# Acquiring and providing access to historical web collections

Daniel Gomes, David Cruz, João Miranda, Miguel Costa, Simão Fontes
Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal
{daniel.gomes, david.cruz, joao.miranda, miguel.costa, simao.fontes}@fccn.pt

## ABSTRACT

Every day, unique valuable information that describes our current days disappears from the web. National archives or libraries have been keeping cultural heritage for centuries by collecting and preserving past generation objects or printed media. Now, it is mandatory to preserve digital cultural heritage in the form of web content. The Portuguese Web Archive project began in 2008. Since then, it has periodically collected live-web content to be preserved but also acquired historical web collections from third-parties previously published. However, storing information before it vanishes from the web is not enough to make web archives useful to societies. Thus, the Portuguese Web Archive developed and made freely available several software tools to enable access to web-archived collections. The Portuguese Web Archive provides a full-text search service to access 1 131 million files archived from the web since 1996 (`www.archive.pt`). It also provides access methods to enable research and development activities over web-archived data.

## Keywords

Web archiving, digital preservation, Portuguese Web Archive

## 1. INTRODUCTION

The web is replacing printed media. Nowadays, we can find all kinds of publication genres transposed to online equivalents: electronic books, photo galleries, personal diary blogs, news articles, discussion forums or social networks. However, all this valuable information that describes our current days quickly disappears from the web [6]. In the same way that national archives or libraries have been preserving information published through printed media, it is now mandatory the creation of web archives to preserve the information published online.

Many web archives spread around the world collect and store web content [5]. However, enabling broad and efficient access to the web archived data is crucial to make web archives useful for societies. Live-web search engines are essential tools to enable access to current information. Web archives should complement live-web search engines by enabling access to past information published online.

The Portuguese Web Archive (PWA) project began in 2008 and aims to preserve information published on the web of main interest to the Portuguese community. Nonetheless, it also preserves content of international interest such as the sites from reputable worldwide organizations. In 2010, the PWA released the first version of a public search service that enables full-text search over its archived information. In March 2013, the PWA held 1 834 million web files collected from the web and integrated from historical collections provided by third-parties. Figure 1 presents the search interface that provides access to 62% (1 131 million) of the web files archived since 1996. This innovative service is publicly available at `archive.pt` and can serve a wide scope of user profiles and use cases. For example, web archive search can be useful to: journalists documenting articles, webmasters recovering lost versions of pages, historians studying digital documents about past events, lawyers obtaining evidence for legal cases, engineers consulting old documentation to fix legacy equipments or common web surfers recovering their broken-link favourites.

The software that supports the PWA was based on the Internet Archive Archive-access project tools [7], which are used by most web archives worldwide [5]. However, we observed that these tools did not fulfill our users requirements. Thus, we enhanced and adapted the Archive-access tools to support our service and made all the developed software available as a free, public open source project that can be reused and improved by other web archivists (available at `code.google.com/p/pwa-technologies/`). The PWA also provides tools and services to enable research over the archived data such as a distributed data processing platform or an OpenSearch API to facilitate the development of web applications that need to access web-archived data.

This demonstration will enable the attendees to discover the services provided by the PWA and discuss with the authors the details about how to develop and maintain a preservation service for web publications.

## 2. ACQUIRING HISTORICAL WEB COLLECTIONS

Since the first steps of the Internet in Portugal that individuals and organizations keep copies of published web content, most of the times for backup purposes. The PWA started collecting information for the web in January 2008 but we also acquired online content previously published to be preserved. We obtained historical content from web data sets gathered by research projects and personal collections that were gently supplied by their authors (see `arquivo.pt/supply` for details). The PWA preserves a total of 175 million web files (2.47 TB) supplied by third-party entities. The majority of these data was obtained from the Internet Archive that provided 124 million (1.9 TB) of data gathered from .PT between 1996 and 2007. Replicating these collections improved their chance of long-term preservation. Plus,

