



Open Data in Practice

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

David Tarrant · @davetaz

Session 2

Data discovery patterns

Outcomes

Identify a number of different sources of open data on the web.

Create search patterns that enable easy discovery of new sources of open data.

Analyse the usability of available data and formulate plans for usage.

Understand the difference between “data on the web” and the “web of data.”

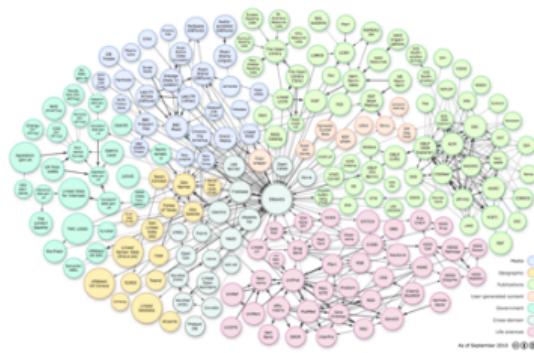


Approaches to publishing data

ON the web



IN the web

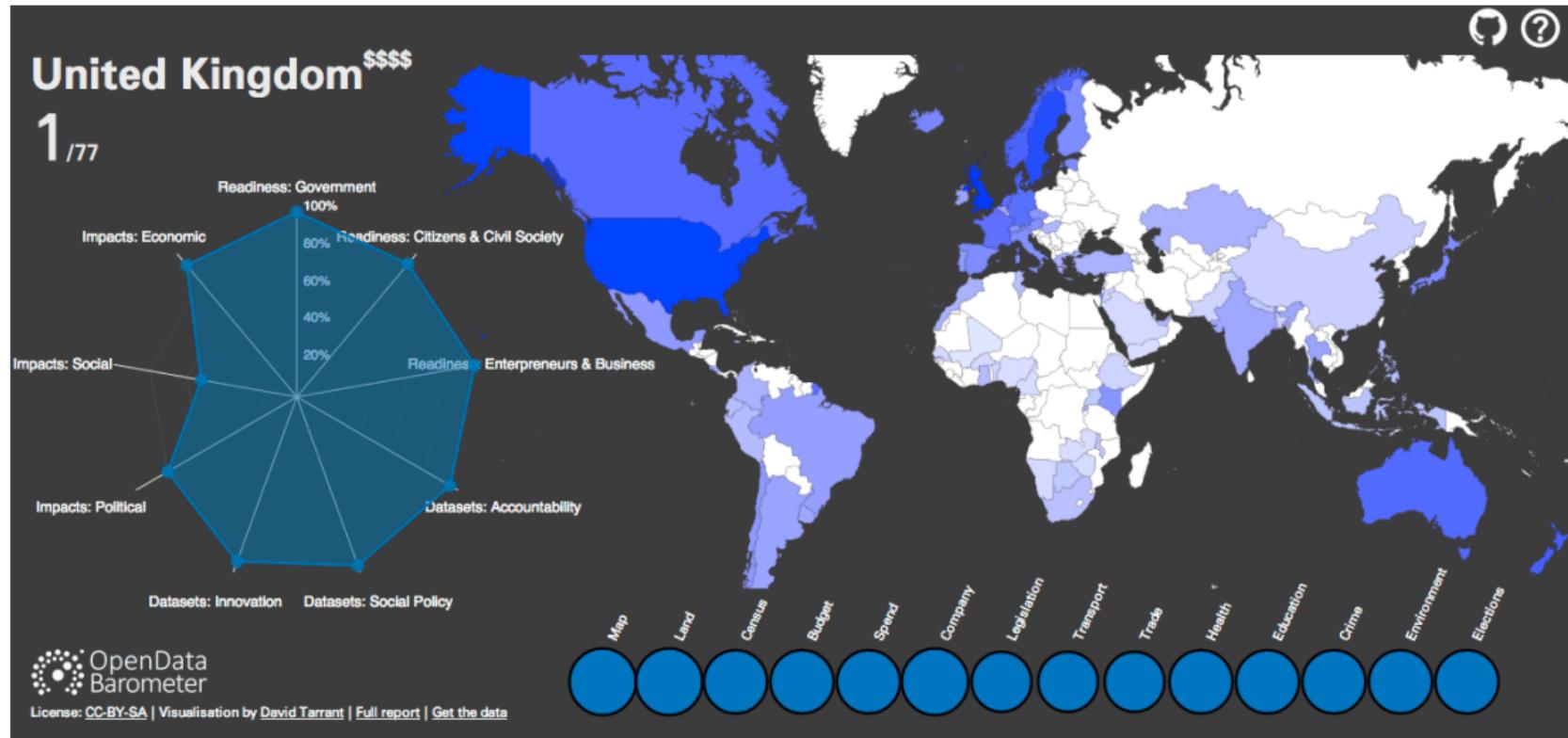


Finding data on the web (of documents)

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



Government data



data.gov.XX

DATA.GOV.UK
Opening up Government

Home Data Apps Interact

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

/ Datasets

Search for data...

19266 Results

Sort by

Show only...
Published datasets (15180)

Live traffic information from the Highways Agency

Highways Agency
Live traffic information data showing traffic information on the strategic road

OpenData Burkina Faso

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

THEMES

- Agriculture
- Éducation
- Diplomatie
- Infrastructures
- Eau et Environnement
- TIC
- Santé
- Collectivités territoriales
- Sécurité Publique
- Tourisme Culture

Tous les thèmes

Open Data Burkina Faso statistiques

