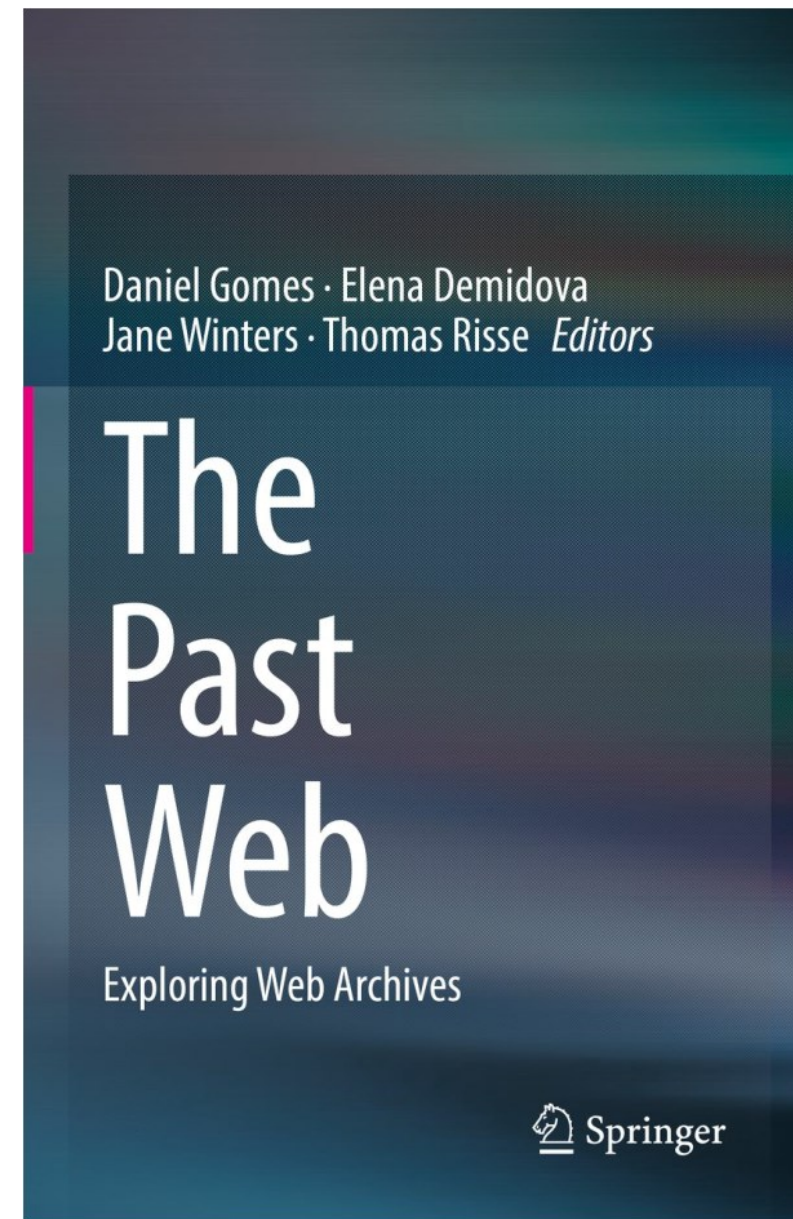


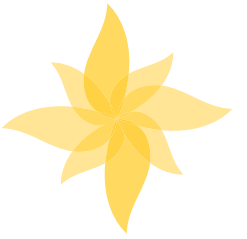
# The Past Web: Exploring Web Archives

## Como nasce um livro?

[Daniel.Gomes@fccn.pt](mailto:Daniel.Gomes@fccn.pt)



# Web archiving workshop 2003



## 3rd ECDL Workshop on Web Archives

August 21<sup>st</sup>, 2003  
Trondheim, Norway

in conjunction with the [7<sup>th</sup> European Conference on Research and Advanced Technologies for Digital Libraries](#)



### Objectives

Following the great success of the first two ECDL Workshops on Web Archiving in Darmstadt, Germany in 2001 ([WebArchiving 2001](#)), and Rome, Italy, in 2002 ([WebArchiving2002](#)), we are happy to invite you to the third Workshop in this series.