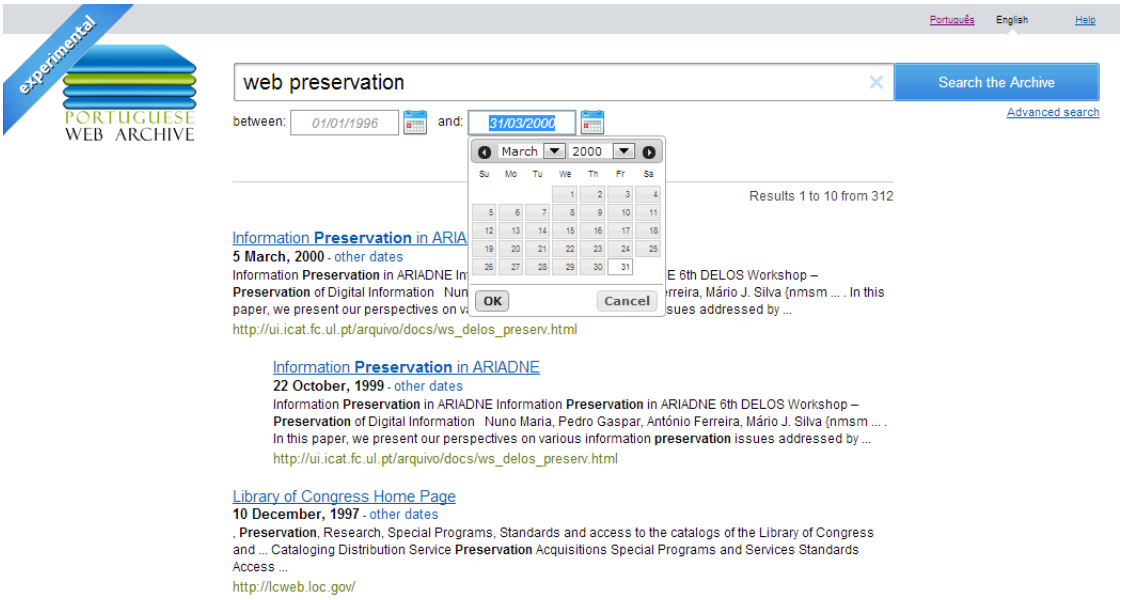


**Figure 1: Result page for a full-text search over the Portuguese Web Archive (www.archive.pt).**

they became full-text searchable through the PWA search service, while on the Internet Archive Wayback Machine the users have to know the exact address of the archived page that contained the desired information.

The acquired historical collections reached our web archive in heterogeneous formats, media support (e.g. CD-ROMs, backup tapes, original source code) and with scarce associated meta-data (e.g. missing original site URL or publication dates). Making this data searchable implied converting it to a uniform archive format so that it could be automatically processed indexed such as ARC or its successor standard WARC. We chose to use the ARC format instead of the official standard because it is the most widely supported by web archiving tools. The historical web collections acquired from the Internet Archive were delivered in the ARC format and were directly integrated in the PWA. The remaining acquired historical web collections were delivered to us in several distinct formats. Thus, we had to create specific integration modules to convert each one of these collections to the ARC format, which imposed a significant effort to integrate a relatively small amount of data. However, these collections provided valuable unique content. For instance, we obtained a version of the Library of Congress homepage dated from April 1996, while the oldest version existent on the Internet Archive is dated from December 1997.

A recurrent situation that we faced during the integration of the historical web collections was that acquired data were site backups made on local file systems with unknown or obsolete software. We observed that the majority of the acquired web collections created by organizations and individuals have been generated using software that was not designed with long-term preservation concerns, such as offline-browsers or through the Save feature of common web browsers. Offline browsers typically do not store meta-data related to each content saved locally, such as the original URL. As consequence, web archives cannot support URL search over these contents. If full-text is supported by a web archive, the contents could still be searchable but link-based algorithms could not be directly applied. Thus, these type of integrations required reverse engineering to model the archive file format and extract content meta-data. For instance, in 1995 a CD-ROM containing a snapshot of the Portuguese web was published as an attachment of a book. The web collection had the original URLs embedded as a reminder within each HTML page. However, non-HTML contents such as images did not have any URL associated. The extraction of the original URLs for these contents was automatized because each site was stored on a different directory and the original URLs were inferred by following relative links from pages. If the page with the URL `site.net/index.html` refered the image located in `./0.jpg`, then the original URL for the image was `site.net/0.jpg`.

