



ODI Learning in Tanzania

Dr David Tarrant | @davetaz | The Open Data Institute

<http://tanzania.learndata.info>



 Content created by
The Open Data Institute

Open data discovery patterns



 Content created by
The Open Data Institute

Outcomes

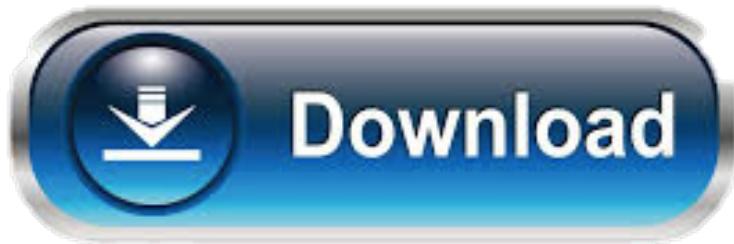
Identify a number of different sources of open data on the web.

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

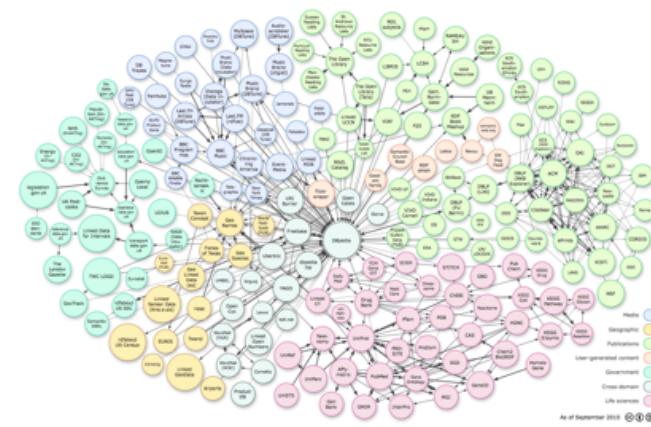


Approaches to publishing data

ON the web



IN the web



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

Government data

Private sector data

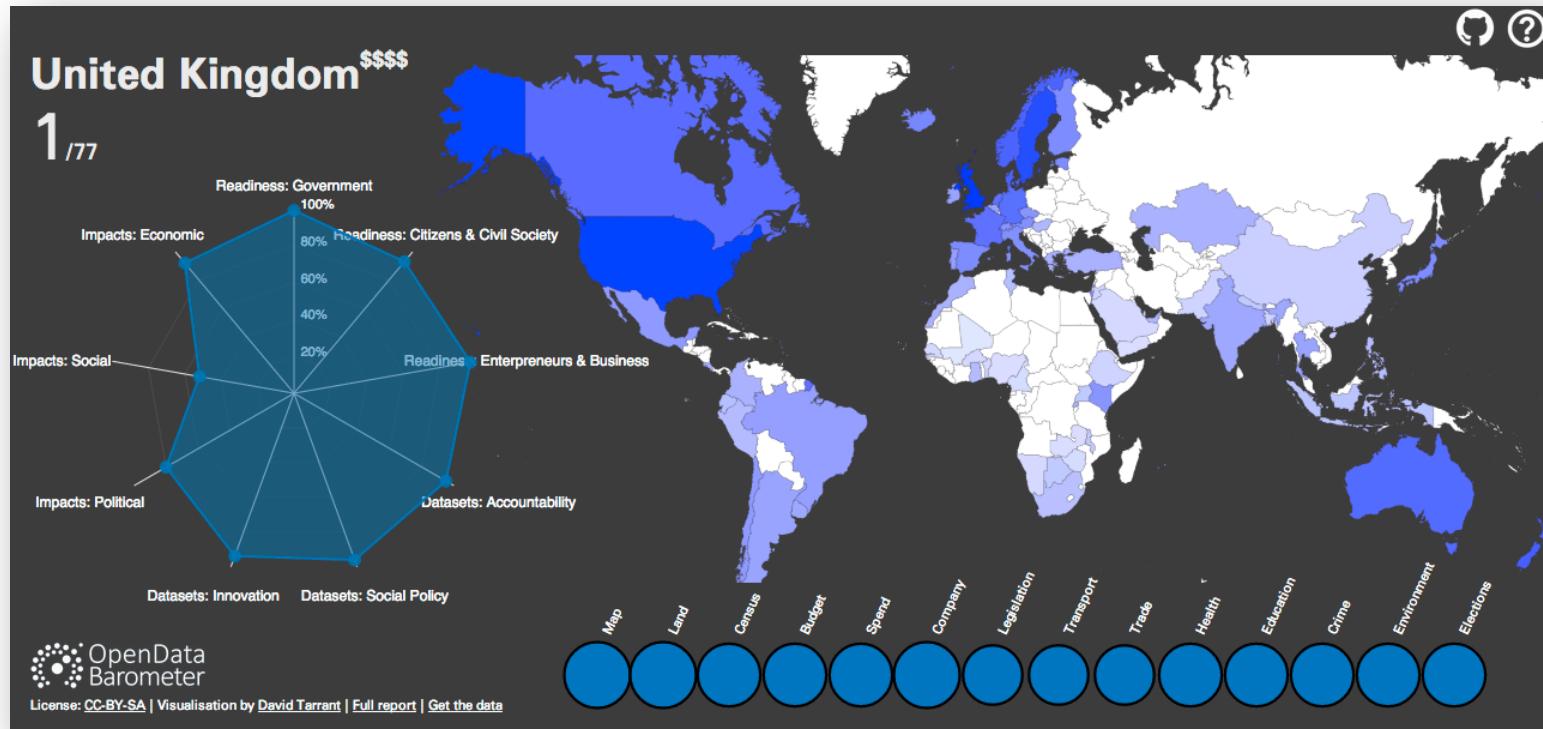
Google advanced

Aggregators and portals

Scraping



Government data



<http://www.opendatabarometer.org/>



data.gov.XX

The screenshot shows the DATA.GOV.UK homepage with a search bar containing "Search for data...". Below it, a section displays "19266 Results" for "Live traffic information from the Highways Agency". A map of the United States and Mexico is visible in the background.



The screenshot shows the DATOS.GOB.MX BETA website. It features a search bar with "Buscar conjuntos de datos..." and a results section titled "127 datasets found". Below this, there's a "RATING SOCIAL PROGRAMS" section with a description of the database. At the bottom, there's a "Latest News and Events" section with three items.

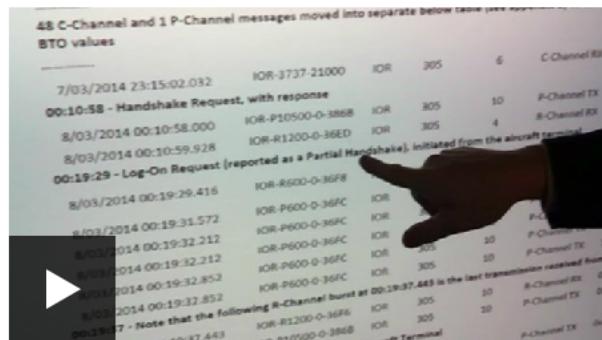


The Open Data Institute

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.

