Open Data Where We Are Where We're Going

International Open Government Conference, Washington DC, July 2012

Rufus Pollock

@okfn[.org]

@rufuspollock[.org]

Open Knowledge Foundation

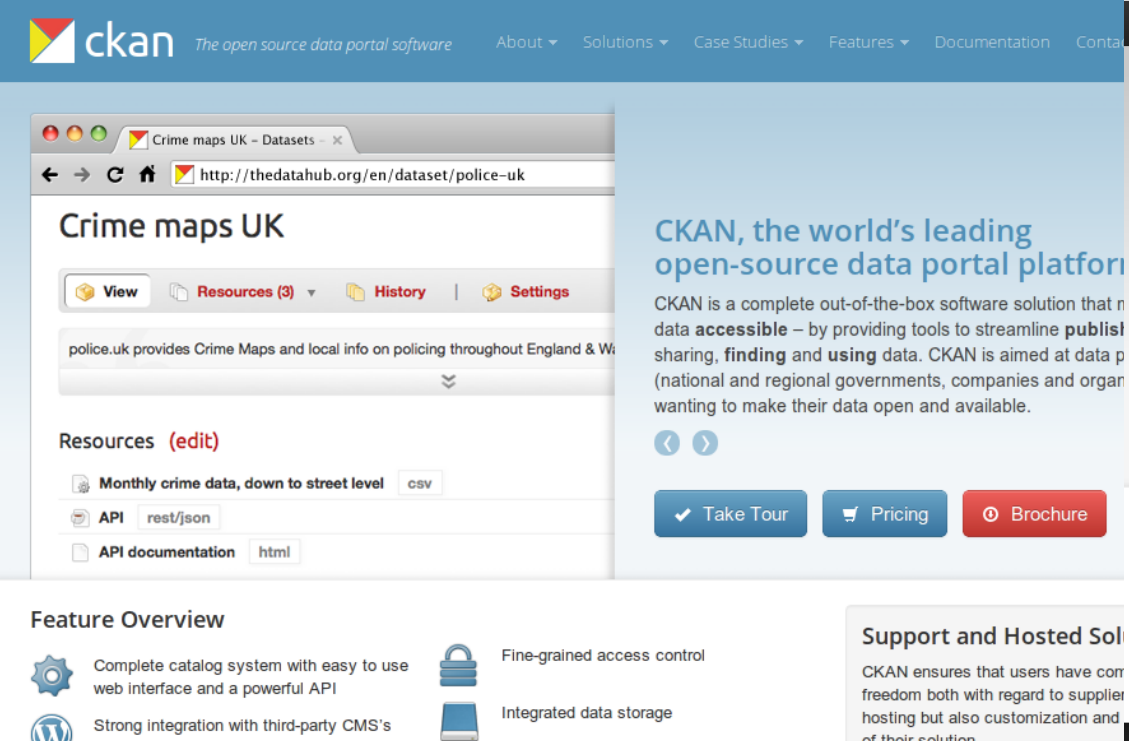
SHUTTLEWORTH FELLOW

Open Knowledge
Foundation

We build **technology** and **communities** to **develop, disseminate** and **use open knowledge** – **content** and **data** that everyone can use, share and build on.

OPEN DATA

OPEN CONTENT


ckan
The open source data portal software
About ▾
Solutions ▾
Case Studies ▾
Features ▾
Documentation
Conta
Crime maps UK – Datasets ×
http://thedatahub.org/en/dataset/police-uk

Crime maps UK

View
Resources (3) ▾
History
Settings

police.uk provides Crime Maps and local info on policing throughout England & Wa

Resources (edit)

Monthly crime data, down to street level	csv
API	rest/json
API documentation	html

CKAN, the world's leading open-source data portal platfor

CKAN is a complete out-of-the-box software solution that m data **accessible** – by providing tools to streamline **publish** sharing, **finding** and **using** data. CKAN is aimed at data p (national and regional governments, companies and organ wanting to make their data open and available.

Feature Overview

Complete catalog system with easy to use web interface and a powerful API

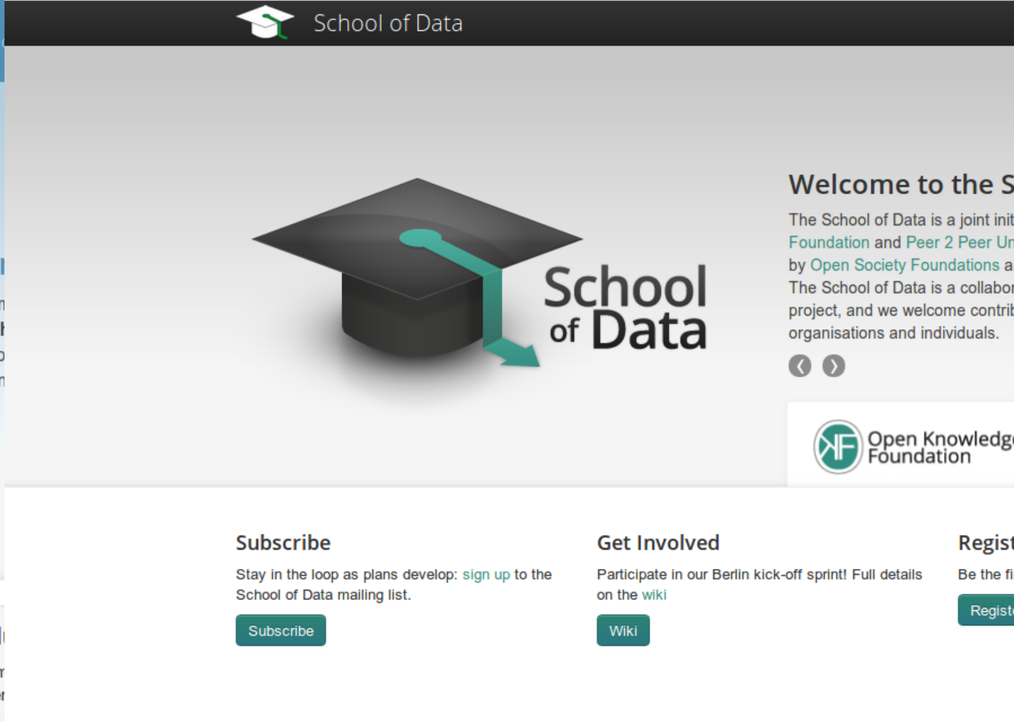
Strong integration with third-party CMS's

Fine-grained access control

Integrated data storage

Support and Hosted Sol

CKAN ensures that users have com freedom both with regard to supplier hosting but also customization and of their solution



Welcome to the S

The School of Data is a joint init Foundation and Peer 2 Peer Un by Open Society Foundations a The School of Data is a collabor project, and we welcome contrib organisations and individuals.

Subscribe

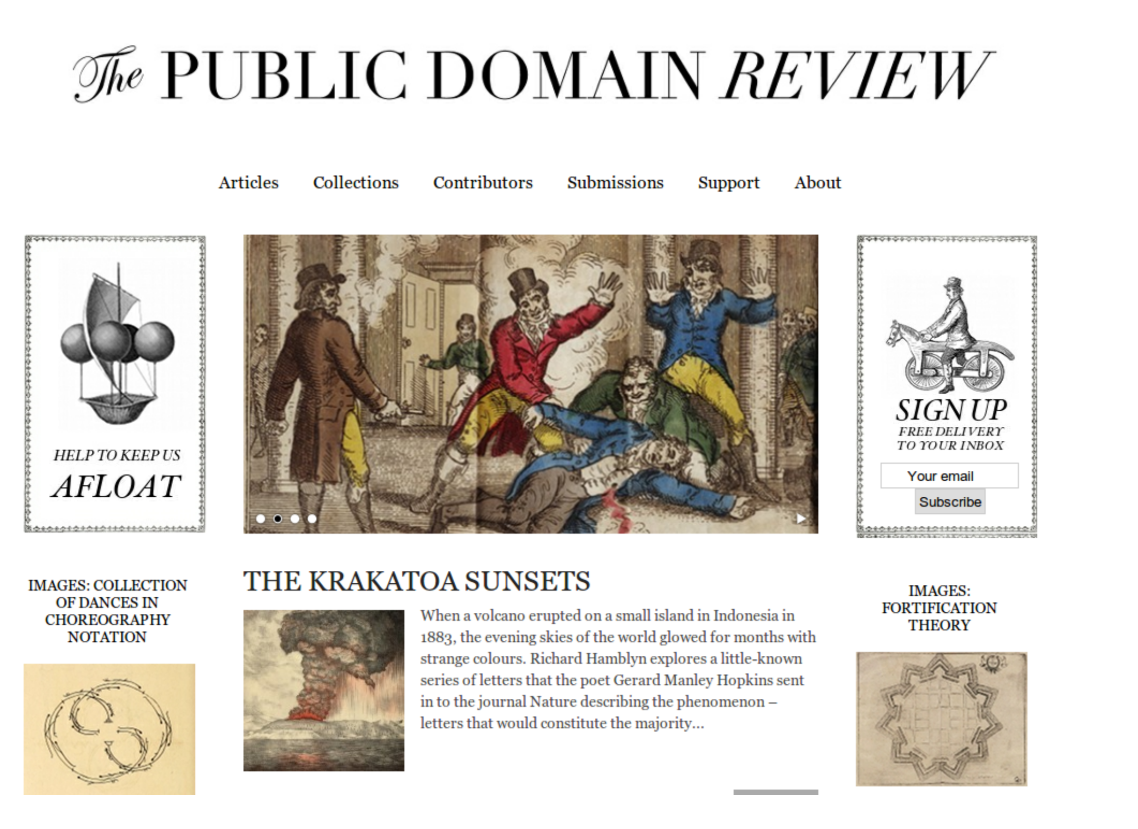
Stay in the loop as plans develop: sign up to the School of Data mailing list.

