

dados abertos

cartilha para desenvolvedores

W3C® WORLD WIDE WEB
Consortium
Escritório Brasil

cgj.br
Comitê Gestor de Internet
no Brasil

nic.br
Núcleo de Informação
e Coordenação do
Ponto BR



Laboratório Brasileiro
de Cultura Digital

Esta publicação é um encarte integrante do “Manual dos dados abertos: desenvolvedores”, resultante do acordo de cooperação técnico-científica entre o Laboratório Brasileiro de Cultura Digital e o Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

dados abertos

Para ser considerado aberto, todo dado público deve ser completo, primário (sem tratamento), atual, compreensível por máquina, não discriminatório, acessível, não proprietário e com licenças que garantam esses princípios e não cerceiem a liberdade de uso. Para quem desenvolve, a peça-chave é “compreensível por máquina”, princípio que possibilita o cruzamento e o reuso dos dados.

scraping

Muitos dos dados disponíveis publicamente não estão realmente abertos. Às vezes, estão disponíveis em tabelas HTML, arquivos em texto plano, em PDF ou em ambientes que pedem um *captcha* ou outra técnica que impeça múltiplas requisições. *Scrapers* são *softwares* que vasculham o *site* e traduzem os dados para formatos estruturados, como JSON ou XML, permitindo visualizações em formatos processáveis por máquinas/programas e/ou que permitam agregar informações complementares. As linguagens mais populares para a implementação de *scrapers* são Python, Ruby e PHP, linguagens multiplataforma de alto nível.

cinco formatos que todo desenvolvedor deve conhecer

HTML (HiperText Markup Language)

É a linguagem de marcação de hipertexto base para publicação na *web*. Na versão 5, uma série de novos elementos e atributos foram adicionados, inclusive alguns que melhor definem um modelo de página, facilitando a identificação de sua composição.

XML (Extensible Markup Language)

É, como a HTML, uma linguagem de marcação, mas extensível para descrever os dados que representa. É amplamente utilizada no intercâmbio de dados por meio da *web*.

CSV (Comma-Separated Values)

Arquivos no padrão CSV são arquivos em formato texto, representando conteúdo tabular separado por vírgulas e organizados sequencialmente por linhas. Por serem muito simples e de fácil reprodução, são amplamente difundidos na *web* e nos *sites* que disponibilizam dados.

RDF (Resource Description Framework)

É um dos principais formatos para a infraestrutura de *web* semântica e para a interoperabilidade de dados em aplicações vinculadas na *web*. Sua grande vantagem é a representação dos dados em estrutura triplificada, com descrição semântica dos campos que podem estar vinculados (*linkados*) a vocabulários.

JSON (JavaScript Object Notation)

É um formato nativo para uso com JavaScript, mas existem bibliotecas simples para uso em quase todas as linguagens de programação.

cinco ferramentas para abrir dados

Dapp Factory
(<http://open.dapper.net>)

O Open Dapper faz *scraping* de quase qualquer página *web* com apenas alguns cliques, e permite exportar os dados em XML e RSS, entre outros.

Scrapewiki
(<http://scrapewiki.com>)

É uma plataforma *on-line*, gratuita e livre para escrever e rodar *scrapers* colaborativamente. Suporta atualmente Python, PHP e Ruby, com diversas bibliotecas para captura e tratamento dos dados e uma boa quantidade de exemplos para se basear.

Google Refine
(<http://code.google.com/p/google-refine>)

É um *software* livre para limpar e vincular diferentes bases de dados, criar RDF triplos RDF e expor *web-services*.

YQL (Yahoo! Query Language,
<http://developer.yahoo.com/yql>)

É uma plataforma do Yahoo! que auxilia e facilita o *mashup* de dados. Com ele, consegue-se ler e parsear tabelas HTML, XML, CSV e vários outros formatos com alguma facilidade e em linguagem próxima do SQL.

Yahoo! Pipes
(<http://pipes.yahoo.com>)

O Pipes permite obter páginas inteiras ou *feeds* RSS e aplicar uma série de regras para produzir seu próprio RSS ou XML. É bem útil para extrair tabelas html.

três ferramentas de visualização de dados

Google Fusion
(<http://tables.googlelabs.com>)

É uma plataforma de visualização de dados do Google com suporte a arquivos grandes (CSV < 100 Mb). Entre suas funções estão o georreferenciamento automático a partir de texto, a criação de *heatmaps* e a exibição de linhas do tempo. Deve-se ter cuidado com o formato dos campos; datas, por exemplo, precisam estar em MM/DD/AA.

Many Eyes
(<http://www-958.ibm.com/>)

O ManyEyes foi criado pela brasileira Fernanda Viegas na época em que ela trabalhava na IBM. É feito em Java e permite várias visualizações interessantes, com destaque para aquelas baseadas em texto, como *tagclouds* e *wordtrees*. No entanto, só permite pequenos *datasets* e é um *software* completamente fechado.

Tableau

(<http://www.tableausoftware.com/>)

É uma suíte de *softwares* para tratamento e visualização de dados. A versão gratuita funciona como SaaS, e permite que várias visualizações e *dashboards* interativos sejam publicados na rede. O aplicativo roda apenas em Windows.

exemplos de listas com dados abertos,
para você começar “fazendo”

TCM-CE

<http://api.tcm.ce.gov.br>

Dados: execução orçamentária e fornecedores dos municípios do Ceará.

Formatos: API com XML e JSON.

CGU

<http://www.portaltransparencia.gov.br>

Dados: execução orçamentária do Governo Federal, repasses, contratos, convênios, etc.

Formato: CSV.

TSE

<http://spce2010.tse.jus.br/spceweb.consulta.prestacaoconta2010>

Dados: resultado de eleições, votos por candidato e região, doações de campanha.

Formatos: a maioria é CSV ou não estruturado.

MEC

http://esfera.mobi/datasets/mec_escolaspublicaseja2005.tar.gz

Dados: lista de escolas com Ensino de jovens e adultos, contendo endereços.

Formato: CSV.

