



Open Data in Practice

<http://training.theodi.org/InPractice>

David Tarrant · @davetaz

Session 2

Data discovery patterns

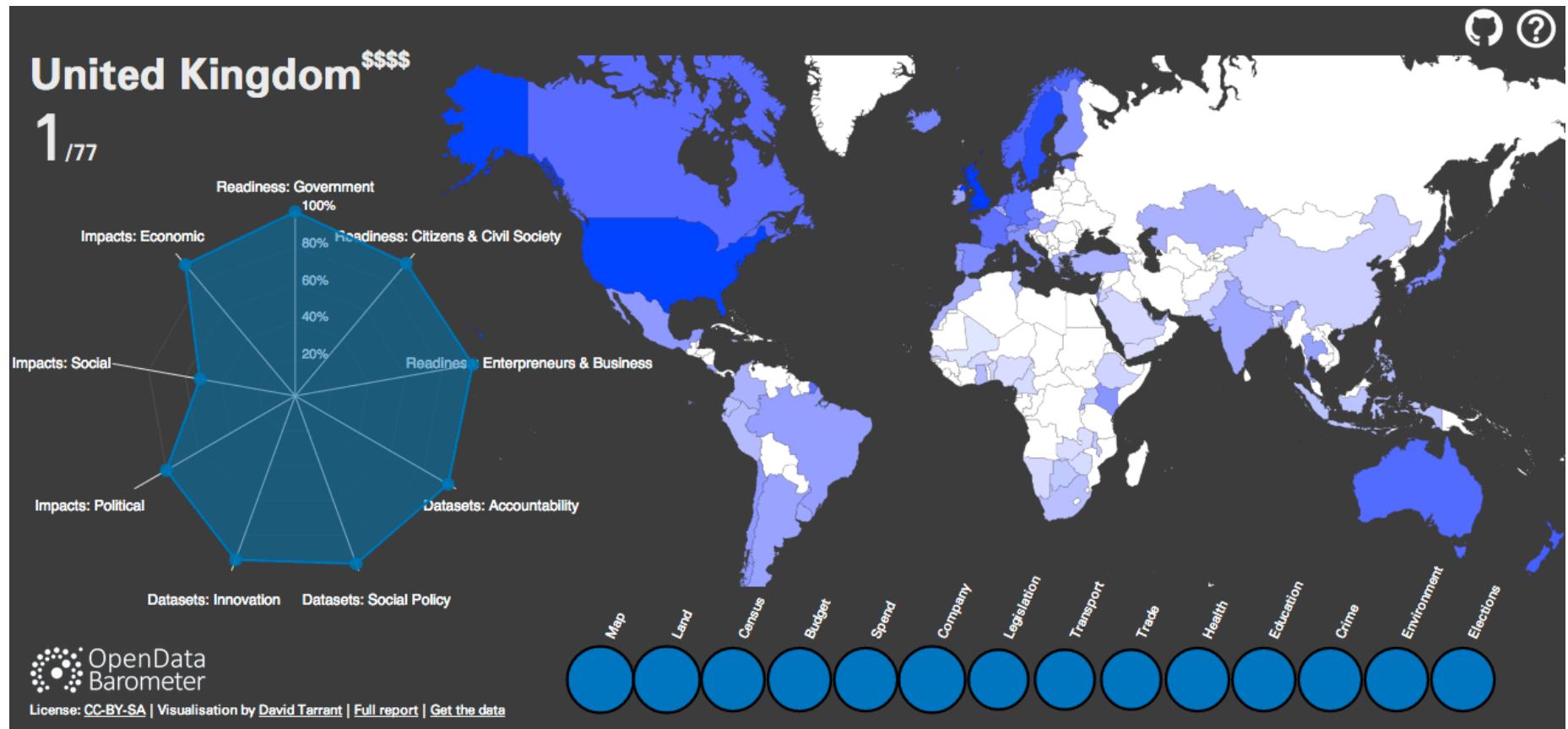


Finding data on the web **(of documents)**

- Government data
- Private sector data
- Google advanced
- Aggregators and portals
- Scraping



Government data



data.gov.XX

DATA.GOV.UK Beta
Opening up Government

Home Data Apps Interact

Datasets Map Search Data Requests Publishers Public Roles & Salaries Spend Reports Site Analytics Reports

Search for data... or conduct map-based search

SHOW ONLY... Published datasets (15186)

Accueil Comprendre l'Open Data Thèmes Producteurs Jeux de données Applications Partenaires A propos

THEMES

Tous les thèmes

Agriculture Éducation Diplomatie Infrastructures Eau et Environnement

TIC Santé Collectivités territoriales Sécurité Publique Tourisme et Culture

Open Data Burkina Faso statistiques

- 78 jeux de données
- 28 organisations
- 11 groupes

Business & Economy Education Energy & Environment Finance

DATOS.GOB.MX BETA
OPEN TO PARTICIPATE

Data Stories Advances

/ datasets

Buscar conjuntos de datos...

