

Preservação de websites: faça-você-mesmo!

Dia Mundial da Preservação Digital - 7 de novembro de 2019

Ricardo Basílio - Curador Digital do Arquivo.pt



A photograph of a beach at low tide. In the foreground, the word "WELCOME" is written in large, capital letters in the sand. The letters are slightly shadowed, giving them a three-dimensional appearance. Behind the text, the ocean waves are breaking onto the shore, creating a white, foamy surf. The water is a deep blue, and the sky is a pale, hazy blue. The overall scene is peaceful and inviting.

WELCOME

Escrever na areia

Escrever um livro

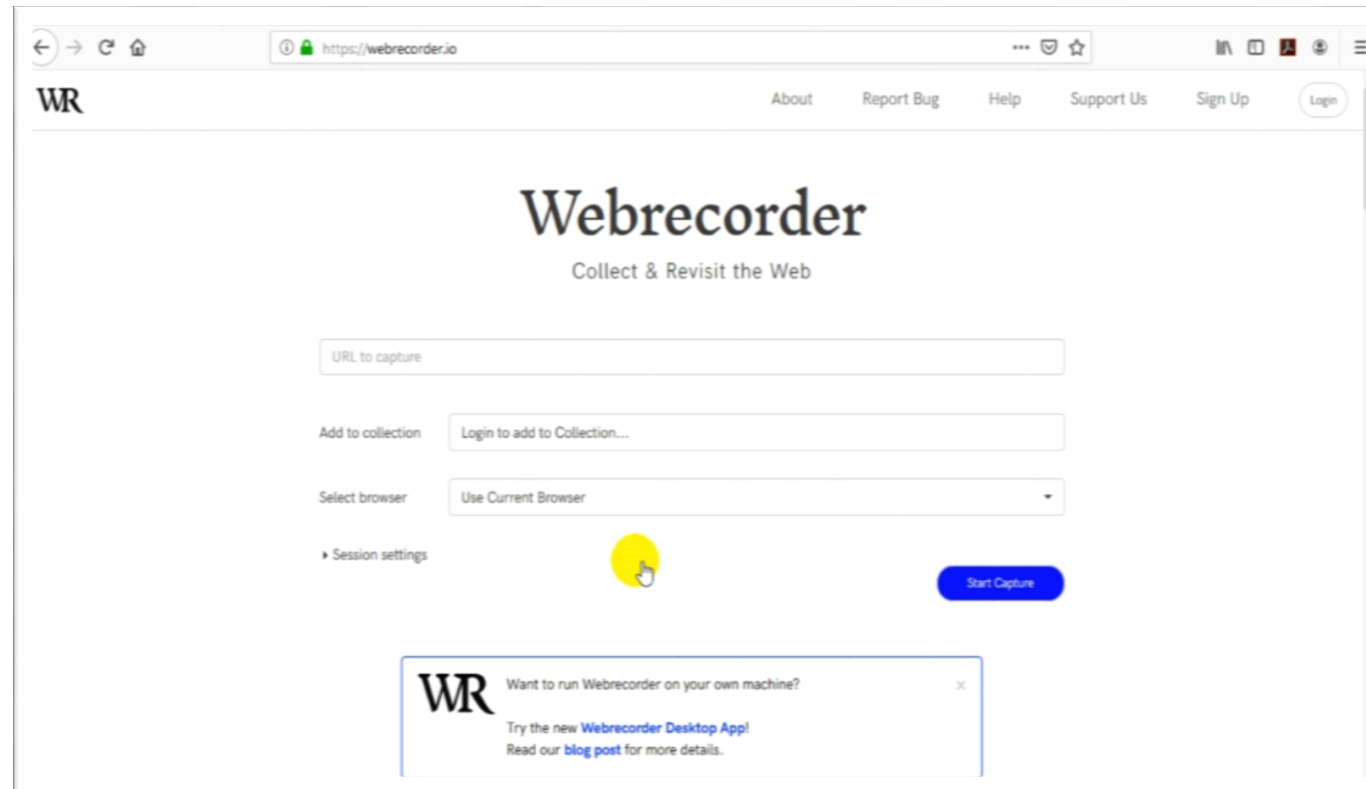
Publicar na Web

Procure aproximar a publicação de conteúdos na Web da experiência de escrever um livro: escrever/produzir para durar muitos anos, escrever/produzir cuidadosamente, etc.

Objetivo - desafio

- Gravar, armazenar e reproduzir uma página Web no formato WARC no próprio computador

Objetivo – desafio: exemplo



Neste breve exemplo, utilizamos o Webrecorder.io na versão online, pronto a utilizar. [\(Ver no Youtube\)](#)

I- Como preservar

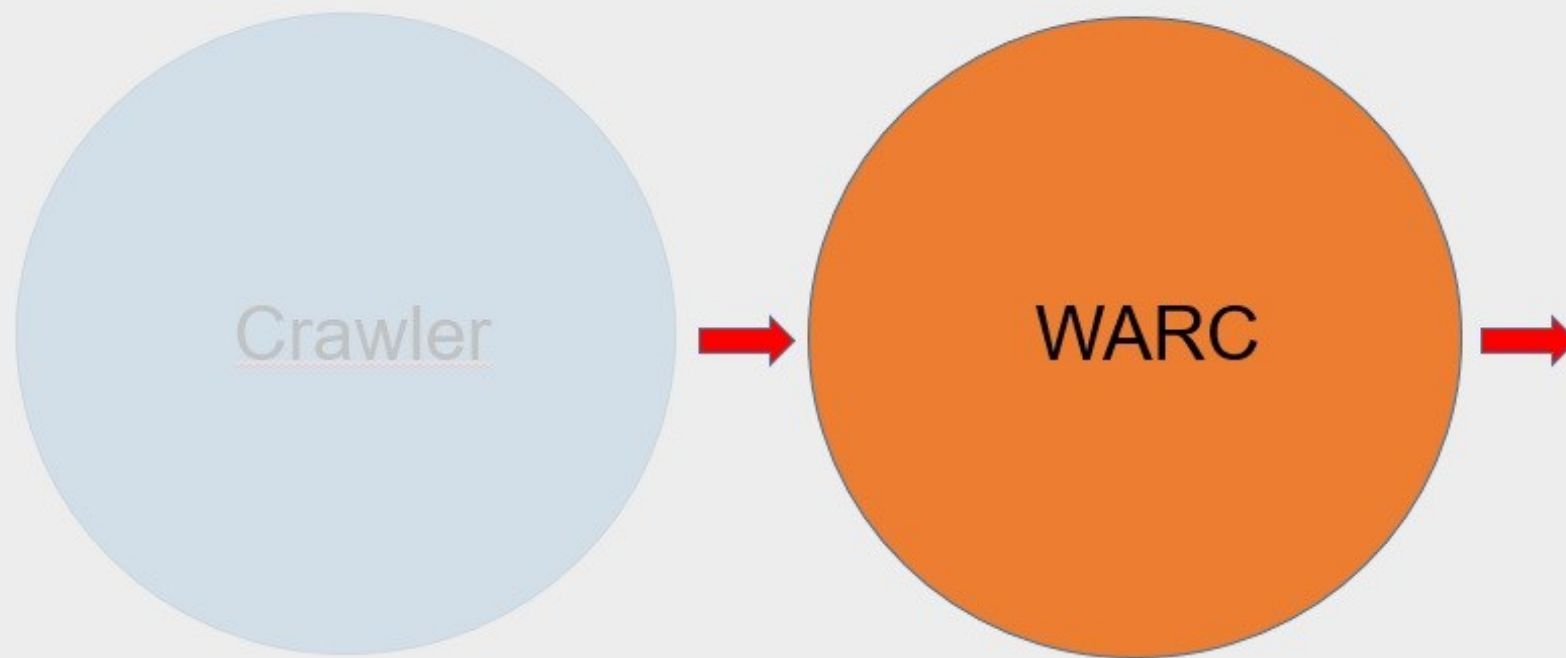
Três fases do processo de preservação



Três fases do processo de preservação



Três fases do processo de preservação



ISO 28500:2017

O formato WARC é compatível com os arquivos da Web. Se tiver páginas gravadas neste formato e as enviar os ficheiros WARC para o Arquivo.pt, por exemplo, estes podem ser integrados nas coleções e passam a ser reproduzidas pelo Arquivo.pt e apresentados nos resultados das pesquisas.

Se armazenar os WARC na sua instituição pode utilizá-los em qualquer altura, localmente e offline, ou usá-los em projetos futuros. É património da instituição. Uma vez produzidos pode descrevê-los, inseri-los num repositório, etc.

Três fases do processo de preservação

O Webrecorder reproduz as páginas de pois de as ter gravado. Reproduz ficheiros WARC externos mediante “importação” desses ficheiros



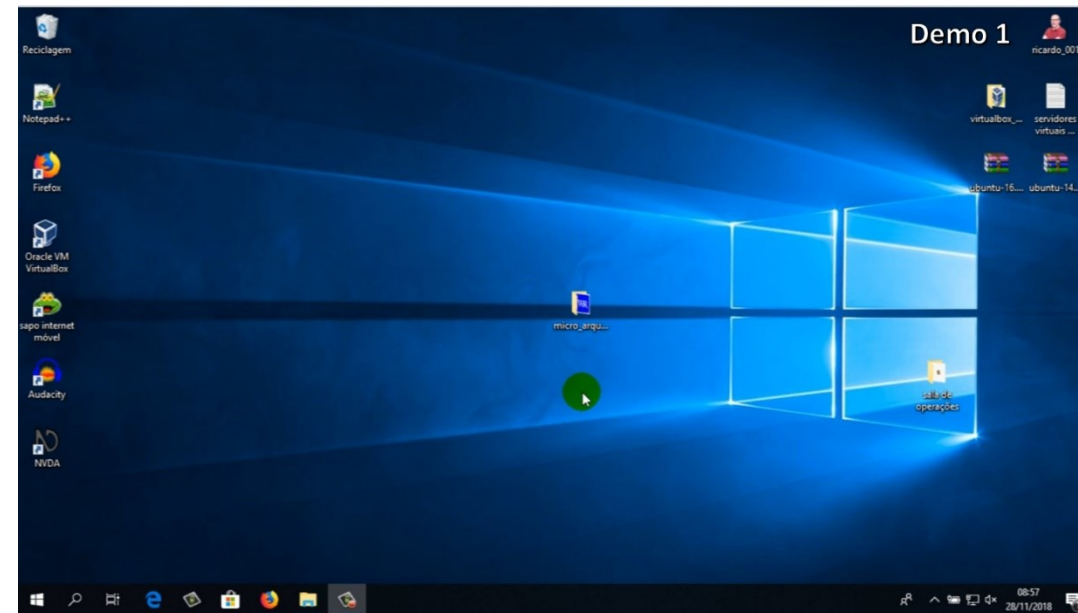
Como preservar

O Webrecorder é uma aplicação que permite gravar páginas Web com alta qualidade e descarregar o conteúdo no formato normalizado dos arquivos da Web, um ficheiro WARC.

