



Open Data in Practice

Dr David Tarrant

@davetaz

The Open Data Institute



Introductions

Your name

What is your favourite example/use of open data?

What do you want to do differently after the course?



Course aim

Build a solid foundation and experience in
publishing, consuming and building a
business in Open Data



Schedule

Day 1: Practical publication

Day 2: Business, the law and open data

Day 3: Enriching and visualising data



Today

The characteristics of open data

Open data discovery patterns

Open data publication

Quick big data break

Practical publication hands-on



Recap session

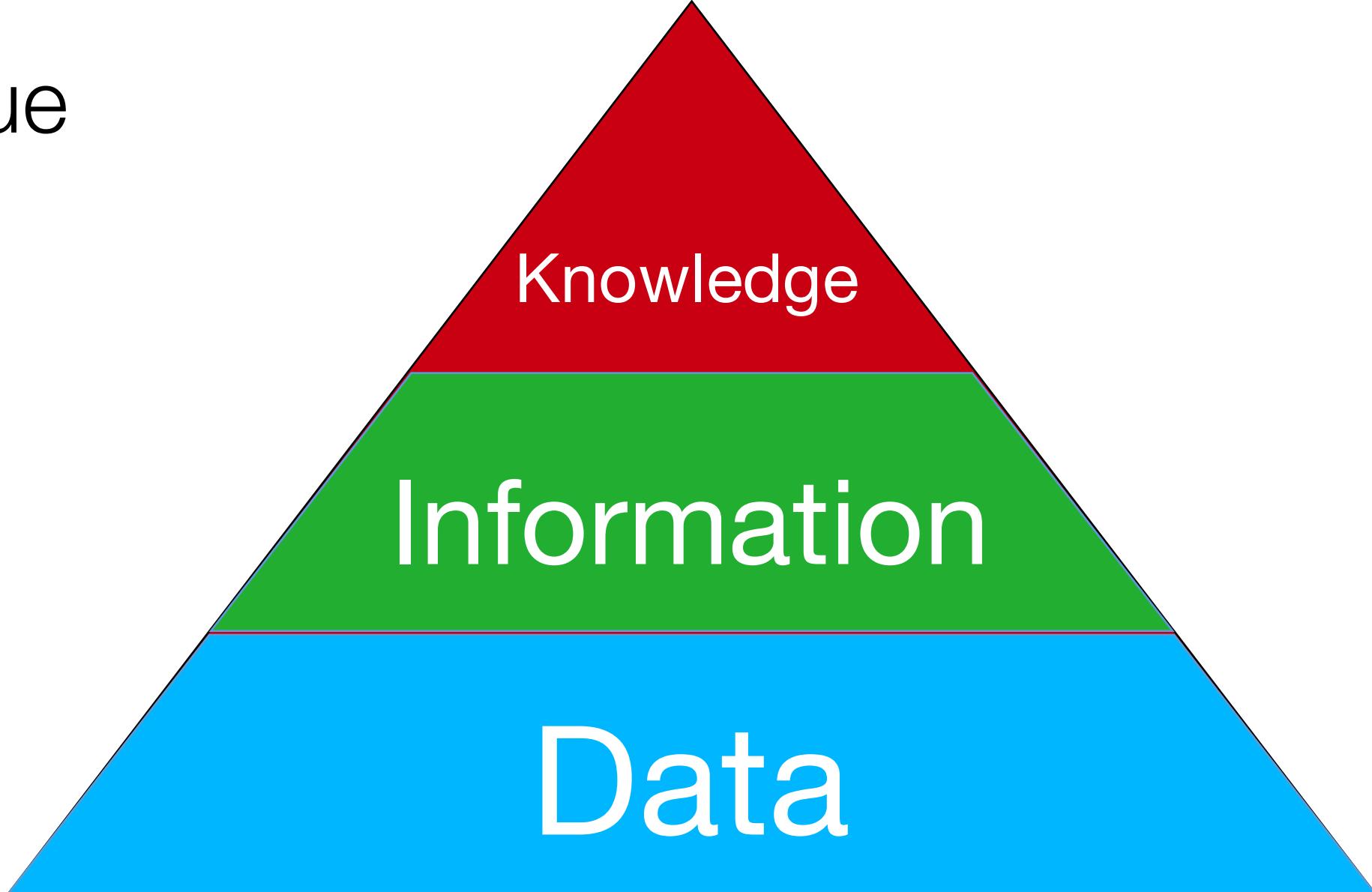


Exercise

What is Data?



Value



Exercise

What is Open Data?



Option A

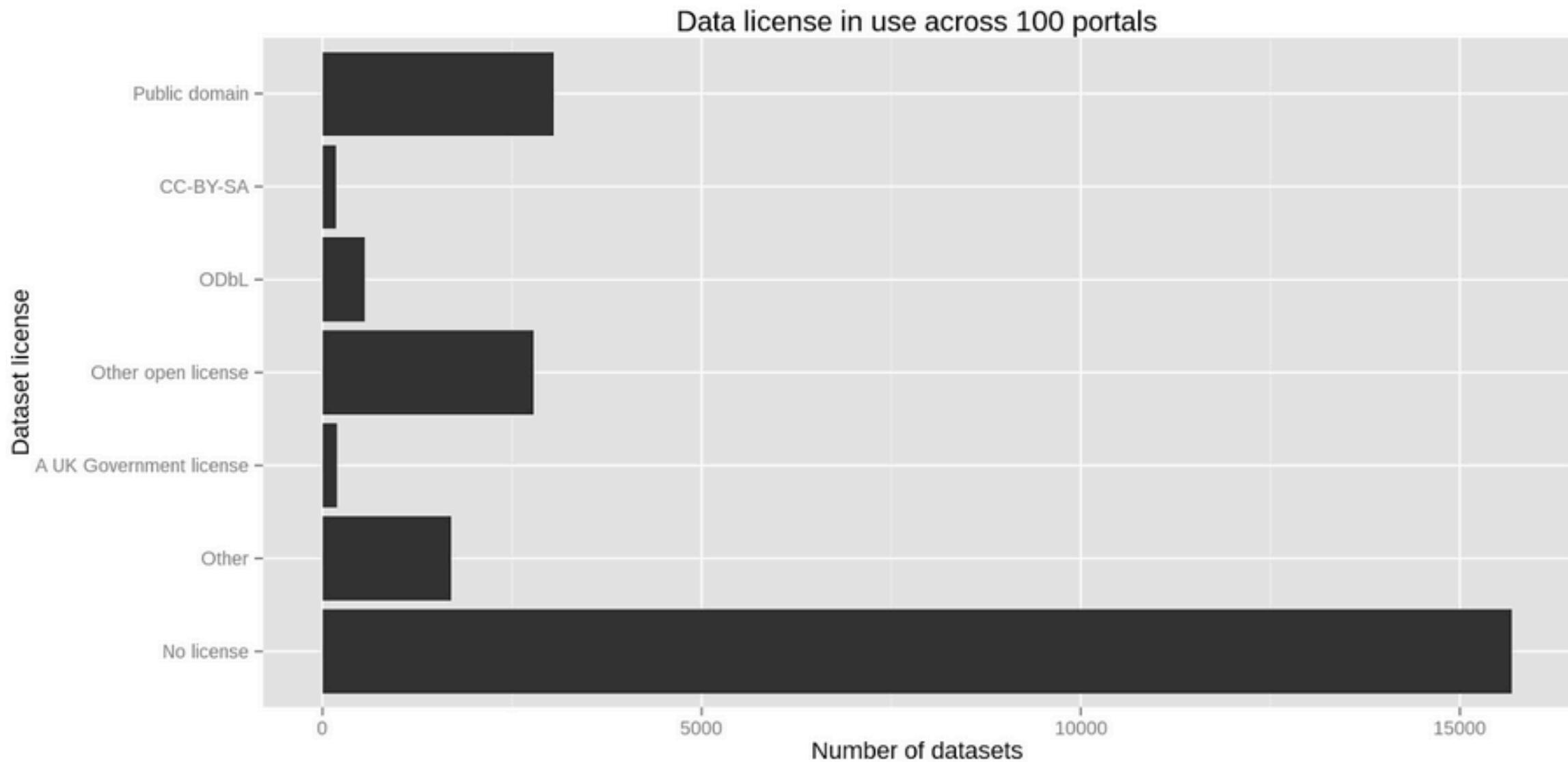
Open data is data that is made **available** by
organisations, businesses and **individuals** for
anyone to **access, use** and **share**.

