14º Congresso Nacional BAD

POSTER

CitationSaver – Preserva as referências Web dos artigos científicos

Ricardo Basílio[a], Pedro Gomes[b]

[a]*Fundação para a Ciência e a Tecnologia, Portugal, ricardo.basilio@fccn.pt*

[b]*Arquivo.pt – FCT-FCCN, Portugal, pedro.gomes@fccn.pt*

Resumo

CitationSaver é uma nova funcionalidade, um novo serviço, criado pelo Arquivo.pt para preservar os conteúdos Web, tais como URLs (endereços Web) referenciados em artigos científicos, no corpo do texto ou na bibliografia final. O processo consiste em identificar automaticamente os endereços Web presentes num qualquer texto, depois extrai-los para serem gravados pelo Arquivo.pt. Ao fazer isto ficará gravada e acessível no Arquivo.pt uma cópia da página Web ou do recurso que um determinado autor pretendeu referenciar para sustentar a sua investigação. O serviço funciona via Web, no site do Arquivo.pt. Qualquer pessoa pode carregar um documento no formato PDF ou simplesmente copiar o texto de onde pretende salvar os endereços Web. Depois o Arquivo.pt faz o resto. Desta forma previne-se o desaparecimento de conteúdos referenciados em artigos, robustecendo o conhecimento científico e servindo a comunidade. O Arquivo.pt conta com a colaboração da comunidade para salvar o maior número de referências a conteúdos da Web.

Palavras-chave: Preservação digital, Ciência Aberta, Curadoria digital, Repositórios científicos

O problema das ligações quebradas

Os artigos científicos fazem referência a conteúdos da Web, tanto no corpo do texto como nas referências bibliográficas e recursos complementares. Muitas dessas ligações quebram-se com o passar do tempo, porque os conteúdos referenciados deixam de estar online ou mudam de endereço.

Ao apresentarem ligações quebradas os artigos perdem parte da sua consistência e do seu valor científico.

Como fazer para preservar as ligações e os conteúdos Web referenciados, de modo a poderem ser consultados no futuro?

Objetivo

Pretende-se apresentar o novo serviço do Arquivo.pt chamado CitationSaver, focando-nos na sua utilização por parte da comunidade.

Metodologia

A criação do CitationSaver para uso da comunidade deriva da forma como o Arquivo.pt quer situar-se no ecossistema da produção de Ciência. A missão do Arquivo.pt é preservar e dar acesso a conteúdos Web. Não é desenvolver *software*. Para criar este serviço começou por identificar o problema das ligações quebrados nas publicações científicas. Comparou e testou o *software* disponível. Definiu um *workflow* em que o processamento é automatizado e invisível para o utilizador. Documentou o desenvolvimento no Github. Criou uma interface de utilizador no site do Arquivo.pt. Lançou o serviço em fase experimental, para obter *feedback*, detetar erros e fazer melhorias no serviço.

Resultados

Como resultado foi desenvolvida uma interface para qualquer pessoa utilizar o CitationSaver. Esta ferramenta preserva o conteúdo das ligações citadas (ex. páginas web citadas num livro) para que possam ser recuperadas mais tarde a partir do Arquivo.pt.

Está acessível em https://arquivo.pt/citationsaver/.

O processo pode representar-se em quatro passos:

1) Utilizador insere o artigo no CitationSaver
2) CitationSaver extrai todos endereços (URLs, tudo o que começa por http)
3) CitationSaver valida e grava automaticamente lista de endereços
4) Arquivo.pt disponibiliza conteúdos gravados

Existem três opções para extrair URLs de um artigo usando o CitationSaver

a) URL - indicar o URL ou endereço Web onde o artigo se encontra publicado;
b) Ficheiro – fazer *upload* do ficheiro PDF ou Word;
c) Texto – copiar e colar texto onde estão referências a conteúdos Web.

Os conteúdos gravados através do CitationSaver ficam disponíveis no Arquivo.pt num período mais curto do que as recolhas gerais do Arquivo.pt, ou seja, algumas semanas. Assim, quem aceder a essa publicação e encontrar uma ligação quebrada pode recuperar o seu conteúdo no Arquivo.pt.

Para testar a capacidade de processamento automático do serviço, foram submetidos 30 mil ficheiros PDF de publicações científicas, das quais foram extraídas 670 mil links (URLs).

O *software* utilizado, a resolução de problemas e demais detalhes técnicos estão documentados no canal Github do Arquivo.pt (https://github.com/arquivo/CitationSaver).

Conclusões

O CitationSaver, tal como está disponibilizado pelo Arquivo.pt, permite aos profissionais incluir esta ferramenta na agenda da preservação digital institucional.

A comunidade BAD tem um papel fundamental para criar uma mentalidade de preservação, incluindo os conteúdos da Web.

O contributo do CitationSaver para a preservação de conteúdos citados nas publicações científicas é tanto mais significativo se o considerarmos em grande escala, aplicado a milhares de documentos.

O CitationSaver enquadra-se na missão do Arquivo que é "promover a preservação de conteúdos disponíveis na Internet nacional, garantindo a disponibilização deste à comunidade científica e ao público em geral" (Decreto-Lei n.º 55/2013, de 17 de abril), mas a sua eficácia é maior com o apoio da comunidade.