CONGRESSO INTERNACIONAL EM TECNOLOGIA E ORGANIZAÇÃO DA INFORMAÇÃO

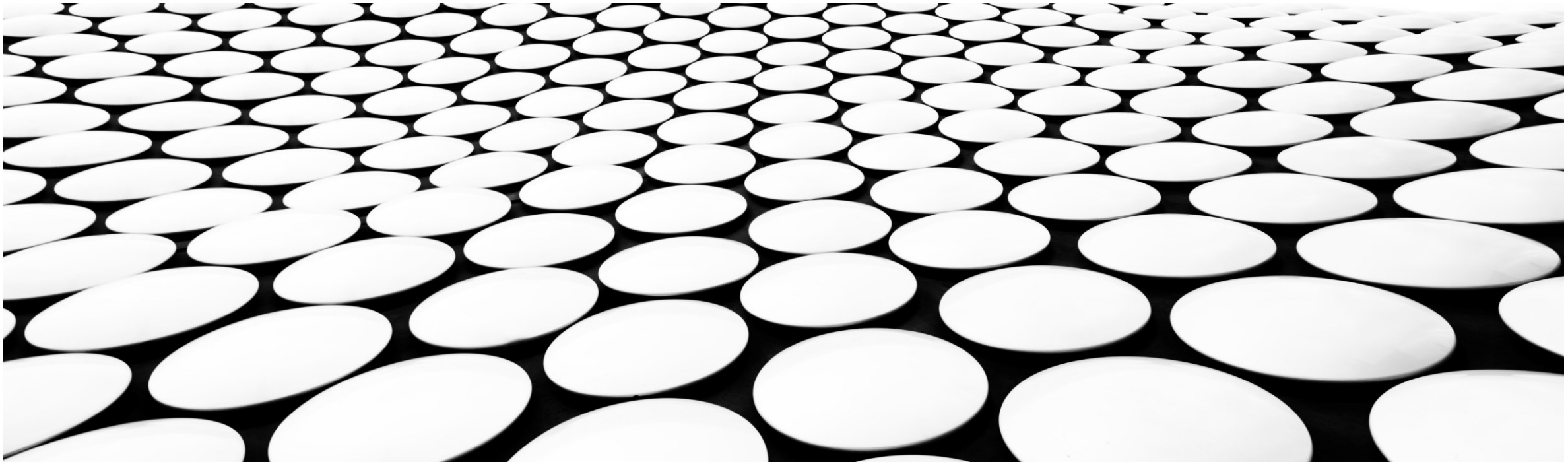
TOI – VI Congresso Internacional em Tecnologia e Organização da Informação

LIBRARY AND INFORMATION SCIENCE CONFERENCE

Experiência do Arquivo.pt: Sites Preserváveis e Ações de Preservação

Ricardo Basílio
contacto@arquivo.pt
Arquivo.pt

I Simpósio Políticas e Estratégias de Preservação de Conteúdo na Web



Cofinanciado por:

Lisb@20 20

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

FCT

QUIVO.PT

NOVA FCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

Agenda

I. O Arquivo.pt – notas históricas e utilização

II. Sites preserváveis - recomendações

III. Ações de preservação – Curadoria digital

O Arquivo.pt

Instalações do Arquivo.pt situadas na Avenida do Brasil, em Lisboa

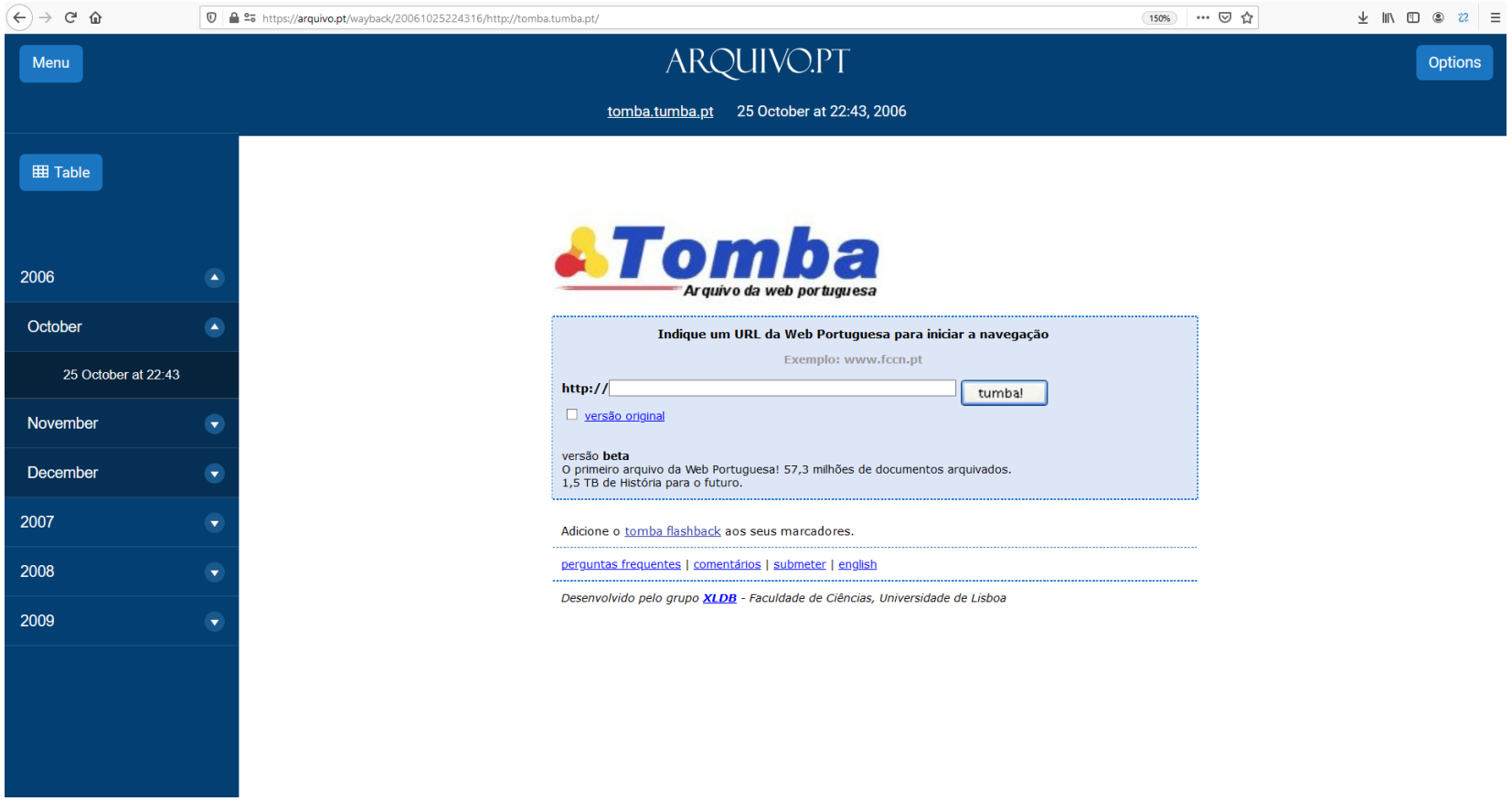
O Arquivo.pt

O Arquivo.pt – um pouco de história

- Serviço público de preservação da Web portuguesa gerido pela Fundação para a Ciência e a Tecnologia
- Iniciou a recolha do domínio .pt em janeiro de 2008
- Possui conteúdos que remontam a 1996
- Foi precedido de projetos experimentais na Faculdade de Ciências da Universidade de Lisboa (TUMBA, 2000; TOMBA, 2006)

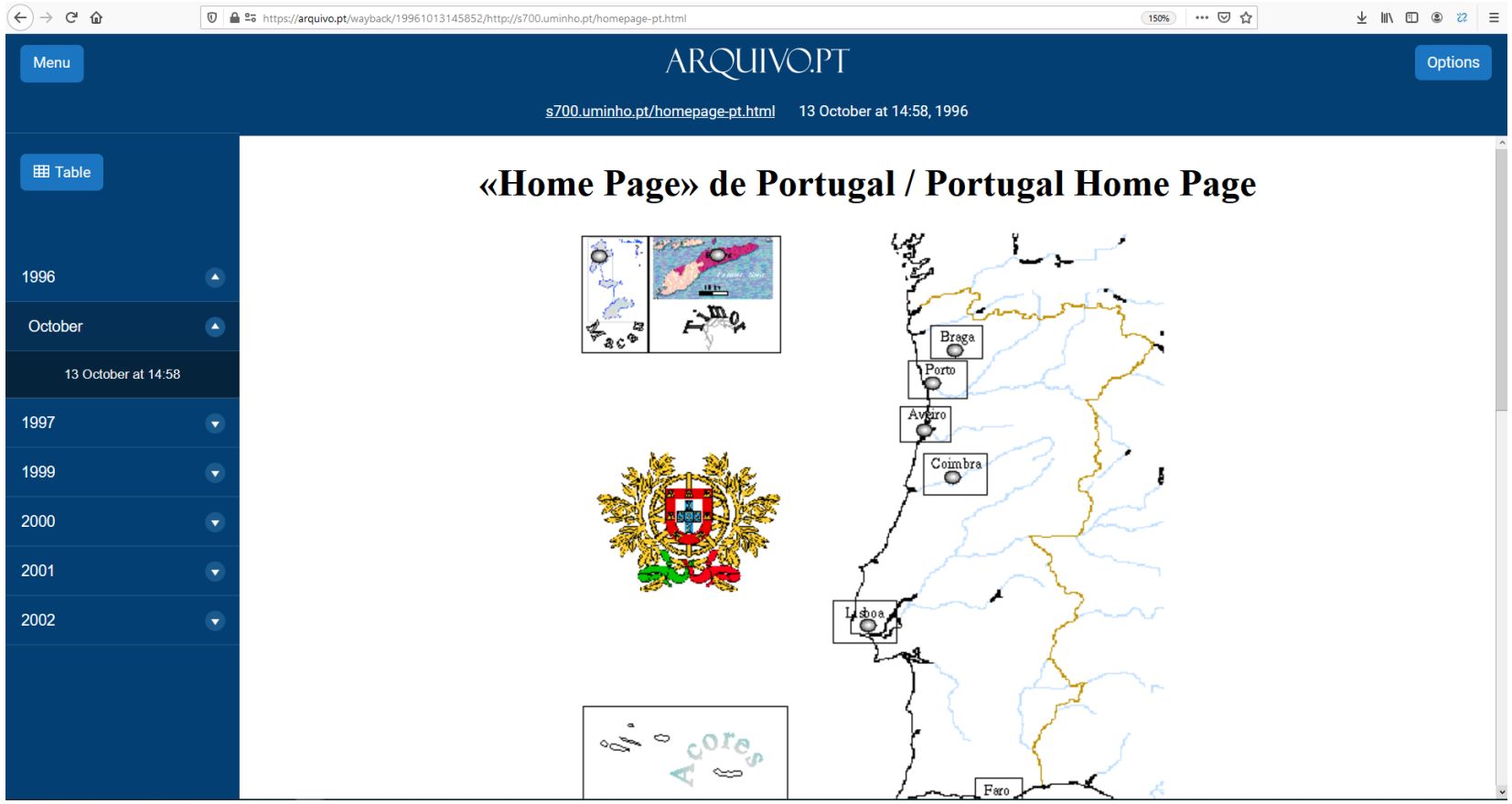
https://sobre.arquivo.pt/pt/arquivo-pt-de-volta-ao-seu-berco-na-fcul-20-anos-depois/

