Arquivo.pt: catálogo de serviços para preservação digital

Daniel Gomes, daniel.gomes@fccn.pt

Fundação para a Ciência e a Tecnologia - Unidade FCCN
Av. do Brasil 101
1700-066 Lisboa
Portugal

Resumo

A informação que rege a vida moderna nasce digital e é disseminada em linha. No entanto, estes objetos digitais de valor inestimável têm sido continuamente perdidos. O Arquivo.pt é uma infraestrutura pública que apoia a preservação de objetos digitais publicados em linha para salvaguardar este legado digital para as gerações futuras. Após 15 anos de investigação e desenvolvimento, o Arquivo.pt lançou um Catálogo de 13 ferramentas inovadoras para apoiar a preservação de conteúdos em linha, desde a sua aquisição até à disseminação (e.g. pesquisa e acesso, APIs, formação, dados abertos cenários, exposições). O Arquivo.pt salvaguarda objetos digitais em linha de interesse mundial para investigação e educação.

Palavras-chave: preservação digital, arquivamento da web, Arquivo.pt, arquivos digitais

Introdução

A transformação digital fez com que os meios de comunicação impressos usados pelas organizações e cidadãos transitassem para informação nascida digital, publicada e disseminada através da Internet. A Web é a maior fonte de informação inventada pela Humanidade e a vida quotidiana nas sociedades da informação modernas é regida por informação digital publicada exclusivamente em linha.

No entanto, a memória desta informação tem sido continuamente perdida. Apesar da informação em linha se tornar imediatamente acessível a milhões de pessoas assim que é publicada, a maioria é irremediavelmente perdida após alguns anos. De acordo com um estudo de 2024 (Chapekis, 2024), 38% das páginas Web que existiam em 2013 já não estão acessíveis passados 10 anos.

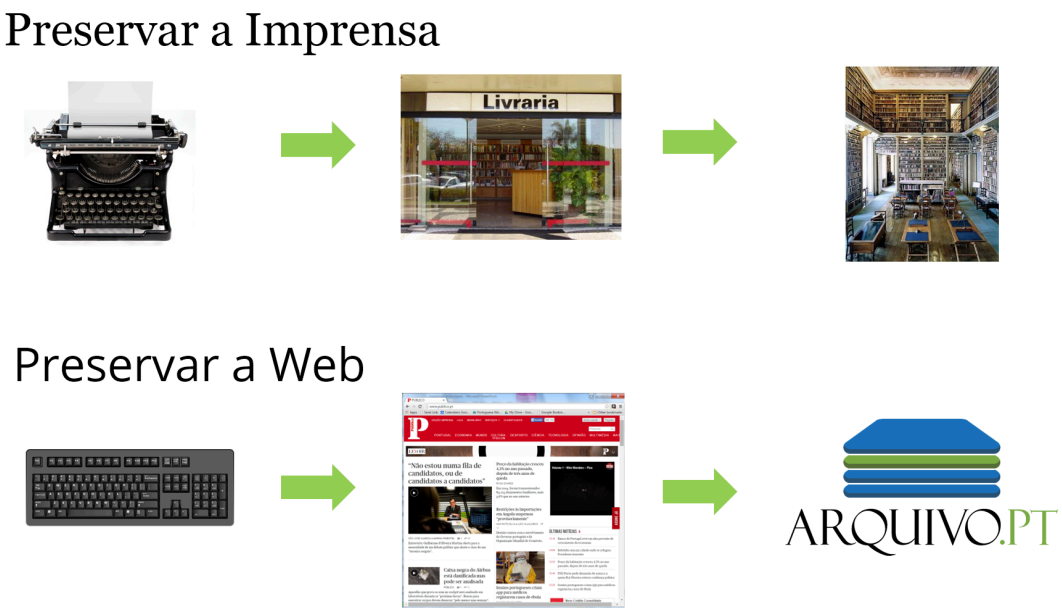

Figura 1. Analogia entre a preservação da informação publicada através de meios impressos e em linha.

Portanto, são necessárias soluções tecnológicas inovadoras prontas a utilizar para salvaguardar a informação publicada em linha, a fim de salvaguardar este legado digital para as gerações futuras e resolver problemas do quotidiano causados pela perda de informação recente, tais como o famoso, mas frustrante erro 404 “Página não encontrada”.

