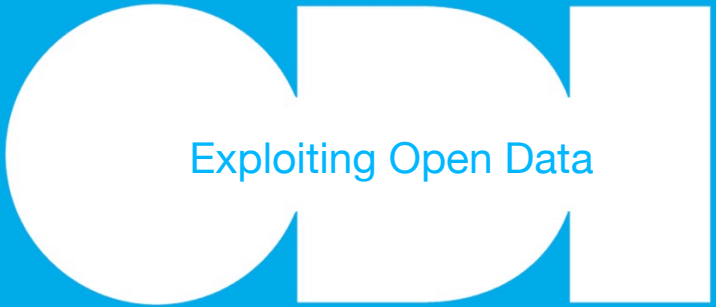


David Tarrant
@davetaz

Toolkit for today

1. Google Chrome Web Browser
   chrome.google.com
2. Open/Google Refine
   openrefine.org
3. A Google Account

4. The Postman extension to Google Chrome (pm only)

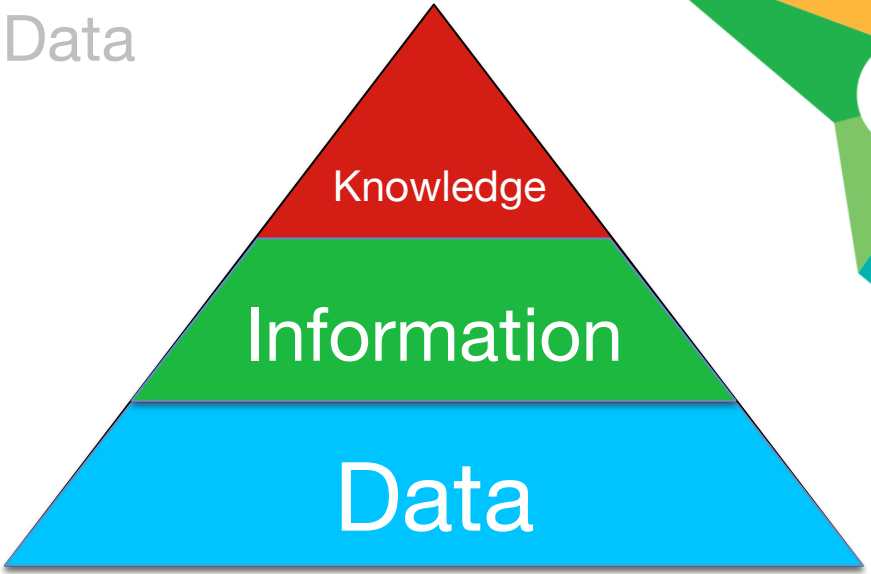
Aims (am)

- Discovering, filtering, combining and analyzing data
- Reconciliation, enriching and linking data
- Small graphs from “Big Data”

Recap

What is Open Data?

Data

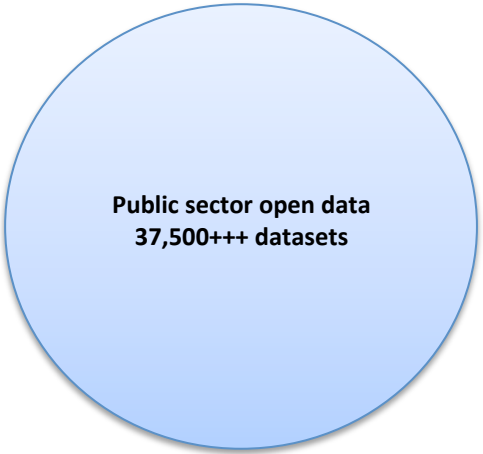


Definition of Open (OKF)

A piece of data or content is open if **anyone** is **free to use, reuse, and redistribute** it — subject only, at most, to the requirement to attribute and/or share-alike.

