ODI Leaning 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

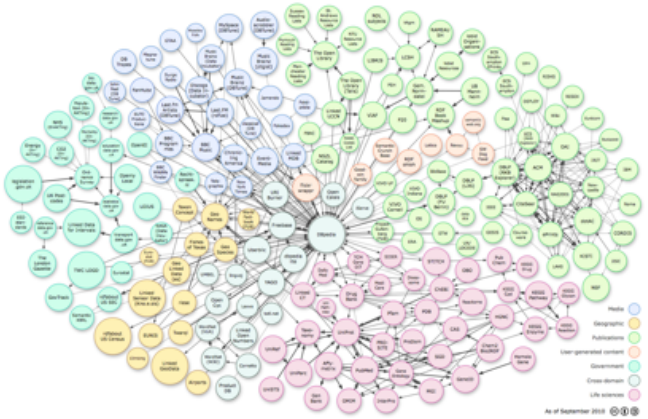
Identify a number of different sources of open data on the web.

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

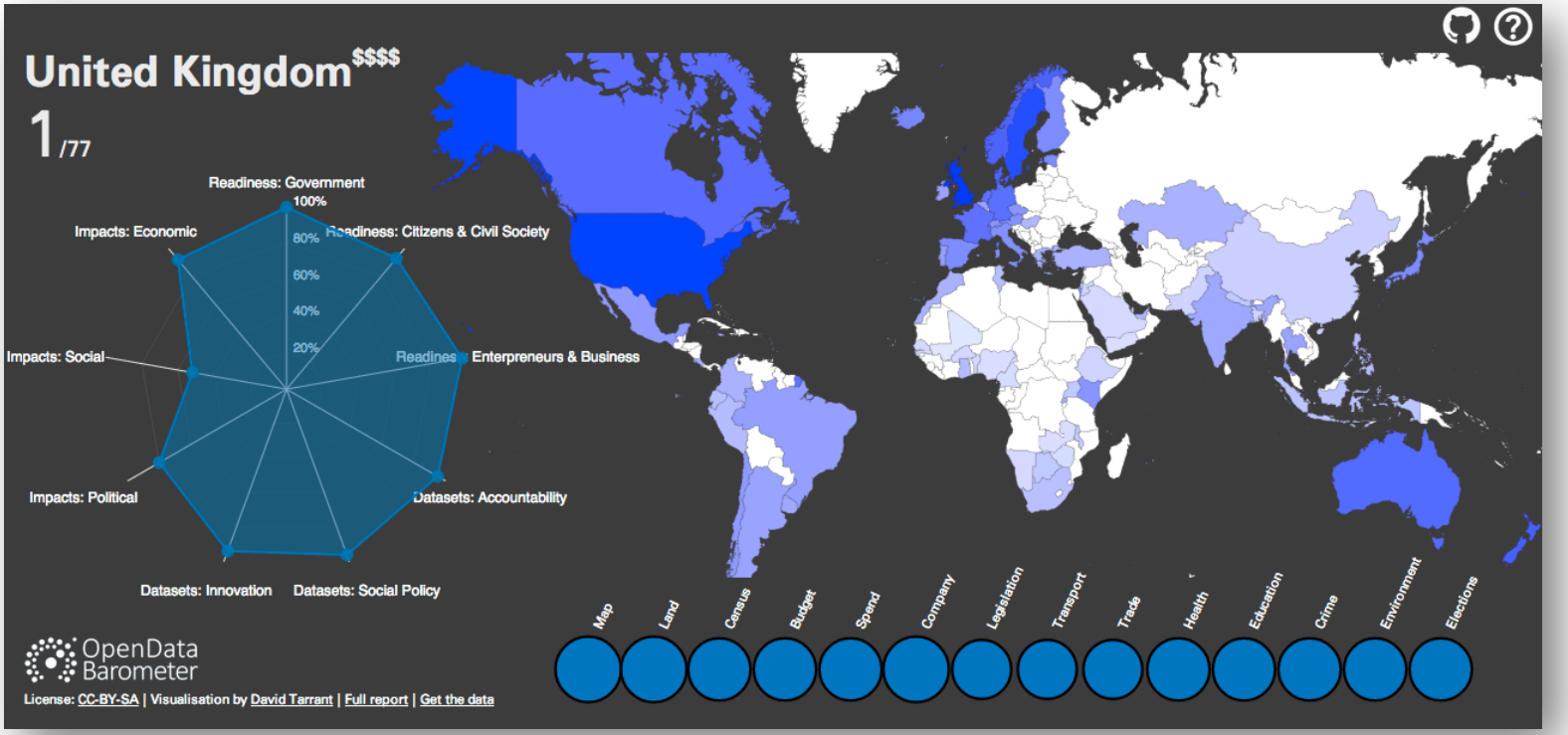
Private sector data

Google advanced

Aggregators and portals

Scraping

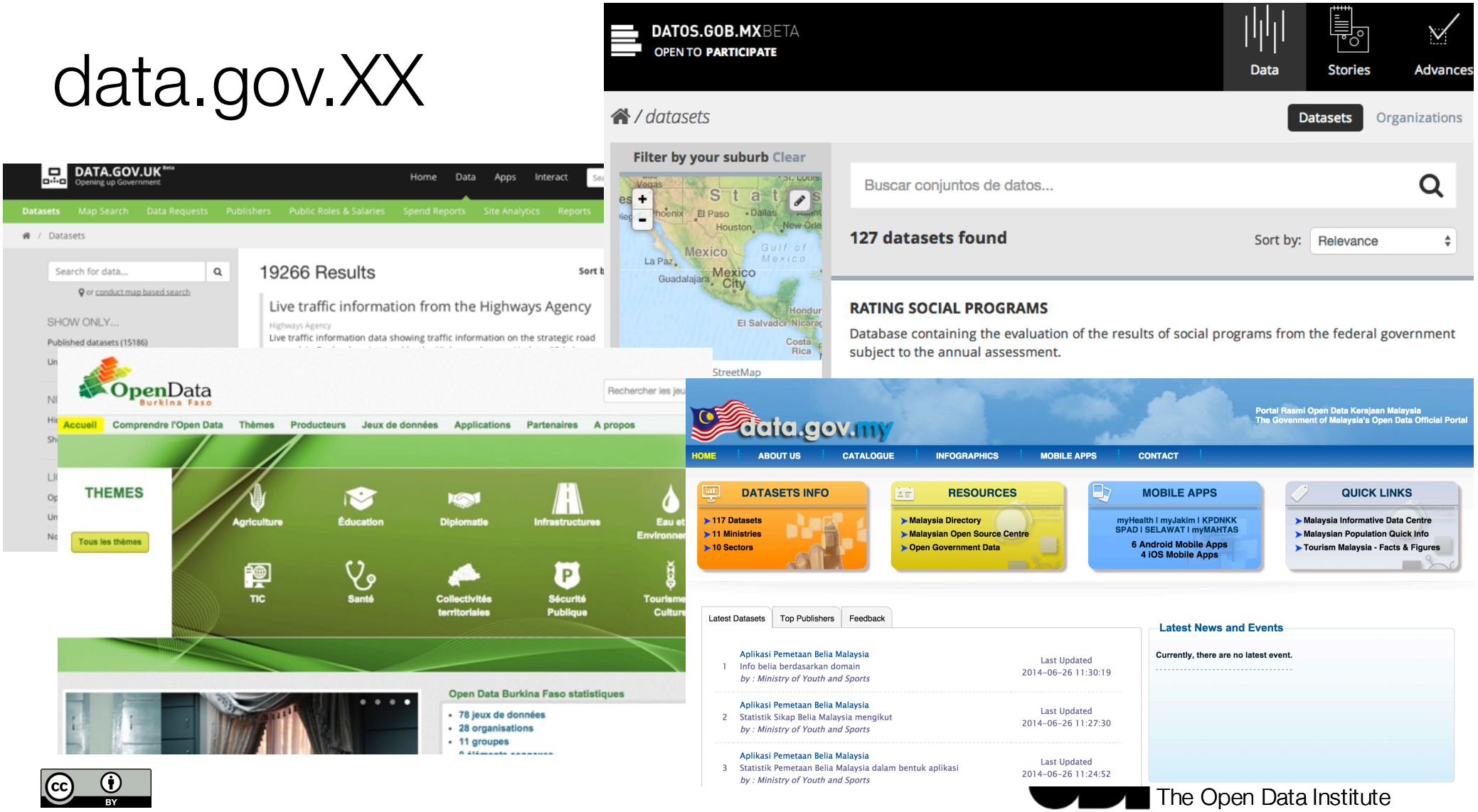
Government data



http://www.opendatabarometer.org/