Figura 2. O primeiro jornal em linha nacional de Portugal desapareceu após 17 anos de atividade e os seus conteúdos informativos deixaram de estar disponíveis.

A Figura anterior apresenta o exemplo do Diário Digital, o primeiro jornal nacional de Portugal em linha que desapareceu após 17 anos de atividade. Sem os arquivos da web, os seus conteúdos informativos, que são fontes únicas para descrever os acontecimentos do século XXI, ter-se-iam perdido irremediavelmente.

Arquivo.pt preserva a informação publicada em linha

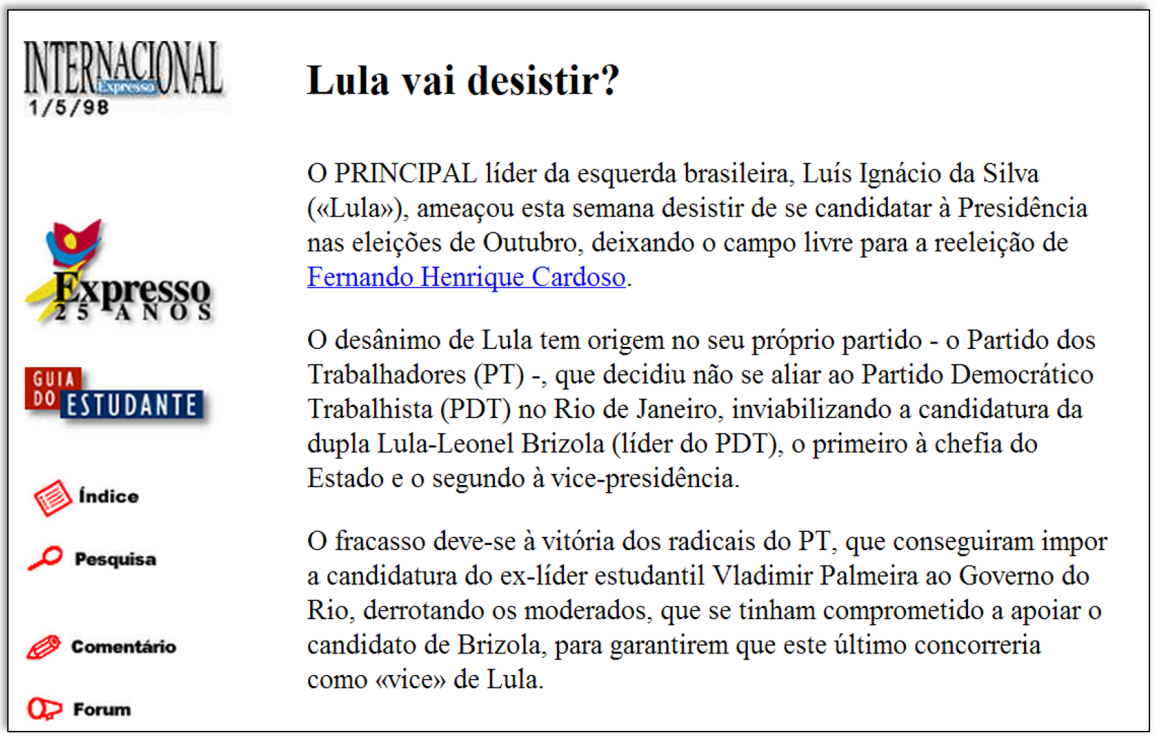

Figura 3. Página arquivada do jornal português "Expresso" no dia 1 de maio de 1998, disponível em: https://arquivo.pt/wayback/19990830180642/http://www.expresso.pt:80/ed1331/i303.asp?i302,i303.

