Web archives as research infrastructure for digital societies: the case study of Arquivo.pt

Daniel Gomes
Fundação para a Ciência e a Tecnologia: Arquivo.pt
daniel.gomes@fccn.pt
http://orcid.org/0000-0002-5447-4581

Keywords

Web archiving, digital preservation, recommendations

Abstract

Humans are the dominant species on Earth. Our advantage comes from our unique capacity of organising at large scale to reach common goals. In digital societies, organising requires communicating information and these days, most of it is published exclusively online. The problem is that online information disappears quickly, after a few months. Humanity's dependence on online information is strong but still recent and the consequences of losing the historical perspective over online data are yet to be seen.

Web archives are digital preservation systems that collect, store and provide access to historical web data. Scientific researchers have been using web archives. However, web archives should also be used by the wider public so that they may serve digital societies. Arquivo.pt is a public web archive started in 2007 that enables search and access to historical information preserved from the Web since the 1990s. This article presents Arquivo.pt as a case study for a research infrastructure that has been developed to serve wider communities at national and international levels. The article shares the main lessons learned so that other web archiving initiatives may arise and be developed at a faster pace. It describes the existing tools and activities which enable exploration of historical web-archived collections. Finally, it presents challenges related to creating web archives and proposes actions to address them.

Introduction

The Web replaced printed media. In 2009, Tim Berners-Lee the inventor of the World Wide Web stated that "when we look at the Web, we don't look at it anymore as connected computers or as connected Web pages. We look at the Web now as humanity connected. Humanity connected by technology"[1]. In 2022, most of the information that constantly rules our societies, from governmental institutions to personal lives, is published exclusively online. Dogmas such as "Information is power", "Data is the new oil" or "Look into the past to see the future" became common jargon across mass media, political agendas or social media. But in reality, societies passively waste most of the online information they strongly invest in producing everyday. Most of the information published online disappears after a few months[2]. Information is the ground for human organisations and this tremendously fast pace of information creation, and loss, is unprecedented in human history. The consequences of losing the historical perspective over the data published online are yet to be seen. However, history taught us that "those who cannot remember the past are condemned to repeat it"[3].

Web archives are digital preservation systems that collect, store and provide access to historical web data. Scientific researchers have been using web archives since their inception. However, as the Internet penetrated all levels of everyday life, web archives must assume their role as research infrastructures useful to a wider public so that they may serve digital societies. This article presents the Arquivo.pt web archive as a case study for a research infrastructure that has been developed to serve wider communities.

Arquivo.pt is a public, free service that enables anyone to search and access historical information preserved from the Web since the 1990s. Arquivo.pt contains billions of files collected from websites in several languages (about half of its users come from outside of Portugal). The search services provided by Arquivo.pt include full-text search, image search, version history listing, advanced search and application programming interfaces (API). The project became an official public service in 2013. Several added-value services that explore this unique source of historical information have been developed, proving that a web archive is a valuable infrastructure for modern societies. This article aims to share the main lessons learned while developing and operating a web archiving service for 15 years, so that more initiatives may arise and be developed at a faster pace.

[1] Berners-Lee Launches, T. 'www Foundation' at Igf 2009, ACM TECHNEWS, Ars Technica November 23, 2009 https://arstechnica.com/tech-policy/2009/11/tim-berners-lee-launches-www-foundation-at-igf-2009/, accessed 31 October 2022.

[2] Gomes, D. and Silva, M.J., 2006, July. Modelling information persistence on the web. In Proceedings of the 6th international conference on Web engineering (pp. 193-200).

[3] George Santayana, Wikiquote, https://en.wikiquote.org/wiki/George_Santayana, accessed 31 October 2022.

Related work and background

Brewster Kahle launched the Internet Archive in 1996[4] and the first research publications about how to build web archives were published during the early 2000s. The first book that documented the main web archiving efforts made so far was edited by Julien Masanès in 2006[5]; Masanès had also been organising web archiving workshops since 2001. During the early 2010s, the European Union funded several research projects in the area of web archiving, such as ARCOMEM, LiWA or Blog4Ever, but in recent years this funding was discontinued. Nonetheless, a research effort regarding web archiving evolution showed that between 2010 and 2014 the web archiving community continued to grow considerably[6]. As web archives developed to become usable and accessible research infrastructure, interest from several research communities, such as the humanities community, increased[7]. Niels Brugger led research in the area of media studies using web archives. Since the early days of the Web, he had been raising awareness that the Internet should be considered an integral part of our culture, rather than just a technical device. The book entitled "Archiving Websites: General Considerations and Strategies", published in 2005, was the first one that specifically addressed preservation of web content from media studies perspective[8]. Since then, Niels Brugger published several books where web archives were used as source to obtain research data such as "Web History"[9], "Web 25: Histories from the First 25 Years of the World Wide Web"[10], "The Web as History: Using Web Archives to Understand the Past and the Present"[11], "The Archived Web Doing History in the Digital Age"[12] or "The Historical Web and Digital

[4] Kahle, B., 1997. Preserving the internet. Scientific American, 276(3), pp.82-83.

[5] Masanes, J., 2006. Web archiving: issues and methods. In Web archiving (pp. 1-53). Springer, Berlin, Heidelberg.

[6] Costa, M., Gomes, D. and Silva, M.J., 2017. The evolution of web archiving. International Journal on Digital Libraries, 18(3), pp.191-205.

[7] Gomes, D. and Costa, M., 2014. The importance of web archives for humanities. International Journal of Humanities and Arts Computing, 8(1), pp.106-123.

[8] Brügger, N., 2005. Archiving Websites. General Considerations and Strategies: General Considerations and Strategies.

[9] Brügger, N. ed., 2010. Web history (Vol. 56). Peter Lang.

[10] Brügger, N., Locatelli, E., Weber, M. and Nanni, F., 2017. Web 25: histories from the first 25 years of the World Wide Web.

[11] Schroeder, R. and Brügger, N., 2017. The Web as History: Using Web Archives to Understand the Past and the Present (p. 296). UCL Press.

[12] Brügger, N., 2018. The archived web: doing history in the digital age. MIT Press.

Humanities"[13]. In 2017, Niels Brugger founded the new journal, called "Internet Histories"[14]. In the history area, Ian Milligan has been leading historians to explore digital sources and web archives in particular. He co-edited the books "Exploring Big Historical Data: The Historian's Macroscope"[15] and "SAGE Handbook of Web History"[16]. He authored the books "History in the Age of Abundance: How the Web is Transforming Historical Research"[17] and "The Transformation of Historical Research in the Digital Age"[18]. Among other works that address web archives, Jane Winters led the "Big UK Domain Data for the Arts and Humanities" project which awarded bursaries to 10 researchers to carry out research in their subject area using the UK web archive. The case studies that they produced showcase the richness of web archives as a source for humanities and other researchers[19]. Valérie Schafer is a Professor in Contemporary European History who has been defending public engagement with web archives[20] and Anat Ben-David is a Professor of Communication who has been showing the value of web archives to analyse social and political events[21],[22].

Added-value tools built on top of web archives have also been developed to aid researchers in their work. The Archives Unleashed project provides a software toolkit that facilitates the extraction and analysis of data sets derived from web-archived content, making large amounts of historical web content accessible to research the recent past[23]. The GLAM Workbench led by Tim Sherratt provides researchers with examples, tools, and documentation to help them

[13] Brügger, N. and Laursen, D. eds., 2019. The historical web and digital humanities: the case of national web domains. Routledge.

[14] Brügger, N., Goggin, G., Milligan, I. and Schafer, V., 2017. Introduction: Internet histories. Internet Histories, 1(1-2), pp.1-7.

[15] Graham, S., Milligan, I., Weingart, S.B. and Martin, K., 2016. Exploring big historical data: the historian's macroscope.

[16] Brügger, N. and Milligan, I. eds., 2018. The SAGE handbook of web history. Sage.

[17] Milligan, I., 2019. History in the age of abundance?: how the web is transforming historical research. McGill-Queen's University Press.

[18] Milligan, I., 2022. The Transformation of Historical Research in the Digital Age. Elements in Historical Theory and Practice.

[19] Winters, J., 2015. „Big UK Domain Data for the Arts and Humanities", Presentation, 2015 International Internet Preservation Coalition General Assembly, April 27-May 1, 2015. Silicon Valley, California, https://buddah.projects.history.ac.uk/.

[20] Schafer, V. and Winters, J., 2021. The values of web archives. International Journal of Digital Humanities, 2(1), pp.129-144.

[21] Ben-David, A. and Amram, A., 2018. The Internet Archive and the socio-technical construction of historical facts. Internet Histories, 2(1-2), pp.179-201.

[22] Ben-David, A., 2019. National web histories at the fringe of the Web: Palestine, Kosovo, and the quest for online self-determination. In The Historical Web and Digital Humanities (pp. 89-109). Routledge.

[23] Ruest, N., Lin, J., Milligan, I. and Fritz, S., 2020, August. The archives unleashed project: Technology, process, and community to improve scholarly access to web archives. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (pp. 157-166), https://archivesunleashed.org/.

explore and use web archives through a large and growing collection of Jupyter notebooks[24]. The Arquivo.pt annual awards were established in 2018; their aim is to showcase innovative works that use preserved historical information from the Web[25]. The works can address any subject as long as they use Arquivo.pt as source of information. Over 5 editions, 127 applications were received and the 15 works that were awarded clearly demonstrate the utility of web archives in digital societies. Studies performed over the historical web data addressed political[26],[27], sociological[28] or health issues[29]. On the other hand, innovative search tools that enable further utilization of web-archived data were also developed, such as Politiquices.pt, Desarquivo (available at https://arquivo.pt/desarquivo), Revisionista.pt or ContaMeHistorias.pt.

In contrast with the significant number of books and works that document research done using web archives, only two books were published that address the process of carrying out web archiving activities namely “Web Archiving” and “The Past Web: exploring web archives”[30]. Research on web archiving tools and algorithms has also been developing at a slow pace. The PhD thesis “Information Search in Web Archives”, published in 2014, was a significant contribution to improve textual search in web archives but no scientific research has been done in this area since then[31]. The SolrWayback has been developed by the community of web archiving practitioners based on the Apache Solr technology[32]. Despite being a useful tool with interesting features, no scientific research has been conducted to study essential aspects required to deliver quality services, like for example ranking search results based on temporal features of historical web data, as opposed to the up-to-date features of live web data. Michael Nelson from The Web Science and Digital Libraries (WS-DL) Research Group at Old Dominion University has been leading research on automatic exploration of web-archived data[33], optimising the operation of web archives[34] or developing innovative services such as Carbondate.cs.odu.edu which automatically estimates the age of web resources[35]. Herbert Van

[24] Sherratt, T. and Jackson, A., 2020. GLAM-Workbench/web-archives, https://glam-workbench.net/web-archives/, accessed 31 October 2022.

[25] Arquivo.pt, Arquivo.pt Awards, https://arquivo.pt/awards, accessed: 31 October 2022.

[26] Arquivo do Parlamento, https://arquivo-parlamento.pt/, accessed: 31 October 2022.

[27] meuParlamento.pt, http://www.meuparlamento.pt/, accessed: 31 October 2022.

[28] Major Minors, http://minors.ilch.uminho.pt/, accessed: 31 October 2022.

[29] Classificação automática de artigos estigmatizantes de doenças mentais em jornais de notícias portugueses online, https://alina-yanchuk02.github.io/estigma/, accessed: 31 October 2022.

[30] Gomes, D., Demidova, E., Winters, J. and Risse, T., 2021. Past Web. Springer International Publishing.

[31] Miguel Costa, Information Search in Web Archives, PhD thesis, Universidade de Lisboa, December 2014

[32] SolrWayback 4.0 release! What’s it all about? Part 2, https://netpreserveblog.wordpress.com/2021/03/04/solrwayback-4-0-release-whats-it-all-about-part-2/, accessed: 31 October 2022.

[33] Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C. and Nelson, M.L., 2011, June. How much of the web is archived?. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (pp. 133-136).

[34] AlSum, A., Weigle, M.C., Nelson, M.L. and Van de Sompel, H., 2014. Profiling web archive coverage for top-level domain and content language. International Journal on Digital Libraries, 14(3), pp.149-166.

[35] SalahEldeen, H.M. and Nelson, M.L., 2013, May. Carbon dating the web: estimating the age of web resources. In Proceedings of the 22nd International Conference on World Wide Web (pp. 1075-1082).

de Sompel led the Memento project that designed one of the most important contributions to openness and interoperability of web archives, the Memento protocol[36],[37]. For instance, this protocol was applied to implement the Time Travel service that enables automatic searching of several web archives[38]. Martin Klein studied the impact of broken links in digital libraries[39] and investigated algorithms based on web archives to recover missing content[40] or generate event collections[41].

Researchers, citizens or institutions sometimes need to archive information from the web on their own, for documentation purposes. The Webrecorder project provides a suite of open-source tools and packages to perform the complete web archiving workflow[42]. ArchiveWeb.page enables any user to archive information from the web using a web browser and store it in the adequate standard WARC format[43], ReplayWeb.page and pywb enable the replay of web-archived content. Browsertrix enables automation of data acquisition from the web, for instance to archive all the pages from a given website. The Webrecorder project was a breakthrough in web archiving because it enables any user or small institution to create their own web archives of selected information using standard formats that allow reuse and ensure interoperability of the web-archived data.

Despite growing interest of the research community in using historical web data, literature discussing the ways to manage web archiving services and establish them as crucial infrastructure in modern societies is not abundant. This article aims to contribute to fill this gap by describing the main features of the Arquivo.pt web archive and share the experience obtained during its development and operation since its original proposal as part of a PhD thesis published in 2007[44].

Short history of Arquivo.pt

Arquivo.pt is a governmental service provided by the Foundation for Science and Technology (Portugal) mandated to preserve publicly accessible information regarding Portugal. However,

[36] Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S. and Shankar, H., 2009. Memento: Time travel for the web. arXiv preprint arXiv:0911.1112.