As new web archiving initiatives arise they face the same challenge of having to integrate past content that is no longer available online and must be acquired from third-parties. Thus, we shared publicly the software developed to integrate our obtained historical web collections. The integration software was developed modularly so that it can partially applied and combined to address recurrent problems in independent collections. The converter of the CD-ROM collection to ARC format is available as an open source project at `code.google.com/p/roteiro2arc/`. HTTrack is a crawler used by web archives that stores content in a specific format [5]. The main purpose of HTTrack is to create site backups and the web collections generated with it contain most of the meta-data required to be successfully converted to the ARC format. The software to convert HTTrack crawls to ARC files is available at `code.google.com/p/httrack2arc/`.

## 3. PROVIDING ACCESS TO WEB-ARCHIVED CONTENT

Web archives contain rich and diverse information about international, regional or even personal events. However, web-archived information must be widely accessible to be useful. Most web archives rely on Archive-access tools to

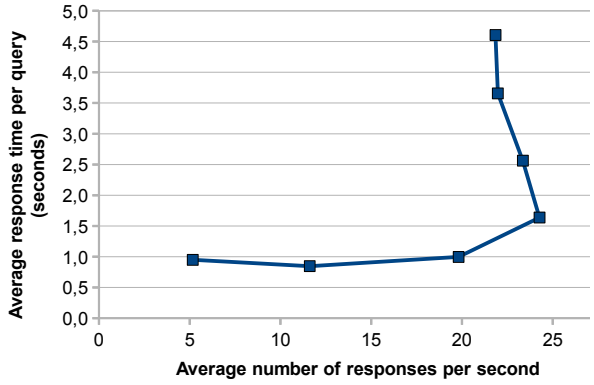


**Figure 2: Experimental search performance results: relation between average response time and workload with load balancing across 7 machines.**

provide access to their collections, in particular on the Wayback Machine for URL search and content visualization and NutchWAX for full-text search [5]. However, NutchWAX did not support the indexing of web collections that contained several versions of URLs harvested across time, which prevented its direct usage to index our acquired historical web collections that contained multi-version content. The performance of NutchWAX was also considered unsatisfactory by web archive stakeholders [5], missed support for internationalization of the user interface and did not include a query suggestions mechanism. We used the NutchWAX (v.0.11) and Wayback Machine (v.1.2.1) software as inception and enhanced it by addressing these drawbacks. The resultant software that supports the PWA search service was made publicly available as a free open-source project (`code.google.com/p/pwa-technologies/`). The introduction of a query suggestion mechanism had great impact on the perceived quality of our web archive because during usability testing we observed that users frequently mistyped queries and blamed the web archive for poor search results, often failing to spot their own mistypes [3]. The developed query suggestion mechanism was based on Hunspell, optimized for the Portuguese language and to the lexicon of our web collections [1] (`code.google.com/p/pwa-technologies/wiki/PwaSpellchecker`).

## 3.1 Search throughput capacity

Multiple, large and heterogeneous web collections must be searchable by queries in a few seconds. Users are used to the fast and high precision results of live-web search engines, such as Google, and expect the same behavior from web archive search systems.

