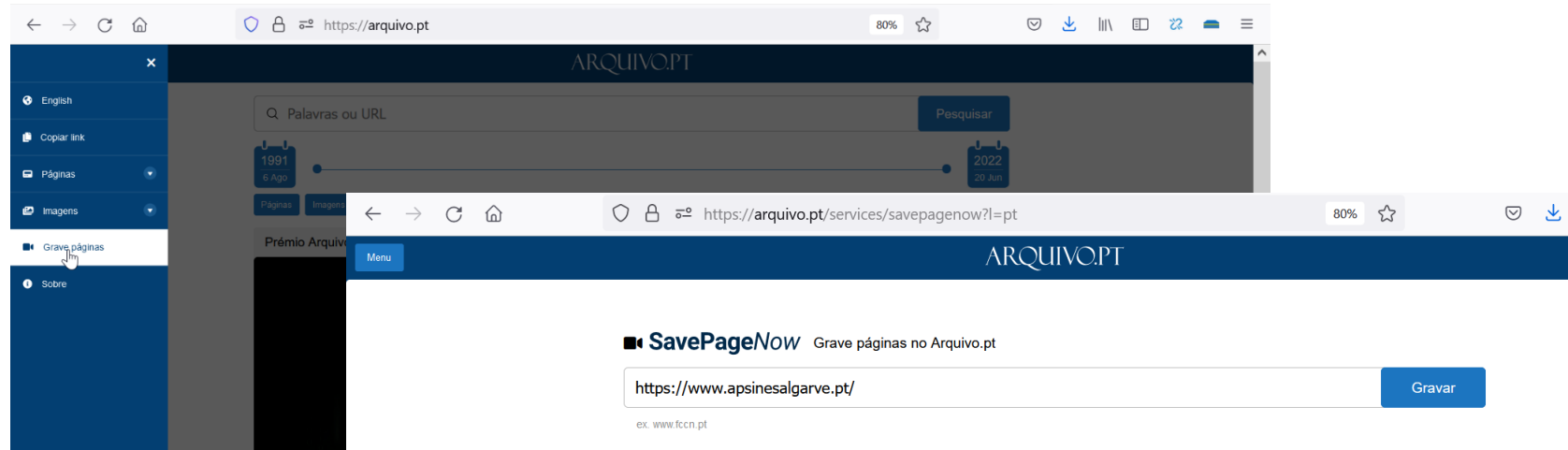
Neste vídeo mostramos exemplificamos o uso SavePageNow

Duração: 01:38

- Copiei o link <https://www.apsinesalgarve.pt/> É o site da Administração dos Portos de Sines
- Site do Arquivo.pt >> Menu lado direito SavePageNow ou Grave páginas
- Colei o link e comecei a “Gravar”
- Sigo as páginas ou os links que quero gravar, tal como no ArchiveWeb.page
- Para gravar de forma controlada, sem me perder na navegação, sigo links abrindo novo separador
- Dica! Sabia que pode abrir página em novo separador com o rato clicando no botão do meio ou na rodinha de scroll
- E pronto. Os conteúdos são imediatamente gravados no Arquivo.pt e disponibilizados em 48 horas.

Atenção! Nem sempre os conteúdos de uma página Web que estamos a ver ficam gravados. O streaming de vídeos, por exemplo, foi limitado no SavePageNow para evitar carregar o sistema.

