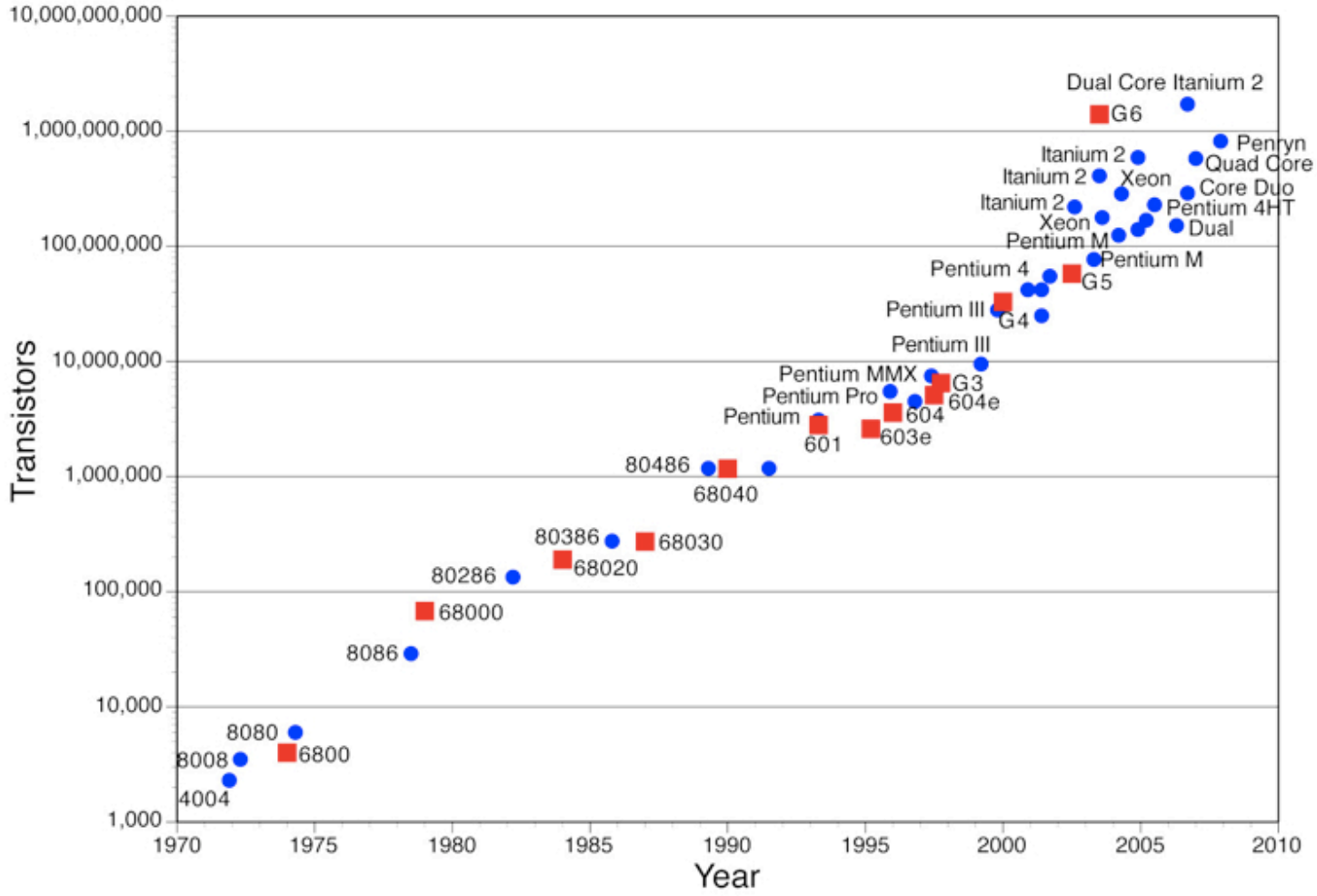
UNIVERSITY OF Southampton
School of Electronics and Computer Science

From Data to Decisions: The Power of Information in the Age of the Web of Linked Data