The workshop will provide a cross domain overview on active research and practice in the emergent domain of web archiving and studies on effective usage of this type of archives.

It is also intended to provide a forum for interaction among librarians, archivists, academic researchers and industrial researchers interested in establishing effective methods and developing improved solution for Web archiving.

### Workshop Proceedings

Download [full Workshop Proceedings as a PDF file](#) (3,5 Mo).  
See below for individual papers and presentations.

### Afternoon session: 14:00 - 18:00 // Case Studies //

#### **A Characterization of the Portuguese Web** (download [paper](#))

*Daniel Gomes, Mário J. Silva*  
Faculty of Sciences, University of Lisbon, Portugal

#### **Building Thematic Web Collections: Challenges and Experiences from the September 11 Web Archive and the Election 2002 Web Archive** (download [paper](#))

*Steven M. Schneider*, SUNY Institute of Technology & WebArchivist.org, USA  
*Kirsten Foot*, Department of Communication, University of Washington & WebArchivist.org, USA  
*Michele Kimpton*, Internet Archive, USA  
*Gina Jones*, Library Services, Library of Congress, USA

#### **Political Communications Web Archiving: Addressing Typology and Timing for Selection, Preservation and Access** (download [paper](#))

*Bernard Reilly, Gretchen Tuchel, James Simon*, Center for Research Libraries, USA  
*Carolyn Palaima, Kent Norsworthy*, Latin American Network Information Center - LANIC, USA  
*Leslie Myrick*, Digital Library Group, New York University Libraries, USA

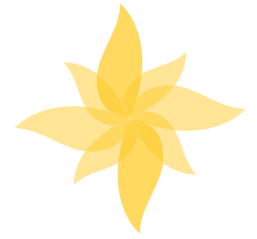
#### **Learning by Doing: the Digital Archive for Chinese Studies (DACHS)** (download [paper](#))

*Jennifer Gross*  
University of Heidelberg, Germany


#### **Archiving the Czech Web: Issues and Challenges** (download [paper](#))

*Petr Žabicka*  
Moravian Library in Brno, Czech Republic

# Livro “Web Archiving”, Springer 2006



SpringerLink



© 2006  
**Web Archiving**  
Authors ([view affiliations](#))  
Julien Masanès

Combines the librarian's application knowledge with the computer scientist's implementation knowledge  
Introduces all aspects from website monitoring to deep Web preservation  
Presents an unbiased view on current standardization and preservation projects

Book | 96 Citations | 11k Downloads

[Download book PDF](#)

[Table of contents \(10 ch...](#) | [About this book](#) | [Reviews](#)

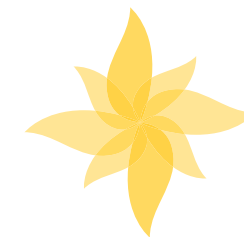
## Introduction

The public information available on the Web today is larger than information distributed on any other media. The raw nature of Web content,



Julien Masanès: o pai dos arquivos da web na Europa

# Arquivo.pt, 2007



Menu

ARQUIVO.PT

arquivo-web.fccn.pt 16 Março às 04:49, 2008

Opções

Mapa do sítio Acessibilidade Contacto

apenas na secção corrente

Entrada Como participar Fui arquivado? Funcionamento Ligações Perguntas frequentes Sobre o Arquivo

Você está aqui: Entrada English Português

## Arquivo da Web Portuguesa

**O Arquivo da Web Portuguesa é um projecto da Fundação para a Computação Científica Nacional que tem como principal objectivo a preservação da informação publicada na web de Portugal.**

Este projecto da [Fundação para a Computação Científica Nacional \(FCCN\)](#) visa a criação de um sistema de arquivo de conteúdos da web portuguesa, que terá como missão recolher periodicamente, armazenar e preservar a informação publicada.

A primeira fase do desenvolvimento do Arquivo teve início em Janeiro de 2008 e prevê-se que termine no prazo de 2 anos. Contudo, a manutenção de um sistema desta natureza e a preservação da informação arquivada é uma tarefa que deverá ser perpetuada posteriormente.