127 datasets found Sort by: Relevance

RATING SOCIAL PROGRAMS

Database containing the evaluation of the results of social programs from the federal government subject to the annual assessment.

Singapore Government Integrity • Service • Excellence Contact Us | Sitemap | Feedback Search this site

data.gov.sg discovering data, inspiring ideas

Home Data Sharing Principles Data Catalogue App Showcase For Developers News & Events

First-stop to Discover Government Data

Search Data Catalogue

By Theme By Government Agency

Home > Data Catalogue

Business & Economy Education Energy & Environment Finance

Government

The screenshot shows the INEGI (Instituto Nacional de Estadística y Geografía) website. The main navigation bar includes links for Statistics, Geography, Home, Contact, INEGI Mobile, and social media follow buttons. The page title is "Statistics by topic - Google Chrome". The URL in the address bar is www3.inegi.org.mx/sistemas/sisept/Default.aspx?t=mdemo148&s=est&c=29192. The main content area displays a table of population data for various Mexican states and the Federal District. A dropdown menu is overlaid on the table, offering export options: "Excel 5.0 (.Xls)" (selected), "Word (.Doc)", and "Format:". At the bottom right of the page, there is a large blue banner with the text "OPEN DATA".

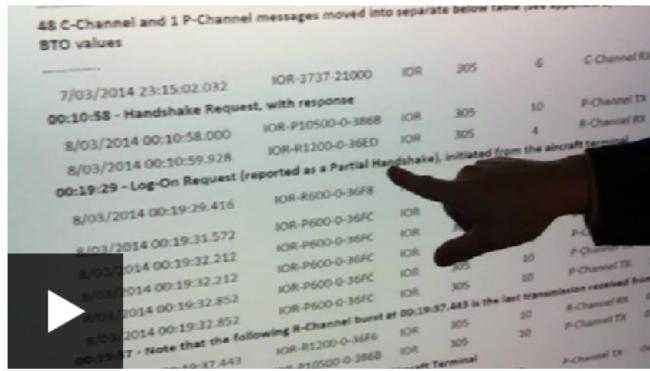
	Federal District	474 860	541 516	720 753	906 063	1229576	1757530	3050442	4870876	6874165	8831079	8235744	8489007	8605239	87
Durango	296 979	370 294	483 175	336 766	404 364	483 829	629 874	760 836	939 208	1182320	1349378	1431748	1448661	15	
Oaxaca	657 182	846 055	1040396	970 003	1084545	1152754	1421513	1727200	2013424	2369076	3019560	3228895	3438765	35	
Puebla	992 426	1021133	1101600	1024955	1150425	1294620	1625830	1973837	2508226	3347685	4126101	4624365	5076686	53	
Querétaro	232 305	232 389	244 663	220 231	234 058	244 737	286 238	355 045	485 523	739 605	1051235	1250476	1404306	15	
Quintana Roo	NA	NA	9109	10,966	10,620	18,752	26,967	50,169	88,150	225 985	493 277	703 536	874 963	11	
San Luis Potosí	571 420	575 432	627 800	445 681	579 831	678 779	856 066	1048297	1281996	1673893	2003187	2200763	2299360	24	
Sinaloa	261 050	296 701	323 642	341 265	395 618	492 821	635 681	838 404	1266528	1849879	2204054	2425675	2536844	26	
Sonora	192 721	221 682	265 383	275 127	316 271	364 176	510 607	783 3							
Tabasco	134 956	159 834	187 574	210 437	224 023	285 630	362 716	496 3							
Tamaulipas	209 106	218 948	249 641	286 904	344 039	458 832	718 167	10241							
Tlaxcala	168 358	172 315	184 171	178 570	205 458	224 063	284 551	346 6							



Government / Private



Flight MH370: Malaysia releases raw satellite data



The BBC's Richard Westcott visited Inmarsat's headquarters to find out what the data tells us about MH370's fate

The Malaysian government has released the raw data used to determine that the missing Malaysia Airlines flight MH370 crashed into the southern Indian Ocean.