SavePageNow para gravar no Arquivo.pt



SavePageNow Grave páginas no Arquivo.pt

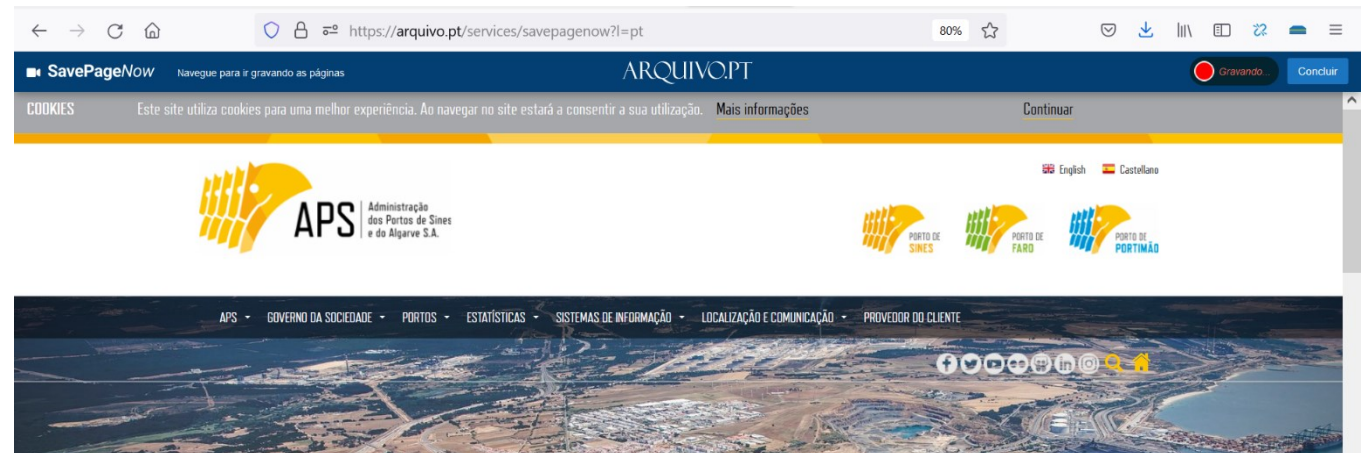
<https://www.apsinesalgarve.pt/>

Gravar

ex. www.fccn.pt

O serviço SavePageNow permite arquivar uma página no exato momento em que o utilizador faz o pedido.

Os conteúdos arquivados serão posteriormente integrados no acervo do Arquivo.pt.



No seu computador ou smartphone crie um atalho para

arquivo.pt/savepagenow

Parte V

Browsertrix

para gravar site inteiro

Browsertrix

Parte V - Browsertrix para gravar site inteiro – introdução

Web crawler

Baseado em Browser

Vários perfis de recolha

- Single page
- Todos os links de um domínio
- Profundidade personalizada

Fácil de implementar para recolhas em pequena escala

Browsertrix

Parte V - Browsertrix para gravar site inteiro - introdução

O Browsertrix serve para gravar um site web inteiro de forma automática. É uma ferramenta do Webrecorder.net.

Para gravar sites com milhares de páginas o ArchiveWeb.page é pouco prático. Teríamos de clicar em todos os links e esperar que todos os conteúdos fossem descarregados. Além de demorado, certamente ficariam muitos conteúdos esquecidos.

O Browsertrix é um crawler e por isso uma das suas funcionalidades é descobrir todos os links existentes num site e proceder à sua gravação, se assim quisermos.

Browsertrix

Parte V - Browsertrix para gravar site inteiro – características, comentário e descrição do vídeo

Dizemos que a forma de gravar a Web é *browser based* ou seja baseada num browser, diferentemente da forma mais antiga de gravar websites na qual apenas se descarregavam ficheiros e em que não era possível gravar certas interações e dinamismos que só acontecem quando se acede a um site num browser.

Destacamos os perfis ou modos de gravação que podemos personalizar. Ou seja,

- Posso gravar apenas a página cujo endereço eu indicar – por vezes é muito útil
- Posso gravar todos os links do mesmo domínio, que corresponde frequentemente a um website
- Posso definir a profundidade, indicando por exemplo, que quero gravar uma página e seguir os links lá existentes apenas uma vez

A implementação é fácil para treino e aprendizagem.

Para usar em produção num arquivo da Web local recomenda-se vivamente o envolvimento dos técnicos informáticos.

Browsertrix

Parte V - Browsertrix para gravar site inteiro – características, comentário e descrição do vídeo

Descrição do vídeo “Recolha de website utilizando o Browsertrix (6 min):

Este vídeo exemplifica o uso do Browsertrix, neste caso, ainda numa versão antiga (2018)