[37] Van de Sompel, H., Nelson, M. and Sanderson, R., 2013. RFC 7089-HTTP framework for time-based access to resource states-Memento. Internet Engineering Task Force (IETF), RFC.

[38] Memento Time Travel, http://timetravel.mementoweb.org/, accessed: 31 October 2022.

[39] Jones, S.M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R. and Grover, C., 2016. Scholarly context adrift: three out of four URI references lead to changed content. PloS one, 11(12).

[40] Klein, M. and Nelson, M.L., 2014. Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. International Journal on Digital Libraries, 14(1), pp.17-38.

[41] Klein, M., Balakireva, L. and Van de Sompel, H., 2018, May. Focused crawl of web archives to build event collections. In Proceedings of the 10th ACM Conference on Web Science (pp. 333-342).

[42] Webrecorder: Web archiving for all!, https://webrecorder.net/, accessed: 31 October 2022.

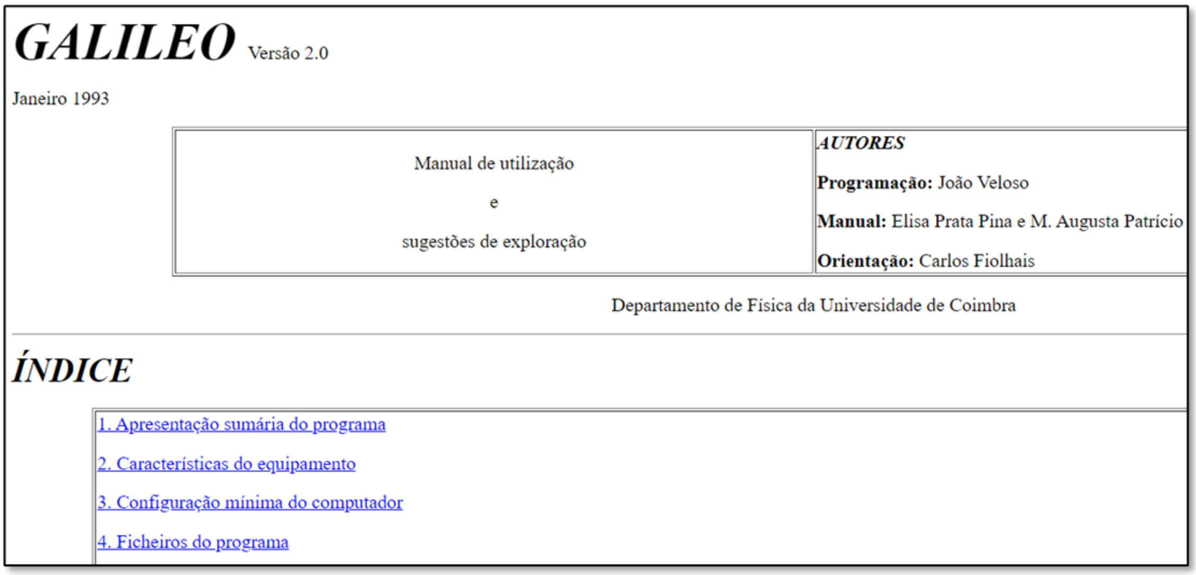
[43] ISO, I., 2009. 28500: 2009 Information and documentation-WARC file format. International Organization for Standardization.

[44] Gomes, D.C., 2006. Web Modelling for Web Warehouse Design (Doctoral dissertation, Universidade de Lisboa (Portugal)).

it also preserves selected online information related to research and education at an international level. Periodically, the Arquivo.pt system automatically collects and stores information published on the web. The Arquivo.pt hardware infrastructure is hosted at its own data-center and it is managed by full-time dedicated staff. The project began in 2007 and became an official running service in 2013 with the publication of the Decree-Law[45] that mandated preservation of content available on the national Internet. However, the original idea emerged in 2001, as part of a research project named TUMBA! (Portuguese initials of "we have an alternative search engine!"), developed at the University of Lisbon. Tumba! enabled full-text search of the most recent crawl of the Portuguese web. Following tumba!, came a web archive prototype called Tomba (named after the Tombo Tower which is the Portuguese National Archive established in 1378). Tomba enabled access to different versions of web pages collected by the tumba! search engine for 4 years (2002-2006). The initial team who worked on the Portuguese web archive project was composed of 3 former researchers of tumba!. The know-how and experience gained from these academic projects were crucial to the development of Arquivo.pt.

Developing a web archive raised significant challenges in areas such as web archive information retrieval, user experience or quality assurance. The members of the Arquivo.pt team have been publishing technical and scientific articles related to web archiving in open-access since 2008[46]. All the developed software is available as free open source projects[47].

Collections and selection



GALILEO Versão 2.0

Janeiro 1993

Manual de utilização

e

sugestões de exploração

AUTORES

Programação: João Veloso

Manual: Elisa Prata Pina e M. Augusta Patrício

Orientação: Carlos Fiolhais

Departamento de Física da Universidade de Coimbra

ÍNDICE

1. Apresentação sumária do programa

2. Características do equipamento

3. Configuração mínima do computador

4. Ficheiros do programa

Figure 1: Oldest page preserved at Arquivo.pt nautilus.fis.uc.pt published in January 1993,

[45] Ministério da Educação e Ciência, Decreto-Lei n.º 55/2013, Diário da República, 1.ª série — N.º 75 — 17 de abril de 2013.

[46] Arquivo.pt, Publications, https://arquivo.pt/publications, accessed: 31 October 2022.

[47] Arquivo.pt,, Arquivo.pt · GitHub, https://github.com/arquivo/, accessed: 31 October 2022.

available at [*https://arquivo.pt/wayback/19950322142012/http://nautilus.fis.uc.pt/softc/programas/manuais/galileo/galileo.htm*](https://arquivo.pt/wayback/19950322142012/http://nautilus.fis.uc.pt/softc/programas/manuais/galileo/galileo.htm)

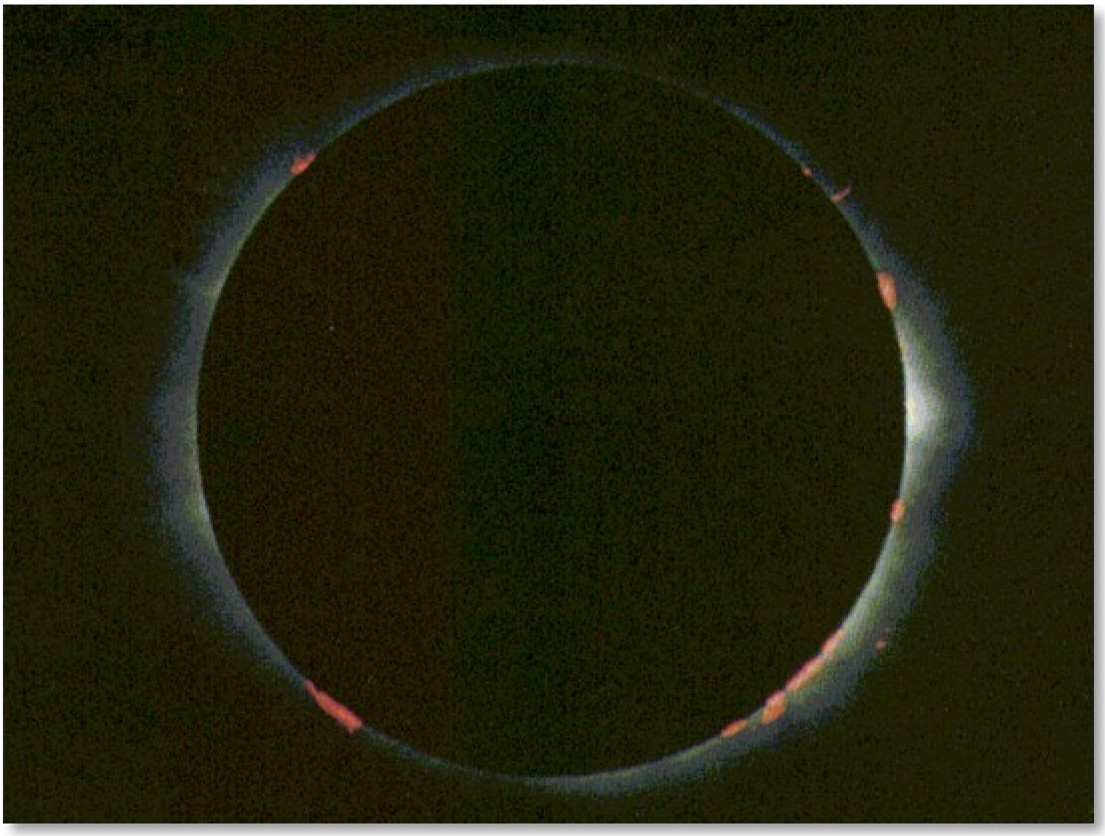
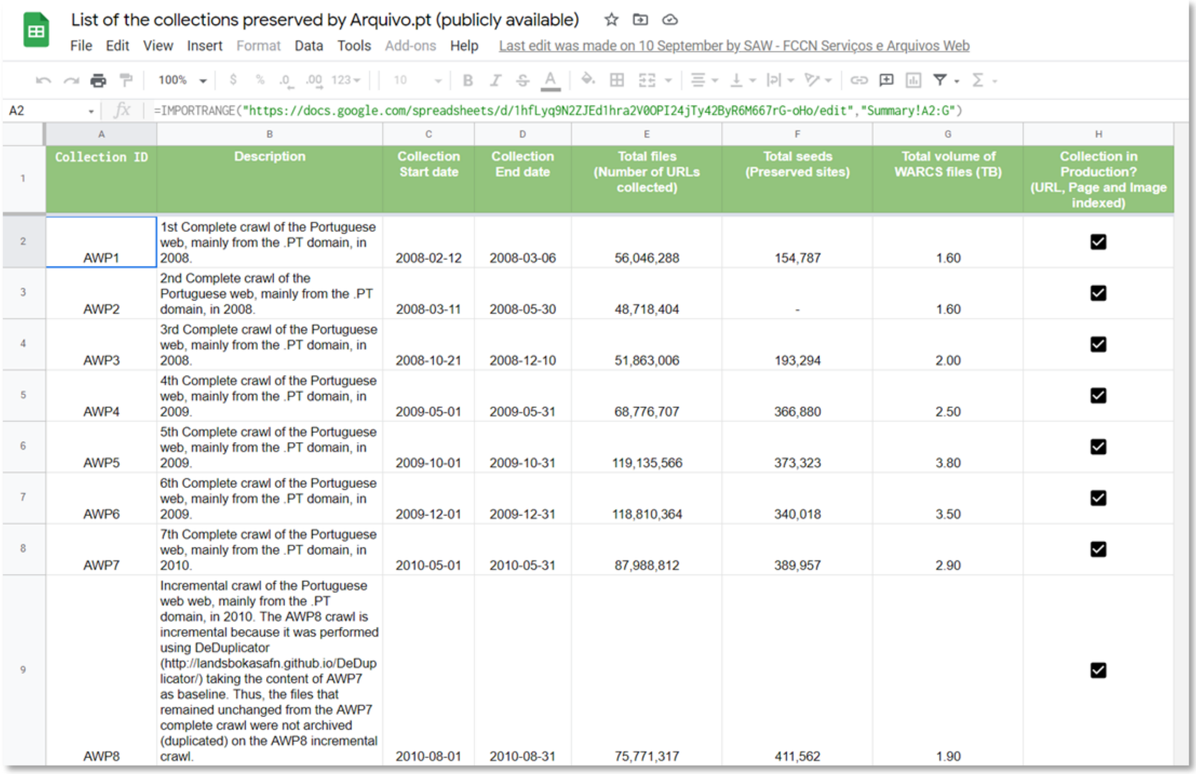


Figure 2. Oldest image preserved at Arquivo.pt spacelink.nasa.gov – published in April 1992, available at [*https://arquivo.pt/wayback/19920415005528im_/http://spacelink.nasa.gov/Instructional.Materials/Curriculum.Support/Space.Science/Our.Solar.System/Our.Sun/Solar.Eclipse-full.gif*](https://arquivo.pt/wayback/19920415005528im_/http://spacelink.nasa.gov/Instructional.Materials/Curriculum.Support/Space.Science/Our.Solar.System/Our.Sun/Solar.Eclipse-full.gif)

Figure 1 presents the oldest page preserved at Arquivo.pt, published in January 1993 by the Physics department of the University of Coimbra. It is a user manual for a computer program that simulates the GALILEO spacecraft trip to planet Jupiter. Figure 2 presents the oldest image preserved at Arquivo.pt, which was published in April 1992 by NASA and linked from a page on the website that hosted the oldest page. These examples that illustrate the preservation of national and international web-content relevant for education were found thanks to a research work led by the Old Dominion University[48]. This research was possible due to the open-access

[48] Alam, S., Weigle, M., Nelson, M., Melo, F., Bicho, D. and Gomes, D., 2019, June. MementoMap framework for flexible and adaptive web archive profiling. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 172-181). IEEE.

of web-archived data and the API based on the Memento international protocol that enabled its automatic processing.