The data was first released to relatives of passengers, who have been asking for greater transparency, before copies were also provided to media.

The document released on Tuesday comprises 47 pages of data, plus notes, from British firm Inmarsat.

Time	Channel Name	Ocean Region	GES ID (octal)	Channel Unit ID	Channel Type	SU Type	Burst Frequency Offset (Hz) BFO	Burst Timing Offset (microseconds) BTO
03/2014 16:00:13.406	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	0x15 - Log-on/Log-off Acknowledge		
03/2014 16:00:13.905	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	0x15 - Log-on/Log-off Acknowledge	103	14820
03/2014 16:00:17.430	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eleven Octet User Data		
03/2014 16:00:17.906	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eleven Octet User Data	103	14740
03/2014 16:00:18.404	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eight Octet User Data	103	14780
03/2014 16:00:18.905	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	0x62 - Acknowledge User Data	103	14820
03/2014 16:00:20.906	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	0x71 - User Data (ISU) - RLS		
03/2014 16:00:20.906	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	Subsequent Signalling Unit		
03/2014 16:00:22.906	IOR-R1200-0-3603	IOR	305	10	P-Channel TX	Subsequent Signalling Unit		
03/2014 16:00:23.407	IOR-R1200-0-3603	IOR	305	8	T-Channel RX	Subsequent Signalling Unit		
03/2014 16:00:23.905	IOR-P10500-0-3859	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:27.741	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:27.901	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:28.061	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:28.221	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:28.405	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:28.541	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	
03/2014 16:00:28.541	IOR-T1200-0-3607	IOR	305	8	T-Channel RX	Subsequent Signalling Unit	9820	

OPEN DATA ?

What we know



Suppliers

The screenshot shows the BP website's 'Statistical Review downloads' page. At the top, there's a navigation bar with links for About BP, Products and Services, Sustainability, Investors, Press, Careers, and Gulf of Mexico restoration. Below the navigation is a breadcrumb trail: Home > About BP > Energy economics > Statistical Review of World Energy 2013 > Downloads. The main content area is titled 'Statistical Review downloads'. It explains that users can download the Excel workbook of historical data, the printed edition, and the launch presentation speech and slides. A large red 'X' is overlaid on the right side of the page, covering the 'OPEN DATA' button and some download links. The 'Statistical Review 2013 workbook' link is specifically highlighted with a red box.

BP: You may not frame this site nor link to a page other than the home page without our express permission.
To find this Google “**bp statistical review**”

BP Global | BP Worldwide

About BP Products and Services Sustainability Investors Press Careers Gulf of Mexico restoration

Home > About BP > Energy economics > Statistical Review of World Energy 2013 > Downloads

◀ Energy economics

Statistical Review of World Energy 2013

Group chief executive's introduction

2012 in review

Review by energy type

Using the Review

Energy charting tool

Statistical Review 1951-2011

Downloads

Energy Outlook

Statistical Review downloads

Use this section to download the Excel workbook of historical data, the printed edition and the launch presentation speech and slides

Statistical Review 2013

Download the printed edition of the Statistical Review of World Energy 2013, the supporting PowerPoint Excel workbook of historical statistical data from 1965-2012

Statistical Review of World Energy 2013 (pdf, 9.6MB)

Statistical Review 2013 workbook (xlsx, 1.5MB) **(highlighted)**

Statistical Review slidepack (pptx, 17MB)

Statistical Review 2013 Speech (pdf, 663.0KB)

Statistical Review 2013 speech slide pack (pptx, 2.9MB)

OPEN DATA



Suppliers (ish)

The screenshot shows the Sky website's 'Bigger Picture' section. The left sidebar lists various data categories. The main content area features a heading 'Environmental impact data' with a sub-section 'Our progress so far'. It includes a photograph of a film crew and a summary of environmental targets and progress. A table titled 'Summary performance against environment targets' is shown, with a large red 'X' overlaid on the right side.