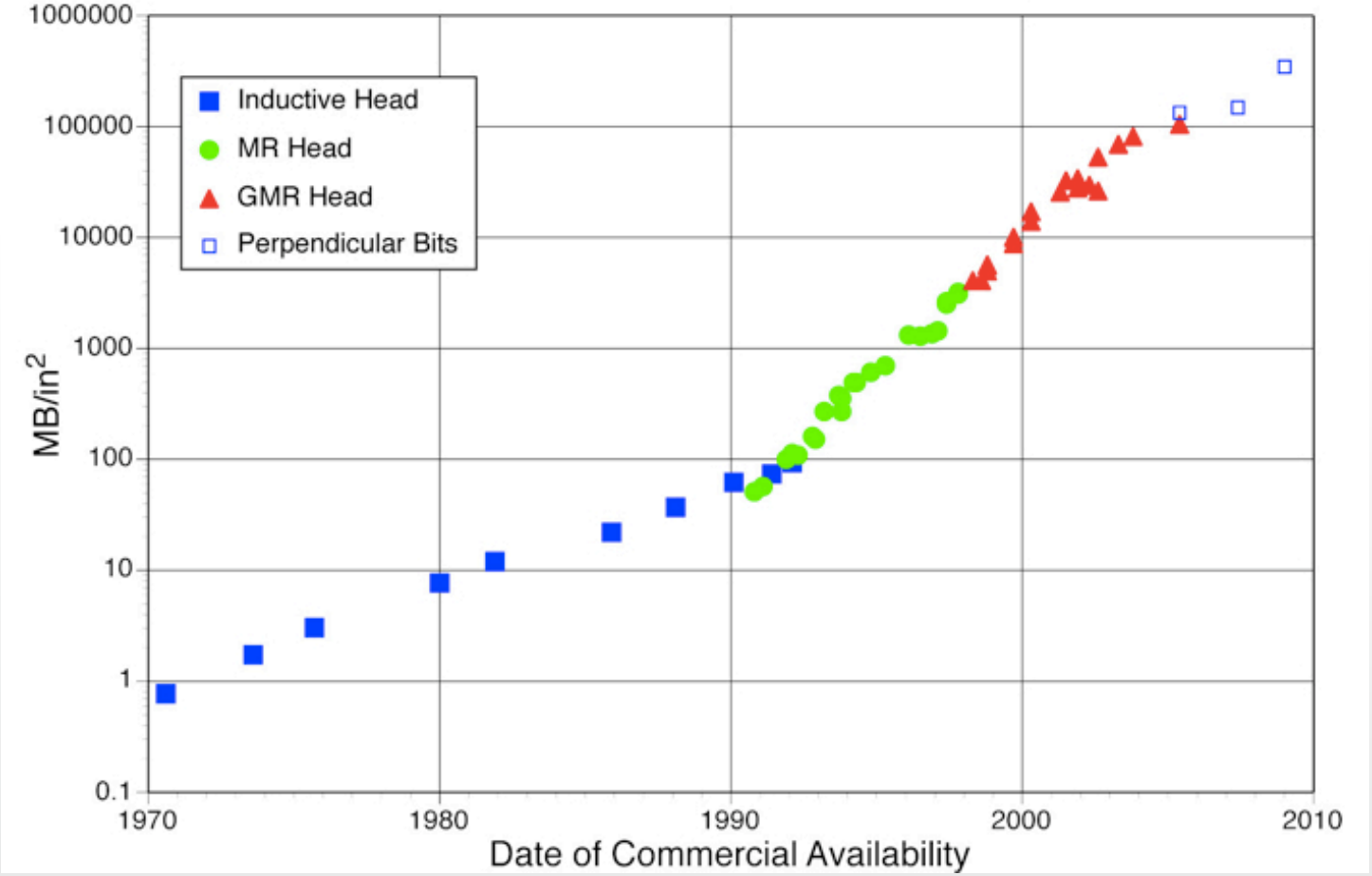
Professor Nigel Shadbolt FREng

HQS Wellington
19th October 2010

Technology: Powers of 10



Technology:
Powers of 10

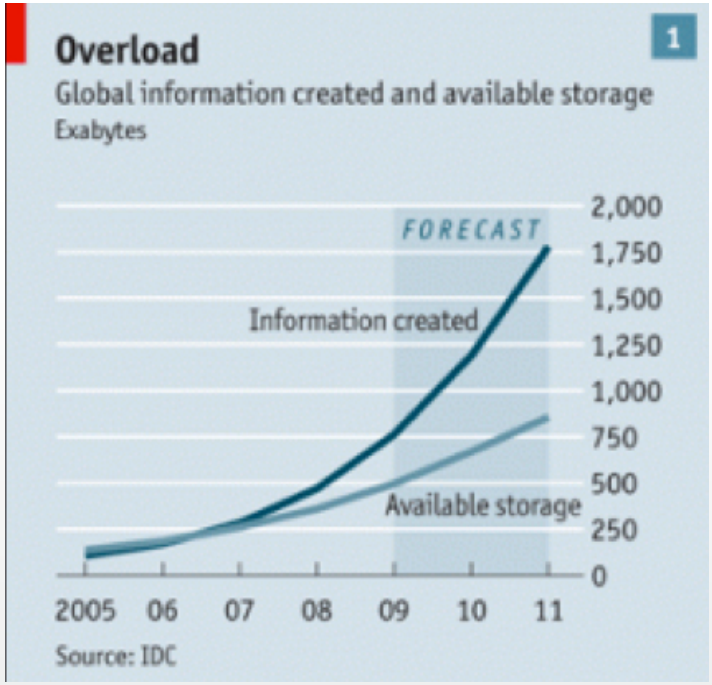


Data:
Powers of 10

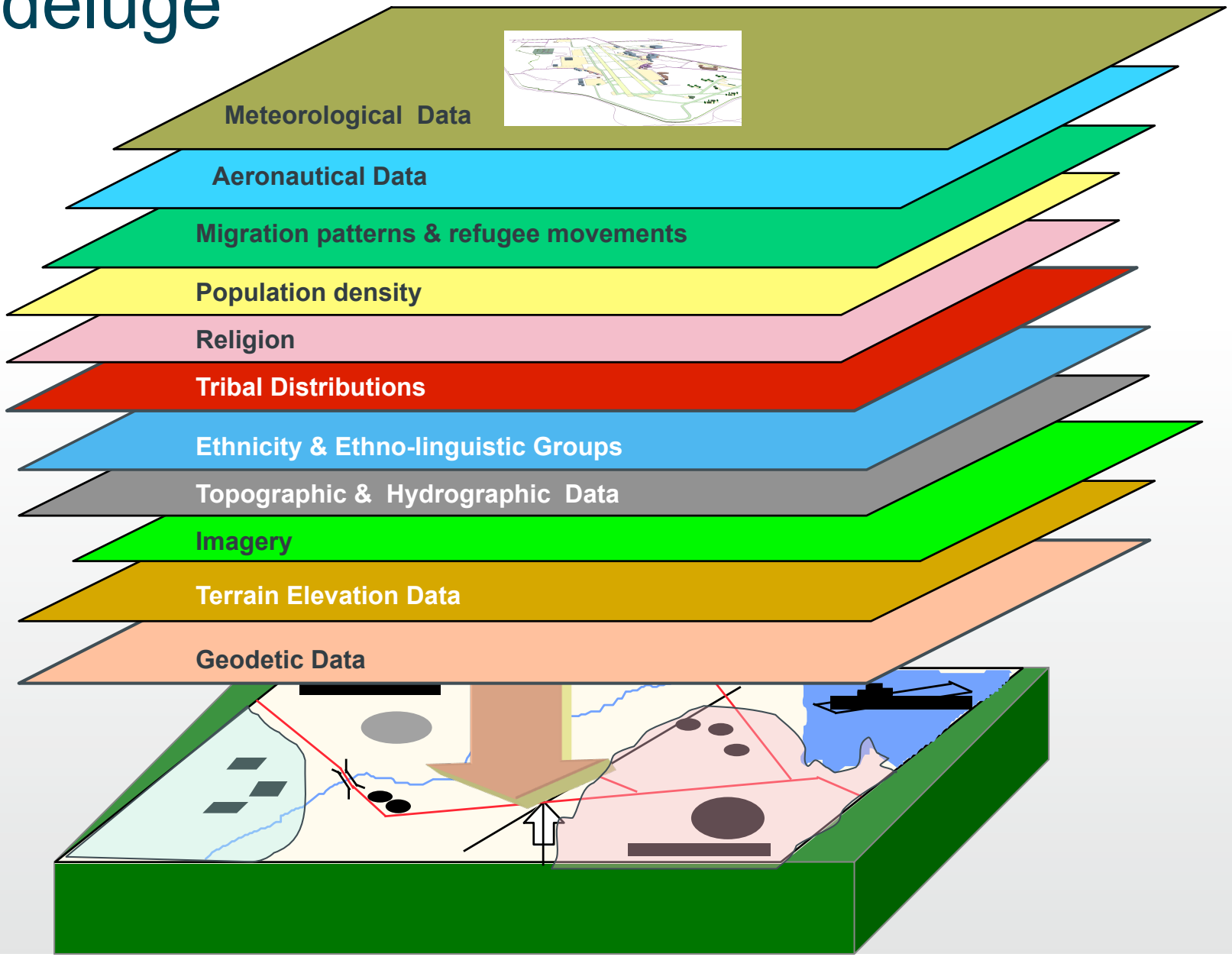
- 2007 – 281 EB (1 exabytes 1 x 10^18 bytes))
- Or 281 trillion digitised novels
- 2010 – 1200 EB
- 2011 – 1.8 ZB

Data:
Powers of 10

2011 amount of digital content created will be twice the amount of available storage