É de utilização gratuita para pequenos projetos.

Para utilizar só precisa de criar uma conta e começar a gravar e a descarregar os ficheiros que produz.

Se preferir, pode acumular ficheiros até 5GB e descarrega-los mais tarde.



Recolher com [Webrecorder.io](https://www.webrecorder.io)
Exemplo, [NOVA-FCSH](https://www.nova-fcsh.org)

Como preservar

As páginas sociais também se podem gravar no formato WARC, que é um formato aberto.

O Facebook, por ex. permite descarregar toda a sua página.

Porém, do ponto de vista da preservação, é uma boa prática gravar em formatos diferentes e, se possível, em formato aberto.

Se está a utilizar o Webrecorder na versão online, é aconselhável gravar a página sem fazer login.

Se quer gravar a página com login, convém fazê-lo com a versão Desktop do Webrecorder para ter a segurança que todos os dados ficam na sua máquina.

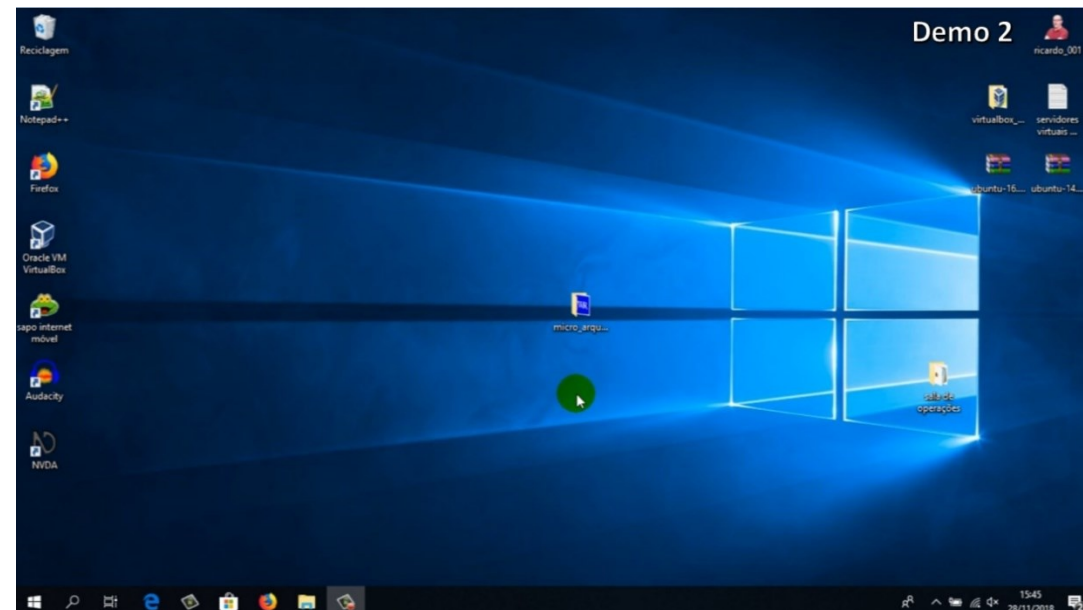
O Webrecorder tem funcionalidades (behaviors) que fazem automaticamente o que faria um humano: clicar, fazer scrool down, pôr vídeos a correr, etc.

Para começar, experimente e carregue no “piloto automático”.

Boa sorte.

Sobre a questão da preservação do Twitter e social media, leia:

<https://tinyurl.com/preservacaotwitter>



[Recolher Rede Social: demo \(ver no Youtube\)](#)

Gravando com o Webrecorder.io

1. Criar conta / login
2. Criar *collection*
3. Criar *session 1, session 2, session 3, etc.*
4. Gravar
5. Descarregar o ficheiro WARC
6. Armazenar em pasta local
7. Reproduzir onde e quando quiser

Se instalar o Webrecorder no seu computador, os ficheiros são gravados automaticamente numa pasta chamada >> Webrecorder-data. Pode encontra-los nos seus >> documentos.

Nesse caso dispensa-se o ponto 5.

Como preservar

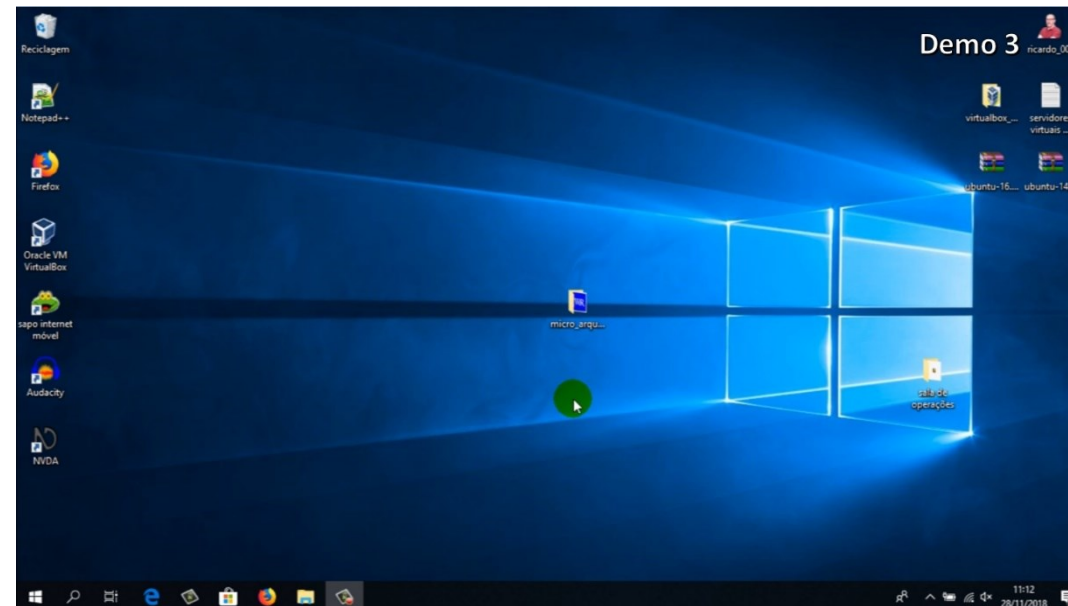
O armazenamento é a parte central do processo de preservação. Conserva-se os conteúdos preservados em contentores, no formato WARC.

Cada vez é mais barato armazenar. O que custa é o processamento. Portanto, não espere para gravar e armazenar. Depois logo vê como reproduzir, mostrar, processar.

Neste caso, pretende-se reforçar o carácter local e autónomo do trabalho do curador.

Descarregue para uma pasta local os ficheiros que grava. Armazene-os e crie o seu micro-arquivo de páginas Web.

Se instalar o Webrecorder no seu computador os ficheiros são gravados automaticamente numa pasta chamada >> Webrecorder-data. Pode encontra-los nos seus >> documentos.



Armazenar ficheiro WARC extraído do
[Webrecorder.io](https://www.webrecorder.io)

Gravando com o Webrecorder.io

Comparação com outros sistemas de recolha

Software de recolha	Aconselhável para	Requisitos / IT
Webrecorder	Recolhas de alta qualidade em algumas páginas escolhidas, “manualmente”	Pronto a utilizar
Browsertrix	Recolha de alta qualidade de um site inteiro, ou vários, automaticamente	Exige apoio técnico
Brozzler	Recolha de alta qualidade de um site inteiro, ou vários, automaticamente	Exige apoio técnico
Heritrix	Recolhas milhares de sites, automaticamente	Exige apoio técnico especializado

Tarefas de um curador digital de websites

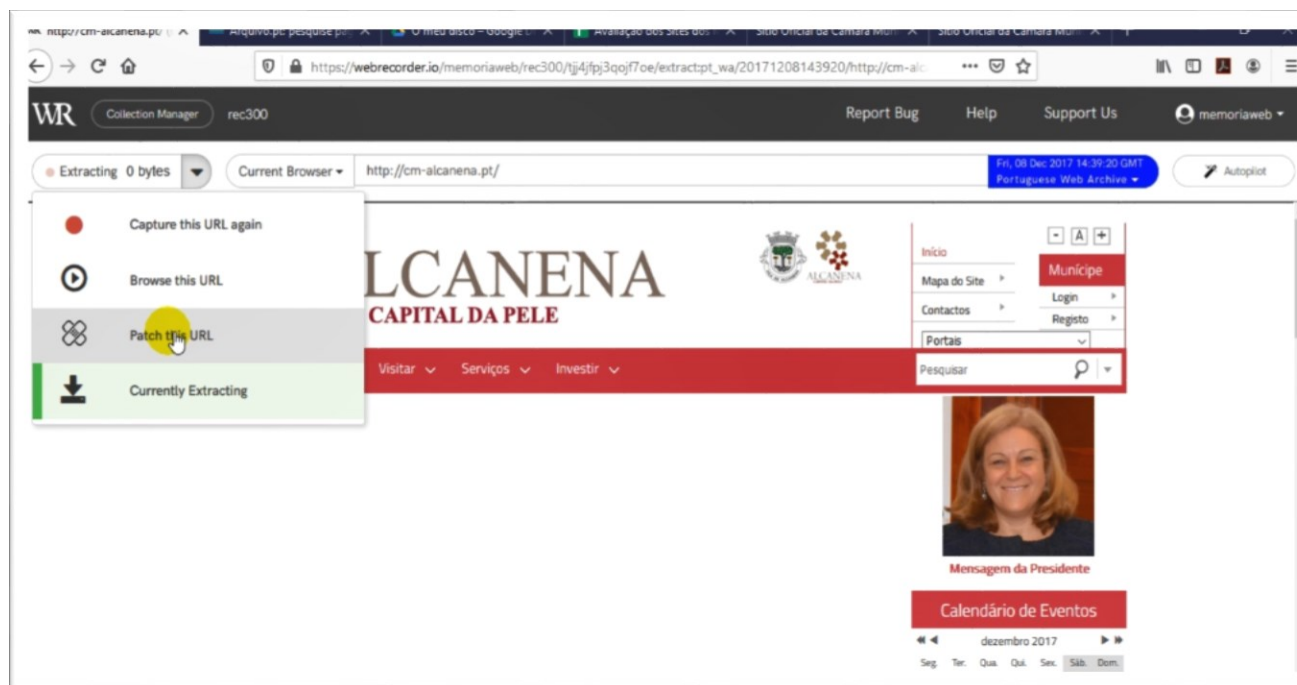
1. Fazer lista de sites
2. Olhar para eles na perspetiva da preservação
3. Garantir que são recolhidos pelo Arquivo.pt
4. Melhorar a preservação
5. Divulgar as páginas preservadas

