O papel do Arquivo.pt na Ciência e Ensino Superior

Daniel Gomes, daniel.gomes@fccn.pt
Agosto de 2020

Arquivo.pt é uma infraestrutura para investigação científica

A missão do Arquivo.pt é a preservação de informação publicada na Web para fins científicos e académicos. Esta infraestrutura de investigação consta no Registry of Research Data Repositories e é utilizada por investigadores internacionais como fonte de dados abertos.

O Arquivo.pt é uma infraestrutura de investigação gerida pela Fundação para a Ciência e a Tecnologia (FCT) que permite pesquisar e aceder a páginas da web arquivadas desde 1996.

A FCT|FCCN é uma unidade orgânica da Fundação para a Ciência e a Tecnologia, I.P. (FCT) que gere infraestruturas digitais para investigação. A FCT|FCCN gere e opera o Arquivo.pt de acordo com o Artigo 3º alínea 2n) do Decreto-Lei 55/2013.

O Arquivo.pt preserva dados científicos internacionais

O Arquivo.pt tem vindo a desenvolver diversas atividades para a identificação de dados online relacionados com projetos de I&D para que sejam preservados de forma sistemática.

Os sites dos projetos de Investigação & Desenvolvimento (I&D) são cada vez mais usados para disponibilizar importante informação científica que complementa a literatura publicada (ex. conjuntos de dados ou documentação, software). Porém, a informação online relativa a projetos de I&D não tem sido exaustivamente documentada.

Por exemplo, a informação referente aos endereços dos sites dos projetos financiados no programa 7º Programa-Quadro (FP7) disponibilizada através do Portal de Dados Abertos da União Europeia (EU Open Data Portal) está omissa para 92% dos projetos. O Arquivo.pt já identificou automaticamente e preservou mais de 52 milhões de ficheiros (7 TB) oriundos de 53 993 sites de projetos de I&D financiados pela União Europeia desde o FP4 (1994).

O Arquivo.pt guarda a memória da Ciência nacional

Ao nível nacional, foram preservados 600 721 ficheiros (72 GB) recolhidos de 7 956 sites relacionados com projetos financiados pela Fundação para a Ciência e a Tecnologia. Desde 2020, a informação online relativa a projetos financiados pela FCT passou a ser documentada nos relatórios de progresso e finais para que passe a ser sistematicamente preservada.

O Arquivo.pt tem realizado recolhas especiais direcionadas para preservar informação científica nacional disponível online citada a partir de publicações científicas em acesso aberto (RCAAP) e currículos científicos (Ciência Vitae).

O serviço Memorial do Arquivo.pt tem preservado websites de eventos, projetos os portais científicos que já não são atualizados, como por exemplo o Degois.pt.

Os websites de unidades de Investigação e Desenvolvimento são periodicamente recolhidos para preservação. Estas atividades visam principalmente manter a validade das referências científicas para recursos online em publicações *peer-reviewed* e CVs académicos.

Formação acerca de preservação de informação *online*

A equipa do Arquivo.pt tem vindo a ministrar um programa de formação cujos objetivos são capacitar os formandos para conseguirem:

- publicar dados abertos online por forma a que possam ser preservados para o futuro;
- preservar os dados fonte das suas investigações online e auto-preservar os resultados científicos derivados que sejam publicados online;
- pesquisar, aceder e reutilizar dados históricos oriundos da web;
- processar de forma automática grandes volumes de dados históricos preservados da web através de Interfaces de Programação de Aplicações (API).

Uma nova fonte de dados para a investigação

O Arquivo.pt tem contribuído para a produção de conjuntos de dados e software em acesso aberto. Todo o software que suporta o serviço Arquivo.pt e as experiências de investigação realizadas estão disponíveis através de uma conta no GitHub.

O Arquivo.pt disponibiliza dados abertos valiosos para investigação tais como os registos históricos de recolhas, de pesquisas temporais por texto e imagem (únicos no mundo) e os dados preservados desde 1996 através de recolha proactiva da web e integração de coleções históricas.

O Prémio Arquivo.pt galardoa trabalhos que utilizem os dados abertos preservados pelo Arquivo.pt e os resultados dos trabalhos são disponilizados em acesso aberto como condição do Regulamento (Licença Creative Commons Atribution By.)

Colaboração com a comunidade científica

Estabelecendo colaborações com organizações da comunidade científica e académica, o Arquivo.pt pode contribuir para:

- Capacitar a comunidade de formação que permita publicar online dados abertos por forma a que possam ser preservados e reutilizados no futuro;

- Preservar sistematicamente a produção online de dados abertos ou de informação acerca dos conjuntos de dados abertos (ex. documentação, apresentações, vídeos, etc.);
- Disponibilizar acesso a dados abertos preservados da web;
- Produzir conjuntos de dados abertos para investigação tais como coleções especiais temáticas acerca do Covid-19, eleições nacionais ou europeias (multi-lingue).

Para saber mais

Projetos internacionais relacionados

- CLEOPATRA - Marie Skłodowska-Curie Innovative Training Network
- RESAW -Research Infrastructure for the Study of Archived Web Materials
- ASAP - Archives sauvegarde attentats Paris
- WARCNet - Web ARChive studies network researching web domains and events
- IIPC collaborative collections
- Internet Archive - non-profit library of millions of free books, movies, software, music, websites, and more.
- Archives Unleashed - aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past
- Alexandria - aims to develop models, tools and techniques necessary to explore and analyze Web archives in a meaningful way.
- Memento - Memento wants to make it as straightforward to access the Web of the past as it is to access the current Web.

Referências adicionais

- Arquivo.pt in a nutshell: overview of services and activities, https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-in-a-nutshell:-overview-of-services-and-activities, 2018.
- Daniel Bicho, Daniel Gomes, Preserving Websites Of Research & Development Projects, International Conference on Digital Preservation, 2016.
- Daniel Gomes, João Miranda, Miguel Costa, A survey on web archiving initiatives, International Conference on Theory and Practice of Digital Libraries 2011, 2011.
- Web Archiving, Julien Masanès, 2006.

- The SAGE Handbook of Web History, Niels Brügger & Ian Milligan, 2018
- The Historical Web and Digital Humanities: The Case of National Web Domains, Niels Brügger, Ditte Laursen, 2018
- History in the Age of Abundance? How the Web Is Transforming Historical Research, Ian Milligan, 2019
- Remembering and Forgetting in the Digital Age, Thouvenin, F., Hettich, P., Burkert, H., Gasser, 2018.
- The Archived Web: Doing History in the Digital Age, Niels Brugger, 2018.

Exemplos de aplicações científicas dos dados abertos preservados pelo Arquivo.pt

- Sawood Alam et al., MementoMap Framework for Flexible and Adaptive Web Archive Profiling, In Proceedings of Joint Conference on Digital Libraries 2019, 2019.
- AlSum, Ahmed, et al., Profiling web archive coverage for top-level domain and content language, *International Journal on Digital Libraries*, 2014.
- Miguel Costa, Information Search in Web Archives, Tese de Doutoramento, 2014.
- A Garzó et al., Cross-lingual web spam classification, Proceedings of the 22nd International Conference on World Wide Web, 2013.
- Rui Lopes et al., Web Not For All: A Large Scale Study of Web Accessibility, W4A '10: Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A), 2010.