List of the collections preserved by Arquivo.pt (publicly available)

Last edit was made on 10 September by SAW - FCCN Serviços e Arquivos Web

A2 =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1hfLyq9N2ZJEd1hra2V0OPI24jTy42ByR6M667rG-oHo/edit","Summary!A2:G")

	A	B	C	D	E	F	G	H
1	Collection ID	Description	Collection Start date	Collection End date	Total files (Number of URLs collected)	Total seeds (Preserved sites)	Total volume of WARCS files (TB)	Collection in Production? (URL, Page and Image indexed)
2	AWP1	1st Complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-02-12	2008-03-06	56,046,288	154,787	1.60	☑
3	AWP2	2nd Complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-03-11	2008-05-30	48,718,404	-	1.60	☑
4	AWP3	3rd Complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-10-21	2008-12-10	51,863,006	193,294	2.00	☑
5	AWP4	4th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-05-01	2009-05-31	68,776,707	366,880	2.50	☑
6	AWP5	5th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-10-01	2009-10-31	119,135,566	373,323	3.80	☑
7	AWP6	6th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-12-01	2009-12-31	118,810,364	340,018	3.50	☑
8	AWP7	7th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2010.	2010-05-01	2010-05-31	87,988,812	389,957	2.90	☑
9	AWP8	Incremental crawl of the Portuguese web web, mainly from the .PT domain, in 2010. The AWP8 crawl is incremental because it was performed using DeDuplicator (http://landsbokasafn.github.io/DeDuplicator/) taking the content of AWP7 as baseline. Thus, the files that remained unchanged from the AWP7 complete crawl were not archived (duplicated) on the AWP8 incremental crawl.	2010-08-01	2010-08-31	75,771,317	411,562	1.90	☑

Figure 3: list of all the collections generated by Arquivo.pt publicly available at arquivo.pt/collections.

As of August 2022, Arquivo.pt held 145 collections containing 13 269 million web files (868 TB of historical web data) obtained from 28.9 million websites. The idealistic objective of Arquivo.pt is to archive all the published online content related to Portugal so that it may be preserved as memory for future access. Despite being a directive goal, this objective is not fully attainable because some pages cannot be archived due to technological limitations, while other pages change so rapidly that Arquivo.pt cannot acquire all of their versions. Therefore, Arquivo.pt generates several types of collections according to their crawl frequency, scope and quality of the acquisition process. A crawler is a system that automatically collects information from the web[49]. Several technologies are applied to collect information from the web, such as Heritrix, Brozzler or Browsertrix, but the speed of crawl is inversely proportional to the quality of the obtained content. If the objective is to thoroughly archive a web page in highest quality possible, including, for instance, all the embedded videos and content generated from AJAX calls, then browser-based crawlers should be used, such as Brozzler or Browsertrix. However,

[49] Gomes, D. and Silva, M.J., 2008. The Viúva Negra crawler: an experience report. Software: Practice and Experience, 38(2), pp.161-188.

browser-based crawlers are slower and demand more resources (e.g. CPU) than conventional crawlers that archive web content by following links on web pages such as Heritrix. Figure 3 presents the list of collections generated by Arquivo.pt. The main types of collections are:

- Daily crawls (prefixed with "FAWP" on column A "Collection ID" of the list): a selection of national online publications is collected every day using a combination of browser-based and conventional crawlers;
- Monthly crawls (prefixed with "MAWP"): a selection of national and international online publications is collected every month using browser-based crawlers;
- Trimestral crawls (prefixed with "AWP"): lists of all the domains registered under the .PT domain is collected every 3 months using conventional crawler due to the large volume of data involved;
- High-quality crawls (prefixed with "RAQ"): a selection of websites that were iteratively archived and curated, in highest quality possible, using the best combination of technologies available;
- Save Page Now crawls (prefixed with "SAWP"): pages web-archived by users in high-quality using the savepagenow service (see Section SavePageNow);
- Complete-page crawls (prefix with "PATCHING"): collections of web resources archived through the collaborative invocation of the Complete page service by users (see Section Complete page) or through its automatic execution in batch to fix selected lists of web pages;
- Donated collections (prefix varies): donated or acquired collections of historical web content integrated in Arquivo.pt (e.g. collection Geocities). The web content is received in varied formats and converted into the standard WARC format;
- Special crawls (prefixed with "EAWP"): selections of pages about a given topic, crawled with varied frequency and methodologies;

The *Special crawls* focus on events, such as elections, which require specific efforts to be identified and collected because the relevant online information is spread across the web and quickly disappears. One approach was to request voluntary contributions from the users to nominate web addresses to be collected (seeds). However, the obtained nominations despite being relevant were few. Therefore, we developed a low-cost methodology, requiring very limited human resources, to automatically select and preserve online information about any given topic [50]. In summary, a human curator creates a data-set of relevant keywords about a target topic, such as candidates and parties running in the elections. Then a software script uses these keywords to automatically search for online content using the APIs of live web search engines. The original motivation to develop this selection methodology was the need to preserve online content about research and development funded by European institutions. Previously, we tried to apply the brute-force approach of crawling the .EU domain to obtain relevant content related to Europe. The obtained results showed that it was inadequate due to the large prevalence of crawler traps, web spam, porn or suspicious websites registered under this top-

[50] Bicho, D. and Gomes, D., 2016. Preserving Websites Of Research & Development Projects. In iPRES.

level domain[51],[52]. The developed methodology has been applied to automatically select content from the web to originate special thematic collections; for instance to create a cross-lingual collection that documents the 2019 European Elections[53] or to document the online results produced by the projects funded under the H2020 programme[54].

Other collections of international interest for research purposes were generated in collaboration with external organisations. The Content Development Group of the International Internet Preservation Consortium coordinated international initiatives to select relevant online information documenting, for instance, the European Refugee Crisis, the Olympic Games, the World War I Commemoration, the Coronavirus (Covid-19) outbreak or International Cooperation Organisations. In 2009, the Archive Team gathered 641 GB of information from Geocities.com, which was the first major "social network" that enabled anyone to publish information on the Web[55]. It was created in 1994 and shut down in 2009. Anat Ben-David proposes an approach for filling-in gaps in existing archives by taking into account platform dynamics and cultural differences in link sharing practices[56]. As case study, the author identified web resources that documented the 2014 War in Gaza. Arquivo.pt obtained and integrated the web data selected for the previous collections so that they may be further explored through the innovative tools provided by Arquivo.pt (e.g. full-text search, image search or API) in areas such as arts, humanities or sociology.

However, as the web continues to quickly evolve, the methods to select and acquire web content to be preserved must also be tuned and adapted to new contexts. Cryptocurrency is a global phenomenon that shapes our current times and must be documented. It is an exclusively digital phenomenon that revolutionised the world by influencing other areas such as finance or art. Due to the particular context of cryptocurrency we had to use specific APIs, such as Coingecko.com API, to select relevant information to be archived[57]. Information (and misinformation) related to cryptocurrency has been constantly published and erased as projects

[51] Arquivo.pt, A first attempt to archive the .EU domain, https://sobre.arquivo.pt/en/a-first-attempt-to-archive-the-eu-domain/, accessed: 31 October 2022

[52] Hockx-Yu, H., Laursen, D. and Gomes, D., 2019. The curious case of archiving. eu. In The Historical Web and Digital Humanities (pp. 64-72). Routledge.

[53] Arquivo.pt, Cross-lingual collection about the 2019 European Elections is available, https://sobre.arquivo.pt/en/cross-lingual-collection-about-the-2019-european-elections-is-available/, accessed: 31 October 2022.

[54] Arquivo.pt, H2020 projects preserved by Arquivo.pt, https://sobre.arquivo.pt/en/h2020-projects-preserved-by-arquivo-pt/, accessed: 31 October 2022.

[55] Arquivo.pt, Search the Geocities history!, https://sobre.arquivo.pt/en/historical-collection-geocities-available-at-arquivo-pt/, accessed: 31 October 2022.

[56] Ben-David, A., 2019. 2014 not found: a cross-platform approach to retrospective web archiving. Internet Histories, 3(3-4), pp.316-342.

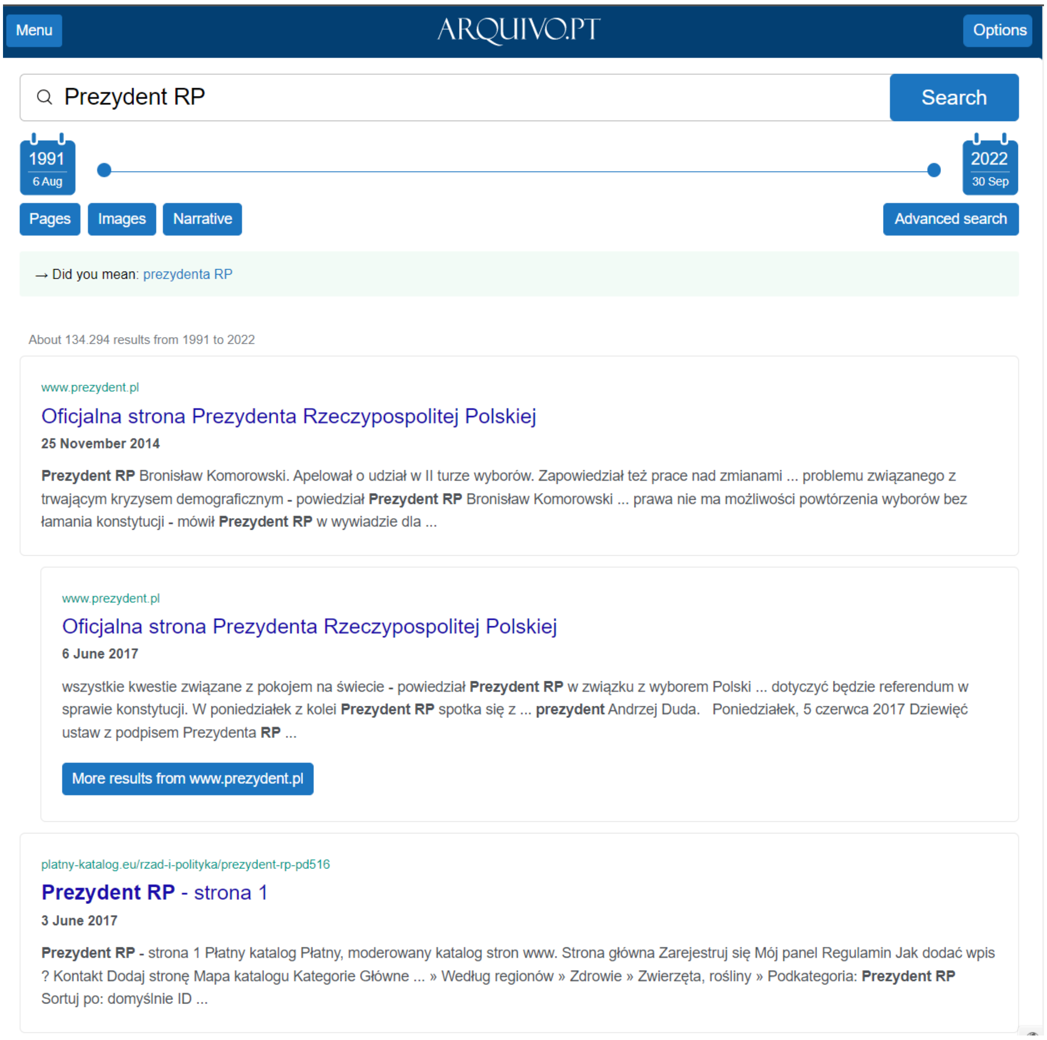
[57] Arquivo.pt, Open dataset about cryptocurrency, https://sobre.arquivo.pt/en/open-dataset-about-cryptocurrency/, accessed: 31 October 2022.

ended, or much more frequently, when scams were detected. Thus, Arquivo.pt created a collection to preserve web information which documents cryptocurrency activities with the objective of supporting future innovative studies in areas such as economy or digital humanities.

Search and access services

Preservation requires maintaining accessibility of information across time. If information cannot be efficiently searched and accessed, it is not preserved, only stored. Also, without the scrutiny of real usage, it is not possible to evaluate the relevance of the information being web-archived nor the quality of the provided services. Common web users are demanding but they must also be the target of web archives, because as the Internet penetrates societies, every human becomes a *web user*, from students to decision-makers. This Section presents the search and access services provided by Arquivo.pt.

Full-text search



→ Did you mean: prezydenta RP

About 134.294 results from 1991 to 2022

www.prezydent.pl

Oficjalna strona Prezydenta Rzeczypospolitej Polskiej

25 November 2014

Prezydent RP Bronisław Komorowski. Apelował o udział w II turze wyborów. Zapowiedział też prace nad zmianami ... problemu związanego z trwającym kryzysem demograficznym - powiedział **Prezydent RP** Bronisław Komorowski ... prawa nie ma możliwości powtórzenia wyborów bez łamania konstytucji - mówił **Prezydent RP** w wywiadzie dla ...

www.prezydent.pl

Oficjalna strona Prezydenta Rzeczypospolitej Polskiej

6 June 2017

wszystkie kwestie związane z pokojem na świecie - powiedział **Prezydent RP** w związku z wyborem Polski ... dotyczyć będzie referendum w sprawie konstytucji. W poniedziałek z kolei **Prezydent RP** spotka się z ... **prezydent** Andrzej Duda. Poniedziałek, 5 czerwca 2017 Dziewięć ustaw z podpisem Prezydenta **RP** ...