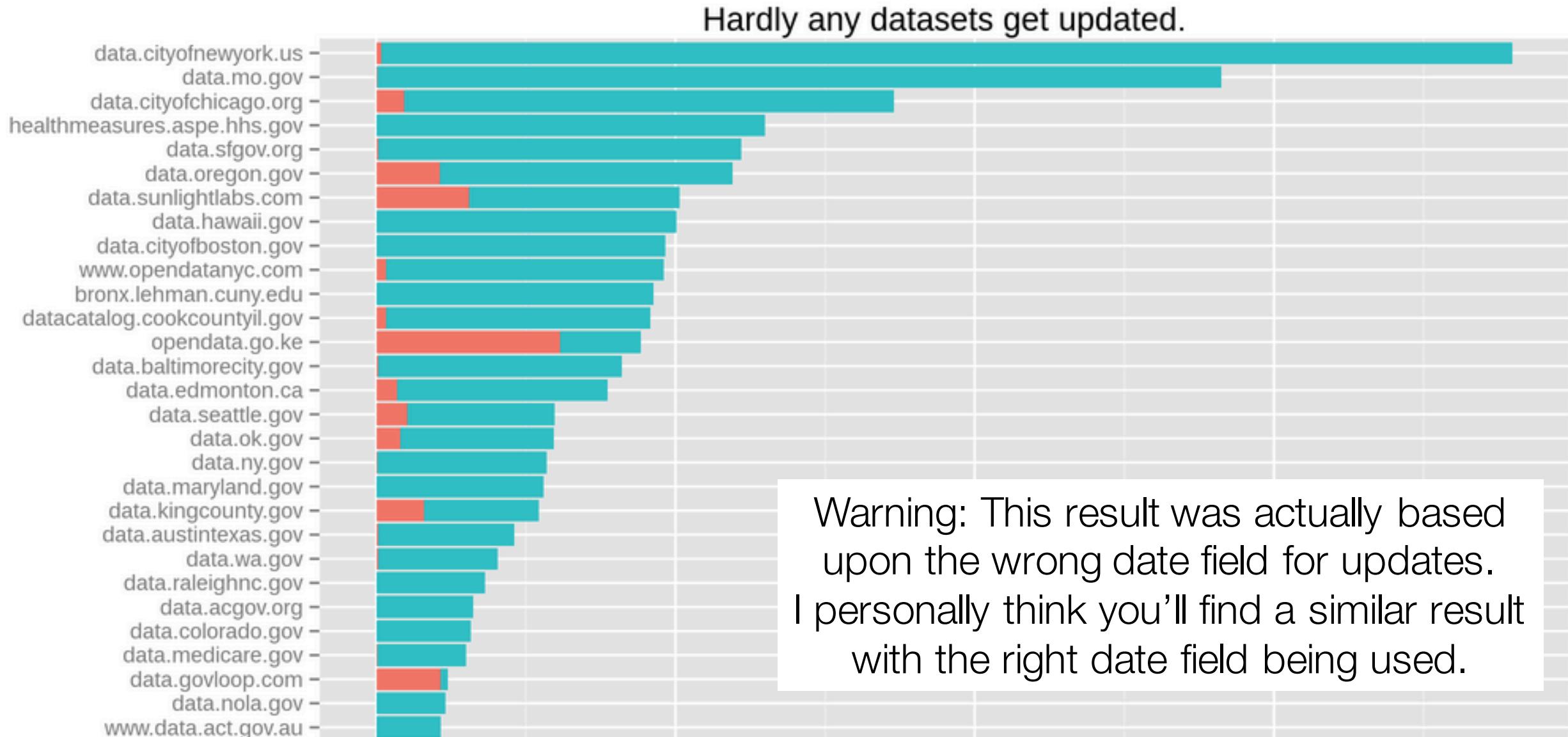
- Open Data Institute
Introduced November 2014



	Open Definition <i>Open Knowledge Foundation</i>	OMB Memo , 2013 <i>The White House Sylvia Burwell et al.</i>	Data.Gov.UK <i>Antonio Acuña</i>	"DBpedia: A Nucleus for a Web of Open Data <i>Sören Auer et al.</i>	Open Data Institute (ODI) <i>Open Data Institute</i>	LinkedGov <i>LinkedGov</i>	McKinsey <i>James Manyika et al.</i>	Open Data Now <i>Joel Gurin</i>	Open Data Barometer <i>Tim Davies</i>	The World Bank <i>The World Bank</i>
Free	✓	✓		✓	✓		✓			
Negligible Cost										
Publicly Available	✓	✓			✓		✓	✓		
Re-usable	✓		✓		✓		✓			
Can be Redistributed	✓			✓						
Non-exclusive (No Restrictions from copyright, patents, etc.)	✓			✓	✓					
Structured for Usability		✓	✓							
Requires "Open" License			✓		✓		✓			
Non Personally Identifiable						✓	✓	✓		
Produced during business operation						✓	✓			
Belongs to the Taxpayer (when not in violation of laws/privacy)						✓	✓			
Accessible in Bulk									✓	

Open data is hardly ever appropriately licensed.

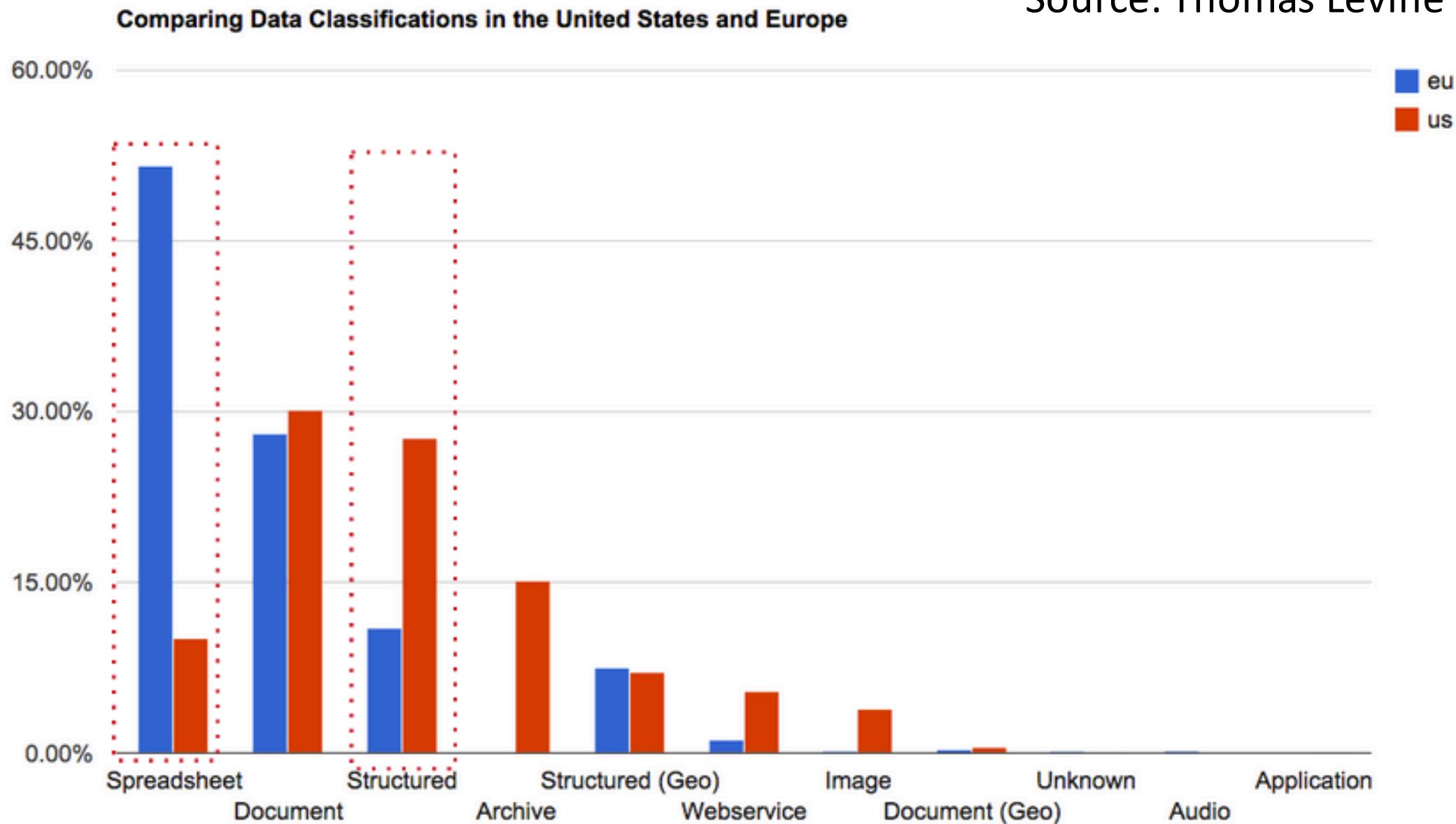




Source: Thomas Levine

Open data is rarely structured.