Content created by
The Open Data Institute

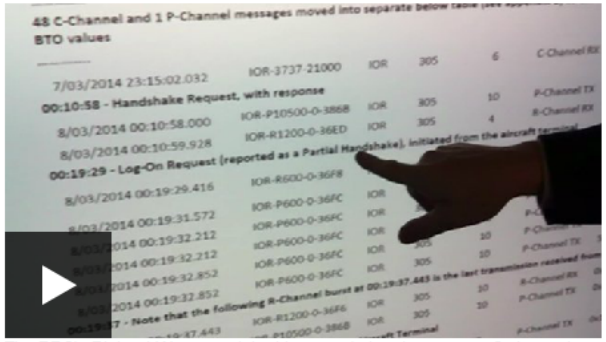
data.gov.XX



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

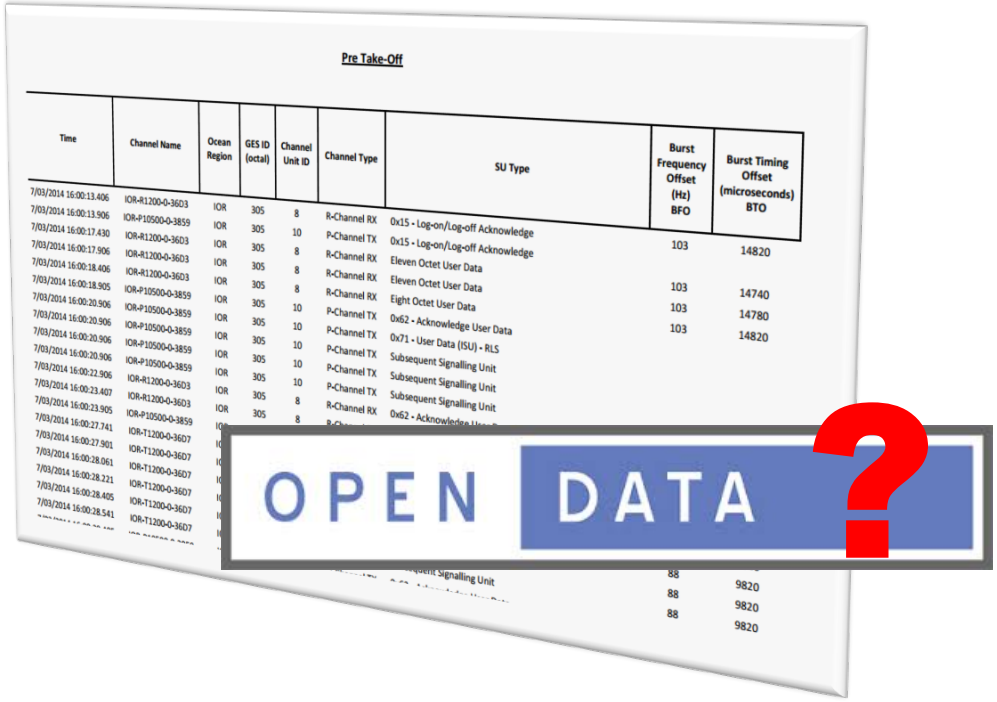
The Malaysian government has released the raw data used to determine that the missing Malaysia Airlines flight MH370 crashed into the southern Indian Ocean.

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.