Get Involved

Participate in our Berlin kick-off sprint! Full details on the wiki

Regis

Be the fi

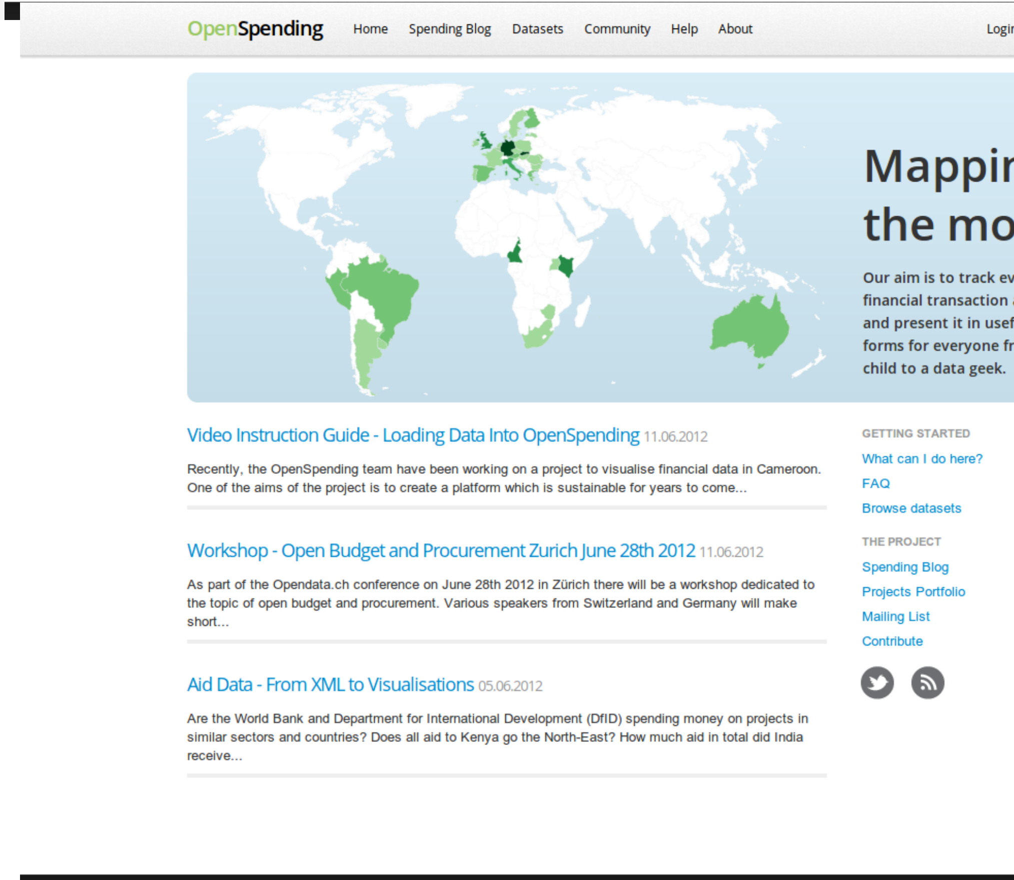


IMAGES: COLLECTION OF DANCES IN CHOREOGRAPHY NOTATION

THE KRAKATOA SUNSETS

When a volcano erupted on a small island in Indonesia in 1883, the evening skies of the world glowed for months with strange colours. Richard Hamblyn explores a little-known series of letters that the poet Gerard Manley Hopkins sent in to the journal Nature describing the phenomenon – letters that would constitute the majority...

IMAGES: FORTIFICATION THEORY



Mappin the mo

Our aim is to track ev financial transaction and present it in usef forms for everyone fr child to a data geek.

Video Instruction Guide - Loading Data Into OpenSpending 11.06.2012

Recently, the OpenSpending team have been working on a project to visualise financial data in Cameroon. One of the aims of the project is to create a platform which is sustainable for years to come...

Workshop - Open Budget and Procurement Zurich June 28th 2012 11.06.2012

As part of the Opendata.ch conference on June 28th 2012 in Zürich there will be a workshop dedicated to the topic of open budget and procurement. Various speakers from Switzerland and Germany will make short...

Aid Data - From XML to Visualisations 05.06.2012

Are the World Bank and Department for International Development (DfID) spending money on projects in similar sectors and countries? Does all aid to Kenya go the North-East? How much aid in total did India receive...

GETTING STARTED
- What can I do here?
- FAQ
- Browse datasets

THE PROJECT
- Spending Blog
- Projects Portfolio
- Mailing List
- Contribute

OKFestival = OGDcamp + OKCon. Helsinki, Finland. 17-22 Sept 2012.

We are delighted to invite you to the world's first Open Knowledge Festival: a week of participatory sessions, keynote lectures, workshops, hackathons and satellite events in Helsinki, organised by diverse communities from across the globe.

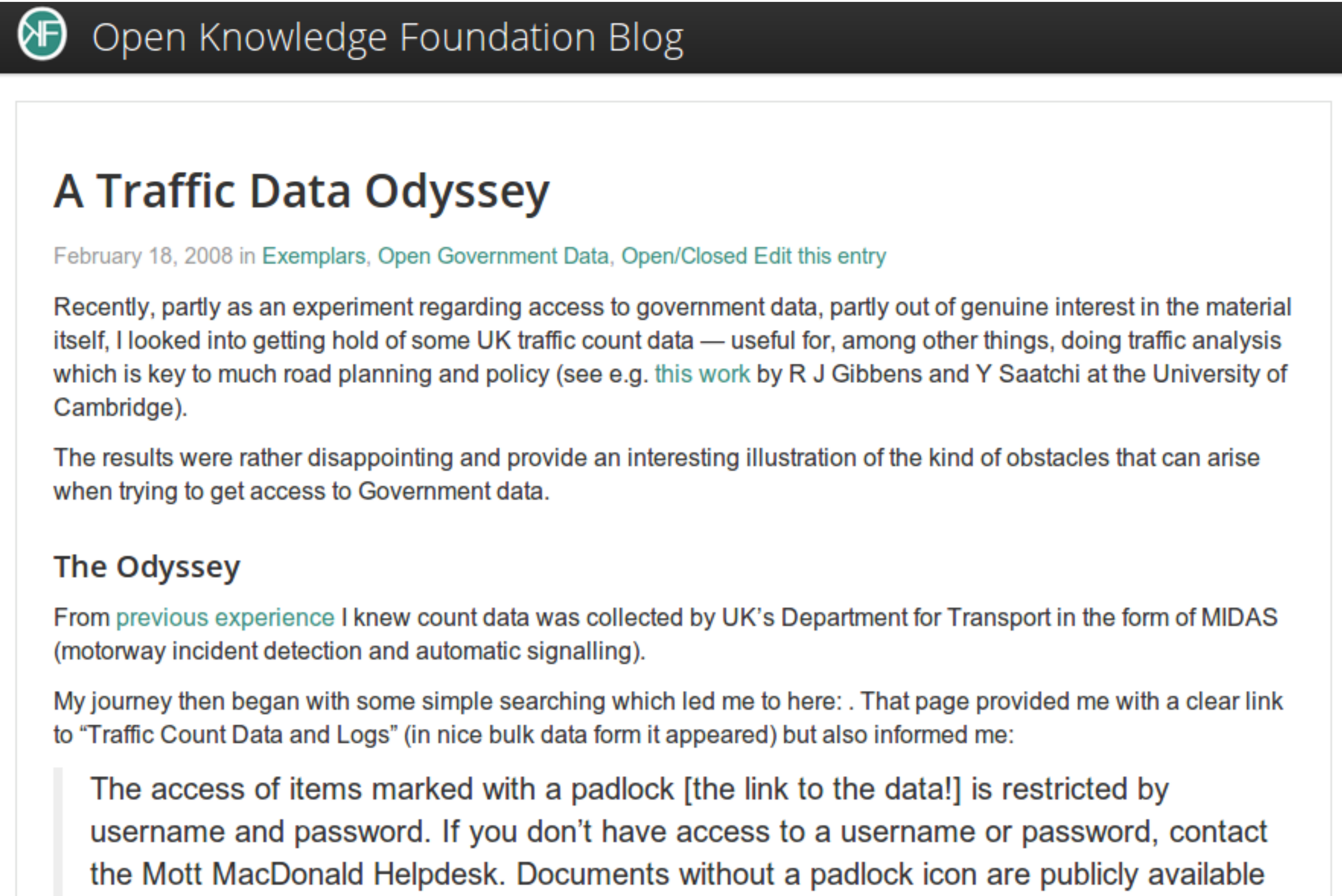
The 2012 theme of OKFestival is *Open Knowledge in Action*, looking at the *value* that can be generated by opening up knowledge, the ecosystems of organisations that can benefit from such sharing, and the impacts that transparency can have in our societies. What kinds of new professions, ideas and community initiatives can emerge within our governments, markets, networks and neighbourhoods as a result of these engagements?