Source: Thomas Levine



Publication phases

Phase 1: Get the data online, in some form. This will help with the trust and transparency and community building.

Phase 2: Increase the usability of the data by potentially publishing differently and keeping it up to date.



Today's mission

To move to phase 2 of publishing open data and
solve some of the phase 1 problems



Guidelines



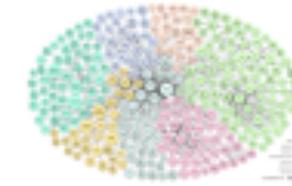
5 - S t a r s



5-Stars



<http://5stardata.info/>



<http://data...>



Open Data Certificate



Bronze level

self certified

GB final

Embed this
on your site

This data has achieved Bronze level on 29 March 2016 which means this data makes a great start at the basics of publishing open data.

Grants to voluntary community and social enterprise organisations

Summary

Type of release

ongoing release of a series of related datasets

Data Licence

UK Open Government Licence

Content Licence

Not Applicable

Verification

self certified

General Information

This data is described at

<http://data.hounslow.gov.uk/View/loc...>

This data is published by

London Borough of Hounslow

The data is published on

<http://data.hounslow.gov.uk/>

Legal Information

This data was

originally created or generated by its curator

The rights statement is at

<http://www.hounslow.gov.uk/index/c...>

This data is available under

UK Open Government Licence

There are

no rights in the content of the data



<http://certificates.theodi.org>



Content created by
The Open Data Institute

Open Refine

Google Refine 2.0 - Introduction (1 of 3) (vide...

Mass edit 2350 cells in column Type of Contract: Undo

Facet / Filter Undo / Redo ↻ Refresh Reset All Remove All

5200 rows Show as: new records Show: 5 10 25 50 rows ↻ First ↻ previous 1 - 10 next ↻ last ↻

Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	End Date	Total value of Contract	Contract Awarded
1. 1938	ASAP SOFTWARE EXPRESS INC DELL MARKETING LP	Microsoft Enterprise Agreement	04/01/2009	04/01/2009	06/03/2011	1,952	yes
2. 1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Ramsey Service Desk Maintenance	04/01/2009	04/01/2009	03/01/2010	0.001	yes
3. 1941	GO CONNECTION INCORPORATED	Cisco SmartNet	05/01/2009	05/01/2009	04/03/2011	0.007	yes
4. 1942	ITS CORPORATION	Time & Materials	12/01/2008	01/01/2009	12/03/2011	20	yes
5. 1943	ISINET INTERNATIONAL CORPORATION	Non Fixed Price	05/01/2009	05/05/2009	07/03/2009	0.04057	yes
6. 1945	IT FEDERAL SALES LIMITED LIABILITY COMPANY	Non Fixed Price	01/06/2009	01/26/2010	08/03/2010	0.708	yes
7. 1946	IT FEDERAL SALES LIMITED LIABILITY COMPANY	Non Fixed Price	10/01/2009	10/01/2009	06/25/2010	0.049	yes
8. 1947		Firm Fixed Price	09/03/2009	10/01/2009	08/02/2010	0.004	yes
9. 1948		Firm Fixed Price	11/05/2009	11/05/2009	05/03/2010	0.002	yes
10. 1949	PICHAWK IT SOLUTIONS LLC	Firm Fixed Price	01/22/2009	01/01/2010	12/01/2010	0.013	yes

YouTube 0:00 / 6:48

<http://openrefine.org>



Session 1

The characteristics of open data



Outcomes

- Identify a number of different characteristics of data
- Explain the justifications for publishing different types of data
- Evaluate the current open data ecosystem and future opportunities



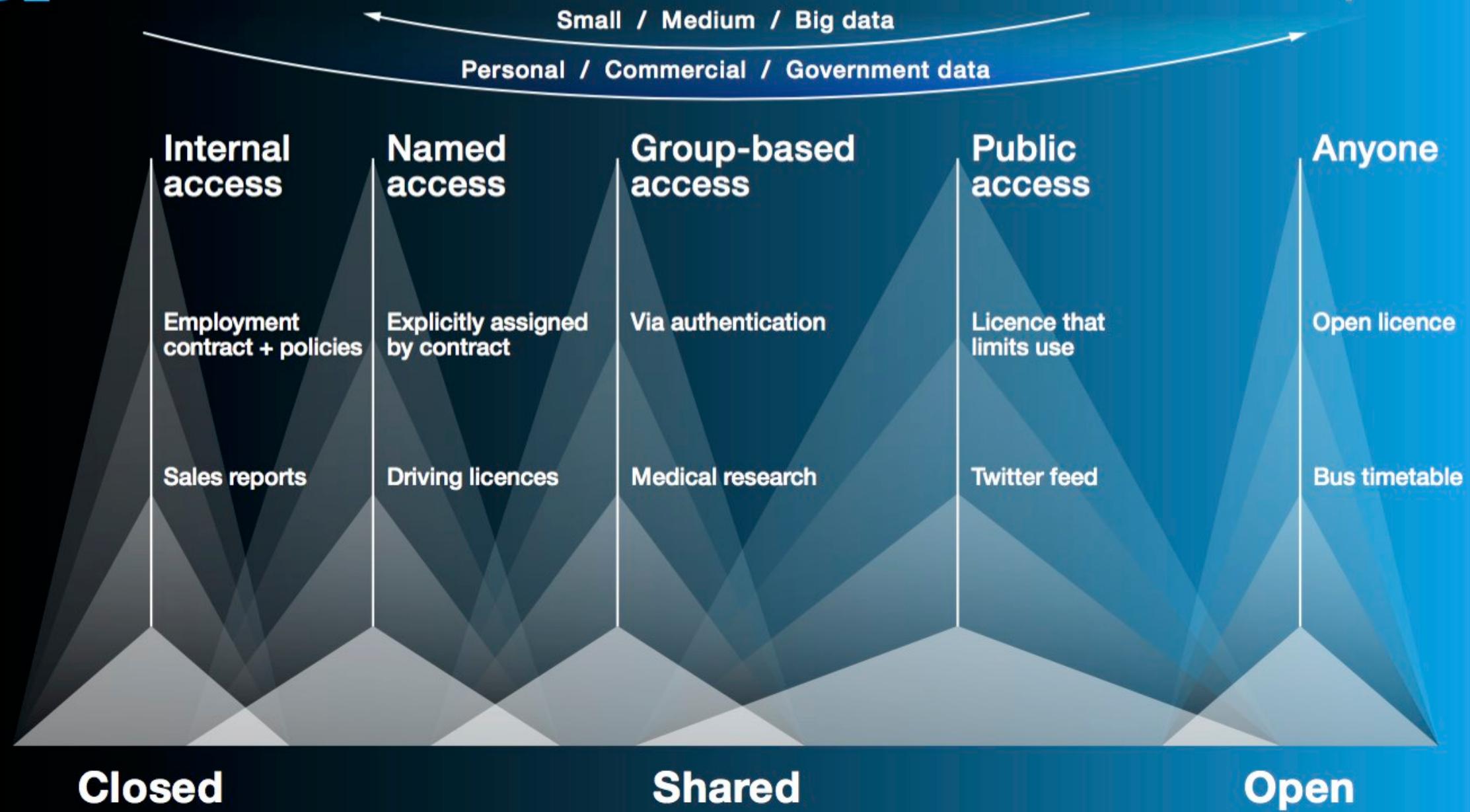
Exercise (part 1)

In your pre-training exercise, you were all asked to identify a dataset.

In your groups briefly discuss each others datasets and write down some key characteristics of each.

Also write the dataset title on a post-it, one per post-it.