Trends in publishing open data

Suppliers examples



Private sector open datasets

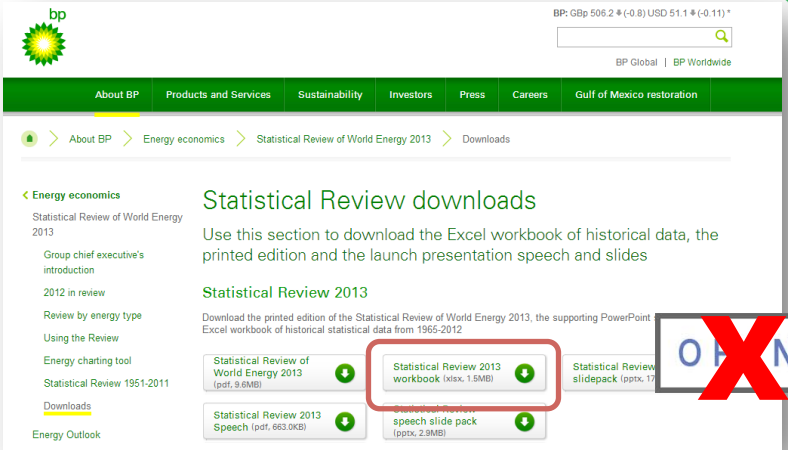
"While there is no central figure on the number of public sector information datasets currently being made available, a review of selected data portals suggests the number could exceed 37,500 from over 750 different publishers with over 2.5 million downloads by October 2012" Deloitte 2012

Note

Open data examples:

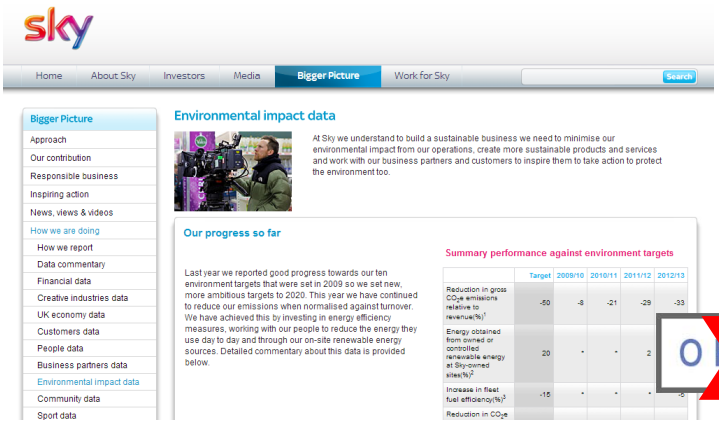
Non open data examples:

Suppliers



BP: You may not frame this site nor link to a page other than the home page without our express permission.
To find this Google "**bp statistical review**"

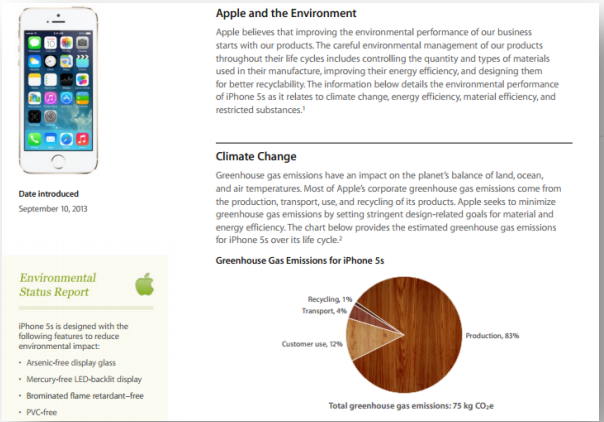
Suppliers (ish)



http://corporate.sky.com/the_bigger_picture/how_we_are_doing/environmental_impact_data

Suppliers (ish)

Product Environment Reports



Apple and the Environment

Apple believes that improving the environmental performance of our business starts with our products. The careful environmental management of our products throughout their life cycles includes controlling the quantity and types of materials used in their manufacture, improving their energy efficiency, and designing them for better recyclability. The information below details the environmental performance of iPhone 5s as it relates to climate change, energy efficiency, material efficiency, and restricted substances.[1]