- 78 jeux de données
- 28 organisations
- 11 groupes

CC BY SA

DATOS.GOB.MX BETA
OPEN TO PARTICIPATE

Data Stories Advances

/ datasets

Buscar conjuntos de datos...

127 datasets found

Sort by: Relevance

StreetMap

RATING SOCIAL PROGRAMS

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

data.gov.my

Portal Rasmi Open Data Kerajaan Malaysia
The Government of Malaysia's Open Data Official Portal

HOME ABOUT US CATALOGUE INFOGRAPHICS MOBILE APPS CONTACT

DATASETS INFO

- 117 Datasets
- 11 Ministries
- 10 Sectors

RESOURCES

- Malaysia Directory
- Malaysian Open Source Centre
- Open Government Data

MOBILE APPS

- myHealth | myJalani | KPONKK SPAD | SELAWAT | myMAHTAS
- 5 Android Mobile Apps
- 4 iOS Mobile Apps

QUICK LINKS

- Malaysia Informative Data Centre
- Malaysian Population Quick Info
- Tourism Malaysia - Facts & Figures

Latest Datasets Top Publishers Feedback

Applikasi Pemetaan Belia Malaysia
1 Info belia berdasarkan domain
by : Ministry of Youth and Sports

Last Updated 2014-06-26 11:30:19

Applikasi Pemetaan Belia Malaysia
2 Statistik Sikap Belia Malaysia mengikut
by : Ministry of Youth and Sports

Last Updated 2014-06-26 11:27:30

Applikasi Pemetaan Belia Malaysia
3 Statistik Pemetaan Belia Malaysia dalam bentuk aplikasi
by : Ministry of Youth and Sports

Last Updated 2014-06-26 11:24:52

Latest News and Events

Currently, there are no latest event.