MH370 mystery

Deep sea challenge

Ocean maps
problem

Costs of the search

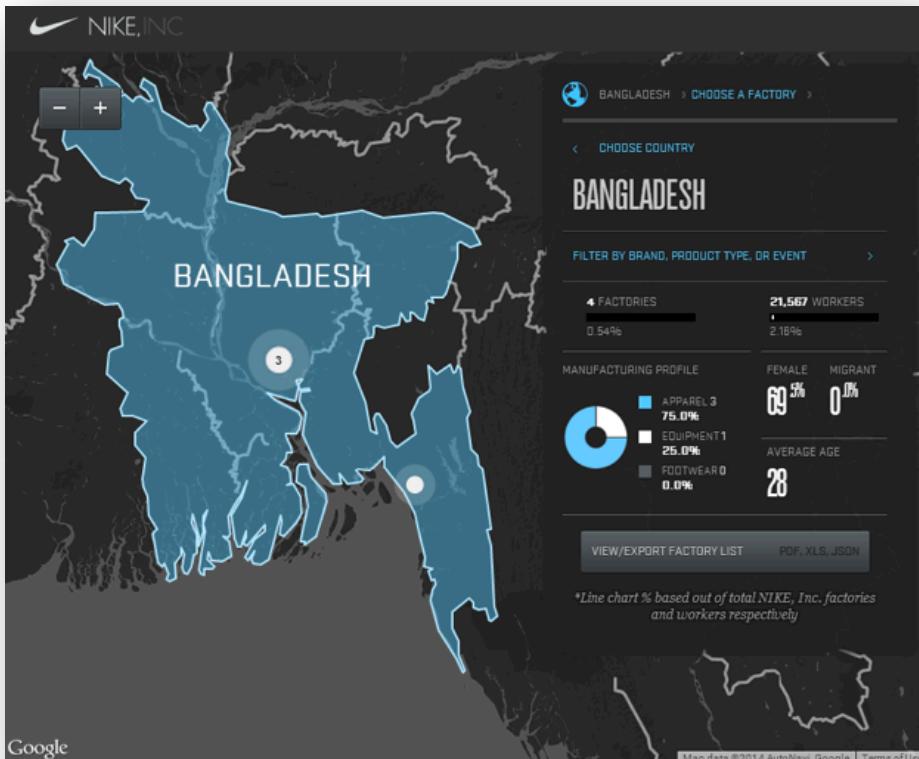
What we know

A screenshot of a document titled "Pre Take-Off". It contains a table of raw satellite data with the same columns as the BBC screenshot. Below the table is a large blue banner with the text "OPEN DATA?" and a large red question mark. The banner also includes the text "Subsequent Signalling Unit" and "88 9820".

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
7/03/2014 16:00:13.406	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	0x15 - Log-on/Log-off Acknowledge		
7/03/2014 16:00:13.506	IOR-P1050-0-3859	IOR	305	10	P-Channel TX	0x15 - Log-on/Log-off Acknowledge	103	14820
7/03/2014 16:00:17.430	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eleven Octet User Data		
7/03/2014 16:00:17.906	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eleven Octet User Data		
7/03/2014 16:00:18.406	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eleven Octet User Data	103	14740
7/03/2014 16:00:18.905	IOR-P1050-0-3859	IOR	305	8	R-Channel RX	Eight Octet User Data	103	14780
7/03/2014 16:00:20.306	IOR-P1050-0-3859	IOR	305	10	P-Channel TX	0x62 - Acknowledge User Data	103	14820
7/03/2014 16:00:20.906	IOR-P1050-0-3859	IOR	305	10	P-Channel TX	0x71 - User Data (ISU) - RLS		
7/03/2014 16:00:20.906	IOR-P1050-0-3859	IOR	305	10	P-Channel TX	Subsequent Signalling Unit		
7/03/2014 16:00:22.906	IOR-P1050-0-3859	IOR	305	10	P-Channel TX	Subsequent Signalling Unit		
7/03/2014 16:00:23.407	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Subsequent Signalling Unit		
7/03/2014 16:00:23.905	IOR-P1050-0-3859	IOR	305	8	R-Channel RX	0x62 - Acknowledge User Data		
7/03/2014 16:00:27.491	IOR-T1200-0-3607	IOR	307	8	R-Channel RX	Subsequent Signalling Unit		
7/03/2014 16:00:27.901	IOR-T1200-0-3607	IOR	307	8	R-Channel RX	Subsequent Signalling Unit		
7/03/2014 16:00:28.061	IOR-T1200-0-3607	IOR	307	8	R-Channel RX	Subsequent Signalling Unit		
7/03/2014 16:00:28.221	IOR-T1200-0-3607	IOR	307	8	R-Channel RX	Subsequent Signalling Unit		
7/03/2014 16:00:28.405	IOR-T1200-0-3607	IOR	307	8	R-Channel RX	Subsequent Signalling Unit		
7/03/2014 16:00:28.541	IOR-T1200-0-3607	IOR	307	8	R-Channel RX	Subsequent Signalling Unit		



Suppliers



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



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



 Content created by
The Open Data Institute

Google advanced

The screenshot shows a Google search results page with the query "site:gov filetype:xls" in the search bar. The results are categorized under "Web". The first result is a link to "Code List or Concept (Acronym)" from www.acquisition.gov/short_codelistsTS.xls. The second result is a link to "Approps - Foreign Assistance.gov" from www.foreignassistance.gov/Full_ForeignAssistanceData.xls. The third result is a link to "TSB Monthly Cash Flow Projection" from www.dia.iowa.gov/tsbsmcashflow.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