Climate Change

Greenhouse gas emissions have an impact on the planet's balance of land, ocean, and air temperatures. Most of Apple's corporate greenhouse gas emissions come from the production, transport, use, and recycling of its products. Apple seeks to minimize greenhouse gas emissions by setting stringent design-related goals for material and energy efficiency. The chart below provides the estimated greenhouse gas emissions for iPhone 5s over its life cycle.[2]

Date introduced

September 10, 2013

Environmental Status Report

iPhone 5s is designed with the following features to reduce environmental impact:

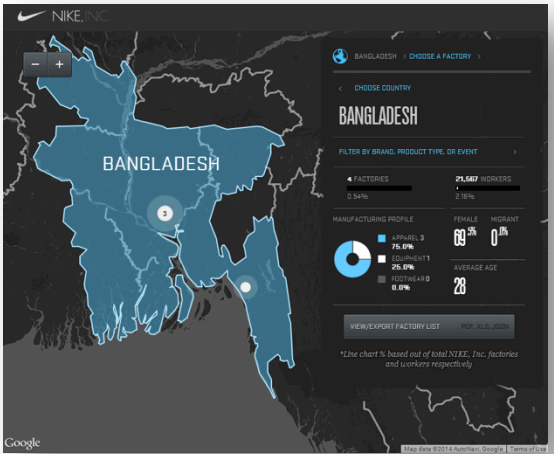
- Arsenic-free display glass
- Mercury-free LED-backlit display
- Brominated flame retardant–free
- PVC-free