A Web possibilita que cada um de nós disponibilize informação e que esteja se torne acessível a todos, sem necessidade de recurso a editoras e meios de impressão tradicionais. Diariamente, são publicados milhões de conteúdos na web como por exemplo, textos, fotografias ou vídeos. A quantidade de informação que é publicada exclusivamente na web tem vindo a aumentar rapidamente nos últimos anos. No entanto, passado relativamente pouco tempo, a grande maioria desta informação deixa de estar acessível *online* e perde-se irremediavelmente.

Surge assim o interesse no arquivo e preservação da informação publicada na Web para que o conhecimento nela contido esteja acessível às gerações futuras.

### O que é a Web portuguesa?

Entende-se por Web portuguesa, todos os conteúdos alojados sob o domínio .pt.

Numa primeira fase, pretende-se arquivar apenas conteúdos alojados sob este domínio nacional, embora posteriormente se possam vir a abranger todas as páginas escritas em língua portuguesa.

### Para que serve o Arquivo da Web?

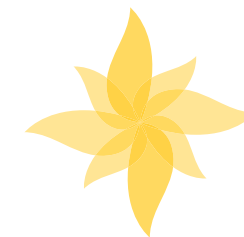
Os serviços a serem prestados pelo Arquivo da Web Portuguesa ultrapassam o âmbito histórico-cultural da preservação de informação digital. A existência de um Arquivo da Web de Portugal poderá:

- Contribuir para a expansão do uso do português enquanto língua para comunicação na Web;
- Disponibilizar conteúdos de interesse às diversas comunidades científicas, por exemplo, na área da História, Sociologia ou Processamento Computacional da Língua Portuguesa;
- Contribuir para o desenvolvimento da capacidade local de tratamento e prospecção de informação publicada na Web, reduzindo a dependência nacional de serviços estrangeiros;
- Fornecer provas em casos judiciais que tenham como base informação publicada na Web.

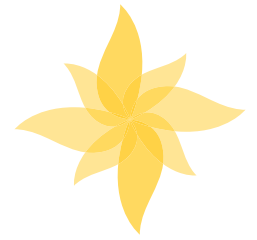
### O Arquivo da Web Portuguesa e os outros arquivos da web

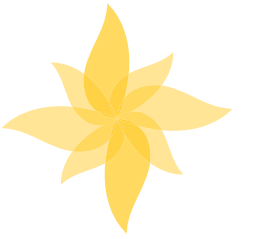


# TPDL 2018@Porto: Tutorial Research the Past Web using Web Archives



# Escrevendo “The Past Web”



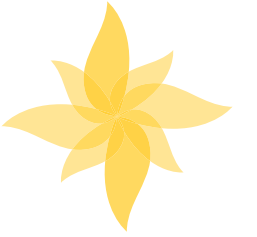


# Objectivos do livro

- Alertar consciências para a **importância de preservar** a informação publicada na Web
- **Novo recurso pedagógico** actualizado acerca do estado da arte em arquivos da web
- Dar a conhecer os **serviços disponibilizados** pelos arquivos da web
- Divulgar **trabalhos de investigação** inspiradores que exploraram a Web do passado