Government / Private



Flight MH370: Malaysia releases raw satellite data

45 C-Channel and 1 P-Channel messages moved into separate below raw BTO values							
7/03/2014 23:15:02.032	IOR-3737-21000	IOR	305	6	C-Channel RX		
00:10:58 - Handshake Request, with response							
8/03/2014 00:10:58.000	IOR-P10500-0-3868	IOR	305	10	P-Channel TX		
8/03/2014 00:10:59.328	IOR-R1200-0-166D	IOR	305	4	P-Channel RX		
00:19:29 - Log-On Request [reported as a Partial Handshake], initiated from the aircraft terminal							
8/03/2014 00:19:29.416	IOR-R6000-0-36F8	IOR	305	10	P-Channel TX		
8/03/2014 00:19:31.572	IOR-R600-0-36FC	IOR	305	10	P-Channel RX		
8/03/2014 00:19:32.212	IOR-R600-0-36FC	IOR	305	10	P-Channel TX		
8/03/2014 00:19:32.212	IOR-R600-0-36FC	IOR	305	10	P-Channel RX		
8/03/2014 00:19:32.852	IOR-R600-0-36FC	IOR	305	10	P-Channel TX		
8/03/2014 00:19:32.852	IOR-R600-0-36FC	IOR	305	10	P-Channel RX		
8/03/2014 00:19:32.852 - Note that the following R-Channel burst at 00:19:32.853 is the last transmission received from the aircraft terminal	IOR-R1200-0-36E6	IOR	305	10	P-Channel TX		
8/03/2014 00:19:32.853	IOR-P10500-0-3868	IOR	305	10	P-Channel RX		

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.

A screenshot of a Microsoft Excel spreadsheet titled "Flight Log". The columns are labeled "Time", "Event Type", "Source IP", "Destination IP", "Protocol", "Source Port", "Destination Port", "Size (bytes)", "Rate (Mbps)", and "Rate (Mbps) (Cumulative)". The data shows numerous entries of "Unknown" events between two IP addresses. A large red question mark is overlaid on the bottom right corner of the spreadsheet area.

Flight Log

Time	Event Type	Source IP	Destination IP	Protocol	Source Port	Destination Port	Size (bytes)	Rate (Mbps)	Rate (Mbps) (Cumulative)
2014-03-08 00:19:32.852	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.853	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.854	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.855	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.856	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.857	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.858	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.859	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.860	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.861	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.862	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.863	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.864	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.865	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.866	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.867	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.868	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.869	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.870	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.871	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.872	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.873	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.874	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.875	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.876	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.877	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.878	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.879	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.880	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.881	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.882	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.883	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.884	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.885	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.886	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.887	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.888	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.889	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.890	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.891	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.892	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.893	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.894	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.895	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.896	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.897	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.898	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.899	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.900	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.901	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.902	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.903	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.904	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.905	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.906	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.907	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.908	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.909	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.910	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.911	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.912	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.913	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.914	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.915	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.916	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.917	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.918	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.919	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.920	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.921	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.922	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.923	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.924	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.925	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.926	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.927	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.928	Unknown	192.168.1.100	192.168.1.101	TCP	12345	12345	100	0.000000	0.000000
2014-03-08 00:19:32.									

Suppliers



X 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)

[XL] [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 ...

[XL] [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, ...

[XL] [TSB Monthly Cash Flow Projection](#) ●
www.dca.state.mn.us/tax/taxflow.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

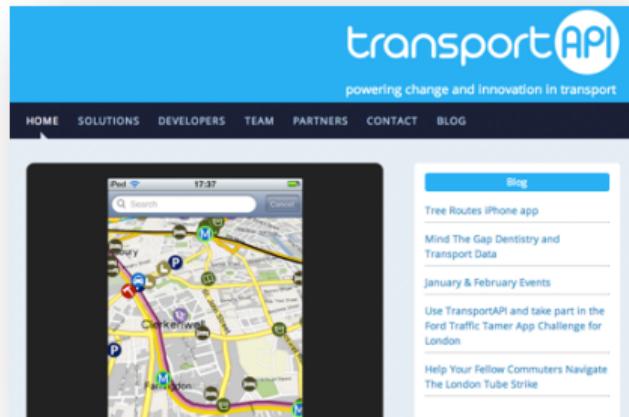


Aggregators and portals

Collect together data from across the web into one place.