Greenhouse Gas Emissions for iPhone 5s

Total greenhouse gas emissions: 75 kg CO_2e

http://www.apple.com/uk/environment/reports/

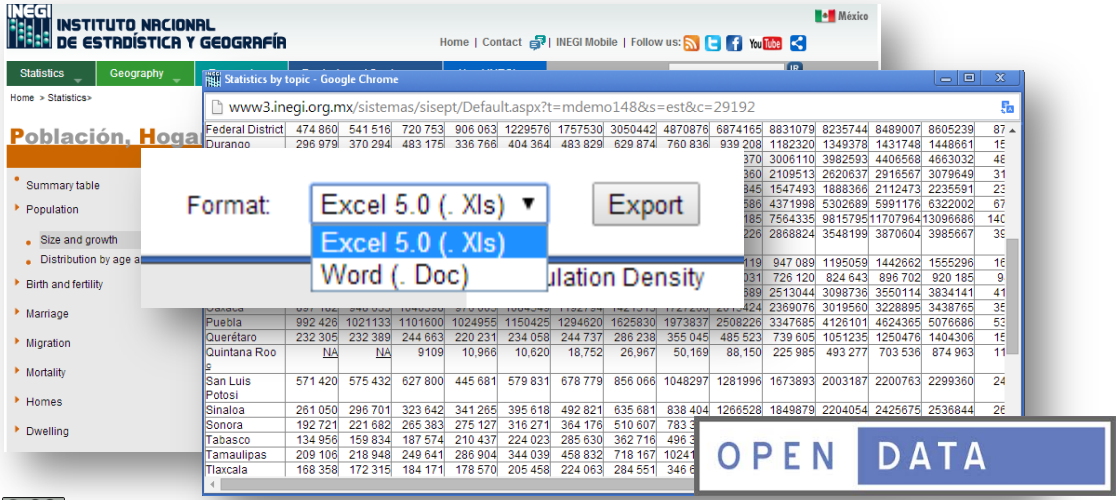
Suppliers



OPEN DATA

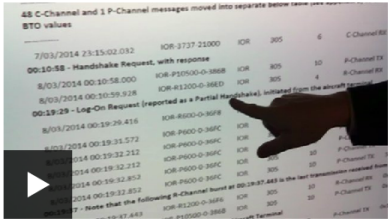
http://manufacturingmap.nikeinc.com/#

Government


INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA
México
Home | Contact | INEGI Mobile | Follow us:
Statistics
Geography
Home > Statistics>
Población, Hoga
Summary table
Population
Size and growth
Distribution by age a
Birth and fertility
Marriage
Migration
Mortality
Homes
Dwelling
Statistics by topic - Google Chrome
www3.inegi.org.mx/sistemas/sisept/Default.aspx?t=mdemo148&s=est&c=29192
Format:
Excel 5.0 (. Xls)
Export
Excel 5.0 (. Xls)
Word (. Doc)
ulation Density
Federal District 474 860 541 516 720 753 906 063 1229576 1757530 3050442 4870876 6874165 8831079 8235744 8489007 8605239
Durango 296 979 370 294 483 175 336 766 404 364 483 829 629 874 760 836 939 208 1182320 1349378 1431748 1448661
Puebla 992 426 1021133 1101600 1024955 1150425 1294620 1625830 1973837 2508226 3347685 4126101 4624365 5076686
Querétaro 232 305 232 389 244 663 220 231 234 058 244 737 286 238 355 045 485 523 739 605 1051235 1250476 1404306
Quintana Roo NA NA 9109 10,966 10,620 18,752 26,967 50,169 88,150 225 985 493 277 703 536 874 963
San Luis Potosí 571 420 575 432 627 800 445 681 579 831 678 779 856 066 1048297 1281996 1673893 2003187 2200763 2299360
Sinaloa 261 050 296 701 323 642 341 265 395 618 492 821 635 681 838 404 1266528 1849879 2204054 2425675 2536844
Sonora 192 721 221 682 265 383 275 127 316 271 364 176 510 607
Tabasco 134 956 159 834 187 574 210 437 224 023 285 630 362 716
Tamaulipas 209 106 218 948 249 641 286 904 344 039 458 832 718 167
Tlaxcala 168 358 172 315 184 171 178 570 205 458 224 063 284 551
OPEN DATA

Government / Private (ish)

Flight MH370: Malaysia releases raw satellite data

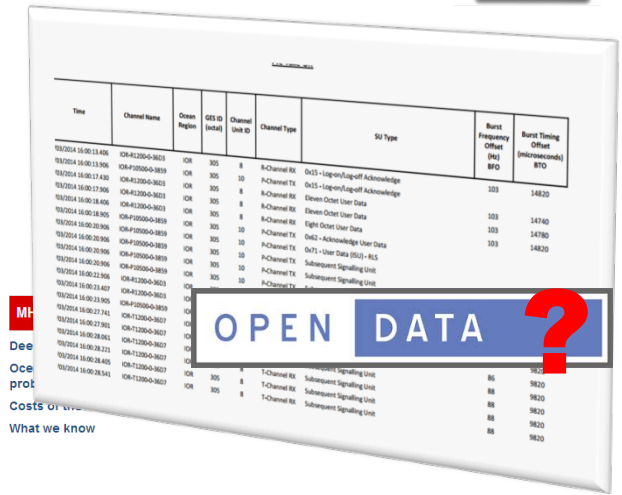


The BBC's Richard Westcott visited Inmarsat's headquarters to find out what the data tells us about MH370's fate

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.

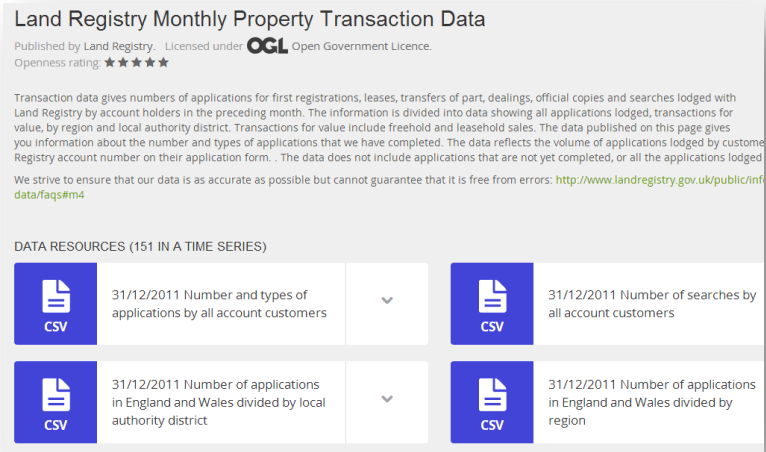
The document released on Tuesday comprises 47 pages of data, plus notes, from British firm Inmarsat.