platny-katalog.eu/rzad-i-polityka/prezydent-rp-pd516

Prezydent RP - strona 1

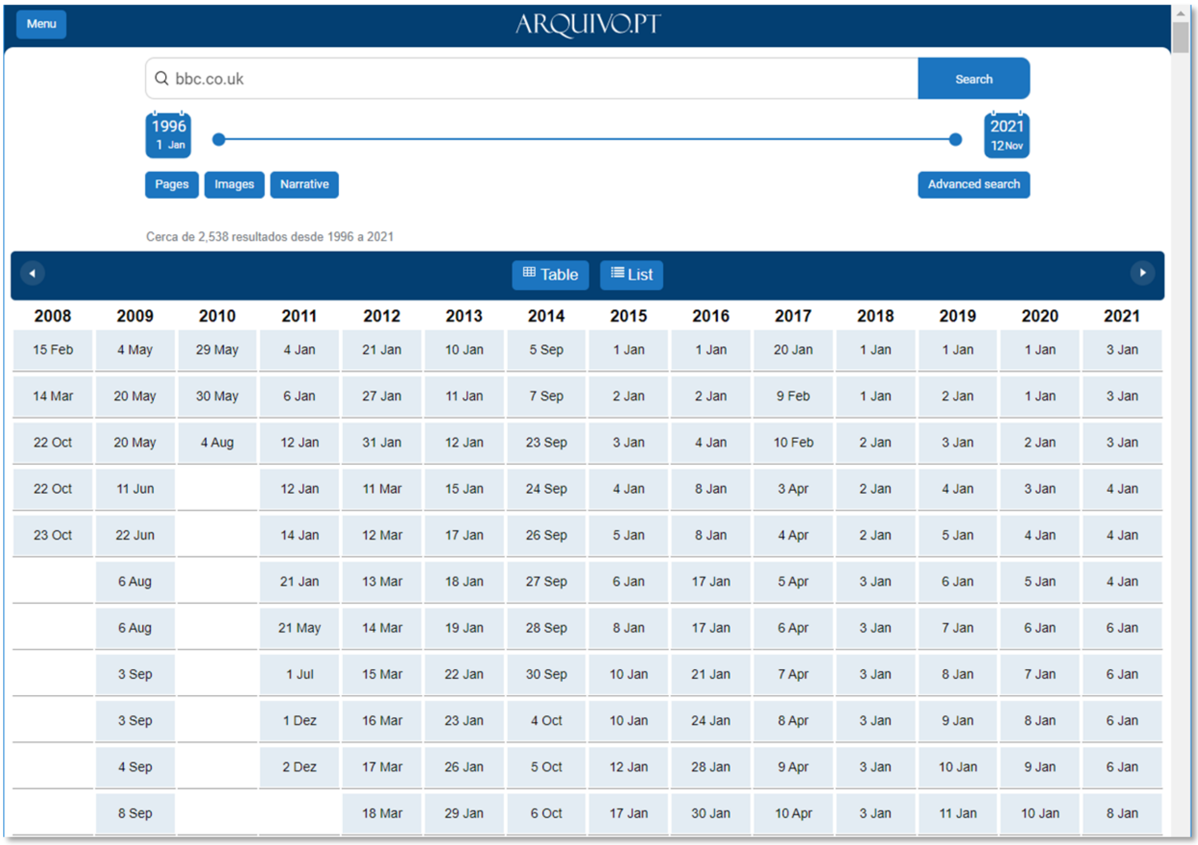
3 June 2017

Prezydent RP - strona 1 Płatny katalog Płatny, moderowany katalog stron www. Strona główna Zarejestruj się Mój panel Regulamin Jak dodać wpis ? Kontakt Dodaj stronę Mapa katalogu Kategorie Główne ... » Według regionów » Zdrowie » Zwierzęta, rośliny » Podkategoria: **Prezydent RP** Sortuj po: domyślnie ID ...

Figure 4: Page search results for query in Polish language "Prezydent RP".

Figure 4 presents the web user interface (UI) that displays results for searching text in any language among web-archived pages. This UI was designed to provide a look-and-feel similar to a live web search engine in order to facilitate its adoption by common web users. However, notice that the date picker components were discreetly integrated. They allow defining a time span for the search, that is, the interval between the date of crawl of the oldest and the most recent page that will be searched for the queried keywords. If the users change the time-span for the same keywords, they will obtain different results. For instance, if users search for a president of a given country, they will find distinct results as they choose different time spans. This is a useful and unique feature enabled by web archives because they preserve web information across time, complementing live web search engines which only support search over the most recent web data. A discussion on how to support full-text search and adapt user interfaces to the context of web archives can be found in previous works [58],[59].

Version history



bbc.co.uk

1996 1 Jan

2021 12 Nov

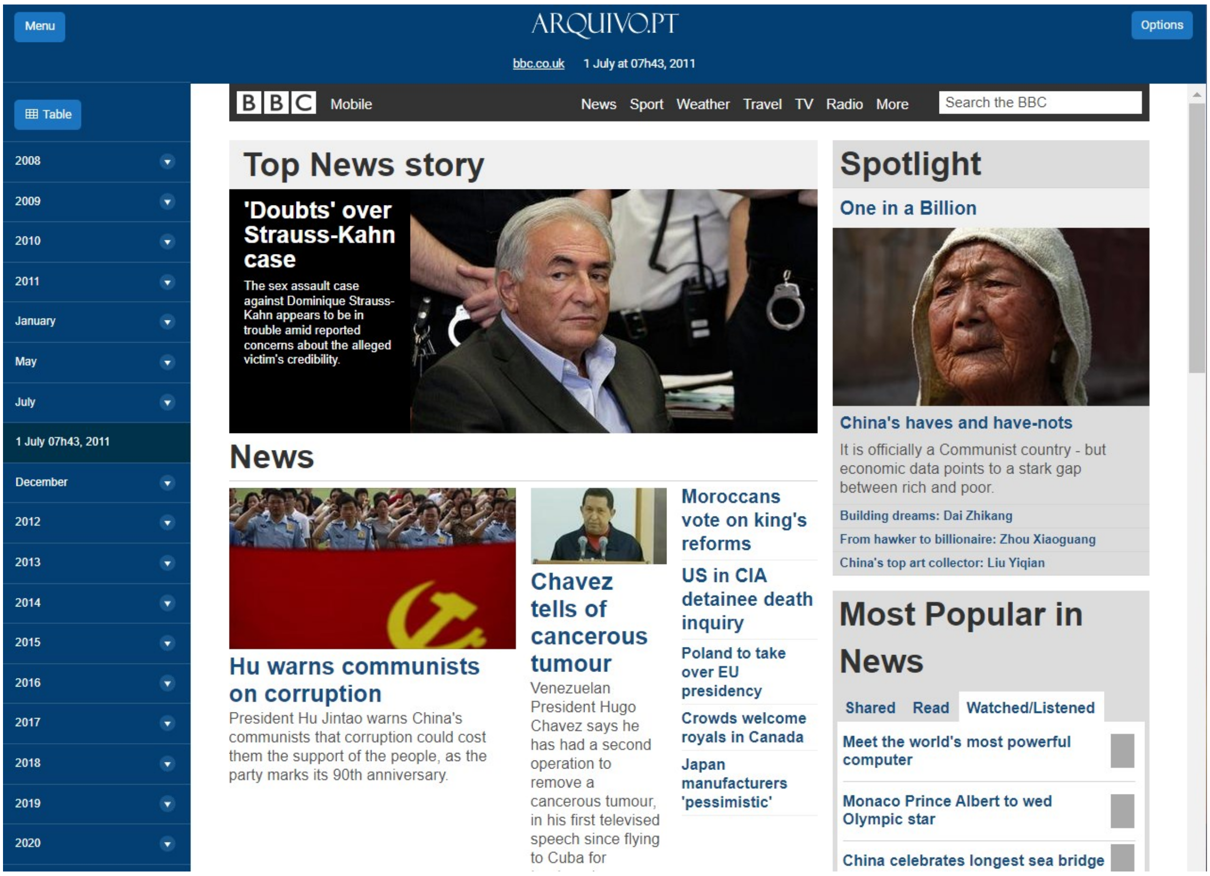
Cerca de 2,538 resultados desde 1996 a 2021

2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
15 Feb	4 May	29 May	4 Jan	21 Jan	10 Jan	5 Sep	1 Jan	1 Jan	20 Jan	1 Jan	1 Jan	1 Jan	3 Jan
14 Mar	20 May	30 May	6 Jan	27 Jan	11 Jan	7 Sep	2 Jan	2 Jan	9 Feb	1 Jan	2 Jan	1 Jan	3 Jan
22 Oct	20 May	4 Aug	12 Jan	31 Jan	12 Jan	23 Sep	3 Jan	4 Jan	10 Feb	2 Jan	3 Jan	2 Jan	3 Jan
22 Oct	11 Jun		12 Jan	11 Mar	15 Jan	24 Sep	4 Jan	8 Jan	3 Apr	2 Jan	4 Jan	3 Jan	4 Jan
23 Oct	22 Jun		14 Jan	12 Mar	17 Jan	26 Sep	5 Jan	8 Jan	4 Apr	2 Jan	5 Jan	4 Jan	4 Jan
	6 Aug		21 Jan	13 Mar	18 Jan	27 Sep	6 Jan	17 Jan	5 Apr	3 Jan	6 Jan	5 Jan	4 Jan
	6 Aug		21 May	14 Mar	19 Jan	28 Sep	8 Jan	17 Jan	6 Apr	3 Jan	7 Jan	6 Jan	6 Jan
	3 Sep		1 Jul	15 Mar	22 Jan	30 Sep	10 Jan	21 Jan	7 Apr	3 Jan	8 Jan	7 Jan	6 Jan
	3 Sep		1 Dez	16 Mar	23 Jan	4 Oct	10 Jan	24 Jan	8 Apr	3 Jan	9 Jan	8 Jan	6 Jan
	4 Sep		2 Dez	17 Mar	26 Jan	5 Oct	12 Jan	28 Jan	9 Apr	3 Jan	10 Jan	9 Jan	6 Jan
	8 Sep			18 Mar	29 Jan	6 Oct	17 Jan	30 Jan	10 Apr	3 Jan	11 Jan	10 Jan	8 Jan

[58] Cruz, D. and Gomes, D., 2013, September. Adapting search user interfaces to web archives. In Proc. of the 10th International Conference on Preservation of Digital Objects (Vol. 17).

[59] Gomes, D., Costa, M., Cruz, D., Miranda, J. and Fontes, S., 2013, May. Creating a billion-scale searchable web archive. In Proceedings of the 22nd International Conference on World Wide Web (pp. 1059-1066).

Figure 5: Table view that displays the versions web-archived from the URL bbc.co.uk over time.



Top News story

'Doubts' over Strauss-Kahn case

The sex assault case against Dominique Strauss-Kahn appears to be in trouble amid reported concerns about the alleged victim's credibility.

Spotlight

One in a Billion

China's haves and have-nots

It is officially a Communist country - but economic data points to a stark gap between rich and poor.

Building dreams: Dai Zhikang

From hawker to billionaire: Zhou Xiaoguang

China's top art collector: Liu Yiqian

News

Hu warns communists on corruption

President Hu Jintao warns China's communists that corruption could cost them the support of the people, as the party marks its 90th anniversary.

Chavez tells of cancerous tumour

Venezuelan President Hugo Chavez says he has had a second operation to remove a cancerous tumour, in his first televised speech since flying to Cuba for

Moroccans vote on king's reforms

US in CIA detainee death inquiry

Poland to take over EU presidency

Crowds welcome royals in Canada

Japan manufacturers 'pessimistic'

Most Popular in News

Meet the world's most powerful computer

Monaco Prince Albert to wed Olympic star

China celebrates longest sea bridge

Figure 6: The replay user interface reproduces the version web-archived from the URL bbc.co.uk on 1 July 07h43m, 2011.

Figure 5 presents the web user interface that enables accessing the history of a given web address (URL). If the users know the address of a given past web page, they can type its URL directly into the search box and they will receive a list of all the versions web-archived across time. If users select any of the listed dates, they will be redirected to the replay user interface presented in Figure 6, showing as an example the home page of the BBC website as archived on the 1th of July, 2011 at 07h43m. On the left sidebar of the replay UI, users can drill down the multiple web-archived versions and browse links as if they were back in the past. Notice that in order to enable this feature, the original code of the web pages must have been preserved, not just a screenshot of their layout. If the users reach a page that was not available in Arquivo.pt, a button "Search in other archives" is presented. If the users click on this button, they will be automatically redirected to an external web archive where the page may have been properly archived.

Image search

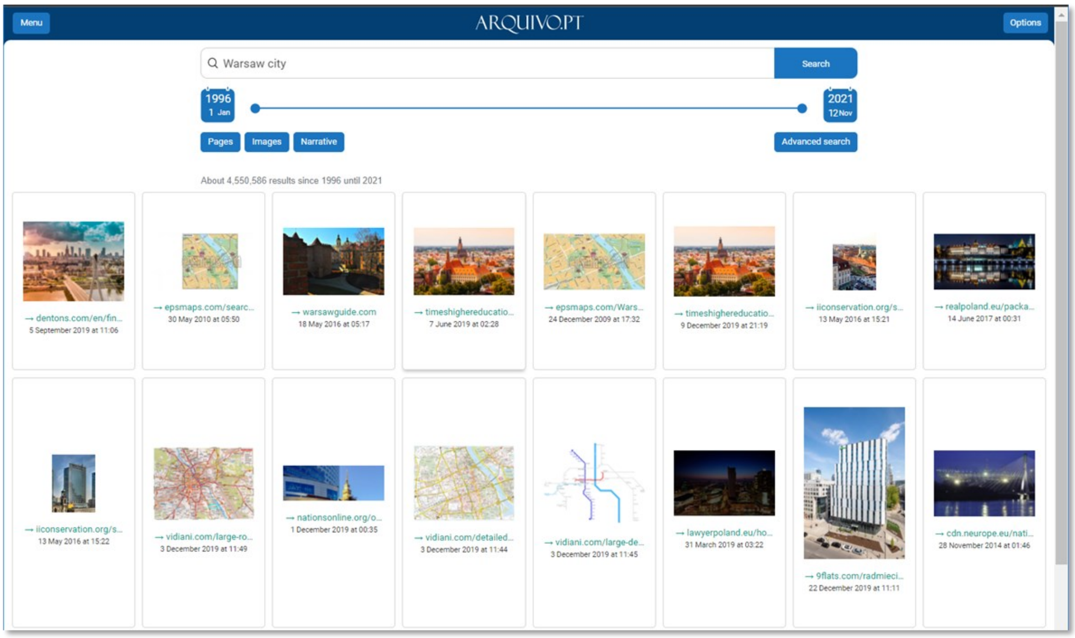
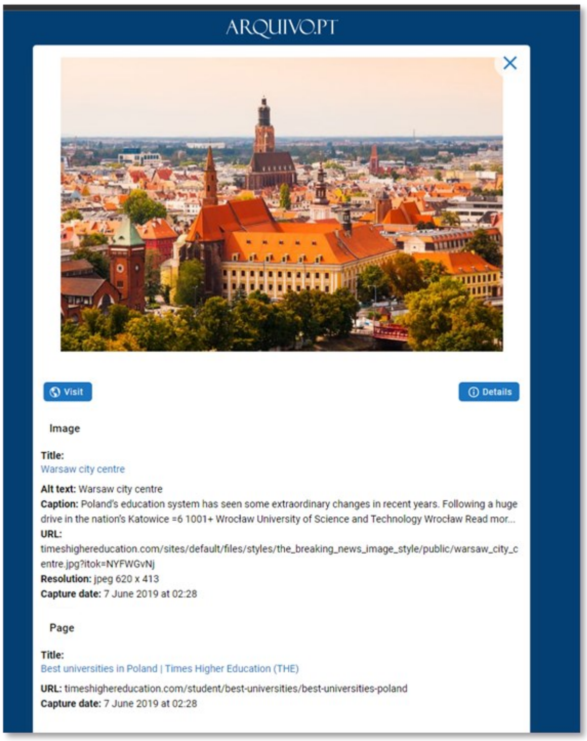


Figure 7: Image search results for query in English language "Warsaw city".



