

# Arquivo.pt CitationSaver: Preserving Citations for Online Documents

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

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

# Arquivo.pt **preserves** historical content **published online**

***GALILEO*** Versão 2.0

Janeiro 1993

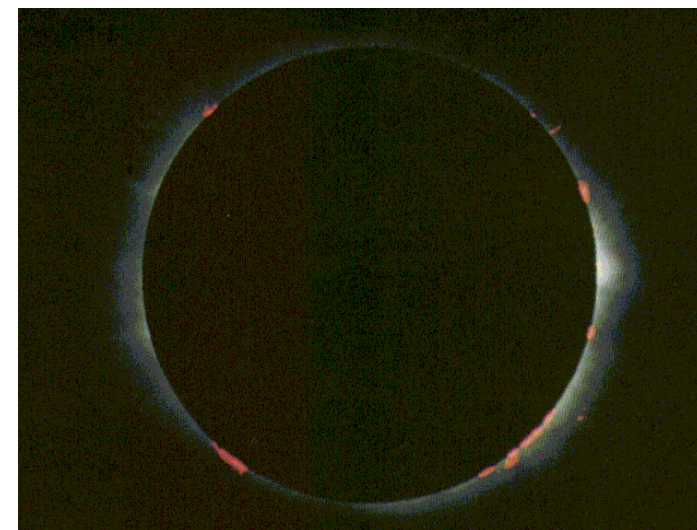
Manual de utilização e sugestões de exploração	<b><i>AUTORES</i></b> <b>Programação:</b> João Veloso <b>Manual:</b> Elisa Prata Pina e M. Augusta Patricio <b>Orientação:</b> Carlos Fiolhais
--	---

Departamento de Física da Universidade de Coimbra

***ÍNDICE***

<a href="#">1. Apresentação sumária do programa</a>
<a href="#">2. Características do equipamento</a>
<a href="#">3. Configuração mínima do computador</a>
<a href="#">4. Ficheiros do programa</a>

[nautilus.fis.uc.pt](http://nautilus.fis.uc.pt)- 1993  
(oldest page)



[spacelink.nasa.gov](http://spacelink.nasa.gov) – 1992  
(oldest image)

# Search for texts from past

Menu ARQUIVO.PT Options

Q Bibliothèque nationale de France Search

1991 6 Aug 2024 3 Apr

Pages Images Narrative Advanced search

About 1.385.409 results from 1991 to 2024

[www.bnf.fr](http://www.bnf.fr)

**Bibliothèque nationale de France**

3 October 2009

Bibliothèque nationale de France Aller au contenu Vie de la Bibliothèque Actualités culturelles ... Découvrir le site web Écrire à la BnF Bibliothèque nationale de France La Bibliothèque Actualités ... Offres d'emploi Presse Lettres d'information Crédits © Bibliothèque nationale de France, 30 ...

[www.bnf.fr](http://www.bnf.fr)

**Bibliothèque nationale de France**

28 June 2009

Bibliothèque nationale de France Aller au contenu Vie de la Bibliothèque Actualités culturelles ... Découvrir le site web Écrire à la BnF Bibliothèque nationale de France La Bibliothèque Actualités ... 'information Crédits © Bibliothèque nationale de France, 25 juin 2009 ...

[More results from www.bnf.fr](#)