The exploration of this theme will not only be visible in the festival's content, but also in its implementation as the first global event of its kind. This year, OKFestival will combine two popular annual events – the ***Open Government Data Camp*** and the ***Open Knowledge Conference.*** This combination allows us to highlight a set of 13 diverse Topic Streams from open development to municipal data, all organised by global teams of Guest Programme Planners. With this collaborative format, we aim to highlight the diversity of open knowledge and data initiatives from around the world. We will bring together civil society representatives, programmers, data wranglers, designers, students, members of government, local communities and citizens for a week of building new things and sharing great ideas.

{Context}

Access and Reuse
A Traffic Data Odyssey

http://blog.okfn.org/2008/02/18/a-traffic-data-odyssey/



Open Knowledge Foundation Blog

A Traffic Data Odyssey

February 18, 2008 in Exemplars, Open Government Data, Open/Closed Edit this entry

Recently, partly as an experiment regarding access to government data, partly out of genuine interest in the material itself, I looked into getting hold of some UK traffic count data — useful for, among other things, doing traffic analysis which is key to much road planning and policy (see e.g. this work by R J Gibbens and Y Saatchi at the University of Cambridge).

The results were rather disappointing and provide an interesting illustration of the kind of obstacles that can arise when trying to get access to Government data.

The Odyssey

From previous experience I knew count data was collected by UK's Department for Transport in the form of MIDAS (motorway incident detection and automatic signalling).

My journey then began with some simple searching which led me to here: . That page provided me with a clear link to "Traffic Count Data and Logs" (in nice bulk data form it appeared) but also informed me:

> The access of items marked with a padlock [the link to the data!] is restricted by username and password. If you don't have access to a username or password, contact the Mott MacDonald Helpdesk. Documents without a padlock icon are publicly available

The Request (Nov 2007)

Request for count data collected by UK’s Department for Transport in the form of MIDAS (motorway incident detection and automatic signalling):

I’m a UK citizen interested in getting access to the Traffic Count Data and Logs dataset linked to from: http://www.midas-data.org.uk/

It appears that a username and password is required from yourselves in order to do this and so I wondered if you could therefore be kind enough to provide me with such a username and password.

The Refusal (Jan 2008)

6 emails later: told these conditions required by Dept for Transport ...

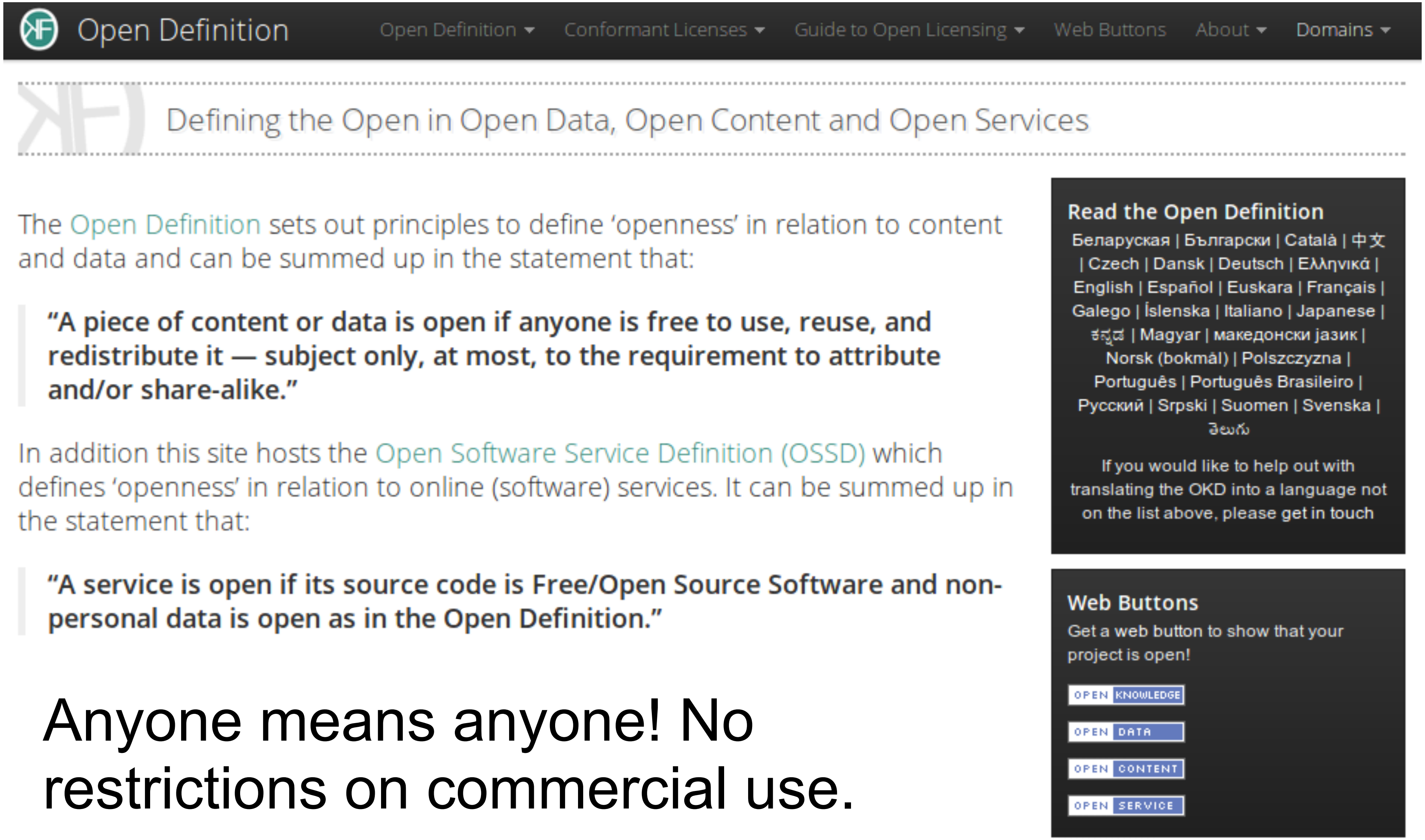
I need your **acceptance of the conditions stated below and some information regarding the research project you are undertaking before we allow you access to the data**. The conditions and information I have requested will allow the Group to **justify the costs associated with supplying this data** [what costs, it's already in a bzip file on a website?] and to **ensure the data is being used appropriately** [why is such paternalism needed?].

Note:- if the project is being undertaken jointly with **another organisation** then that organisation will **also be required to supply the information requested**. Please ensure **all grant and contract holders, staff and students** associated with the grant and project are **made aware of the conditions** contained within this letter.

Conditions

1. The data may not be copied to any other persons or organisations without the prior approval of the Highways Agency. The data may only be copied to another person or organisation after that person or organisation has confirmed with the HA the purpose for which the data is required and accepted the conditions laid down in this letter.
2. The data may not be used for any other purpose within your organisation without the prior written approval of the Highways agency.
3. **The data must not be sold or used for commercial gain.**
4. **The data will not be used to contradict or challenge any research project, works or statement made by the Government, the Department of Transport or the The Highways Agency as a result of analysis of the data by them or their agents.**
5. the Highways Agency will be provided, upon publication and free of charge, with: annual progress reports; any interim reports describing significant findings; a complete copy of the final report; and any technical papers resulting from the research.

OPEN DATA



Defining the Open in Open Data, Open Content and Open Services

The Open Definition sets out principles to define 'openness' in relation to content and data and can be summed up in the statement that:

> **"A piece of content or data 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."**

In addition this site hosts the Open Software Service Definition (OSSD) which defines 'openness' in relation to online (software) services. It can be summed up in the statement that:

> **"A service is open if its source code is Free/Open Source Software and non-personal data is open as in the Open Definition."**

Read the Open Definition

Беларуская | Български | Català | 中文 | Czech | Dansk | Deutsch | Ελληνικά | English | Español | Euskara | Français | Galego | Íslenska | Italiano | Japanese | ಕನ್ನಡ | Magyar | македонски јазик | Norsk (bokmål) | Polszczyzna | Português | Português Brasileiro | Русский | Srpski | Suomen | Svenska | తెలుగు

If you would like to help out with translating the OKD into a language not on the list above, please get in touch

