



associação portuguesa de
bibliotecários, arquivistas e documentalistas

Desafios para os profissionais da informação

Alexandra Lourenço



web archive – porquê?



- Processo de recolha e armazenamento de dados disponibilizados na World Wide Web, garantindo a sua preservação e disponibilização para pesquisas futuras.
- Volatilidade da Web, a maioria da informação deixa de estar disponível passado pouco tempo. Após um ano, apenas 20% dos endereços se mantêm válidos.



A abordagem arquivística tradicional aplica-se à organização e descrição do “arquivo” da web?

Web archive

Propósito da recolha:

- Conservar somente o conteúdo intelectual
- Garantir a autenticidade e integridade
- Preservar o valor legal

Condiciona políticas e procedimentos

Conteúdo
Contexto
Estrutura

Princípio da proveniência
Princípio da ordem original

Web Archive

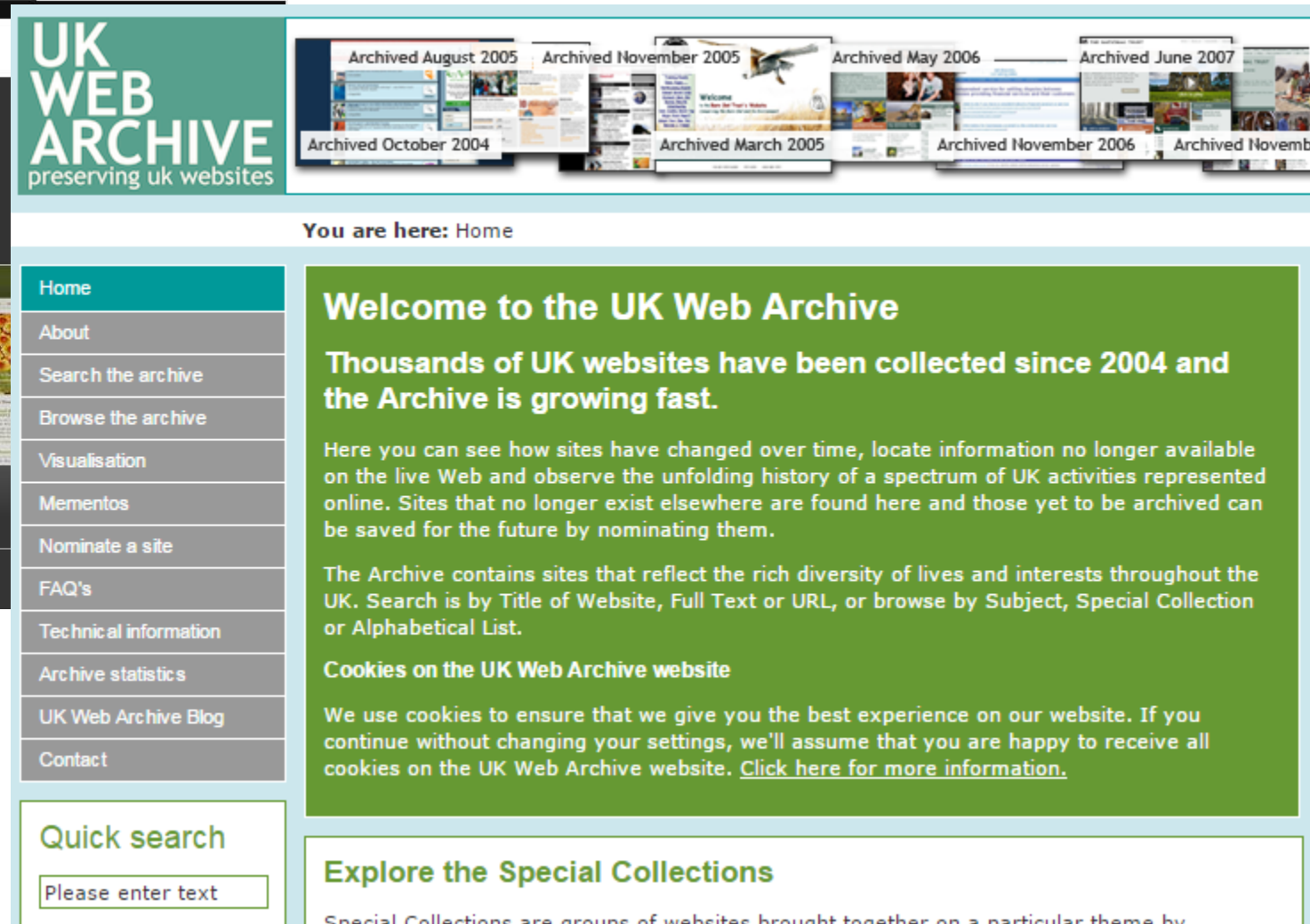


You are here: [Home](#) > [Help with your research](#) > UK Government Web Archive