O Arquivo.pt



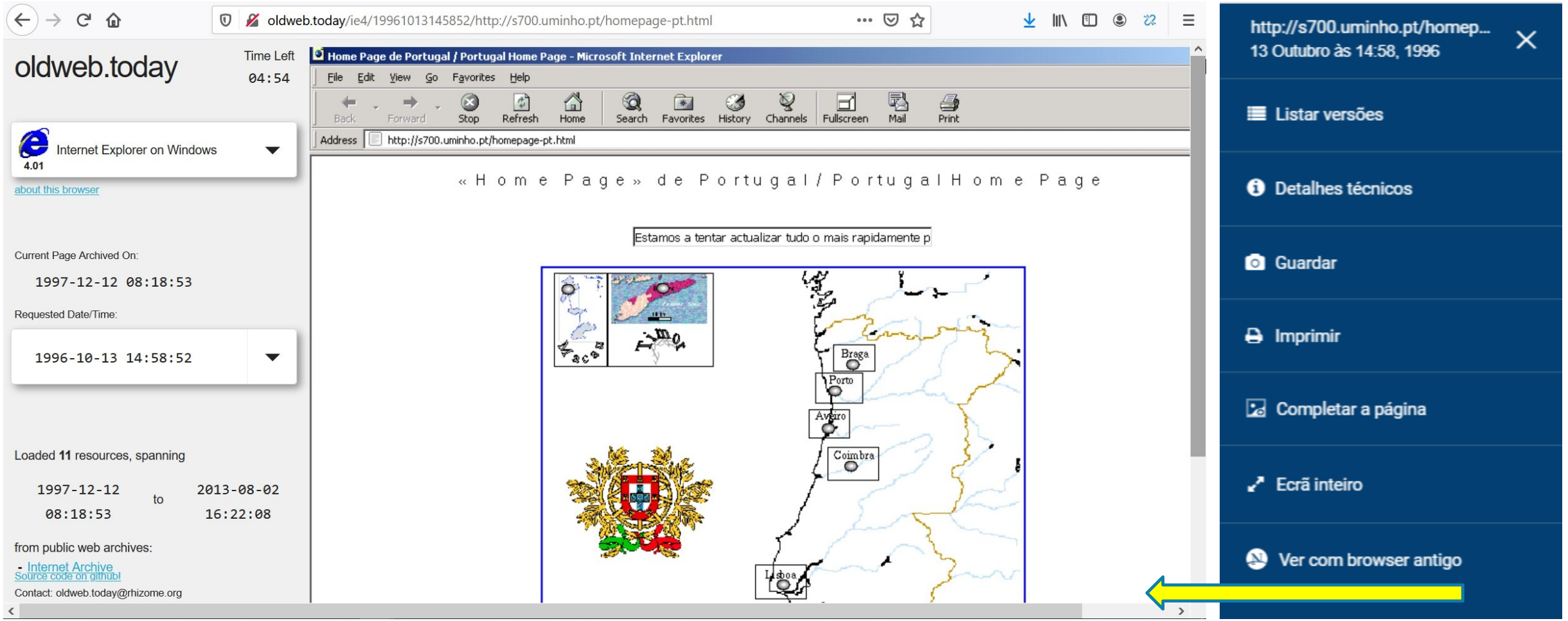
https://arquivo.pt/wayback/20061025224316/http://tomba.tumba.pt/

O Arquivo.pt



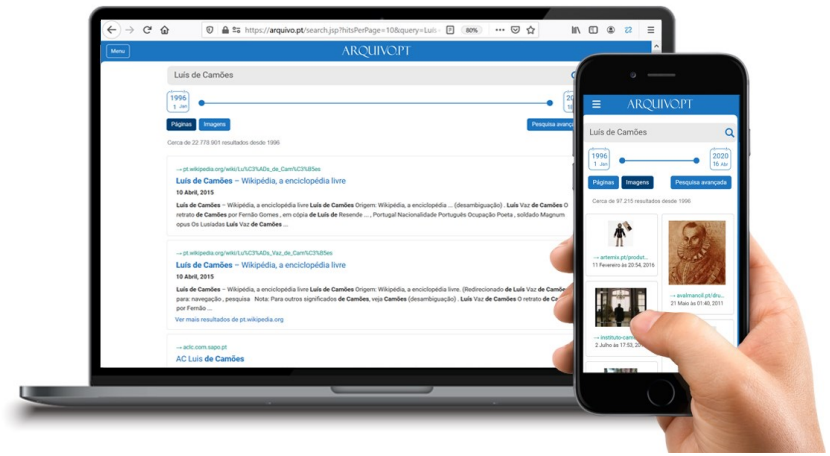
"Home page" de Portugal, em 1996
https://arquivo.pt/wayback/19961013145852/http://s700.uminho.pt/homepage-pt.html

O Arquivo.pt

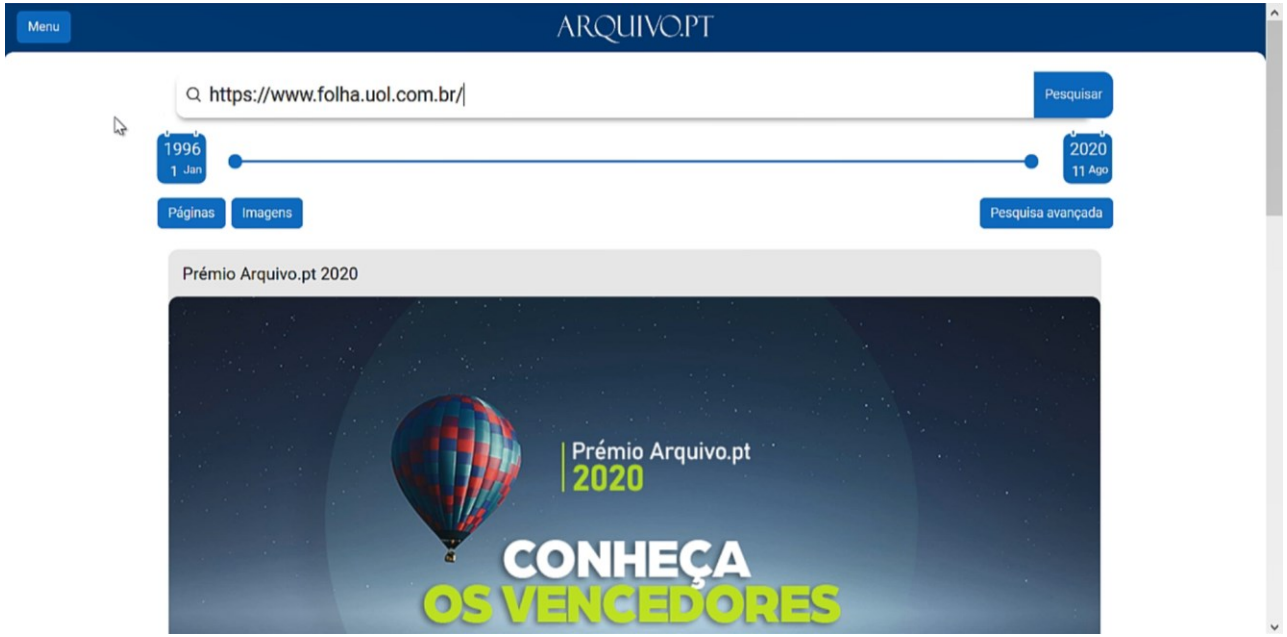


"Home page" de Portugal, em 1996 reproduzida por um browser antigo, através do serviço Oldweb.today, a partir do Arquivo.pt

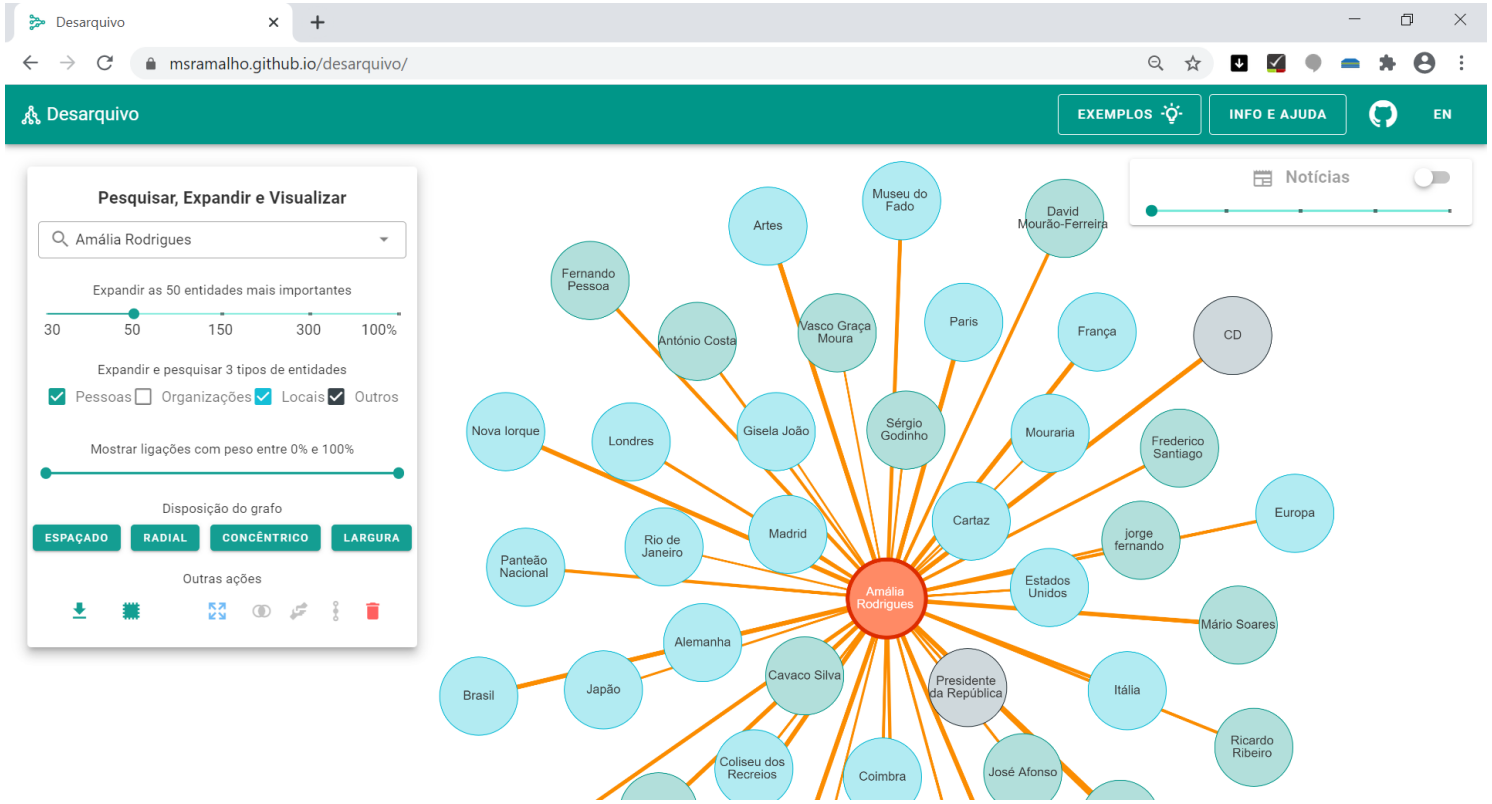
Utilização do Arquivo.pt



O Arquivo.pt


Menu
ARQUIVO.PT
https://www.folha.uol.com.br/
Pesquisar
1996
1 Jan
2020
11 Ago
Páginas
Imagens
Pesquisa avançada
Prémio Arquivo.pt 2020
Prémio Arquivo.pt
2020
CONHEÇA
OS VENCEDORES

O Arquivo.pt



Caso de uso: [Trabalho vencedor do Prémio Arquivo.pt 2020 “Desarquivo”], desenvolvido utilizando as APIs do Arquivo.pt. É uma forma alternativa de pesquisar.