# Os Editores



Elena Demidova



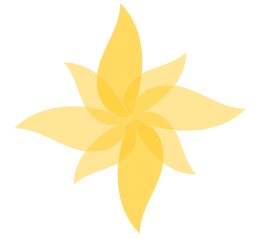
Jane Winters



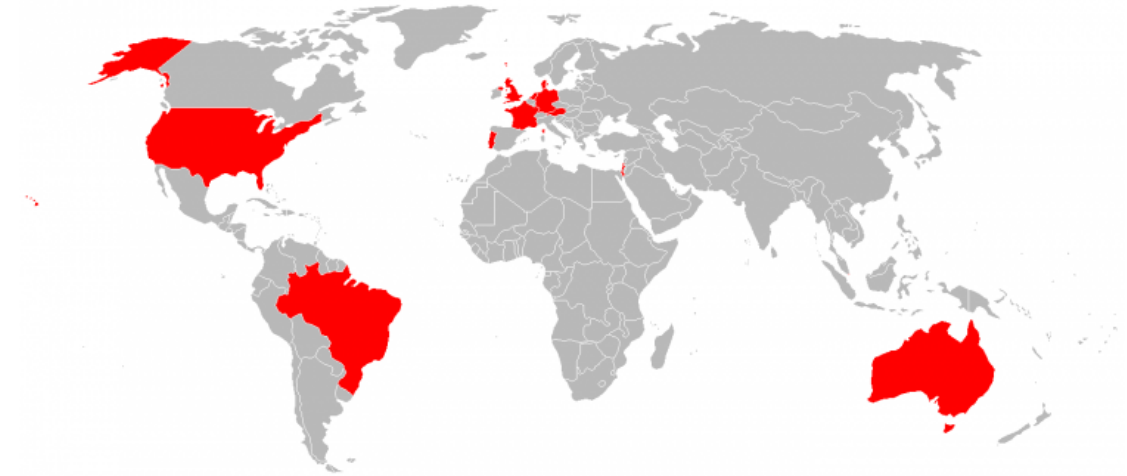
Thomas Risse

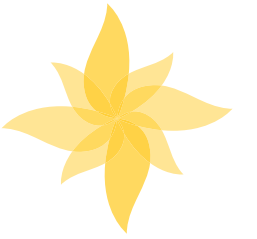


# 40 autores de 12 países do mundo



Peter Webster  
Shawn M. Jones  
Ivy Huey Shin Lee  
Saskia Huc-Hepher  
Elena Demidova  
Jane Winters  
Paul Koerbin  
Martin Klein  
Ilya Kreymer  
Miguel Won  
Alípio Mário Jorge  
Adam Jatowt  
Naomi Wells  
Dragan Espenschied  
Herbert Van de Sompel  
Shereen Tay  
Michele C. Weigle  
Sebastian Diering  
Matthias Springstein  
Janne Nielsen  
Ralph Ewerth  
Fernando van der Vliet  
Eric Müller-Budack  
Vítor Mangaravite  
Anne Helmond  
Anat Ben-David  
Thomas Dugeon  
Kader Pustu-Iren  
Miguel Costa  
Ricardo Campos  
Helge Holzmann  
Niels Brügger  
Thomas Risse  
Daniel Gomes  
Arian Pasquali  
Zeynep Pehlivan  
Daniela Major  
Michael L. Nelson

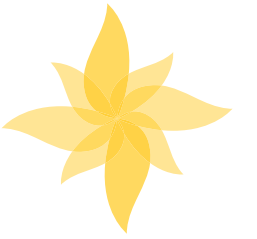




# A quem se destina?

- **Professores**  
para ensinarem acerca de preservação da web
- **Investigadores**  
para estudarem o passado através da Web
- **Informáticos**  
para criarem novas aplicações para explorar o passado
- **Profissionais da informação**  
para preservarem também a informação online
- **Cidadãos**  
que utilizam a Internet (quem não?)

# Parte 1: Porquê preservar a web?



## The Era of Information Abundance and Memory Scarcity

---

Front Matter

Pages 1-3

---

[The Problem of Web Ephemera](#)

Daniela Major

Pages 5-10

---

[Web Archives Preserve Our Digital Collective Memory](#)

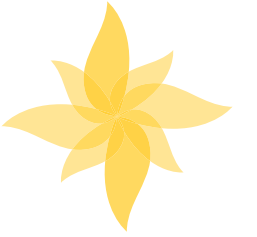
Daniela Major, Daniel Gomes

Pages 11-19



# Parte 2:

## Como obter a informação da web?



### Collecting before it vanishes

Front Matter

Pages 21-22

[National Web Archiving in Australia: Representing the Comprehensive](#)

Paul Koerbin

Pages 23-32

[Web Archiving in Singapore: The Realities of National Web Archiving](#)

Ivy Huey Shin Lee, Shereen Tay

Pages 33-42

[Archiving Social Media: The Case of Twitter](#)