enigma.io

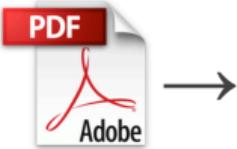


transportAPI



Scraping

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



#	Category Name	Total Sales	Total Margins	Total Profit	% Margin
1	Apparel	48,041,436	48,166,642	160,447	2.61%
2	Automotive	1,000,000	1,000,000	0	0.00%
3	Books	104,891,438	147,202,111	1,358,481	1.35%
4	Butcher/Petite	223,047,498	233,070,000	1,226,121	1.07%
5	Clothing	122,000,000	122,000,000	0	0.00%
6	Confectionery	336,426	336,426	1,000	0.00%
7	Electronics	122,000,000	122,000,000	0	0.00%
8	Furniture	36,467	36,467	1,000	0.00%
9	Grocery	142,000,000	142,000,000	0	0.00%
10	Home Goods	5,763	5,763	200	0.00%
11	Leisure Brand	7,794	7,794	200	0.00%
12	Marketing	410,442	410,442	800	0.00%
13	Meat	51,000	51,000	100	0.00%
14	Merchandise	112,000,000	112,000,000	30,000	0.00%
15	Software	1,000,000	1,000,000	0	0.00%
16	Services	112,000,000	112,000,000	30,000	0.00%

“excellent, so excited beyond description”
George Ofosu, Doctoral Student, UCLA

pdftables.com

A screenshot of the import.io web scraping tool's interface. At the top, there's a search bar with the placeholder "Enter a URL for a list page" and a pink "Extract Data" button. Below the search bar, there are six examples of websites being scraped: Reseller Ratings, Zoopla, 500px, Growth Hackers, Udemy, and Stack Exchange. Each example shows a preview of the website's content and the data being extracted.