Types of Data



Reference data

“things”

Transaction data

“stats involving things”



Exercise

Categorize your data into reference and transactional data.

If they are all in one category you have 2 minutes to add some new datasets to the empty category.

When done, put a “T” or and “R” on each dataset post-it.



Types of Data



Reference data

“things”

People Facilities Places
Books Buildings

Transaction data

“stats involving things”

Expenditure Weather
Consumption
Observation



Update frequency

Static

In frequent updates

Frequent updates

Live



Exercise

Categorize your data into **frequency of updates**

If they are all in one category you have 2 minutes to add some new datasets to the empty category/ies

Put a number on your post-its representing the frequency of updates.

0 = static, 1 = In frequent, 2 = Frequent, 3 = Live

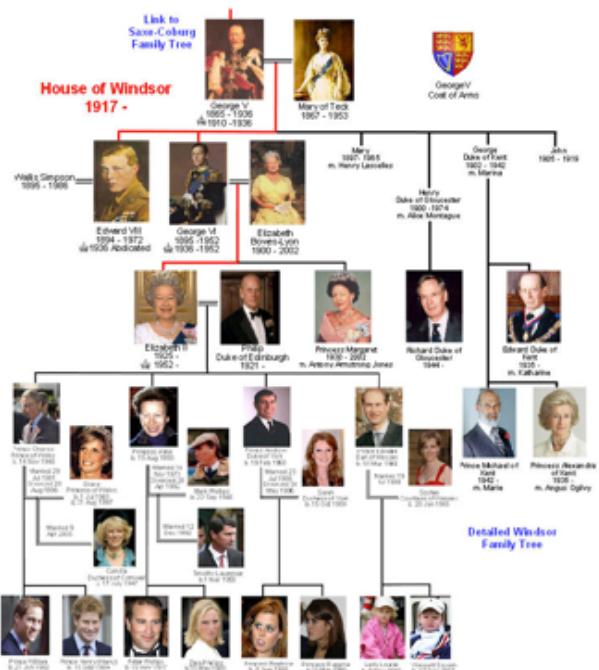


Data Representations

Tabular

Region	Production			YTD Production (billion MT)
Country	Production (thousand MT)	Change from last year	Change from 5 year average	YTD Production (billion MT)
Brazil	57289	-4.05%	+2.00%	
Mato Grosso	18,008	0.90%	6.17%	
Parana	9,571	-19.55%	-9.08%	
Rio Grande do Sul	7,844	0.88%	9.35%	
Goiás	6,820	4.23%	5.27%	
Mato Grosso do Sul	4,218	-7.60%	-1.97%	
Minas Gerais	2,667	5.12%	2.41%	
Bahia	2,512	-8.58%	4.84%	
Sao Paulo	1,392	-3.77%	-6.81%	
Maranhão	1,087	-13.93%	0.58%	
Santa Catarina	1,039	9.81%	13.35%	
Tocantins	902	-0.90%	7.05%	
Piauí	856	4.49%	23.75%	
Para	194	-3.13%	1.02%	
Distrito Federal	155	1.37%	-1.11%	
Roraima	22	-54.10%	-41.70%	

Hierarchical



Network/Graph



Exercise

Categorize your data into **tabular, hierarchical (tree) and graph (network)**

If they are all in one category you have 2 minutes to add some new datasets to the empty category.

Add the word “**tab**”, “**tree**” or “**net**” to your post-its to represent the different structures.



Justifications

Trust and
Transparency

Enabling the
economy



One more

Categorize your data into **transparent** and **enabling**.



Summing up

Do you have any obvious grouping of your datasets?

Is this reflective of the whole open data ecosystem?



Policy paper

G8 Open Data Charter and Technical Annex

Published 18 June 2013

Contents

1. Principle 1: Open Data by Default
2. Principle 2: Quality and Quantity
3. Principle 3: Usable by All
4. Principle 4: Releasing Data for Improved Governance
5. Principle 5: Releasing Data for Innovation
6. Technical annex

Exercise

Pick one “group” of datasets that share similar colours and come up with a data publication strategy for getting these datasets online and usable.

What are the publication requirements on the human publisher?

What are the requirements on potential users?



Outcomes

- Identify a number of different characteristics of data
- Explain the justifications for publishing different types of data
- Evaluate the current open data ecosystem and future opportunities



Session 2

Open 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.”