Bigger Picture

- Approach
- Our contribution
- Responsible business
- Inspiring action
- News, views & videos
- How we are doing
- How we report
- Data commentary
- Financial data
- Creative industries data
- UK economy data
- Customers data
- People data
- Business partners data
- Environmental impact data**
- Community data
- Sport data

Environmental impact data

At Sky we understand to build a sustainable business we need to minimise our environmental impact from our operations, create more sustainable products and services and work with our business partners and customers to inspire them to take action to protect the environment too.

Our progress so far

Last year we reported good progress towards our ten environment targets that were set in 2009 so we set new, more ambitious targets to 2020. This year we have continued to reduce our emissions when normalised against turnover. We have achieved this by investing in energy efficiency measures, working with our people to reduce the energy they use day to day and through our on-site renewable energy sources. Detailed commentary about this data is provided below.

	Target	2009/10	2010/11	2011/12	2012/13
Reduction in gross CO ₂ e emissions relative to revenue(%) ¹	-50	-8	-21	-29	-33
Energy obtained from owned or controlled renewable energy at Sky-owned sites(%) ²	20	-	-	2	-
Increase in fleet fuel efficiency(%) ³	-15	-	-	-	-5
Reduction in CO ₂ e					



http://corporate.sky.com/the_bigger_picture/how_we_are_doing/environmental_impact_data

Suppliers (ish)

Product Environment Reports



Date introduced
September 10, 2013

Environmental Status Report

iPhone 5s is designed with the following features to reduce environmental impact:

- Arsenic-free display glass
- Mercury-free LED-backlit display
- Brominated flame retardant-free
- PVC-free

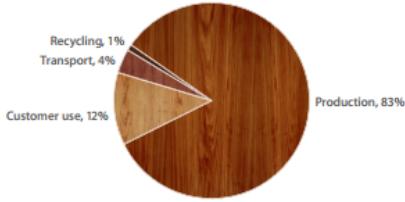
Apple and the Environment

Apple believes that improving the environmental performance of our business starts with our products. The careful environmental management of our products throughout their life cycles includes controlling the quantity and types of materials used in their manufacture, improving their energy efficiency, and designing them for better recyclability. The information below details the environmental performance of iPhone 5s as it relates to climate change, energy efficiency, material efficiency, and restricted substances.¹

Climate Change

Greenhouse gas emissions have an impact on the planet's balance of land, ocean, and air temperatures. Most of Apple's corporate greenhouse gas emissions come from the production, transport, use, and recycling of its products. Apple seeks to minimize greenhouse gas emissions by setting stringent design-related goals for material and energy efficiency. The chart below provides the estimated greenhouse gas emissions for iPhone 5s over its life cycle.²

Greenhouse Gas Emissions for iPhone 5s



Stage	Percentage
Production	83%
Customer use	12%
Transport	4%
Recycling	1%

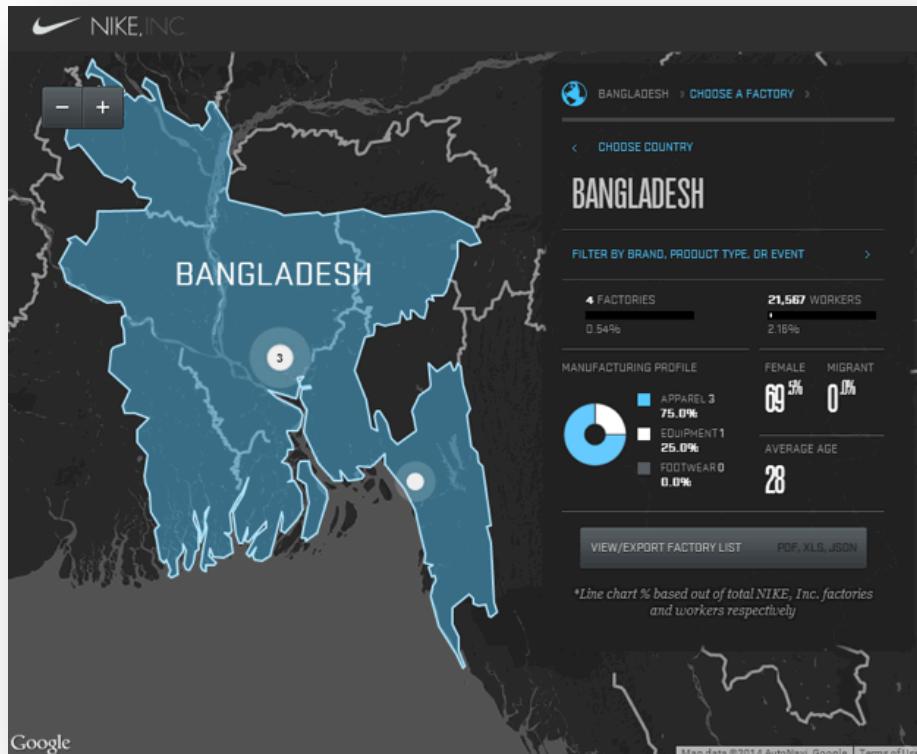
Total greenhouse gas emissions: 75 kg CO₂



