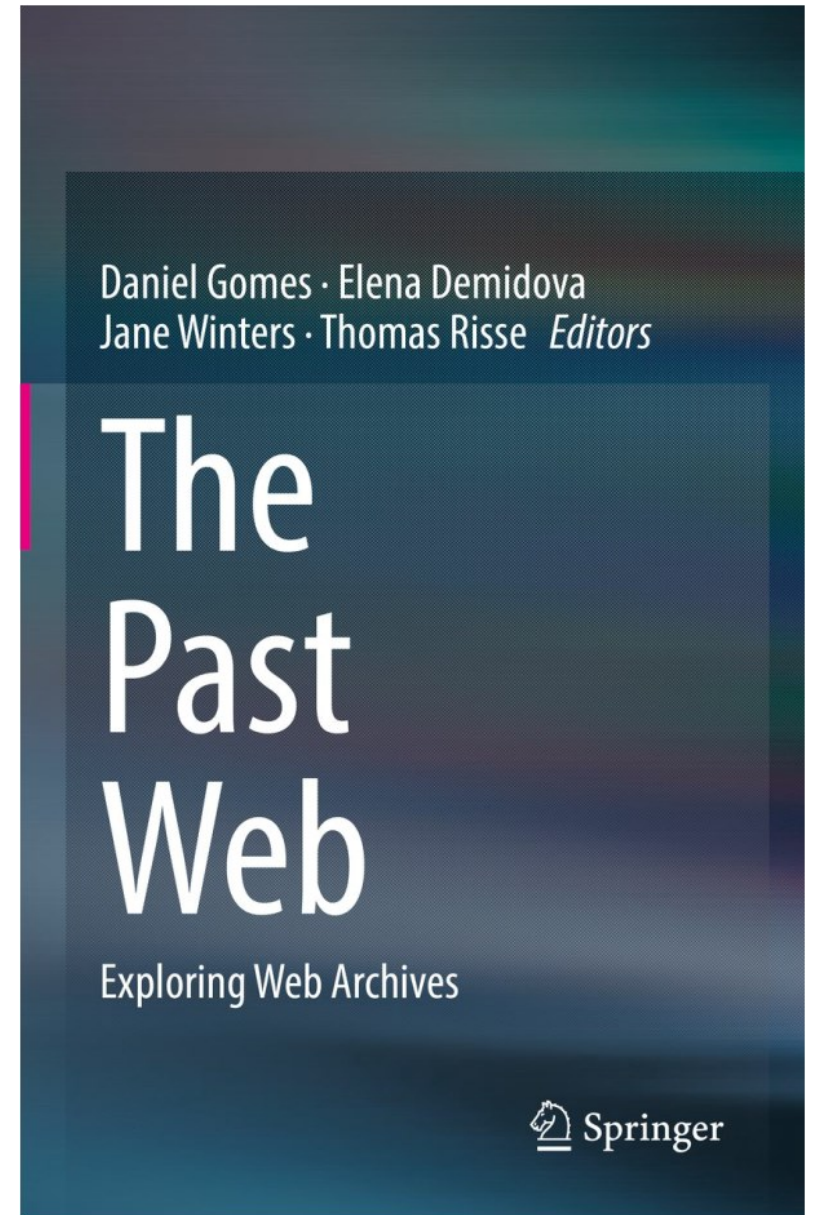


# The Past Web a book to support web archiving

[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

# Book “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	11k
Citations	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:  
the father of web archiving in Europe

# Arquivo.pt, 2007

Menu

Tabela

2008

Março

16 Março às 04:52

ARQUIVO.PT

arquivo-web.fccn.pt/portuguese-web-archive?set\_language=en16 Março às 04:52, 2008

Opções

Site MapAccessibilityContact

Search

☐ only in current section

HomeCrawlerTeam

You are here: HomeEnglish Português

Portuguese Web Archive

Welcome to the Tomba project: the Portuguese web archive

Publishing tools, such as Blogger, enabled people with limited technical skills to become web publishers. Never before in the history of mankind so much information was published. However, it was never so ephemeral. Web documents such as news, blogs or discussion forums are valuable descriptions of our times, but most of them will not last longer than one year.