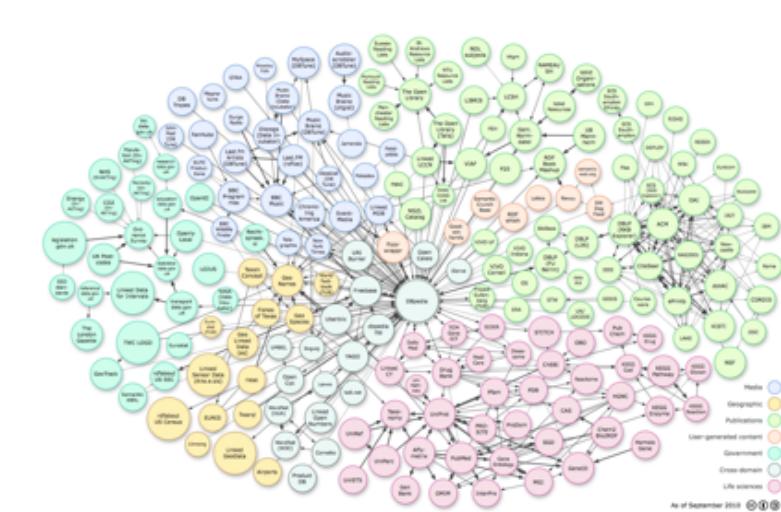


Approaches to publishing data

ON the web



IN the web



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

Government data

Private sector data

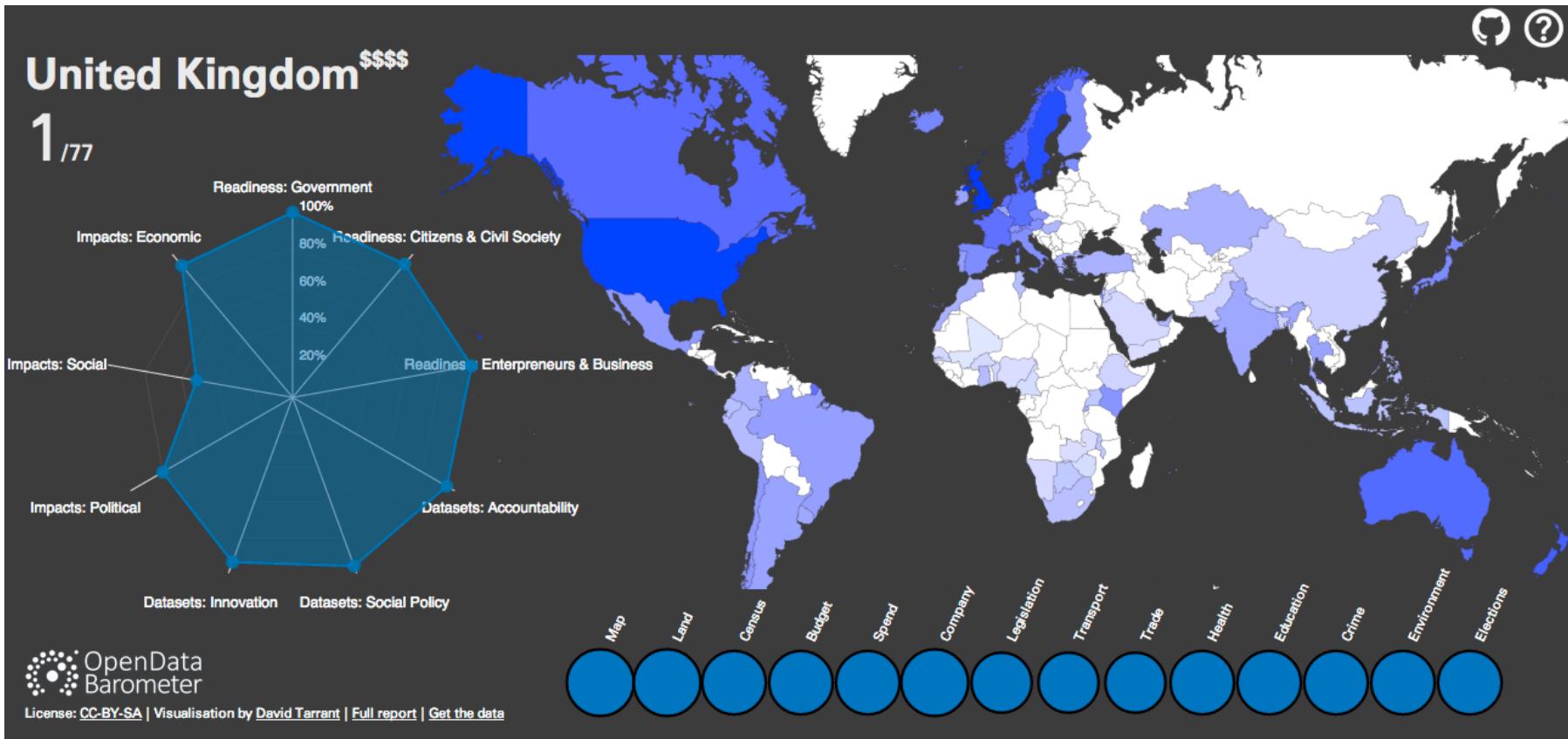
Google advanced

Aggregators and portals

Scraping



Government data



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



data.gov.XX

The screenshot shows the DATA.GOV.UK homepage with a search bar and navigation menu. A sidebar on the left lists categories like Datasets, Map Search, Data Requests, Publishers, Public Roles & Salaries, Spend Reports, Site Analytics, and Reports. The main content area displays a search result for "19266 Results" related to "Live traffic information from the Highways Agency". Below this, there's a map titled "StreetMap" showing locations in the United States and Mexico. The footer features the OpenData Burkina Faso logo and a Creative Commons BY license icon.

The screenshot shows the DATOS.GOB.MX BETA website. At the top, there's a navigation bar with icons for Data, Stories, and Advances, and tabs for Datasets and Organizations. The main search bar has the placeholder "Buscar conjuntos de datos...". Below it, a section titled "127 datasets found" is shown. To the right, there's a section titled "RATING SOCIAL PROGRAMS" with a brief description. The footer includes the Open Data Institute logo and the text "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](#)

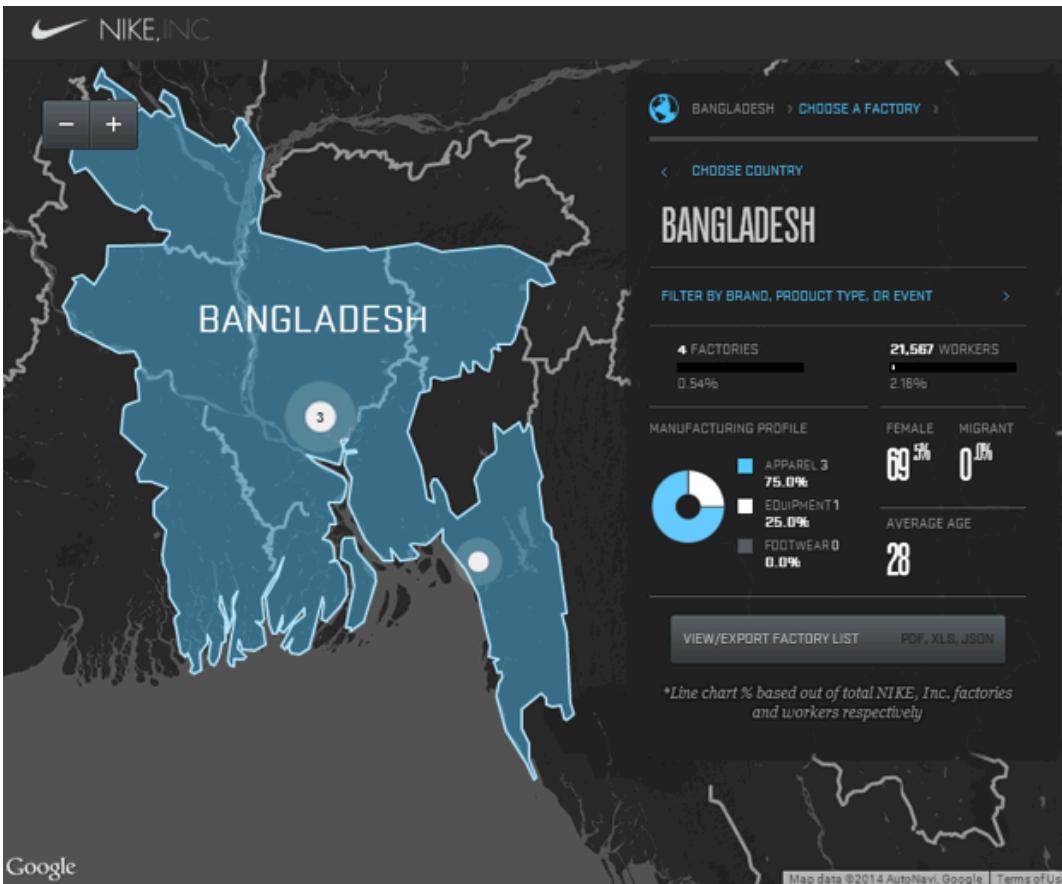
[What we know](#)

<u>Pre Take-Off</u>								
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 16:00:13.906	IOR-P10500-0-3859	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	103	14740
7/03/2014 16:00:18.406	IOR-R1200-0-3603	IOR	305	8	R-Channel RX	Eight Octet User Data	103	14780
7/03/2014 16:00:18.905	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	0x62 - Acknowledge User Data	103	14820
7/03/2014 16:00:20.906	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	0x71 - User Data (ISU) + RLS		
7/03/2014 16:00:20.906	IOR-P10500-0-3859	IOR	305	10	P-Channel TX	Subsequent Signalling Unit		
7/03/2014 16:00:22.906	IOR-R1200-0-3603	IOR	305	10	P-Channel TX	Subsequent Signalling Unit		
7/03/2014 16:00:23.407	IOR-R1200-0-3603	IOR	305	8	P-Channel TX	Subsequent Signalling Unit		
7/03/2014 16:00:23.905	IOR-P10500-0-3859	IOR	305	8	R-Channel RX	0x62 - Acknowledge User Data		
7/03/2014 16:00:27.741	IOR-T1200-0-3607	IOR	88	88	R-Channel RX			
7/03/2014 16:00:27.901	IOR-T1200-0-3607	IOR	88	88	R-Channel RX			
7/03/2014 16:00:28.061	IOR-T1200-0-3607	IOR	88	88	R-Channel RX			
7/03/2014 16:00:28.221	IOR-T1200-0-3607	IOR	88	88	R-Channel RX			
7/03/2014 16:00:28.405	IOR-T1200-0-3607	IOR	88	88	R-Channel RX			
7/03/2014 16:00:28.541	IOR-T1200-0-3607	IOR	88	88	R-Channel RX			

OPEN DATA?