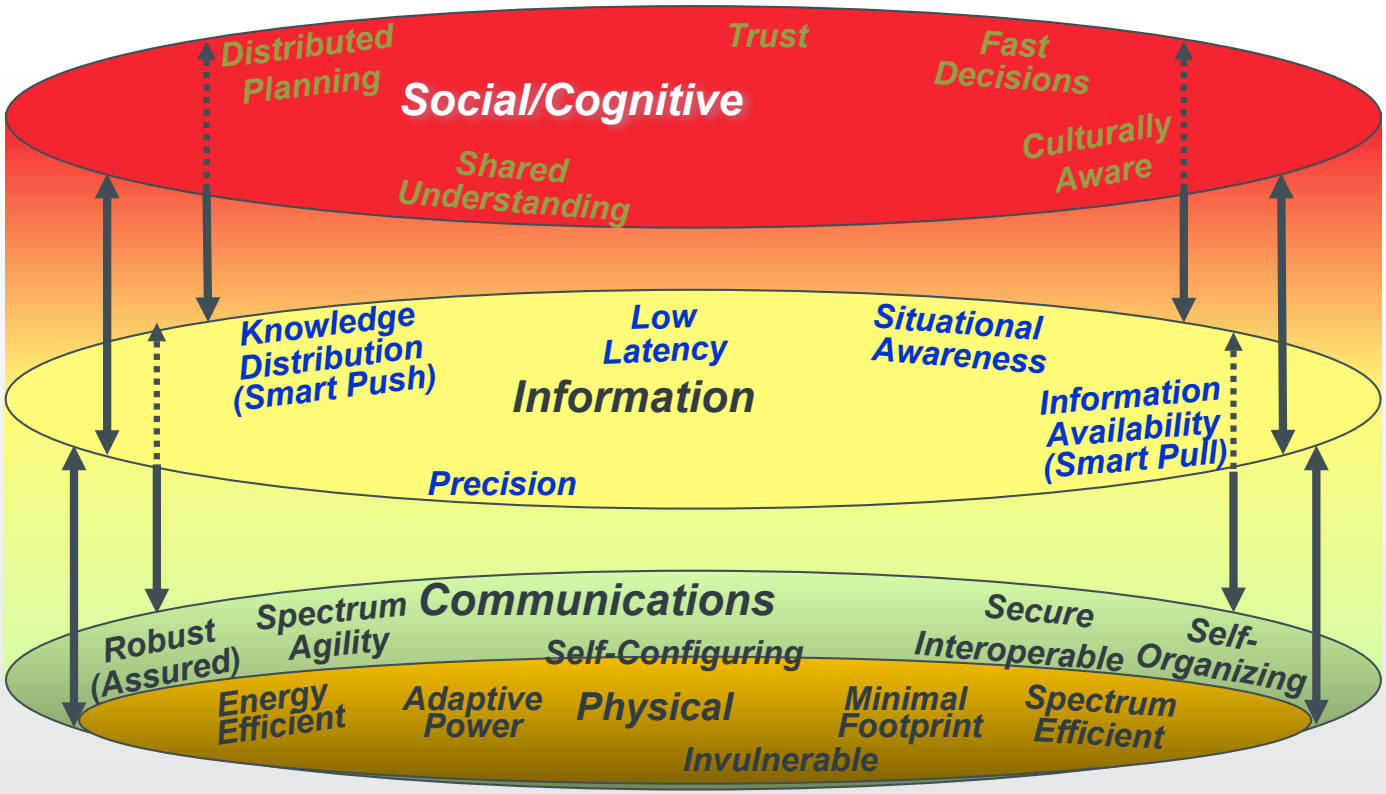


Data:
A deluge



Networks:
A plethora

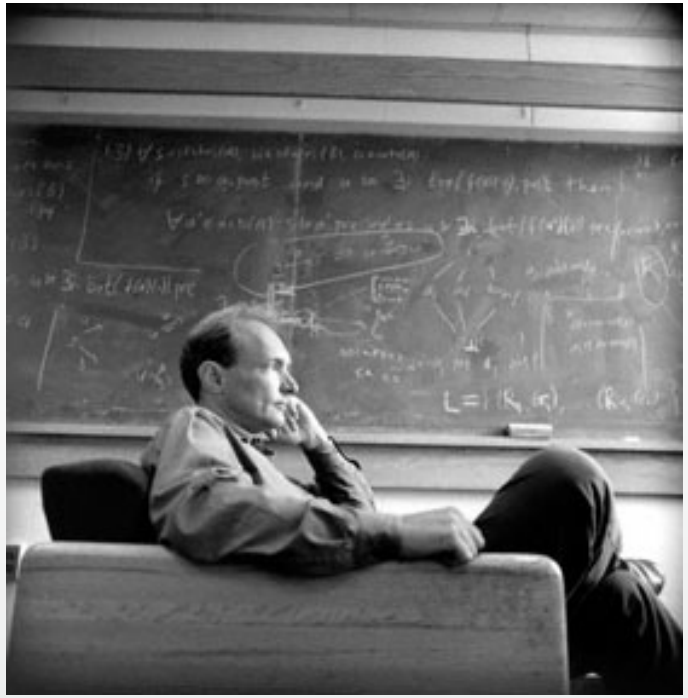
Command and Control ⇌ Collaborate and Connect

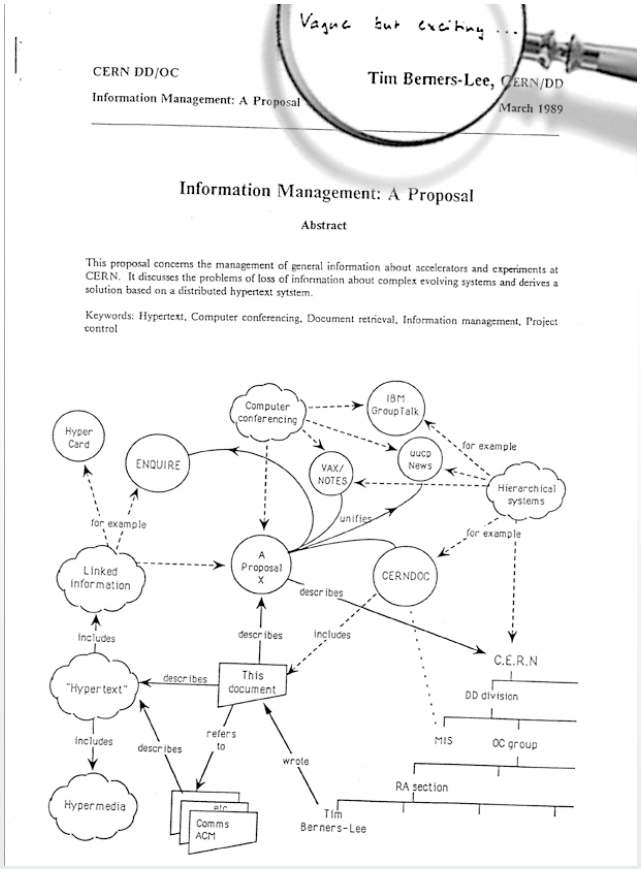


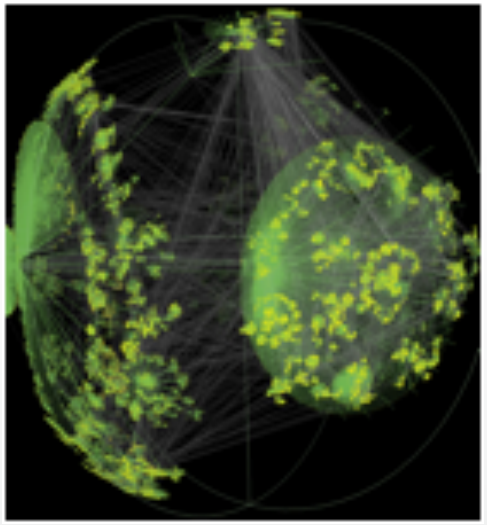
UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