Um curador digital de sítios Web não tem de ser um informático. As tarefas consistem em verificação humana, manual, qualitativa.

1 hora/mês, numa instituição de pequena e média dimensão, é suficiente para melhorar significativamente o estado de preservação dos sítios Web e canais institucionais. Siga um fluxo de trabalho:

[Fluxo de Preservação de sítios Web institucionais](https://bit.ly/2ounZhh). Acessível em: <https://bit.ly/2ounZhh>

Gravando com o Webrecorder - patching



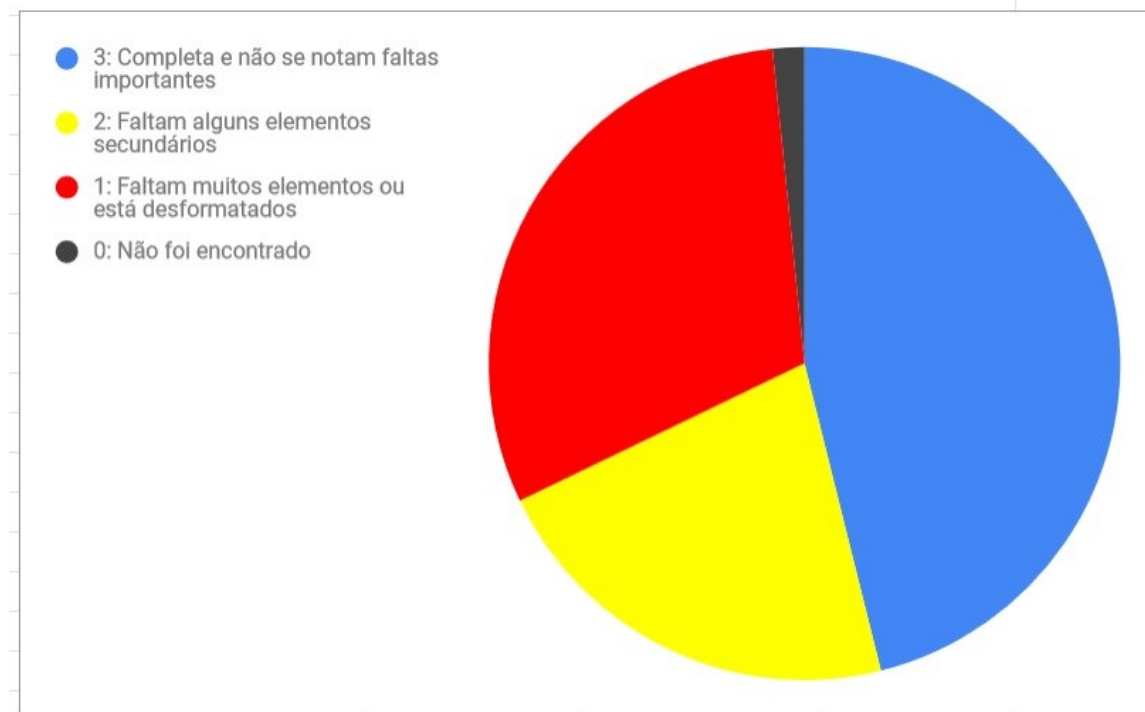
[Demo: remendar páginas passadas ou *patching* \(no Youtube\)](#)

Ex., página principal do Município de Alcanena de dezembro de 2017

Gravando com o Webrecorder

Aplicável a conjuntos de websites para melhorar qualidade

Gráfico 1: Qualidade dos sites dos municípios no Arquivo.pt em Dezembro de 2017



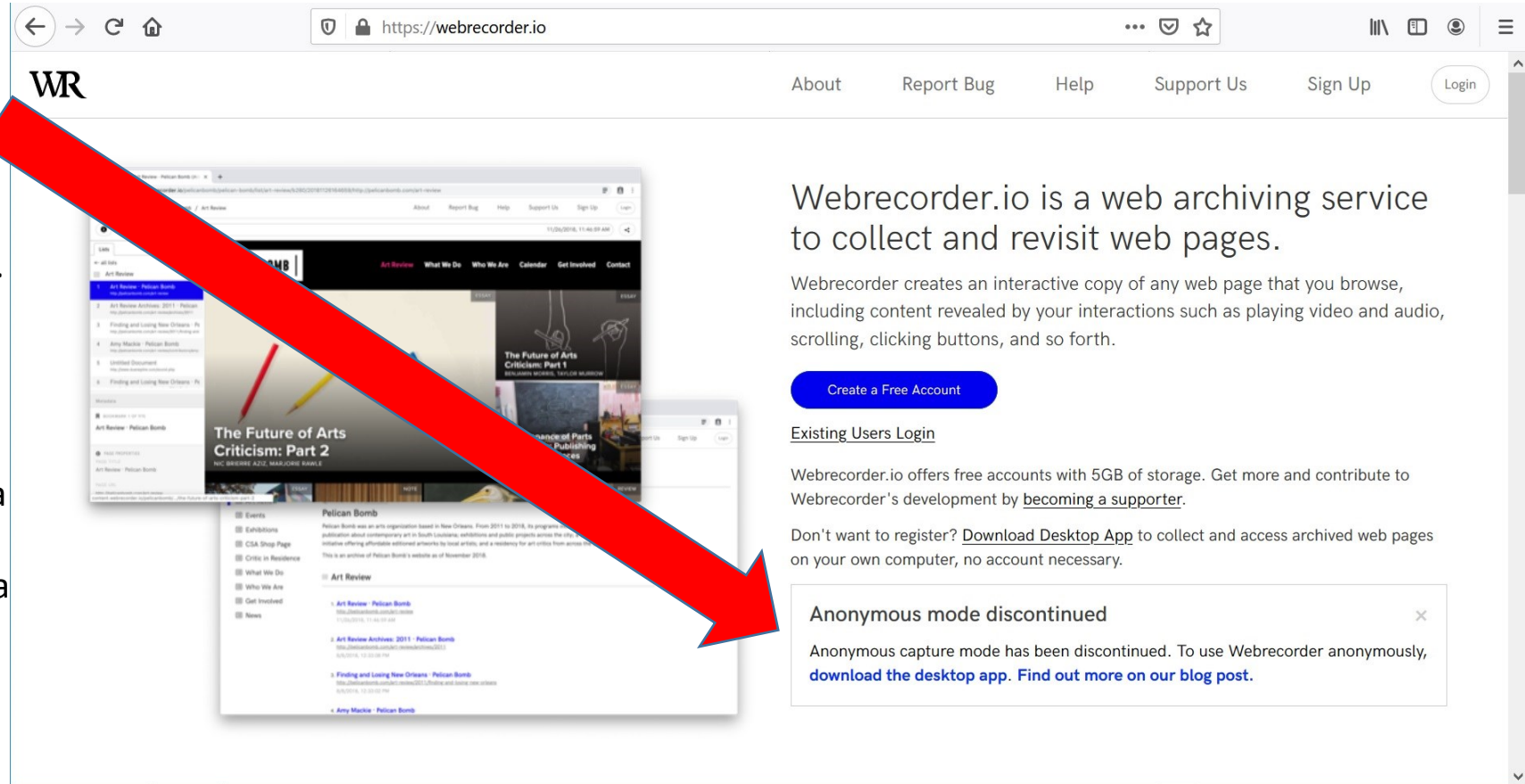
Análise da qualidade dos sítios Web dos municípios: Acessível em: <https://bit.ly/2MCkHjV>
Estudo de caso: <https://www.bad.pt/publicacoes/index.php/arquivosmunicipais/article/view/1978>

Gravando com o Webrecorder localmente

Webrecorder Desktop app

Instale a app no seu computador. Os WARC's ficam na sua máquina. Vão para a pasta uma pasta local (documents/Webrecorder-data). Funciona como na versão online.

A versão Desktop é adequada para gravar páginas em que faz login (ex. Facebook, Twitter, páginas da Intranet, etc.) – os WARC's resultantes destas gravações em ambiente interno devem ficar para uso interno.