O Arquivo.pt é um serviço público que preserva informação publicada em linha (Gomes, 2022). O Arquivo.pt preserva mais de 20 000 milhões de objetos digitais (1,3 PB) em múltiplos formatos e idiomas, adquiridos a partir de websites de todo o mundo. O perfil público dos objetos digitais salvaguardados está disponível (https://arquivo.pt/collections). O Arquivo.pt detém 176 coleções compostas por 20 041 milhões de ficheiros web (1 311 TB de dados históricos web) obtidos a partir de 47,9 milhões de websites. Inicialmente, o objetivo do Arquivo.pt era o de preservar o legado digital em linha relacionado com Portugal, como memória nacional para as gerações futuras. No entanto, as informações históricas preservadas ao longo do tempo têm um interesse de âmbito internacional. Por exemplo, a Figura anterior apresenta um exemplo de uma notícia do website do jornal português "Expresso" publicada no dia 1 de maio de 1998 que se encontra preservada no Arquivo.pt. Esta notícia intitulada "Lula vai desistir" fornece informação acerca de como as eleições, que são um evento de

âmbito nacional no Brasil, são descritas fora do seu país. A Figura seguinte apresenta como exemplo a página arquivada do jornal brasileiro Estadão no dia 21 de maio de 2009 que descreve a morte do mais famoso cantor de música pop Michael Jackson.

Figura 4. Página arquivada do jornal brasileiro "Estadão" no dia 21 de maio de 2009, disponível em: https://arquivo.pt/wayback/20090626105349/http://www.estadao.com.br/home/index.shtm.

Com o passar do tempo, o Arquivo.pt evoluiu para se tornar uma infraestrutura mundial para a preservação digital de objetos em linha. Atualmente, o Arquivo.pt produz e mantém vários tipos de coleções de acordo com a sua abrangência, frequência de ingestão e qualidade do processo de aquisição.

O desenvolvimento do Catálogo exigiu um esforço significativo de Investigação e Desenvolvimento que originou contribuições de longo prazo para a preservação digital, como

artigos técnicos e científicos disponíveis em acesso aberto (https://arquivo.pt/publica) ou software de código aberto (https://github.com/arquivo/). Todos os objetos digitais preservados estão disponíveis em acesso aberto através de múltiplos métodos para apoiar a sua ampla reutilização ao longo do tempo. O Arquivo.pt disponibiliza o acesso a humanos e máquinas para suportar atividades de preservação digital. Os objetos preservados também foram replicados nas coleções do Internet Archive para aumentar sua longevidade (https://archive.org/details/portuguese-web-archive). Todo o software aplicado no desenvolvimento do Arquivo.pt é gratuíto e de código-aberto, tendo sido desenvolvido principalmente pelas organizações WebRecorder.net, Internet Archive e Apache Software Foundation.

Catálogo de serviços para preservação digital

Após 15 anos de Investigação e Desenvolvimento, o Arquivo.pt lançou um Catálogo de 13 ferramentas inovadoras para apoiar a preservação de informação em linha. Estas ferramentas estão ao dispor dos cidadãos e organizações para que processos de transformação digital, como por exemplo a renovação de um website institucional, possam ser realizados de forma mais eficaz e eficiente, evitando perdas de informação. Qualquer cidadão pode armazenar, pesquisar e aceder a informação histórica preservada da Web desde a década de 1990. Os serviços do catálogo do Arquivo.pt (https://arquivo.pt/catalogo) são os seguintes:

- Pesquisa e acesso (arquivo.pt): inclui pesquisa sobre os textos e imagens das páginas, listagem de histórico de versões arquivadas de uma determinada página, pesquisa avançada, geração automática de narrativas e reprodução de conteúdo arquivado com 6 opções complementares (ex. “Detalhes técnicos”, “Completar a página” ou “Ver com browser antigo”);
- Interfaces de programação para aplicações (arquivo.pt/api): facilitam o desenvolvimento de aplicações de valor acrescentado por terceiros que utilizem automaticamente os serviços de pesquisa e acesso (API Arquivo.pt, API Image Search, API CDX-server, API Memento);
- Sugerir websites (arquivo.pt/sugerir): qualquer cidadão pode sugerir websites para passarem a ser preservados para memória futura. Apenas é necessário submeter o endereço da página inicial. Opcionalmente, podem fornecer um email para que sejam notificados quando o website sugerido estiver disponível no Arquivo.pt, e possam avaliar a qualidade do conteúdo arquivado;