O Arquivo.pt

Maquinaria e Software

- 77 servidores
- 15 TB de memória RAM
- 1 544 vCPU
- 3,5 PetaBytes de espaço em disco rígido

https://sobre.arquivo.pt/pt/imprensa/o-arquivo-pt-em-numeros/

II - Sites Preserváveis

Recomendações para **Publicar** Informação **Preservável**

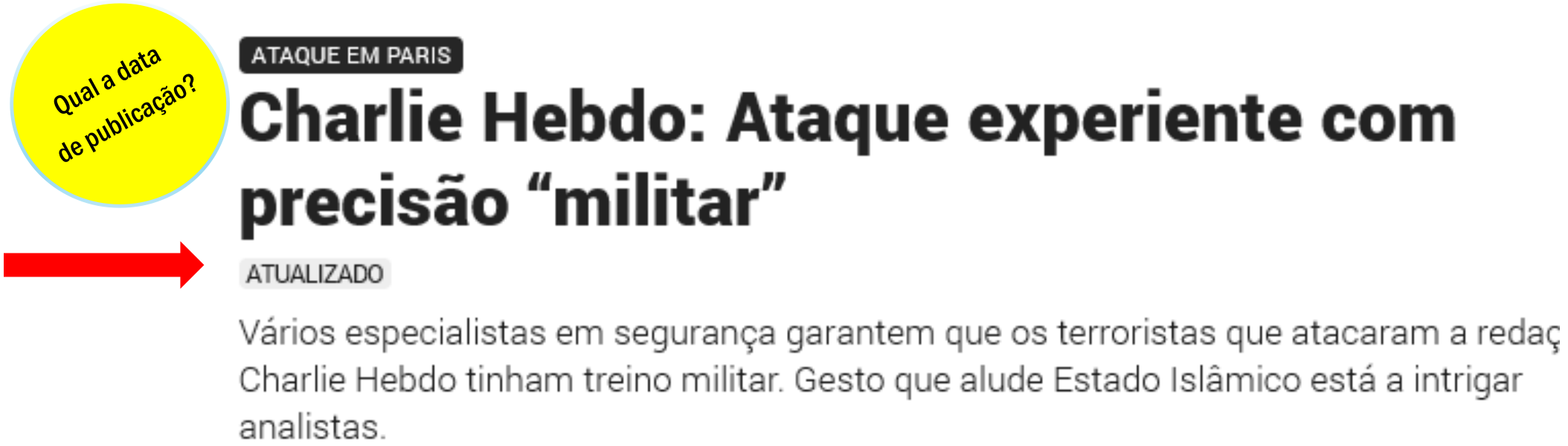
1

Identifique corretamente a data de publicação

Identifique corretamente a data de publicação

Caso de um site de notícias em que o sistema não apresenta a data da publicação de uma. Publicações sem data prejudicam a preservação no seu contexto.

Identifique corretamente a data de publicação



Onde costuma ser apresentada a data, o sistema de edição do jornal apresenta o estado de edição (“atualizado”). Para o consumo imediato de informação serve, mas para a preservação da página Web pode trazer problemas de contextualização temporal.

Identifique corretamente a data de publicação

- Configure o sistema para apresentar a data de publicação
- Introduza manualmente a data, sempre que o sistema o não permitir

2

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

```
User-agent: *

Disallow:
```

O dono do site apresenta para os robots de acesso automático ou crawlers (*user agent)* as condições de acesso.
Neste exemplo, o dono de um site utilizou apenas duas linhas que significam:

- As diretrizes dirigem-se a todos os robots ou crawlers; o asterisco (*) significa, neste caso, todos.
- Não há qualquer limite há recolha, nenhum diretório ou tipo de ficheiro vedado; “Disallow:” em branco significa nada a limitar.

Para saber mais sobre a sintaxe deste protocolo consulte: http://www.robotstxt.org/robotstxt.html

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

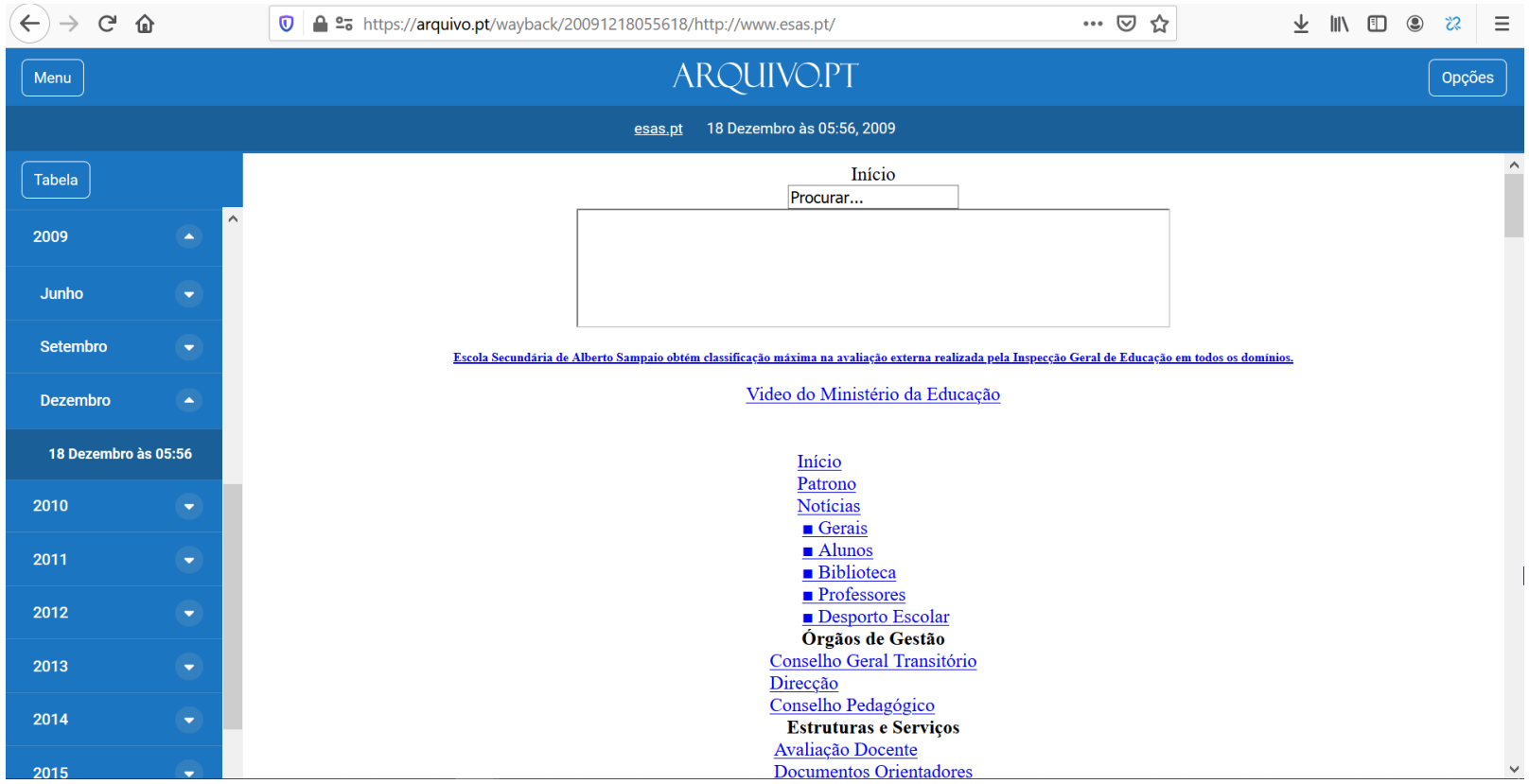
```
User-agent: Arquivo-web-crawler

Disallow:
```

O Arquivo.pt tem o seu crawler devidamente identificado pelo nome “Arquivo-web-crawler”.
Assim, se o dono de um site quiser permitir ao Arquivo.pt o acesso sem limitações pode fazê-lo.

Neste exemplo, o “Arquivo-web-crawler” não tem restrições à recolha de conteúdos.

O Robots Exclusion Protocol é uma declaração do dono do site que não é impeditiva do acesso. No entanto, o Arquivo.pt respeita essas diretrizes nas recolhas gerais à Web portuguesa. Assim, se os robots.txt restringiram o acesso a conteúdos importantes, tais como imagens, ficheiros CSS) a recolha do site fica incompleta. Para saber mais, consulte o Arquivo.pt

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*



Exemplo de uma página recolhida sem ter havido acesso acesso aos ficheiros CSS (Cascading Style Sheet) que produzem o estilo da página.

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

Neste caso, foi possível reconstituir o aspeto da página através da funcionalidade “Completar página”, também designada por *patching*, que o Arquivo.pt integrou no seu serviço.

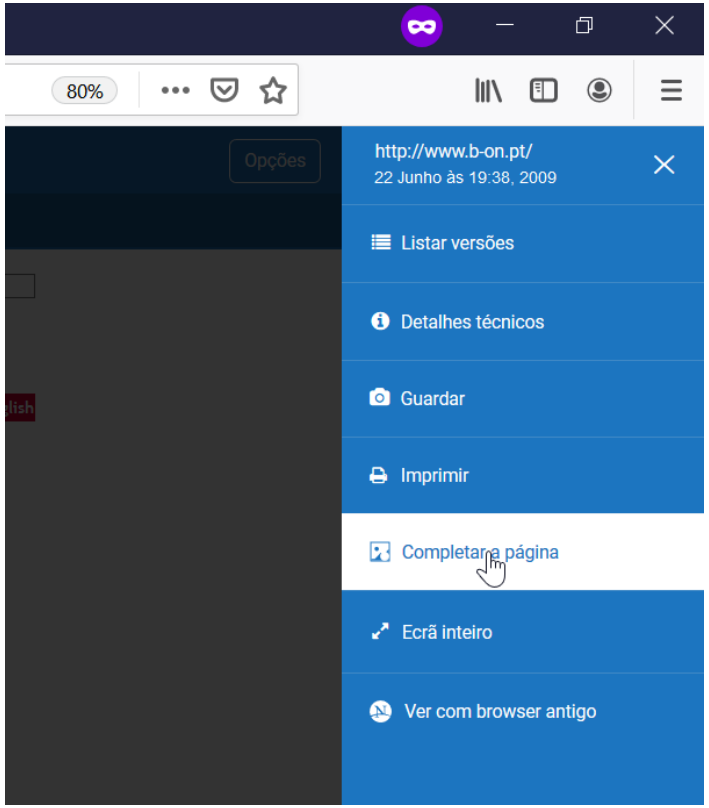
Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

A nova funcionalidade “Completar página”, foi disponibilizada pelo Arquivo.pt em julho de 2020.

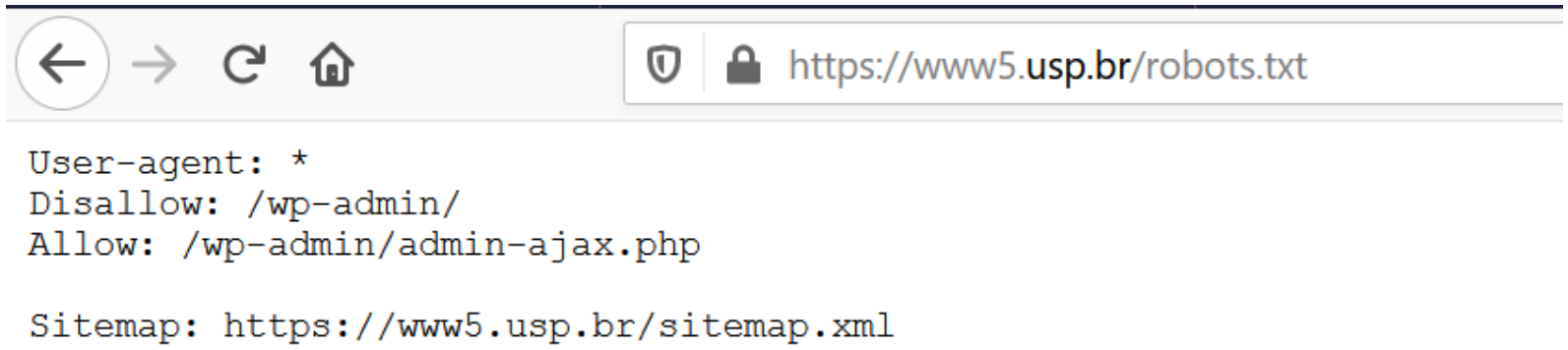
Ao utilizá-la sobre uma página “quebrada”, com muitos elementos em falta, é feita uma busca pelo Internet Archive e pela *live Web* para encontrar os elementos em falta.

Em alguns casos, o “Completar página” resolve o problema. Em muitos outros, sobretudo em páginas mais antigas, os conteúdos em falta são irrecuperáveis.

Daí a importância de uma recolha o mais completa possível no início do processo de preservação



Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*



https://www5.usp.br/robots.txt

Aceda ao robots.txt na raiz de um website (depois de /) para ver se encontra declarado alguma restrição que pode comprometer uma boa recolha por parte de *web crawler* para efeitos de preservação.