Web Buttons

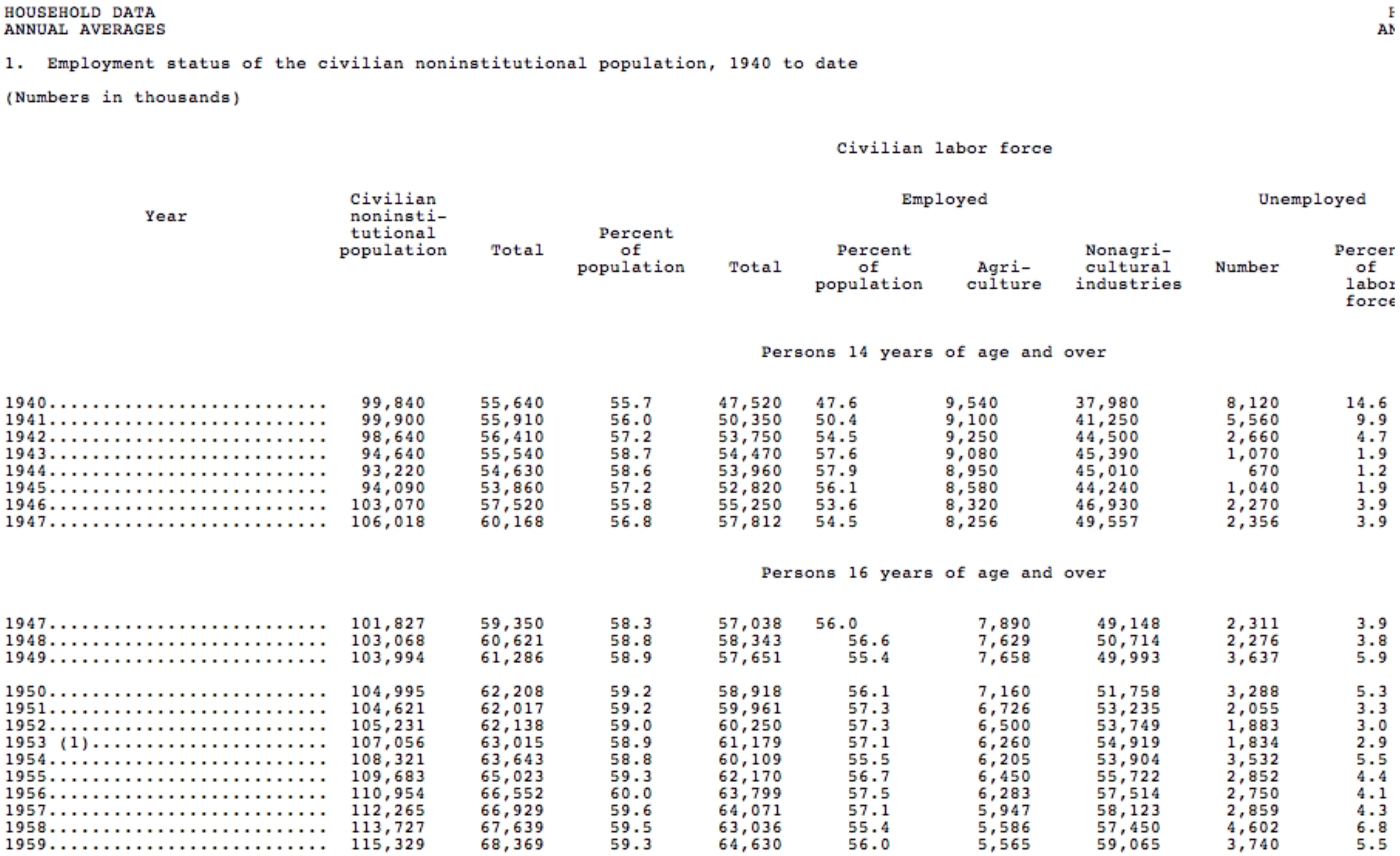
Get a web button to show that your project is open!

Anyone means anyone! No restrictions on commercial use.

Open != Creative Commons. Many CC licenses NOT open (and most not appropriate for data).