UK National Archives

UK British Library



associação portuguesa de
bibliotecários, arquivistas e documentalistas

Seleção

Critérios de seleção por a recolha de web sites::

- abordagem global (Internet Archive "Wayback Machine")
- abordagem parcial (Ex.: tipo de mídia (áudio, vídeo ou textual) e domínio (por exemplo, .gov, org, pt).
- “on-demand” - programas e serviços comerciais de arquivamento (webcite)
- Destacar um tema ou evento - necessário julgamento humano (avaliação) e/ou seleção manual
 - Envolver especialistas
 - Envolver a comunidade
- Amostragem

Utilização de
web crawlers

<http://bad@bad.pt>

BROWSE HISTORY

<http://bad@bad.pt>

Saved **27 times** between Junho 22, 2009 and Março 11, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



PLEASE DON

1996 1997 1998 1999 2000

2014-09-24



associação portuguesa de
bibliotecários, arquivistas e documentalistas

© BAD 2004-2014

Quem Somos

Eventos e
Notícias

Profissão

Intervenção
Política e Social

Associados

Cooperação

Formação

Edições

Biblioteca

Bolsa de Emprego

Del. Regionais

Secções

Pesquisar no Site



powered by



Contactos



Enviar uma Mensagem



Mapa do Site

BAD nas Redes Sociais



Partilhamos no Twitter



Conversamos no Facebook



Imagens no Flickr

Declaração
Universal dos
arquivos
aprovada pela
UNESCO >>>

Procedimentos
Concursais >>>

Mensagem do

Eleições para os
Órgãos NACIONAIS
Triénio 2014-2016
Convocatória da
Assembleia Geral >>>
Lista Candidata >>>
Notas Biográficas >>>
Plano de Ação >>>
Lista Associados
Proponentes >>>

Campanha de
Associados
2014/2015
Fique a conhecer as
vantagens de se tornar
associado da BAD! >>>

Cadernos BAD
Já está disponível em
linha o n.º 1 (2014) >>>

Jornada "Gestão,
preservação e acesso à
informação digital" -
GT-GDA
11 setembro 2014, Ponte
de Lima
Programa >>>
Inscrições >>>

Conferência "Advocacy
for Libraries"
16 e 17 setembro 2014,

Prémio Raul
Proença
As candidaturas ao
Prémio Raul
Proença 2013
estão abertas até
31 de Julho de
2014. Consulte o
Regulamento
atualizado e os
procedimentos da
candidatura >>>

Formação
FORMAÇÃO CONTÍNUA
- CALENDÁRIO 2014
>>>
CURSOS DE INTRODUÇÃO 2014
>>>
SEMINÁRIOS 2014
>>>
WORKSHOPS 2014
>>>
WEBINARS 2014
>>>

← → ↻ arquivo.pt/nutchwax/search.jsp?query=http%3A%2F%2Fbad.pt

http://bad.pt

entre:

01/01/1996



e:

31/12/2015



Pretendia ver os resultados das páginas que contêm o texto: <http://bad.pt>?

Versões da página web guardadas no arquivo

Foram gravadas no arquivo 7 versões da página <http://bad.pt> entre 1 Janeiro, 1996 e 6 Abril, 2016.

9	2000 0	2001 0	2002 0	2003 0	2004 0	2005 0	2006 0	2007 0	2008 0	2009 2	2010 0	2011 1	2012 0	2013 2	2014 2	2015 0
										22 Jun		22 Mai		6 Nov	24 Set	
										18 Dez				7 Nov	30 Set	