Neste caso, o sítio Web da Universidade de São Paulo autoriza o acesso automático aos conteúdos públicos e ainda disponibiliza um mapa do site (sitemap.xml) que facilita a descoberta dos links necessários para a recolha.

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

Os bloqueios no robot.txt, à parte usos abusivos, prejudicam a visibilidade de um website em motores de busca.
Por esse motivo, tem havido recomendações no sentido de reduzir ao mínimo os bloqueios.

Porque continua a haver bloqueios no Robots.txt ?

"**Disallowing** crawling of **Javascript** or **CSS** files in your site's robots.txt directly **harms** how well our algorithms render and index **your content** and can result in suboptimal rankings."

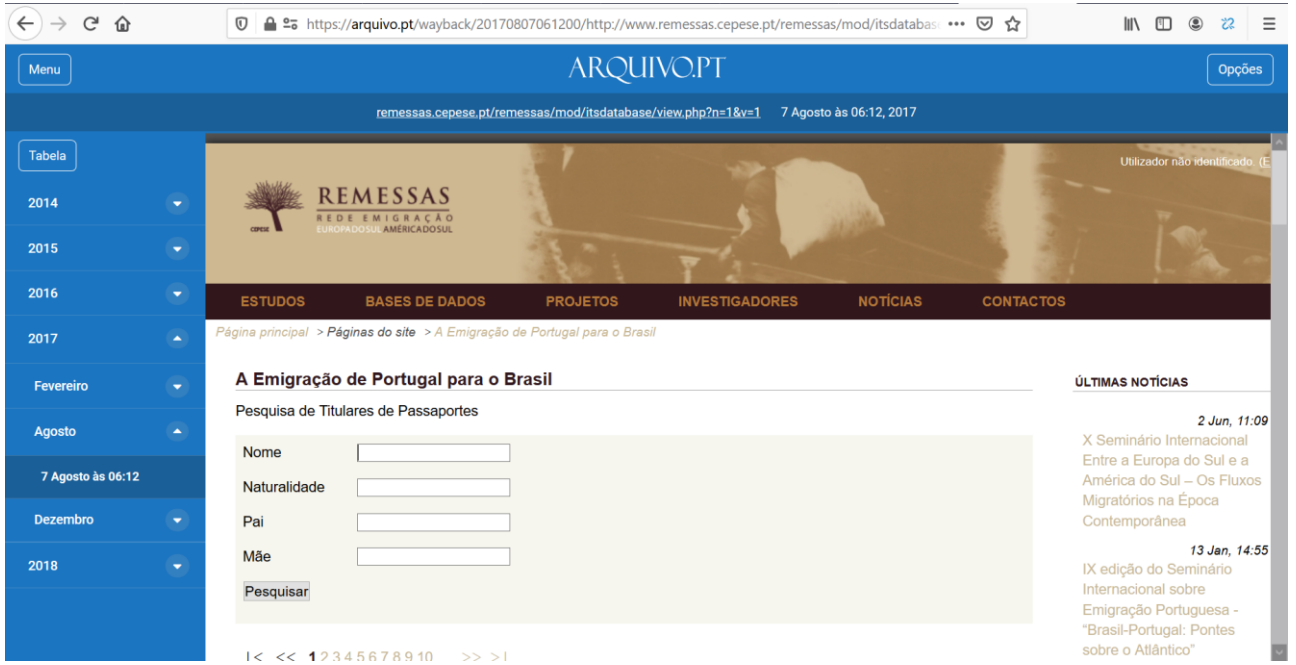
https://webmasters.googleblog.com/2014/10/updating-our-technical-webmaster.html?m=1

3

Utilize um endereço para cada conteúdo

Utilize um endereço para cada conteúdo

- Conteúdos atrás de formulários escapam às recolhas



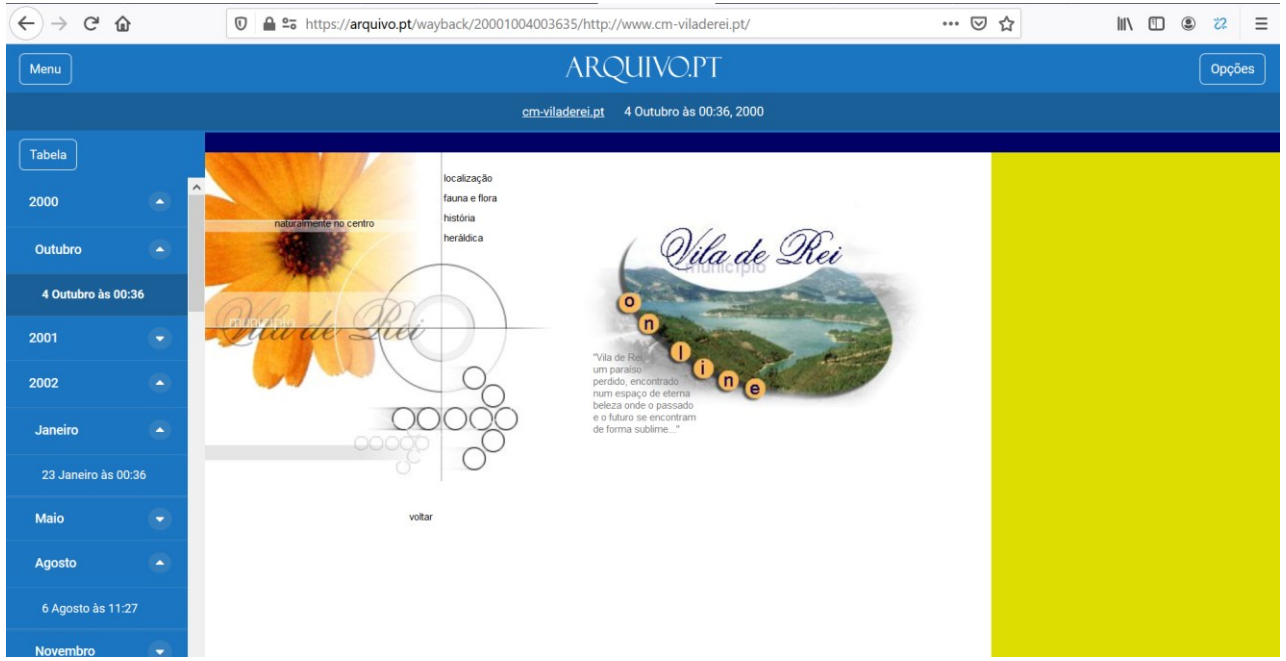
Conteúdos atrás de formulários escapam às recolhas do Arquivo.pt. Websites alimentados por uma base de dados necessitam que se preencha um ou vários campos de formulário para que os conteúdos sejam apresentados na página Web.

Neste caso, um formulário no website do CEPESE, Porto. O Arquivo.pt não preenche formulários automaticamente e por isso não vai chegar a um endereço específico para um determinado registo ou conteúdo. Assim, estes conteúdos escapam às suas recolhas.

Sites preserváveis devem ter explícitos um endereço para cada conteúdo. Os restantes, em base de dados, devem valer-se de um sistema de preservação próprio.

Utilize um endereço para cada conteúdo

- Conteúdos não identificados por um endereço único escapam às recolhas



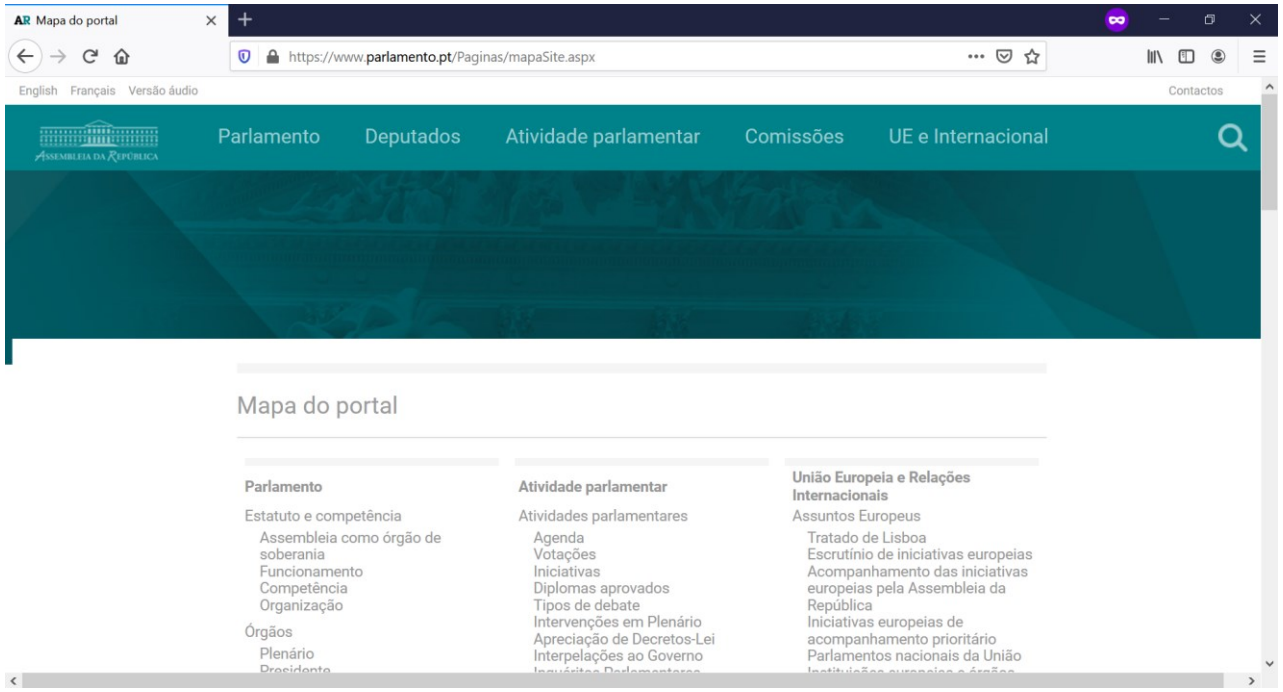
Neste caso, o Website do Município de Vila de Rei, Portugal, em Outubro de 2000, utilizou a tecnologia Flash para publicar conteúdos na Web.

Esta tecnologia não é apropriada para a Web, pois não identifica os diversos conteúdos com um endereço (URL). Assim, muitos conteúdos ficam escondidos atrás do *frame* principal, não sendo possível serem referenciados inequivocamente.

A imagem a que se refere este slide é do *frame* relativo à "localização, fauna, flora, heráldica". Como referenciar esta página se o endereço na barra do browser é o mesmo em todas as páginas?

Utilize um endereço para cada conteúdo

- Mapa do Site facilita acesso a pessoas (usabilidade) e máquinas (SEO)



Neste caso, o Website da Assembleia da República de Portugal, disponibilizou um mapa do site, útil para humanos e também para máquinas. Um *robot* ou *web crawler* pode facilmente identificar os links para os conteúdos importantes do website.

4

Mantenha o mesmo endereço ao longo do tempo

Mantenha o mesmo endereço ao longo do tempo

Não podemos evitar que as organizações mudem de nome ao longo do tempo e que os nomes dos seus websites mudem.

Cada mudança de nome num website dá origem a uma tabela de versões.

Assim, os websites que mudam de nome têm a suas versões ao longo do tempo representadas em mais do que uma tabela.

Neste exemplo, o "site antigo" apenas inclui as versões até 2010, sem ligação às seguintes do "novo-site".

Mude o nome do seu website quando estritamente necessário para não quebrar a representação das versões ao longo do tempo.

