The Award for Safeguarding the Digital Legacy

Arquivo.pt Catalog of tools for digital preservation

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.

Information that rules modern-day lives is born-digital and disseminated online. However, invaluable digital objects published online have been continuously lost.
Arquivo.pt is a public infrastructure which supports the preservation of digital objects published online to safeguard this digital legacy for future generations. Thus, in October 2023 after 15 years of research and development, Arquivo.pt released a Catalog of 13 innovative tools to support the preservation of at-risk online content, from acquisition to dissemination (e.g. search and access, APIs, training, open data sets, exhibitions). Arquivo.pt safeguards online digital objects of worldwide interest for research and education.

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.

The Web is the largest and most widely used source of digital objects. Despite these objects becoming immediately accessible to millions of people as soon as they are published, most of them are solely hosted at their original source and are at-risk of being irremediably lost. Therefore, ready-to-use tools and services to safeguard digital objects published online are required to safeguard this invaluable digital legacy for future generations.

Arquivo.pt is a public digital preservation infrastructure that enables anyone to store, search and access historical digital objects preserved from the Web since the 1990s. It contains over 20 billion digital objects (1.3 PB) in multiple formats and languages, acquired from websites from all over the world. About half of Arquivo.pt users come from outside of Portugal.

The main objective of the Arquivo.pt Catalog is to support the preservation of the born-digital information published online that rules modern societies by providing a toolkit of services freely accessible to a broad scope of users so that any Internet user can contribute to the digital preservation lifecycle of objects published online. As different users may have

different needs regarding digital preservation, providing a comprehensive Catalog of tools potentiates the fulfilment of most requirements.

The Arquivo.pt Catalog of tools for digital preservation was officially launched in October 2023 after 15 years of iterative development. It is composed of 13 running tools/services listed at https://arquivo.pt/catalog to support the preservation of online digital objects from their acquisition to dissemination:

- Search and access (arquivo.pt): includes full-text search, image search, version history listing, advanced search, automatic generation of narratives and replay of web-archived content with 6 complementary options (e.g. Technical details, Complete page or Reply with old browser);
- Application programming interfaces (arquivo.pt/api): facilitates the development of added-value applications by third parties to support search over URL, full-text and images (Arquivo.pt API, Image Search API, CDX-server API, Memento API);
- Suggest websites (arquivo.pt/suggest): any Internet user can suggest websites to be safeguarded. 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;
- SavePageNow (arquivo.pt/savepagenow): allows users to immediately perform high-quality archiving of a set of web pages in Arquivo.pt using a browser-based crawler. The users only need to enter a page's address and start browsing so that all the visited content is preserved, which facilitates the complete archiving of a small website to be carried out autonomously by the users;
- Integration of historical web data collections (arquivo.pt/donate): Arquivo.pt began archiving web content in January 2008. However, donations of historical web content previously published have been received from external sources to be safeguarded;
- Training (arquivo.pt/training): is a free training programme that aims to raise awareness about the importance of preserving the digital legacy published online. It is composed of four modules: "New ways of searching the past", "Well publish to well preserve", "Automatic processing of information preserved from the Web" and "Web archiving: Do-it-yourself!". This training programme was the seed for the publication of the book "The Past Web: Exploring Web Archives" which aggregated contributions from worldwide experts in web preservation and research;
- Open data (arquivo.pt/dadosabertos): are datasets containing metadata about the preserved digital objects useful for third parties, such as lists of URLs that document elections. These datasets have been reused and improved by other organisations also interested in preserving this digital legacy (e.g. Museums);
- CitationSaver (arquivo.pt/citationsaver): extracts links from documents and safeguards the targeted digital objects so that they can be later retrieved from Arquivo.pt. Conventional documents meant to be printed (e.g. in PDF format) cite online digital objects by referencing their URLs. However, when these links become inaccessible, even printed documents lose their integrity because their citations become useless;
- Arquivo404 (arquivo.pt/arquivo404): presents preserved digital objects instead of error messages (e.g. "Error 404: Page Not Found"). Webmasters just need to insert one single line of code in the page that generates the 404 error message. 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 page preserved in a configurable set of web archives using the Memento protocol;

- Memorial (arquivo.pt/memorial): safeguards the digital objects which compose a website after its deactivation. Costs grow as websites become older because of the obsolescence of supporting technologies and consequent dangerous security vulnerabilities. The Memorial offers high quality preservation of historical web-content that enables maintaining the original domain name of the deactivated website, keeping its content searchable through live-web search engines and avoiding broken links to its pages;
- High-quality archive (on-demand): conventional crawlers quickly collect large amounts of information but sometimes miss rich media, such as embedded videos. This service enables high-quality archiving of selected websites which are iteratively archived and curated using the best combination of technologies available;
- Creation of collections and thematic exhibitions (arquivo.pt/expos): are online exhibitions of safeguarded web pages organised by theme curated in collaboration with external entities which are field experts such as press, radio, municipalities, R&D units, schools or museums. Each exhibition is followed by dissemination campaigns promoted by the external organisations which expand the awareness to the importance of digital preservation to new audiences;
- Itinerant exhibition of posters at external institutions (arquivo.pt/posters): the down-side of preserving exclusively born-digital artefacts is that it becomes a challenge to catch the attention of potential new users in the physical world. Many digital preservation initiatives rely on digital methods to preserve printed documents. We reversed this strategy and printed a set of posters featuring historical digital objects published online to raise awareness about the pertinence of preserving born-digital legacy.

