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.

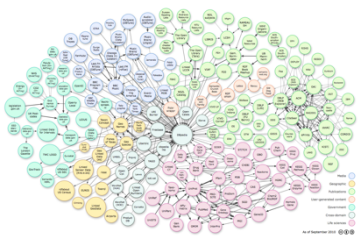
Analyse the usability of available data and formulate plans for usage.

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

Approaches to publishing data

ON the web

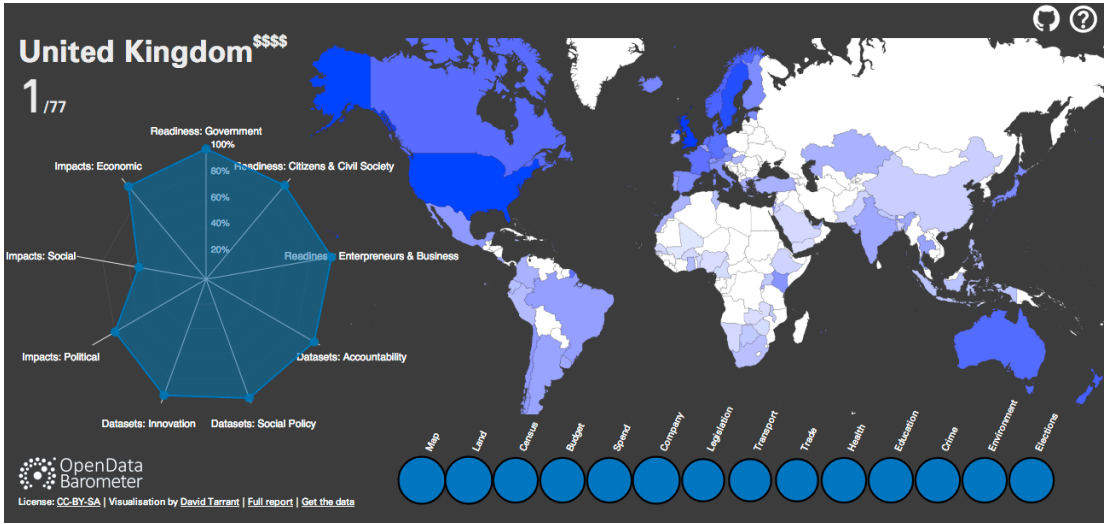
IN the web



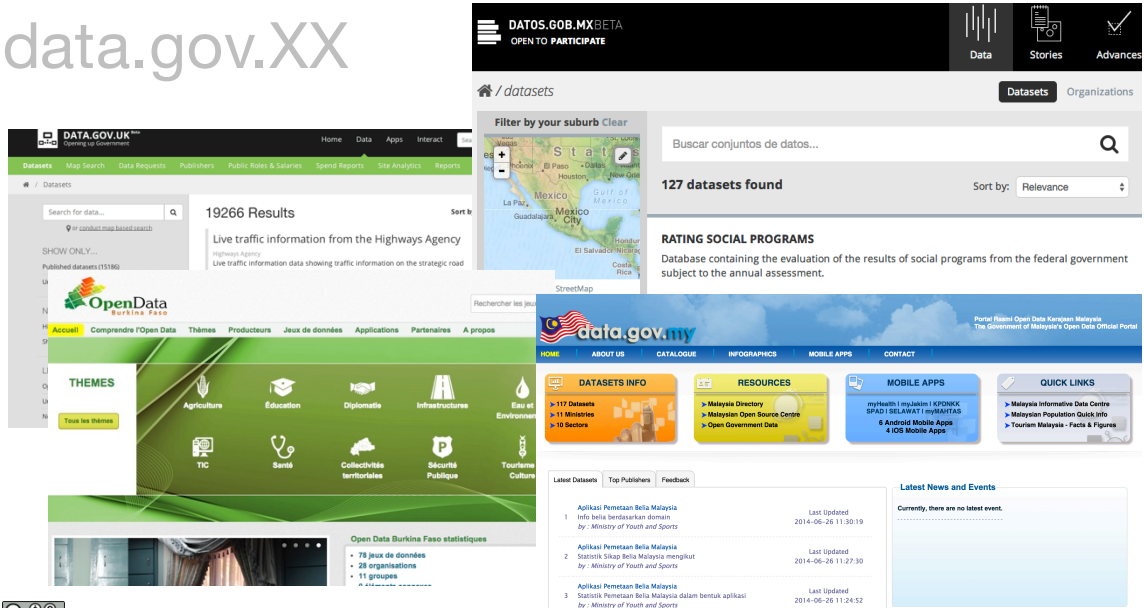
Finding data on the web **(of documents)**