No caso de haver mudança é útil, do ponto de vista da preservação, fazer o redirecionamento do antigo para o novo domínio. Ver slide seguinte (caso do ISCTE, Lisboa)

Problema: quebra de histórico devido a mudança de endereço do site

Mantenha o mesmo endereço ao longo do tempo

Tabela de versões

460 versões de iscte.pt

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
22 Jan	28 Jan	3 Fev	7 Jan	13 Jan	4 Jan	6 Jan	20 Mai	31 Mai	21 Jan	21 Jan	5 Nov	26 Set
25 Mar	2 Fev	21 Mar	12 Jan	2 Fev	9 Jan	18 Jan	20 Mai	6 Jun	20 Mai	22 Jan	7 Nov	30 Set
25 Mai	18 Fev	21 Abr	13 Jan	6 Fev	13 Jan	26 Jan	21 Mai	5 Aug	22 Mai			
27 Mai	24 Mar	5 Jun	1 Fev	7 Fev	19 Jan	28 Jan	23 Jun					
5 Jun	27 Mar	11 Jun	4 Fev	2 Mar	21 Jan	6 Fev	24 Set					
24 Set	9 Abr	11 Jun	4 Fev	2 Abr	24 Jan	15 Fev	26 Set					
27 Set	10 Abr	12 Jun	8 Fev	25 Abr	27 Jan	11 Mar	17 Dez					
13 Nov	10 Abr	14 Jun	8 Fev	27 Abr	5 Fev	14 Mar	18 Dez					
26 Nov	31 Mai	23 Jun	9 Fev	2 Jun	20 Fev	21 Out						

http://iscte.pt → http://iscte-iul.pt

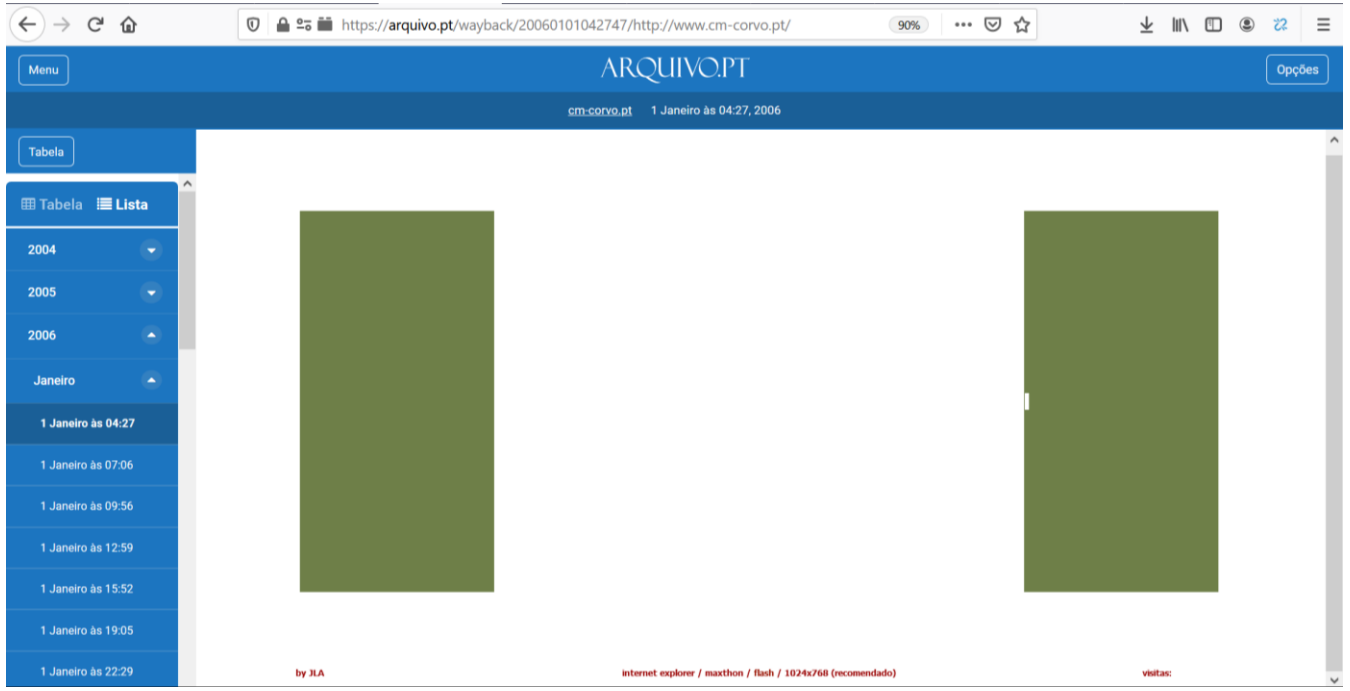
Redirecionamento

5

Utilize formatos adequados para preservação

Utilize formatos adequados para preservação

- No passado, o Flash como exemplo de formato pouco preservável
- E no presente? Há formatos pouco preserváveis nos websites?



https://arquivo.pt/wayback/20060101042747/http://www.cm-corvo.pt

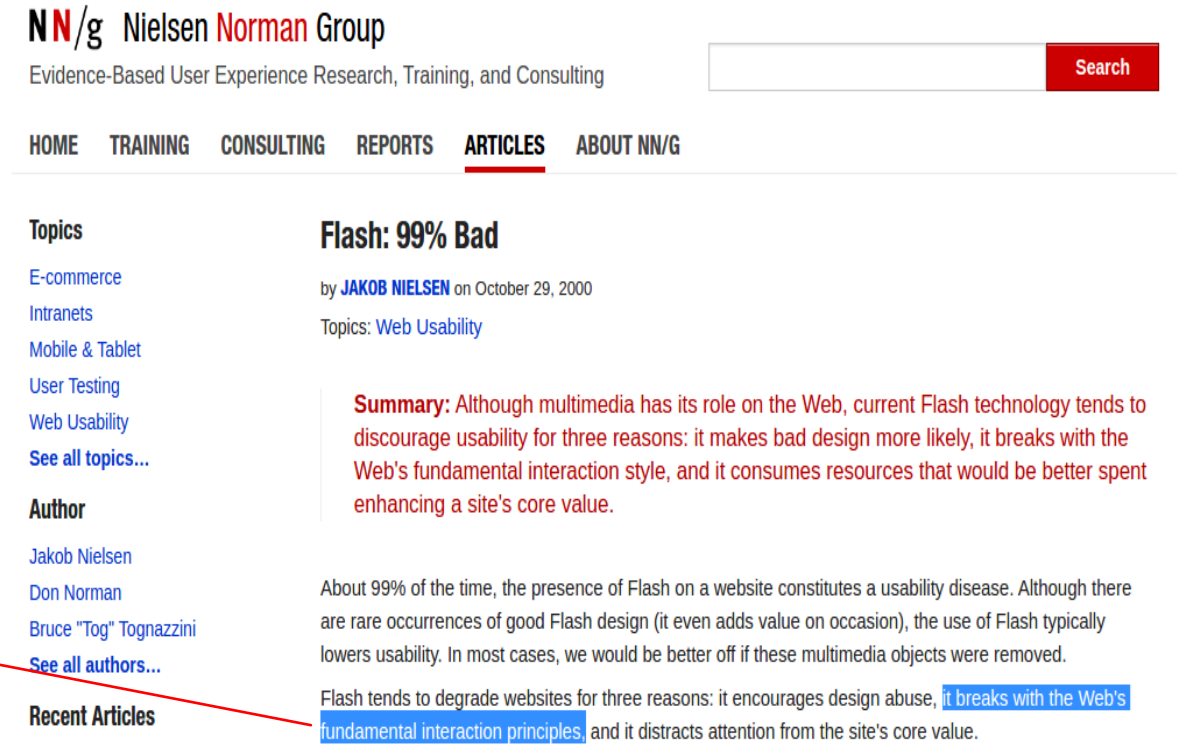
Um formato adequado para preservação deve ser aberto, não proprietário, compatível com as políticas do consórcio W3C.

Neste caso, o website do Município do Corvo, Açores, Portugal, 2006, em Flash, não pôde ser reproduzido pelo Arquivo.pt, ficando o seu conteúdos inacessível.

Utilize formatos adequados para preservação

- Caso da tecnologia Flash
  - Não foi criado especificamente para a Web
  - No entanto, foi utilizado por desenvolvedores

"It **breaks** with the Web's **fundamental** interaction principles"



Em 2000, Nielsen Norman tinha alertado para a quebra dos princípios da publicação na Web que originam perdas de informação.

As tecnologias mudaram, mas continua a utilizar-se formatos e técnicas de geração de conteúdos que quebram os princípios da Web e que podem dificultar a sua preservação. O principal princípio é que a tecnologia seja aberta.

Utilize formatos adequados para preservação

- Condições de **licenciamento** que permitam a sua utilização
- **Normas** emitidas por um organismo oficial (W3C)
- Documentados abertamente através de uma **especificação pública**
- Lidos e escritos por múltiplas plataformas de software, incluindo código-aberto
- Amplamente usados

Utilize formatos adequados para preservação

- Texto
  - HTML, XHTML ou XML
  - Open Document Text (.odt)
  - PDF/A-1 segundo a norma ISO 19005-1 (.pdf)
- Imagem
  - PNG (.png)
  - JPEG2000
- Video
  - AVI sem compressão (.avi)

Utilize formatos adequados para preservação

- Texto
  - HTML, XHTML ou XML
  - Open Document Text (.odt)
  - PDF/A-1 segundo a norma ISO 19005-1 (.pdf)
- Imagem
  - PNG (.png)
  - JPEG2000
- Video
  - AVI sem compressão (.avi)

6

Utilize metadados para descrever os conteúdos

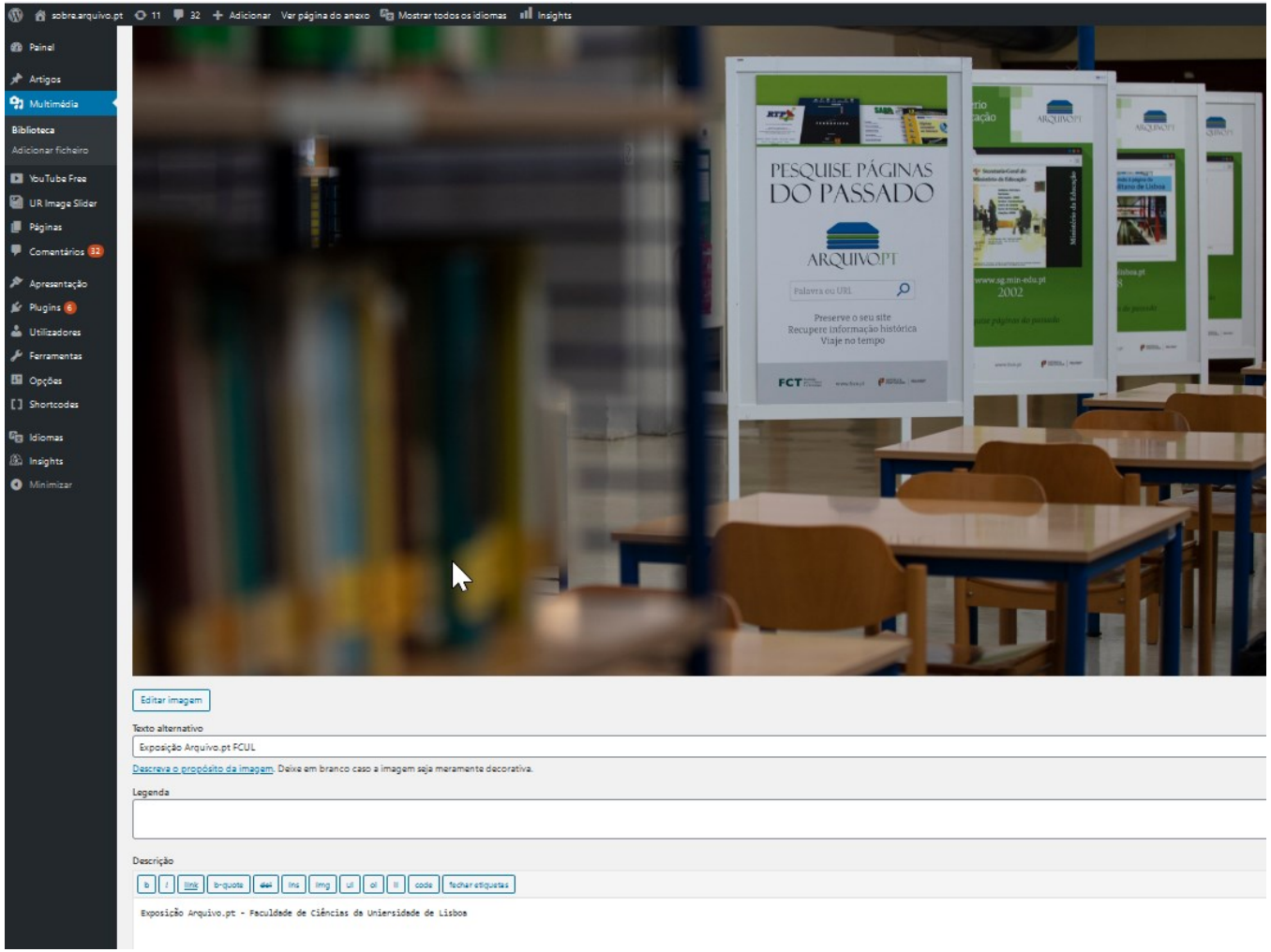
Utilize metadados para descrever os conteúdos

Neste exemplo, refere-se a plataforma Wordpress, por ser muito utilizada, a qual disponibiliza campos para descrever conteúdos (imagens, documentos PDF, etc).