<http://www.apple.com/uk/environment/reports/>



Suppliers



OPEN DATA

<http://manufacturingmap.nikeinc.com/#>

You agree not to change or delete any ownership notices from materials downloaded or printed from the Platform. You agree not to modify, copy, translate, broadcast, perform, display, distribute, frame, reproduce, republish, download, display, post, transmit or sell any Intellectual Property or Content appearing on the Platform



Google advanced

Google site:gov filetype:xls

Web Images Maps Shopping More ▾ Search tools

About 4,150,000 results (0.22 seconds)

[XLS] [Code List or Concept \(Acronym\)](#) ↗
www.acquisition.gov/short_codelistsTS.xls Share
File Format: Microsoft Excel - View as HTML
A, B, C, D, F, G, H, 1, Code List or Concept (Acronym), Definition, Authoritative or Maintenance Agency, Current Version, Update Frequency and/or ...

[XLS] [Approps - Foreign Assistance.gov](#) ↗
www.foreignassistance.gov/Full_ForeignAssistanceData.xls
File Format: Microsoft Excel
A, B, C, D, E, F, G, H, 1, Planning Data, 2, 3, 4, Fiscal Year, Fiscal Year Type Account Name, Agency Name, Operating Unit, Category, Sector, Amount ...

[XLS] [TSB Monthly Cash Flow Projection](#) ↗
www.dia.iowa.gov/tsb/cashflow.xls

site: Get results only from certain sites or domains

link: Find pages that link to a certain page

related: Find sites similar to one you already know

filetype: Find certain file types only

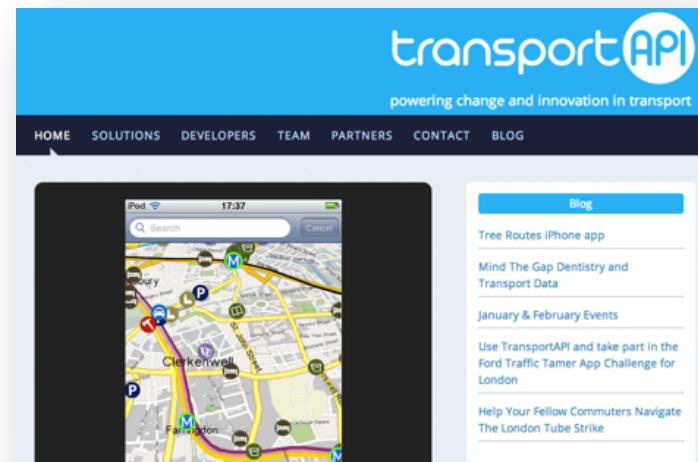


Aggregators and portals

Collect together data from across the web into one place.