- SavePageNow (arquivo.pt/savepagenow): permite aos cidadãos arquivar imediatamente páginas web no Arquivo.pt. Apenas necessitam de introduzir o endereço de uma página e iniciar a navegação para que todo o conteúdo visitado seja preservado. Permite por exemplo, preservar todas as páginas de um pequeno website de forma autónoma;
- Integração de coleções históricas de dados da web (arquivo.pt/doar): o Arquivo.pt iniciou a preservação de informação publicada na web em Janeiro de 2008. No entanto, fontes externas têm doado conteúdos históricos anteriormente publicados para serem salvaguardados;
- Formação (arquivo.pt/forma): é um programa de formação gratuito que visa conscientizar acerca da importância de preservar o legado digital e disseminar boas práticas de publicação e preservação digital nas Tecnologias de Informação e Comunicação. É composto por quatro módulos: "Arquivo.pt: uma nova ferramenta para pesquisar o passado", "Bem publicar, para bem preservar", "Acesso e processamento automático de informação preservada da Web através de APIs" e "Arquivar a Web: faça-você-mesmo!";
- Dados abertos (arquivo.pt/dadosabertos): são conjuntos de dados que contêm metadados sobre os objetos digitais preservados, como listas de URLs que documentam eleições. Estes conjuntos de dados foram reutilizados e melhorados por outras organizações também interessadas em preservar este legado digital, como por exemplo, Museus. O Arquivo.pt é fornecedor oficial do Portal de Dados Abertos da Administração Pública;
- CitationSaver (arquivo.pt/citationsaver): extrai links de documentos e preserva os objetos digitais citados para que possam vir a ser posteriormente recuperados a partir do Arquivo.pt. Os documentos convencionais são criados para serem impressos, por exemplo em formato PDF, mas citam objetos digitais em linha referenciando os seus URLs. Porém, quando esses links se tornam inacessíveis, mesmo os documentos impressos perdem a integridade porque suas citações ficam inacessíveis;
- Arquivo404 (arquivo.pt/arquivo404): apresenta páginas preservadas em vez de mensagens de erro (ex. "Erro 404: Página não encontrada"). Os webmasters só necessitam de inserir uma única linha de código na página que gera a mensagem de erro 404. Quando um visitante do website tenta aceder a uma página que já não está disponível, o Arquivo404 verifica automaticamente se existe uma versão arquivada daquela página. Se existir, apresenta um link para a página arquivada para que o

visitante possa aceder a esta informação, em vez de desistir or ir procurá-la noutro website;

- Memorial (arquivo.pt/memorial): preserva a informação publicada num website após a sua desativação. Os custos de manutenção aumentam à medida que os websites envelhecem devido à obsolescência das tecnologias de suporte e às consequentes perigosas vulnerabilidades de segurança. O Memorial oferece preservação de alta qualidade do conteúdo histórico de um website desativado. O nome do domínio original é mantido, não ocorrem links quebrados para as páginas do website e todo o conteúdo mantém-se pesquisável através dos motores de busca (ex. Google);
- Arquivo de alta qualidade (a-pedido): permite a preservação em alta-qualidade de websites selecionados que são arquivados e curados em colaboração com os seus donos usando a melhor combinação de tecnologias disponíveis;
- Criação de coleções e exposições temáticas (arquivo.pt/expos): são exposições em linha de páginas web preservadas, organizadas por tema e com curadoria realizada em colaboração com instituições especialistas na área (ex. imprensa, rádio, municípios, unidades de I&D, escolas ou museus). Cada exposição é seguida de campanhas de divulgação promovidas pelas instituições parceiras que amplificam a consciência para a importância da preservação da informação digital;
- Exposição itinerante de cartazes em instituições externas (arquivo.pt/posters): a desvantagem de preservar exclusivamente artefatos nascidos digitalmente é que se torna um desafio atrair a atenção de potenciais novos utilizadores no mundo físico. Muitas iniciativas de preservação digital dependem de métodos digitais para preservar documentos impressos. Invertemos esta estratégia e imprimimos um conjunto de cartazes com páginas da web históricas (ex. a primeira página da web portuguesa) para sensibilizar sobre a pertinência de preservar o legado nado-digital.

A Fundação para a Ciência e a Tecnologia, Instituto Público é responsável pela sustentabilidade económica do serviço público Arquivo.pt (Decreto-Lei 55/2013).

