Você está em: Link > **.pt em arquivo**

# .pt em arquivo

- 26 de abril de 2010|
- 19h11|
- Tweet este Post

Por Tatiana de Mello Dias

Em Portugal, tudo o que é publicado sob o domínio .pt está sendo guardado para sempre. Posts em blogs, notícias, sites públicos e privados. O conteúdo está sendo guardado no projeto Arquivo da Web Portuguesa, que já tem quase 15 TB de dados armazenados.

O projeto, capitaneado pela Fundação para a Computação Científica Nacional, conta com a ajuda dos internautas para armazenar toda essa informação. O rARC (replicador de ficheiros ARC) permite que pessoas comuns doem espaço em seus computadores para proteger os sites. É um sistema parecido com o utilizado no Internet Archive, projeto que inspirou o Arquivo.

O maior problema, porém, não é guardar – mas, sim, conseguir um motor de busca eficiente para a geração Google. A busca no Arquivo já está funcionando em versão experimental. O Link conversou com o coordenador do projeto, Daniel Gomes, para saber mais sobre a iniciativa.


*Daniel Gomes*

**Como surgiu o Arquivo da Web?**
Começou em 2001 como um projeto de pesquisa. Foi uma parceria entre a Biblioteca Nacional e a Faculdade de Ciências. Era um sistema experimental, com apenas algumas publicações relacionadas. Nós verificávamos se estava tudo bem guardado para depois ser aquivado definitivamente. Em 2006 nós criamos um primeiro protótipo que permitia pesquisar sobre a informações históricas entre 2001 e 2006. E em 2007, quando eu acabei o meu doutorado, fui convidado pra iniciar o Arquivo da Web de uma forma oficial.

**O que fica armazenado? Há uma seleção do que é relevante?**

Nós não fazemos qualquer tipo de avaliação. Tentamos fazer o arquivo mais exaustivo possível. Nós

arquivamos tudo o que conseguimos encontrar sob o domínio .pt, mas também sites de fora que tenham sido sugeridos pelos utilizadores. Trabalhamos também com a Biblioteca Nacional para obtenção desses sites. Basicamente todas as pessoas podem contribuir com sugestões.

**Como funciona a parte técnica? O volume de dados deve ser imenso, não?**

Não, não é muito grande. Espaço comprimido ocupa muito menos espaço. Porque parte dos arquivos que estão na web são textos, e os textos têm o tamanho reduzido quando comprimidos. No fim de 2009 tinha 14,5 TB arquivados. Temos o grande objetivo de arquivar grandes publicações todos os dias. E, nesse contexto, o volume de informação vai ser muito maior. Mas essa não é a parte mais complicada, porque hoje em dia temos um bom acesso à internet e muitos computadores. A parte mais complicada é conseguir fazer um sistema de pesquisa que funcione rapidamente sobre toda essa informação. Não adianta: para guardar informação, você tem que conseguir pesquisar. Essa é a parte
complicada porque as pessoas estão acostumadas ao Google e querem a resposta em poucos segundos. Temos que conseguir dar as respostas, senão os utilizadores ficam frustrados.

**Vocês se inspiraram em algum projeto pra fazer o arquivo?**

Sim. No Archive Access, do Internet Archive, onde existem uma série de ferramentas pra fazer o Arquivo da Web. Quando começamos o nosso projeto, baixamos essas ferramentas. São gratuitas e de código aberto. Começamos a tentar fazer o nossos sitemas a partir disso. Só que essas ferramentas não permitiam dar os tempos de resposta que nós precisávamos. Uma pesquisa com cerca de 40 mil documentos demorava vários segundos. Então nós começamos a melhorar os códigos para termos tempos melhores. E conseguimos ter respostas bantante boas. Agora na nossa versão experimental o tempo médio de resposta é abaixo de cinco segundos.

**A internet é efêmera. Não estamos acostumados à memória dela. Qual é a importância de guardar a historia da web?**

A resposta que eu costumo dar a isso: é tão importante guardar as publicações que estão na web quanto é importante nós guardarmos todas as publicações impressas. Quando guardamos as publicações impressas, por exemplo no jornal, também está lá o correio do leitor, onde as pessoas mandam suas opiniões tal como hoje em dia enviam suas opinões
para blogs. É um meio de comunicação diferente em que acontece tudo o que acontecia antes, só que em uma escala muito maior. Pode servir como fonte de pesquisa e também dá para verificar a evolução da opinião acerca de determinado assunto.

**A Biblitoeca do Congresso Americano adquiriu o acervo do Twitter. E aí entram algumas questões de privacidade. As pessoas que postaram coisas no Twitter não sabiam que aquilo ficaria guardado pra eternidade. Vocês lidam com essas questões também?**

Nós ainda não tivemos queixas. Mas o que eu posso dizer acerca disso é que há uma falta de conscientização

das pessoas. As pessoas não têm ideia que quando publicam algo na internet estão publicando em uma escala muito superior ao que se via antes em papel. Por exemplo: uma carta de leitor no jornal. O jornal tem um alcance muito limitado. Hoje em dia, quando você publica algo na web, o alcance é mundial. E, portanto, qualquer pessoa dos milhões de utilizadores pode fazer uma cópia, guardar e pronto. Não há nada que você possa fazer pra que aquela informação desapareça.
Em relação ao Arquivo da Web, essa questão de privacidade é um pouco discutível. Porque como a pessoa que publica uma informação para o mundo inteiro ver depois diz que essa informação é privada? Isso é um pouco estranho. Hoje há um problema que a legislação de direitos de autor varia de país pra país. A pessoa de qualquer maneira tem o direito sobre aquela informação que ela publicou. É a autora e, portanto, não podemos retirar esse direito.

Nós guardamos tudo, damos acesso a tudo, mas se algum autor disser que quer que o acesso a informação seja bloqueado, respondemos esse pedido e bloqueamos a informação. É a mesma coisa que o Internet Archive faz.
O que acontece é que é preciso provar que aquela informação é mesmo de minha autoria. Além disso, nós respeitamos um protocolo que se chama Robots Exclusion Protocol em que o autor do site quando publica uma informação pode escrever no ficheiro dele ‘não quero que o meu site seja arquivado’, ou ‘não quero que essa parte seja arquivada’. Isso também funciona nas buscas. O autor também pode dizer ‘quero que essa informção seja recolhida pelo Arquivo da Web Portuguesa mas não quero que seja recolhida pelo Google’. Já houve casos em tribunal com essa situação. Nos EUA foi dado razão aos motores de busca porque a pessoa não indicou que aquela informação era
pra ser excluida. Foi entendido que era pública.