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

Government data

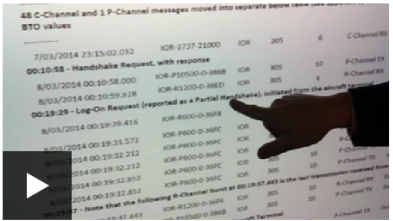


data.gov.XX


DATOS.GOB.MX BETA
OPEN TO PARTICIPATE
Data
Stories
Advances
/ datasets
Datasets
Organizations
Filter by your suburb Clear
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.
DATA.GOV.UK
Opening up Government
Home
Data
Apps
Interact
Datasets
Map Search
Data Requests
Publishers
Public Roles & Salaries
Spend Reports
Site Analytics
Reports
Search for data...
19266 Results
Live traffic information from the Highways Agency
SHOW ONLY...
Published datasets (15186)
OpenData
Burkina Faso
Accueil
Comprendre l'Open Data
Thèmes
Producteurs
Jeux de données
Applications
Partenaires
A propos
THEMES
Tous les thèmes
Agriculture
Éducation
Diplomatie
Infrastructures
TIC
Santé
Collectivités territoriales
Sécurité Publique
Open Data Burkina Faso statistiques
78 jeux de données
28 organisations
11 groupes
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 | myJakim | KPDNKK
SPAD | SELAWAT | myMAHTAS
6 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
Aplikasi Pemetaan Bella Malaysia
Info bella berdasarkan domain
by : Ministry of Youth and Sports
Last Updated
2014-06-26 11:30:19
Aplikasi Pemetaan Bella Malaysia
Statistik Sikap Bella Malaysia mengikut
by : Ministry of Youth and Sports
Last Updated
2014-06-26 11:27:30
Aplikasi Pemetaan Bella Malaysia
Statistik Pemetaan Bella 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

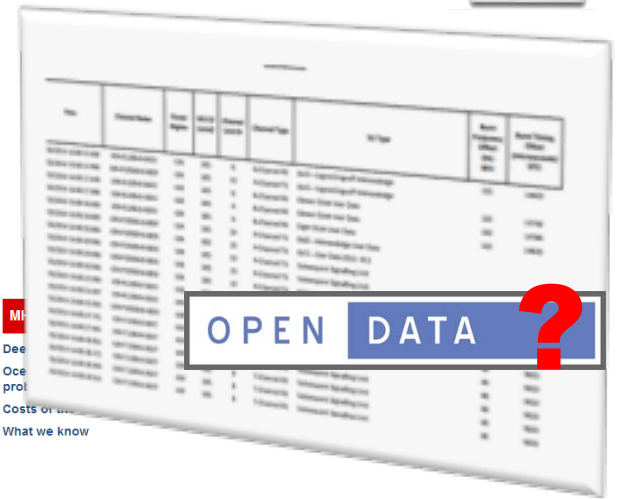


The BBC's Richard Westcott visited Inmarsat's headquarters to find out what the data tells us about MH370's fate

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.



