The Award for Research and Innovation

Application form

Searching preserved web-images

Summary of the initiative (up to 100 words).

This short description will be used within publicity material for the awards. Care should be taken to ensure it is accessible to a lay audience.

Images published online are precious digital assets that document contemporary times for future generations. This initiative describes the research and development of an innovative image search system that enables the discovery and access to billions of preserved images acquired from the web since the 1990s. This research was applied to enhance the Arquivo.pt web archive with an image search service publicly available to any Internet user, officially launched in August 2022. The resulting scientific and technical publications are available in open-access and the developed software is available as free open-source software to be reused and enhanced by the community.

Long description (up to 1000 words)

This should set out clearly and concisely the key points of the work and the main achievements. It will be used in publicity materials, so care should be taken to ensure that it is written for a lay audience.

Images published online are valuable digital assets that document our contemporary societies. However, 80% of web content, including web images, is not available at its original source after only one year because they were updated or deleted. This situation constitutes a concerning continuous loss of historical data which consequently undermines the sustainability of the future for Humanity's digital assets. Digital assets that were archived but cannot be effectively discovered and accessed are not preserved, just stored. Digital preservation requires effective methods to ensure the long-term accessibility of the objects archived in various formats, including images.

Live-web search engines monopolise the discovery and access methods to the digital assets published online. It is concerning that they constantly update their systems to replace historical versions of images with the most recent information available online, erasing preserved digital assets or making them unavailable to the public. Web archives complement live-web search engines because they provide a temporal perspective over web data. However, creating a practical Google-like service that enables any Internet user to submit a text query, and immediately receive a list of historical preserved images raises challenging research questions not addressed when dealing with the live-web. How to index web images collected over multiple years? How to deal with the resulting duplicated content? Should web images and pages which prevail over time be considered redundant? Moreover, as most existing tools focus on supporting search over text (e.g. web pages) there was no publicly available system specifically developed to search and access preserved web images.

The goal of this project was to research and develop an innovative open-source system focused on supporting search and access over web-archived images. The developed system addresses the challenges raised by the inherent temporal features of historical web-archived data, but at the same time provides a familiar look-and-feel similar to a live-web image search engine such as Google Images to facilitate its adoption by common Internet users.

Challenges addressed by this research include extracting image-specific metadata for web assets spanning decades, dealing with multiple versions of the same image over time (captured more than once or embedded in multiple pages), processing and indexing large amounts of preserved images for real-time search.

For extracting image metadata, we developed a novel image caption extraction technique that identifies the most relevant textual information in the web page about the embedded image. To deal with multiple versions of the historical images preserved over time, we found the balance between preserving the evolution of the image metadata over time, versus keeping the volume of information manageable for large-scale processing. To index the resulting hundreds of million images, we deployed a distributed architecture that is able to handle over 50 concurrent users per second, with a median query response time below 500 milliseconds.

The developed system was deployed in the Arquivo.pt web archive to enhance it with an innovative image search service. The user inputs a textual query and a set of image results are displayed in a grid layout to facilitate a quick choice of relevant images. When clicking on an image result, an image viewer is displayed, showing the image in a larger format together with metadata about the image and web page where it was found. Advanced search for images provides filters over date range, image format, size or website. For technical users, the image search service provides an Image Search API which supports large scale automated processing. The Image Search API is documented on GitHub at https://arquivo.pt/api/imagesearch.

The main research contributions of this project arise from the innovative techniques developed to:

- extract image metadata by identifying relevant textual content in web pages;
- reduce index sizes by de-duplicating web-archived images,
- handle the large volume of information (1,862 million web images)

This project originated one peer-reviewed scientific publication entitled “Searching images in a web archive” at the 10th IEEE International Conference on Data Science and Advanced Analytics 2023, one master thesis entitled “Automatic Identification of Not Suitable For Work images”, two technical reports and several invited talks. The deployment of this system in the Arquivo.pt web archive represents a practical and innovative contribution to digital preservation, as it currently enables the discovery and access of billions of images preserved from the web since the 1990s for present and future generations.

SECTION FOUR: SUPPORTING STATEMENTS

Nominees are invited to review their projects against the criteria for the awards.

Clarity of purpose: what were your main aims and objectives? (up to 100 words)

The main goal of this image search project was to develop a free and open-source system, that can be reused and enhanced by the Digital Preservation community, to enable Internet users to discover and access millions of historical images preserved from the web over decades in a simple way. This system was aimed to be deployed in production as a practical innovative image-search service running in the Arquivo.pt web archive. Providing a free Image Search API aimed to enable external entities to autonomously develop additional added value service for Digital Preservation (e.g. format conversion or new discovery methods).

Effectiveness of methodology: why was this methodology chosen and was it effective? (up to 100 words)

The chosen methodology was to iteratively research techniques to validate user needs. First, we processed a small amount of data end-to-end and created a fully fledged prototype, validating user needs and goals. After the success of the original prototype, we focused on improving the quality of the information extracted according to user feedback and built a scalable system to process the petabytes of images preserved by Arquivo.pt. R&D for this project began in 2018, it was elected as an official service of Arquivo.pt in 2022 and finished with the release of the associated research article in 2023.