Image

Title:
Warsaw city centre

Alt text: Warsaw city centre
Caption: Poland's education system has seen some extraordinary changes in recent years. Following a huge drive in the nation's Katowice =6 1001+ Wrocław University of Science and Technology Wrocław Read mor...
URL:
timeshighereducation.com/sites/default/files/styles/the_breaking_news_image_style/public/warsaw_city_centre.jpg?itok=NYFWGvNj
Resolution: jpeg 620 x 413
Capture date: 7 June 2019 at 02:28

Page

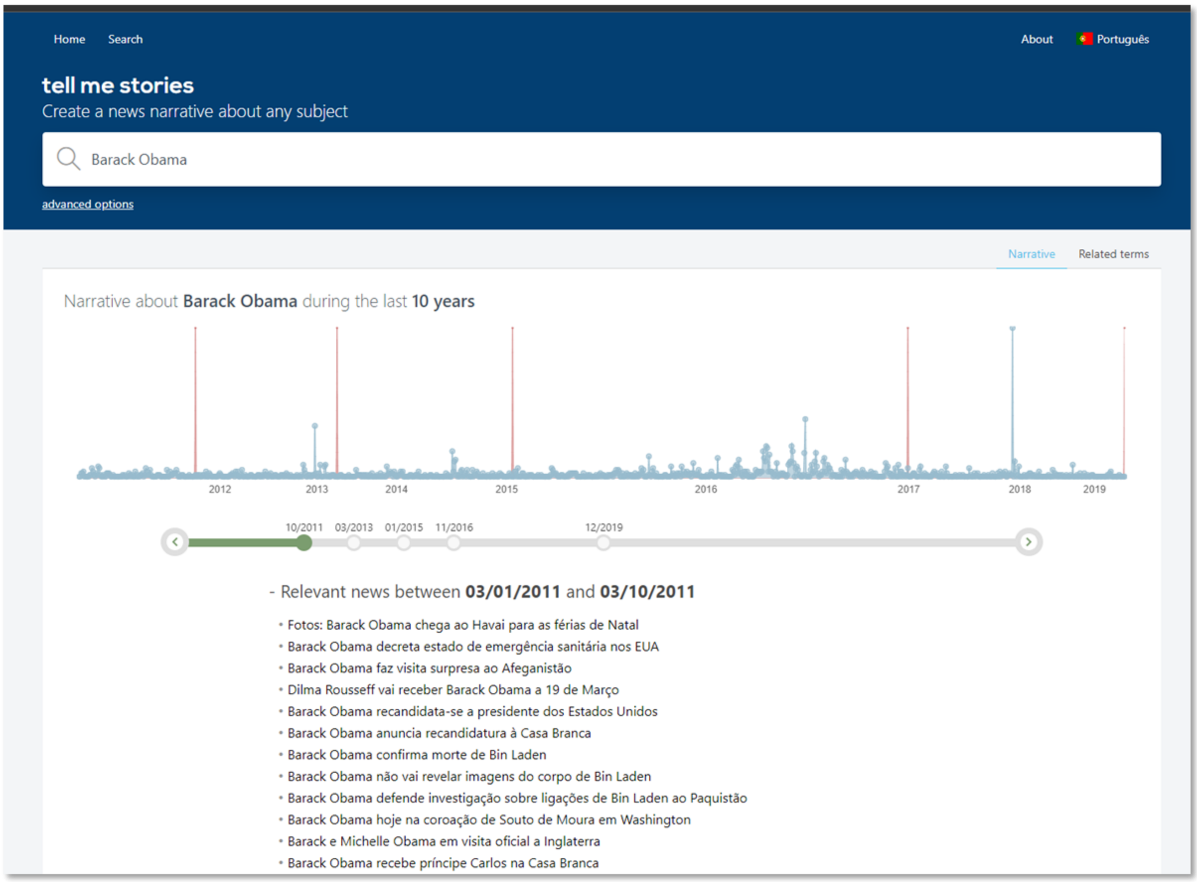
Title:
Best universities in Poland | Times Higher Education (THE)

URL: timeshighereducation.com/student/best-universities/best-universities-poland
Capture date: 7 June 2019 at 02:28

Figure 8: User interface that presents the details about an image returned on the image search results about "Warsaw city".

Figure 7 presents the web user interface that supports searching images from the past web. It shows, as an example, the web-archived images found for the query "Warsaw city". If users select one of the image results, they will receive related metadata, which was automatically generated for the image (Figure 8), including a link to the web-archived page that embedded the image, which is very important to interpret the context of its publication. The design and implementation of this innovative image search system over historical web-archived images is described in detail in a previous work[60].

Automatic Programming Interfaces and Narratives of news



tell me stories

Create a news narrative about any subject

Narrative about **Barack Obama** during the last **10 years**

- Relevant news between **03/01/2011** and **03/10/2011**

- Fotos: Barack Obama chega ao Havai para as férias de Natal
- Barack Obama decreta estado de emergência sanitária nos EUA
- Barack Obama faz visita surpresa ao Afeganistão
- Dilma Rousseff vai receber Barack Obama a 19 de Março
- Barack Obama recandidata-se a presidente dos Estados Unidos
- Barack Obama anuncia recandidatura à Casa Branca
- Barack Obama confirma morte de Bin Laden
- Barack Obama não vai revelar imagens do corpo de Bin Laden
- Barack Obama defende investigação sobre ligações de Bin Laden ao Paquistão
- Barack Obama hoje na coroação de Souto de Moura em Washington
- Barack e Michelle Obama em visita oficial a Inglaterra
- Barack Obama recebe príncipe Carlos na Casa Branca

Figure 9: Automatically generated narratives about any subject, based on online news from the past. Winners of the Arquivo.pt Award 2018 (https://arquivo.pt/winners2018).

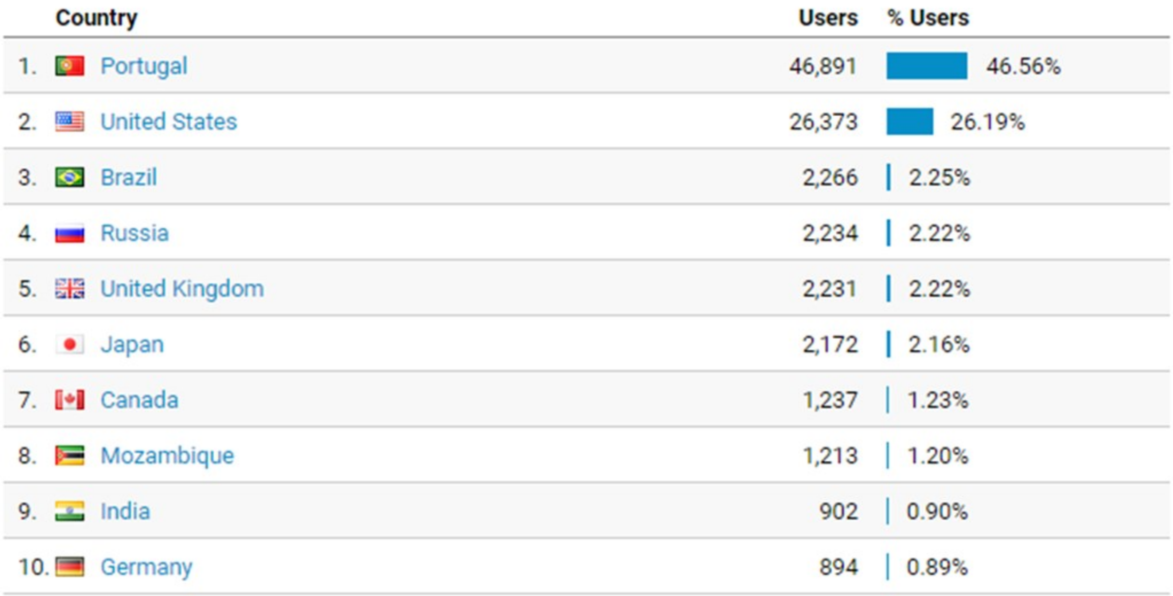
Arquivo.pt is a research infrastructure and it will never be able to fulfil all the specific requirements of the users from distinct research areas. Thus, it provides application programming interfaces so that external developers and researchers can develop their own specific applications to process the historical web data preserved in Arquivo.pt. It provides two

[60] Mourão, A. and Gomes, D., 2021, The Anatomy of a Web Archive Image Search Engine-Technical Report, https://sobre.arquivo.pt/wp-content/uploads/The_Anatomy_of_a_Web_Archive_Image_Search_Engine_tech_report-1.pdf

APIs which are specifically designed to enable full exploration of its services and features (Arquivo.pt API and Image Search API)[61]. Arquivo.pt also supports the Memento API which is based on the Memento protocol and the CDX-server API which is not based on a formal international protocol, but is nonetheless widely supported by most web archives. This way, Arquivo.pt can interoperate with other web archives and the applications developed to process data from other archives can quickly be reused.

Figure 9 presents an example of one application developed using the Arquivo.pt API named ContaMeHistoria.pt (Tell Me Stories), an online service that automatically generates narratives about any subject based on online news from the past. This service helps users to quickly grasp news information about any entity or event across time. This service was developed by external researchers who won the Arquivo.pt Award in 2018. Later in 2021, the TellmeStories service was connected to the Arquivo.pt website, through the integration of the "Narrative" button on the page search user interface (see Figure 4).

International usage statistics



	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%
9.	India	902	0.90%
10.	Germany	894	0.89%

Figure 10: Top countries of origin of Arquivo.pt users in 2021.

Figure 10 presents the top 10 countries from which the users of Arquivo.pt originated in 2021, and it shows that 53% came from countries other than Portugal. Providing user interfaces and documentation also in English contributes to this international usage. The development and widespread usage of automatic translation tools, such as Google Translate, also contributed to breaking down the language barriers, enabling international users to interpret web-archived

[61] Arquivo.pt, Arquivo.pt Application Programming Interfaces (APIs), https://arquivo.pt/api, accessed 31 October 2022

information originally published in languages they could not understand, such as Portuguese. For instance, a Polish researcher can access a web-archived page about the Polish elections, originally written in Portuguese, and apply automatic translation to study how Polish elections were documented from a Portuguese perspective. Arquivo.pt received 1 million users and served 10 million pages in the period from January 2012 until August 2022. These absolute figures may seem humble in comparison to other online services. However, we must consider that web archive users recover unique historical online information that otherwise would have been irremediably lost forever.

Complementary services to expand the scope of web archives use

Web archives have the noble mission of preserving historical web documents for future access. However, if web archives are to be established as infrastructure for digital societies, it is crucial that they also provide general-purpose services that would enable citizens and organisations to easily take advantage of web archives for their own benefit in solving everyday life situations. This section presents additional services that Arquivo.pt has been launching to engage users in four tasks of web data preservation: acquisition (SavePageNow), curation (Complete page), storing (Memorial) and access (Arquivo404).

SavePageNow: archive a web page immediately

The process of acquiring web data to be preserved is mainly automatic due to its large volume and transience. However, fully automated acquisition leads to two problems. The first problem is that even with most carefully configured, monitored and tuned crawlers, web archives waste resources storing web content that is not of primary interest for preservation of cultural heritage, such as spam websites, link farms or domain parking pages. The second problem is that at the same time, information that would have been very valuable to be preserved is not found in time by the automatic processes used by web archives and disappears before it can be acquired. Therefore, Arquivo.pt provides a public form where any web user can suggest a website to be preserved[62]. The users only need to submit the address of the homepage and optionally provide an email, so that they can be notified when the suggested website becomes available at Arquivo.pt, and assess the quality of the web-archived content. These user suggestions are precious and instrumental in identifying relevant websites hosted outside the .PT domain or public pages on social networks, for instance those belonging to governmental institutions. However, the process of crawling a website after it was suggested is not immediate and may take months before it is executed. Moreover, the suggested websites are acquired using conventional crawlers which quickly collect large amounts of information but sometimes miss rich media, such as embedded videos, or fail to access information published through complex

[62] Arquivo.pt, Suggest websites to be preserved - Collaborate, https://arquivo.pt/suggest accessed 31 October 2022.

web applications such as social network platforms. Meanwhile the information that the users wanted to preserve may have changed or even disappeared. Thus, Arquivo.pt experimentally launched the SavePageNow, which allows users to immediately save a set of web pages in Arquivo.pt using a browser-based crawler, which despite being slower in comparison to conventional crawlers, enables a higher quality acquisition of web data. The users only need to enter a page's address and start browsing for all the visited content to be archived. The Arquivo.pt SavePageNow[63] was inspired by the Internet Archive Save Page Now[64] and implemented using the recording feature of pywb[65]. The Internet Archive service offers more features, such as saving outgoing links or screenshots and provides immediate access to the web-archived content. In turn, Arquivo.pt SavePageNow archives all the pages browsed during a session, which facilitates complete archivisation of a small website, carried out autonomously by the users. The experimental launch of the Arquivo.pt Save Page Now service began on the 30th of November 2021 and lasted for 10 months. During this period, the service was accessed 7 119 times (average of 711,9 accesses per month) and its users selected and archived a total of 34.250 million URLs (3.7 TB in WARC format). Please note that each time a user saves one webpage, its several embedded URLs are also automatically archived.