MH

Dee

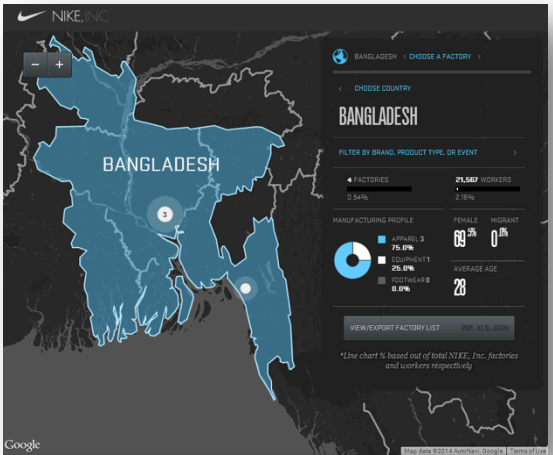
Oce

prol

Costs of th

What we know

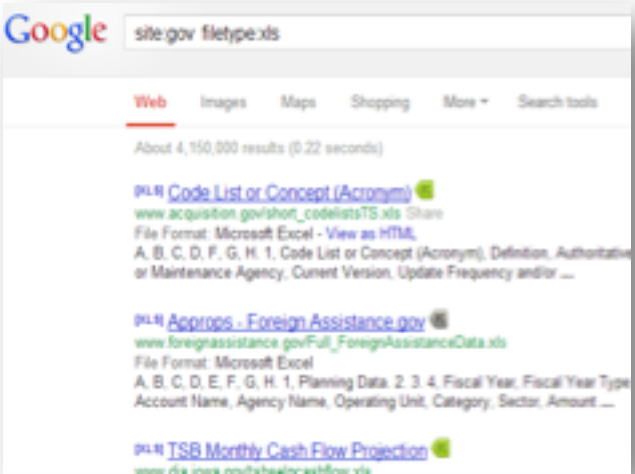
Suppliers



OPEN DATA

http://manufacturingmap.nikeinc.com/#

Google advanced



site: Get results only from certain sites or domains

link: Find pages that link to a certain page

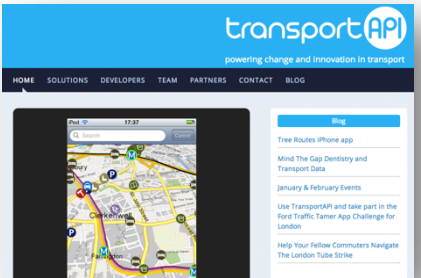
related: Find sites similar to one you already know

filetype: Find certain file types only

Aggregators and portals

Collect together data from across the web into one place.