If we do not archive the current web contents, the future generations could witness an information gap in our days.

The [Internet Archive](#) collects and stores contents from the world-wide web. However, it is difficult for a single organization to archive the web exhaustively while satisfying all needs, because the web is permanently changing and many contents disappear before they can be archived.

As a result, several countries are creating their own national archives to ensure the preservation of contents of historical relevance to their cultures.

Portugal is now beginning its national web archiving initiative with the Tomba project at [FCCN](#) (National Foundation for Scientific Computing).

Send this — Print this —



Fundação para a Computação Científica Nacional



UMIC  
Agência para a Sociedade do Conhecimento

The Plone® CMS — Open Source Content Management System is © 2000-2008 by the Plone Foundation et al.  
Plone® and the Plone logo are registered trademarks of the Plone Foundation. Distributed under the GNU GPL license.

Powered by PloneValid XHTMLValid CSSSection 508WCAG



# TPDL 2018@Porto:

## Tutorial Research the Past Web using Web Archives



# The Editors



Elena Demidova



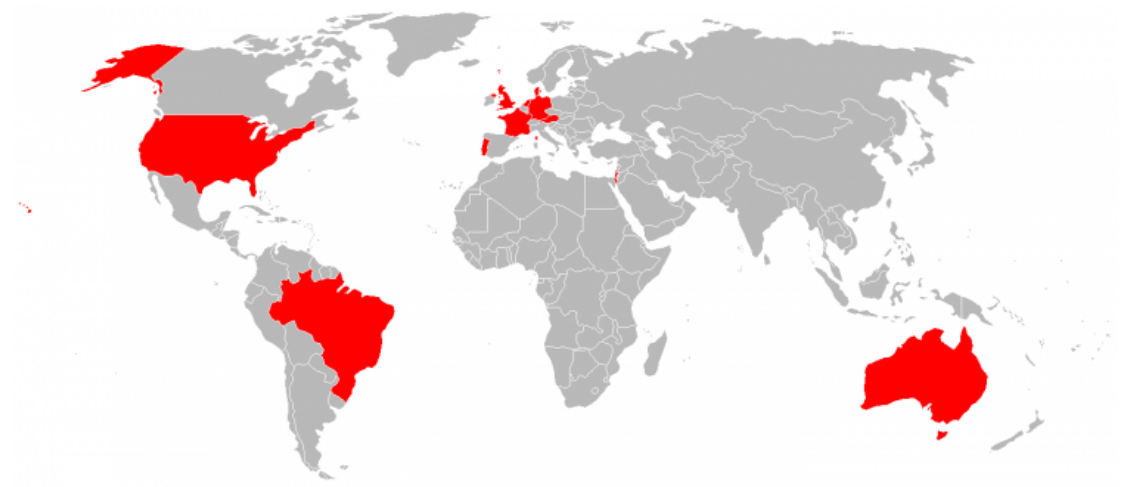
Jane Winters



Thomas Risse

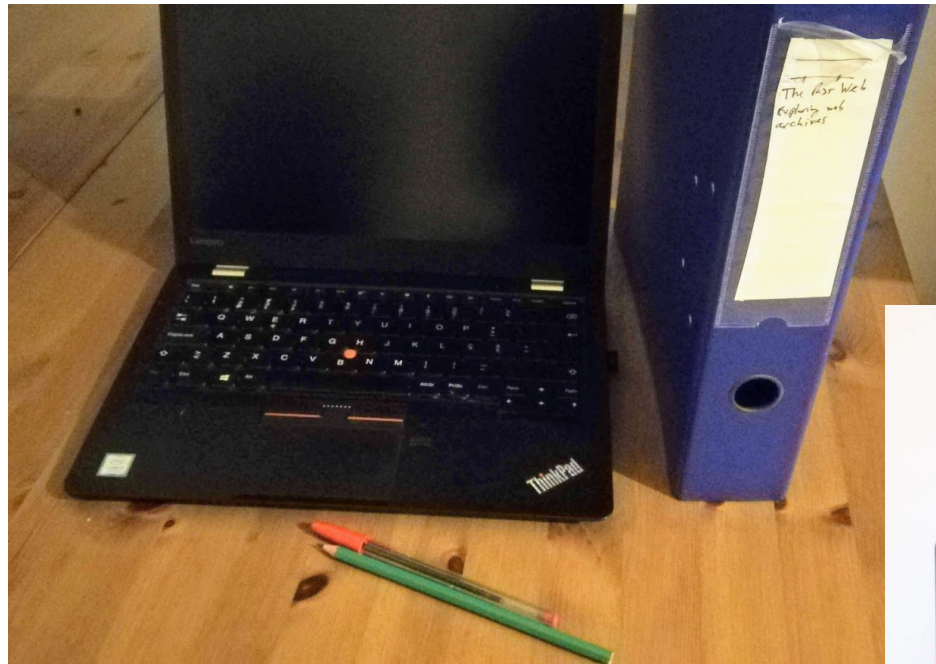
# 40 authors from 12 countries around the World

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





# Writing “The Past Web” took 4 years





A book to support web archiving

Raise awareness about the **importance of web archiving**

New pedagogical resource to **train new web archivists**

Disseminate **services offered by web archives**

Share inspiring **research work that used web archives**

# Besides web archivists, who is this book for?

- **Professors**  
to teach about web preservation
- **Researchers**  
to study the past through the Web
- **Computer scientists**  
to develop new applications to explore historical web data
- **Information professionals**  
to also preserve online information
- **Citizens who use Internet (who doesn't?)**  
to make the past as accessible as the present