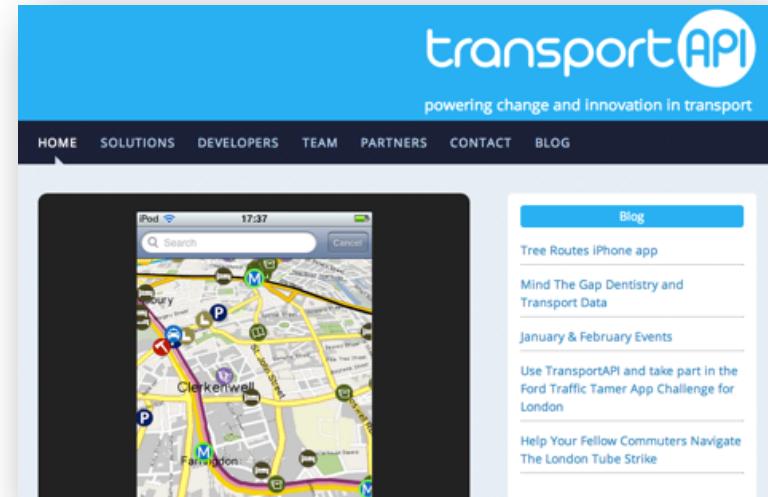


Aggregators and portals

Collect together data from across the web into one place.