We executed performance workload simulations to measure the throughput capacity and average response time of our web archive to search queries. These performance experiments were executed in a laboratory controlled environment to enable their reproducibility. The test collection was composed by 147 million web files gathered from 1996 to 2007. The experimental setup was composed by one load balancer that distributed the queries among 7 replicated search servers. The load balancing mechanism was implemented using the Linux Virtual Cluster software. Each search server supported queries over the full collection. The search servers shared the data storage device through a Storage Area Network that held the index. Each machine had 2 Xen Quad-core CPUs, 32 GB of memory and ran Linux. An increasing number of queries were submitted in parallel during a fixed interval of 5 minutes using several instances of the JMeter software and it was measured the time taken by PWA search system to respond to each query. The query set used to simulate the workload was composed by 300 000 queries obtained from a Portuguese web search engine [10] because a representative and structured web archive query log data set was not available at the time. Figure 2 presents the relation between workload and response time supported by the system. The obtained results show that up to an average workload of 20 responses per second, the system is able to maintain an average response time of approximately 1 second. However, when the workload reaches 25 responses per second, the average response time increases to 1.5 seconds and the system reaches its exhaustion point. From this point, we continued to increase the number of queries issued to the system but it was unable to respond to them. Thus, the system entered a thrashing state caused by overload and the average number of served responses per second decreased while the average response time increased.

The obtained results showed that our search software was able to support a throughput of 25 responses/second with an average response time of 2 seconds using load balancing across 7 machines. However, our search software may have to be installed on a smaller number of servers due to budget restrictions. Hence, we repeated the experiments and evaluated our search software without using any load balancing mechanism. We concentrated all software components and data structures on a single machine and the obtained results showed that the maximum supported throughput was 5 responses/second with an average response times of 2 seconds. Adding one replica and balancing the queries among these two machines, the response throughput increased to 10 responses/second with the same average response times of 2 seconds. We concluded that our search software is able to provide a satisfactory performance even when installed on limited hardware infrastructures.

The experimental setup previously described was deployed to production and we analyzed the logs of the queries issued by real users between May 2010 and July 2011 over a web collection of 187 million web files. Table 1 presents the response time distribution for full-text and URL queries. Around 87.7% and 89.9% of the full-text and URL queries, respectively, were responded in less than 3 seconds.

| Response time (s) | %full-text queries | %URL queries |
|---|---|---|
| $[0, 1[$ | 62.9% | 71.7% |
| $[1, 2[$ | 14.9% | 11.7% |
| $[2, 3[$ | 9.9% | 6.5% |
| $[3, 4[$ | 4.5% | 1.4% |
| $[4, 5[$ | 2.3% | 2.2% |
| $[5, \infty[$ | 5.5% | 6.5% |

**Table 1: Response time distribution derived from query log analysis (seconds).**

## 3.2 Search results relevance

Measuring the relevance of web archive search results usage requires test collections to obtain representative and reproducible results. However, existing test collections from

evaluation campaigns, such as the Text REtrieval Conference (TREC), do not address web archive requirements. For instance, their data sets are not composed by historical web collections gathered across time and the query sets are not focused on temporal queries that reflect the needs of web archive users. Therefore, we made a first effort to evaluate the relevance of our search results by performing a user click-through analysis derived from the query logs of the production environment of the PWA search service gathered from June to December 2010. The obtained results showed that 66% of the clicks were made on the first page of results, 23% of the clicks were made by the users on the first result presented by the system and 12% on the second result. These results are similar to those presented on web search engine studies [2]. Thus, they are a positive indicator of relevance. Another positive indicative is that only 2% of the URL search sessions did not receive any click by the users.

On the other hand, we obtained two negative quality indicators. The first one, is that 31% of the full-text search sessions did not receive any click by users. This abandonment rate suggests that users quit search before finding what they needed. The second negative indicative is that 85% of users identified by IP address did not revisit the web archive during the seven months period. One possible explanation for the non-revisit figure is that most users do not have a frequent need to search for historical web contents as they do for current information. Hence, the interval of time for users to revisit a web archive tends to be longer than for search engines. A longer time interval between revisits also reduces the probability of the same user revisiting the web archive using the same IP address.