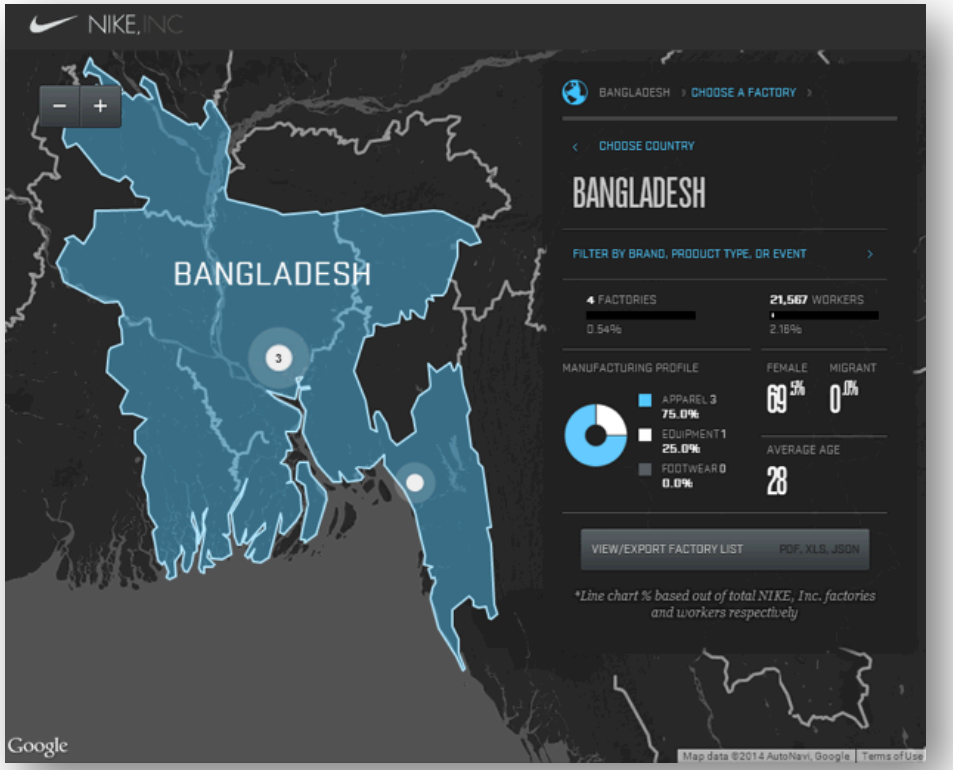
MH370 mystery

- Deep sea challenge
- Ocean maps problem
- Costs of the search
- What we know



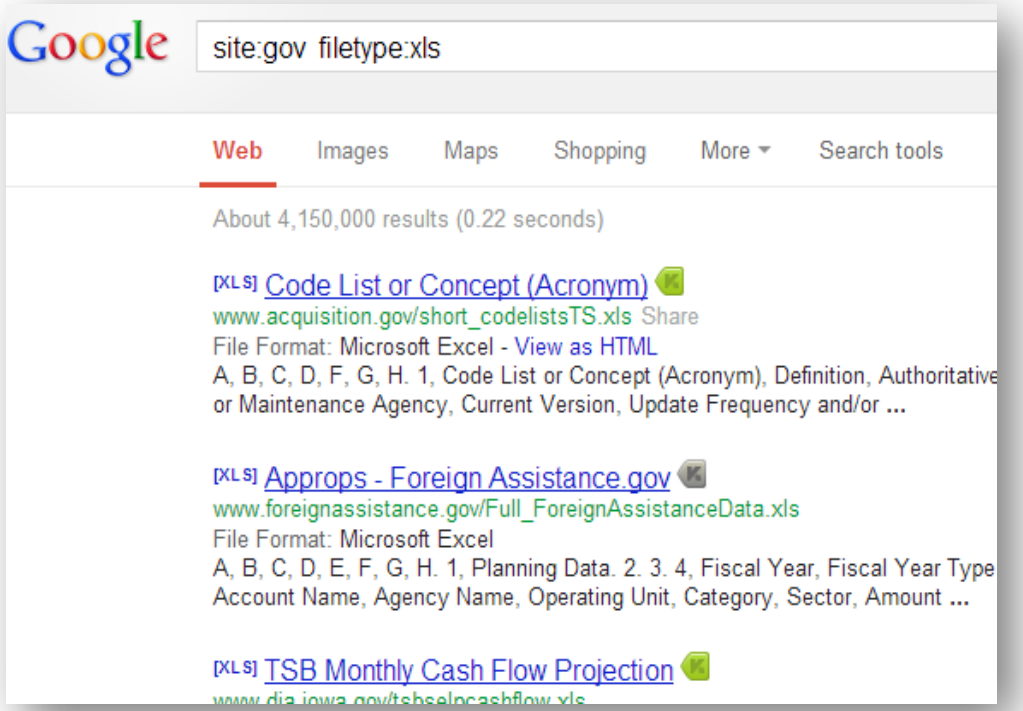
Content created by
The Open Data Institute