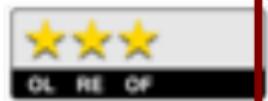
magic.import.io

5-Stars



<http://5stardata.info/>

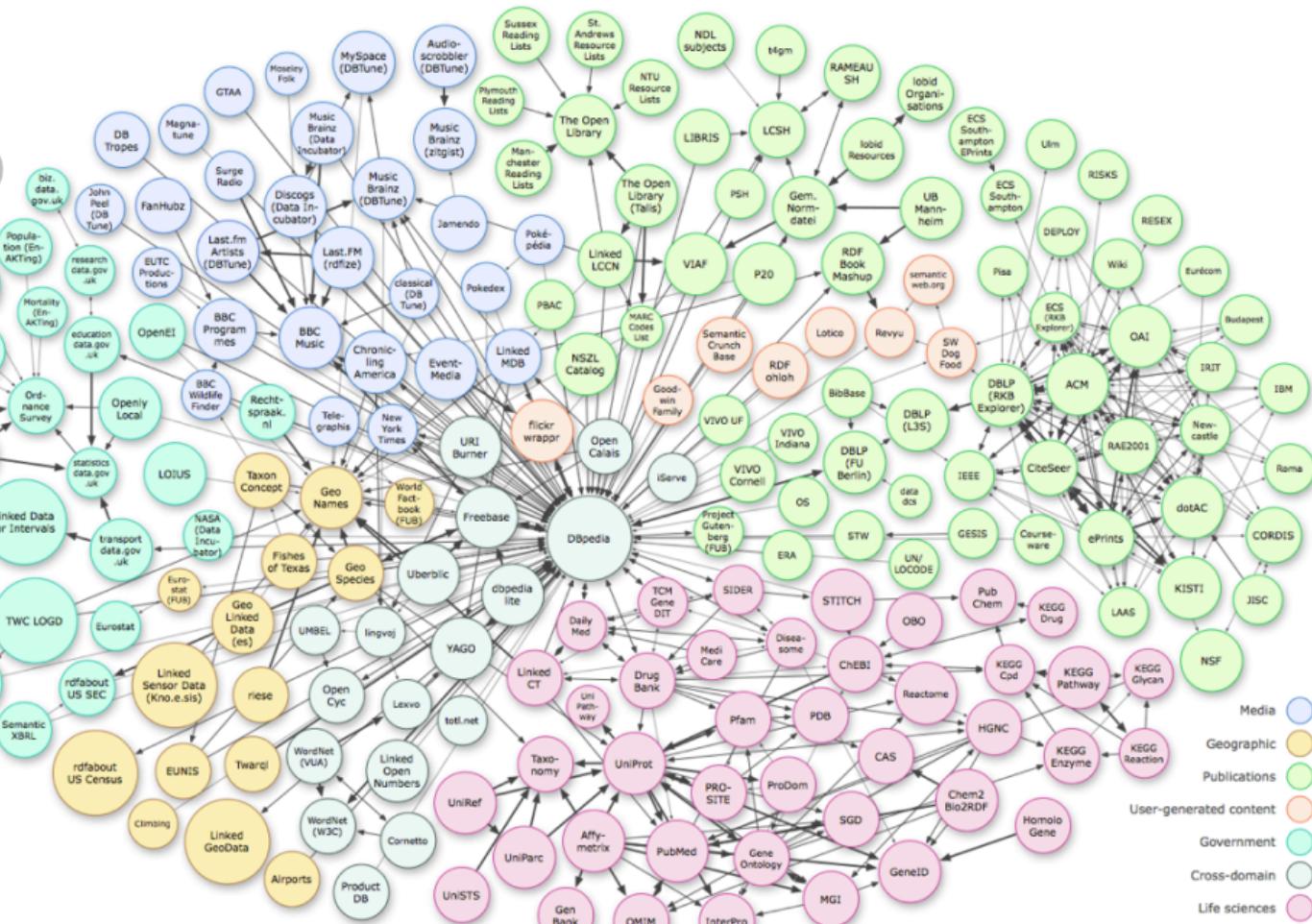
ON THE WEB



IN THE WEB



Data IN the web



Linked data

Amazing but hard to publishing and use.

EEE Building

<http://id.southampton.ac.uk/building/32> ← This is the URI

Detail	Facilities	Services	Energy
--------	------------	----------	--------

Site: Highfield Campus
Construction: 2006
Architect: John McAslan & Partners
Features: Building 32 is non-residential

[View Disability Report for this Building](#)

Occupants

Electronics & Computer Science
Southampton Education School
Agents, Interactions & Complexity
Web & Internet Science
Leadership School Improve & Effectiveness
Lifelong & Work-Related Learning
Mathematics & Science Education
Social Justice & Inclusive Education
Teaching Only Staff
Deanery



©2010 Francois-Xavier Beckers (CC-BY)

ton.ac.uk/building/32
→ rooms:Building, <http://id.southampton.ac.uk/ns/UoSBuilding>
→ "EEE Building"
cupant → Electronics & Computer Science, Southampton Education School, Agents, Interactions Complexity, Web & Internet Science, ... show 8 more...
tion → "32" → <http://id.southampton.ac.uk/ns/building-code-scheme>
tions:within → Highfield Campus
→ <http://www.soton.ac.uk/estates/ourestate/buildings/highfield/32.html>
→ "50.9364157" → [ad:float](#)
→ "-1.395905" → [ad:float](#)
[southampton.ac.uk/ns/disabledGoPage](#) → <http://www.disabledgo.com/en/access-guide/building-32>
tions:easting → "442544" → [ad:integer](#)
tions:northing → "115392" → [ad:integer](#)
Organization → University of Southampton
[southampton.ac.uk/ns/ombilName](#) → "Bldg 32 (EEE)"
feature → Building 32 is non-residential
[southampton.ac.uk/ns/buildingDate](#) → "2006"
[southampton.ac.uk/ns/buildingArchitect](#) → John McAslan & Partners
patial → "POLYGON((-1.3961073411331264 50.93683868764933, -1.3958347895092957 50.9368567227702, -1.3956958407975968 50.93605737417, -1.3959558923017397 50.93603859197583, -1.3961073411331264 50.93683868764933))"
[southampton.ac.uk/ns/electricityTimeSeries](#) → "elec/b32/ekw"
← is spatialrelations:within ← 32 / 3077, 32 / 1015, Physical and Applied Science Faculty Deanery, Social and Human Sciences Faculty Deanery, ... show 54 more...
← is foaf:depicts of ← <http://data.southampton.ac.uk/image-archive/buildings/raw/32.jpg>, <http://data.southampton.ac.uk/image-archive/buildings/1000/32.jpg>, <http://data.southampton.ac.uk/image-archive/buildings/800/32.jpg>, ... show 5 more...
← is event:place of ← AeS Solent Branch Christmas Special Lecture - The Red Arrows