Como preservar

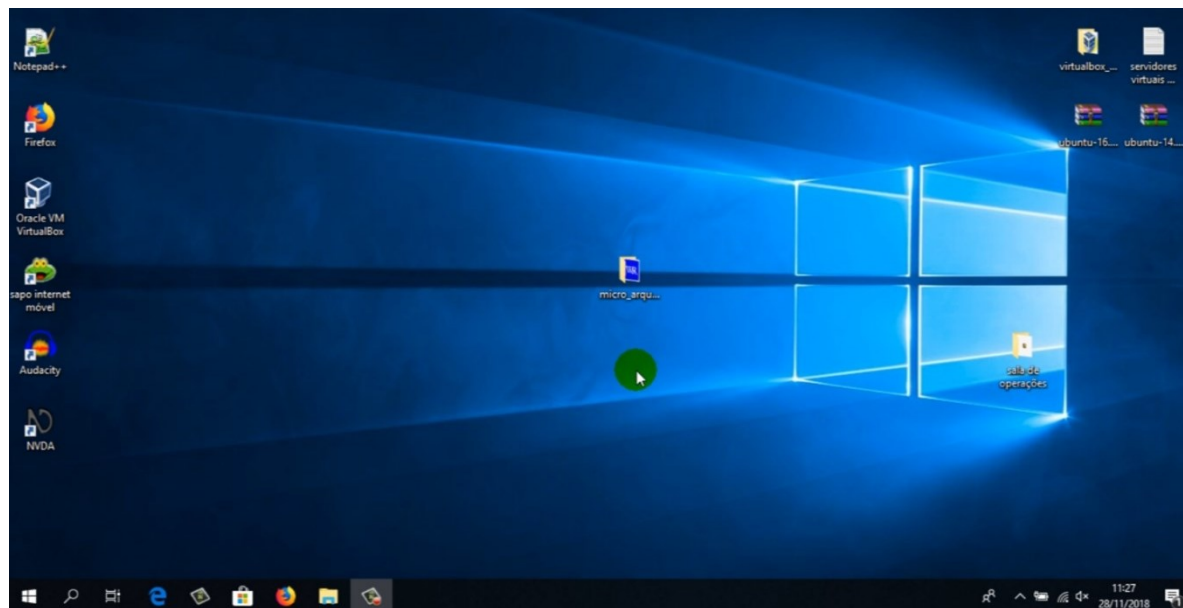
Veja um WARC por dentro:

Neste vídeo que pode [ver no Youtube](#) mostra-se que tipo de informação contém um ficheiro WARC. Para isso, usamos um editor de texto simples – Notepad, bloco de notas.

O ficheiro WARC é um container semelhante a um ZIP. Por isso precisa de o “desembrulhar”, descompactar, com costuma fazer com um ZIP ou um RAR.

Depois de extrair o ficheiro, pode >> abrir como >> ficheiro de texto com Bloco de Notas ou Notepad++

Repare como ficou registada toda a transação entre o servidor e o cliente, o seu browser. Esses dados reforçam a integridade dos conteúdos preservados.



Ficheiro WARC - ISO 28500:2017

Para primeira abordagem, veja especificações do formato numa versão *draft* (2016) disponível em : <http://bibnum.bnf.fr/warc/>

II - Razões para preservar Sítios Web

Sítios Web contêm memórias únicas

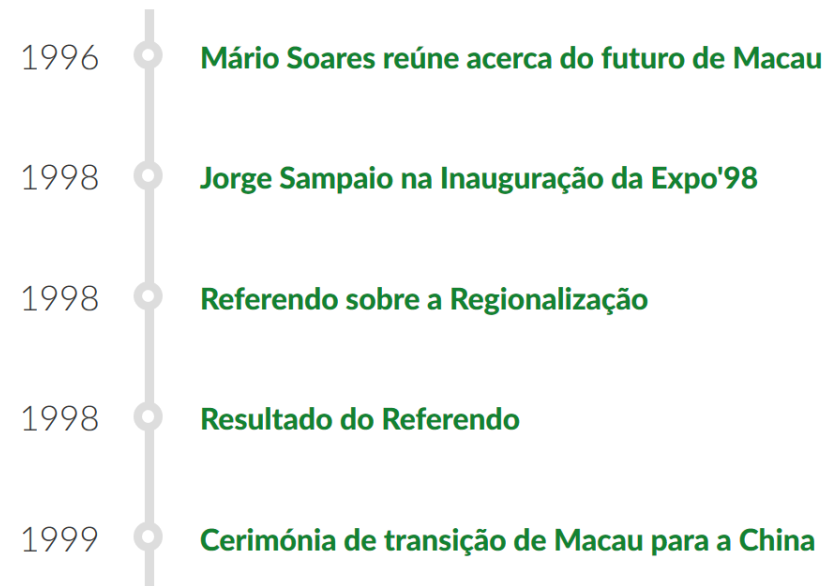


[Site da Presidência da República, 2011](#)

Páginas do passado valorizam o presente

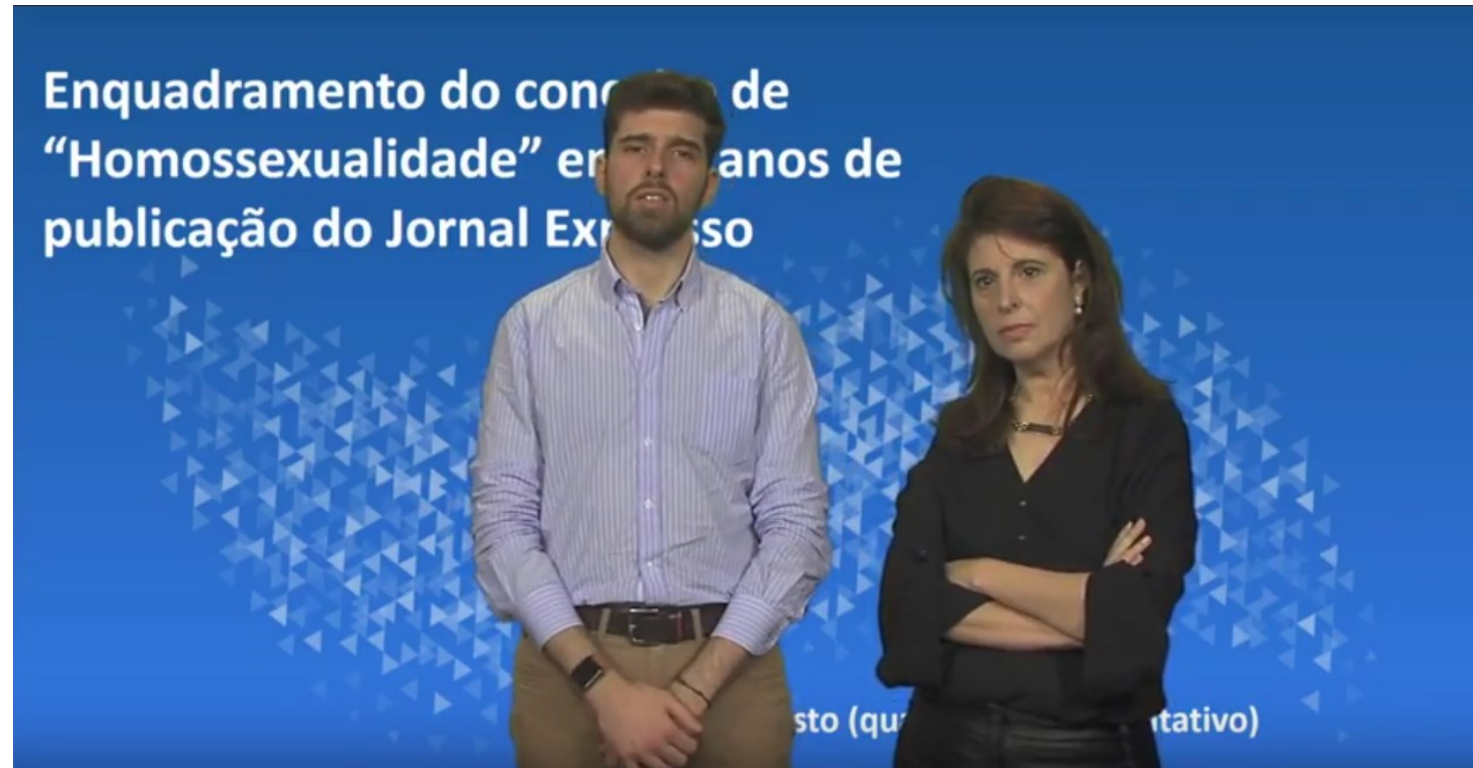
Viagem no tempo com os **Presidentes da República**

Memória web dos mandatos presidenciais usando o [Arquivo.pt](http://arquivo.pt).



Colaboração com o Museu da Presidência da República: arquivo.pt/presidentes

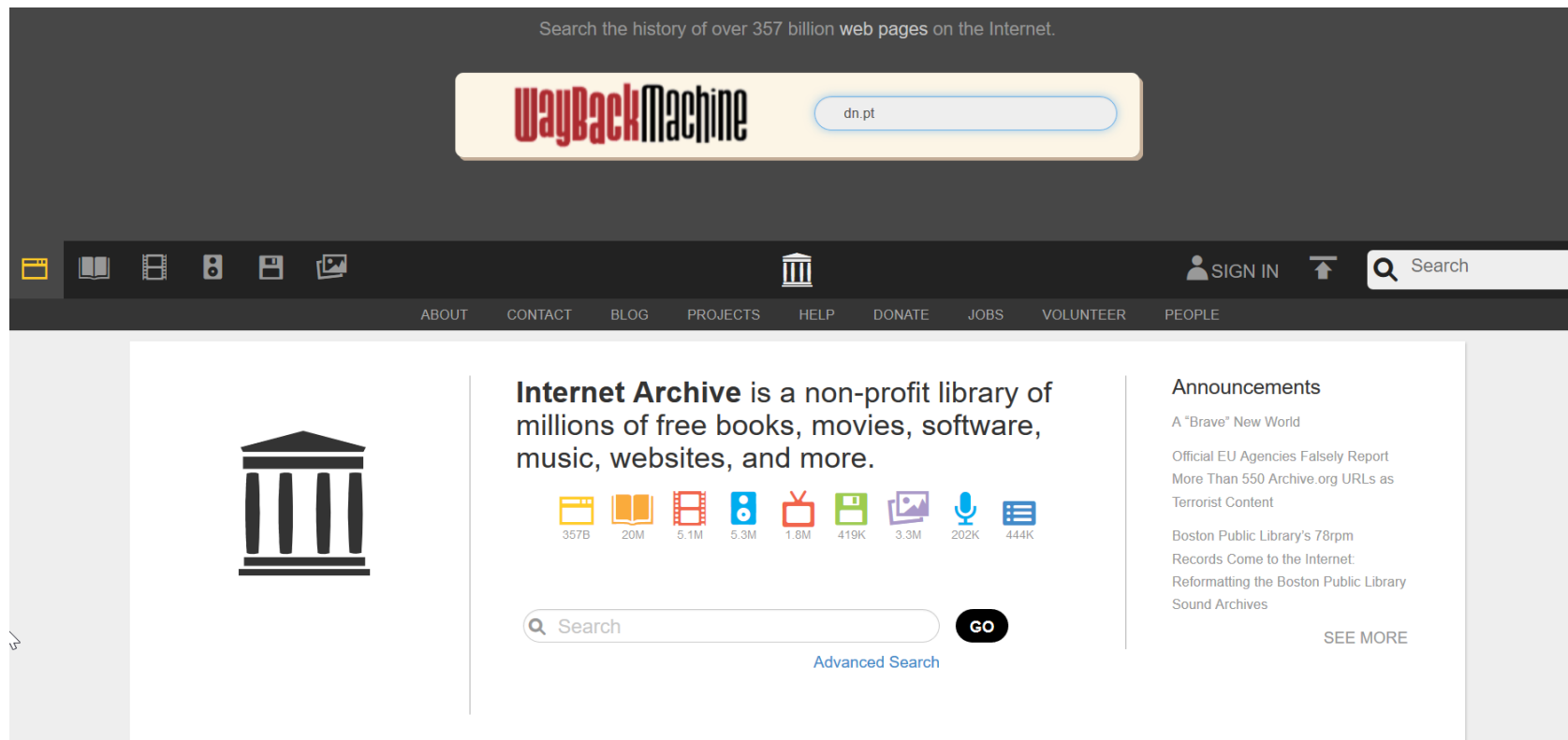
Páginas do passado são material para o estudo e a investigação