Os serviços do Arquivo.pt são utilizados por cidadãos e instituições de todo o mundo

Country	Users	% Users
1. Portugal	72,493	44.96%
2. United States	23,006	14.27%
3. Brazil	6,356	3.94%
4. United Kingdom	5,394	3.35%
5. Japan	4,231	2.62%
6. Germany	3,460	2.15%
7. Canada	2,667	1.65%
8. India	2,494	1.55%
9. Russia	2,414	1.50%
10. Spain	2,340	1.45%

Figura 5. Distribuição geográfica da origem dos utilizadores do Arquivo.pt.

O principal objetivo da criação do catálogo de serviços do Arquivo.pt foi apoiar a preservação digital, disponibilizando um conjunto de serviços de acesso gratuito a um vasto leque de utilizadores, para que qualquer utilizador da Internet possa gerir de forma eficaz e eficiente o ciclo de vida da informação digital publicada em linha. Cerca de metade dos utilizadores do Arquivo.pt são internacionais (ver Figura anterior).

Como diferentes utilizadores têm diferentes necessidades em relação à preservação digital, fornecer um catálogo abrangente de ferramentas potencia a resposta à maioria dos requisitos dos diferentes utilizadores. Desta forma, arquivistas, investigadores, especialistas em Informática ou utilizadores comuns da Internet podem contribuir para a preservação digital de objetos em linha, utilizando as ferramentas do Arquivo.pt para selecionar, adquirir, armazenar, aceder, reutilizar e divulgar informações históricas valiosas publicadas em linha ao longo dos últimos 30 anos.

O Arquivo.pt é também um catalisador de inovação tecnológica. A quantidade de casos de uso de um arquivo da web é vasta. Os investigadores são utilizadores assíduos e existem pelo menos 800 trabalhos científicos relacionados com o Arquivo.pt. O Prémio Arquivo.pt distingue anualmente trabalhos inovadores que utilizaram o Arquivo.pt. Ao longo de 8 edições, foram recebidas 194 candidaturas e os 29 trabalhos premiados demonstram claramente como os arquivos web beneficiam amplamente o legado digital, abrangendo todas as áreas do conhecimento, como Saúde, Humanidades Digitais ou Ciências da Computação.

A lista dos vencedores do Prémio Arquivo.pt está disponível em https://arquivo.pt/vencedores. As ferramentas do Catálogo têm sido utilizadas e ampliadas pelos candidatos aos prémios Arquivo.pt e utilizadas para produzir conjuntos de dados para investigação.

As exposições em linha têm sido promovidas por instituições ligadas ao Património para complementar as suas coleções (ex. Museu do Turismo) ou celebrar eventos através da criação de "Viagens no Tempo" (ex. Museu da Presidência da República). O serviço SavePageNow recebeu 45 000 solicitações em 2023 e foi usado pela Agência de Imprensa Alemã (DPA) para proteger informações de verificação de factos ou por utilizadores da Wikipedia para proteger links para citações externas. Em 2023, as ferramentas do Arquivo.pt forneceram acesso a 470 TB de informação preservada (100 milhões de pedidos à API). Outras estatísticas em tempo-real acerca do Arquivo.pt podem ser consultadas em linha em https://arquivo.pt/numeros.

Arquivo.pt é inovador a nível mundial

O Arquivo.pt foi o primeiro arquivo da web do mundo público a suportar pesquisa de páginas e imagens sobre todo o seu acervo. O desenvolvimento do Arquivo.pt levantou desafios em áreas como Information Retrieval, User Experience ou Machine Learning (Inteligência Artificial) que tiveram de ser superados autonomamente para criar o serviço atual adaptando as melhores práticas de TIC (ex. SRE, DevOps, ITIL). Todo o software desenvolvido está disponível em regime de código-aberto gratuito (https://github.com/arquivo/) para que possa ser aplicado noutras instituições ou países. Toda informação preservada está disponível em acesso aberto através de múltiplos métodos de acesso para permitir a sua ampla reutilização ao longo do tempo. As ferramentas do Arquivo.pt que suportam o acesso tanto por humanos quanto por máquinas, incluem interfaces web de utilização, APIs e suporte para *downloads* em massa (arquivo.pt/api#bulk) para apoiar outras atividades de processamento (ex. Big Data).