← → ↻ arquivo.pt/

http://bad.pt

entre: 01/01/1996


Pretendia ver os resu

Versões da página

Foram gravadas no arquivo 7

9	2000	2001	2002	2003	20
0	0	0	0	0	

2014-09-24



associação portuguesa de
bibliotecários, arquivistas e documentalistas

© BAD 2004-2014

Associados

Cooperação

Formação

Edições

Biblioteca

Bolsa de Emprego

Del. Regionais

Secções

Pesquisar no Site

powered by Google

Contatos

Enviar uma Mensagem

Mapa do Site

BAD nas Redes Sociais

Partilhamos no Twitter

Conversamos no Facebook

Imagens no Flickr

Declaração
Universal dos
arquivos
aprovada pela
UNESCO >>>

Procedimentos
Concursais >>>

Mensagem do
CDN sobre a
abertura de
concursos sem
respeito pelos
requisitos

Eleições para os
Órgãos NACIONAIS
Triénio 2014-2016
Convocatória da
Assembleia Geral >>>
Lista Candidata >>>
Notas Biográficas >>>
Plano de Ação >>>
Lista Associados
Proponentes >>>

"a minha biblioteca é
a tua biblioteca"
Programa de
Mobilidade em
Bibliotecas de Ensino

Campanha de
Associados
2014/2015
Fique a conhecer as
vantagens de se tornar
associado da BAD! >>>

Cadernos BAD
Já está disponível em
linha o n.º 1 (2014) >>>

Convite à apresentação
de textos para
publicação.
Brevemente será emitido

Jornada "Gestão,
preservação e acesso à
informação digital" -
GT-GDA
11 setembro 2014, Ponte
de Lima
Programa >>>
Inscrições >>>

Conferência "Advocacy
for Libraries"
16 e 17 setembro 2014,
Lisboa
Programa >>>
Inscrições >>>

11º Encontro Nacional

Prémio Raul
Proença
As candidaturas ao
Prémio Raul
Proença 2013
estão abertas até
31 de Julho de
2014. Consulte o
Regulamento
atualizado e os
procedimentos da
candidatura >>>

"Notícia BAD" é o
jornal em linha dos
profissionais da
Inscrição

Formação
FORMAÇÃO CONTÍNUA
- CALENDÁRIO 2014
>>>
CURSOS DE INTRODUÇÃO 2014
>>>
SEMINÁRIOS 2014
>>>
WORKSHOPS 2014
>>>
WEBINARS 2014
>>>

PROJETO TRANS E-FACILITADOR
30 de Setembro
Apresentação no Porto
Inscrição



associação portuguesa de
bibliotecários, arquivistas e documentalistas

© BAD 2004-2014

http://bad.pt



Insira o Endereço Web:

Tudo ▾

Aliases: ☐

Pesquisar

[Pesquisa Avançada](#)

Recurso não está no Arquivo

O Recurso que requisitou não está no arquivo. Verifique se inseriu o endereço correcto.

[Início](#) | [Ajuda](#)



União Europeia
FEDER

[Site](#)

Declaração
Universal dos
arquivos
aprovada pela
UNESCO >>>

Procedimentos
Concursais >>>

Mensagem do
CDN sobre a
abertura de
concursos sem
respeito pelos
requisitos

Eleições para os
Órgãos NACIONAIS
Triénio 2014-2016
Convocatória da
Assembleia Geral >>>
Lista Candidata >>>
Notas Biográficas >>>
Plano de Ação >>>
Lista Associados
Proponentes >>>

"a minha biblioteca é
a tua biblioteca"
Programa de
Mobilidade em
Bibliotecas de Ensino

Campanha de
Associados
2014/2015
Fique a conhecer as
vantagens de se tornar
associado da BAD! >>>

Cadernos BAD
Já está disponível em
linha o n.º 1 (2014) >>>

Convite à apresentação
de textos para
publicação.
Brevemente será emitido

preservação e acesso à
informação digital" -
GT-GDA
11 setembro 2014, Ponte
de Lima
Programa >>>
Inscrições >>>

Conferência "Advocacy
for Libraries"
16 e 17 setembro 2014,
Lisboa
Programa >>>
Inscrições >>>

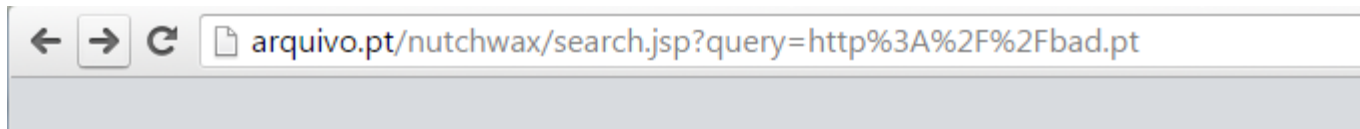
11º Encontro Nacional

Prémio Raul
Proença
As candidaturas ao
Prémio Raul
Proença 2013
estão abertas até
31 de Julho de
2014. Consulte o
Regulamento
atualizado e os
procedimentos da
candidatura >>>