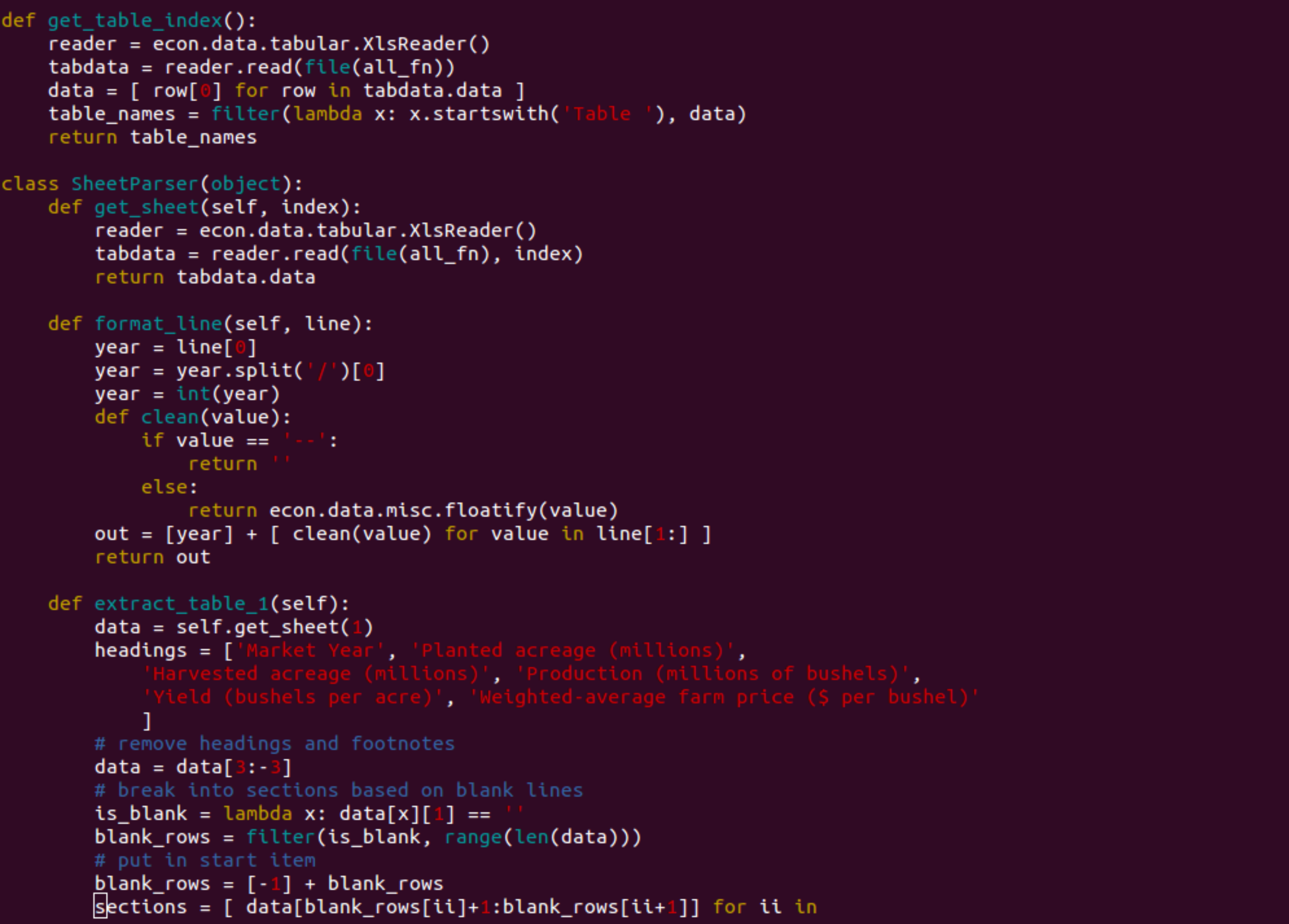
Machine Access

US Unemployment Stats

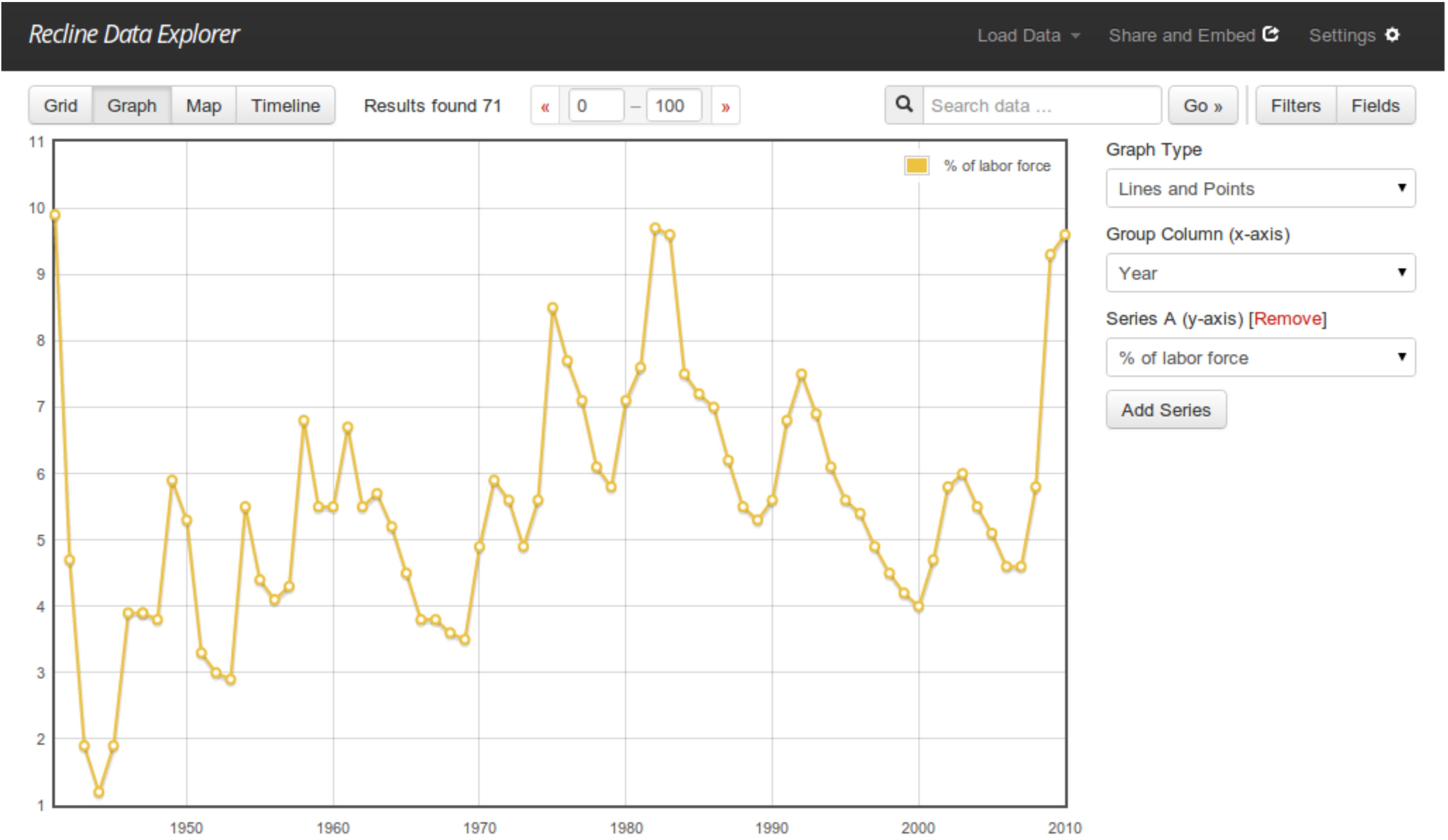


```
HOUSEHOLD DATA                                                                                            F
ANNUAL AVERAGES                                                                                           AN

1.  Employment status of the civilian noninstitutional population, 1940 to date

(Numbers in thousands)

                                                                Civilian labor force

                                                                             Employed                     Unemployed
            Year                 Civilian
                                 noninsti-
                                 tutional                Percent
                                 population     Total      of                Percent               Nonagri-               Percer
                                                        population   Total      of       Agri-     cultural     Number     of
                                                                             population  culture   industries             labor
                                                                                                                          force

                                                          Persons 14 years of age and over


1940..........................   99,840      55,640     55.7      47,520   47.6       9,540     37,980       8,120      14.6
1941..........................   99,900      55,910     56.0      50,350   50.4       9,100     41,250       5,560       9.9
1942..........................   98,640      56,410     57.2      53,750   54.5       9,250     44,500       2,660       4.7
1943..........................   94,640      55,540     58.7      54,470   57.6       9,080     45,390       1,070       1.9
1944..........................   93,220      54,630     58.6      53,960   57.9       8,950     45,010         670       1.2
1945..........................   94,090      53,860     57.2      52,820   56.1       8,580     44,240       1,040       1.9
1946..........................  103,070      57,520     55.8      55,250   53.6       8,320     46,930       2,270       3.9
1947..........................  106,018      60,168     56.8      57,812   54.5       8,256     49,557       2,356       3.9

                                                          Persons 16 years of age and over


1947..........................  101,827      59,350     58.3      57,038   56.0          7,890     49,148     2,311       3.9
1948..........................  103,068      60,621     58.8      58,343      56.6       7,629     50,714     2,276       3.8
1949..........................  103,994      61,286     58.9      57,651      55.4       7,658     49,993     3,637       5.9

1950..........................  104,995      62,208     59.2      58,918      56.1       7,160     51,758     3,288       5.3
1951..........................  104,621      62,017     59.2      59,961      57.3       6,726     53,235     2,055       3.3
1952..........................  105,231      62,138     59.0      60,250      57.3       6,500     53,749     1,883       3.0
1953 (1)......................  107,056      63,015     58.9      61,179      57.1       6,260     54,919     1,834       2.9
1954..........................  108,321      63,643     58.8      60,109      55.5       6,205     53,904     3,532       5.5
1955..........................  109,683      65,023     59.3      62,170      56.7       6,450     55,722     2,852       4.4
1956..........................  110,954      66,552     60.0      63,799      57.5       6,283     57,514     2,750       4.1
1957..........................  112,265      66,929     59.6      64,071      57.1       5,947     58,123     2,859       4.3
1958..........................  113,727      67,639     59.5      63,036      55.4       5,586     57,450     4,602       6.8
1959..........................  115,329      68,369     59.3      64,630      56.0       5,565     59,065     3,740       5.5
```

Human but not machine readable ASCII!
Note lovingly word-wrapped columns in plain text



```
def get_table_index():
    reader = econ.data.tabular.XlsReader()
    tabdata = reader.read(file(all_fn))
    data = [ row[0] for row in tabdata.data ]
    table_names = filter(lambda x: x.startswith('Table '), data)
    return table_names

class SheetParser(object):
    def get_sheet(self, index):
        reader = econ.data.tabular.XlsReader()
        tabdata = reader.read(file(all_fn), index)
        return tabdata.data

    def format_line(self, line):
        year = line[0]
        year = year.split('/')[0]
        year = int(year)
        def clean(value):
            if value == '--':
                return ''
            else:
                return econ.data.misc.floatify(value)
        out = [year] + [ clean(value) for value in line[1:] ]
        return out

    def extract_table_1(self):
        data = self.get_sheet(1)
        headings = ['Market Year', 'Planted acreage (millions)',
            'Harvested acreage (millions)', 'Production (millions of bushels)',
            'Yield (bushels per acre)', 'Weighted-average farm price ($ per bushel)'
            ]
        # remove headings and footnotes
        data = data[3:-3]
        # break into sections based on blank lines
        is_blank = lambda x: data[x][1] == ''
        blank_rows = filter(is_blank, range(len(data)))
        # put in start item
        blank_rows = [-1] + blank_rows
        sections = [ data[blank_rows[ii]+1:blank_rows[ii+1]] for ii in
```



Results found 71

0 – 100

Graph Type

Lines and Points

Group Column (x-axis)

Year

Series A (y-axis) [Remove]

% of labor force

Machine Readable
Bulk Data

OPEN DATA

PDFs are not enough!
APIs are not enough!

{Where We Are}

Challenge and an Opportunity

Challenge: Exploding Information Complexity

In 1820s all UK bank clearing done in a single room in London once a day. Today, billions of transactions a minute.

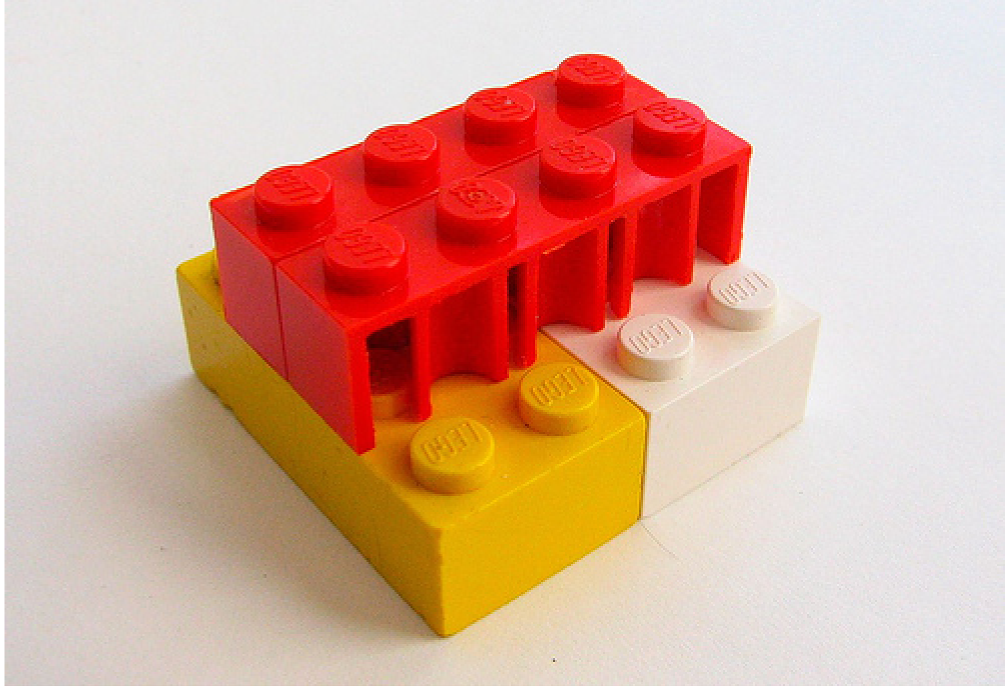
=> componentization to divide and conquer complexity

Opportunity: Info Technology

1TB of storage is around $100, in 1994 this would have cost ~ $400,000. Your smartphone is more powerful than a mainframe 20y ago

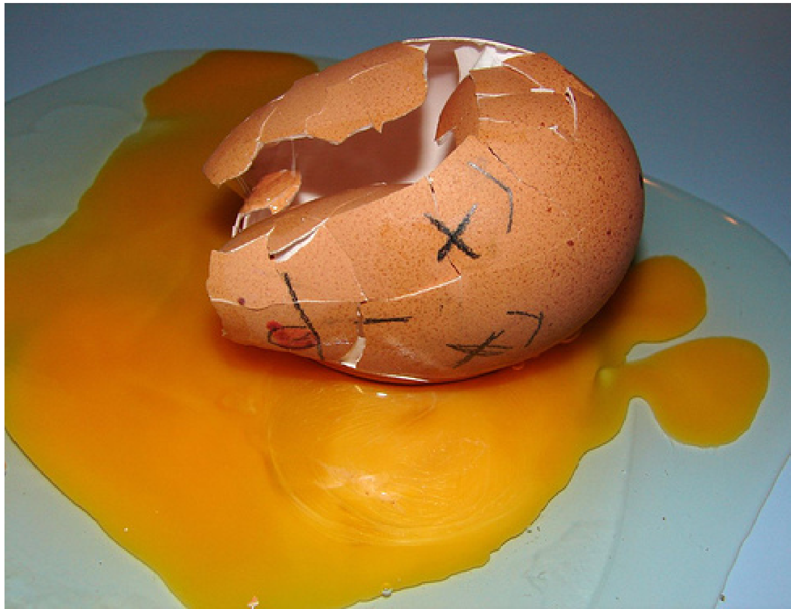
=> Mass participation in information access, processing and production. Decentralization.