Suppliers



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.

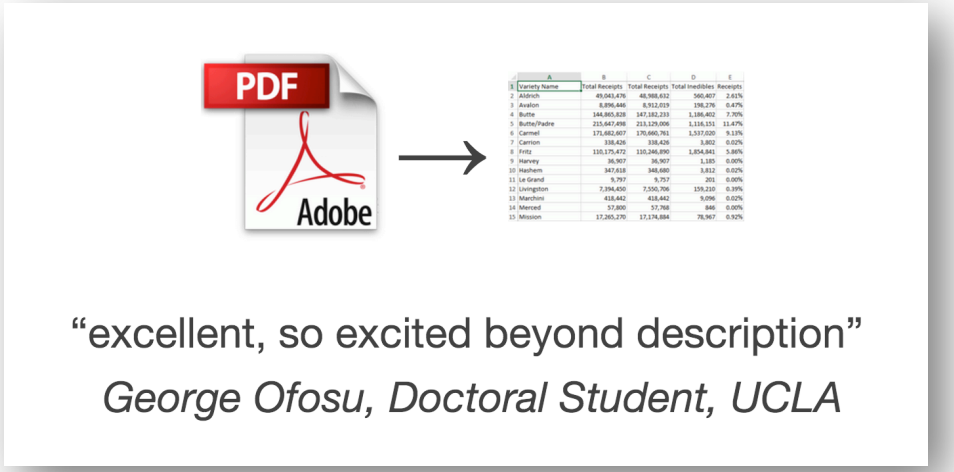
enigma.io

transportAPI

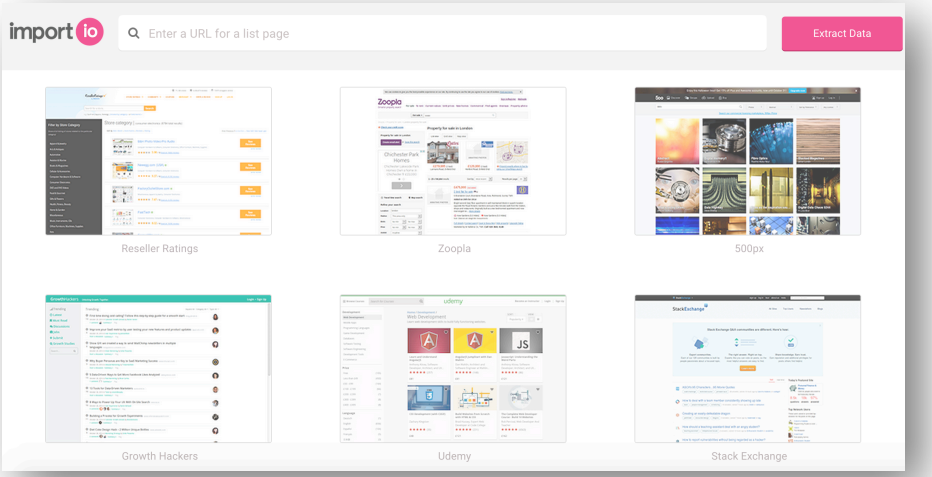
Content created by
The Open Data Institute