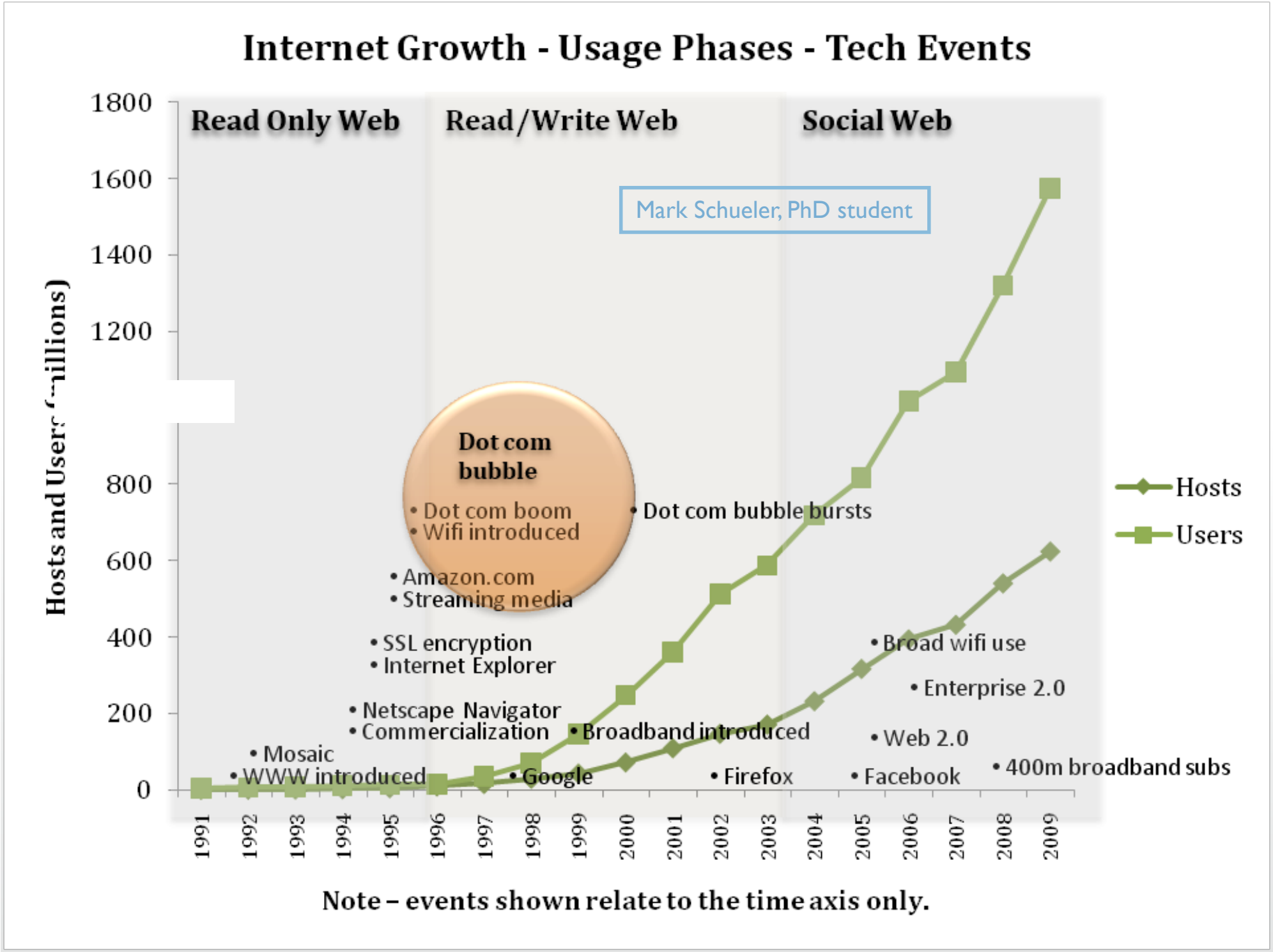
How to adapt?