Além de desenvolver, manter e operar o serviço, os membros da equipa Arquivo.pt publicaram mais de 30 artigos científicos em acesso-aberto tendo em vista a partilha de experiências, incluindo o livro "The Past Web: Exploring Web Archives", e mais de 20 relatórios técnicos que incluem teses de mestrado e doutoramento. O Arquivo.pt é um dos arquivos da web de referência mundial.

Arquivos da web para sociedades digitais robustas e resilientes

A informação publicada em linha é património intelectual e a sua perda contínua compromete a sustentabilidade económica das organizações. O PIB de Portugal em 2023 foi de 267 384 milhões de euros. Estima-se que foram investidos 516 000 milhões de euros na produção da informação preservada no Arquivo.pt, que teriam sido desperdiçados se não existisse este serviço.

O Arquivo.pt é uma ferramenta para a segurança da informação para que esta não se perca e pode ser usada para áreas tão importantes como a Cibersegurança. O Memorial preserva a informação de websites antigos que já não são atualizados, evitando vulnerabilidades de segurança. Os dados-web históricos suportam investigações forenses e as APIs permitem análises automáticas em larga escala. O SavePageNow guarda evidências para posterior investigação. O Arquivo.pt contribui também para reagir a ataques maliciosos através da redireção temporária de tráfego para a informação arquivada quando os websites sucumbem a ataques maliciosos como aconteceu no ataque ao Instituto Politécnico de Leiria (Ferreira 2023).

A reutilização de informação preservada da web também contribui para a sustentabilidade ambiental. O Memorial poupa recursos ao manter acessível a informação de websites históricos e o Arquivo404 mostra páginas preservadas em vez de “páginas não encontradas”. Ambos os exemplos evitam que se tenham de desperdiçar recursos para recriar informação que já havia sido produzida no passado.

O Arquivo.pt é uma fonte de informação única que gera economias de escala e tem permitido derivar tendências, treinar modelos grandes de linguagem (LLM) para Inteligência Artificial (Lopes, 2024) ou recuperar trabalhos julgados perdidos. O programa de formação do Arquivo.pt ministra boas práticas de preservação digital que agilizam o desenvolvimento de sistemas de informação em linha e reduzem custos de manutenção.

Conclusões

Em apenas 30 anos, os objetos digitais que circulam em linha tornaram-se o meio de comunicação dominante, enquanto os meios impressos, anteriormente prevalentes,

tornaram-se um produto de luxo. Perder o legado digital dos objetos em linha coloca em risco a sustentabilidade das organizações porque representam a grande maioria da informação utilizada para organizar a Humanidade. Objetos digitais em linha preservados também são uma fonte única de informação para derivar tendências evolutivas, treinar modelos de Inteligência Artificial ou recuperar obras consideradas perdidas.

A preservação digital exige a aquisição de objetos digitais em risco antes que desapareçam, armazenando-os com segurança e mantendo-os acessíveis para que continuem úteis aos cidadãos do presente e do futuro. O Arquivo.pt foi o primeiro serviço de preservação digital a suportar um catálogo tão abrangente de ferramentas para salvaguardar o legado digital de objetos em linha. As suas ferramentas abordam todas as etapas da preservação digital. Por exemplo, o SavePageNow e o CitationSaver suportam métodos alternativos para a aquisição de objetos digitais em linha. Os serviços de pesquisa, o Memorial e as coleções e exposições temáticas suportam o acesso aos objetos preservados. Em 2023, as ferramentas do Catálogo do Arquivo.pt receberam 1,7 milhões de utilizadores únicos e forneceram acesso a 470 TB de objetos digitais preservados (100 milhões de solicitações de API).

Sem memória, não é possível sustentar sociedades humanas a longo prazo. O Arquivo.pt é uma infraestrutura de memória para as sociedades digitais.

Referências

Chapekis A., Bestvater S., Remy E., Rivero G., When Online Content Disappears, Pew Research Center Report, 17 de maio de 2024.

Gomes D., Web archives as research infrastructure for digital societies: the case study of Arquivo.pt, Archeion 123, 2022.

Lopes T, Magalhães J., Semedo D., GlórIA: A Generative and Open Large Language Model for Portuguese, Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics, 2024.

Ferreira M. L., Polícia Judiciária investiga ataque informático no Instituto Politécnico de Leiria, Jornal Público, 4 de maio de 2023.