Scraping

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



pdftables.com

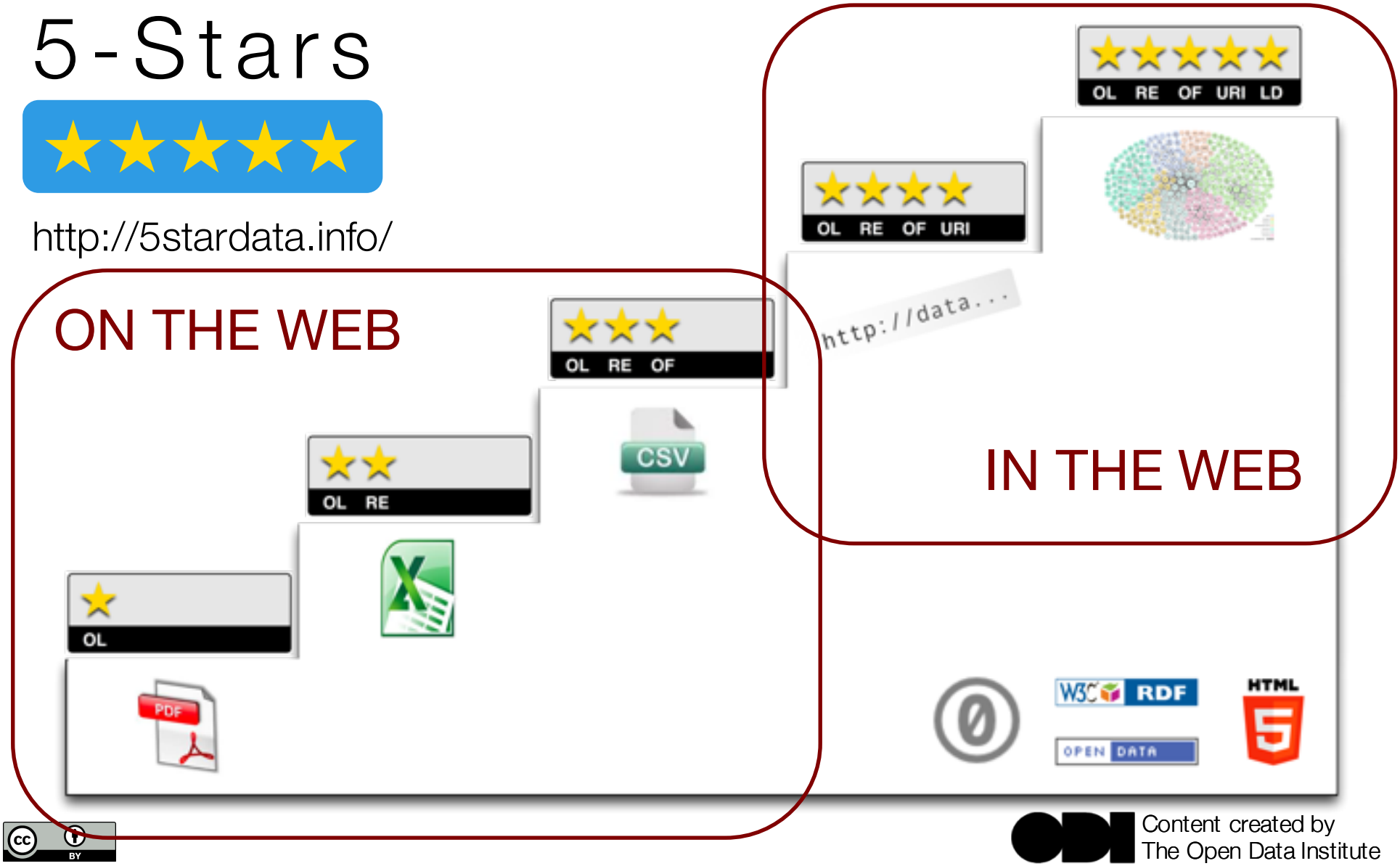


magic.import.io

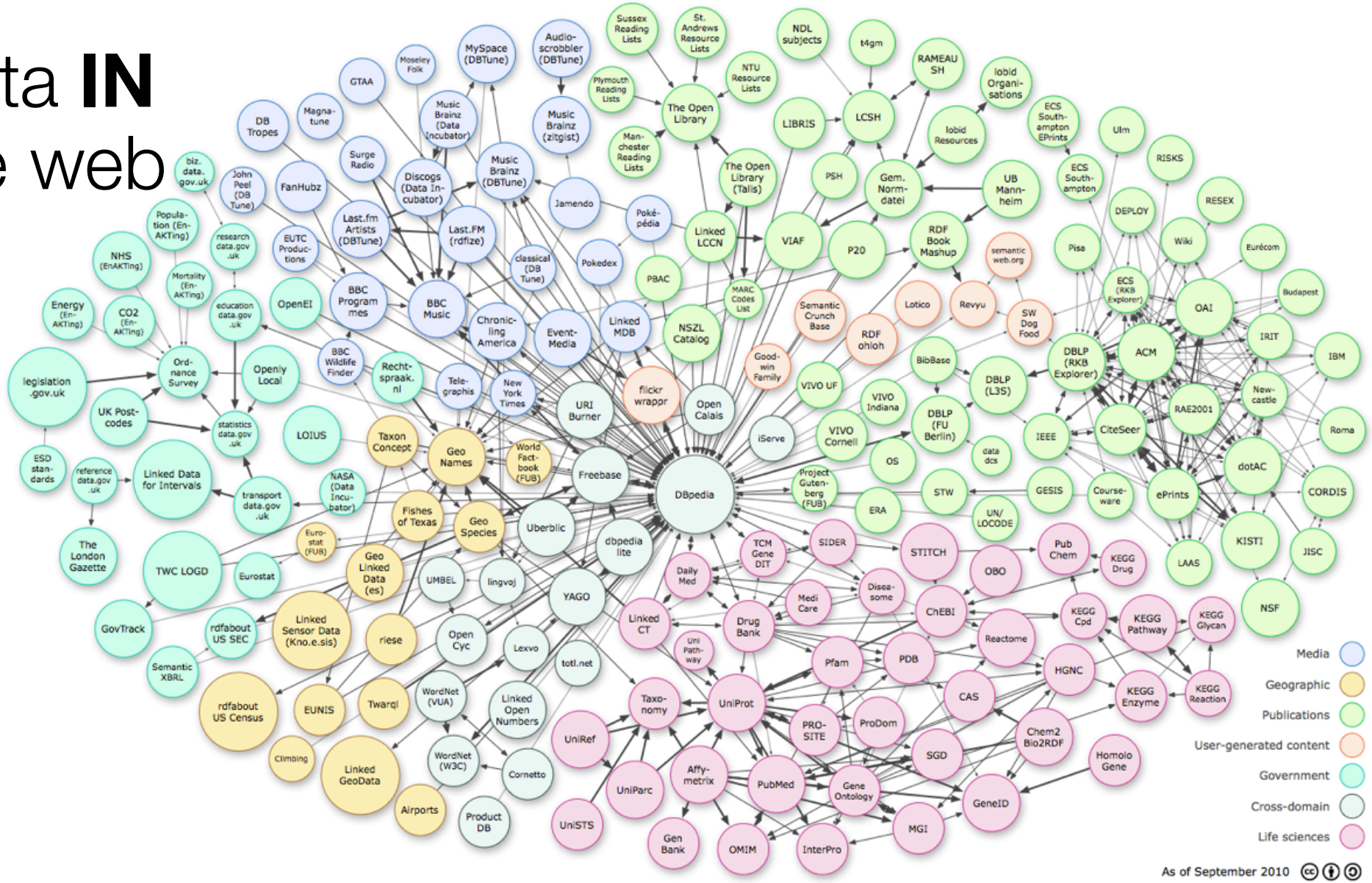
Content created by
The Open Data Institute

5-Stars

http://5stardata.info/