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		
7/03/2014 16:00:18.406	IOR-R1200-0-3603	IOR	305	8	R-Channel Rx	Eight Octet User Data	103	14740
7/03/2014 16:00:18.905	IOR-P10500-0-3859	IOR	305	10	P-Channel Tx	Eight Octet User Data	103	14780
7/03/2014 16:00:20.906	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 16:00:21.906	IOR-P10500-0-3859	IOR	305	10	P-Channel Tx	Subsequent Signalling Unit		
7/03/2014 16:00:23.407	IOR-R1200-0-3603	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit		
7/03/2014 16:00:23.905	IOR-R1200-0-3603	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit		
7/03/2014 16:00:27.741	IOR-P10500-0-3859	IOR	305	10	P-Channel Tx	Subsequent Signalling Unit		
7/03/2014 16:00:27.901	IOR-T1200-0-3607	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit		
7/03/2014 16:00:28.061	IOR-T1200-0-3607	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit		
7/03/2014 16:00:28.221	IOR-T1200-0-3607	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit		
7/03/2014 16:00:28.405	IOR-T1200-0-3607	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit		
7/03/2014 16:00:28.541	IOR-T1200-0-3607	IOR	305	8	T-Channel Rx	Subsequent Signalling Unit	86	9820
							88	9820
							88	9820
							88	9820
							88	9820
							88	9820

MH
Dee
Oce
prob
Costs of th
What we know

Government



Land Registry Monthly Property Transaction Data

Published by Land Registry. Licensed under OGL Open Government Licence.
Openness rating: ★★★★★

Transaction data gives numbers of applications for first registrations, leases, transfers of part, dealings, official copies and searches lodged with Land Registry by account holders in the preceding month. The information is divided into data showing all applications lodged, transactions for value, by region and local authority district. Transactions for value include freehold and leasehold sales. The data published on this page gives you information about the number and types of applications that we have completed. The data reflects the volume of applications lodged by custome Registry account number on their application form. . The data does not include applications that are not yet completed, or all the applications lodged

We strive to ensure that our data is as accurate as possible but cannot guarantee that it is free from errors: http://www.landregistry.gov.uk/public/inf data/faqs#m4

DATA RESOURCES (151 IN A TIME SERIES)

CSV 31/12/2011 Number and types of applications by all account customers

CSV 31/12/2011 Number of searches by all account customers

CSV 31/12/2011 Number of applications in England and Wales divided by local authority district

CSV 31/12/2011 Number of applications in England and Wales divided by region

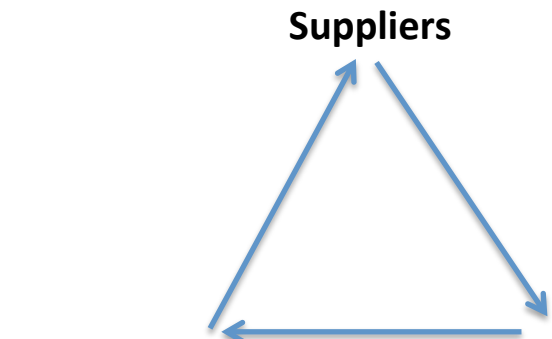
Exercise

Analyse a number of different suppliers of data and critique the different methods through which the data is available. What are the benefits and drawbacks of each?