"Notícia BAD" é o
jornal em linha dos
profissionais da
Associação

Formação
FORMAÇÃO CONTÍNUA
- CALENDÁRIO 2014
>>>
CURSOS DE INTRODUÇÃO 2014
>>>
SEMINÁRIOS 2014
>>>
WORKSHOPS 2014
>>>
WEBINARS 2014
>>>

PROJETO TRANS E-FACILITATOR
30 de Setembro
Apresentação no Porto
Inscrição



http://bad.pt

entre: e:


Pretendia ver os resultados das páginas q

Versões da página web g

Foram gravadas no arquivo 7 versões da página

9	2000	2001	2002	2003	2004	2005	2006
	0	0	0	0	0	0	0

arquivo.pt/wayback/20090622183043/http://www.bad.pt/

 **Arquivo da Web Portuguesa** - ligações exteriores, formulários e caixas de pesquisa poderão não funcionar corretamente. URL: <http://www.bad.pt/> Data: 18:30:43 22 Junho, 2009 [[esconder](#)]

Este é o sítio: **bad.pt**

Este domínio foi atribuído à empresa titular, no âmbito da **iniciativa Empresa na Hora** do Ministério da Justiça em colaboração com a **FCCN - Fundação para a Computação Científica Nacional**.



Para a gestão DNS do seu domínio vá a:

<https://online.dns.pt/site/publico>

(Utilize o login e password que recebeu por carta ou virá a receber)



Você está aqui: [Entrada](#) → [Notícias](#) → Recolhemos os sites das Eleições Presidenciais de 2016!

Recolhemos os sites das Eleições Presidenciais de 2016!

O Arquivo.pt realizou 2 recolhas de sites relacionados com as Eleições Presidenciais de 2016.

Apelámos à comunidade para contribuir sugerindo sites relacionados com as eleições presidenciais de 2016, para que os arquivássemos.

Realizaram-se 2 recolhas, durante e após o período de campanha eleitoral, que partiram da lista de 284 sites sugeridos pela comunidade, resultando no arquivo de 551 672 ficheiros recolhidos da Web, que ocupam 7 GB.

Recolhemos sites tais como os oficiais dos candidatos, sites dos partidos dos candidatos, notícias na comunicação social, blogs e artigos de opinião.

O Arquivo.pt respeita um [período de embargo de 1 ano](#), pelo que esta coleção só estará disponível para consulta no final de 2016.

Contudo, pode consultar já alguns sites das Eleições Presidenciais anteriores, tais como:

- [Presidenciais de 1996](#)
- [Presidenciais de 2001](#)
- [Presidenciais de 2006](#)

Agradecemos sinceramente a todos os voluntários que colaboraram com esta iniciativa.

CANDIDATOS



GARCIA PEREIRA > [garciapereira-presidenciais2006.net](#)



CAVACO SILVA > [www...](#)



FRANCISCO LOUÇÃ > [www...](#)



MANUEL ALEGRE > [www...](#)



JERÓNIMO DE SOUSA > [www...](#)



MÁRIO SOARES > [www...](#)

← → ↺ arquivo.pt/wayback/20091219134021/http://www.garciapereira-presidenciais2006.net/frame.aspx?u=http%3a



Insira o Endereço Web:

Tudo ▾

Aliases: ☐

[Pesquisa Avançada](#)

Recurso não está no Arquivo

O Recurso que requisitou não está no arquivo. Verifique se inseriu o endereço correcto.

[Início](#) | [Ajuda](#)

Aquisição

Competência legal para recolha e preservação

- Depósito legal

Período de embargo / solicitação de
autorização / solicitação de exclusão

(Ex.: National Library of New Zealand,
UK National Archives)

Métodos de aquisição



- Envio regular ou a pedido de **snapshot** para entidades de preservação patrimonial
- Utilização de **web crawlers**, programas que realizam o processo de recolha automática de acordo com critérios de seleção

Dependentes de:

- escala da recolha
- relação entre o arquivo da web e os proprietários do site
- natureza do conteúdo web



NATIONAL ARCHIVES

[Blogs](#) | [Bookmark/Share](#) | [Contact Us](#)

GO

[Research Our Records](#)

[Veterans Service Records](#)

[Teachers' Resources](#)

[Our Locations](#)

[Shop Online](#)

Records Managers

[Home](#) > [Records Management](#) > [Basics](#) > [Memorandum to Chief Information Officers:](#)

Records Management Resources

[E-mail Management](#)