Suppliers



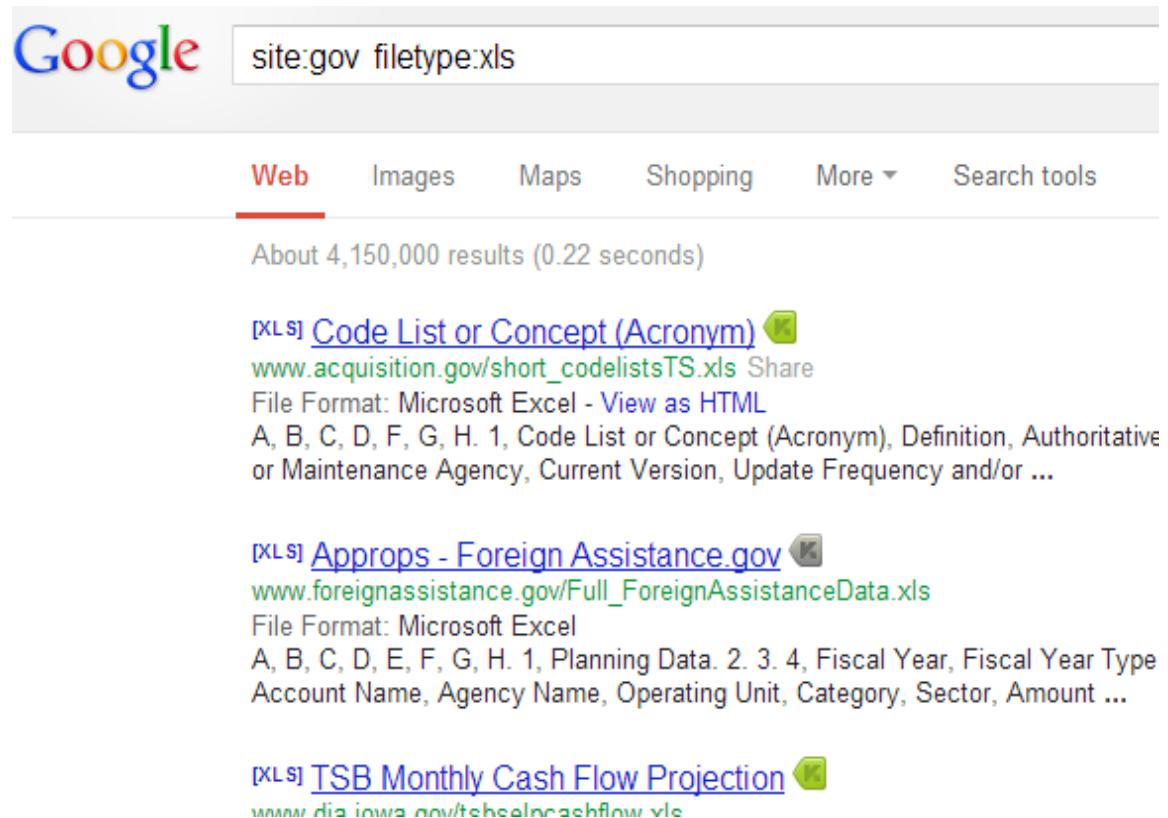
You agree not to change or delete any ownership notices from materials downloaded or printed from the Platform. You agree not to modify, copy, translate, broadcast, perform, display, distribute, frame, reproduce, republish, download, display, post, transmit or sell any Intellectual Property or Content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



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



Google advanced



Google search results for "site:gov filetype:xls". The results show three links related to government Excel files:

- [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/tech/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



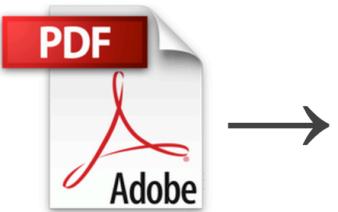
transportAPI



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.



→

Variety Name	Total Receipts	Total Receipts	Total Inedibles	Receipts
Aldrich	49,043,476	48,900,332	560,403	2.61%
Aviation	8,399,120	8,322,233	136,796	7.79%
Butte	144,863,828	147,182,233	1,186,402	7.70%
Butte/Padre	215,647,498	213,129,006	1,118,151	11.47%
Carmel	171,682,007	170,460,781	1,571,020	9.13%
Canyon	138,180,120	137,500,000	1,680,120	1.21%
Fritz	110,175,472	110,246,890	1,854,841	5.86%
Harvey	36,907	36,907	1,181	0.00%
Hashem	347,618	348,680	3,812	0.02%
LeGrand	8,797	8,557	261	0.00%
Livingston	7,394,450	7,500,200	159,210	0.39%
Marchini	418,442	418,442	9,096	0.02%
Mered	57,800	57,768	846	0.00%
Mission	17,265,270	17,174,884	78,967	0.92%

A screenshot of the import.io web 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 four main examples: "Reseller Ratings" (a screenshot of a ResellerRatings.com page), "Zoopla" (a screenshot of a Zoopla property listing page), "500px" (a screenshot of a 500px photo gallery page), and "Growth Hackers" (a screenshot of a GrowthHackers.com blog post). Further down, there are two more examples: "Udemy" (a screenshot of a Udemy course page) and "Stack Exchange" (a screenshot of a Stack Exchange question page).

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

pdftables.com

magic.import.io



5-Stars



<http://5stardata.info/>

ON THE WEB



IN THE WEB

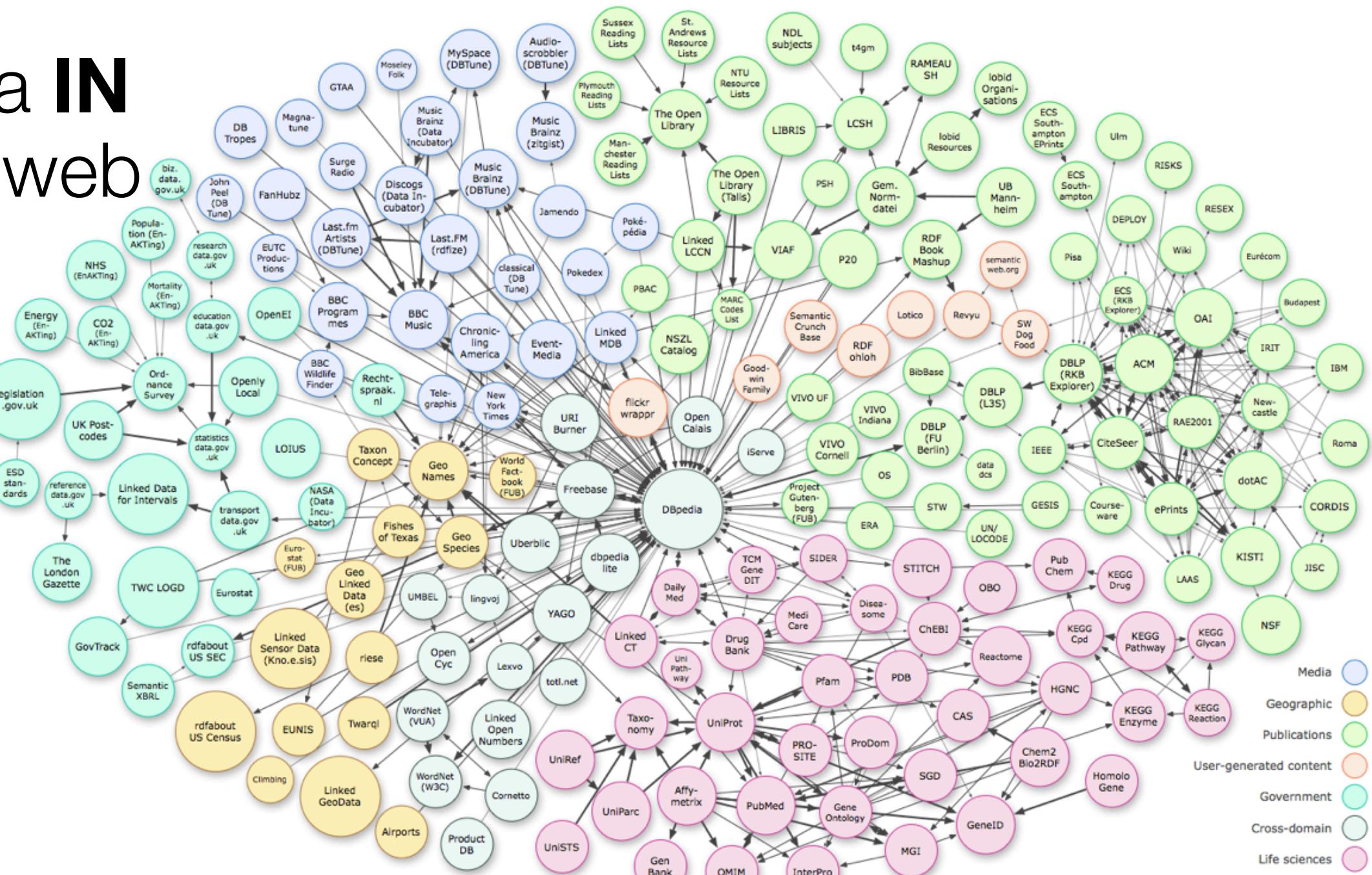


OPEN DATA



Content created by
The Open Data Institute

Data IN the web



Linked data

Amazing but challenging to publish and use

EEE Building

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

Detail	Facilities	Services	Energy
<p>Site: Highfield Campus</p> <p>Construction: 2006</p> <p>Architect: John McAslan & Partners</p> <p>Features: Building 32 is non-residential</p> <p>View Disability Report for this Building</p>			


©2010 Francois-Xavier Beckers (CC-BY)

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



[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
- [southampton.ac.uk/ns/disabledGoPage](#) → <http://www.disabledgo.com/en/access-guide/building->
- locations:easting → "442544"^^xsd:integer
- locations:northing → "115392"^^xsd:integer
- Organization → University of Southampton
- [southampton.ac.uk/ns/ombielName](#) → "Bldg 32 (EEE)"
- feature → Building 32 is non-residential
- [southampton.ac.uk/ns/buildingDate](#) → "2006"
- [southampton.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))"
- [southampton.ac.uk/ns/electricityTimeSeries](#) → "elec/b32/ekw"
- ← is spatialrelations:within of ← 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 ← RaeS Solent Branch Christmas Special Lecture - The Red Arrows


Google
Infoterra Ltd & Bluesky, The GeoInformation Group



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 GOV.UK Trade Tariff homepage. At the top, there's a navigation bar with the GOV.UK logo and a search bar. Below it, a breadcrumb trail reads "Home > Business and self-employed > Imports and exports". The main content area is titled "Trade Tariff". It features a search bar with placeholder "Search the tariff name or code" and a "Search" button. Below the search bar, text indicates the tariff is for 6 August 2014 with a "change date" link. There are also links to "View all sections" and "A-Z Index". A table lists tariff sections I through IX with their respective chapter ranges and titles.