We Compentize to Scale



We Want and Need to Integrate

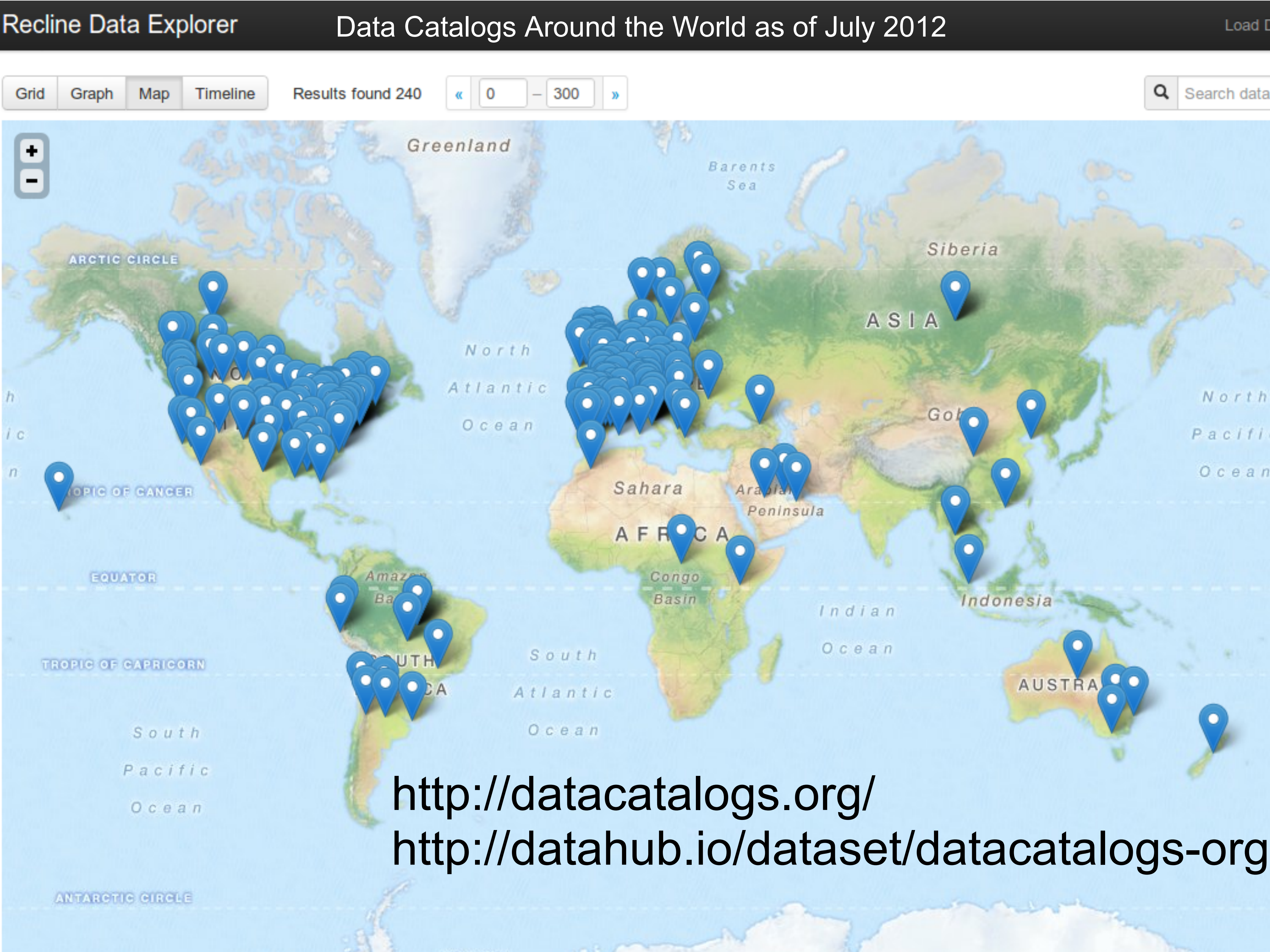
Without Open Data this will Fail!



Huge Growth in Open Data in Last Few Years

Especially for Government Data

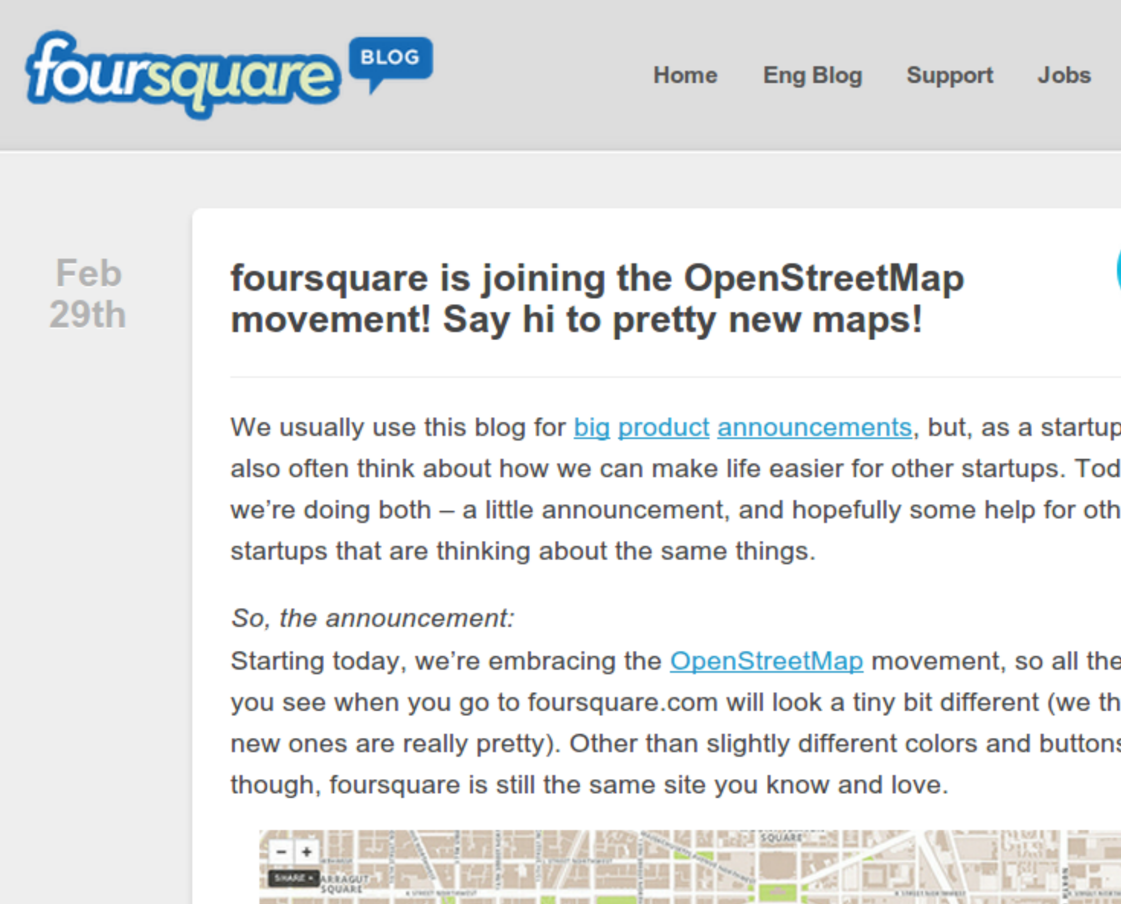
Data Catalogs Around the World as of July 2012



http://datacatalogs.org/

http://datahub.io/dataset/datacatalogs-org

Controladoria-Geral da União
GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA
Open Government Partnership
Parceria para Governo Aberto
Conferência Anual
Annual Meeting



Feb
29th

foursquare is joining the OpenStreetMap movement! Say hi to pretty new maps!

We usually use this blog for big product announcements, but, as a startup also often think about how we can make life easier for other startups. Tod we're doing both – a little announcement, and hopefully some help for oth startups that are thinking about the same things.

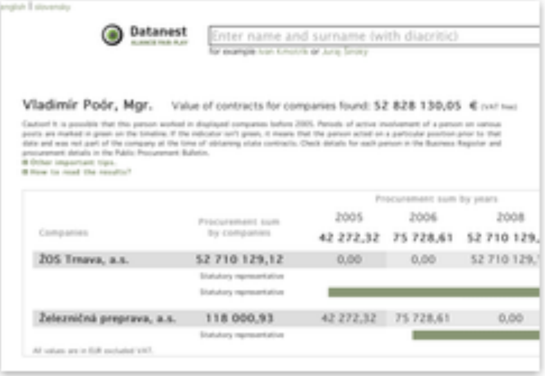
So, the announcement:
Starting today, we're embracing the OpenStreetMap movement, so all the you see when you go to foursquare.com will look a tiny bit different (we th new ones are really pretty). Other than slightly different colors and button though, foursquare is still the same site you know and love.