# Search for images from past















Menu ARQUIVO.PT Options

Q  Search

1991 ● ● 2024  
6 Aug  3 Apr

Pages Images Narrative Advanced search

About 12.663.508 results from 1991 to 2024

 → <a href="http://www.crl-bourgogne...">http://www.crl-bourgogne...</a> 15 September 2009	 → <a href="http://www.google.com/p...">http://www.google.com/p...</a> 23 December 2010	 → <a href="http://www.europeana-ne...">http://www.europeana-ne...</a> 22 November 2014	 → <a href="http://www.photo-passion...">http://www.photo-passion...</a> 9 July 2009	 → <a href="https://www.sgd.l.org/">https://www.sgd.l.org/</a> 14 September 2019	 → <a href="http://www.europeanareg...">http://www.europeanareg...</a> 26 November 2014	 → <a href="https://br.depositphotos.c...">https://br.depositphotos.c...</a> 14 January 2020
 → <a href="https://www.ekium.eu/fr/a...">https://www.ekium.eu/fr/a...</a> 25 September 2019	 → <a href="http://www.google.com/p...">http://www.google.com/p...</a> 23 December 2010	 → <a href="http://www.lexilogos.com...">http://www.lexilogos.com...</a> 9 July 2009	 → <a href="http://www.camertola.pl/...">http://www.camertola.pl/...</a> 22 January 2012	 → <a href="https://www.genesispark....">https://www.genesispark....</a> 11 September 2019	 → <a href="http://www.photo-passion...">http://www.photo-passion...</a> 9 July 2009	 → <a href="http://www.europeana-ne...">http://www.europeana-ne...</a> 22 November 2014

# Search the **history** of an URL

Menu ARQUIVO.PT

Search  Search

1991 6 Aug 2024 3 Apr

Pages Images Narrative Advanced search

About 202 results from 1991 to 2024

← Table List →

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
							14 Mar	28 Jun					6 Sep		9 Jan	6 Jun	6 Apr	14 Mar	12 Jan	11 Jan	19 Jan	24 Jan	
							22 Oct	2 Oct					24 Nov		15 May	20 Jun	12 Jul	14 Mar	12 Jan	11 Jan	25 Jan	24 Jan	
								3 Oct					26 Nov		20 May	3 Aug	22 Jul	18 Mar	23 Jan	14 Jan	25 Jan	12 Feb	
																	19 Sep	20 Mar	23 Jan	18 Jan	25 Jan	15 Feb	
																	27 Sep	20 Mar	1 Feb	4 Mar	25 Jan	18 Feb	
																	28 Sep	27 Mar	5 Feb	4 Mar	11 Apr	18 Feb	
																	1 Oct	27 Mar	5 Feb	30 Apr	16 Apr	21 Mar	
																	4 Oct	2 Apr	7 Feb	30 Apr	16 Apr	21 Mar	
																	5 Nov	17 Apr	14 Feb	30 Apr	16 Apr	21 Mar	
																		17 Apr	7 Mar	2 Jun	24 Apr		

# Arquivo.pt is an **international** and **interdisciplinary** service

## Half of the **users** are **international**

Country	Users	% Users
1.  Portugal	46,891	46.56%
2.  United States	26,373	26.19%
3.  Brazil	2,266	2.25%
4.  Russia	2,234	2.22%
5.  United Kingdom	2,231	2.22%
6.  Japan	2,172	2.16%
7.  Canada	1,237	1.23%
8.  Mozambique	1,213	1.20%

## Source of data for **Artificial Intelligence**

### Generating a European Portuguese BERT Based Model Using Content from Arquivo.pt Archive

[Nuno Miquelina](#) , [Paulo Quaresma](#) & [Vitor Beires Nogueira](#)

Conference paper | [First Online: 21 November 2022](#)

608 Accesses

Part of the [Lecture Notes in Computer Science](#) book series (LNCS, volume 13756)

## More than 1,000 **academic references**

"Arquivo.pt"|"Portuguese web archive"

Articles Page 6 of about 1,020 results (0.05 sec) My profile

**Any time** **Major Minors-Ontological representation of minorities by newspapers** [\[PDF\] uminho.pt](#)  
Since 2023 [PJP Martins](#), [LJAD Costa](#), [JG Ramalho](#) - 2021 - repositorium.uminho.pt  
Since 2022 The stigma associated with certain minorities has changed throughout the years, yet there's no central data repository that enables a concrete tracking of this representation. Published ...  
Since 2019 [☆ Save](#) [🔗 Cite](#) Cited by 2 [Related articles](#) [All 9 versions](#) [🔗](#)  
Custom range...

**Sort by relevance** **Automatic Classification of Stigmatizing Articles of Mental Illness: The Case of Portuguese Online Newspapers**  
**Sort by date** [JL Oliveira](#) - New Trends in Database and Information Systems ..., 2022 - books.google.com  
The stigma related to mental health continues to be present in online newspapers, where mental diseases are often used metaphorically to refer to entities or situations outside the ...