Data **IN** the web



Linked data

Amazing but challenging to publish and use

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

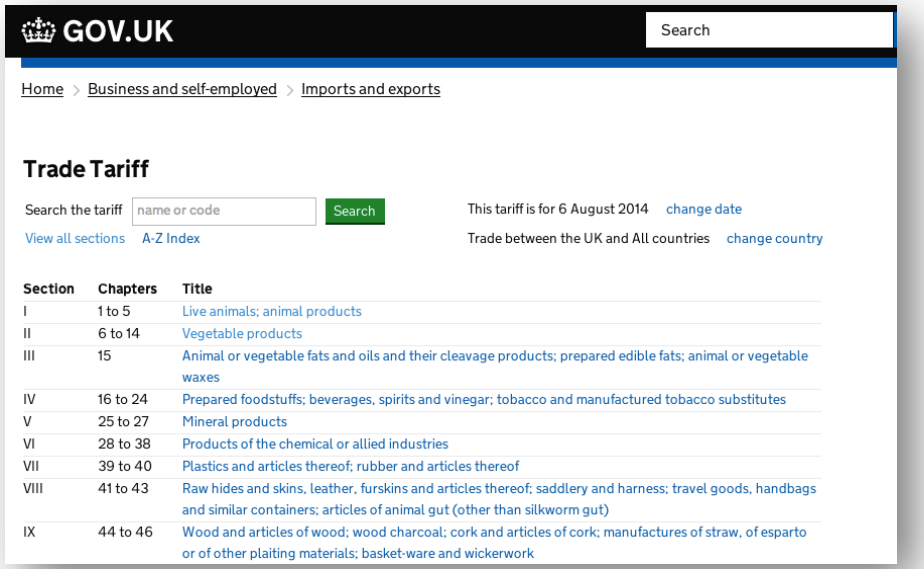
How the web should work, but people forgot that Tim put this in when he invented it!