[Records Management FAQs](#)

[Memorandums to Agency Records Officers](#)

[Federal Records Centers \(FRC\)](#)

[Guidance and Policy for Accessioning](#)

[Electronic Records Management \(ERM\)](#)

[Toolkit for Managing Electronic Records](#)

Memorandum to Chief Information Officers: Snapshot of Agency Public Web Sites

The National Archives and Records Administration (NARA) will be preserving a one-time snapshot of agency public web sites as they exist on or before January 20, 2001, as an archival record in the National Archives of the United States. The web site snapshot initiative outlined in this memorandum will ensure that we are able to document at least in part agency use of the Internet at the end of the Clinton Administration. To accomplish this, we ask that agencies take the following actions:

- Take a snapshot of your agency's public web site(s) following the instructions in the attachment to this memorandum.
- Within 60 days, send the snapshot and related documentation, also described in the attachment, to NARA.

Aquisição:



Limitações da recolha automática:

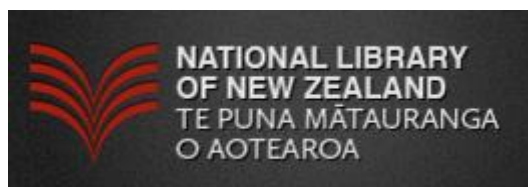
- Protocolos de exclusão
- Acesso à *Deep Web*
- Conteúdos dinâmicos (Links, relações entre páginas e sites)
- Elementos complexos (video, audio, texto)
- Bases de dados



Aquisição:

Propriedade intelectual

- Período de embargo / solicitação de autorização / solicitação de exclusão



Archived Websites Frequently Asked Questions

Q13: How does copyright law affect the archiving of websites?

A: Copyright law and Legal Deposit legislation are different. Legal Deposit legislation allows the National Library of New Zealand to make a copy of the publicly available pages of a New Zealand website for preservation purposes and to provide access to the archived copy of the website. However copyright is still retained by the copyright holder. Anyone wanting to copy or download pages of the archived website must comply with copyright law.

Organização e descrição



- Por sessão
- Estrutura hierárquica multinível
- Abordagem bibliográfica



WEB ARCHIVING

[Print](#) [Subscribe](#) [Share/Save](#)

- [Web Archiving Home](#)
- [About the Program](#)
- [Cataloging](#)
- [Collections in Progress](#)
- [FAQs](#)
- [Contact](#)

Resources

- > [Digital Preservation](#)
- > [Library of Congress Web Archives](#)

Cataloging

Archived Web sites were cataloged using the Metadata Object Description Schema (MODS). Preliminary keyword, title, and subject metadata were extracted from the archived Web sites to create preliminary MODS records that were subsequently reviewed and/or enhanced by catalogers who assigned controlled subjects from Library of Congress Subject Headings (LCSH) or Thesaurus of Graphic Materials (TGM). A Lucene search interface was developed to search the MODS records both within and across the archived collections.

Collection-level:

In addition, a MARC record for each **collection** is available in the Library of Congress Online Catalog so that the collection can be found along with other Library materials in the catalog.

Metadata included in **collection**-level records in Library of Congress Online Catalog:

245 \$a Collection title \$h [electronic resource].
520 \$a General description of the collection content and number of Web sites and date range when Web sites were captured
6XX \$a Collection-level subject heading (usually several 6XX fields)
856 \$a <http://hdl.loc.gov/loc.natlib/collnatlib.12345678>

Web site level:

MODS data included in record for each archived Web site:

TITLE INFO

<titleInfo><title> - Title extracted by system from HTML title tag (when available) and reviewed by cataloger, otherwise supplied by cataloger

<titleInfo type="alternative"><title> - Alternative Title supplied by cataloger if different and useful.

[Department for Business, Enterprise and Regulatory Reform \(BERR\)](#)

[Department for Business, Enterprise and Regulatory Reform \(BERR\) - Automotive Sector](#)

[Department for Business, Enterprise and Regulatory Reform \(BERR\) - Consumer Issues](#)

[Department for Business, Enterprise and Regulatory Reform \(BERR\) - Energy and Climate Security Panel](#)

[Department for Business, Enterprise and Regulatory Reform \(BERR\) - Statistics and Analysis](#)

[Department for Business, Enterprise and Regulatory Reform \(BERR\) - YouTube Archive](#)

[Department for Business, Innovation and Skills \(BIS\) \(<http://www.bis.gov.uk/>\)](#)