**Any type**  
**Review articles** [☆ Save](#) [🔗 Cite](#) [Related articles](#)

## Services catalog of the Arquivo.pt web archive

Last updated on October 26th, 2023 at 05:47 pm

---

### Public services

- [Search and access](#) web-archived data since the 1990s
- [Application Programming Interfaces \(APIs\)](#)
- [Suggest websites](#) to be preserved
- [SavePageNow](#): immediately archive web pages
- [Integration of historical web data collections](#)
- [Training on web preservation](#)
- [Open data listing](#) archived web information on various topics
- [CitationSaver](#): extracts links from documents and archives the correspondent web pages
- [Arquivo404](#): presents web-archived pages instead of “pages not found”

### For partners

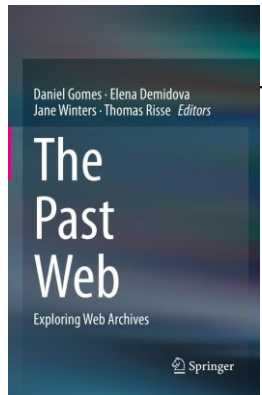
- [Memorial](#) preserves your old website information before deactivating it
- High-quality archive of websites ([on-demand](#))
- [Creation of collections and thematic exhibitions](#)
- [Itinerant exhibition of posters at your facilities](#)

All services at  
[arquivo.pt/catalog](https://arquivo.pt/catalog)

New Service  
**CitationSaver**



# Problem: **scientific citations** for online resources can get **broken**



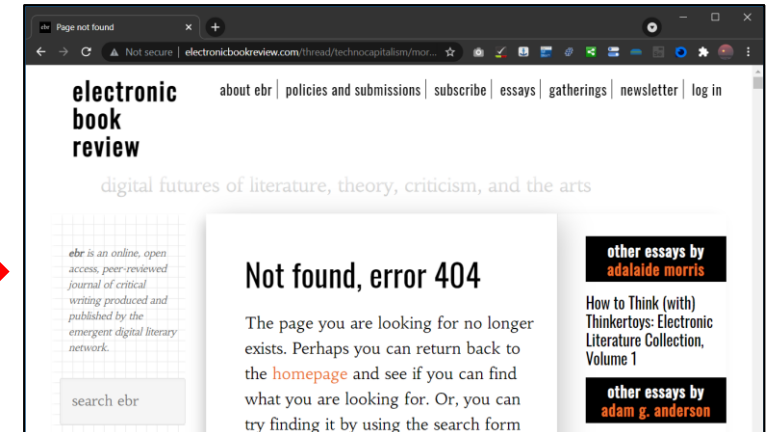
214

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>



Publications lose scientific value because it becomes impossible to reproduce experiments.

# A serious problem that affects all areas of science

Malaysian Journal of Library & Information Science, Vol. 16, no. 3, December 2011: 17-29

## The FASEB Journal

The Journal of the Federation of American Societies for Experimental Biology

HOME | CURRENT ISSUE | NEW ARTICLES | ARCHIVE | ALERTS | RSS | SUBMIT | HELP

### Unavailability of online supplementary scientific information from articles published in major journals

Evangelos Evangelou<sup>\*,1</sup>, Thomas A. Trikalinos<sup>\*,†</sup> and John P.A. Ioannidis<sup>\*,†‡</sup>

Author Affiliations

Correspondence: <sup>1</sup>Correspondence: Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. E-mail: me01760@cc.uoi.gr

« Previous | Next Article »  
Table of Contents

#### This Article

doi:  
10.1096/fj.05-47841sf  
December 2005  
The FASEB Journal  
vol. 19 no. 14 1943-1944

» Abstract  
Full Text  
Full Text (PDF)  
+ Classifications  
+ Services

## Death of web citations: a serious alarm for authors

### The decay and failures of web references

Full Text: [Html](#) [PDF](#)