# Part 1:

## Why preserve the 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

# Part 2:

## How to collect information from the 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 Drugeon

Pages 43-56

---

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

Elena Demidova, Thomas Risse

Pages 57-67

---



# Part 3:

## How to access historical web data?

### 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

# Part 4:

## How to research the past 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

---

# Part 5:

## How to make web archives become infrastructures?

---

### 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

# Part 6:

## The future of web archiving

### 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 preserved the cited links

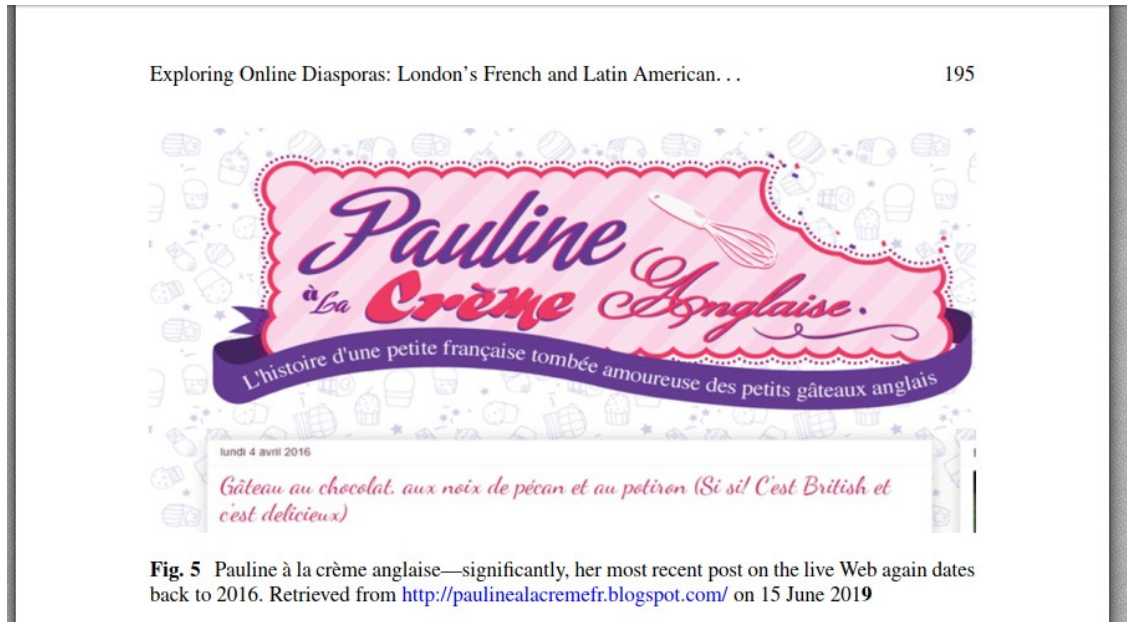
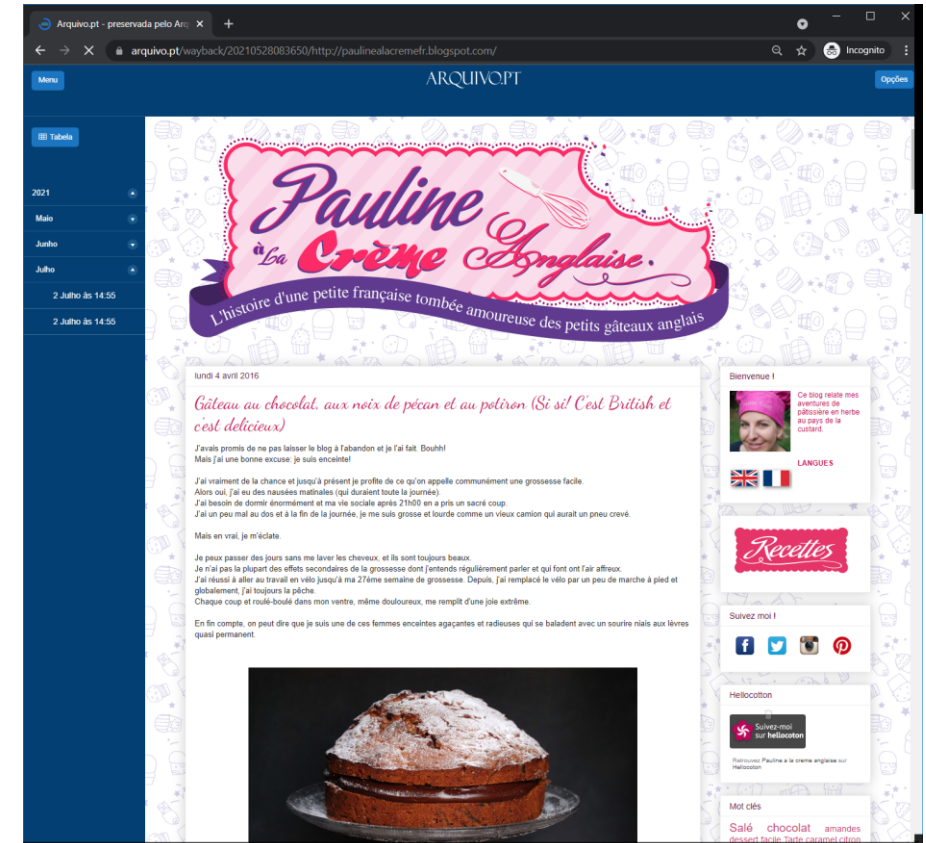
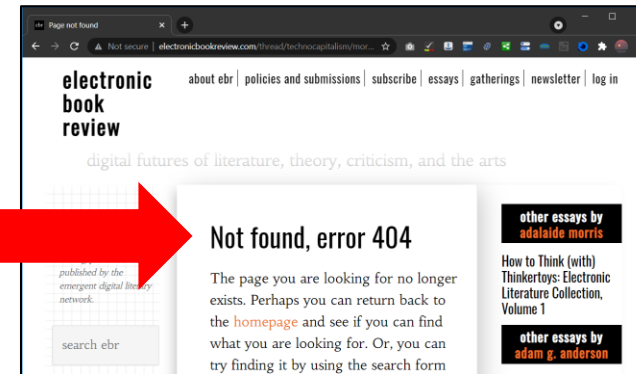
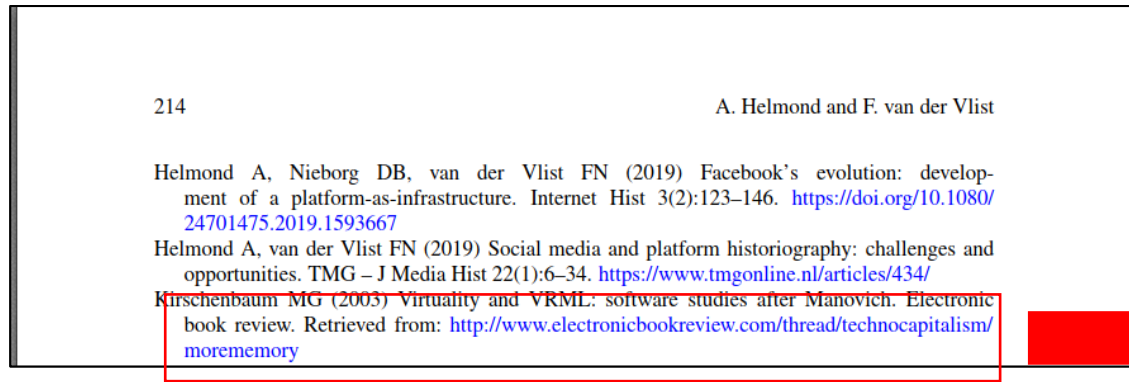