[Prémio Arquivo.pt 2018 – 2º classificado](#)

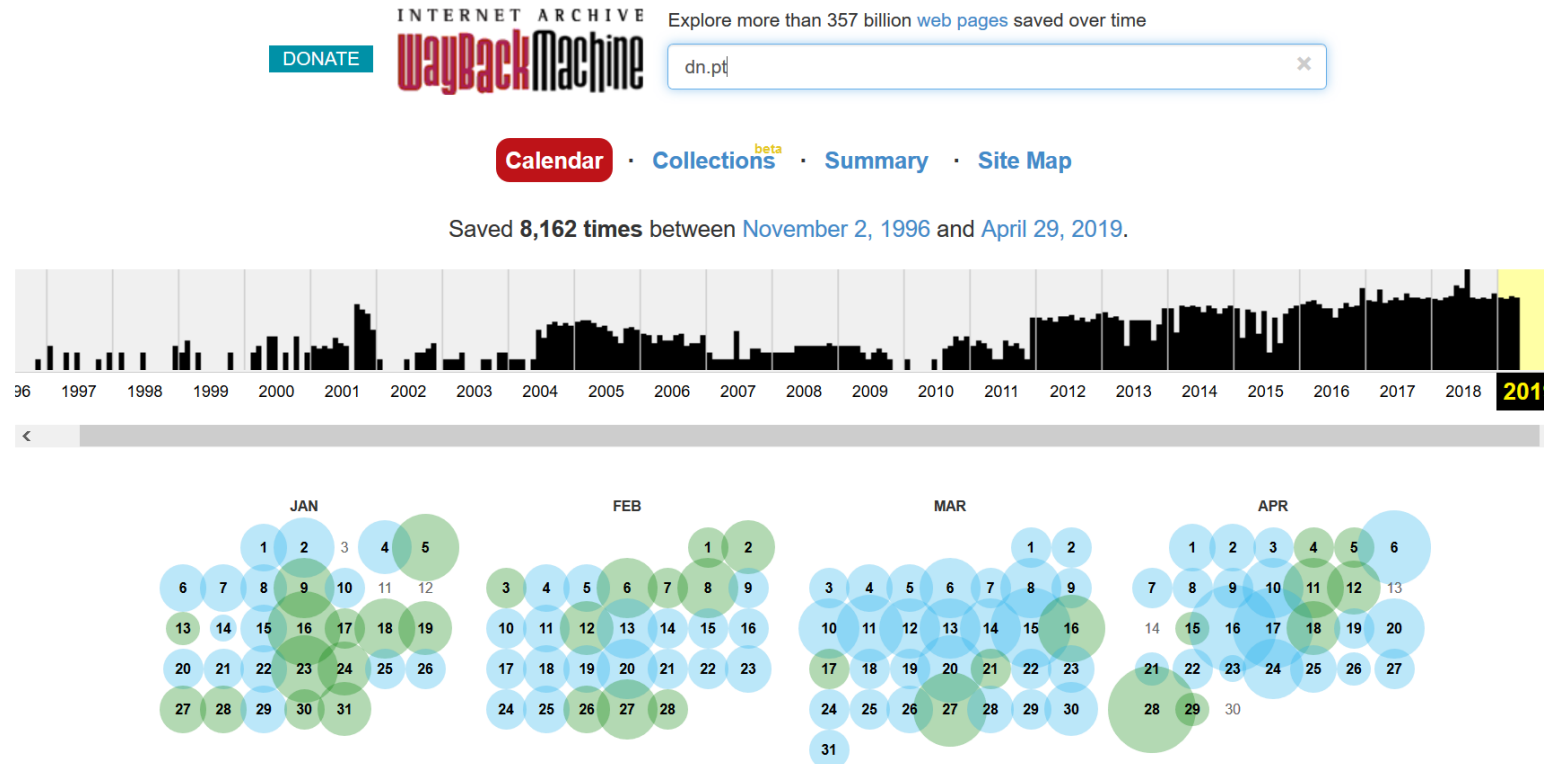
III - Referências e boas práticas

O maior arquivo da Web (Califórnia - EUA)



Internet Archive: <https://archive.org/>

O maior arquivo da Web (Califórnia – EUA)



Internet Archive: <https://archive.org/>

Um serviço para recolhas institucionais (EUA)



Stanford University, Social Sciences Resource Group

Archive-It Partner Since: Jun, 2007
Organization Type: Colleges & Universities
Organization URL: <http://ssrc.stanford.edu>

Description: Stanford Library's Social Sciences Resource Group supports research, instruction and collection development within the social sciences, including subject areas like political science, economics, sociology, anthropology, psychology, communication, environmental and public policy. The SSRC also includes CA state, local, US Federal and International government information, data and statistics. SSRC's Archive-it collections will focus on San Francisco bay area governments, IGO/NGO working papers, and other areas of interest to Stanford social scientists.

Narrow Your Results

Subject Sort By: Count | (A-Z)

- Society & Culture (8)
- Government - US Federal (4)
- Government (3)
- Government - National (2)
- Government - US States (2)

[More ▼](#)

Creator Sort By: Count | (A-Z)

- James Jacobs (8)
- James Jabobs (1)
- Stanford University Libraries. Government Information and Social Sciences Resource

Sites and collections from this organization are listed below. Narrow your results at left, or enter a search query below to find a collection, site, specific URL or to search the text of archived webpages.

Collections Sites Search Page Text

Page 1 of 1 (13 Total Results)

Sort By: Collection Name (A-Z) | Collection Name (Z-A)

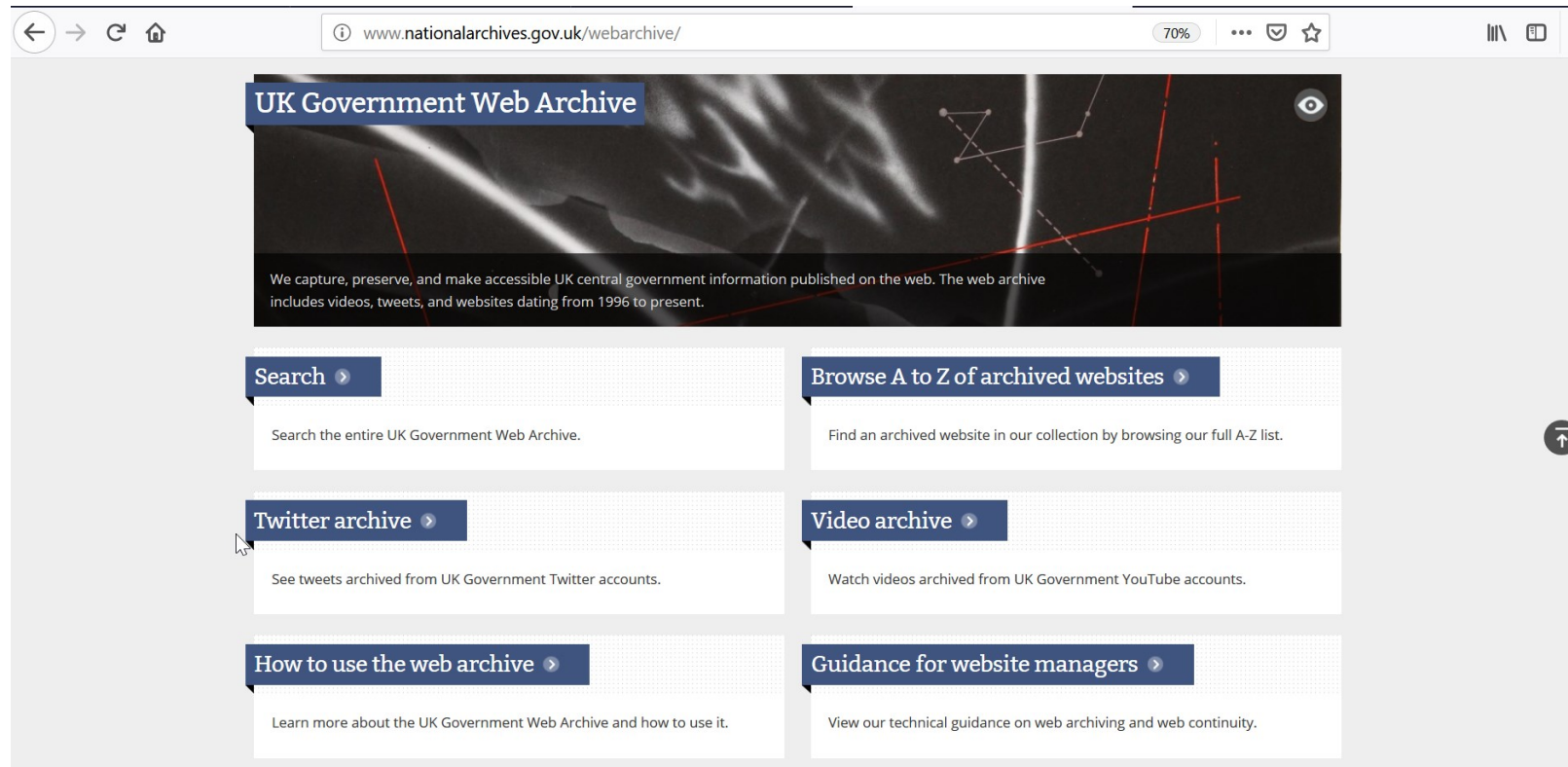
Bay Area Governments

Archived since: Nov, 2007

Description: This collection provides access to the archived Web sites and publications of the special districts, local and regional governments geographically centered around the San Francisco bay area.

