Converting data to friendly formats

Tools for scraping data from the web

PDFs

Scanned Images

Computer-generated (searchable)

Tables

Complex formats

Tables

- Copying and pasting into Excel does not usually work

Free online tools:

- cometdocs.com
- pdftoexcelonline.com
- zamzar.com
- pdftoexcel.org

Free desktop software

- tabula.nerdpower.org

Let's try it!

- Open the File: Country Profile President's Malaria Initiative (PMI) or Maternal Index
- Make sure you understand the data: read all codes, descriptions and legends
- Convert file using one of the free online programs
- Review xls or csv file and evaluate whether the resulting tables are usable
- Clean up table, eliminating unnecessary information and note data source

Complex Formats

- Columns, data divided among various lines, mixed with text
- Automatic tools probably won’t work
- The PDFs will have to be extracted and a programmer will use programming language to create a program to extract the data that you want

Scanned Images

Optical Character Recognition Software

- For documents that you can't search or select text
- Quality depends on quality of image, clarity of text
- Adobe Acrobat Professional: paid
- Google Docs: free
- Document Cloud: free. Advantage is that it allows you to annotate and publish your document as part of a multimedia publication. Your media outlet must request an account

Scanned Images

Optical Character Recognition Software

- For documents that you can't search or select text
- Quality depends on quality of image, clarity of text
- Adobe Acrobat Professional: paid
- Google Docs: free
- Document Cloud: free. Advantage is that it allows you to annotate and publish your document as part of a multimedia publication. Your media outlet must request an account

Let’s try it!

- Open Specimen Financial Statement
- Open Google Drive
- Select “Upload”
- Choose “Convert text from PDF and image files to Google Documents”
- Review quality
- Make corrections by comparing original file to OCR output

Scraping data from the web

- Browser scrapers:
  - Dafizilla Table2Clipboard for Mozilla Firefox
  - Scraper Extension for Chrome
- Allow you to select tables on a website, including specific rows and columns
- Copy the data tables and either create a Google spreadsheet or paste into a blank Excel file

Try it!

- Google: “demographics of India”
- Open wikipedia entry
- Highlight table
- Right click
- In Chrome, select “Scrape similar” and “Export to Google Docs”
- In Firefox, select “Table2Clipboard,” “Copy whole table” and paste in Google Doc or Excel

ScraperWiki: more complicated!

Webpage Screenshot

ScraperWiki

PROFESSIONAL SERVICES HELP BLOG LOG IN

Liberate your data with ScraperWiki!

ScraperWiki helps you do data science on the web.

Get, clean, analyse, visualise and manage your data, with simple tools or custom-written code.

ScraperWiki for businesses | Sign up – Free

I usually explore data in Excel, but I want to take my data analysis to the next level.

I want a place to write and run code that does awesome stuff with data.

I'm a business and I need professional data extraction, analysis and management.

Super-charge your data analysis with tools including…

https://scraperwiki.com/

Scraping for Journalism: A Guide for Collecting Data by ProPublica

Webpage Screenshot

The ProPublica Nerd Blog

Secrets for Data Journalists and Newsroom Developers

Scraping for Journalism: A Guide for Collecting Data

by Dan Nguyen
ProPublica, Dec. 30, 2010, 5:23 p.m.

17 Comments | k

Our Dollars for Docs news application lets readers search pharmaceutical company payments to doctors. We've written a series of how-to guides explaining how we collected the data.

Photo by Dan Nguyen/ProPublica

Most of the techniques are within the ability of the moderately experienced programmer. The most difficult-to-scrape site was actually a previous Adobe Flash incarnation of Eli Lilly's disclosure site. Lilly has since released their data in PDF format.

The News Apps Team

Our Tools and Style Guides

transcribable
Drop in crowdsourcing for your Rails app.

Need help scraping data?

- Ask the School of Data
- Join the data driven journalism listserve
- Join the Hacks/Hackers Nairobi Google group and go to the next meeting or post your question and ask for volunteers
- Visit the Internews data journalism team for support through the whole process, from finding and scraping data to visualization and storytelling