Prémio Transformação Digital 2024 - APDSI

A. **Identificação do prémio ao qual se candidata**: Prémio Transformação Digital 2024.

B. **Identificação da categoria onde o projeto se enquadra**: Categoria Promoção da Sociedade mais Inovadora e Digital.

C. **Identificação da denominação do Projeto**: Arquivo.pt - uma infraestrutura de memória para as sociedades digitais.

D. **Identificação dos autores com indicação clara do porta-voz do grupo de trabalho e respetivo contacto de e-mail**: Daniel Gomes, daniel.gomes@fccn.pt

E. **Identificação da entidade onde o projeto foi desenvolvido**: Fundação para a Ciência e a Tecnologia I.P. - Unidade FCCN

F. **Breve descrição da entidade representada onde conste a respetiva missão, enquadramento geográfico**: A FCCN, serviços digitais da FCT – Fundação para a Ciência e a Tecnologia, tem como propósito contribuir para o desenvolvimento da Ciência, Tecnologia e Conhecimento em Portugal.

Sumário executivo

Transição digital causa perda de informação importante

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 é regida por informação digital publicada exclusivamente online.

No entanto, a memória desta informação de valor inestimável tem sido continuamente perdida. Apesar da informação online 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 recente do Pew Research Center, 38% das páginas Web que existiam em 2013 já não estão acessíveis passados 10 anos.

Portanto, são necessárias soluções tecnológicas inovadoras prontas a utilizar para salvaguardar a informação publicada online, 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”.

Arquivo.pt preserva a informação publicada online

O Arquivo.pt é um serviço público que preserva informação publicada online. 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. Cerca de metade dos utilizadores do Arquivo.pt são internacionais.

Após 15 anos de Investigação e Desenvolvimento, o Arquivo.pt lançou um Catálogo de 13 ferramentas inovadoras para apoiar a preservação de informação online. 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 (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 Reproduzir com *browser* antigo);
- Interfaces de programação para aplicações (arquivo.pt/api): facilita 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, pode fornecer um email para que seja notificado quando o website estiver disponível no Arquivo.pt e possa 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 TIC. 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 úteis para terceiros, 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 (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 destinados a serem impressos (por exemplo, em formato PDF) citam objetos digitais online 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 ao visitante para que 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 de alta qualidade de websites selecionados que são arquivados e curados em colaboração com os donos dos websites usando a melhor combinação de tecnologias disponíveis;
- Criação de coleções e exposições temáticas (arquivo.pt/expos): são exposições online de páginas web preservadas, organizadas por tema e com curadoria 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

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 online.

Como diferentes utilizadores podem ter diferentes necessidades em relação à preservação digital, fornecer um catálogo abrangente de ferramentas potencia o cumprimento da maioria dos requisitos. Desta forma, investigadores, profissionais da informação, especialistas em TI ou utilizadores comuns da Internet podem contribuir para a preservação digital de objetos online, utilizando estas ferramentas para selecionar, adquirir, armazenar, aceder, reutilizar e divulgar informações históricas valiosas publicadas online 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 enorme. 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 (vencedores em arquivo.pt/vencedores). As ferramentas do Catálogo têm sido utilizadas e ampliadas pelos candidatos aos prémios Arquivo.pt ou para produzir conjuntos de dados de investigação.

As exposições online 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 acerca do Arquivo.pt podem ser consultadas online (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 (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 suportam o acesso tanto por humanos quanto por máquinas, incluindo interfaces web de utilização, APIs e download 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 promoção de 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 online 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 possa ser usada para 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 colapsam (ex. ataque ao IP Leiria).

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 de Inteligência Artificial (Large Language Models) 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 online e reduzem custos de manutenção.

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