The Web – most successful information architecture in history

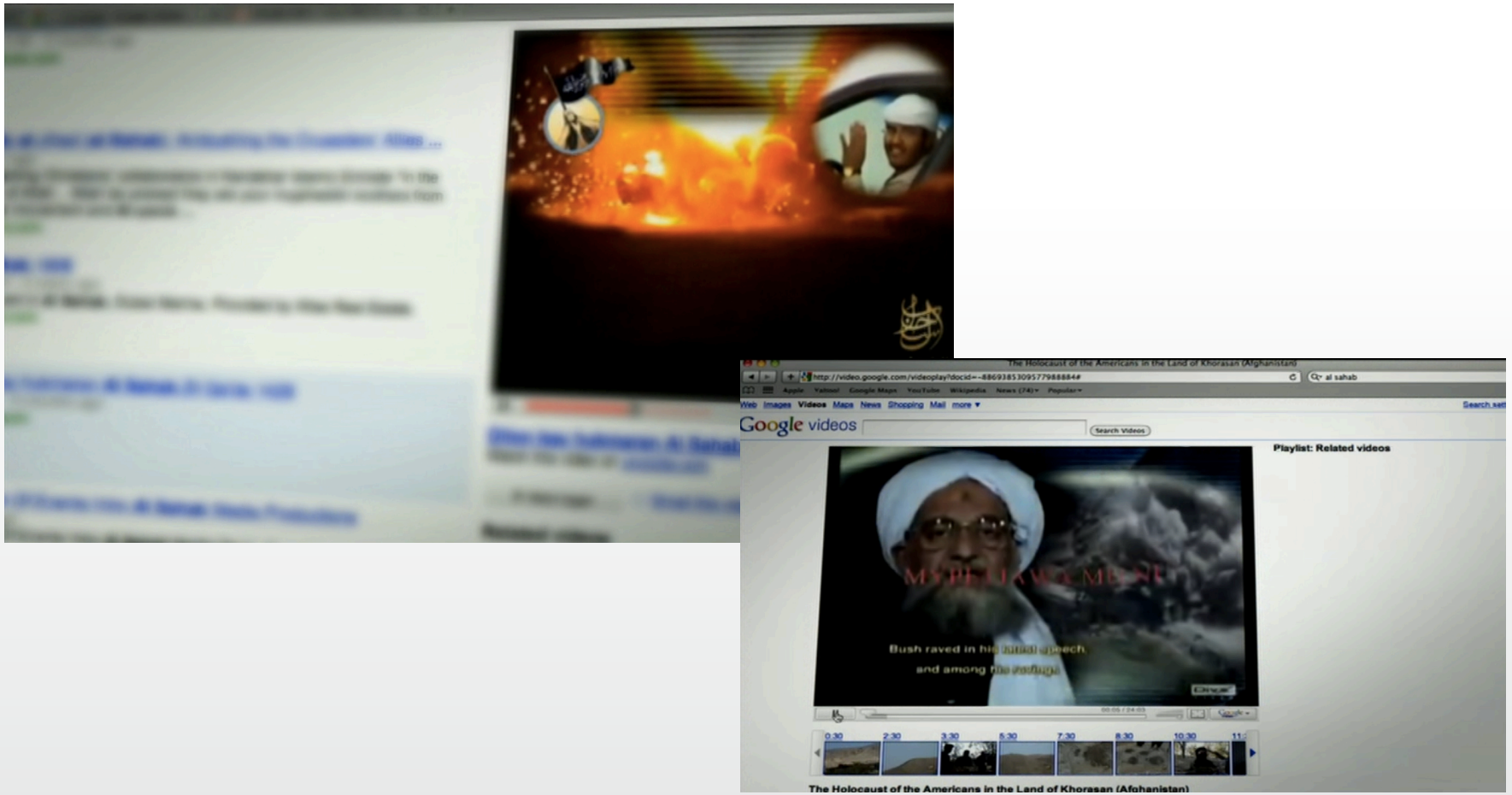




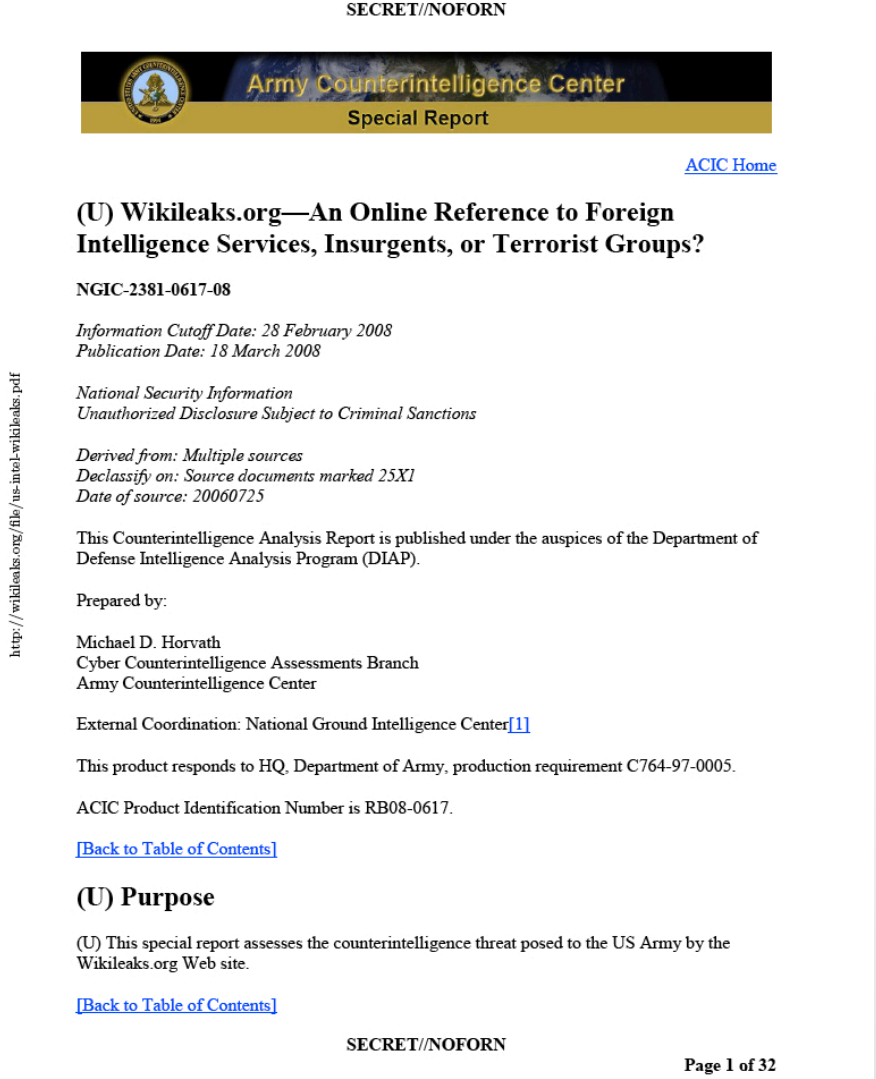



Internet Growth - Usage Phases - Tech Events
Read Only Web
Read/Write Web
Social Web
Mark Schueler, PhD student
Hosts and Users (millions)
1800
1600
1400
1200
800
600
400
200
0
Dot com bubble
• Dot com boom
• Wifi introduced
• Amazon.com
• Streaming media
• SSL encryption
• Internet Explorer
• Netscape Navigator
• Commercialization
• Mosaic
• WWW introduced
• Google
• Broadband introduced
• Dot com bubble bursts
• Firefox
• Broad wifi use
• Enterprise 2.0
• Web 2.0
• Facebook
• 400m broadband subs
Hosts
Users
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Note – events shown relate to the time axis only.