[Department for Business, Innovation and Skills \(BIS\) \(<http://old.berr.gov.uk/>\)](#)

[Department for Business, Innovation and Skills \(BIS\) - 2009 Innovation Value Added Scoreboard](#)

[Department for Business, Innovation and Skills \(BIS\) - About BIS \(<http://www.berr.gov.uk/about>\)](#)

[Department for Business, Innovation and Skills \(BIS\) - About BIS \(<http://www.dius.gov.uk/about>\)](#)

[Department for Business, Innovation and Skills \(BIS\) - Backing Small Business](#)

[Department for Business, Innovation and Skills \(BIS\) - Blogs](#)

[Department for Business, Innovation and Skills \(BIS\) - Budget 2012](#)

[Department for Business, Innovation and Skills \(BIS\) - Consultation Hub](#)

[Department for Business, Innovation and Skills \(BIS\) - Economics and Statistics \(<http://alookup.bis.gov.uk/>\)](#)

[Department for Business, Innovation and Skills \(BIS\) - Economics and Statistics \(<http://stats.bis.gov.uk/>\)](#)

[Department for Business, Innovation and Skills \(BIS\) - GOV.UK](#)

[Department for Business, Innovation and Skills \(BIS\) - Impact Assessment Library](#)

[Department for Business, Innovation and Skills \(BIS\) - Innovation - Self Assessment](#)

Descrição e metadata



Criação de metadata:

- Geração automática – capturada pelos crawlers (timestamp, tamanho em bytes, o URI, o tipo MIME, tags das páginas HTML)
- Criação manual (apenas para pequena escala)
- Combinação de captura automática com a manual



[Discover](#)[Services](#)[Visit](#)[Education](#)[Connect](#)[About](#)

All Formats ▼

Search Loc.gov

GO

[Print](#) [Subscribe](#) [Share/Save](#) [Give Feedback](#)[Archived Web Sites](#)

Official Campaign Web Site - Arvin Vohra



Captured from June 21, 2012 to December 4, 2014

About this Item

Title

Official Campaign Web Site - Arvin Vohra

Contributor Names

Vohra, Arvin

Subject Headings

- Vohra, Arvin
- Political candidates -- United States
- Elections -- United States
- Politics and government -- United States
- United States Elections, 2012

Part of...

- [United States Elections 2014 \(1,224\)](#)
- [United States Elections 2012 \(1,246\)](#)
- [United States Elections Web Archive \(5,407\)](#)
- [Humanities and Social Sciences Division \(9,670\)](#)

Format[Archived Web Sites](#)

Acesso



- Acessibilidade depende de permissões legais:.
 - Depósito legal - acesso à cópia arquivada de todos os website disponíveis (Ex.: [Biblioteca Nacional da Nova Zelândia](#)).
 - Acesso do público às páginas cujos produtores tenham dado permissão (Ex.: [Biblioteca do Congresso](#))
 - Acesso local (Ex.: [Biblioteca Nacional da França](#))
 - Período de embargo (Ex.: [FCCN – arquivo.pt](#))



Who can access the archives?

The Web archives are accessible to authorized users of the BnF, in the reading rooms of the Research Library only. This restriction is the same as that which applies to all legal deposit collections. As of June 22th, 2009, the BnF offers 350 computers to consult its Web archives across all its sites, in Paris and in Avignon.

Acesso



- Capacidade de pesquisa depende dos metadados extraídos e das ferramentas de pesquisa utilizadas:.
 - Controlo de autoridades (Biblioteca do Congresso)
 - Pesquisa por URL (Internet Archive)
 - Pesquisa em texto livre (os que utilizam o motor de pesquisa NutchWax)



Desafios



- Repetibilidade de experiência de navegação atual
- Continuidade semântica e ontológica
- Preservação de web sites autênticos
 - Certificação de autenticidade e integridade
 - Garantia no longo prazo de autenticidade, integridade, usabilidade



Desafios para as entidades produtoras

- Pensar a preservação na conceção
- Assegurar a autenticidade no longo prazo
- Avaliar (Tabelas de seleção para websites)

Ver recomendações do NARA <http://www.archives.gov/records-mgmt/pdf/managing-web-records-index.pdf>

Ver recomendações do arquivo.pt
<http://sobre.arquivo.pt/colabore/recomendacoes-para-autores-de-sitios-web/recomendacoes-para-autores-de-sitios-da-web>

Desafios



Formação dos profissionais da informação

Projetos colaborativos – Envolver a comunidade