enigma.io

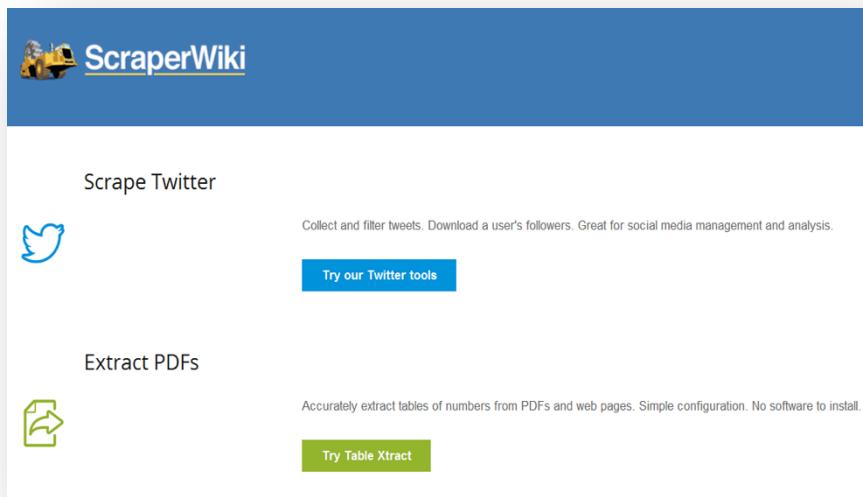


transportAPI



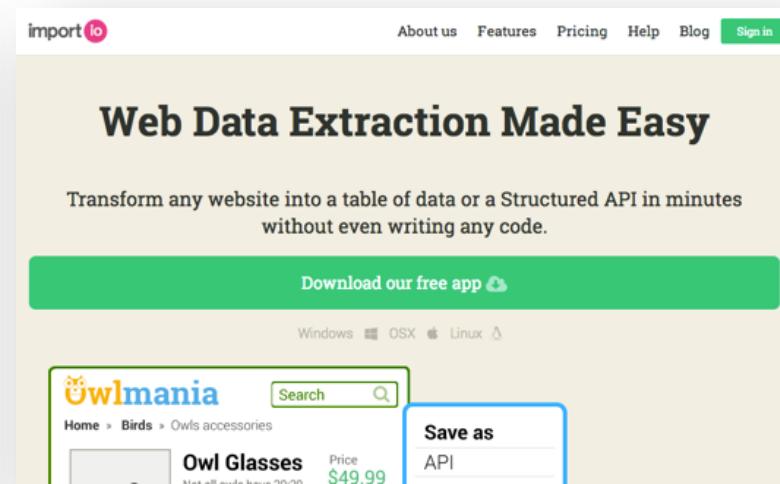
Scraping

If you can't obtain usable data (csv, xls) then you may have to resort to scraping.



The ScraperWiki homepage features a blue header with the logo 'ScraperWiki' and a yellow icon of a bulldozer. Below the header, there are two main sections: 'Scrape Twitter' and 'Extract PDFs'. The 'Scrape Twitter' section includes a Twitter icon, a brief description, and a 'Try our Twitter tools' button. The 'Extract PDFs' section includes a PDF icon, a brief description, and a 'Try Table Xtract' button.

scraperwiki.com



The import.io homepage has a light beige background. At the top, it displays the 'import.io' logo and a navigation bar with links to 'About us', 'Features', 'Pricing', 'Help', 'Blog', and 'Sign in'. The main heading is 'Web Data Extraction Made Easy', followed by a subtext: 'Transform any website into a table of data or a Structured API in minutes without even writing any code.' A large green button says 'Download our free app' with a download icon. Below this, there's a screenshot of a web page from 'Owlmania' showing an 'Owl Glasses' item with a price of '\$49.99'. A 'Save as API' button is overlaid on the screenshot. At the bottom, there are download links for 'Windows', 'OSX', and 'Linux'.

import.io



Finding data on the web (of data)

1. Add random extensions (.xml, .json, .csv etc)
2. Look for alternative links (rss feeds etc)
3. Look for embedded data
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)

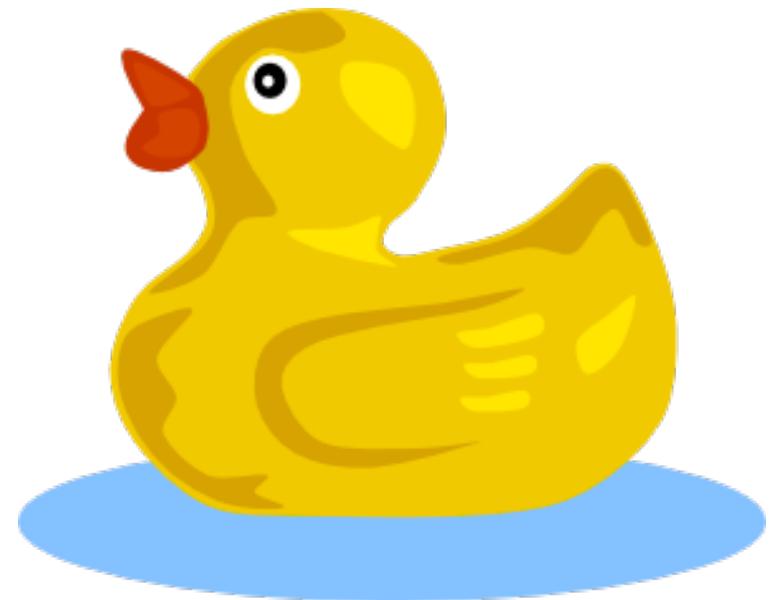


How the web should work,
but people forgot that Tim
put this in when he
invented it!