The Web in your World

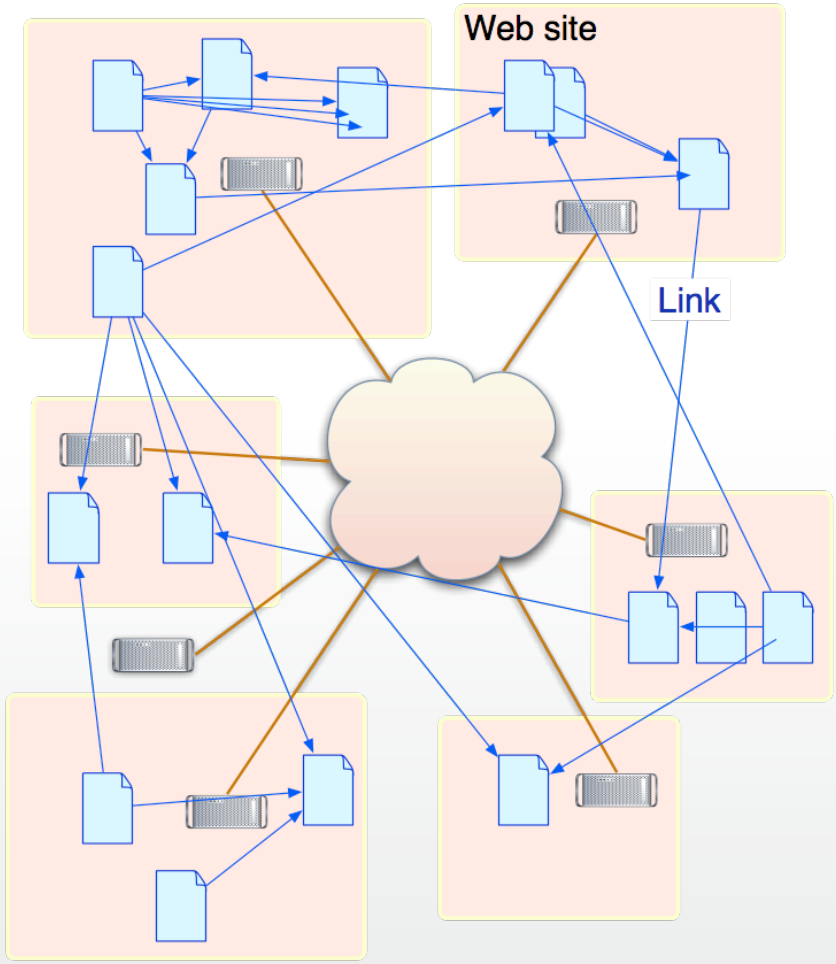


The Web in your World

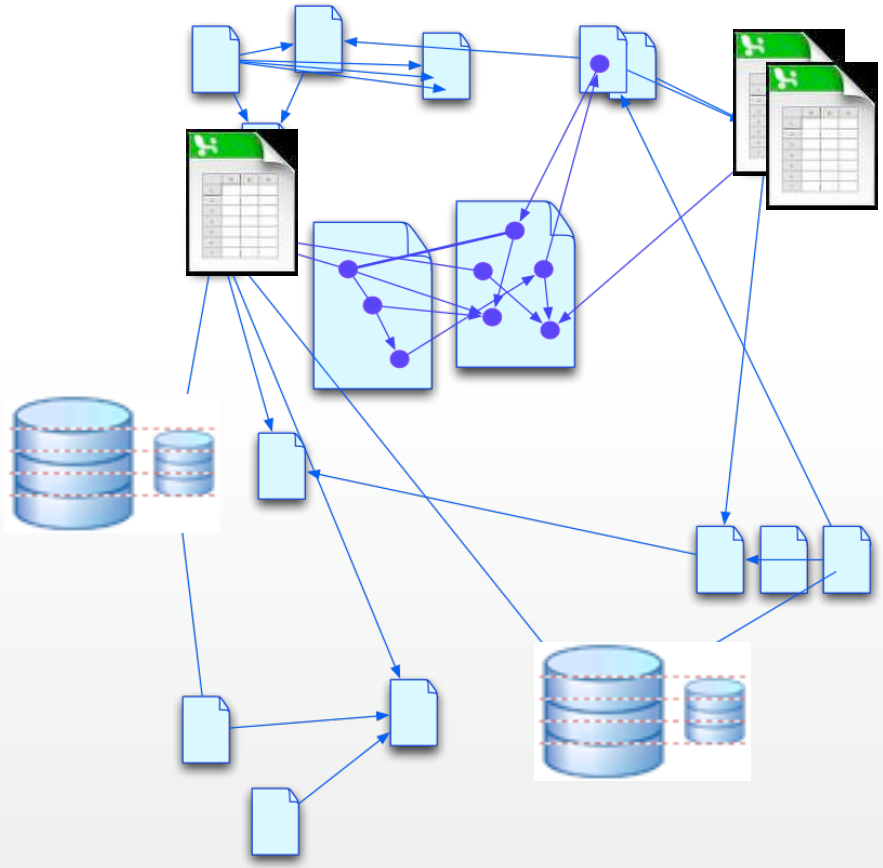


The Web evolves fast

A Web of documents…



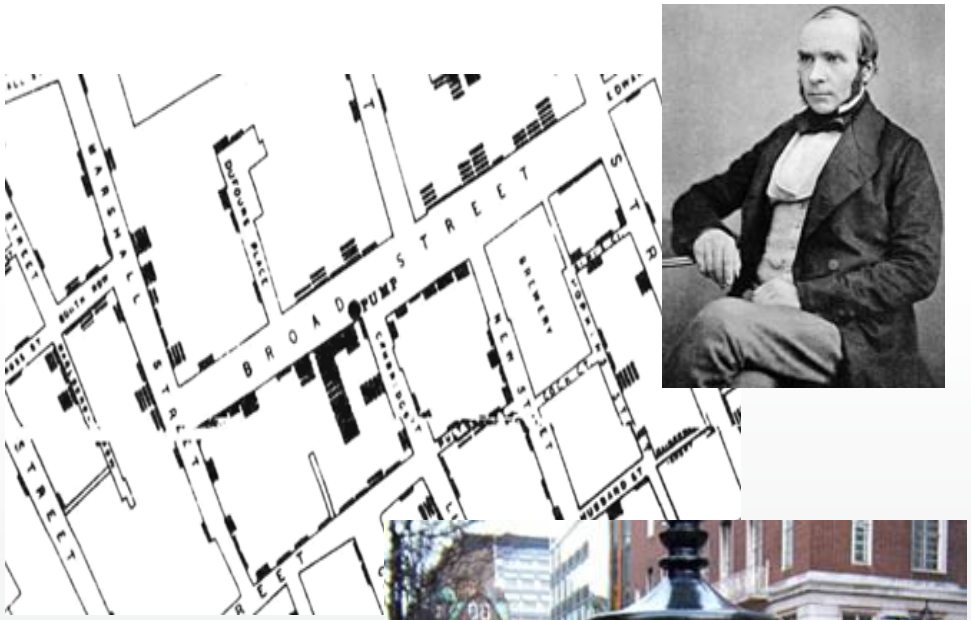
Also a Web of data…



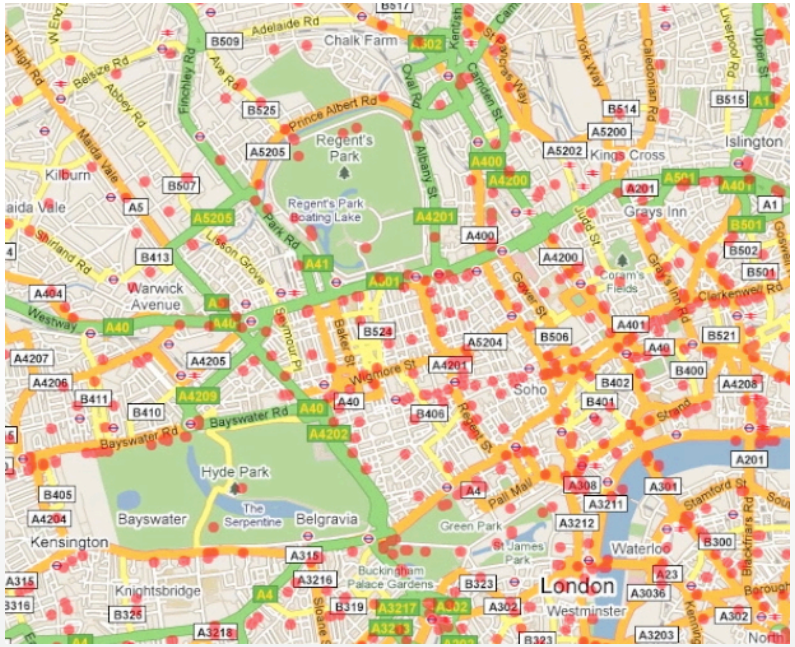
UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