{Where Next}

Solving Problems
Building Applications

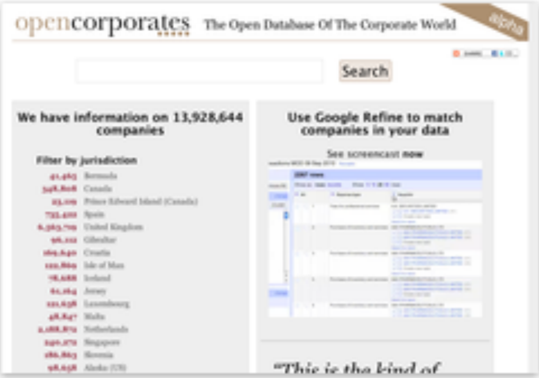
Not about accumulating more and more data!

eatured Applications


Datanest
Enter name and surname (with diacritic)
Vladimír Poór, Mgr.
Value of contracts for companies found: 52 828 130,05 €
Procurement sum by years
2005
2006
2008
ZOS Trnava, a.s.
52 710 129,12
Železničná preprava, a.s.
118 000,93

ZNasichDani / From Our Taxes

ZNasichDani.sk uncovers who are influential persons (owners, managers, statutories) standing behind companies successful in securing contracts with the state, thus helping...


opencorporates
The Open Database Of The Corporate World
alpha
Search
We have information on 13,928,644 companies
Filter by jurisdiction
Use Google Refine to match companies in your data
See screencast now
"This is the kind of

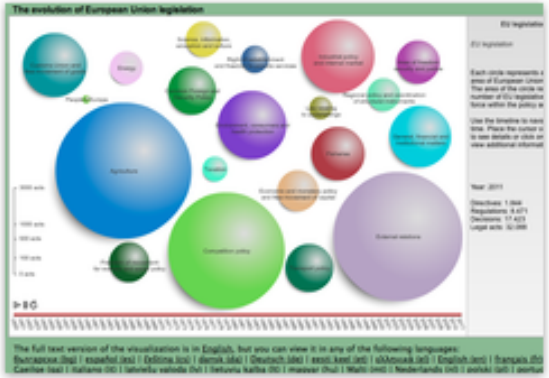
OpenCorporates

OpenCorporates has taken one of the most important global datasets -- companies, and government data relating to them -- and for the first time exposed it on the web in an open,...

Live train map for the London Underground, by Matthew Somerville
What is this?
How does it
Who did this

Live London Underground tube map

It plots the current positions of all London Underground trains on a map, and updates the map in real time. It provides a stunning visualisation of the sheer amount going on...



Evolution of European Union legislation

Explaining the legislative activity of the European Union in time and within different policy areas. It

Bike Share Map

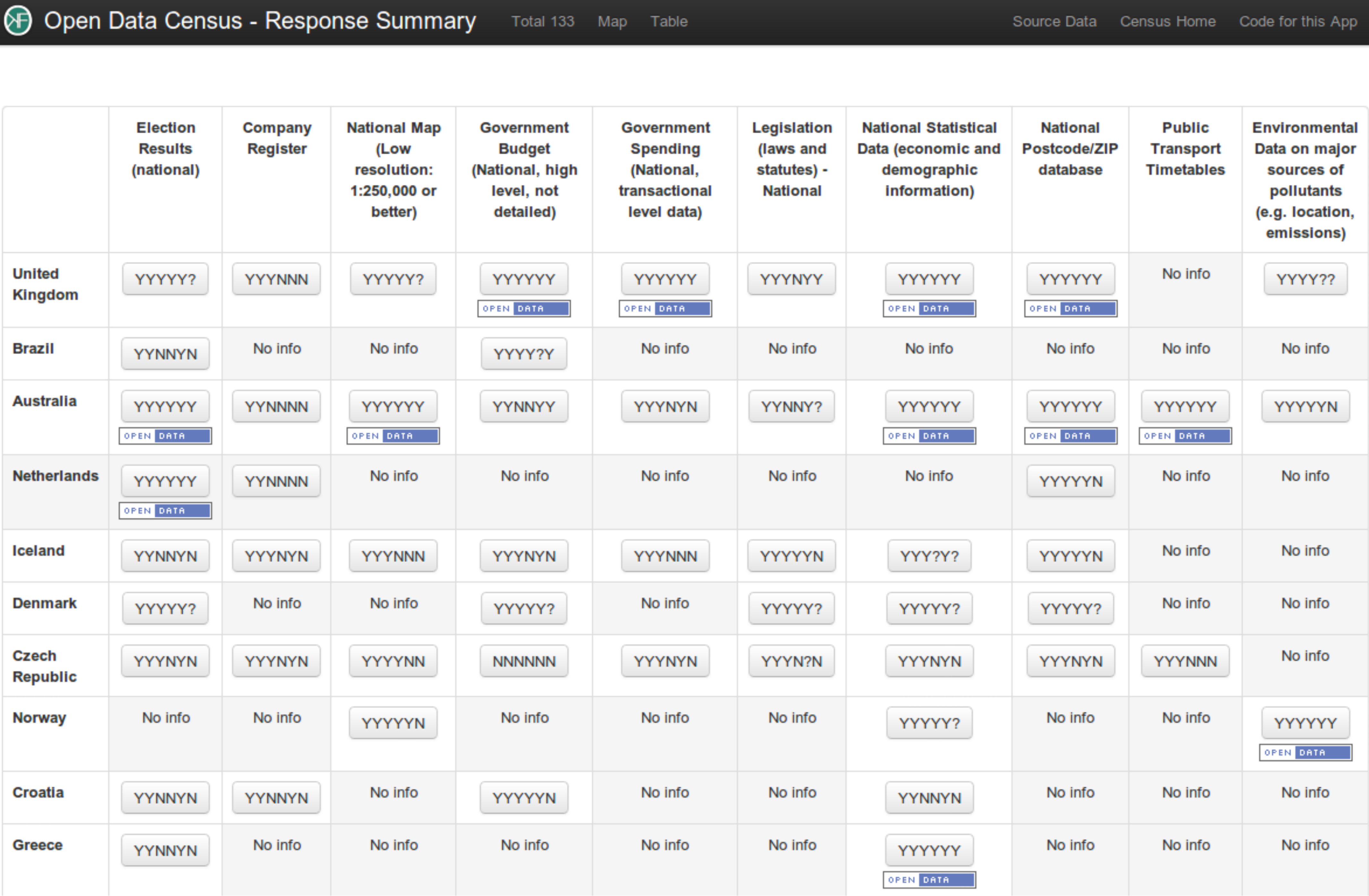
Showing the current state of bike share systems in over 30 cities around the world - from London to Barcelona, from Bordeaux to Vienna from and

Europe's carbon dioxide emissions

Our emissions map (http://www.sandbag.org.uk/maps/emissions/) shows how much carbon dioxide is emitted by