Duck typed data

If it looks like a duck
and quacks like a duck,
then it's probably a duck.

Basically, keep an eye out for tables,
lists and other stuff that looks like data.



1. Adding random extensions

GOV.UK

Search

Home > Business and self-employed > Imports and exports

Trade Tariff

Search the tariff name or code

This tariff is for 6 August 2014 [change date](#)

View all sections [A-Z Index](#)

Trade between the UK and All countries [change country](#)

Section	Chapters	Title
I	1 to 5	Live animals; animal products
II	6 to 14	Vegetable products
III	15	Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes
IV	16 to 24	Prepared foodstuffs; beverages, spirits and vinegar; tobacco and manufactured tobacco substitutes
V	25 to 27	Mineral products
VI	28 to 38	Products of the chemical or allied industries
VII	39 to 40	Plastics and articles thereof; rubber and articles thereof
VIII	41 to 43	Raw hides and skins, leather, furskins and articles thereof; saddlery and harness; travel goods, handbags and similar containers; articles of animal gut (other than silkworm gut)
IX	44 to 46	Wood and articles of wood; wood charcoal; cork and articles of cork; manufactures of straw, of esparto or of other plaiting materials; basket-ware and wickerwork

UK Trade Tariff



Try using the following: .csv .json .xml .rss .rdf

one

DOCTOR WHO

Home Episodes Clips Galleries Latest News Characters Monsters Fun and Games More

On iPlayer
This programme will be available shortly after broadcast



It's Tomorrow... Get the Latest on the Launch!
What's happening and how to follow the action during tomorrow's big launch in Cardiff.

On TV



The Day of the Doctor
SATURDAY 19:00
BBC THREE
All upcoming
(0 NEW AND 1 REPEAT)

BBC Music and Programmes

2. Look for alternative links



Business Insight - NEWSASIA

NEWS TV WATCH LIVE

Wed, Aug 06 2014

ASIA PACIFIC SINGAPORE WORLD BUSINESS SPORT ENTERTAINMENT TECHNOLOGY HEALTH LIFESTYLE VIDEOS WEATHER MORE ▾

CHANGELIVES LUMINARY AWARDS START-UP

APORE STORIES

Raise of up to 12% for Home Team officers, with sign-on bonuses of up to S\$30,000

al Day Award 2014

SP (2) Ng

(SP)

MEDIACORP

Officers from the Home Team, both past and present were recognised at the Home Team National Day Observance Ceremony on Wednesday.

9 hours ago

Pay rise, special bonus for about 23,000 nurses

10 hours ago

50,000 openings on Jobs Bank for Singaporeans, PRs

1 hour ago

NUS University Town identified as a high-risk dengue cluster

10 hours ago

Scroll down!

LIFESTYLE

VIDEOS

A large black arrow points downwards from the headline "Raise of up to 12% for Home Team officers, with sign-on bonuses of up to S\$30,000" towards the "LIFESTYLE" and "VIDEOS" sections at the bottom of the page.



2. Look for alternative links



 CHANNEL NEWSASIA MediaCorp News Group. © 2014 MediaCorp Pte Ltd. All Rights Reserved. Terms and Conditions Privacy Policy About MediaCorp Pte Ltd	NEWS Asia Pacific Singapore World Business Sport Entertainment Technology Health Lifestyle Videos Photos Special Reports Archives	TV Live TV TV Videos TV Schedule SERVICES Weather ADVERTISE WITH US Online Advertising Mobile Advertising TV Advertising Contact Sales	ABOUT US About Channel NewsAsia Our Logo Our Coverage Our Tagline Presenters and Correspondents Contact Us GET OUR NEWS 
---	---	---	---



RSS



Finding data on the web

(
Techniques 3-5 are not
covered in this session. Please
ask your trainer for more
information if there is time.

1. Look for JSON (json, .csv etc)
2. Look for RSS feeds (feeds etc)
3. Look for embedded data
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)



Exercise

Find a data set using one of the routes we've just looked at.....

Ask yourself – (and discuss in groups)

- . Is it usable?
- . What makes it usable?
- . What more do you need to know?