Zeynep Pehlivan, Jérôme Thièvre, Thomas Durgeon

Pages 43-56

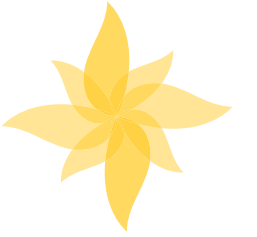
[Creating Event-Centric Collections from Web Archives](#)

Elena Demidova, Thomas Risse

Pages 57-67

# Parte 3:

## Como aceder à informação arquivada?



### Access methods to analyse the Past web

#### Front Matter

Pages 69-70

#### Full-Text and URL Search Over Web Archives

Miguel Costa

Pages 71-84

#### A Holistic View on Web Archives

Helge Holzmann, Wolfgang Nejdl

Pages 85-99

#### Interoperability for Accessing Versions of Web Resources with the Memento Protocol

Shawn M. Jones, Martin Klein, Herbert Van de Sompel, Michael L. Nelson, Michele C. Weigle

Pages 101-126

#### Linking Twitter Archives with Television Archives

Zeynep Pehlivan

Pages 127-139

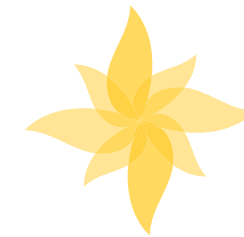
#### Image Analytics in Web Archives

Eric Müller-Budack, Kader Pustu-Iren, Sebastian Diering, Matthias Springstein, Ralph Ewerth

Pages 141-151

# Parte 4:

## Como investigar o passado da web?



### Researching the Past Web

Front Matter

Pages 153-154

[Digital Archaeology in the Web of Links: Reconstructing a Late-1990s Web Sphere](#)

Peter Webster

Pages 155-164

[Quantitative Approaches to the Danish Web Archive](#)

Janne Nielsen

Pages 165-179

[Critical Web Archive Research](#)

Anat Ben-David

Pages 181-188

[Exploring Online Diasporas: London's French and Latin American Communities in the UK Web Archive](#)

Saskia Huc-Hepher, Naomi Wells

Pages 189-201

[Platform and App Histories: Assessing Source Availability in Web Archives and App Repositories](#)

Anne Helmond, Fernando van der Vlist

Pages 203-214



# Parte 5:

## Como tornar os arquivos da web em infraestruturas de investigação?

---

### Web Archives as Infrastructures to Develop Innovative Services

---

Front Matter

Pages 215-216

---

#### The Need for Research Infrastructures for the Study of Web Archives

Niels Brügger

Pages 217-224

---

#### Automatic Generation of Timelines for Past-Web Events

Ricardo Campos, Arian Pasquali, Adam Jatowt, Vítor Mangaravite, Alípio Mário Jorge

Pages 225-242

---

#### Political Opinions on the Past Web

Miguel Won

Pages 243-252

---

#### Oldweb.today: Browsing the Past Web with Browsers from the Past

Dragan Espenschied, Ilya Kreymer

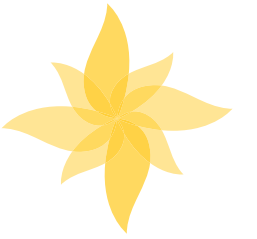
Pages 253-269

---

#### Big Data Science Over the Past Web

Miguel Costa, Julien Masanès

Pages 271-282



# Parte 6: O futuro dos arquivos da web

## A Look into the Future

Front Matter

Pages 283-283

[The Past Web: A Look into the Future](#)