Toy vs Core Datasets

Location of park benches vs National Map



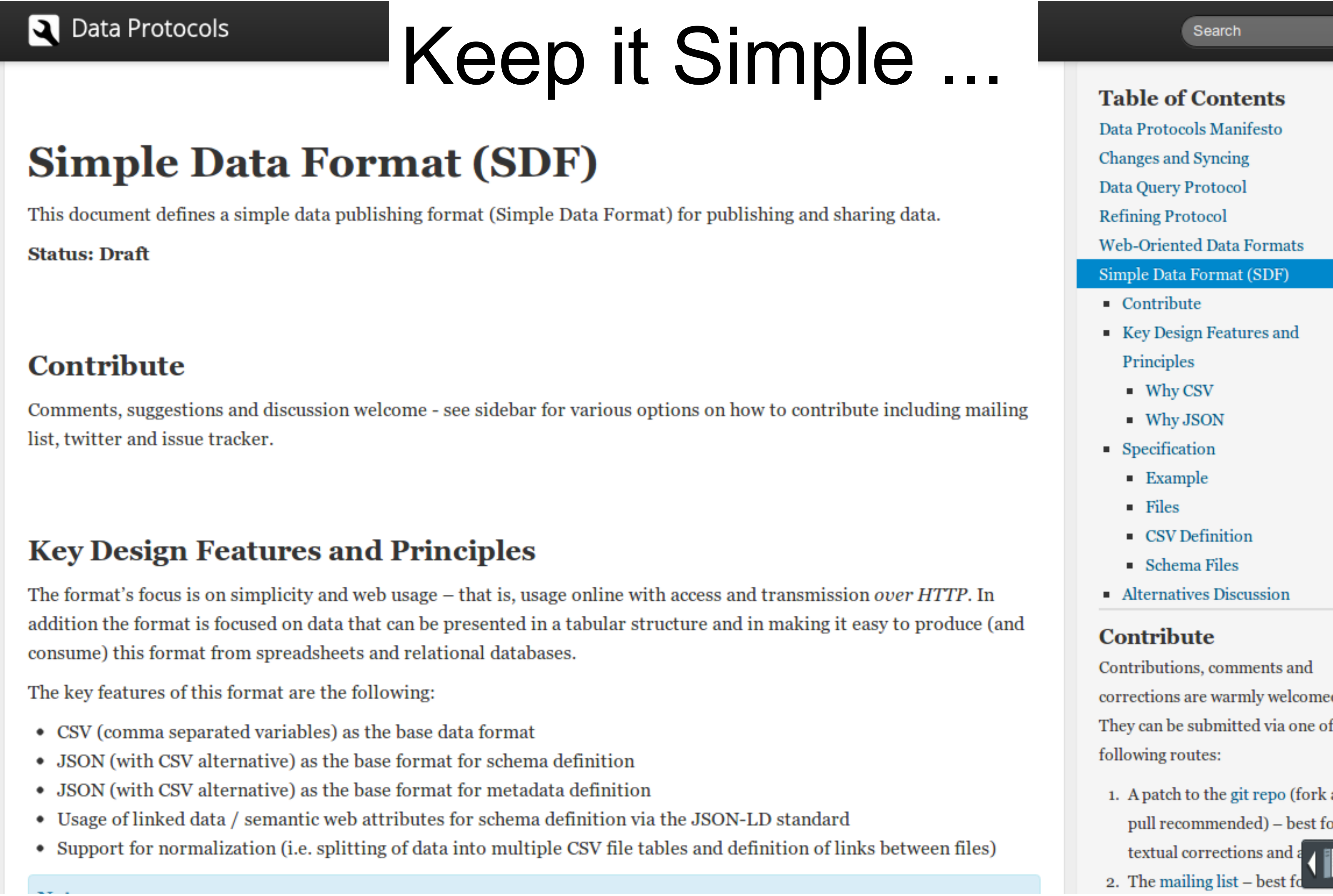
Open Data Census - Response Summary

	Election Results (national)	Company Register	National Map (Low resolution: 1:250,000 or better)	Government Budget (National, high level, not detailed)	Government Spending (National, transactional level data)	Legislation (laws and statutes) - National	National Statistical Data (economic and demographic information)	National Postcode/ZIP database	Public Transport Timetables	Environmental Data on major sources of pollutants (e.g. location, emissions)
United Kingdom	YYYYY?	YYYNNN	YYYYY?	YYYYYY OPEN DATA	YYYYYY OPEN DATA	YYYNYY	YYYYYY OPEN DATA	YYYYYY OPEN DATA	No info	YYYY??
Brazil	YYNNYN	No info	No info	YYYY?Y	No info	No info	No info	No info	No info	No info
Australia	YYYYYY OPEN DATA	YYNNNN	YYYYYY OPEN DATA	YYNNYY	YYYNYN	YYNNY?	YYYYYY OPEN DATA	YYYYYY OPEN DATA	YYYYYY OPEN DATA	YYYYYN
Netherlands	YYYYYY OPEN DATA	YYNNNN	No info	No info	No info	No info	No info	YYYYYN	No info	No info
Iceland	YYNNYN	YYYNYN	YYYNNN	YYYNYN	YYYNNN	YYYYYN	YYY?Y?	YYYYYN	No info	No info
Denmark	YYYYY?	No info	No info	YYYYY?	No info	YYYYY?	YYYYY?	YYYYY?	No info	No info
Czech Republic	YYYNYN	YYYNYN	YYYYNN	NNNNNN	YYYNYN	YYYN?N	YYYNYN	YYYNYN	YYYNNN	No info
Norway	No info	No info	YYYYYN	No info	No info	No info	YYYYY?	No info	No info	YYYYYY OPEN DATA
Croatia	YYNNYN	YYNNYN	No info	YYYYYN	No info	No info	YYNNYN	No info	No info	No info
Greece	YYNNYN	No info	No info	No info	No info	No info	YYYYYY OPEN DATA	No info	No info	No info

http://census.opengovernmentdata.org/

Machine Readable

Keep it Simple ...



Simple Data Format (SDF)

This document defines a simple data publishing format (Simple Data Format) for publishing and sharing data.

Status: Draft

Contribute

Comments, suggestions and discussion welcome - see sidebar for various options on how to contribute including mailing list, twitter and issue tracker.

Key Design Features and Principles

The format's focus is on simplicity and web usage – that is, usage online with access and transmission *over HTTP*. In addition the format is focused on data that can be presented in a tabular structure and in making it easy to produce (and consume) this format from spreadsheets and relational databases.

The key features of this format are the following:

- CSV (comma separated variables) as the base data format
- JSON (with CSV alternative) as the base format for schema definition
- JSON (with CSV alternative) as the base format for metadata definition
- Usage of linked data / semantic web attributes for schema definition via the JSON-LD standard
- Support for normalization (i.e. splitting of data into multiple CSV file tables and definition of links between files)

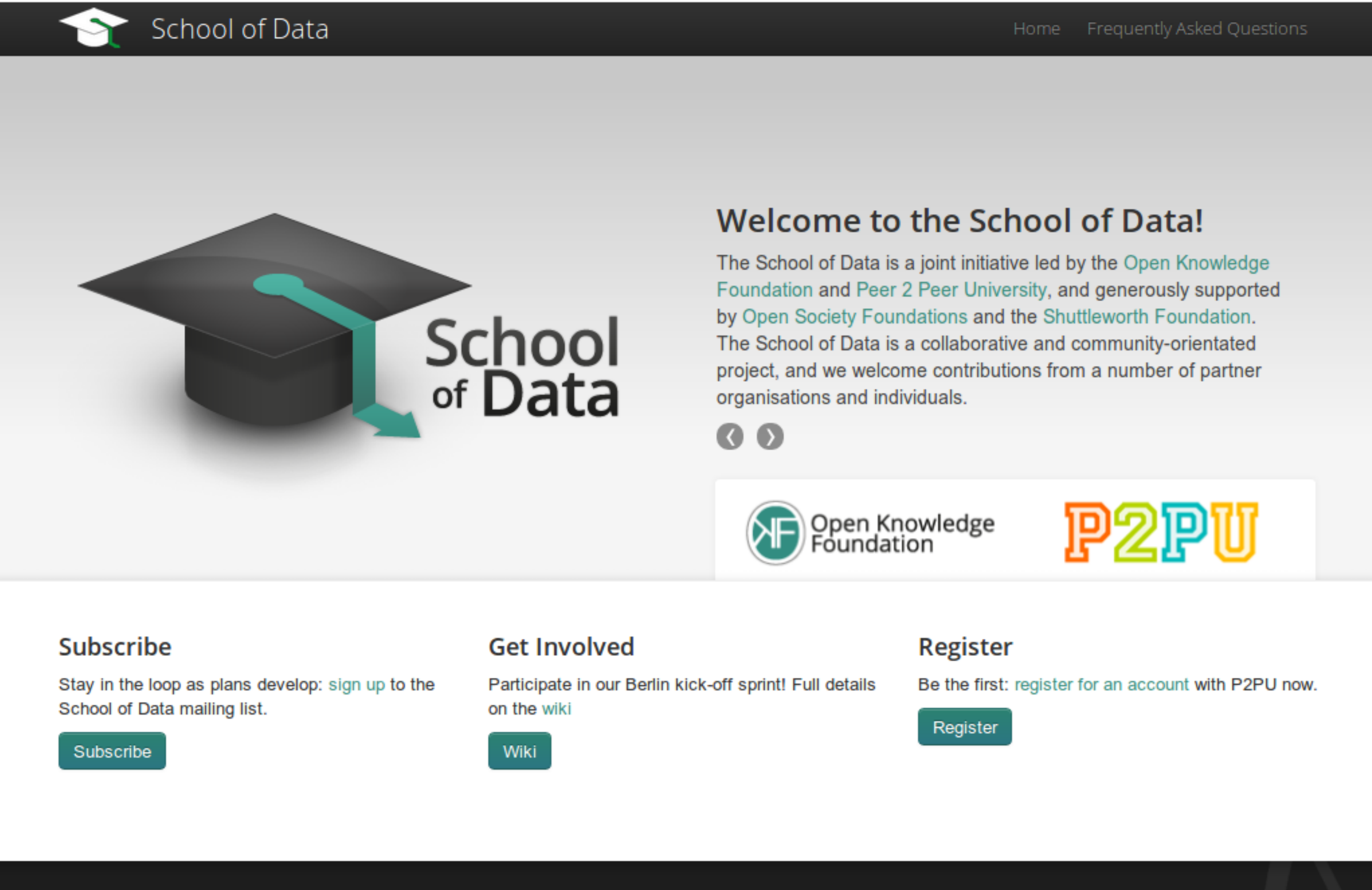
Table of Contents

Contribute

Contributions, comments and corrections are warmly welcome
They can be submitted via one of
following routes:

1. A patch to the git repo (fork a
pull recommended) – best fo
textual corrections and a
2. The mailing list – best fo

Education and Skills



Welcome to the School of Data!

The School of Data is a joint initiative led by the Open Knowledge Foundation and Peer 2 Peer University, and generously supported by Open Society Foundations and the Shuttleworth Foundation. The School of Data is a collaborative and community-orientated project, and we welcome contributions from a number of partner organisations and individuals.

Subscribe

Stay in the loop as plans develop: sign up to the School of Data mailing list.

Get Involved

Participate in our Berlin kick-off sprint! Full details on the wiki

Register

Be the first: register for an account with P2PU now.

Small Data
vs
Big Data

It's about small pieces loosely joined not one ring to rule them all!

{Conclusion}

Increasing Amounts of Data

But need to ensure
It is *really* open
Of reasonable quality
(good enough not perfect)

Open Data is Platform not a Commodity

Let's Build on It, Not Sell It!

Be Problem and Application Driven

(Rather than Data and Technology Driven)

Remember Faraday's Baby

Open Knowledge Foundation

@okfn[.org]
@openspending[.org]
CKAN.org
DataHub.io
PublicDomainReview.org
@SchoolOfData[.org]
~
Rufus Pollock @rufuspollock[.org]

SHUTTLEWORTH
FELLOW