É muito importante enriquecer os conteúdos descrevendo-os. No caso das imagens, por exemplo, o título, a legenda ou a descrição permitem uma recuperação e contextualização.

O Arquivo.pt tem uma pesquisa por imagens. A experiência tem mostrado que a falta de metadados associados às imagens dificulta a sua recuperação.

Alguns sistemas integram esquemas de metadados descritivos, como por exemplo como o Dublin Core. Todos os metadados são uma mais valia do ponto de vista da preservação. Servem para extrair de forma mais especializada informação de um arquivo da Web.



7

Torne-se curador dos seus Websites

Torne-se curador dos seus Websites

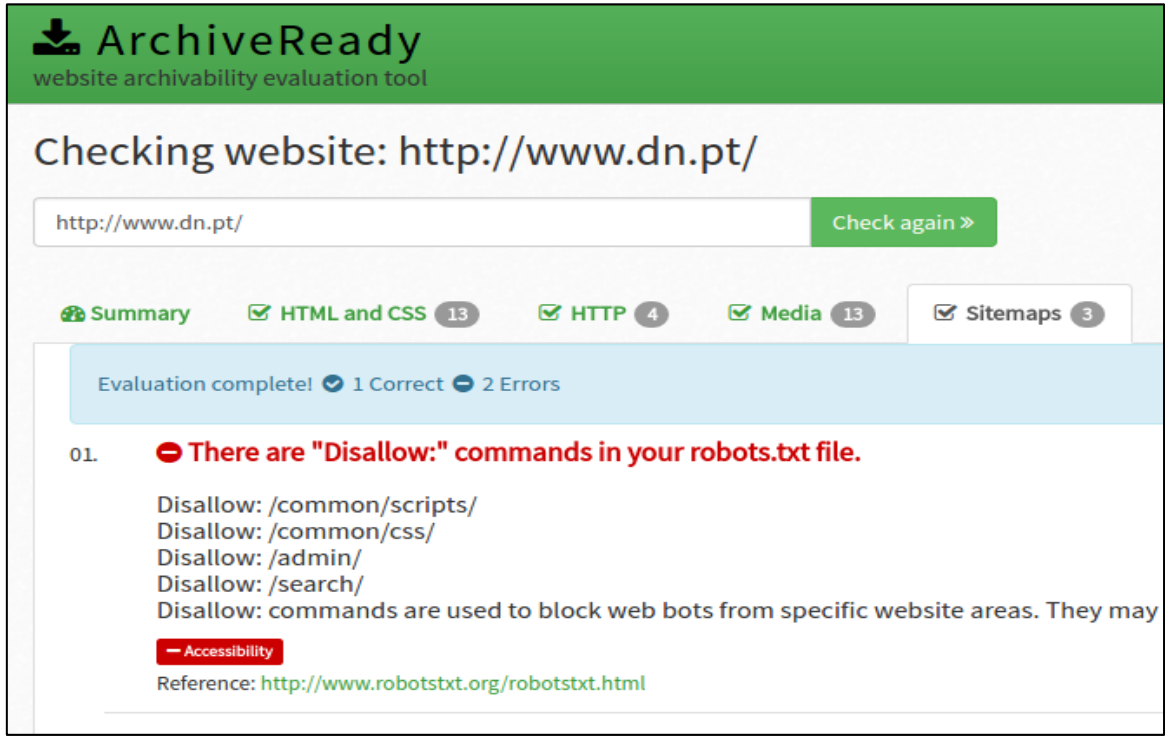
A ferramenta ArchiveReady é útil para considerar um Website sob o ponto de vista da preservabilidade.

Neste exemplo, verificamos que o Diário de Notícias (Lisboa) apresenta alguns bloqueios à recolha do seu website, nomeadamente ao scripts e aos CSS. É grande a probabilidade de resultar em preservação de páginas incompletas.

ArchiveReady contém referências formativas, muito úteis para um curador digital sensibilizar a sua organização para a preservação.

Um curador digital, mais propriamente curador da Web, numa organização, é alguém incumbido de vigiar pela preservação dos websites e canais socais dessa mesma organização.

Antes mesmo de haver uma política ou um fluxo de preservação, a sua função concretiza-se em pequenas ações de vigilância, sugestão de melhorias e identificação de páginas para serem preservadas em arquivos da Web existentes (ex. Internet Archive, Arquivo.pt, outros)



http://archiveready.com

Recomendações do Arquivo.pt

1. Identifique corretamente a data de publicação
2. Use corretamente o protocolo de exclusão de robots
3. Use um endereço para cada conteúdo
4. Mantenha o mesmo endereço ao longo do tempo
5. Utilize formatos adequados para preservação
6. Publique metadados para descrever os conteúdos
7. Torne-se curador dos seus websites

arquivo.pt/recomenda

III - Ações de Preservação

Curadoria Digital

Funcionamento do Arquivo.pt

- Consideramos 3 fases principais

1

Recolha

ARQUIVO.PT

2

Armazenamento

3

Reprodução

Terminologia associada

- Terminologia associada a cada fase

1

Recolha

Crawler

ARQUIVO.PT

2

Armazenamento

WARC
ARC

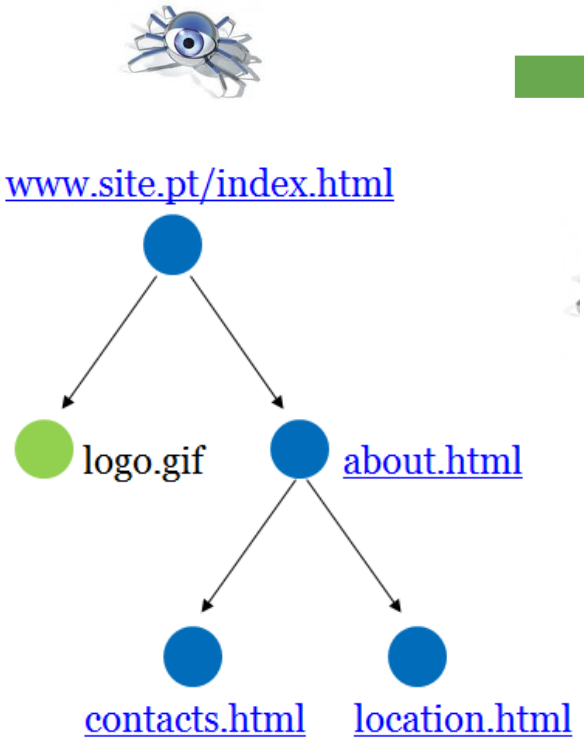
3

Reprodução

Wayback

Funcionamento do Arquivo.pt

- Representação da recolha automática: “crawlers” ou “robots” seguem links



ARQUIVO.PT

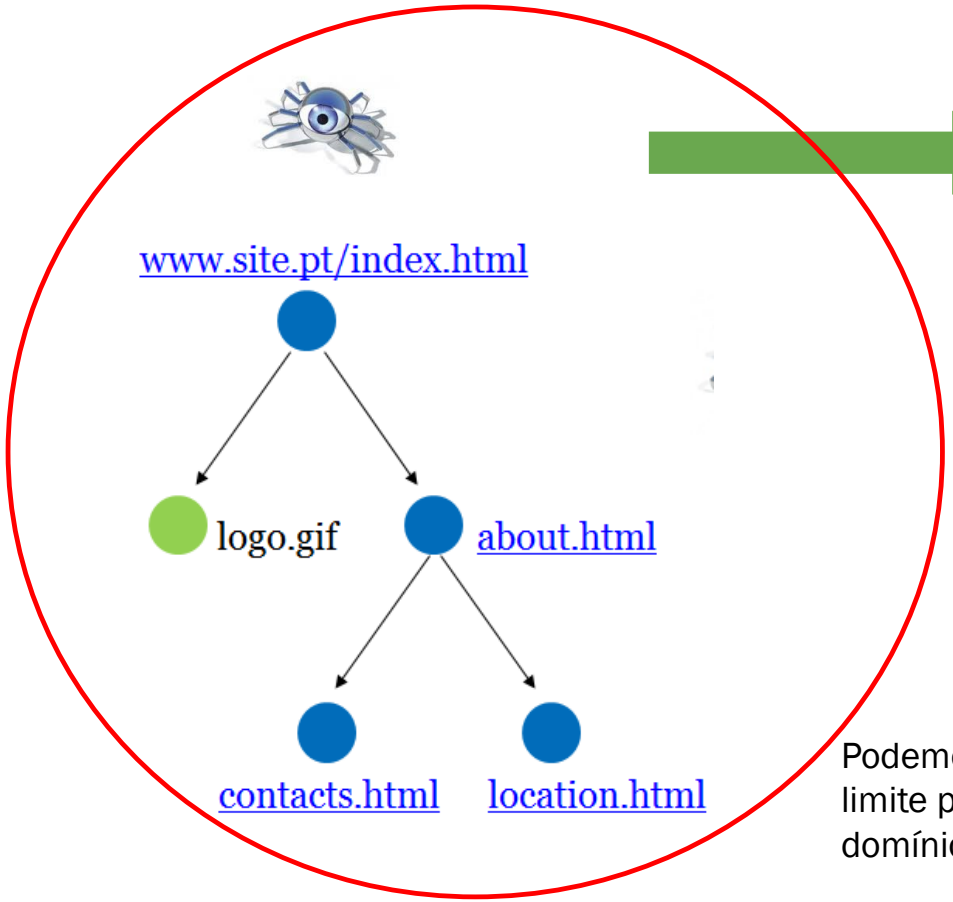
2

3

Num arquivo da Web um *crawler* ou *robot* é um software utilizado para identificar *links*, segui-os e identificar outros *links*, sucessivamente e de uma forma não hierárquica, permitindo assim localizar os conteúdos que vão ser gravados. Para as recolhas automáticas o Arquivo.pt utiliza o Heritrix e o Brozzler.

Funcionamento do Arquivo.pt

- Representação dos limites impostos ao crawler enquanto recolhe, âmbito,"scope"



ARQUIVO.PT

2

3

Podemos determinar os limites, âmbito, ou *scope* para a ação do *crawler*. Por exemplo, um limite por domínio significa que o *crawler* identifica todos os links num determinado domínio e não sai fora. Neste exemplo, estamos perante um *domain crawl*.