Complete page: recover missing information from external sources

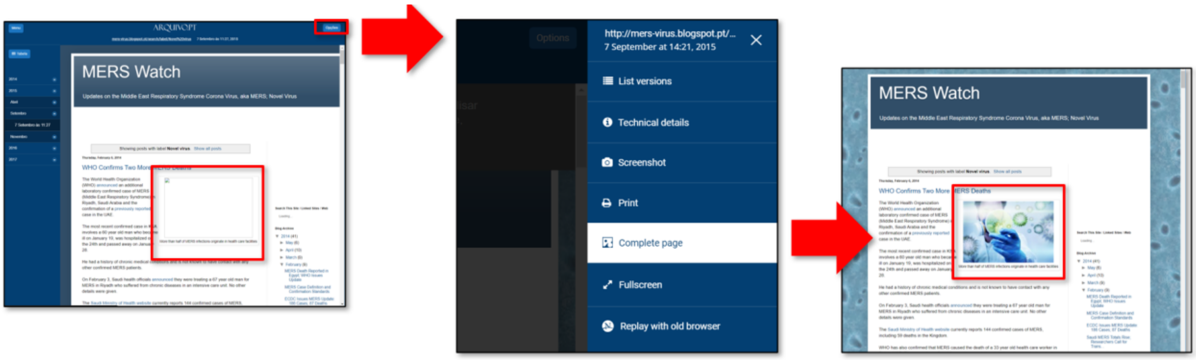

MERS Watch
Updates on the Middle East Respiratory Syndrome Corona Virus, aka MERS, Novel Virus
http://mers-virus.blogspot.pt/...
7 September at 14:21, 2015
List versions
Technical details
Screenshot
Print
Complete page
Fullscreen
Replay with old browser
MERS Watch
Updates on the Middle East Respiratory Syndrome Corona Virus, aka MERS, Novel Virus

Figure 11. The "Complete page" function available as an option at the replay user interface automatically looks for missing information in external web archives and the live web. *The obtained information is later integrated in Arquivo.pt collections and becomes available for all the users.*

Web archives try to exhaustively acquire all the content required to enable reproducing the archived web pages as faithfully to the original ones as possible, including for example embedded images or CSS stylesheets. Still, frequently some content is missing. Therefore, continuous curation of the archived web pages is necessary to evaluate the quality of acquired web data and identify missing content. Please note, however, that even if a resource is missing

[63] Arquivo.pt, SavePageNow, https://arquivo.pt/savepagenow, accessed 31 October 2022.

[64] Internet Archive, Wayback Machine Save Page Now, https://web.archive.org/save/

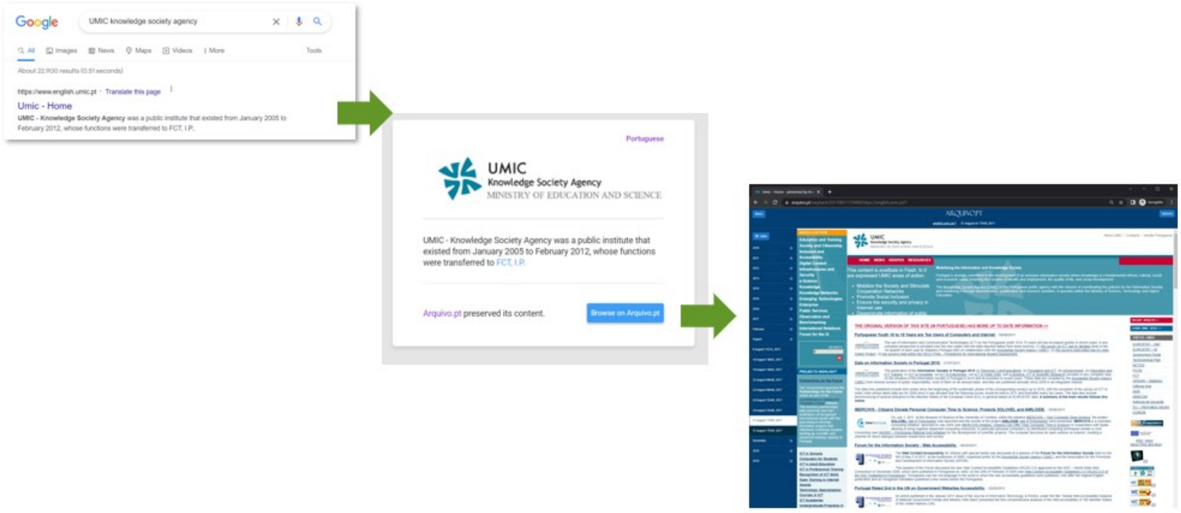
[65] Pywb, Configuring the Web Archive — pywb 2.0 documentation, https://pywb.readthedocs.io/en/latest/manual/configuring.html#recording-mode, accessed 31 October 2022.

in a web archive, it could still be available on the original website if it is recent, or at an external web archive if it is older. The process of looking for missing information can be automated through the Memento protocol. Thus, Arquivo.pt provides the “Complete page” option, available at the replay user interface described in Figure 11. When the user selects this option, Arquivo.pt automatically looks for the information missing on the reproduced web-archived page in external web archives and on the live web. The obtained information is subsequently integrated in Arquivo.pt collections and becomes available for all the users. The process of completing pages cannot be automatically applied to all the web-archived pages because looking for information in external sources is slow and a high number of automatic requests could overload these sources. “Complete page” engages the users in curation of the web-archived collections by collaboratively selecting relevant pages with missing content and trying to enhance only these ones. However, this process of completing web-archived pages can result in temporal incoherence issues, for instance by integrating an embedded image that was published on the given web page later, leading to reproduction of a combination of a web-archived page with an embedded image that was not available online at the time of archiving. It is impossible to entirely prevent such issues from ever occurring, but they can be identified by inspecting the metadata of the web-archived resources, available through the “Technical details” option of the replay interface or by applying automatic algorithms[66],[67]. The “Complete page” function was launched in November 2016 and as of August 2022 it enabled integration of 15.2 million missing files (603.4 GB in WARC format).

[66] Spaniol, M., Mazeika, A., Denev, D. and Weikum, G., 2009, September. Catch me if you can: Visual analysis of coherence defects in web archiving. In 9th International Web Archiving Workshop (IWAW 2009), Corfu, Greece (pp. 27-37).

[67] Ainsworth, S.G., Nelson, M.L. and de Sompel, H.V., 2015. Evaluating the Temporal Coherence of Archived Pages.

Memorial: preserving information from old websites



About 22,900 results (0.51 seconds)

https://www.english.umic.pt · Translate this page

Umic - Home

UMIC - Knowledge Society Agency was a public institute that existed from January 2005 to February 2012, whose functions were transferred to FCT, I.P..

UMIC - Knowledge Society Agency was a public institute that existed from January 2005 to February 2012, whose functions were transferred to FCT, I.P.

Arquivo.pt preserved its content.

Figure 12: The website of the UMIC - Knowledge Society Agency (english.umic.pt) was deactivated in 2017 but its content was preserved in the Arquivo.pt Memorial and continues to be searchable through live web search engines.

There are many historical websites that provide valuable information but are no longer updated (for example websites of finished projects) and organisations must invest significant amounts of resources to keep them online. However, as websites become older, costs increase because of obsolescence of supporting technologies and consequent dangerous security vulnerabilities. The Arquivo.pt Memorial offers high-quality storing of websites' content with the possibility of maintaining their original domains[68]. Figure 12 describes an example of a website preserved in the Arquivo.pt Memorial. The website https://english.umic.pt belonged to a governmental organisation named "UMIC – Knowledge Society Agency" that terminated its activities in 2012. Arquivo.pt stored the information published online by this website, and then it was shut down by its owners.

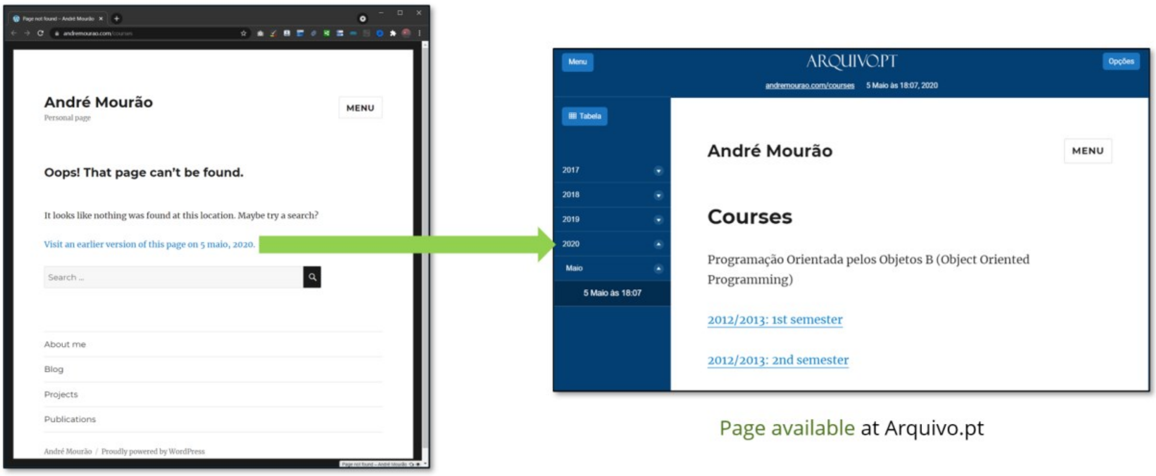
The Arquivo.pt Memorial service preserves the content of a website in high quality by applying an iterative curation process and several acquisition technologies. By maintaining the original domain name of the website (e.g. english.umic.pt), the website contents remain searchable through live web search engines (e.g. search for "UMIC – Knowledge Society Agency"). The links to internal pages on the website are also redirected to the correspondent web-archived

[68] Arquivo.pt, Arquivo.pt Memorial: preserves information of historical websites, https://arquivo.pt/memorialen, accessed 31 October 2022.

pages to avoid occurrence of broken links from external pages. For example, the links to the inner webpage “UMIC Education and Training”[69] are resolved to its web-archived version[70].

By using the Arquivo.pt Memorial, website owners can preserve their online information and don’t have to continuously invest in infrastructure (e.g. servers, electricity, content management system) nor allocate IT staff to maintain their historical websites. They only have to maintain the ownership of the domain name. Notice that it must be made clear to the website owners that the Memorial is not a hosting service. It stores the content of the website and does not allow any updates afterwards. Integrating a website in the Memorial is done collaboratively with the website owners, as it requires an iterative process of acquiring and validating whether the web-archived content was stored satisfactorily, or whether it requires further data acquisition efforts. The “Save Page Now” and “Complete page” functions contribute to this process because they enable the website owners to autonomously acquire missing content to be archived.

Arquivo404: fix broken links



André Mourão
Personal page

Oops! That page can’t be found.

It looks like nothing was found at this location. Maybe try a search?

Visit an earlier version of this page on 5 maio, 2020.

About me
Blog
Projects
Publications

André Mourão / Proudly powered by WordPress

andremourao.com/courses 5 Maio às 18:07, 2020

André Mourão

Courses

Programação Orientada pelos Objetos B (Object Oriented Programming)

2012/2013: 1st semester

2012/2013: 2nd semester

Page available at Arquivo.pt

Page not found at live website

Figure 13: Arquivo404 is a Javascript code to be installed on the "404 - Page not found" error pages that mitigates broken links. If a link is broken, Arquivo404 generates a message that suggests an alternative link to a corresponding web-archived version.

Link rot has been a prevalent problem since the early days of the web[71]. As websites evolve over time, some of their URLs, which used to reference valid information, become broken.

[69] http://www.english.umic.pt/index.php?option=com_content&task=section&id=10&Itemid=86

[70] https://arquivo.pt/wayback/20170831190220/http://www.english.umic.pt/index.php?option=com_content&task=section&id=10&Itemid=86

[71] Gomes, D. and Silva, M.J., 2006, July. Modelling information persistence on the web. In Proceedings of the 6th international conference on Web engineering (pp. 193-200).

When this happens, links, users' bookmarks, paper citations and other online references start returning a 404 error page instead. The result is that the users of websites get frustrated when they receive a dead-end error message instead of the page they desired to visit. The Arquivo404 service aims to mitigate this problem (see Figure 13)[72]. It is a free, open-source project that improves soft-error pages on any website by providing the link to a web-archived version of the missing page. The website owner just needs to insert a single line of code in the page that generates the 404 error message. Arquivo404 uses the Memento protocol to check for archived versions, allowing any Memento compliant web archives to also be searched. When a user tries to access a page that is no longer available on a website, Arquivo404 automatically checks if there is a version of that webpage preserved in Arquivo.pt. If the webpage is archived in Arquivo.pt, a link is presented so that the user may visit this version. If it was not archived, the pre-defined error page is displayed. We believe that Arquivo404 is an enhancement to any website, benefiting owners and users at virtually no cost. At the same time, it raises awareness of importance of web archiving to a wider public. Arquivo404 was experimentally tested in production for a period of 8 months from November 30th, 2021 until September 30th, 2022. During this period, the Arquivo404 service was accessed on average 1 849 times per month and no malfunctions were detected.

Explore the Past Web

Arquivo.pt preserves an ever growing number of millions of websites accessible and searchable by text, image and URL. However, open-access and searchability by themselves do not raise community's interest. Web archivists must demonstrate the potential and even the charm of historical web information. It is necessary to capture the attention of target communities to use web archives for research and valorisation of historical web-data. This section presents 3 examples of concrete initiatives undertaken to promote the exploration of historical web information.