Why is this an exemplary or innovative use of digital preservation techniques and principles? (up to 100 words)

The digital preservation process starts with the acquisition of the assets to be stored. However, there is no digital preservation without practical and effective methods that assure the accessibility of the stored digital assets over time. This project originated an innovative system that enables search and access over large amounts of images acquired from the web, contributing to secure a sustainable future for the images published online and preserved in web archives. Adopting open-access and open-source policies for all the software, publications, preserved data and services is an exemplary practice of Digital Preservation principles.

Clarity and practicality of benefits: what were the benefits; who were the beneficiaries? (up to 100 words)

A system optimised to enable search over historical images preserved from the web is a unique and precious tool that opens a new window to the past, demonstrating in practice how online query intents are prone to temporal contexts. For instance, searching for images of the "President of the USA" in 2000 vs. 2016 yields distinct results. Searching preserved web-images appeal ranges from everyday users who are curious about images from the past, to professional users which use it through human curation or large scale automated processing through the provided API.

Longevity of contribution: how have you made a lasting contribution to digital preservation? (up to 100 words)

The resulting publications are available in open-access and the software as free open-source to be reused and enhanced by the Digital Preservation community. In the age of AI-generated images, a practical service that enables access to preserved images before the emergence of these machine-generated images is a potential ground-truth resource for current and future researchers and practitioners. The system supports access both by humans and machines, including web user interfaces, APIs and bulk download (arquivo.pt/api#bulk), to support further digital preservation activities. The deployment of the system in Arquivo.pt contributes to the accessibility of its preserved images.

Extensibility of benefits: have others been able to use your tools or techniques? (up to 100 words)

The image-search system is available for everyday users in a free and open access model.
The source code and tools developed are open source and available for reuse by other initiatives with similar goals. The Image Search API was used for academic research projects and by projects that applied to the Arquivo.pt Awards which promote the usage of preserved web data (arquivo.pt/awards). The Image-search system deployed in Arquivo.pt was launched as an official service in August 2022 and responded to 2 million requests in 2023.

Cost/benefit analysis: what resources did you consume? This could refer to financial investment, human resources and effort (up to 100 words)[1]

Arquivo.pt team is composed of four members and two of them worked on this Image Search project full time for 2 years. The deployment of the image search service relied on the Arquivo.pt infrastructure of over 70 servers. We estimate that €500 000 were invested in the production of the image search system, split between computing infrastructure costs (about 250000€ for about one year of mostly exclusive usage) and five man-years of work, split across the team.

Significance: what is the significance of the objects preserved, or of the work undertaken for the digital preservation community? (up to 100 words)

The images published online are precious digital assets that document contemporary times for future generations. However, online images are being continuously lost or stored without practical tools to be searched and accessed. Live-web search engines such as Google support effective search over currently available web images but ignore preserved assets. This work contributed with an innovative system to ensure the accessibility of historical web images, so that current and future generations may open a new window into the past. The deployment of this work in Arquivo.pt improved the accessibility of 14 028 million preserved documents since the 1990s.

Fit to audience: how did you assess audience needs and how did you fit them? (up to 100 words)

After the creation of the original prototype, we identified two classes of users: everyday users who want to search images about a specific past event or person (e.g. find photos) and advanced users who need a broad access method for processing large amounts of past web images (e.g. to train an AI model). The image search UI with advanced filtering and the Image Search API support both audiences. The image search system was iteratively assessed for 5 years, by systematically gathering user needs through user-centred methodologies (e.g. usability tests) and monitoring system performance (e.g. log analysis).

[1] The cost/benefit analysis was introduced by the judges in 2010 as a mechanism to help compare large well-funded projects with small unfunded ones. In principle, the former would always deliver more impact than the latter, which if followed to a logical conclusion would mean that only large projects would ever win the award. This is to be avoided. In practice therefore, the judges will assess the cost-effectiveness of your initiative based on a description of the resource used. The category is entirely about the effort used to develop the solution, not the effort that the solution might save once deployed. You may choose to illustrate the latter point under the 'Clarity and practicality of benefits' section.

Assessment of quality by peers: how have others reacted to your work? (up to 100 words)

The reception of this project has been widely positive. The research and development work of this project resulted in one peer-reviewed scientific publication, one master thesis, two technical reports and 6 invited talks at scientific and training events. As a didactic tool, it can change how one perceives the world by showing how queries return different images when searched over distinct time periods. The image-search service deployed in Arquivo.pt responded to 2 million requests in 2023.

Public profile: what is the public profile of the digital objects you have safeguarded? (up to 100 words)[2]

The public profile of the digital objects safeguarded is available at arquivo.pt/collections. Arquivo.pt holds 176 collections comprising 14 028 million documents obtained from 47.9 million websites. Image search augments Arquivo.pt's objective of preserving online digital legacy for future generations, by providing practical access to its preserved images. Arquivo.pt expanded to become part of an international digital preservation effort that collaborates to preserve digital content about events across the world. Currently, this work enables search and access over 4 088 million web images preserved from the web since the 1990s.

[2] The judges would like to know if/why the content you are preserving would be of interest to the public? Is it recognizable as having belonged to someone of scientific, cultural or historic importance for example? Is it the first of a kind? Did it herald a change in scientific, historic or cultural process? E.g. David Bowie's email collection, the first data produced by the Large Hadron Collider.