Archive-it: <https://archive-it.org>

Um arquivo público (Reino Unido)



UK Government Web Archive: [nationalarchives.gov.uk/webarchive](https://www.nationalarchives.gov.uk/webarchive)

Coleções de sítios Web em biblioteca (EUA)



Everything

Search Loc.gov



Subscribe Share/Save

ARCHIVED WEB SITE

JL8



Library of Congress [Web Archiving Collections](#)

Citar sítios Web na investigação (Library of Congress - EUA)

Cite This Item

! Citations are generated automatically from bibliographic data as a convenience, and may not be complete or accurate.

Chicago citation style:

Ridge, Thomas J, and United States Dept. Of Homeland Security. *DHS | Department of Homeland Security | DHS Home Page*. 2003. Archived Web Site. <https://www.loc.gov/item/lcwa00010111/>.

APA citation style:

Ridge, T. J. & United States Dept. Of Homeland Security. (2003) *DHS | Department of Homeland Security | DHS Home Page*. [Archived Web Site] Retrieved from the Library of Congress, <https://www.loc.gov/item/lcwa00010111/>.

MLA citation style:

Ridge, Thomas J, and United States Dept. Of Homeland Security. *DHS | Department of Homeland Security | DHS Home Page*. 2003. Archived Web Site. Retrieved from the Library of Congress, <www.loc.gov/item/lcwa00010111/>.

Citar um Sítio Web Preservado: [exemplo da Library of Congress](#)

IV- Ideias para a valorização da memória institucional

URL de uma versão preservada

<https://arquivo.pt/wayback/20010208131804/http://uac.pt/>

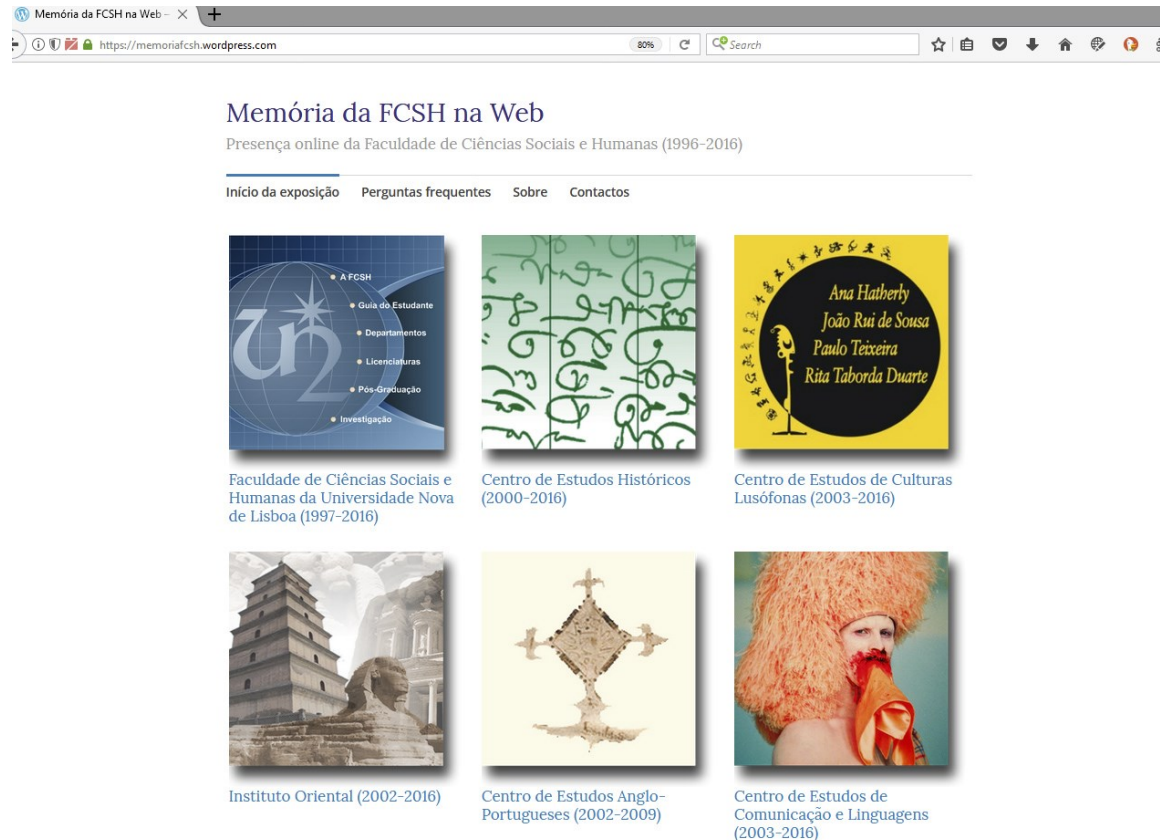
URL de uma versão preservada

Timestamp

2001-02-08.13:18'04''

Ano Mês Dia Hora Min. Seg.

Exposição online: memoriaFCSH



[memoriaFCSH.wordpress.com](https://memoriafcsch.wordpress.com)

Preservação de websites: faça-você-mesmo!

ricardobasilio@fcsch.unl.pt

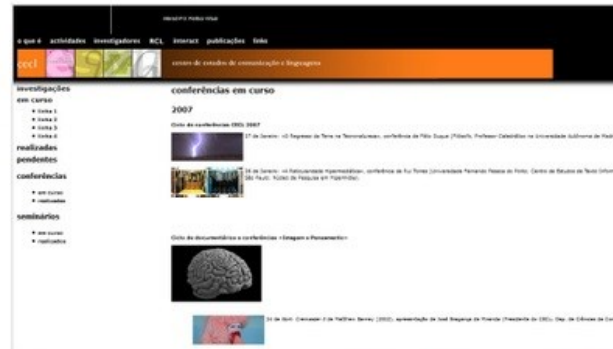
Exposição online: memoriaFCSH

Memória da FCSH na Web

Presença online da Faculdade de Ciências Sociais e Humanas (1996-2016)

[Início da exposição](#) [Perguntas frequentes](#) [Sobre](#) [Contactos](#)

Centro de Estudos de Comunicação e Linguagens (2003-2016)



Pormenor da página "Actividades" (cecl.unl.com.pt, 2003)

Exposição online: memoriaFCSH

Memória do sítio Web

A [primeira versão preservada é de 2003](#). Alguns elementos gráficos estão em Flash, o que limita a sua reprodução em vários dispositivos actuais. Em [2008](#), surge com um *design* ligeiramente diferente.

Em 2014, sítio Web, já com um *design* diferente, o sítio Web informa que o CECL [passou a integrar o centro inter-universitário CIC.DIGITAL](#), tornando-se seu pólo na FCSH.

O sítio Web do CECL foi um ponto de acesso para a Revista de Comunicação e Linguagens, a revista online Interact e o sítio Working papers, recursos que tiveram destaque na sua página de entrada. Foi também um suporte para registo de actividades correntes e passadas, projectos em curso ou finalizados.

Endereço

- cecl.com.pt, [2003-2016](#)

Momentos para recordar

- [Revista Comunicação e Linguagens, 2006](#)
- [Working papers, 2008](#)
- [CD-ROM Cultura Digital numa notícia do site, 2008](#)

Exposição online: memoriaFCSH

Memória da Imprensa Portuguesa

Exposição de páginas preservadas dos jornais portugueses

[Home](#) [Casos de uso](#) [Sobre](#)



Diário Digital (1997-2017)



Sábado (2004-2017)



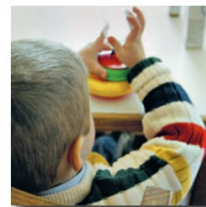
Visão (2001-2017)



Tribuna da Madeira (2004-2017)



Semanário (2003-2017)



Global Noticias (2004-2010)

Memória Web da Rádio em Portugal

Páginas Web antigas das rádios (1996-2018)

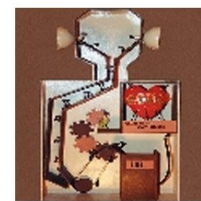
[Início da exposição](#) [Lista de rádios em Portugal](#) [Sobre a coleção](#) [Contacto](#)



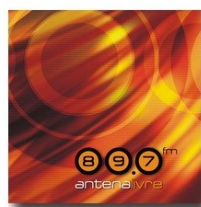
Antena 1 (1998-2018)



Antena 2 (1998-2018)



Antena 3 (1998-2018)



Antena Livre (2004-2018)



Antena Minho (2001-2018)

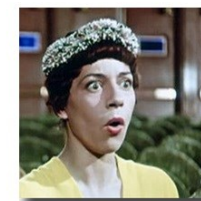


Antena Sul (2014-2018)

Memória da Televisão em Portugal

Páginas Web antigas das Televisões em Portugal (1996-2018)