Online exhibitions

Thematic exhibitions and collaborative collections are important to illustrate the utility of web archives as a source of historical documentation. Arquivo.pt has been creating online exhibitions organised by theme, such as press, radio, municipalities, R&D units, schools or museums[73]. Each exhibition is followed by a dissemination campaign to motivate collaborations with external organisations related to the exhibition theme. For example, the

[72] Arquivo.pt, Put an end to "page not found" on your website, https://arquivo.pt/arquivo404en, accessed 31 October 2022.

[73] Arquivo.pt, Exhibitions, https://arquivo.pt/exhibitions/, accessed 31 October 2022.

exhibition “Memory of Art Festivals and Events”[74] is the result of a collaboration with the Calouste Gulbenkian Foundation Art Library, a leading institution on Art in Portugal, and ROSSIO, a Digital Humanities infrastructure led by the Faculty of Social Sciences of the New University of Lisbon. This initiative generated interaction with the artists' community and contributed to the improvement of the preservation of Art events and their corresponding web documents. It adopted low-cost tools, such as a free account on Wordpress.com, accessible to any person or organisation.

Publication of open data sets of derived results

During the operation of Arquivo.pt activities, metadata that can be useful for third parties is generated, such as lists of URLs concerning given topics. Let’s consider the example of websites about Research and Development (R&D) projects which are increasingly being used to publish important scientific information that complements published literature (e.g. data sets, documentation or software). After the end of the projects the corresponding websites usually disappear, resulting in permanent loss of unique and valuable scientific information. Moreover, the references to online information about R&D projects are not being fully documented. For example, the URLs for projects funded by the 7th Framework Program (FP7) available at the European Union’s Open Data Portal were missing for 92% of the projects. Arquivo.pt identified and created a list of 53 993 URLs that document R&D projects financed since the FP4 (1994). This list was applied to the data sets published through the European Open Data Portal to complement the missing information regarding project URLs and the completeness of the FP7 data set was improved by 86.6%. All the resulting data sets were made publicly available so that they can be improved and reused by other organisations also interested in preserving this digital heritage.

The Administrative Modernization Agency, IP (AMA) is the public institute that carries out the duties in the areas of administrative modernization and simplification, and electronic administration. Arquivo.pt has been collaborating with AMA with the aim of improving the preservation of Public Administration websites. AMA recognized Arquivo.pt as a public service and open data provider and awarded its certification seal on the Open Data Portal. So far, Arquivo.pt published 39 open datasets derived from its activities so that they can be reused[75].

[74] Biblioteca de Arte da Fundação Calouste Gulbenkian, arteparasempre.wordpress.com…, accessed 31 October 2022.

[75] Dados.gov.pt - Portal de dados abertos da Administração Pública, Arquivo.pt - pesquise páginas do passado, https://arquivo.pt/dadosabertos, accessed 31 October 2022.

Training in scope of web preservation

Arquivo.pt issued a set of recommendations with regard to developing preservable sites[76] and has been promoting a free training programme[77]. The main objectives of the training programme are to raise awareness about the importance of preserving the digital heritage published online and maximise users' productivity when using the services provided by Arquivo.pt. Acquisition of skills regarding web preservation by professionals, such as web developers, enables them to deliver higher quality solutions to their clients or users, for example by developing websites resistant to link rot by implementing the Arquivo404 mechanism. The training programme comprises 4 modules:

- *New ways of searching the past (module A)*: presents the services provided by Arquivo.pt so that any Internet user may take advantage of historical web data for professional and personal usage;
- *Well publish to well preserve (module B)*: provides recommendations and techniques for web publishing so that online information is not lost and can be preserved for future access;
- *Automatic processing of information preserved from the Web (module C)*: presents methods and technologies to develop web applications that automatically access and process information preserved from the Web, for instance using Arquivo.pt APIs;
- *Web archiving: Do-it-yourself! (module D)*: teaches how anyone, not being IT experts nor archivists, can adequately acquire, store and replay web-content.

The training programme was launched in 2018 and has been delivered as in-person or remote sessions. Throughout 2021, 27 sessions were delivered to at least 1 242 trainees (we missed counting in some sessions). The average satisfaction, measured through anonymous questionnaires filled in by the trainees, was on a level of 86%. All the teaching materials, including slides and videos, are freely available for reuse. The Youtube training playlist includes 71 videos in Portuguese and 30 in English.

It has been easy to demonstrate usefulness of web archives because every live web use case has the potential of becoming a web-archive use case when we apply a historical perspective. For instance, users access the live web to get information about a current event but must access a web archive to get information about similar past events. Therefore, the amount of potential use cases that a web archive may support is as vast as those that the live web supports. Web archivists should explore this fact to present potential use cases meaningful to different audiences of trainees, to emphasize the importance of web archiving. Web archivists must identify and engage their existing communities and address their needs so that they also become ambassadors of web archiving within their own organisations and specific communities. The

[76] Arquivo.pt, Recommendations for authors to enable web archiving, https://arquivo.pt/recommendations, accessed 31 October 2022.

[77] Arquivo.pt, Training courses, https://arquivo.pt/training, accessed 31 October 2022.

Arquivo.pt training programme produced documentation that can be easily reused or adapted by educators; it includes presentations, texts and videos. Recent books contributed to launching the web archiving educational efforts, but they still need to be contextualised. Our strategy has been to start by reaching the education community, teach the teachers, so that they may teach their students about the importance of preserving online information and the existing web archiving services.

Challenges and recommendations regarding the creation of a web archive

Web archiving is a difficult task that must address problems in multiple areas. This section discusses broad challenges inherent in commencing web archiving activities and proposes guidelines to address them.

Lack of awareness of the value of preserving online information

The first and most daunting challenge is the lack of awareness with regard to the value of online information. Humans are the dominant species on Earth but they are not physiologically superior in comparison to other species, as discussed in the book Sapiens[78]. Our advantage comes from our unique capacity of collaborating at large scale to reach common goals even with distant individuals that we have never met. Humans invented ways to organise themselves to reach common goals through information sharing. However, notice that most of the information used to organise our digital societies is published exclusively online. Researchers are aware of the value of historical information. The problem is how to direct the attention of common citizens and decision-makers to the problems that arise from not preserving online information. Notice that without this awareness, one may ask why are web archives required when people are already overwhelmed by the tremendous amount of available online information? Web archives must first invest in raising awareness of web data transience problem, and only then, of the web archiving services they provide. In a nutshell, advertising web archives is not enough by itself, as most people are not even aware that online information, despite being widely available, is extremely ephemeral.

It would be extremely useful if an economics scholar systematically calculated the value of online information that has been vanishing from the web, so that web archivists could present monetary values to decision makers. However, despite the pertinence of this research topic, I could not find any such works. Thus, I'll take the liberty to present an oversimplified but meaningful estimation of the investment made by societies, including governments, companies and individuals, to produce the information preserved in Arquivo.pt. The objective is to quantitatively illustrate the value of preserving online information, or the value wasted by not

[78] Harari, Y.N., 2014. Sapiens: A brief history of humankind. Random House.

doing it. A website is basically a front-end used to publish information and make it available worldwide. Each website preserved in Arquivo.pt comprises on average 55 web pages[79]. Upwork is one of the largest freelance marketplace platforms in the world. According to Upwork, the average cost of building a medium size website (25 to 75 pages) in 2022 ranged from 10 000 to 35 000 USD[80]. Notice that in the early 2000s these values were much higher because powerful web publication platforms such as Wordpress did not exist, and expert software engineers had to be hired to develop any simple HTML website. For simplicity, let's assume that it was necessary to invest 20 000 USD to build each website preserved in Arquivo.pt. In 2022, Arquivo.pt held 28 million websites published since the 1990s. Based on our previous assumptions, the investment made to build these websites was 560 billion USD (the Gross Domestic Product of Portugal in 2021 amounted to 250 billion USD). 560 billion USD is a rough estimation of the investments made to produce the web data preserved by Arquivo.pt, which otherwise would have been lost forever. The real economic value of web-archived information is complex to estimate. While some information loses value as it becomes obsolete, other information becomes more valuable with time, as it provides unique historical perspectives that enable observation of trends. However, it is undeniable that digital societies strongly invest in producing online information. The important question that must be constantly raised is: can the societies afford to continue wasting their online information?

Hire, train and retain web archivists

A web archive is a web-based information system. A web archive collects information from the web, archives it and then provides access to it, typically through web user interfaces. Therefore, it requires hiring experts in web technologies, who are not abundant in the work market. Moreover, web archiving is Big Data, so web archives must compete with the Internet giants (e.g. Google, Facebook, Amazon) to hire some of these professionals. Also consider that the Web keeps evolving along with the technologies that support it, so web archive staff must have knowledge about both current and past web technologies. Web archiving staff requires cutting-edge web experts with historian mentality, and such mindset is not taught at any course. Therefore, hiring skilled staff to develop and operate a web archive is a major challenge.

Web archiving is complex and requires full-time dedication. Part-time workers in web archiving will struggle to achieve the required expertise to create a running service. Start with a small but autonomous web archiving team so that it does not have to internally compete for resources and can be fully dedicated to learn about web archiving. As web-archiving is not widely taught, you must prepare to efficiently provide training to newbies or external parties. Your trainees may become your next employees. Hiring and training staff to become web archivists is hard. Retaining such staff for a long time is even harder. Thus, accept staff

[79] Miranda, J. and Gomes, D., 2009, November. Trends in Web characteristics. In 2009 Latin American Web Congress (pp. 146-153). IEEE.

[80] Upwork, How Much Does It Cost To Build a Website? (2022 Data), https://www.upwork.com/resources/how-much-does-it-cost-to-build-website, accessed 31 October 2022.

rotation and proactively prepare for it. Make small but solid steps in development by producing thorough technical documentation, constantly reviewing it and establishing test and maintenance procedures.

Concerns about Intellectual Property Rights

Despite the efforts to manage digital rights through initiatives such as Creative Commons, the problem of addressing Intellectual Property Rights (IPR) of online artefacts, such as copyright, is far from being solved. Obtaining initial information to address IPR, such as who is the author or which usage rights apply, is an impossible task for most resources available online. Web archives, as the digital preservation services that they are, preserve the original artefacts as faithfully as possible along with all their virtues and problems. Thus, IPR issues about online artefacts persist after they are web-archived and it becomes even harder to address them as time passes. Maybe in the future, societies will develop effective models to manage IPR for online resources. However in 2022, this is the scenario that web archivists must address, and it may cause difficulties in countries or institutions which do not have supportive legal frameworks.

Nonetheless, if someone wants to start a web archiving initiative while minimising concerns with regard to legal issues, there is a solution. Start by archiving public administration websites, as by default, their published information is and must be openly available. In Europe, preserving the information published online is an obligation for public sector institutions. The Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information[81] stipulates that the documents, including those published on websites, must be openly available including for automatic processing and reuse. The data published in public sector websites is also mostly relevant to be preserved because it focuses on official government communications or was produced by public institutions such as universities. This fact saves initial effort on selection activities because automatic criteria can be applied, such as selecting all websites under a specified governmental domain (e.g. .GOV.PL), and prevents falling into spam websites that cause waste of precious resources (i.e. time, bandwidth or disk space). Notice that IPR rights applicable to public sector information published on official websites may differ from those applicable to the same information published on private platforms (e.g. social network platforms) because these platforms impose their own Terms and Conditions of usage, may be subject to the legislation of a foreign country or may disclose non-public information such as users' comments. Once again, this is a current societal problem that web archives cannot solve. However, web archivists should organise web-archived information so that public-access websites can be easily and automatically segregated from information published in private platforms. As an example of how to establish official

[81] Publications Office of the EU, Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, Official Journal of the European Union, http://data.europa.eu/eli/dir/2019/1024/oj, accessed 31 October 2022.

collaborations between public sector institutions and web archives, Arquivo.pt has been collaborating by providing training to public servants and preserving public administration websites. The collaboration during the lifecycle of a public sector website and Arquivo.pt involves following preservation recommendations when developing a new website, suggesting the website to start being web-archived when it is launched and at the end of its life, preserving its content in the Memorial. This way, any citizen can access the open data from these historical archives and, for example, search for official information published by successive governments across time.

Full preservation workflow

Web archiving, even at a small scale, requires a significantly large amount of resources. So, it's best to start by web-archiving a short list of websites, possibly even just the website of your own organisation, but perform the full preservation cycle: collect, store, provide access, disseminate and evaluate obtained results. Start small but do it all. In order to request additional resources to advance to the next stage, it is necessary to clearly demonstrate results. Currently, there are tools available (e.g. webrecorder.net) or services (e.g. Internet Archive Archive-it) that facilitate development of small or medium scale web archives. Use them and avoid the temptation of optimising before you carry out a complete preservation workflow. Keep in mind that the most important asset of a web archive is its data, so organise it the best way possible. Despite the particularities of web archives, the best practices in the development and operation of web-based information systems must be learned and applied to achieve success, so learn from the technical best practices of the Internet Giants and adapt them to your reality. Notice that procedures to maintain large services are applicable to smaller ones, while the opposite does not apply. Providing a list of these best practices is beyond the scope of this article, so I will only provide a few hints as examples:

- Adopt Shared-Nothing architectures for hardware and software so that the failure of a single piece of equipment cannot jeopardise the entire service; also, segregate development from production environments;
- Apply replication policies across different media and geographical locations to the web-archived data; test recovery procedures;
- Monitor services systematically using different tools. Induce faults to test recovery procedures and the monitoring tools themselves because they also fail. Current web users are very intolerant to unavailability;
- Maximise automation of testing tasks, using appropriate tools to perform functional, regression, security or workload tests;
- Collaborate with user experience researchers to focus the efforts on the real needs of your users/societies, and quantitatively evaluate how efectively they are being met.