Author: [Diomidis Spinellis](#) Athens University of Economics and Business in Greece

Published in:

Magazine  
Communications of the ACM [CACM Homepage](#) [archive](#)  
Volume 46 Issue 1, January 2003  
Pages 71-77  
ACM New York, NY, USA  
[table of contents](#) doi > [10.1145/802421.602422](#)



2003 Article

Bibliometrics

- Downloads (6 Weeks): 0
- Downloads (12 Months): 28
- Downloads (cumulative): 1,406
- Citation Count: 20

## JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

Research Article

### Link decay in leading information science journals

Dion Hoe-Lian Goh and Peng Kin Ng

Article first published online: 3 NOV 2006

DOI: 10.1002/asi.20513

Copyright © 2007 Wiley Periodicals, Inc., A Wiley Company



Journal of the American Society for Information Science and Technology  
Volume 58, Issue 1, pages 15-24, 1 January 2007

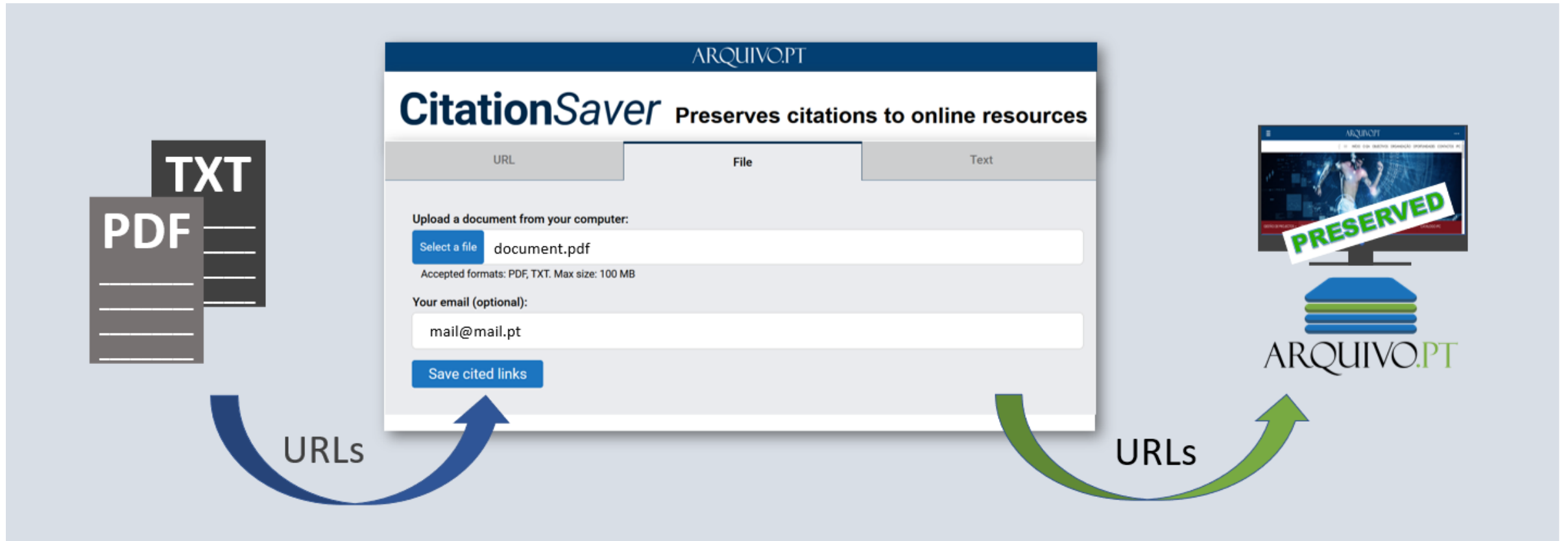
## Persistence and Decay of Web Citations Used in Theses and Dissertations Available at the Sokoine National Agricultural Library, Tanzania

ACADEMIC JOURNAL ARTICLE

By Sife, Alfred S.; Bernard, Ronald

International Journal of Education and Development using Information and Communication Technology, Vol. 9, No. 2, July 1, 2013

# CitationSaver preserves citations in documents



[arquivo.pt/services/citationsaver?l=en](http://arquivo.pt/services/citationsaver?l=en)

## CitationSaver Preserves citations to online resources

Documents cite online content that quickly disappear.