Image screenshot of part of a web page

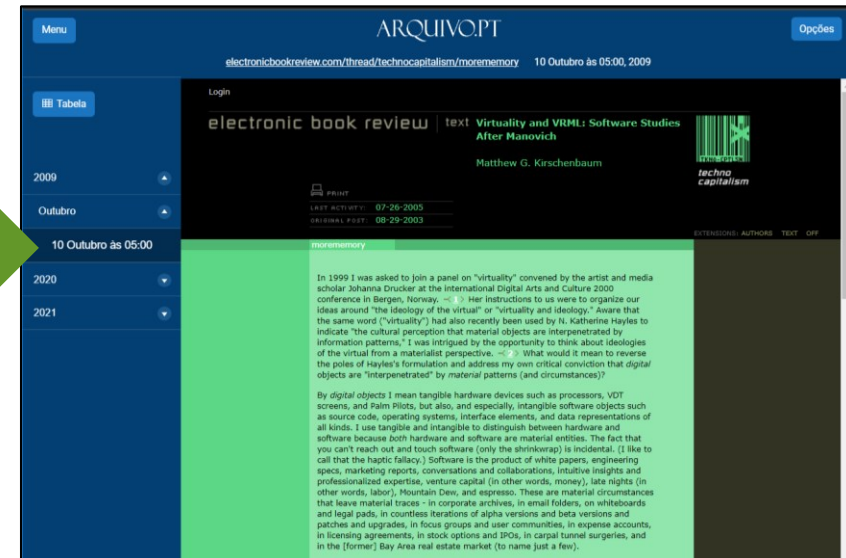


Archived web page, complete and browsable

Some citations of the book to online content are no longer available on the live-web!




But are available on the Past Web preserved by Arquivo.pt



# Promising initial results

“The Past Web”:  
18 000 downloads since July 2021

“Web archiving”:  
11 000 downloads since 2006



© 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

7

79

18k

Citations Mentions Downloads





© 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

11k

Citations Downloads

# The Past Web a book to support web archiving!

Read and disseminate it among your communities!

Book available at:

[www.springer.com/gp/book/9783030632908](http://www.springer.com/gp/book/9783030632908)

Preprint version available in open access at:

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