The 5 ‘O’s

The Power of Open Data



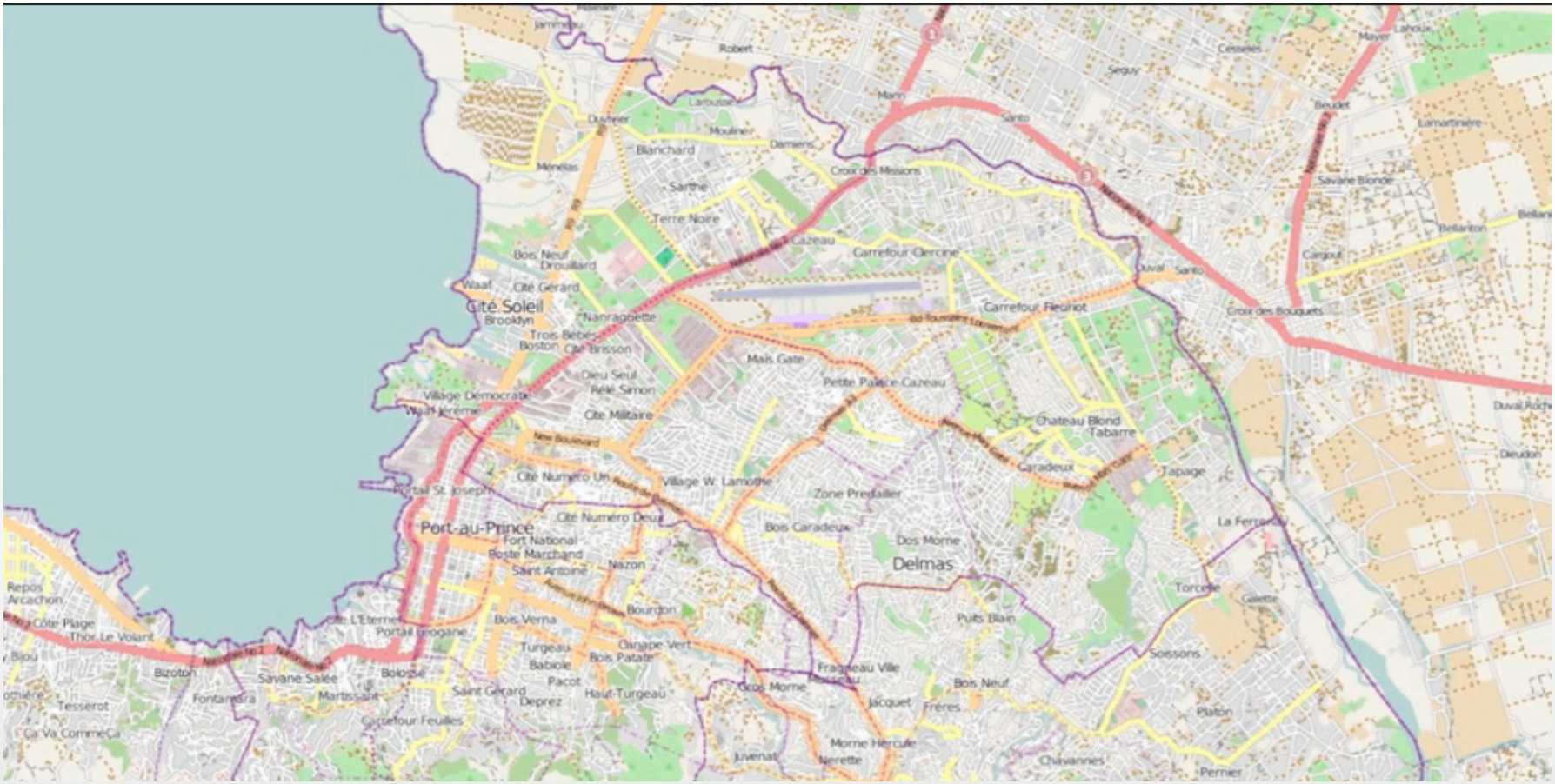
Cholera cases



Bicycling traffic accidents

The Map of Port-au-Prince 13th Jan 2010
The Power of Open Date - from this…

to this… in 13 days



How did this happen?

- Open Data
- Open Standards
- Open Source
- Open Licences
- Open Minds

- What lessons might you draw from this?
- Web can be understood as extranet or intranet – http, html, etc.…

UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

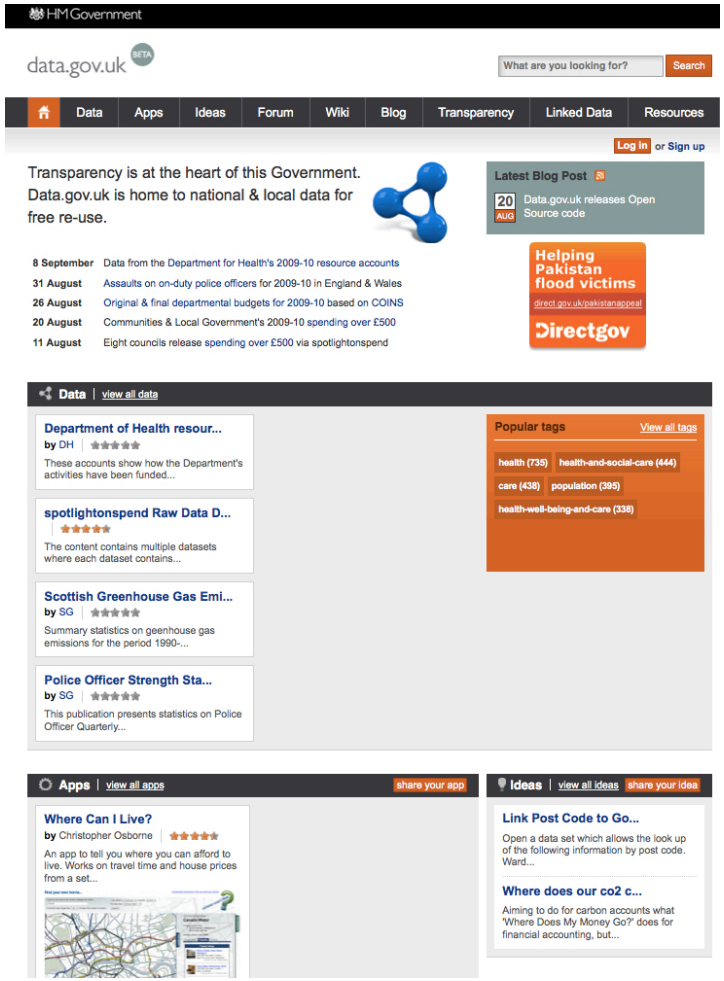
Open Data

Open Data Publishing – Star Quality

★	Put your data on the Web – (any format)
★★	Make it available as structured data (e.g. Excel, CSV, instead of PDF)
★★★	Use open, standard formats (e.g. XML, RDF)
★★★★	Use URLs to identify things (so people and machines can point at your data)
★★★★★	Link your data to other people's data