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. At the top, there's a navigation bar with the "one" BBC logo and links for "Home", "Episodes", "Clips", "Galleries", "Latest News", "Characters", "Monsters", "Fun and Games", and "More". The main banner features the Doctor Who logo against a yellow background. Below the banner, there are two columns of content. The left column features a thumbnail image of the Doctor and a companion, with text encouraging users to get the latest on the launch. The right column features a thumbnail image of the Doctor and a companion, with text about the Day of the Doctor broadcast on Saturday at 19:00 on BBC THREE, with links to all upcoming episodes.

BBC Music and Programmes





2. Look for alternative links

The screenshot shows the NewsAsia website homepage from Wednesday, August 06, 2014. The top navigation bar includes 'Business Insight' logo, 'NEWSASIA' tab, 'NEWS', 'TV', 'WATCH LIVE' button, and the date 'Wed, Aug 06 2014'. Below the navigation is a horizontal menu with categories: ASIA PACIFIC, SINGAPORE, WORLD, BUSINESS, SPORT, ENTERTAINMENT, TECHNOLOGY, HEALTH, LIFESTYLE, VIDEOS, WEATHER, and MORE. A large blue rectangular box highlights the 'NEWS' tab, and a large black arrow points downwards from it towards the main content area. The main content features a large image of two men shaking hands at a ceremony. Headlines include: '12% for Home Team officers, bonuses of up to S\$30,000', 'Pay rise, special bonus for about 23,000 nurses', '50,000 openings on Jobs Bank for Singaporeans, PRs', and 'NUS University Town Identified as a high-risk dengue cluster'. On the left sidebar, there is a vertical list of news categories: SIM, NEWS, WORLD, BUSINESS, SPORT, ENTERTAINMENT, HEADLINES, LIFESTYLE, and VIDEOS.

Scroll down!

12% for Home Team officers, bonuses of up to S\$30,000

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





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

The screenshot shows the 'Hidden data extractor' tool. At the top, there's a dark header bar with the text 'Hidden data extractor' in blue on the left and 'ODI Experiment' in orange on the right. To the right of the header is the ODI logo. Below the header is a light gray main area. In the center, it says 'Hidden data extractor' and 'Enter the URL of any webpage to see what JSON data is hidden within it.' There is a large input field for pasting a URL, followed by a blue 'Submit' button. Below the input field, there's a section titled 'Try these' with two links: 'Products from Marks and Spencer UK' and 'Products from ASOS'.

<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?



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.”



Session 3

Open data publication platforms



Outcomes

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

Evaluate a number of different approaches for publishing open data.

Publish a dataset on one of the many available platforms.



Types

Specialist Solution

- + Easy to get setup and maintain.
- + Open Data focused
- + Clear workflows for publishing open data
- + Visualisation tools
- + Data mashing tools
- + Best for transactional data

Integrated Solution

- + No new platform to learn
- + Data is provided in parallel to web pages
- + No separation from authoritative data
- + Easy discovery of data
- + Best for reference data
- + Best for Linked Open Data



Key characteristics of specialist solution

Separate from your main org website

Designed to publish open data, not to fulfill other org goals

Examples: data.gov.uk | data.cityofchicago.org | data.sncf.com



Key characteristics of integrated solution

It is your main website

Publishes data alongside everything else that the organisation does

Examples: bbc.co.uk/programmes | southampton.ac.uk | gov.uk



Merging specialist and integrated

Method 1: *Build the functionality of your current website into a new open data platform.*

Example: [West Sussex Country Council](#)

Method 2: *Hide the specialist solution behind your main website and use it as a loosely coupled CMS.*

Example: [southampton.ac.uk](#)



The sliding scale of specialist solutions

1. Catalogue: Point to data (leave it at source)
2. Re-present: Provide data services (leave it at source)
3. Host the data: Be the source
4. Control the data: Be the authority
5. Be the hub: Host the data and processor



Specialist Solutions



comprehensive knowledge
archive network

<http://www.flickr.com/photos/okfn>



OpenDataSoft



Socrata



 Content created by
The Open Data Institute



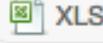
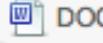
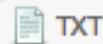
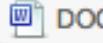
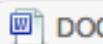
Overview

Licence

Open Government Licence **OGL**

OPEN DATA

Data Resources 6

-  Tariff index
-  Tariff data
-  Tariff data - overview
-  Tariff section index
-  Tariff section notes
-  Tariff chapter notes

Open Knowledge Foundation Supported

Data Catalogue

Open Source

Feels like a record manager

Simple API and search

Lots of community tools

<http://demo.ckan.org/>



Content created by
The Open Data Institute



Evolution of CKAN (v1.4)

Early) Dataset catalogue (data.gov.uk)

no data hosted or searched

Mid) Data and dataset catalogue

no data hosted but it is searchable

Now) Integrated data driven web site

data platform is integrated with data, search and content





Features

Publish, Store and Manage Data and Metadata

Visual and Geospatial

Social

Full Stored History

Federate Your Data With Other Organizations

Rich RESTful JSON API for Developers



Overview

Prix des carburants - J-7

Information Table Map Analyze Export API Share ▾

Nom	Marque	Carburant	Prix Gazole	Prix SP95	Prix SP98
1 MME POGGIO	Total	Gazole	1.35	1.56	1.63
2 RELAIS HAGONDANGE	Total	Gazole/E10	1.367		
3 AGIP CAVAILLON CLEMENCEAU	Agip	Gazole/E10/SP98	1.36		1.59
4 Carrefour Market	Carrefour Market	Gazole/SP95	1.337	1.589	1.58
5 INTERMARCHE RABASTENS	Intermarché	Gazole	1.317	1.519	1.539
6 DAINVILDIS	Leclerc	SP95		1.476	
7 LES BERJALLIENS	Total	Gazole/E10/SP98	1.36		1.59
8 Avia lecourbe	Avia	Gazole	1.46	1.7	1.72
9 Sarl STATION KLEBER	Elan	Gazole/SP95/SP98	1.79	1.94	1.96
10 BP PARIS PAUL DOUMER	BP	Gazole/E10	1.439		
11 RELAIS TOTAL MARECHAL DAVOUT	Total	Gazole/E10/SP98	1.365		1.648
12 ESSO DU MIN	Esso Express	E10	1.222		1.502
13 INTERMARCHE LILLEBONNE	Intermarché	Gazole/SP95/SP98	1.259	1.499	1.529
14 CARREFOUR GRUCHET	Carrefour	Gazole/SP95	1.189	1.429	1.529
15 SARL PICHON	Avia	Gazole/SP95/SP98	1.34	1.53	1.57



Data as a Service (DaaS)

Hosted enterprise solution

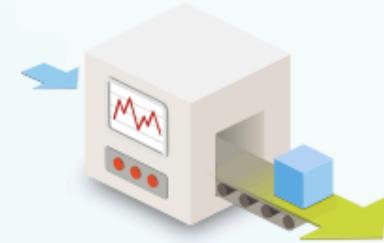
Query based API (3-Star)

EU Based

Closed source



Features



Prepare data

- Collect from any source
- Read every kind of formats
- Understand all data types
- Make datasets findable & reusable



Reuse data

- Full-featured secured API
- Standard access formats
- Interactive & shareable visualization
- Web extensions & open source



Scale services

- Big data technologies
- 100% cloud-based
- API factory
- Monitoring

Overview

The screenshot shows the Socrata platform interface. At the top is a search bar with a magnifying glass icon. Below it is a 'View Types' section with a list of icons and labels: Datasets (orange), Charts (yellow), Maps (green), Calendars (teal), Filtered Views (blue), External Datasets (green), Files and Documents (orange), Forms (purple), and APIs (red). At the bottom is a navigation bar with icons for Manage (with a red '9' badge), More Views, Filter, Visualize, Export, Discuss, Embed, and About.