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



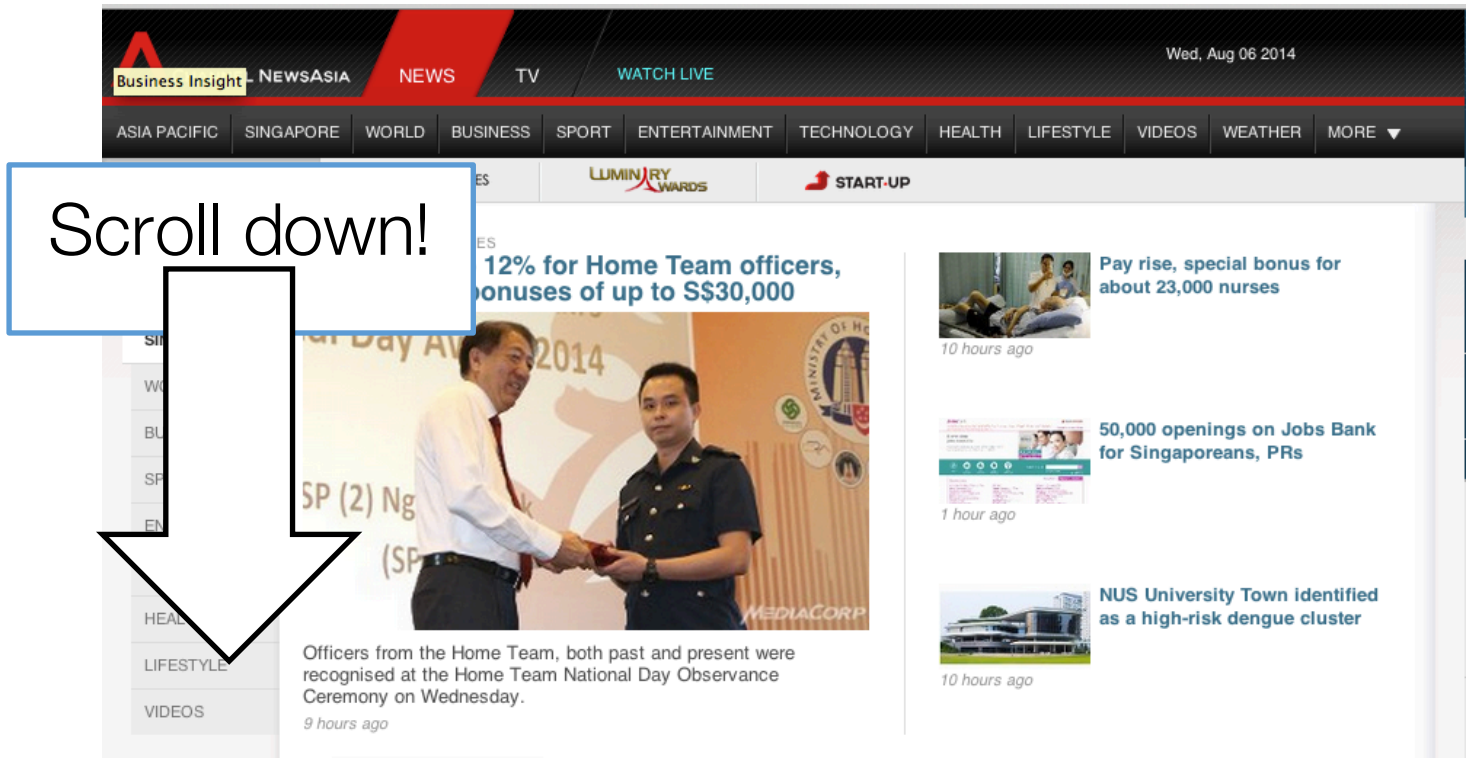
UK Trade Tariff

BBC Music and Programmes

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

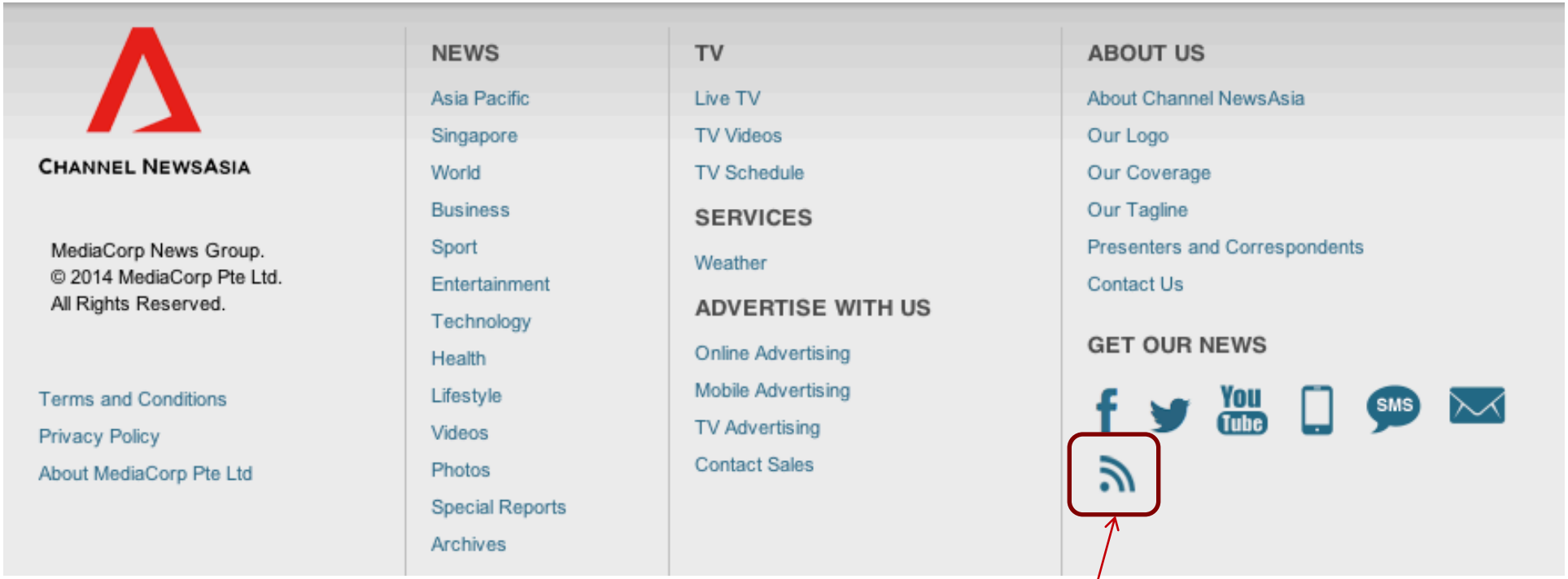
Content created by
The Open Data Institute