[Julien Masanès](#), Daniela Major, Daniel Gomes

Pages 285-291

# Arquivo.pt preservou os links citados

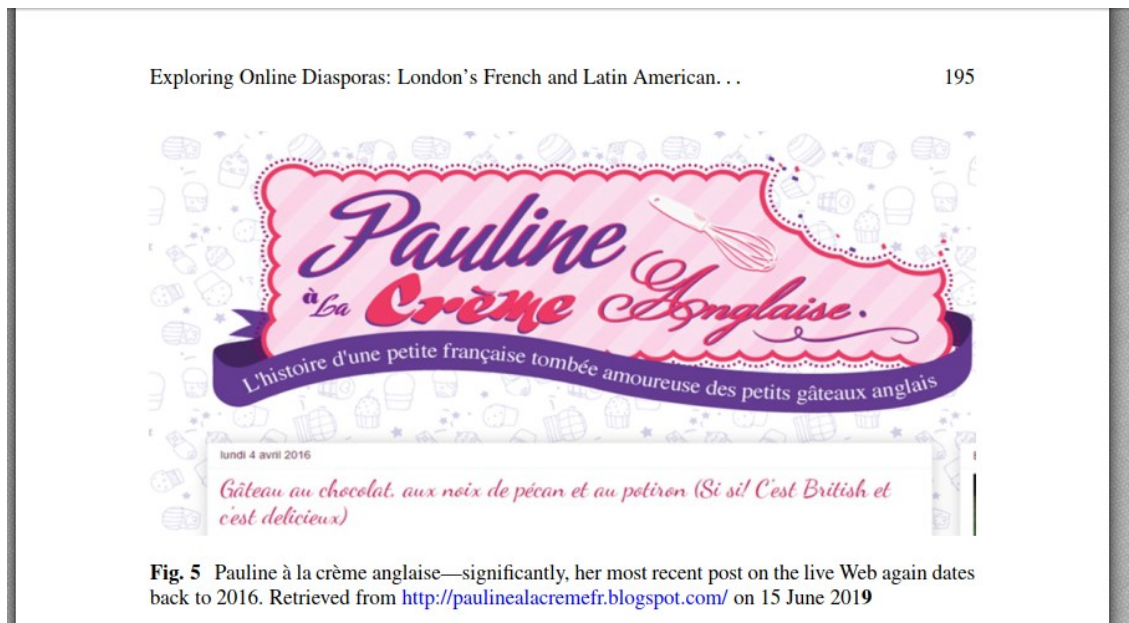
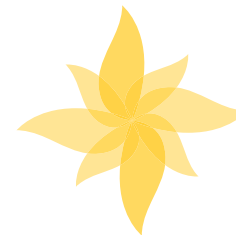
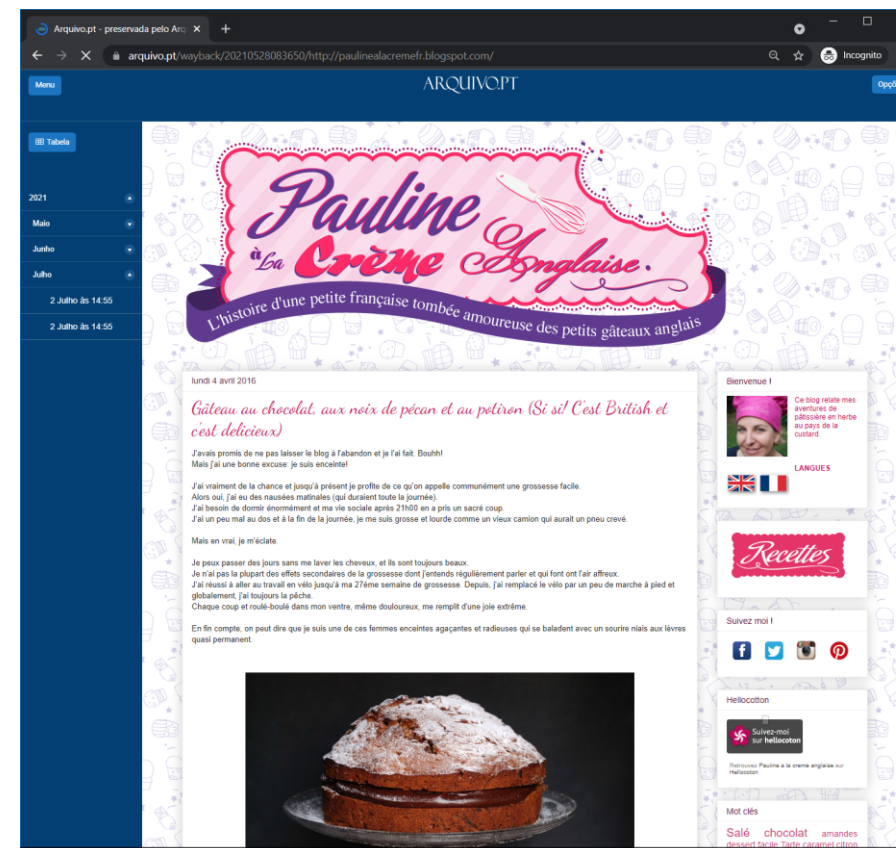