CitationSaver preserves the content of cited links (e.g. web pages cited in a book) so that they can be later recovered from Arquivo.pt.

[Learn more](#)

Submit a document and CitationSaver will preserve its cited links:

URL	File	Text
<b>Insert the URL to the document:</b>		
<input type="text" value="https://sobre.arquivo.pt/wp-content/uploads/SearchingImagesWebArchiveDSAA-202final.pdf"/>		
Accepted formats: PDF, TXT, HTML. Max size: 100 MB		
<b>Your email (optional):</b>		
<input type="text" value="pedro.gomes@fccn.pt"/>		
<input type="button" value="Save cited links"/>		

## CitationSaver Preserves citations to online resources

Documents cite online content that quickly disappear.

CitationSaver preserves the content of cited links (e.g. web pages cited in a book) so that they can be later recovered from Arquivo.pt.

[Learn more](#)

Submit a document and CitationSaver will preserve its cited links:

URL	File	Text
<p data-bbox="198 782 843 818"><b>Upload a document from your computer:</b></p> <div data-bbox="198 829 2186 933"><input data-bbox="206 833 415 929" type="button" value="Select a file"/> Plan.pdf</div> <p data-bbox="224 943 958 979">Accepted formats: PDF, TXT, HTML. Max size: 100 MB</p> <p data-bbox="198 1025 545 1061"><b>Your email (optional):</b></p> <div data-bbox="198 1072 2186 1182"><input data-bbox="198 1072 573 1182" type="text" value="pedro.gomes@fccn.pt"/></div> <div data-bbox="198 1215 588 1296"><input data-bbox="198 1215 588 1296" type="button" value="Save cited links"/></div>		

## CitationSaver Preserves citations to online resources

Documents cite online content that quickly disappear.

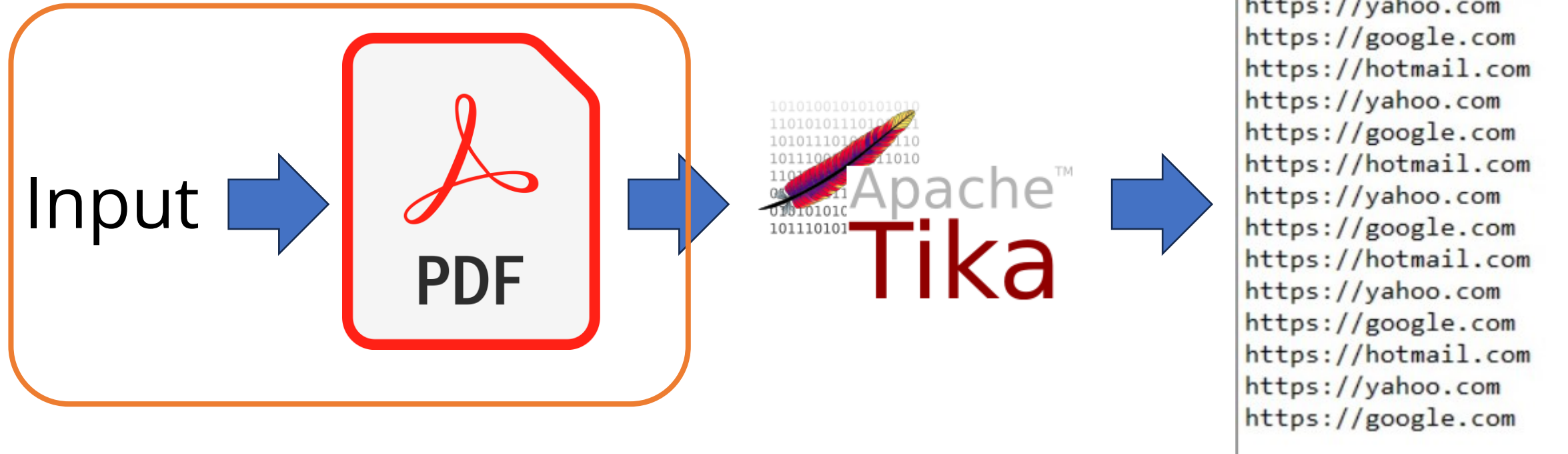
CitationSaver preserves the content of cited links (e.g. web pages cited in a book) so that they can be later recovered from Arquivo.pt.