The International Internet Preservation Consortium organises an annual Web Archiving Conference, which is an excellent opportunity to meet international peers and learn from their experience. It is also worthwhile to reflect on the lessons learned since the early days of the web, concerning the evolution, operation, sustainability and future of web archives[82].

Conclusions

The Web replaced printed media as the main means of communication and most of the information used to organise our current digital societies is published exclusively online. However, it disappears quickly, which causes problems when a need arises to access historical information. Web archives preserve historical web data and are unique sources of information used by academics to research the early years of the 21st century. However, both organisations and ordinary citizens are still not aware of the existence of web archives.

Arquivo.pt is a web archive born in 2007 as a digital infrastructure to support scientific research. It has been widening its scope so as to provide services to societies in general. Its web-archived data is open-access. The search services provided by Arquivo.pt include full-text search, image search, version history listing, advanced search and application programming interfaces (API) that facilitate development of added-value applications by third parties. The collections preserved in Arquivo.pt are created through a combination of automatic and human selection processes. They include information written in several languages, on issues of national and international interest (e.g. national or European elections). Arquivo.pt also provides complementary services that help solving everyday life situations, such as fixing broken links in websites by presenting the correspondent web-archived pages as alternatives. It also produced illustrative examples about how to explore the past-web by offering training in web preservation, publishing open data sets of derived results and creating online exhibitions based on past web pages.

Despite all the progress achieved so far, web archiving remains a challenging task that must address problems such as the lack of awareness about the importance of preserving online information, the difficulties with hiring web archivists or addressing Intellectual Property Rights in the fast-paced online world. Nonetheless, there are powerful tools and services available that enable any citizen or organisation to immediately start their own web archive. There are also international organisations, such as the International Internet Preservation Consortium, that connect international experts in web archiving and provide training. I recommend starting a web archive by performing the full preservation cycle for one single website: collect, store, provide access, disseminate and evaluate the obtained results. Doing something is infinitely more than doing nothing. With results on hand, it becomes easier to request additional resources to progress.

For centuries archivists have been doing their best to preserve information in printed media. Archivists could not preserve every single piece of information produced by humanity, but they

[82] Masanès, J., Major, D. and Gomes, D., 2021. The Past Web: A Look into the Future. In The Past Web (pp. 285-291). Springer, Cham.

preserved enough to keep the memory, required to sustain societies until our day. This work cannot stop just because a new medium of communication has emerged. The Internet is already more than 25 years old and digital societies must quickly pick up the pace and start preserving their own online information.

Bibliography

Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C. and Nelson, M.L., 2011, June. How much of the web is archived?. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (pp. 133-136).

Ainsworth, S.G., Nelson, M.L. and de Sompel, H.V., 2015. Evaluating the Temporal Coherence of Archived Pages.

Alam, S., Weigle, M., Nelson, M., Melo, F., Bicho, D. and Gomes, D., 2019, June. MementoMap framework for flexible and adaptive web archive profiling. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 172-181). IEEE.

AlSum, A., Weigle, M.C., Nelson, M.L. and Van de Sompel, H., 2014. Profiling web archive coverage for top-level domain and content language. International Journal on Digital Libraries, 14(3), pp.149-166.

Arquivo do Parlamento, https://arquivo-parlamento.pt/, accessed: 31 October 2022.

Arquivo.pt, A first attempt to archive the .EU domain, https://sobre.arquivo.pt/en/a-first-attempt-to-archive-the-eu-domain/, accessed: 31 October 2022

Arquivo.pt, Arquivo.pt Application Programming Interfaces (APIs), https://arquivo.pt/api, accessed 31 October 2022

Arquivo.pt, Arquivo.pt Awards, https://arquivo.pt/awards, accessed: 31 October 2022.

Arquivo.pt, Arquivo.pt Memorial: preserves information of historical websites, https://arquivo.pt/memorialen, accessed 31 October 2022.

Arquivo.pt, Cross-lingual collection about the 2019 European Elections is available, https://sobre.arquivo.pt/en/cross-lingual-collection-about-the-2019-european-elections-is-available/, accessed: 31 October 2022.

Arquivo.pt, Exhibitions, https://arquivo.pt/exhibitions/, accessed 31 October 2022.

Arquivo.pt, H2020 projects preserved by Arquivo.pt, https://sobre.arquivo.pt/en/h2020-projects-preserved-by-arquivo-pt/, accessed: 31 October 2022.

Arquivo.pt, Open dataset about cryptocurrency, https://sobre.arquivo.pt/en/open-dataset-about-cryptocurrency/, accessed: 31 October 2022.

Arquivo.pt, Publications, https://arquivo.pt/publications, accessed: 31 October 2022.

Arquivo.pt, Put an end to “page not found” on your website, https://arquivo.pt/arquivo404en, accessed 31 October 2022.

Arquivo.pt, Recommendations for authors to enable web archiving, https://arquivo.pt/recommendations, accessed 31 October 2022.

Arquivo.pt, SavePageNow, https://arquivo.pt/savepagenow, accessed 31 October 2022.

Arquivo.pt, Search the Geocities history!, https://sobre.arquivo.pt/en/historical-collection-geocities-available-at-arquivo-pt/, accessed: 31 October 2022.

Arquivo.pt, Suggest websites to be preserved - Collaborate, https://arquivo.pt/suggest accessed 31 October 2022.

Arquivo.pt, Training courses, https://arquivo.pt/training, accessed 31 October 2022.

Arquivo.pt,, Arquivo.pt · GitHub, https://github.com/arquivo/, accessed: 31 October 2022.

Ben-David, A. and Amram, A., 2018. The Internet Archive and the socio-technical construction of historical facts. Internet Histories, 2(1-2), pp.179-201.

Ben-David, A., 2019. 2014 not found: a cross-platform approach to retrospective web archiving. Internet Histories, 3(3-4), pp.316-342.

Ben-David, A., 2019. National web histories at the fringe of the Web: Palestine, Kosovo, and the quest for online self-determination. In The Historical Web and Digital Humanities (pp. 89-109). Routledge.

Berners-Lee Launches, T. 'www Foundation' at Igf 2009, ACM TECHNEWS, Ars Technica November 23, 2009 https://arstechnica.com/tech-policy/2009/11/tim-berners-lee-launches-www-foundation-at-igf-2009/, accessed 31 October 2022.

Biblioteca de Arte da Fundação Calouste Gulbenkian, arteparasempre.wordpress.com…, accessed 31 October 2022.

Bicho, D. and Gomes, D., 2016. Preserving Websites Of Research & Development Projects. In iPRES.

Brügger, N. and Laursen, D. eds., 2019. The historical web and digital humanities: the case of national web domains. Routledge.

Brügger, N. and Milligan, I. eds., 2018. The SAGE handbook of web history. Sage.

Brügger, N. ed., 2010. Web history (Vol. 56). Peter Lang.

Brügger, N., 2005. Archiving Websites. General Considerations and Strategies: General Considerations and Strategies.

Brügger, N., 2018. The archived web: doing history in the digital age. MIT Press.

Brügger, N., Goggin, G., Milligan, I. and Schafer, V., 2017. Introduction: Internet histories. Internet Histories, 1(1-2), pp.1-7.

Brügger, N., Locatelli, E., Weber, M. and Nanni, F., 2017. Web 25: histories from the first 25 years of the World Wide Web.

Classificação automática de artigos estigmatizantes de doenças mentais em jornais de notícias portugueses online, https://alina-yanchuk02.github.io/estigma/, accessed: 31 October 2022.

Costa, M., Gomes, D. and Silva, M.J., 2017. The evolution of web archiving. International Journal on Digital Libraries, 18(3), pp.191-205.

Cruz, D. and Gomes, D., 2013, September. Adapting search user interfaces to web archives. In Proc. of the 10th International Conference on Preservation of Digital Objects (Vol. 17).

Dados.gov.pt - Portal de dados abertos da Administração Pública, Arquivo.pt - pesquise páginas do passado, https://arquivo.pt/dadosabertos, accessed 31 October 2022.

George Santayana, Wikiquote, https://en.wikiquote.org/wiki/George_Santayana, accessed 31 October 2022.

Gomes, D. and Costa, M., 2014. The importance of web archives for humanities. International Journal of Humanities and Arts Computing, 8(1), pp.106-123.

Gomes, D. and Silva, M.J., 2006, July. Modelling information persistence on the web. In Proceedings of the 6th international conference on Web engineering (pp. 193-200).

Gomes, D. and Silva, M.J., 2008. The Viúva Negra crawler: an experience report. Software: Practice and Experience, 38(2), pp.161-188.

Gomes, D., Costa, M., Cruz, D., Miranda, J. and Fontes, S., 2013, May. Creating a billion-scale searchable web archive. In Proceedings of the 22nd International Conference on World Wide Web (pp. 1059-1066).

Gomes, D., Demidova, E., Winters, J. and Risse, T., 2021. Past Web. Springer International Publishing.

Gomes, D.C., 2006. Web Modelling for Web Warehouse Design (Doctoral dissertation, Universidade de Lisboa (Portugal)).

Graham, S., Milligan, I., Weingart, S.B. and Martin, K., 2016. Exploring big historical data: the historian's macroscope.

Harari, Y.N., 2014. Sapiens: A brief history of humankind. Random House.

Hockx-Yu, H., Laursen, D. and Gomes, D., 2019. The curious case of archiving. eu. In The Historical Web and Digital Humanities (pp. 64-72). Routledge.

Internet Archive, Wayback Machine Save Page Now, https://web.archive.org/save/

ISO, I., 2009. 28500: 2009 Information and documentation-WARC file format. International Organization for Standardization.

Jones, S.M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R. and Grover, C., 2016. Scholarly context adrift: three out of four URI references lead to changed content. PloS one, 11(12).

Kahle, B., 1997. Preserving the internet. Scientific American, 276(3), pp.82-83.

Klein, M. and Nelson, M.L., 2014. Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. International Journal on Digital Libraries, 14(1), pp.17-38.

Klein, M., Balakireva, L. and Van de Sompel, H., 2018, May. Focused crawl of web archives to build event collections. In Proceedings of the 10th ACM Conference on Web Science (pp. 333-342).

Major Minors, http://minors.ilch.uminho.pt/, accessed: 31 October 2022.

Masanes, J., 2006. Web archiving: issues and methods. In Web archiving (pp. 1-53). Springer, Berlin, Heidelberg.

Masanès, J., Major, D. and Gomes, D., 2021. The Past Web: A Look into the Future. In The Past Web (pp. 285-291). Springer.

Memento Time Travel, http://timetravel.mementoweb.org/, accessed: 31 October 2022.

meuParlamento.pt, http://www.meuparlamento.pt/, accessed: 31 October 2022.

Miguel Costa, Information Search in Web Archives, PhD thesis, Universidade de Lisboa, December 2014

Milligan, I., 2019. History in the age of abundance?: how the web is transforming historical research. McGill-Queen's University Press.

Milligan, I., 2022. The Transformation of Historical Research in the Digital Age. Elements in Historical Theory and Practice.

Ministério da Educação e Ciência, Decreto-Lei n.º 55/2013, Diário da República, 1.ª série — N.º 75 — 17 de abril de 2013.

Miranda, J. and Gomes, D., 2009, November. Trends in Web characteristics. In 2009 Latin American Web Congress (pp. 146-153). IEEE.

Mourão, A. and Gomes, D., 2021, The Anatomy of a Web Archive Image Search Engine-Technical Report, https://sobre.arquivo.pt/wp-content/uploads/The_Anatomy_of_a_Web_Archive_Image_Search_Engine_tech_report-1.pdf

Publications Office of the EU, Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, Official Journal of the European Union, http://data.europa.eu/eli/dir/2019/1024/oj, accessed 31 October 2022.

Pywb, Configuring the Web Archive — pywb 2.0 documentation, https://pywb.readthedocs.io/en/latest/manual/configuring.html#recording-mode, accessed 31 October 2022.

Ruest, N., Lin, J., Milligan, I. and Fritz, S., 2020, August. The archives unleashed project: Technology, process, and community to improve scholarly access to web archives. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (pp. 157-166), https://archivesunleashed.org/.

SalahEldeen, H.M. and Nelson, M.L., 2013, May. Carbon dating the web: estimating the age of web resources. In Proceedings of the 22nd International Conference on World Wide Web (pp. 1075-1082).

Schafer, V. and Winters, J., 2021. The values of web archives. International Journal of Digital Humanities, 2(1), pp.129-144.

Schroeder, R. and Brügger, N., 2017. The Web as History: Using Web Archives to Understand the Past and the Present (p. 296). UCL Press.

Sherratt, T. and Jackson, A., 2020. GLAM-Workbench/web-archives, https://glam-workbench.net/web-archives/, accessed 31 October 2022.

SolrWayback 4.0 release! What's it all about? Part 2, https://netpreserveblog.wordpress.com/2021/03/04/solrwayback-4-0-release-whats-it-all-about-part-2/, accessed: 31 October 2022.

Spaniol, M., Mazeika, A., Denev, D. and Weikum, G., 2009, September. Catch me if you can: Visual analysis of coherence defects in web archiving. In 9th International Web Archiving Workshop (IWAW 2009), Corfu, Greece (pp. 27-37).

Upwork, How Much Does It Cost To Build a Website? (2022 Data), https://www.upwork.com/resources/how-much-does-it-cost-to-build-website, accessed 31 October 2022.

Van de Sompel, H., Nelson, M. and Sanderson, R., 2013. RFC 7089-HTTP framework for time-based access to resource states-Memento. Internet Engineering Task Force (IETF), RFC.

Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S. and Shankar, H., 2009. Memento: Time travel for the web. arXiv preprint arXiv:0911.1112.

Webrecorder: Web archiving for all!, https://webrecorder.net/, accessed: 31 October 2022.

Winters, J., 2015. „Big UK Domain Data for the Arts and Humanities", Presentation, 2015 International Internet Preservation Coalition General Assembly, April 27-May 1, 2015. Silicon Valley, California, https://buddah.projects.history.ac.uk/.