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 **IN THE WEB**
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)

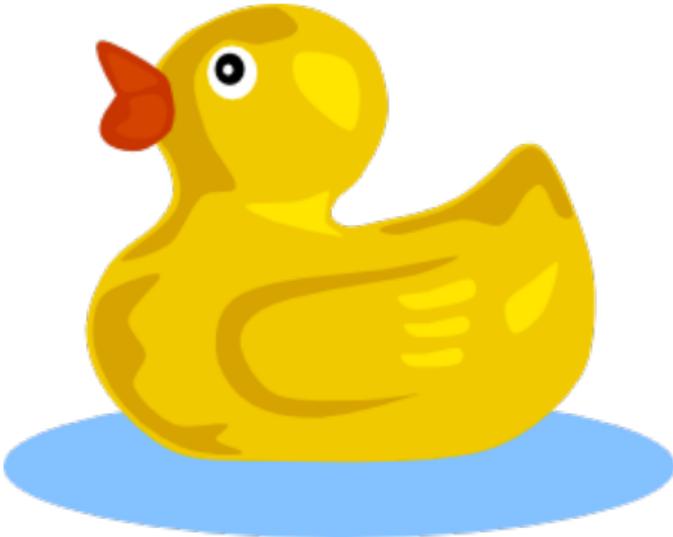


How the web should work,
but people forgot that Tim
put this in when he
invented it!

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

The screenshot shows the GOV.UK Trade Tariff website. At the top, there's a search bar with placeholder text "Search". Below it, a navigation bar includes "Home", "Business and self-employed", and "Imports and exports". A large section titled "Trade Tariff" follows, featuring a search bar for "name or code" and a "Search" button. To the right, a note says "This tariff is for 6 August 2014" with a "change date" link. Below the search area, there are links for "View all sections" and "A-Z Index". The main content area is divided into sections labeled I through IX, each with a chapter range and a title. For example, Section I covers chapters 1 to 5 and includes entries for "Live animals; animal products" and "Vegetable products". Section IX covers chapters 44 to 46.

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

The screenshot shows the BBC Music and Programmes website for the TV show "Doctor Who". The header features the BBC logo and the word "one". The main title "DOCTOR WHO" is prominently displayed with the TARDIS logo. Below the title, a navigation menu includes "Home", "Episodes", "Clips", "Galleries", "Latest News", "Characters", "Monsters", "Fun and Games", and "More". The page is divided into several sections: "On iPlayer" (with a thumbnail of the Doctor and a woman), "On TV" (with a thumbnail of the Doctor and another character), and a central section for the "Launch" event. The "Launch" section includes a thumbnail of the Doctor, a headline "It's Tomorrow... Get the Latest on the Launch!", and a brief description about the broadcast.

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

On TV

The Day of the Doctor

SATURDAY 19:00
BBC THREE

All upcoming
(0 NEW AND 1 REPEAT)

BBC Music and Programmes

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



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 ▾

CHANGINGLIVES LUMINARY AWARDS START-UP

Scroll down!

SINGAPORE STORIES

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

National 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

LIFESTYLE VIDEOS



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



3. Look for embedded data

ODI Experiment

Hidden data extractor

open
data
institute

Hidden data extractor

Enter the URL of any webpage to see what JSON data is hidden within it.

Submit

Try these

[Products from Marks and Spencer UK](#)

[Products from ASOS](#)

 CC BY SA

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 **IN THE WEB**
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!

Finding data on the web (of data)

Techniques 4-5 are not covered in this session. Please ask your trainer for more information if there is time.

1. json, .csv etc)
2. needs etc)
- 3.
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?



Outcomes

Identify a number of different sources of open data on the web.

Create search patterns that enable easy discovery of new sources of open data.

Analyse the usability of available data and formulate plans for usage.

Understand the difference between “data on the web” and the “web of data.”