Ver no Youtube com narração: <https://youtu.be/7JNoMo302NA>

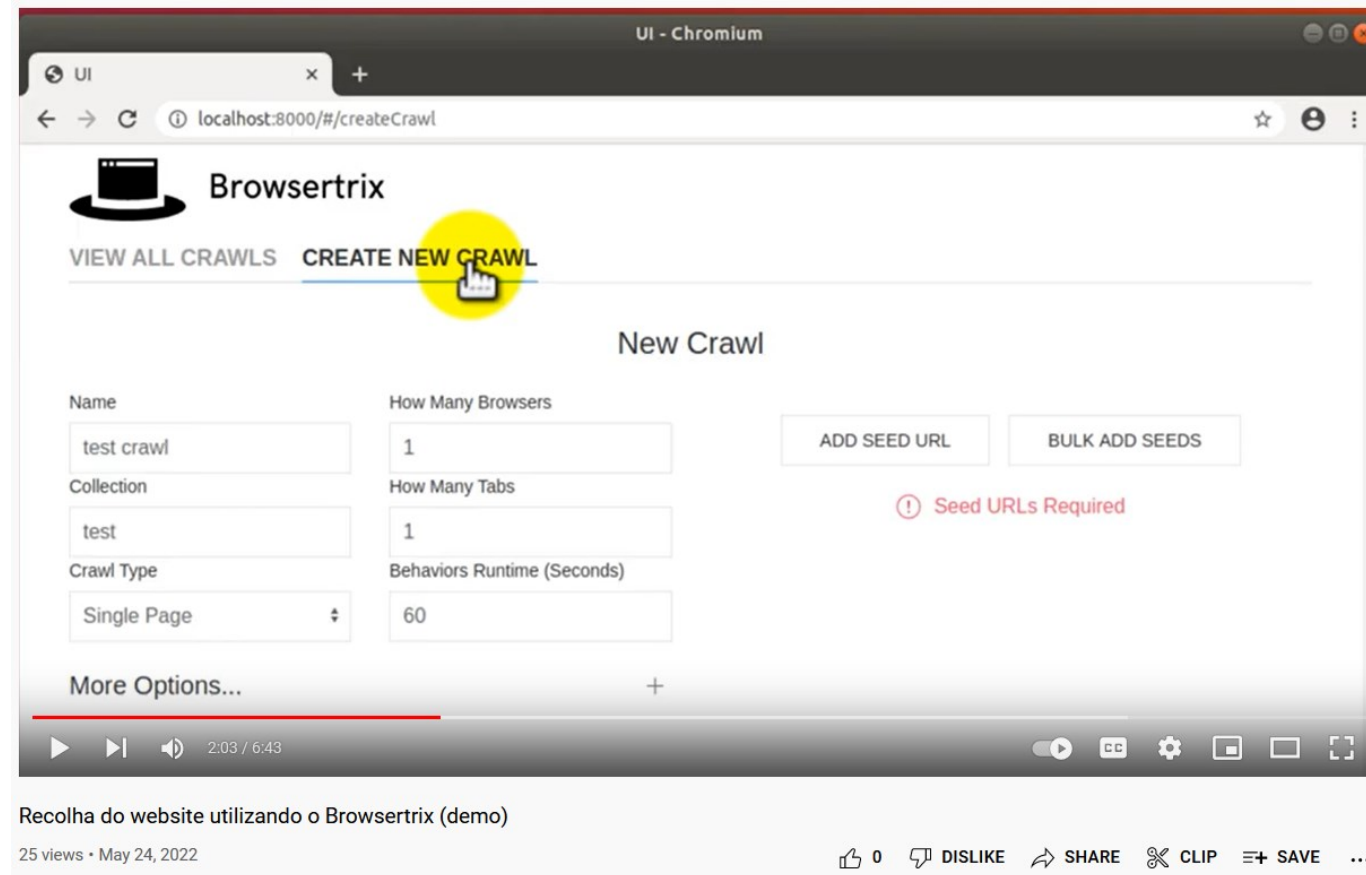
O Browsertrix foi instalado num simples computador portátil para treino e aprendizagem.

O site da Câmara Municipal de Lisboa foi o exemplo utilizado na gravação:

Guia para não especialistas: <https://tinyurl.com/instalar-browsertrix>

Presentemente, há uma nova versão para Desktop e um serviço experimental em Cloud:

Documentação oficial: <https://github.com/webrecorder/browsertrix>



Gravação do site da Câmara Municipal de Lisboa com o Browsertrix: (demo):
www.youtube.com/watch?v=7JNoMo302NA

Conclusão

- Grave conteúdos em formato normalizado
- Reutilize e analise em contexto local
- Contribua (juntamente com os arquivos da Web) para a preservação da memória histórica

Obrigado.

contacto@arquivo.pt

arquivo.pt/inscrever