enigma.io

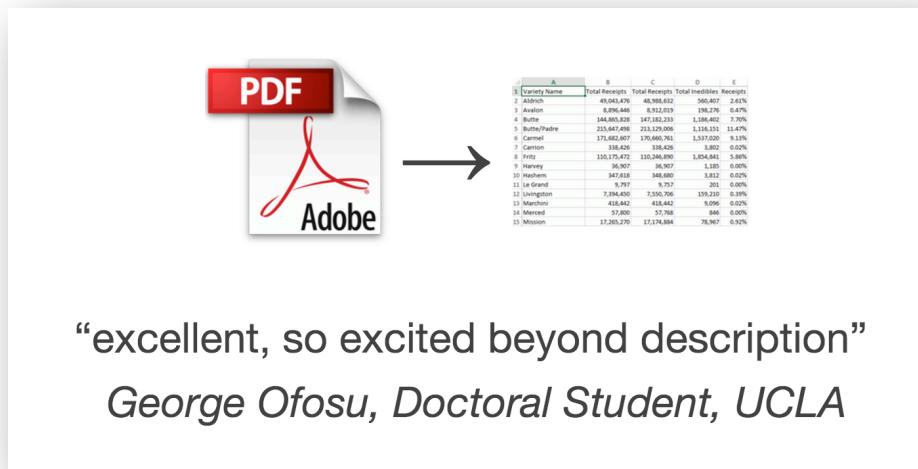


transportAPI

 Content created by
The Open Data Institute

Scraping

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

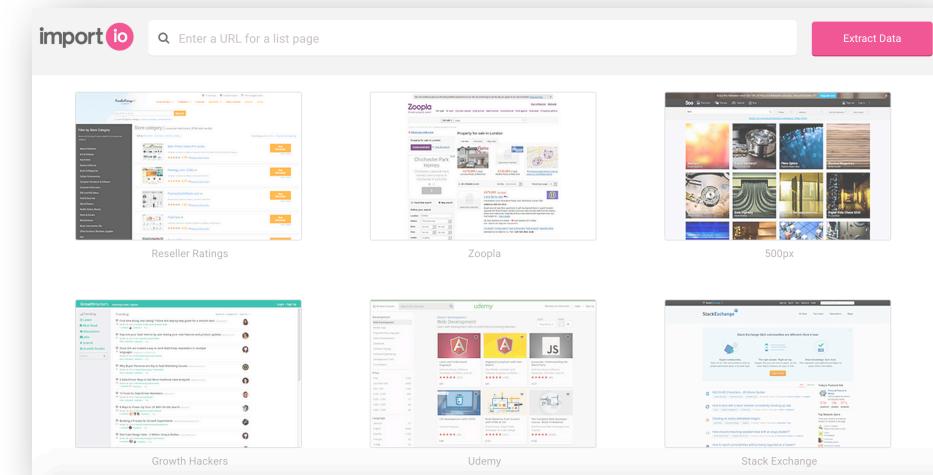


A diagram illustrating the process of data extraction from a PDF. On the left, there is a white document icon with a red 'PDF' button at the top and the Adobe logo at the bottom. An arrow points from this icon to a screenshot of a Microsoft Excel spreadsheet. The spreadsheet contains a table with columns labeled A through E. Column A lists various fruit names, while columns B through E show their respective total receipts and percentages. The data includes entries for Almond, Avocado, Butte, Butter/Padre, Carrizo, Carron, Frits, Harvey, Hashem, Le Grand, Lodi/Merced, Marchin, Merced, and Mission.

Variety Name	Total Receipts	Total Receipts	Total Inedible Receipts	%
Almond	4,986,476	4,986,476	560,496	2.81%
Avocado	8,986,216	8,986,216	1,476,476	16.37%
Butte	144,865,828	147,186,233	1,184,402	7.70%
Butter/Padre	235,647,498	233,122,793	1,116,151	4.77%
Carrizo	171,661,700	170,981,700	1,680,000	1.00%
Carron	338,426	338,426	3,802	0.02%
Frits	120,172,472	120,346,426	1,874,954	3.86%
Harvey	30,800	30,800	1,385	0.00%
Hashem	347,818	348,680	3,812	0.02%
Le Grand	1,797	1,797	201	0.00%
Lodi/Merced	7,394,650	7,395,296	198,246	2.63%
Marchin	418,442	418,442	9,046	0.02%
Merced	57,800	57,760	846	0.00%
Mission	17,265,170	17,174,684	76,967	0.32%

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

pdftables.com



A screenshot of the import.io web interface. At the top, there is 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 smaller screenshots of different websites that can be scraped: "Reseller Ratings", "Zoopla", "500px", "Growth Hackers", "Udemy", and "Stack Exchange". Each screenshot shows a different type of data structure or list that can be extracted.