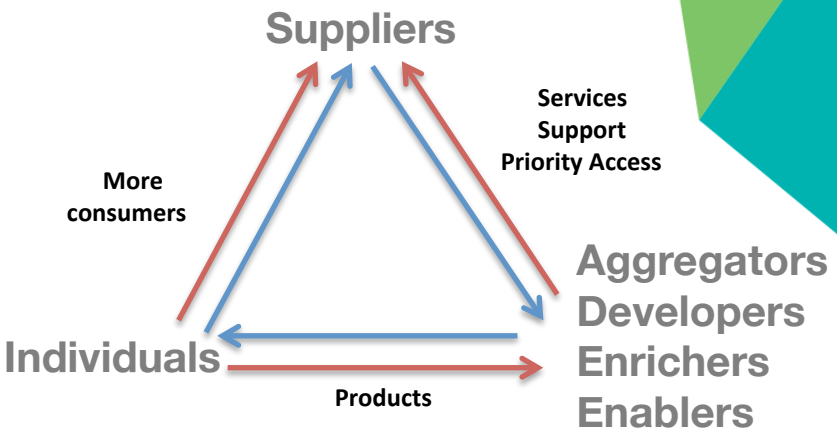
Current companies

- Suppliers
- Aggregators
- Developers
- Enrichers
- Enablers

Value chain



Money go round?



Easy win: Publishing ESG (environmental, social, governance) data

More companies are deciding that the best strategy is to operate sustainably and release open data that shows it

- Sustainable practices as a sign of good corporate governance & predictor of long-term profitability
- Reduction in investment risk and helps attract new investment
- Good for branding and recruiting
- Can improve operations
- Be a step ahead of Government regulation in this area

Attention to environmental, social and governance factors is good for business

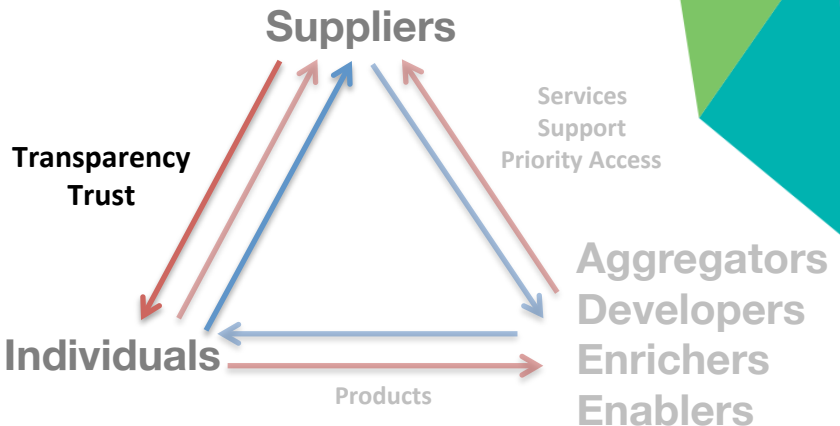
Where can businesses start using data and publishing it openly?

Increase transparency / trust / reputation

Operate more efficiently and make more informed decisions

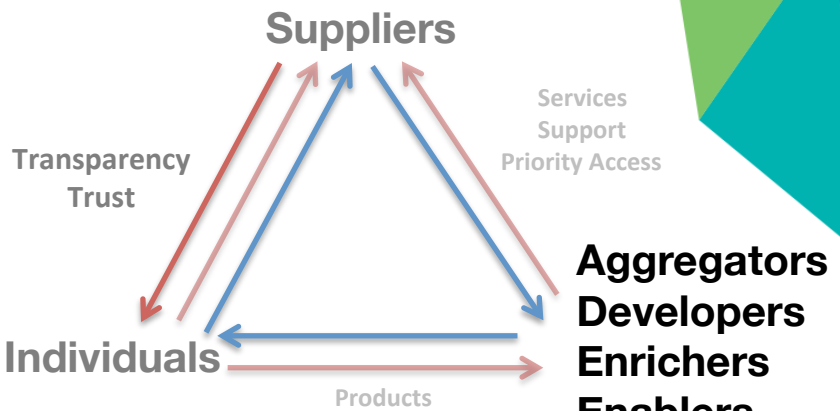
Innovate to deliver new products, services and ways of working

Money go round?