[Learn more](#)

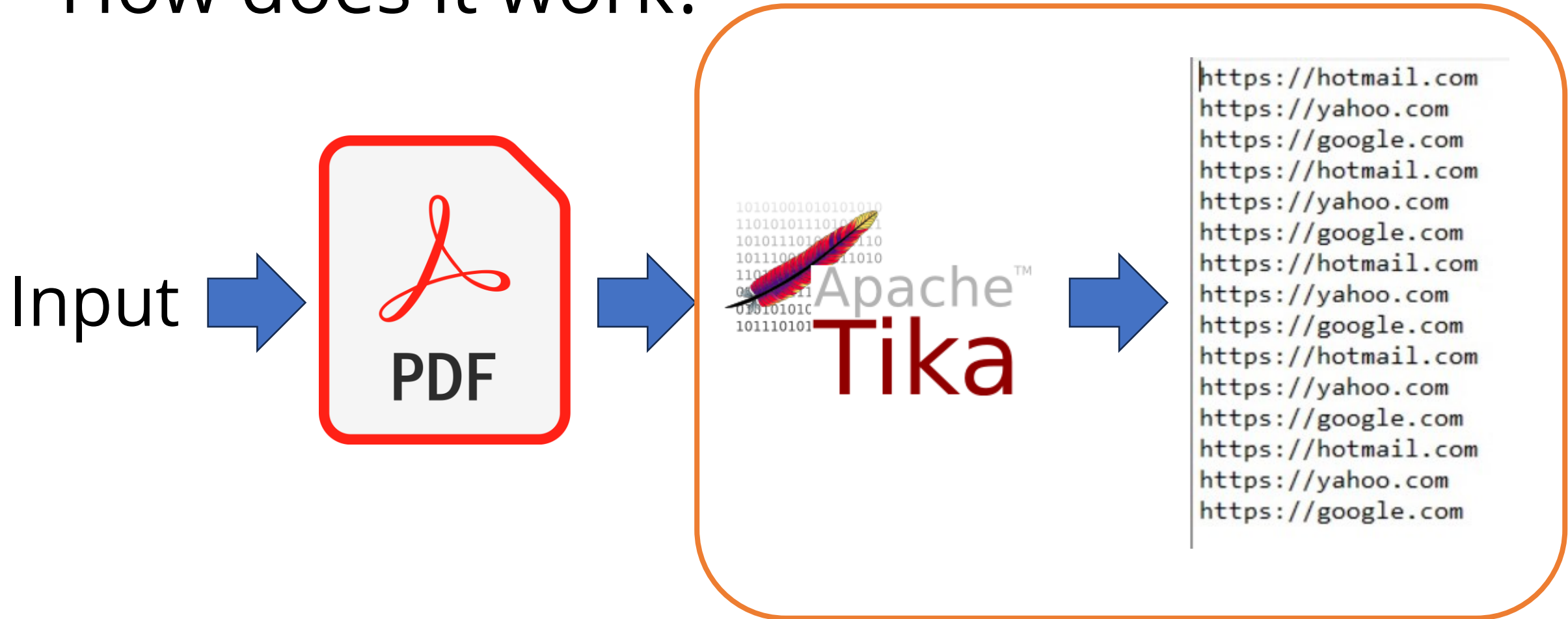
Submit a document and CitationSaver will preserve its cited links:

URL	File	Text
<p data-bbox="468 639 835 668">Insert a text containing URLs:</p> <div data-bbox="468 679 2040 1090" style="border: 1px solid black; padding: 10px;"><p data-bbox="486 704 639 725">REFERENCES</p><p data-bbox="486 732 1335 815">[1] D. <u>Gomes</u>, "Web archives as research infrastructure for digital societies: the case study of <u>Arquivo.pt</u>," <u>Archeion</u>, vol. 123, pp. 46-85, Nov. 2022. [Online]. Available: <u><a href="https://www.ejournals.eu/Archeion/2022/123/art/22601/">https://www.ejournals.eu/Archeion/2022/123/art/22601/</a></u></p><p data-bbox="486 822 1274 905">[2] S. <u>Brin</u> and L. <u>Page</u>, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," <u>Computer Networks</u>, vol. 30, pp. 107-117, 1998. [Online]. Available: <u><a href="http://www-db.stanford.edu/backrub/google.html">http://www-db.stanford.edu/backrub/google.html</a></u></p><p data-bbox="486 912 1327 1022">[3] L. A. <u>Barroso</u>, J. <u>Dean</u>, and U. <u>Holzle</u>, "Web search for a planet: The Google cluster architecture," <u>IEEE Micro</u>,</p></div> <p data-bbox="468 1136 741 1165">Your email (optional):</p> <input data-bbox="468 1176 2040 1259" type="text" value="pedro.gomes@fccn.pt"/> <p data-bbox="468 1286 774 1350"><a href="#">Save cited links</a></p>		

# How does it work?



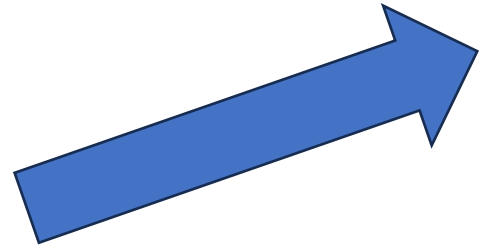
# How does it work?





# How does it work?

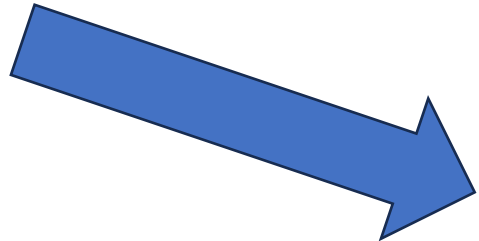
```
https://hotmail.com  
https://yahoo.com  
https://google.com  
https://hotmail.com  
https://yahoo.com  
https://google.com  
https://hotmail.com  
https://yahoo.com  
https://google.com  
https://hotmail.com  
https://yahoo.com  
https://google.com  
https://hotmail.com  
https://yahoo.com  
https://google.com
```



Heritrix



Brozzler



Browsertrix

# CitationSaver Limitations

- We are depended on users to find scientific documents.
- Only one document can be submitted at a time.