magic.import.io

 Content created by
The Open Data Institute

5-Stars



<http://5stardata.info/>

ON THE WEB



IN THE WEB



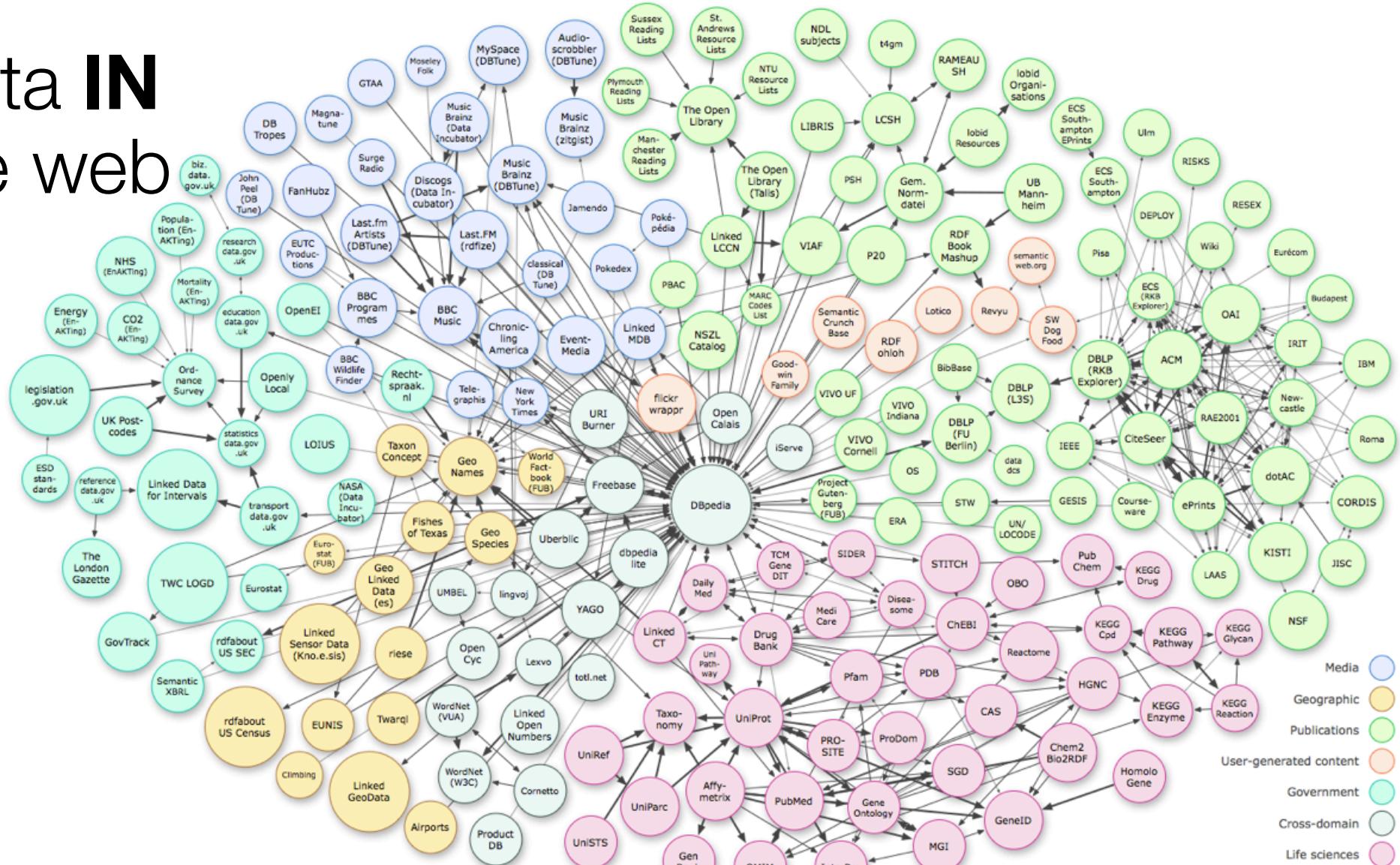
OPEN DATA



Content created by
The Open Data Institute



Data IN the web



As of September 2010 

Linked data

Amazing but challenging to publish and use

EEE Building

[http://id.southampton.ac.uk/building/32 ← This is the URI](http://id.southampton.ac.uk/building/32)

[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)



<http://soton.ac.uk/building/32>

→ rooms:Building, [http://id.southampton.ac.uk/ns/UoSBuilding](#)
→ "EEE Building"

occupant → [Electronics & Computer Science](#), [Southampton Education School](#), [Agents, Interactions & Complexity](#), [Web & Internet Science](#), ...show 8 more...

location → "32"^^[http://id.southampton.ac.uk/ns/building-code-scheme](#)

locations:within → [Highfield Campus](#)

→ <http://www.soton.ac.uk/estates/ourestate/buildings/highfield/32.html>

→ "50.9364157"^^xsd:float
→ "-1.395905"^^xsd:float

[soton.ac.uk/ns/disabledGoPage](#) → <http://www.disabledgo.com/en/access-guide/building-32>

locations:easting → "442544"^^xsd:integer
locations:northing → "115392"^^xsd:integer

Organization → [University of Southampton](#)

[soton.ac.uk/ns/ombielName](#) → "Bldg 32 (EEE)"

feature → [Building 32 is non-residential](#)

[soton.ac.uk/ns/buildingDate](#) → "2006"

[soton.ac.uk/ns/buildingArchitect](#) → [John McAslan & Partners](#)

spatial → "POLYGON((-1.3961073411331264 50.93683868764933, -1.3958347895092957 50.9368567227702, -1.3956958407975968 50.936065737417, -1.3959558923017397 50.93603859197583, -1.3961073411331264 50.93683868764933))"

[soton.ac.uk/ns/electricityTimeSeries](#) → "elec/b32/ekw"

← is spatialrelations:within of ← 32 / 3077, 32 / 1015, [Physical and Applied Science Faculty](#), [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>, <http://data.southampton.ac.uk/image-archive/buildings/600/32.jpg>, ...show 5 more...