## 3.3 Access tools for research

Web-archived information is a precious and abundant source of raw data for research. The PWA provides access tools and services to enable research over its archived data. The PWA has collaborated with researchers by providing data sets and access to its computing platform based on Hadoop. For instance, research has been performed to analyze the evolution of web characteristics [9], evaluate cross-lingual web classification algorithms [4] or to measure the accessibility of the web to people with disabilities [8]. To enable the automatic measurement of web page accessibility, we developed a software library that facilitates the development of distributed applications to process archived data. The PwaProcessor library interacts with the Hadoop framework for the distribution and the NutchWAX code for reading the content of ARC files (`code.google.com/p/pwa-technologies/wiki/PwaProcessor`).

Recently, the PWA published the first version of a test collection to support research on web archive information retrieval. It is composed by three parts: (1) a corpus representative of the documents' versions encountered in a real search environment; (2) a set of topics describing users' information needs; and (3) relevance judgments (a.k.a. qrels) indicating the degree of relevance of each document retrieved for each topic (`code.google.com/p/pwa-technologies/wiki/TestCollection`).

The PWA provides an OpenSearch access API that facilitates the development of web applications that need to access web-archived data (documentation at `code.google.com/p/pwa-technologies/wiki/OpenSearch`). This API has been frequently used by Computer Science students of the University of Lisbon to develop their academic projects.

# 4. CONCLUSIONS AND FUTURE WORK

Web archives already hold historical information spanning decades. However, making all these data widely accessible is still an open challenge. The Portuguese Web Archive collects information from the live web since 2008 and acquired historical web collections previously created by third-party entities.

The PWA search service enables full-text and URL search over 1 131 million files archived since 1996 (`archive.pt`). All the software developed to support this service was made publicly and freely available as an open source project (`code.google.com/p/pwa-technologies/`) so that it can be collaboratively enhanced and used as a baseline to develop more sophisticated accessible web archives in the future. The obtained experimental results showed that our search software is able to support a significant workload even when installed on limited hardware infrastructures and provide relevant search results. The PWA also provides services and data sets specially designed to support research and development activities.

# 5. REFERENCES


[1] M. Costa, J. Miranda, D. Cruz, and D. Gomes. Query suggestion for web archive search. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres 2013)*, September 2013.

[2] M. Costa and M. J. Silva. Characterizing search behavior in web archives. In *Proc. of the 1st International Temporal Web Analytics Workshop*, 2011.

[3] D. Cruz and D. Gomes. Adapting search user interfaces to web archives. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres 2013)*, September 2013.

[4] A. Garzó, B. Daróczy, T. Kiss, D. Siklósi, and A. Benczúr. Cross-lingual web spam classification. In *The 3rd Joint WICOW/AIRWeb Workshop on Web Quality in conjunction with WWW 2013*, WICOW 2013, May 2013.

[5] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. In *International Conference on Theory and Practice of Digital Libraries 2011*, Berlin, Germany, September 2011.

[6] D. Gomes and M. J. Silva. Modelling information persistence on the web. In *ICWE '06: Proceedings of the 6th international conference on Web engineering*, pages 193–200, New York, NY, USA, 2006. ACM Press.

[7] Internet Archive. Nutchwax - Home Page. `http://archive-access.sourceforge.net/`, March 2008.

[8] R. Lopes, D. Gomes, and L. Carriço. Web not for all: A large scale study of web accessibility. In *W4A: 7th ACM International Cross-Disciplinary Conference on Web Accessibility*, Raleigh, North Carolina, USA, April 2010.

[9] J. Miranda and D. Gomes. Trends in Web characteristics. In *7th Latin American Web Congress (LA-Web 2009)*, Merida, Mexico, November 2009.

[10] M. J. Silva. The Case for a Portuguese Web Search Engine. In P. Isaias, editor, *Proceedings of IADIS International Conference WWW/Internet 2003*, Algarve, Portugal, November 2003.