Data as a Service (DaaS)

Hosted enterprise solution

Allows user created content

Full linked API (5-Star)

US Based

Closed source



Features



Data Publishing, Optimized for Business Users

Flexible Metadata Management

Federate Your Data With Other Organizations

Metrics of the Success of Your Initiative in Real-time

Anyone Can Create **Maps and Charts**

Data Becomes **Social**

Developers Are Supported Every Step of the Way



The sliding scale of specialist solutions

1. Catalogue: Point to data (leave it at source)
2. Re-present: Provide data services (leave it at source)
3. Host the data: Be the source
4. Control the data: Be the authority
5. Be the hub: Host the data and processor



Integrated solutions

Integrated solutions expose data using the current infrastructure (web pages).

Data driven web site

Best for reference and live data



The developers secret



[Home](#) > [Business and self-employed](#) > [Imports and exports](#)

Trade Tariff

Search the tariff [Search](#)

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

```
--  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
  
{"id": 1,  
 "position": 1,  
 "title": "Live animals; animal products",  
 "numeral": "I",  
 "chapter_from": "01",  
 "chapter_to": "05",  
 "section_note_id": 1}  
  
{"id": 2,  
 "position": 2,  
 "title": "Vegetable products",  
 "numeral": "II",  
 "chapter_from": "06",  
 "chapter_to": "14",  
 "section_note_id": 6}  
  
{"id": 3,  
 "position": 3,  
 "title": "Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes",  
 "numeral": "III",  
 "chapter_from": "15",  
 "chapter_to": "15",  
 "section_note_id": null},  
{  
 "id": 4,  
 "position": 4,  
 "title": "Prepared foodstuffs; beverages, spirits and vinegar; tobacco and manufactured tobacco substitutes",  
 "numeral": "IV",  
 "chapter_from": "16",  
 "chapter_to": "24",  
 "section_note_id": null},  
{  
 "id": 5,  
 "position": 5,  
 "title": "Mineral products",  
 "numeral": "V",  
 "chapter_from": "25",  
 "chapter_to": "27",  
 "section_note_id": null},  
{  
 "id": 6,  
 "position": 6,  
 "title": "Products of the chemical or allied industries",  
 "numeral": "VI",  
 "chapter_from": "28",  
 "chapter_to": "38",  
 "section_note_id": null},  
{  
 "id": 7,  
 "position": 7,  
 "title": "Plastics and articles thereof; rubber and articles thereof",  
 "numeral": "VII",  
 "chapter_from": "39",  
 "chapter_to": "40",  
 "section_note_id": null},  
{  
 "id": 8,  
 "position": 8,  
 "title": "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)",  
 "numeral": "VIII",  
 "chapter_from": "41",  
 "chapter_to": "43",  
 "section_note_id": null},  
{  
 "id": 9,  
 "position": 9,  
 "title": "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",  
 "numeral": "IX",  
 "chapter_from": "44",  
 "chapter_to": "46",  
 "section_note_id": null}
```



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
<p>Site: Highfield Campus</p> <p>Construction: 2006</p> <p>Architect: John McAslan & Partners</p> <p>Features: Building 32 is non-residential</p> <p>View Disability Report for this Building</p>			


©2010 Francois-Xavier Beckers (CC-BY)

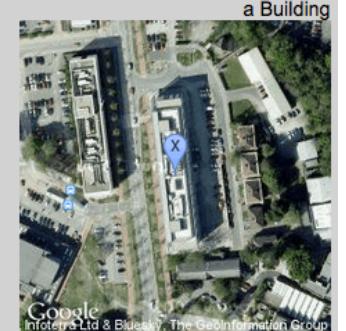
Occupants

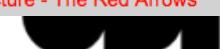
- Electronics & Computer Science
- Southampton Education School
- Agents, Interactions & Complexity
- Web & Internet Science
- Leadership School Improve &Effectiveness
- Lifelong & Work-Related Learning
- Mathematics & Science Education
- Social Justice & Inclusive Education
- Teaching Only Staff
- Deanery



Properties and Values

- 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...
- construction → "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
- southampton.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
- southampton.ac.uk/ns/ombielName → "Bldg 32 (EEE)"
- feature → Building 32 is non-residential
- southampton.ac.uk/ns/buildingDate → "2006"
- southampton.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))"
- southampton.ac.uk/ns/electricityTimeSeries → "elec/b32/ekw"
- ← is spatialrelations:within of ← 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 ← RaeS Solent Branch Christmas Special Lecture - The Red Arrows


a Building
Google
Infoxis Ltd & Bluesky, The GeoInformation Group

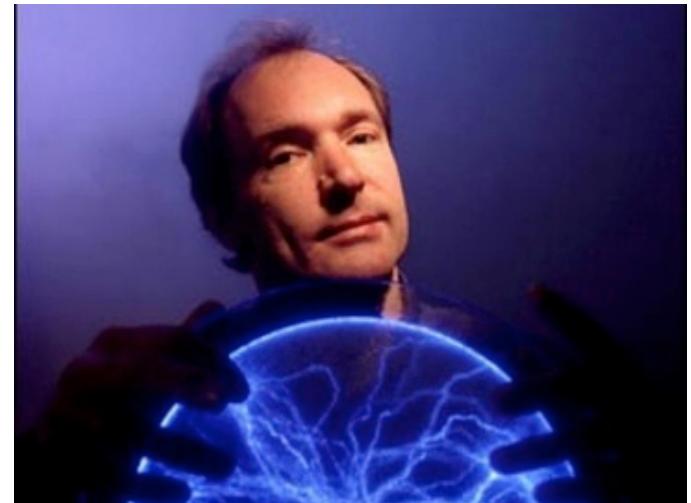


The Open Data Institute

Formats, Structures and Files

Data formats are complex and have ~~suffered~~ benefited from years of development in many different domains.

One thing is bringing them all together.



Recap

Specialist Solution

- + Easy to get setup and maintain.
- + Open Data focused
- + Clear workflows for publishing open data
- + Visualisation tools
- + Data mashing tools
- + Best for transactional data

Integrated Solution

- + No new platform to learn
- + Data is provided in parallel to web pages
- + No separation from authoritative data
- + Easy discovery of data
- + Best for reference data
- + Best for Linked Open Data

Both great for open data

Integrated solutions more suited for building a web of linked data



Exercise

Each table will have one of the publication platforms to explore.

Your task is to publish a prescribed dataset on this platform, with an associated open data certificate.



Outcomes

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

Evaluate a number of different approaches for publishing open data.

Publish a dataset on one of the many available platforms.



Recap session



Today

The characteristics of open data

Open data discovery patterns

Open data publication

Quick big data break

Practical publication hands-on



The characteristics of open data

Identify a number of different characteristics of data

Explain the justifications for publishing different types of data

Evaluate the current open data ecosystem and future opportunities



Open data discovery patterns

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.”



Open data publication

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

Evaluate a number of different approaches for publishing open data.

Publish a dataset on one of the many available platforms.





Open Data in Practice

Dr David Tarrant

@davetaz

The Open Data Institute



Day 1 review:

What are the steps in publishing data as open data?

5-6 steps from nothing to perfection



Step 1

Is open for me?

Do I have personal or sensitive data?



Step 2

Define objective

Why are we doing this? Who are the stakeholders?



Step 3

Define boundaries

What gets published?

What license?

Where are we publishing?



Step 4

Choose methodology

Tools?

Platforms?

Processes?



Step 5

Publish!



Step 6

Showcase / Iterate /

Support

See Day 3



Publication phases

Phase 1 (Steps 1-5): Get the data online, in some form. This will help with the trust and transparency and community building.

Phase 2 (Steps 1-5 + 6): Increase the usability of the data by potentially publishing differently and keeping it up to date.