Funcionamento do Arquivo.pt

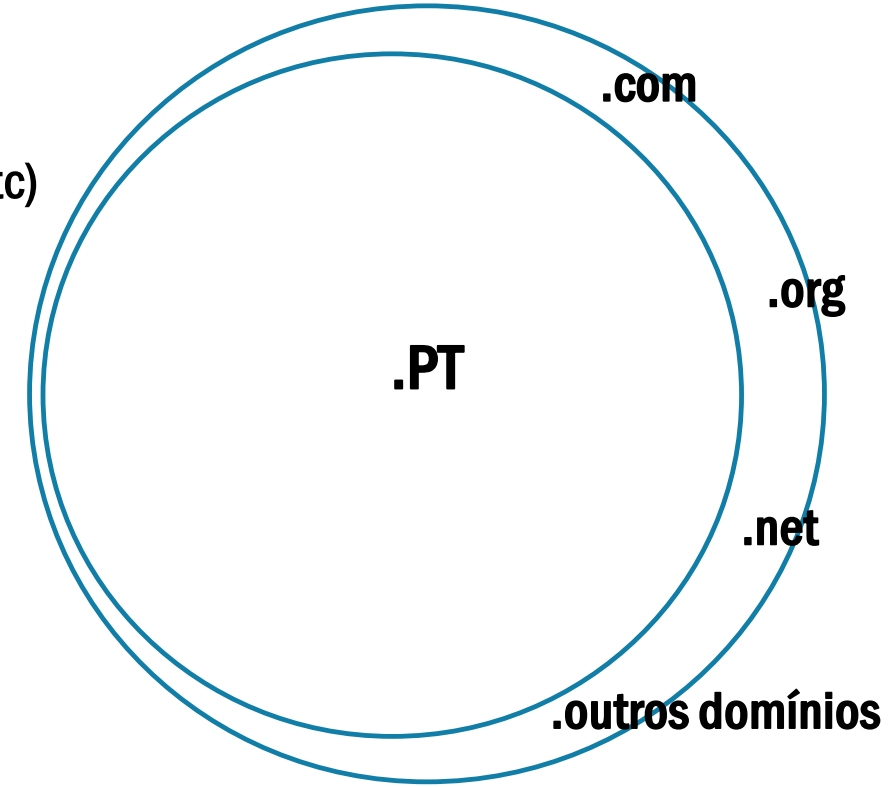
- Representação dos limites impostos ao crawler enquanto recolhe, âmbito,"scope" (exemplo, 1 salto fora do domínio)

Em alguns casos, é necessário alargar o âmbito da recolha e sair fora do website para seguir links importantes referenciados no website, mas que se encontram fora do domínio. Neste caso, estamos perante um *domain crawl* acrescido de 1 salto para fora do domínio.

Âmbito da recolha do Arquivo.pt (“scope”)

- Websites com domínio .pt
- Websites portugueses de outros domínios (ex. .org, .com, .net, .art, etc)
- Websites de interesse nacional (ex. União Europeia)
- Websites sugeridos pela comunidade de utilizadores -

  https://arquivo.pt/sugerir
- Websites no âmbito de colaborações internacionais



O domínio .PT é o foco das recolhas do Arquivo.pt, sem excluir websites qualquer outro domínio.

Periodicidade de recolha no Arquivo.pt

- Periodicidade de recolha no Arquivo.pt
  - Diária (conjunto de ~150 Websites)
  - Mensal (conjunto de ~90 Websites)
  - Trimestral (> 1 milhão de Websites)
  - Especial (ex. recolha sobre as Eleições Europeias de 2019, ~9000)

Profundidade da recolha no Arquivo.pt

- 10 mil links por domínio nas recolhas genéricas
- 5 “saltos”
- Ficheiros até 10 MB

Estes limites são removidos em recolhas especiais ou em recolhas de alta qualidade

A profundidade tem a ver com o caráter mais ou menos extenso da recolha. Dez mil links é o suficiente para a maior parte dos websites. Cinco “saltos”, ou seja, cinco pontos de partida num website para encontrar novos links basta para cobrir todas as páginas de um website.

No entanto, em portais que, por sua vez contêm dezenas de sub-sites no seu interior é necessário aumentar o número de links (até 100 mil) ou de saltos (até 10) para recolher com profundidade todos os conteúdos.

O Arquivo.pt, num compromisso entre qualidade e recursos disponíveis, colocou um limite de 10 MB para a recolha de certos recursos. Assim, por exemplo, vídeos que ultrapassam frequentemente os 10 MB não são recolhidos. O mesmo acontece com ficheiros demasiado grandes.

Para recolher esses conteúdos, em casos especiais, o Arquivo.pt faz o que designou por “recolha de alta qualidade”.

Software de recolha

- Heritrix
- Brozzler
- Webrecorder
- Browsertrix

O Arquivo.pt utiliza o Heritrix para as recolher gerais ou para grande número de *seeds.*
O Brozzler é utilizado para recolhas de alta qualidade, por vezes como complemento às recolhas do Heritrix.

O Webrecorder é utilizado no Arquivo.pt pelo curador digital para recolhas manuais de alta qualidade, tais como vídeos, páginas do Facebook e páginas especiais que não podem esperar pelas recolhas mensais ou trimestrais. É recomendado nas ações de formação como ferramenta útil para a curadoria de páginas institucionais e pessoais.

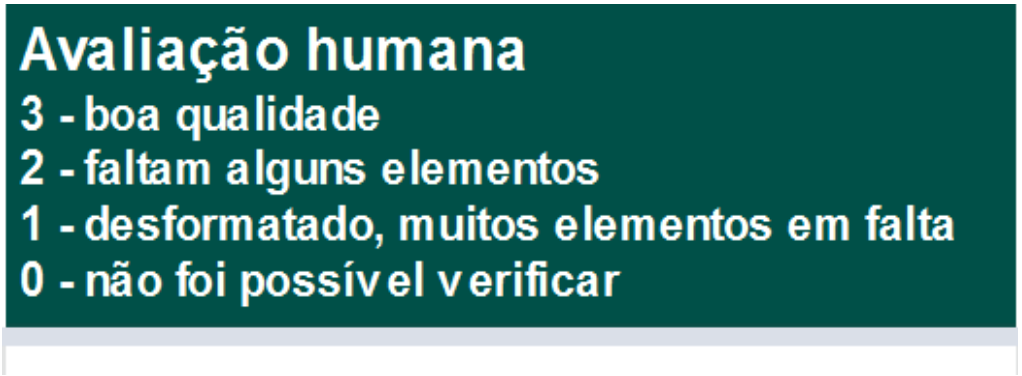
Qualquer pessoa pode utilizar o Webrecorder para recolher de forma personalizada páginas da sua instituição. Pode utilizar o Webrecorder no serviço online Conifer ou no seu próprio computador descarregado o Webrecorder Desktop. Uma página Web é gravada/armazenada num formato normalizado (WARC) e pode ser integrada num arquivo da Web.

A periodicidade das recolhas do Internet Archive ou do Arquivo.pt por vezes não coincidem com as necessidades de preservação de páginas importantes para as instituições. O Webrecorder utilizado localmente pode ajustar-se ao ritmo de cada instituição.

O Browsertrix é uma automatização do Webrecorder para gravar sites inteiros. Exige apoio técnico, mas faz o fundamental que é a gravação automática de conteúdos num formato normalizado. Para instalar siga um tutorial.

Avaliação de qualidade

- Processo manual pelo qual o curador verifica a qualidade de conjuntos delimitados de sites
- Da avaliação decorrem ações de melhoria no processo de preservação



A avaliação manual de pequenos conjuntos de websites segue um processo simples de pontuação com o objetivo de identificar os que estão em pior estado, ou seja, com pior aparência, sem imagens, sem estilo, etc.

Habitualmente, o tratamento de pequenas coleções de websites visa divulgar o Arquivo.pt junto de um determinado grupo de utilizadores. Quanto melhor for a qualidade da preservação mais fácil é expor websites antigos. Ver por exemplo, as exposições online do Arquivo.pt.

A avaliação da qualidade por processos automáticos pode ser feita para alguns aspetos, tais como detetar páginas que têm ou não CSS. No entanto, não existe ainda um processo que avalie a qualidade de forma tão completa como a observação humana.

Avaliação de qualidade

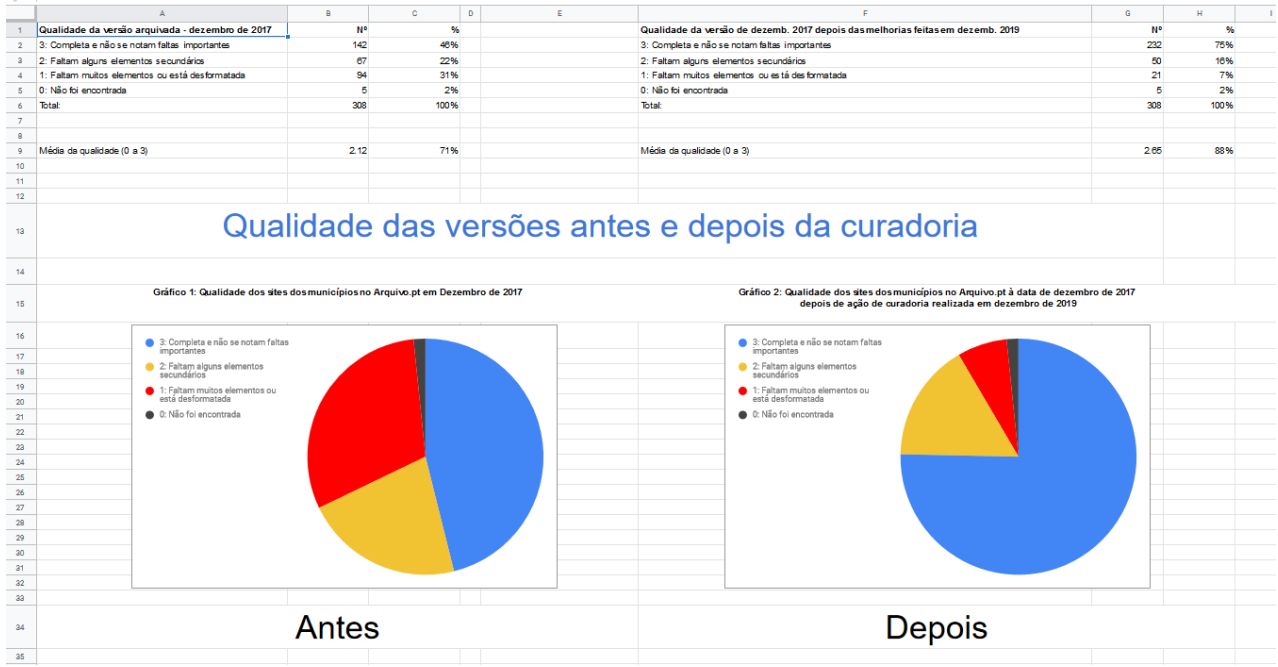
Neste exemplo de curadoria, foi feita uma avaliação a 308 websites dos 308 Municípios Portugueses.

Começou-se por avaliar a versão mais próxima da data 31 de dezembro de 2017, foi atribuída uma pontuação de 1 a 3 para a qualidade do website, sendo 1=mau e 3=bom.

Foram identificadas as causas para o mau estado de preservação de algumas páginas. Procedeu-se a melhorias, nomeadamente, o processo de “Completar página” ou *patching*. Voltou-se a avaliar, depois das melhorias.

Mostrou-se assim que a curadoria manual dos websites melhora grandemente a reprodução de páginas preservadas e que essa ação não requer especialização em tecnologia, podendo ser feita por bibliotecários, arquivistas ou pessoas da comunicação e imagem.

O mesmo procedimento foi aplicado a diversos conjuntos de websites (jornais, televisão, rádio, festivais de música).