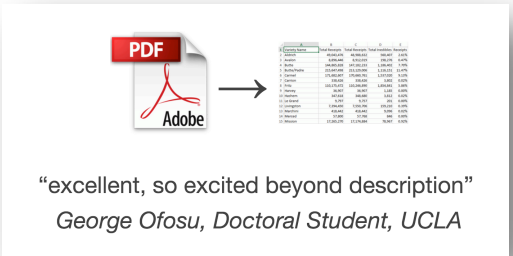
enigma.io



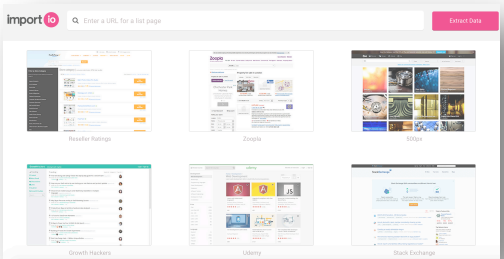
transportAPI

Scraping

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



pdftables.com



magic.import.io

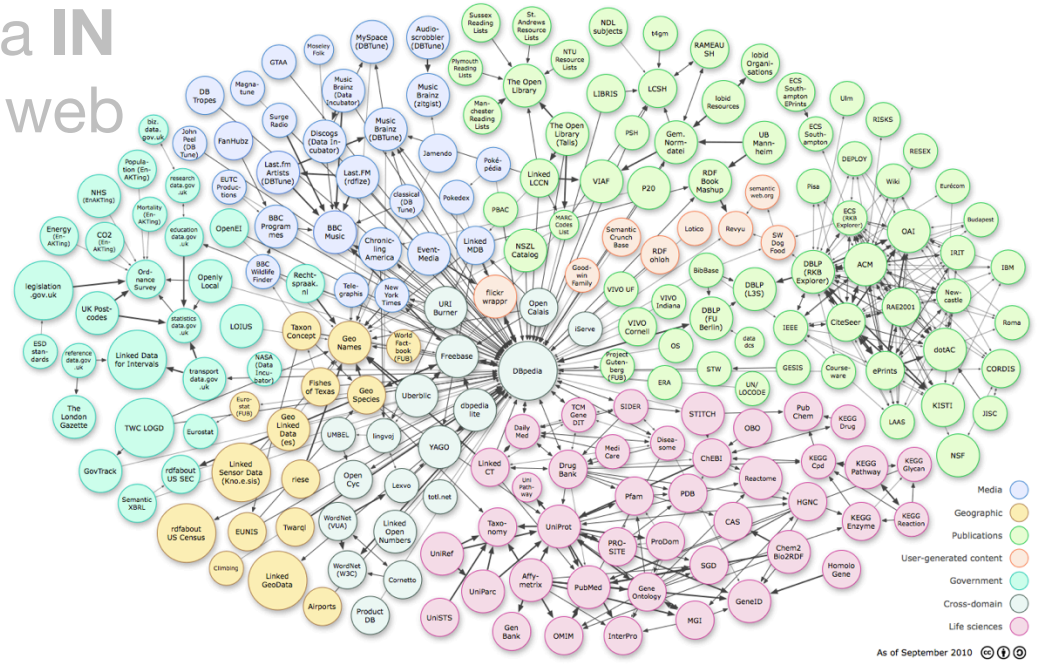
5-Stars


http://5stardata.info/


ON THE WEB
OL
OL RE
OL RE OF
CSV

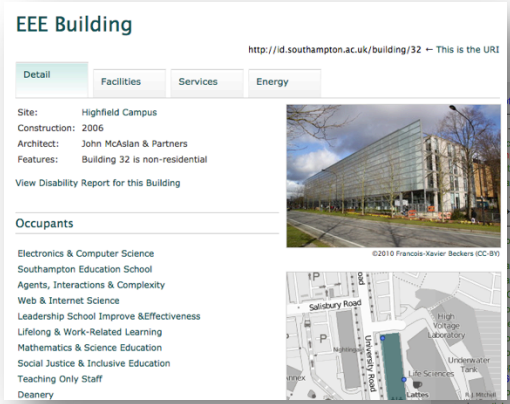

OL RE OF URI LD
OL RE OF URI
http://data...
IN THE WEB
RDF
HTML

Data **IN** the web



Linked data

Amazing but hard to publishing and use.



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)

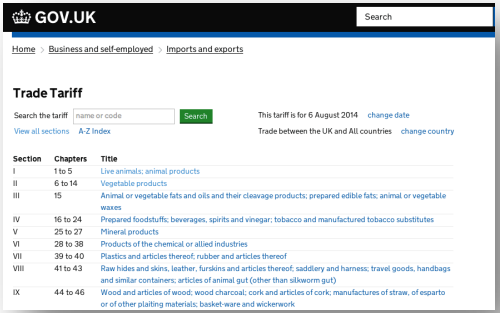
How the web should work, but people forgot that Tim put this in when he invented it!