To solve the limitations  
Arquivo.pt integrated external APIs

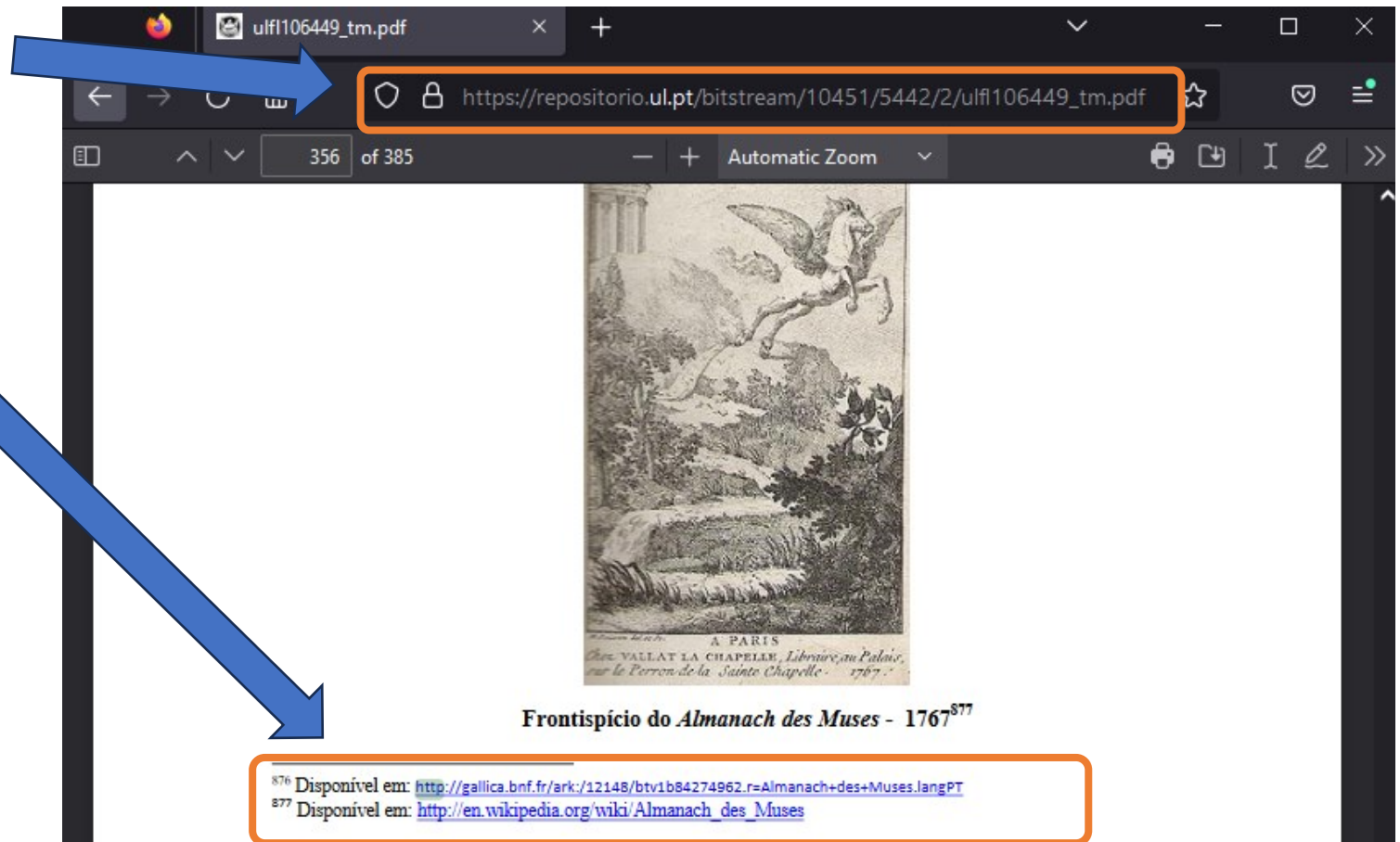
**Arquivo.pt**

**RCAAP API**  
**PTCRIS SciProj API**  
**CienciaVitae API**



# RCAAP (Open Access Scientific Repository of Portugal) provides an API with access to PDFs

1. With the API, Get the links to the PDFs.
2. Download the PDFs
3. Extract the URLs using CitationSaver Service
4. Crawl the URLs



# Extracted and preserved

bibliotheca Augustana - presen. X

ARQUIVO.PT

hs-augsburg.de/~harsch/Chronologia/Lsante01/Horatiu/hor\_c000.html 20 March at 23h45, 2016

Menu Options

Table

2016

March

20 March 23h45, 2016

2022

[<<< introductio](#)

BIBLIOTHECA AUGUSTANA

Q. Horatius Flaccus  
65 - 8 a. Chr. n.

Carminum libri IV

23 a. Chr. n.

Textus:  
Horaz, Oden und Epoden  
ed. B. Kytzler, Stuttgart 1978

Liber primus

[Carmen I](#) Maecenas atavis edite regibus  
[Carmen II](#) Iam satis terris nivi atque dirae  
[Carmen III](#) Sic te diva potens Cypri

# Data obtained from the RCAAP API

- 151 466 PDFs
- 2 197 292 URLs
- 9.1 TB Stored



[arquivo.pt/services/citationsaver?l=en](http://arquivo.pt/services/citationsaver?l=en)

[contacto@arquivo.pt](mailto:contacto@arquivo.pt)