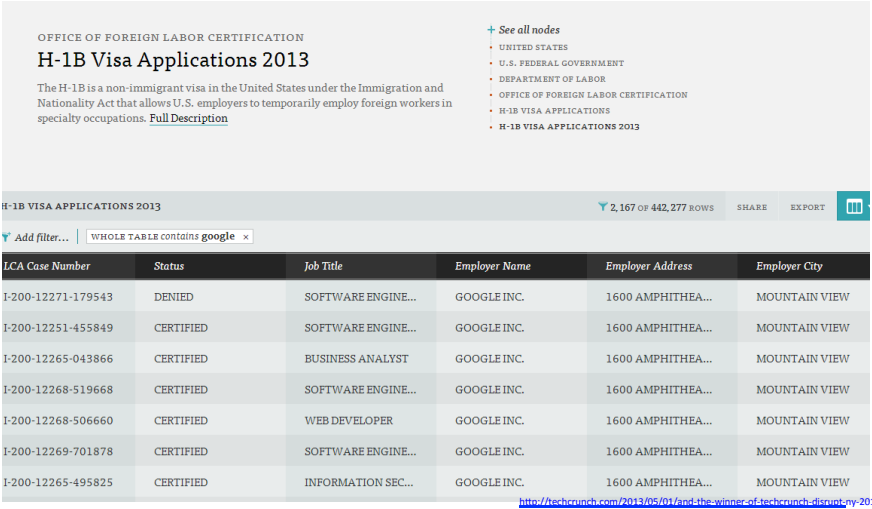


Data is the raw material of the new industrial revolution

Money go round?



enigma.io Aggregator/Enabler



OFFICE OF FOREIGN LABOR CERTIFICATION

H-1B Visa Applications 2013

The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations. Full Description

+ *See all nodes*
- UNITED STATES
- U.S. FEDERAL GOVERNMENT
- DEPARTMENT OF LABOR
- OFFICE OF FOREIGN LABOR CERTIFICATION
- H-1B VISA APPLICATIONS
- **H-1B VISA APPLICATIONS 2013**

H-1B VISA APPLICATIONS 2013 — 2,167 OF 442,277 ROWS — SHARE — EXPORT

Add filter… | WHOLE TABLE *contains* google ×

LCA Case Number	Status	Job Title	Employer Name	Employer Address	Employer City
I-200-12271-179543	DENIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12251-455849	CERTIFIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12265-043866	CERTIFIED	BUSINESS ANALYST	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12268-519668	CERTIFIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12268-506660	CERTIFIED	WEB DEVELOPER	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12269-701878	CERTIFIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12265-495825	CERTIFIED	INFORMATION SEC...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW

http://techcrunch.com/2013/05/01/and-the-winner-of-techcrunch-disrupt-ny-2013-is-enigma/

Exercise

- Enigma.io
- Selecting, filtering and cleaning data
- Using pivot tables and open refine.

Application Programming Interfaces (APIs)

Exercise

What is an API?

Any examples?

What is an API

Defines how one application can **consistently** interact with another.

Teaching nerds passion

```
function makeOut(passionLevel, partsOfBody) {
 for (each partOfBody in partsOfBody) {
   partOfBody.kiss(passionLevel);
   lookIntoEyes();
   sighDeeply();
 }
 moanDaintily();
 sleep();
}
```


Thanks to Pamela Fox (blog.pamelafox.org)

Making your call

```
function makeOut(passionLevel, partsOfBody) {
 for (each partOfBody in partsOfBody) {
   partOfBody.kiss(passionLevel);
   lookIntoEyes();
   s
 }
 moanDaintily();
 sleep();
}
```

```
makeOut(10, ["neck", "ear", "mouth"]);
```