2. Look for alternative links



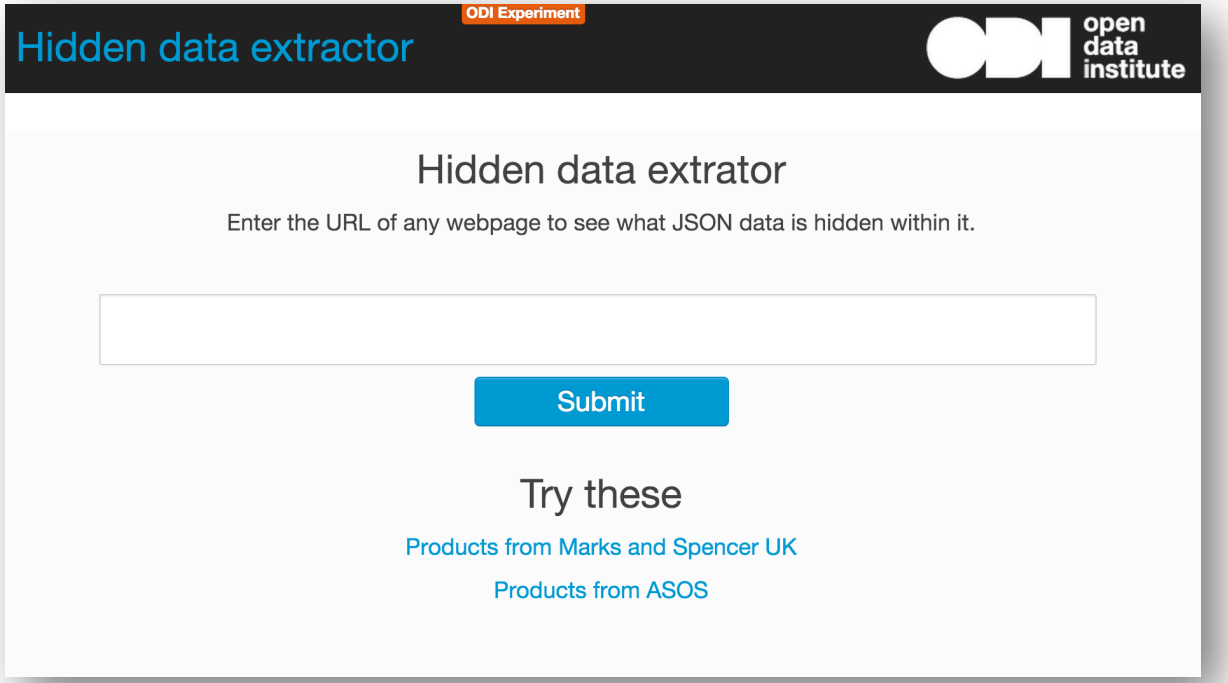
Content created by
The Open Data Institute

2. Look for alternative links



RSS

3. Look for embedded data



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!

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?