Qualidade da versão arquivada - dezembro de 2017	Nº	%
3: Completa e não se notam faltas importantes	142	46%
2: Faltam alguns elementos secundários	67	22%
1: Faltam muitos elementos ou está desformatada	94	31%
0: Não foi encontrada	5	2%
Total:	308	100%
Média da qualidade (0 a 3)	2.12	71%

Qualidade da versão de dezemb. 2017 depois das melhorias feitas em dezemb. 2019	Nº	%
3: Completa e não se notam faltas importantes	232	75%
2: Faltam alguns elementos secundários	50	16%
1: Faltam muitos elementos ou está desformatada	21	7%
0: Não foi encontrada	5	2%
Total:	308	100%
Média da qualidade (0 a 3)	2.65	88%

Qualidade das versões antes e depois da curadoria

Curadoria digital sobre os Websites dos Municípios Portugueses

Avaliação de qualidade

	A	B	C	D	E	F	G	H	I
1	Distrito	Município	URL	Tabela de versões - Todas as versões no Arquivo.pt	Qualidade 1 - desformatada, muitos element 2 - alguns elementos em falta 3 - boa	Em falta	Em falta	Notas 1	Notas 2
2	SANTARÉM	Abrantes	http://www.cm-abrantes.pt	https://arquivo.pt/wayback/*/http://www.cm-abrantes.pt	1	css	img	robotstxt	_
3	AVEIRO	Águeda	http://www.cm-agueda.pt	https://arquivo.pt/wayback/*/http://www.cm-agueda.pt	3	_	_	_	_
4	GUARDA	Aguiar da Beira	http://www.cm-aguiardabeira.pt	https://arquivo.pt/wayback/*/http://www.cm-aguiardabeira.pt	1	css	img	robotstxt	
5	ÉVORA	Alandroal	http://www.cm-alandroal.pt	https://arquivo.pt/wayback/*/http://www.cm-alandroal.pt	1	_	_	_	outro
6	AVEIRO	Albergaria-a-Velha	http://www.cm-albergaria.pt	https://arquivo.pt/wayback/*/http://www.cm-albergaria.pt	2	_	_	_	_
7	FARO	Albufeira	http://www.cm-albufeira.pt	https://arquivo.pt/wayback/*/http://www.cm-albufeira.pt	3	_	_	_	_
8	SETÚBAL	Alcácer do Sal	http://www.cm-alcacerdosal.pt	https://arquivo.pt/wayback/*/http://www.cm-alcacerdosal.pt	3	_	_	_	_
9	SANTARÉM	Alcanena	http://www.cm-alcanena.pt	https://arquivo.pt/wayback/*/http://www.cm-alcanena.pt	1	css	img	robotstxt	_
10	LEIRIA	Alcobaça	http://www.cm-alcobaca.pt	https://arquivo.pt/wayback/*/http://www.cm-alcobaca.pt	2	css	img	_	_
11	SETÚBAL	Alcochete	http://www.cm-alcochete.pt	https://arquivo.pt/wayback/*/http://www.cm-alcochete.pt	3	_	_	_	_
12	FARO	Alcoutim	http://www.cm-alcoutim.pt	https://arquivo.pt/wayback/*/http://www.cm-alcoutim.pt	2	_	img	_	_
13	LISBOA	Alenquer	http://www.cm-alenquer.pt	https://arquivo.pt/wayback/*/http://www.cm-alenquer.pt	2	_	_	_	flash
14	BRAGANÇA	Alfândega da Fé	http://www.cm-alfandegadafe.pt	https://arquivo.pt/wayback/*/http://www.cm-alfandegadafe.pt	3	_	_	_	_
15	VILA REAL	Alijó	http://www.cm-alijo.pt	https://arquivo.pt/wayback/*/http://www.cm-alijo.pt	3	_	_	_	_
16	FARO	Aljezur	http://www.cm-aljezur.pt	https://arquivo.pt/wayback/*/http://www.cm-aljezur.pt	3	_	_	_	_
17	BEJA	Aljustrel	http://www.mun-aljustrel.pt	https://arquivo.pt/wayback/*/http://www.mun-aljustrel.pt	2	_	_	_	desformat
18	SETÚBAL	Almada	http://www.m-almada.pt	https://arquivo.pt/wayback/*/http://www.m-almada.pt	1	css	img	_	ver
19	GUARDA	Almeida	http://www.cm-almeida.pt	https://arquivo.pt/wayback/*/http://www.cm-almeida.pt	3	_	_	_	_
20	SANTARÉM	Almeirim	http://www.cm-almeirim.pt	https://arquivo.pt/wayback/*/http://www.cm-almeirim.pt	2	css	_	_	_
21	BEJA	Almodôvar	http://www.cm-almodovar.pt	https://arquivo.pt/wayback/*/http://www.cm-almodovar.pt	3	_	_	_	
22	SANTARÉM	Alpiarça	http://www.cm-alpiarca.pt	https://arquivo.pt/wayback/*/http://www.cm-alpiarca.pt	3	_	_	_	_
23	PORTALEGRE	Alter do Chão	http://www.cm-alter-chao.pt	https://arquivo.pt/wayback/*/http://www.cm-alter-chao.pt	1	_	_	robots.txt	_
24	LEIRIA	Alvaiázere	http://www.cm-alvaiazere.pt	https://arquivo.pt/wayback/*/http://www.cm-alvaiazere.pt	3	_	_	_	_

Folha de trabalho - Curadoria digital sobre os Websites dos Municípios Portugueses

Conclusão

Conclusão

1. O Arquivo.pt é um serviço público de preservação dos conteúdos para a memória e pesquisa
2. A criação de Websites preserváveis exige o envolvimento da comunidade
3. As ações de preservação são ajustadas a conjuntos diversificados de Websites
4. O curador digital contribui para a validação da qualidade da preservação

A existência de um curador digital dos sítios Web institucionais, em cada organização, contribuiria para a melhoria da qualidade da preservação

Torne-se curador dos seus Websites!

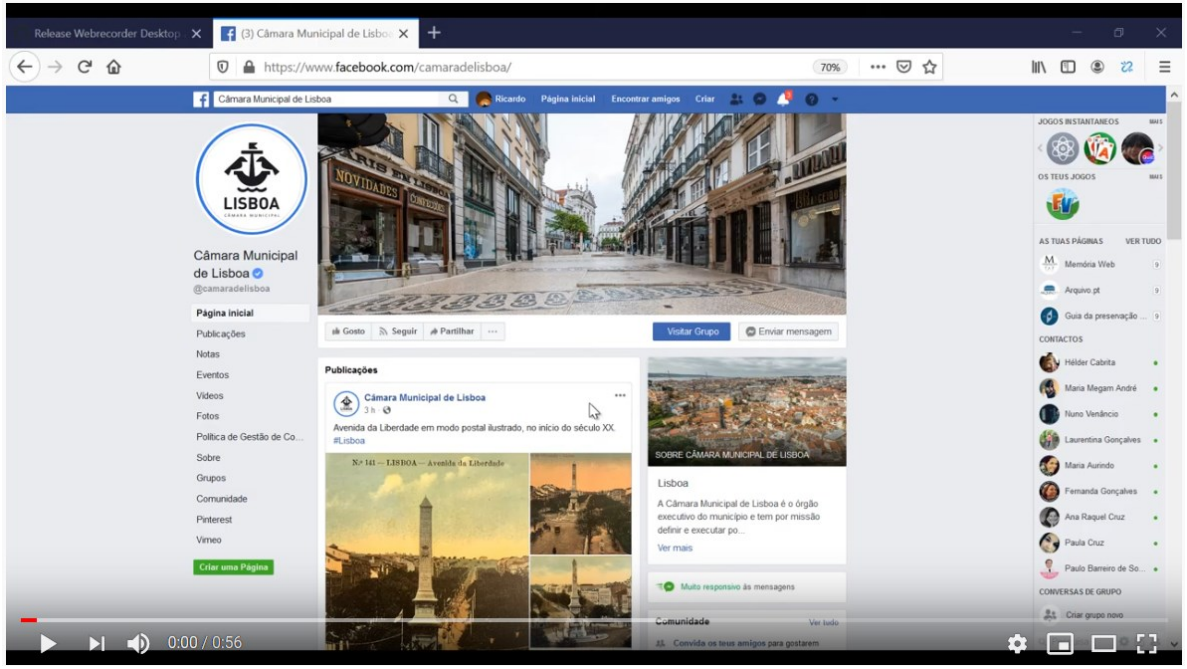
Torne-se curador dos seus Websites!

Torne-se curador dos seus Websites e das suas páginas sociais!

Use o Webrecorder e compreenda melhores alguns aspetos da preservação da Web. Tutorial.

Se tiver apoio técnico aprofunde utilizando o Browsertrix. Tutorial.

É recomendável recolher páginas do Facebook sem fazer o *login* para obter apenas conteúdos visíveis publicamente. A recolha de conteúdos institucionais em redes fechadas é uma questão ainda por resolver. Uma abordagem prática pode ajudar a ponderar que tipo de preservação se pode aplicar a este tipo de canais.

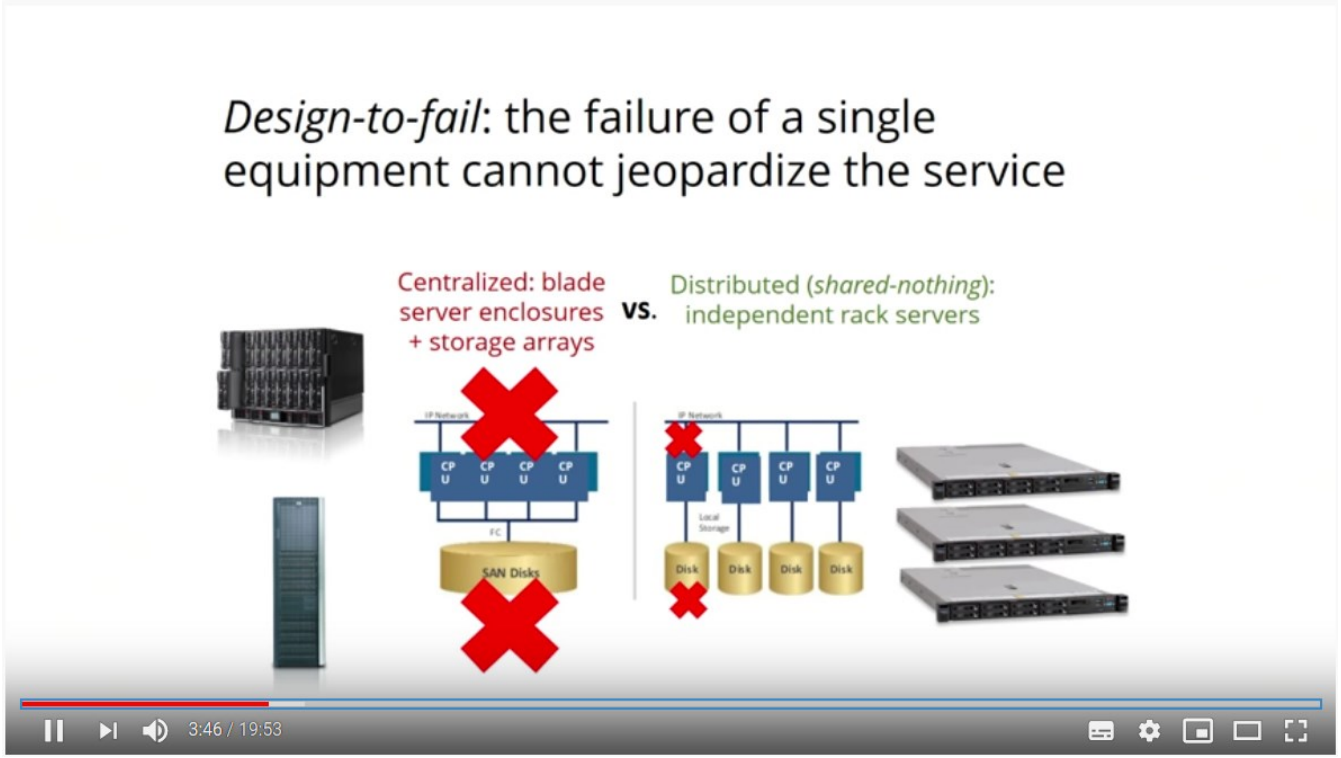


Demo de gravação com o Webrecorder (versão Desktop) de página social do Município de Lisboa

O principal obstáculo/desafio!

A mudança constante das tecnologias é o maior obstáculo/desafio para manter um serviço de arquivo da Web .
Tem de ser robusto para não quebrar.

Veja as lições aprendidas com a quebra de serviço em 2013, por Daniel Gomes, Gestor do Arquivo.pt.



Improving the robustness of the Arquivo.pt web archive

Contacto:

contacto@arquivo.pt

Para receber novidades do Arquivo.pt

https://arquivo.pt/inscrever

Sugerir Websites para preservação

https://arquivo.pt/sugerir

Modelo Biografia: Ricardo Basílio, Mestre em Ciências da Documentação e da Informação pela Faculdade de Letras da Universidade de Lisboa (FLUL), trabalhou na Biblioteca de Arte da Fundação Calouste Gulbenkian e no Centro de Documentação Investigação e Doutoramentos na Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa (NOVA FCSH). Atualmente, é curador digital no Arquivo.pt, o serviço de preservação da Web portuguesa, atividade desenvolvida no âmbito do projeto de Humanidades Digitais da NOVA FCSH, denominado Infraestrutura ROSSIO. A sua área de interesse é a preservação digital.

email: ricardobasilio@fcsh.unl.pt

É permitida a reprodução parcial ou total desta obra, desde que citada a fonte e autoria, proibindo qualquer uso para fins comerciais.
Licença Creative Commons