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

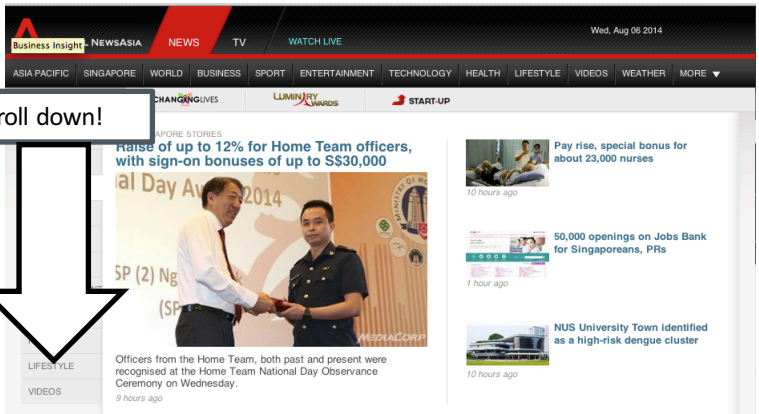


UK Trade Tariff

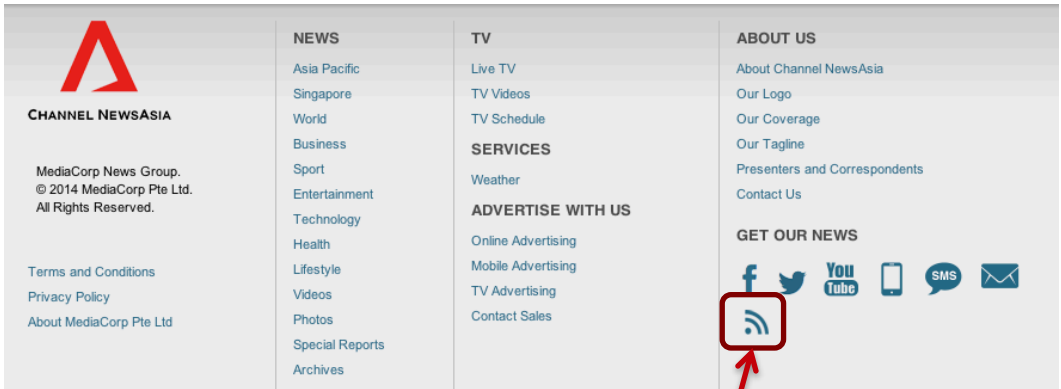
BBC Music and Programmes

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

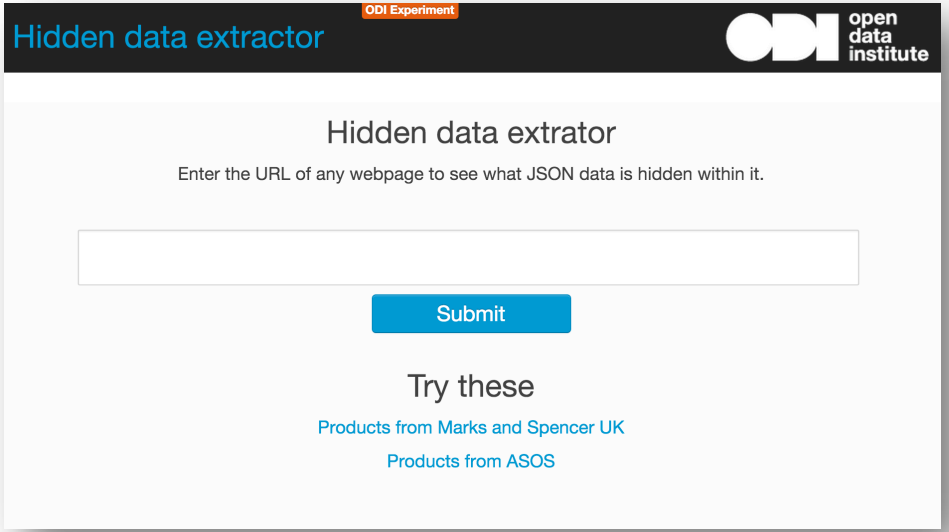
2. Look for alternative links



2. Look for alternative links



3. Look for embedded data



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)

1. [illegible]json, .csv etc)
2. [illegible]eeds etc)
3. [illegible]
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)

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

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.