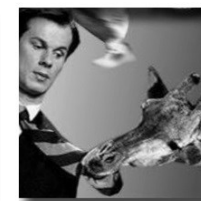
[Início da exposição](#) [Página antigas dos canais de TV](#) [Sobre a coleção](#)



RTP Memória (2010-2018)



MTV Portugal (2003-2018)



RTP (1998-2018)



Porto Canal (2013-2018)



Canal Parlamento (2003-2018)



SIC (2003-2018)

<https://sobre.arquivo.pt/colecoes>

Preservação de websites: faça-você-mesmo!

ricardobasilio@fcs.unl.pt

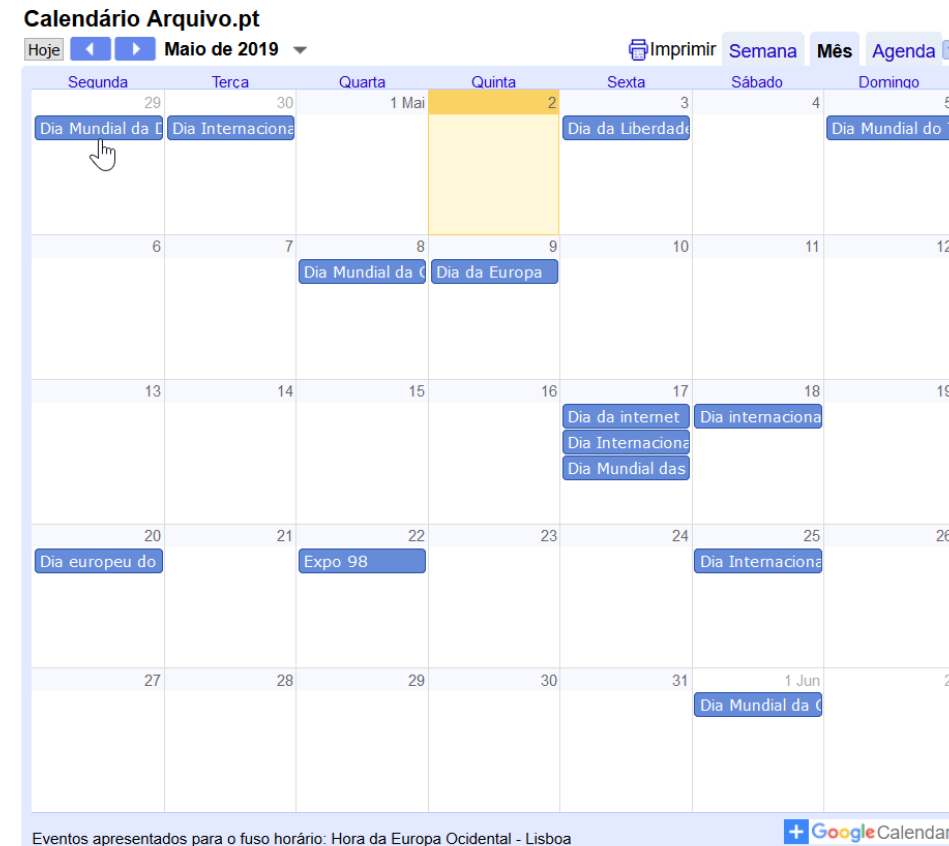
Exposição física de páginas preservadas



Exposição na NOVA-FCT (Monte da Caparica - Almada)

Calendário de efemérides

- Calendário Arquivo.pt
- Sugira uma efeméride
- Crie um calendário adaptado à sua instituição



Calendário de efemérides nas redes sociais

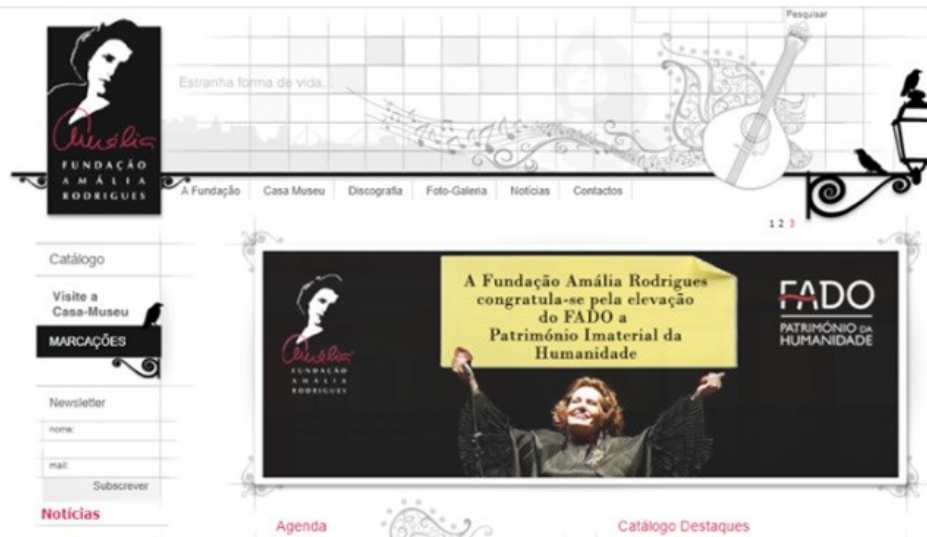
Publicações



Arquivo.pt

Publicado por Luiz Alberto Moura [?] · 22 de Novembro às 14:00 ·

No **#DiadoMusico**, visitamos o site da grande Amália Rodrigues em 2012 com o Arquivo.pt.
<https://bit.ly/2zlnR4>

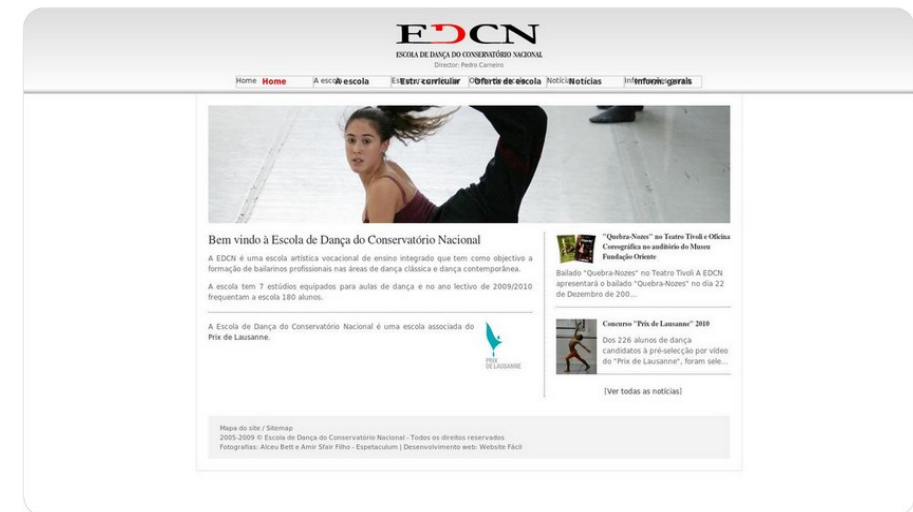


Arquivo.pt @ArquivoWeb_PT · 29 de abr

Hoje é o **#DiaMundialdaDança**. Celebre o trabalho da Escola de Dança do Conservatório Nacional com o [Arquivo.pt](https://arquivo.pt)

arquivo.pt/wayback/200912...

#Dança #arte



Preservação de websites: faça-você-mesmo!