← is event:place of ← [RAeS Solent Branch Christmas Special Lecture - The Red Arrows](#)



Google



The Open Data Institute

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)

IN THE WEB

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



Content created by
The Open Data Institute

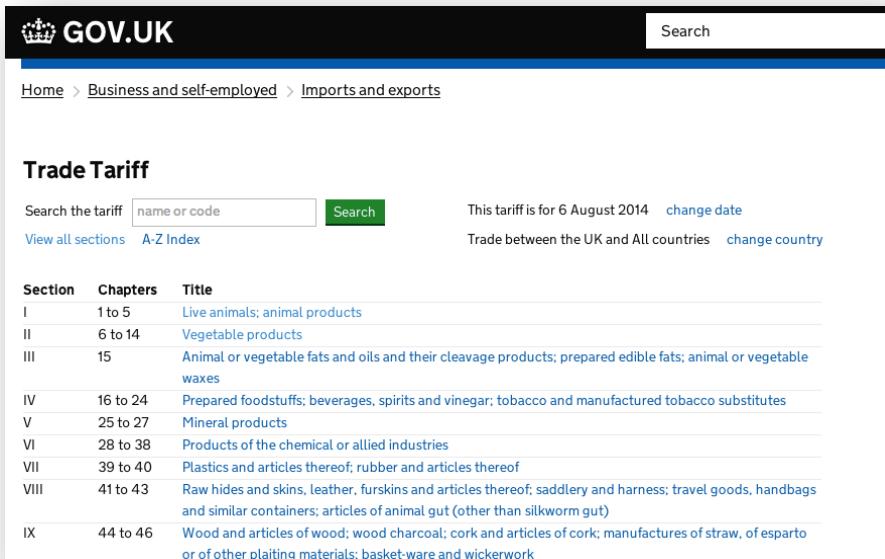
Duck typed data

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 'Trade Tariff' section of the GOV.UK website. At the top, there's a search bar and navigation links for 'Home', 'Business and self-employed', and 'Imports and exports'. Below this, the title 'Trade Tariff' is displayed. A search bar with placeholder 'name or code' and a 'Search' button are present. 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 is a table titled 'Section Chapters Title' with nine rows, each listing a chapter number, its range, and a brief description of the products it covers. For example, Chapter I covers 'Live animals; animal products'.

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



The screenshot shows the BBC Doctor Who website on BBC One. The header features the BBC One logo and the show's name 'DOCTOR WHO' with the TARDIS icon. Below the header, there are several navigation links: Home, Episodes, Clips, Galleries, Latest News, Characters, Monsters, Fun and Games, and More. The main content area has two sections. On the left, there's a thumbnail of the Doctor and a woman, with the text 'It's Tomorrow... Get the Latest on the Launch!' and a description of the event. On the right, there's another thumbnail of the Doctor and a woman, with the text 'On TV' and 'The Day of the Doctor' along with broadcast details: SATURDAY 19:00 BBC THREE. Below this, there's a link to 'All upcoming (0 NEW AND 1 REPEAT)'. The background of the page is a yellow and orange gradient.

BBC Music and Programmes

 Content created by
The Open Data Institute



2. Look for alternative links

The screenshot shows the NewsAsia website homepage. A large blue box highlights the headline "12% for Home Team officers, bonuses of up to \$30,000". A black arrow points from this box down to the main content area. The main content area features a photograph of two men shaking hands at a ceremony. Below the photo is a caption: "Officers from the Home Team, both past and present were recognised at the Home Team National Day Observance Ceremony on Wednesday." A timestamp "9 hours ago" is visible. To the right of the main article, there are three other news items:

- "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)





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



Content created by
The Open Data Institute



3. Look for embedded data

The screenshot shows a web application titled "Hidden data extractor". At the top left is the title "Hidden data extractor" in blue. To its right is a small orange box containing the text "ODI Experiment". On the far right is the "open data institute" logo, which consists of the letters "odi" in white inside a dark square, followed by the words "open data institute" in a smaller, sans-serif font.

The main area of the page has a light gray background. In the center, the title "Hidden data extractor" is displayed in a large, dark font. Below it is a sub-instruction: "Enter the URL of any webpage to see what JSON data is hidden within it." A large, empty input field is provided for this purpose. At the bottom of this section is a blue "Submit" button with white text.

Below the input field, there is a section titled "Try these" in bold black text. It contains two links: "Products from Marks and Spencer UK" and "Products from ASOS", both in blue text.

<http://odinprac.theodi.org/hidden-data-extractor/>



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)

IN THE WEB

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



Content created by
The Open Data Institute

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?