The Arquivo.pt Catalog of tools for digital preservation is an innovative and comprehensive toolkit to safeguard digital legacy published online for future generations available to anyone.

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 aim of the Arquivo.pt Catalog is to safeguard the born-digital objects published online that rule contemporary societies by providing a toolkit of services freely accessible to a broad scope of users. This way, researchers, information professionals, IT experts or common Internet users can contribute to the digital preservation of online objects by using these tools to select, acquire, store, access, reuse and disseminate valuable historical information published online over the past 30 years. It was estimated that €516 000 000 were invested in the production of the safeguarded digital objects, which otherwise had been lost.

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

The chosen methodology consisted of the iterative research and development of a Catalog of digital preservation tools which support the full preservation cycle of online digital objects. These tools were iteratively developed by being publicly launched to gather the real-world feedback which drove their enhancement towards the needs of users and avoided over-optimising before carrying out a complete preservation workflow. Despite the particularities of web archives, the effectiveness methodologies applied in the development and operation of large web-based information systems had to be followed to achieve success (e.g. Quality assurance, Site Reliability Engineering, Software development).

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

Digital preservation requires acquiring at-risk digital objects before they disappear, storing them safely and keeping them as accessible so that they remain useful to citizens of the present and the future. The Arquivo.pt Catalog is innovative because its tools address all the stages of digital preservation. For instance, SavePageNow, CitationSaver or the Integration of collections address several paths for acquiring online digital objects. Search, Memorial and the Exhibitions support the accessibility of the preserved objects. Arquivo.pt was the first digital preservation service to support such a comprehensive Catalog of tools to safeguard the digital legacy of online objects.

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

The Arquivo.pt Catalog provides ready-to-use tools to support the preservation of online objects, which benefit from digital preservation practitioners to common Internet users. Researchers are power-users and there are at least 757 related works, including the production of derived research datasets. The Arquivo.pt Award annually distinguishes innovative works which used the tools that now compose the Catalog. Over 7 editions, 167 applications were received and the 23 works that were awarded clearly demonstrate how web archives broadly benefit digital legacy, spanning all areas of knowledge such as health, digital humanities or computer science (winners at arquivo.pt/awards).

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

The development of the Catalog required a significant Research & Innovation effort which originated long-term contributions to digital preservation such as technical and scientific articles available in open-access or open-source software (github.com/arquivo/). All the safeguarded digital objects are available in open access through multiple methods to support their broad reuse over time. The tools support 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 preserved objects were also donated to the Internet Archive collections to increase their longevity (https://archive.org/details/portuguese-web-archive).

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

The Catalog tools have been used and extended by the applicants for the Arquivo.pt awards or to produce research data sets. Online exhibitions have been extended by GLAM organisations to complement their collections (e.g. Museum of Tourism) or celebrate events by creating “time travels” (e.g. Museum of Presidency of the Republic). SavePageNow received 45000 requests in 2023 and has been used by the German Press Agency (DPA) to safeguard fact-checking information or by Wikipedia users to safeguard links to external citations. In 2023, the tools provided access to 470 TB of digital objects preserved (100M API requests).

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

The average annual cost to support the Catalog is €250 000 and the team has been composed of 4 members. The data centre is composed of 77 servers (18 TB of memory, 2180 vCPUs, 1234 hard drives).

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

Within merely 30 years, digital objects which circulate online became the most significant means of communication, while the previously dominant printed media became a luxury commodity. Losing the digital legacy of online objects jeopardises the sustainability of organisations because they represent the large majority of the information used to organise Humanity. The Catalog provides tools and best practices to support the activities of the digital preservation community. Preserved online digital objects are also a unique source of information to derive some trends, train AI models (LLMs) or recover works thought to be lost.

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

Different audiences have different needs regarding digital preservation (DP). DP professionals may focus on safeguarding at-risk objects before they disappear, while academics are most interested in accessing the previously safeguarded objects. The Catalog of tools was iteratively assessed over 15 years, by systematically gathering user needs through user-centred methodologies (e.g. usability tests) and monitoring system performance (e.g. log analysis). During this process, some tools were abandoned while others were incrementally enhanced until we reached the current Catalog. In 2023, the online tools in our Catalog received 1.7 million unique users.

[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 research and development work that produced the Catalog originated the publication of over 31 peer-reviewed scientific articles, 21 technical reports and over 200 communications across national and international events in the area of web preservation and research. There are at least 1000 citations to our publications. In 2023, the Training service received 901 attendees that evaluated it with an overall satisfaction rate of 94%. The Arquivo.pt Catalog was considered the best Digital Service of 2022, entered the Honour Roll for Cybersecurity in Portugal and was considered one of the top 3 government digital services in Portugal 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 20 041 million web files (1 311 TB of historical web data) obtained from 47.9 million websites. The initial objective of Arquivo.pt was to preserve online digital legacy related to Portugal as a national memory for future generations. However, the Catalog of tools evolved the service to become a worldwide infrastructure for the digital preservation of online objects. Currently, Arquivo.pt generates several types of collections according to their scope, crawl frequency and quality of the acquisition process.

[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.