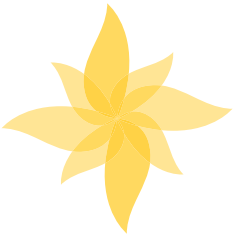
Imagem de um excerto da página da Web



Página da Web completa e navegável



# Algumas citações do livro já não estão disponíveis na web actual!



214 A. Helmond and F. van der Vlist

Helmond A, Nieborg DB, van der Vlist FN (2019) Facebook's evolution: development of a platform-as-infrastructure. *Internet Hist* 3(2):123–146. <https://doi.org/10.1080/24701475.2019.1593667>

Helmond A, van der Vlist FN (2019) Social media and platform historiography: challenges and opportunities. *TMG – J Media Hist* 22(1):6–34. <https://www.tmgonline.nl/articles/434/>

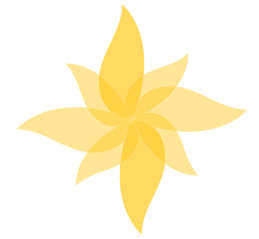
Kirschenbaum MG (2003) *Virtuality and VRML: software studies after Manovich*. *Electronic book review*. Retrieved from: <http://www.electronicbookreview.com/thread/technocapitalism/morememory>

A screenshot of a web browser showing a 404 error page. The page title is "Not found, error 404". The main text says: "The page you are looking for no longer exists. Perhaps you can return back to the homepage and see if you can find what you are looking for. Or, you can try finding it by using the search form". The website header includes "electronic book review" and navigation links like "about ebr", "policies and submissions", "subscribe", "essays", "gatherings", "newsletter", and "log in".


Mas estão disponíveis na Web do passado preservada pelo Arquivo.pt

A screenshot of the archived version of the page from Arquivo.pt. The page title is "Virtuality and VRML: Software Studies After Manovich" by Matthew G. Kirschenbaum. The page content includes a paragraph starting with "In 1999 I was asked to join a panel on 'virtuality' convened by the artist and media scholar Johanna Drucker at the International Digital Arts and Culture 2000 conference in Bergen, Norway." and another paragraph starting with "By digital objects I mean tangible hardware devices such as processors, VDT screens, and Palm Pilots, but also, and especially, intangible software objects such as source code, operating systems, interface elements, and data representations of all kinds." The page is dated "10 Outubro às 05:00" and includes a "Tabela" (Table of Contents) on the left side.

# Resultados iniciais animadores



“The Past Web”:  
8 400 downloads desde julho 2021



© 2021  
**The Past Web**  
Exploring Web Archives

Editors ([view affiliations](#))  
Daniel Gomes, Elena Demidova, Jane Winters, Thomas Risse

Provides practical information about web archives, offers inspiring examples for web archivists and shares recent research results about access methods for exploring preserved information

Targets academics and advanced professionals in digital humanities, social sciences, history, media studies, and information or computer science

Serves as an initial reference for students in various areas of knowledge by introducing how to explore online history through web archives

Book | 8.4k Downloads



“Web archiving”:  
11 000 downloads desde 2006



© 2006  
**Web Archiving**

Authors ([view affiliations](#))  
Julien Masanés

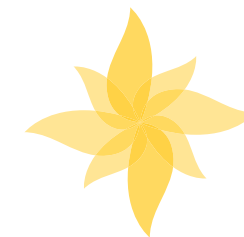
Combines the librarian's application knowledge with the computer scientist's implementation knowledge

Introduces all aspects from website monitoring to deep Web preservation

Presents an unbiased view on current standardization and preservation projects

Book | 96 Citations | 11k Downloads

Leiam e divulguem pelas vossas comunidades!



Descarregue já antes que esgote!!!

[arquivo.pt/livro](http://arquivo.pt/livro)