Thanks to Pamela Fox (blog.pamelafox.org)

The Webs API

```
class Photos {
    function search(api_key, tags) {
        if (!validKey(api_key)) {return(401);}
        ...
        ...
        return results;
    }
}
```

Using the Webs API

http://api.flickr.com/services/rest/?method=flickr.photos.search&api_key=a8c42ef2ac8d88aed7e351f95e93160f&tags=fox

Evolution of Web APIs

- Long URLs are not easy to understand
- Flickr API using query parameters to define everything!

```
?method=…&auth=…&query=…
```

- There are better ways

Web APIs (before)

http://api.flickr.com/services/rest/?method=flickr.photos.comments.getList&api_key=a8c42ef2ac8d88aed7e351f95e93160f&photo_id=13992401185

A better solution?

```
https://www.flickr.com/photos/{user}/{photo}/comments
                                                /tags
                                                /licence

https://www.flickr.com/search?tags=…
```

API Recap

A **promise** by one application to service another.

The underlying code can change separately to the API…

So far we have looked at READ only

Aggregator/Enricher

- Take data from hundreds of transport companies
- Bring it together
- Align it under one easy to use API

opencorporates

The Open Database Of The Corporate World

Aggregator/Enabler

We have information on 70,597,888 companies

SEARCH

search companies search officers

Filter by jurisdiction

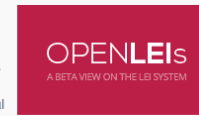
1,298	Abu Dhabi (UAE)
144,755	Alaska (US)
40,157	Albania
899,455	Arizona (US)
46,537	Aruba
165,582	Bahamas
99,185	Bahrain
88,563	Bangladesh

Just released: OpenCorporates API v0.3

Corporate network data, financial accounts, complex filters, and more. Read more

Announcing Open LEIs

Today, OpenCorporates announces a new sister website, Open LEIs, a user-friendly interface on the emerging Global Legal Entity Identifier System. Read more



New! Just added: Open corporate network data

Read more about this important new feature

Exercise

Enriching data

Taking the FCO dataset and using open corporates in Refine

Discussion

What's the biggest problem with APIs like flickr and transport API?

Common data retrieval methods

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.

Getting hold of data

1. Add random extensions (.xml, .json, .csv, .rdf etc)
2. Look for alternative links
3. Look for embedded data
4. Do some content negotiation
5. Spot the API
6. Abandon hope…(search google again)

Getting hold of data

1. Add random extensions (.xml, .json, .csv, .rdf etc)
2. Look for alternative links
3. Look for embedded data
4. Do some content negotiation
5. Spot the API
6. Abandon hope…(search google again)

1. Add random extensions

(.xml, .json, .csv, .rdf etc)

Try out a number of the links listed on the website to see if you can find data.

- When did you?
- What formats?
- Machine readable license?
- Is the URL an identifier?

2. Look for link alternates

These are in the source of the HTML page

```
<link rel="alternate" type="text/csv" href="http…"
title="Alternate for …"/>
```

Find these on dbpedia, CKAN, gov.uk, eprints.

3. Look for embedded data

Also in the source of the HTML page

Harder to spot as there are many different variants. Watch out for a number of common namespace declarations in the <html> tag:

```
<html xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:rdfs="http://
www.w3.org/2000/01/rdf-schema#" xmlns:dc="http://purl.org/dc/
terms/" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#" xmlns:dcat="http://www.w3.org/ns/dcat#" </html>
```

4. Do some content negotiation

This is the **correct answer** and we will come back to it later when we talk about the web.

Basically Tim had the answer in 1994 and we all forgot about it and invented a load of other crap instead!

Recap

- The supply of open data
- Tools for aggregating, processing and enriching data.
- Advanced data discovery
- APIs, extensions, embedded data and links.

One more thing

The promise of “Big” data.

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Exercise

David Tarrant
@davetaz