ricardobasilio@fcsn.unl.pt

Agregações temáticas ou institucionais

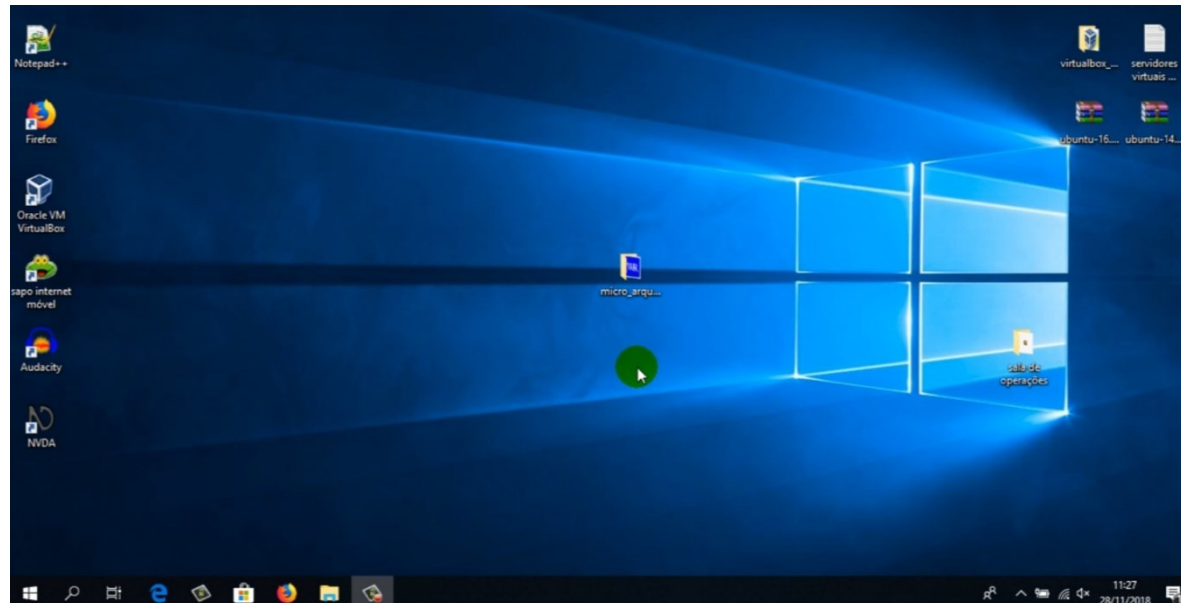
Memória Web

Para Todos

facebook.com/memoriawebparatodos



Tutorial

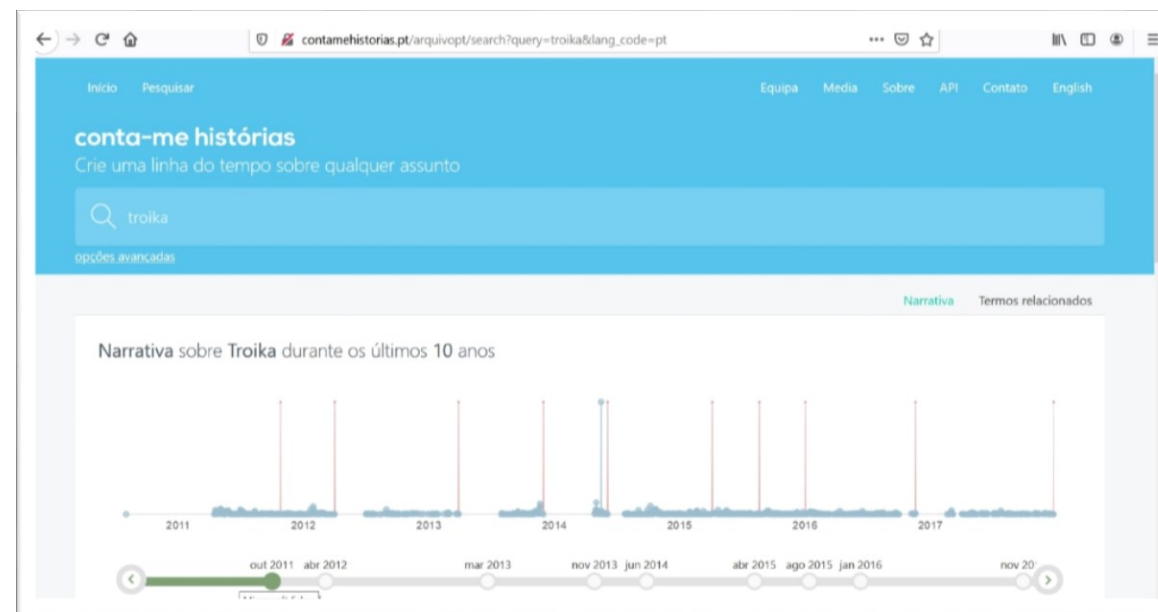


[Como publicar no Facebook conteúdos do Arquivo.pt \(ver no Youtube\)](#)

Casos de uso do Arquivo

Conta-me histórias

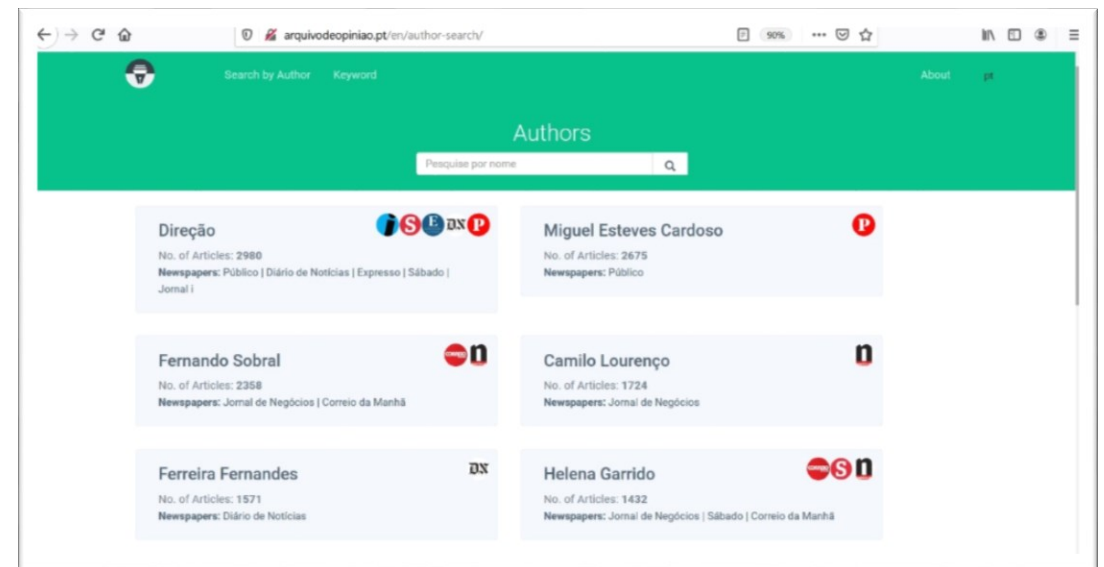
- Aplicação
- Notícias de 24 jornais
- URL: <http://contamehistorias.pt>



Casos de uso do Arquivo

Arquivo de Opinião

- Aplicação
- Colunas de opinião de alguns jornais
- URL: <http://arquivodeopiniao.pt>



Casos de uso do Arquivo

meuParlamento

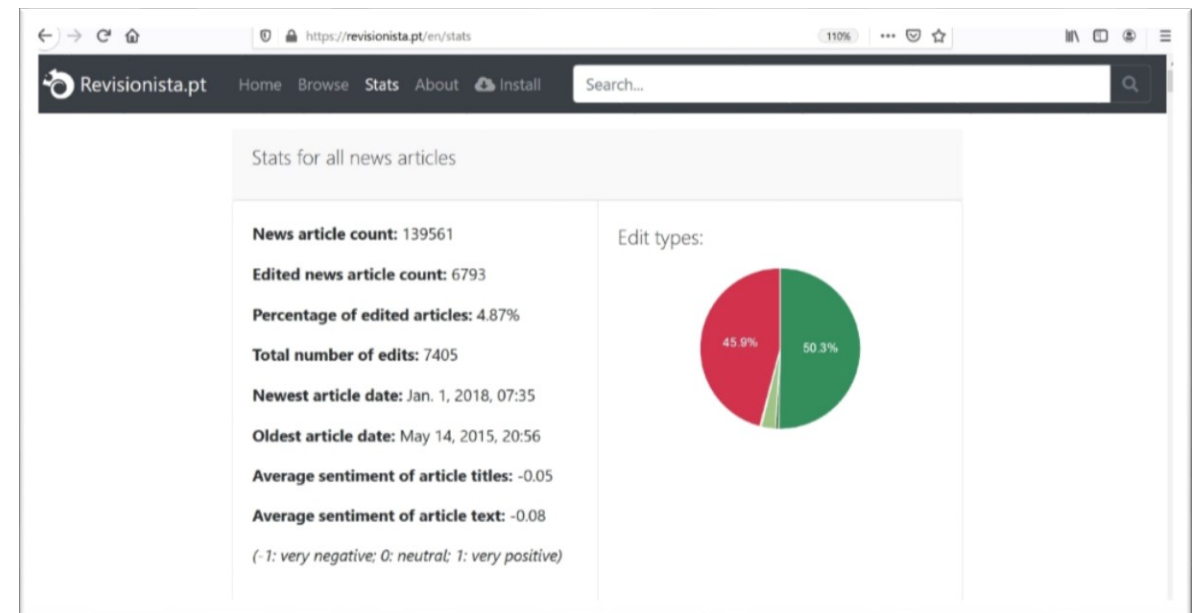
- Aplicação móvel
- Política, legislação, cidadania
- URL: <http://www.meuparlamento.pt>



Casos de uso do Arquivo

Revisionista.pt

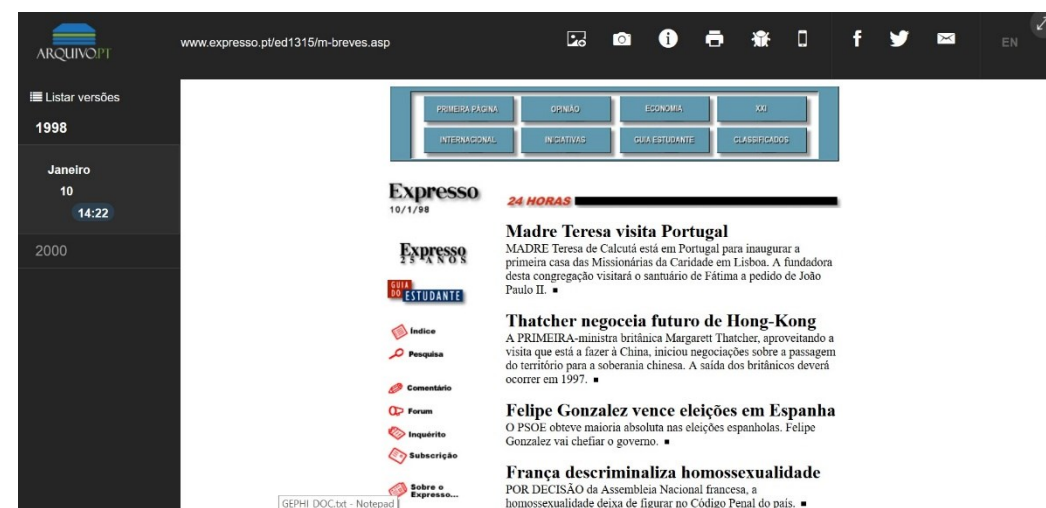
- Extensão de browser
- Imprensa, notícias online
- URL: <https://revisionista.pt>



Casos de uso do Arquivo

Enquadramento do conceito de “homossexualidade” em 20 anos de publicação do Jornal Expresso

- Estudo
- Género, semanário Expresso
- tinyurl.com/3premioarquivopt



Casos de uso do Arquivo

Discursos públicos sobre violência em privado

- Estudo
- Violência doméstica, notícias de jornais
- tinyurl.com/3premioarquivopt



Casos de uso do Arquivo

O Straight-Edge na área metropolitana de Lisboa

- Estudo
- Cultura popular, música
- <http://arquivo.pt/straightedge>



Conclusão - desafio

- Sabemos como gravar, armazenar e reproduzir uma página Web no formato standard WARC, localmente e offline

Desafio

- Consolide a sua aprendizagem repetindo o processo em páginas à escolha e envie os ficheiros WARC para o curador digital do Arquivo.pt.

Não perca a memória!

Sugira para o Arquivo.pt preservar: arquivo.pt/sugerir

Preserve você mesmo: webrecorder.io

ricardobasilio@fcsb.unl.pt

twitter: @ricardobasilio_

Ou

contacto@arquivo.pt