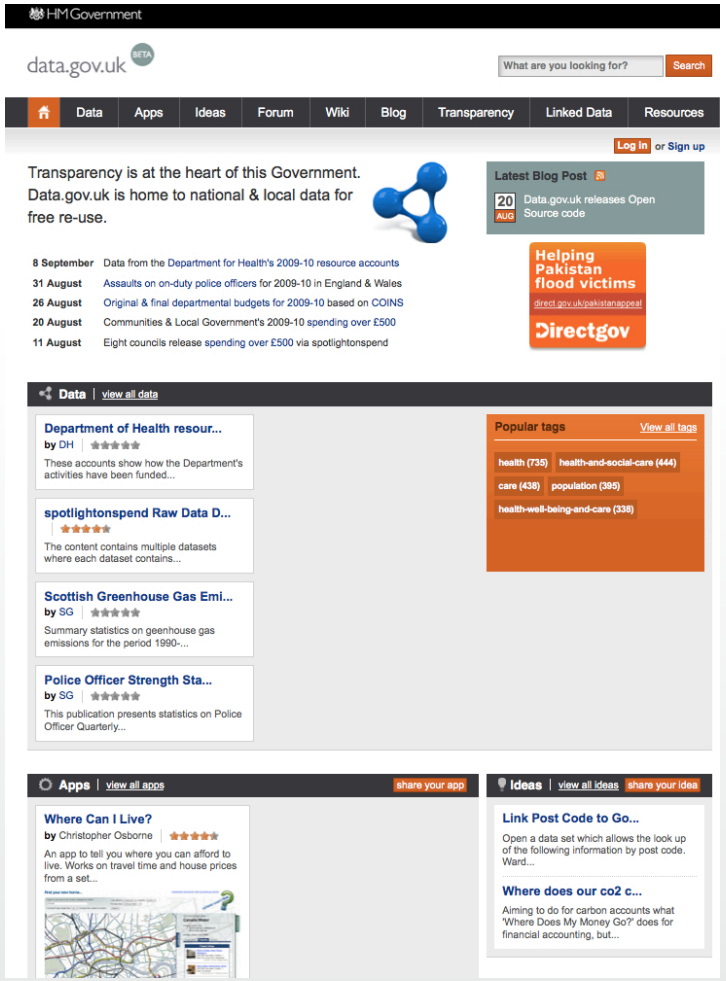
OGD – poster child for open data

- Tim Berners-Lee and myself appointed 9th June 2009 and set about creating data.gov.uk
- Reappointed to Public Sector Transparency Board June 2010
- Open Government Data (OGD) is taking hold
- Governments, local authorities, cities releasing data



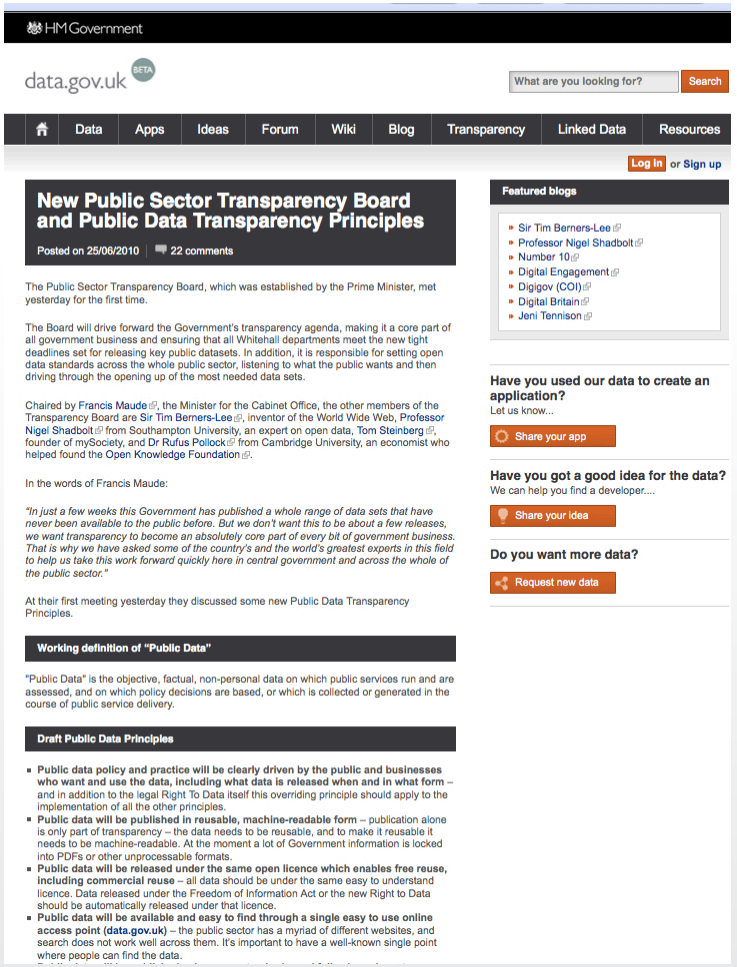
data.gov.uk

- data.gov.uk itself (now with over 4000 datasets)
  - open source, open standards, open licence
  - key data sets released inc OS OpenData, COINS etc.
  - applications from outside HMG
- overcoming the many objections to transparency of data
- community of data users and developers
- political and technical leadership
- establishing public data principles



Public Data and the Public Data Principles

"Public Data" is the objective, factual, non-personal data on which public services run and are assessed, and on which policy decisions are based, or which is collected or generated in the course of public service delivery.



HM Government

data.gov.uk BETA

What are you looking for? Search

Data | Apps | Ideas | Forum | Wiki | Blog | Transparency | Linked Data | Resources

Log in or Sign up

New Public Sector Transparency Board and Public Data Transparency Principles

Posted on 25/06/2010 | 22 comments

The Public Sector Transparency Board, which was established by the Prime Minister, met yesterday for the first time.

The Board will drive forward the Government's transparency agenda, making it a core part of all government business and ensuring that all Whitehall departments meet the new tight deadlines set for releasing key public datasets. In addition, it is responsible for setting open data standards across the whole public sector, listening to what the public wants and then driving through the opening up of the most needed data sets.

Chaired by Francis Maude, the Minister for the Cabinet Office, the other members of the Transparency Board are Sir Tim Berners-Lee, inventor of the World Wide Web, Professor Nigel Shadbolt from Southampton University, an expert on open data, Tom Steinberg, founder of mySociety, and Dr Rufus Pollock from Cambridge University, an economist who helped found the Open Knowledge Foundation.

In the words of Francis Maude:

"In just a few weeks this Government has published a whole range of data sets that have never been available to the public before. But we don't want this to be about a few releases, we want transparency to become an absolutely core part of every bit of government business. That is why we have asked some of the country's and the world's greatest experts in this field to help us take this work forward quickly here in central government and across the whole of the public sector."

At their first meeting yesterday they discussed some new Public Data Transparency Principles.

Working definition of "Public Data"

"Public Data" is the objective, factual, non-personal data on which public services run and are assessed, and on which policy decisions are based, or which is collected or generated in the course of public service delivery.

Draft Public Data Principles

- **Public data policy and practice will be clearly driven by the public and businesses who want and use the data, including what data is released when and in what form** – and in addition to the legal Right To Data itself this overriding principle should apply to the implementation of all the other principles.
- **Public data will be published in reusable, machine-readable form** – publication alone is only part of transparency – the data needs to be reusable, and to make it reusable it needs to be machine-readable. At the moment a lot of Government information is locked into PDFs or other unprocessable formats.
- **Public data will be released under the same open licence which enables free reuse, including commercial reuse** – all data should be under the same easy to understand licence. Data released under the Freedom of Information Act or the new Right to Data should be automatically released under that licence.
- **Public data will be available and easy to find through a single easy to use online access point (data.gov.uk)** – the public sector has a myriad of different websites, and search does not work well across them. It's important to have a well-known single point where people can find the data.

Featured blogs

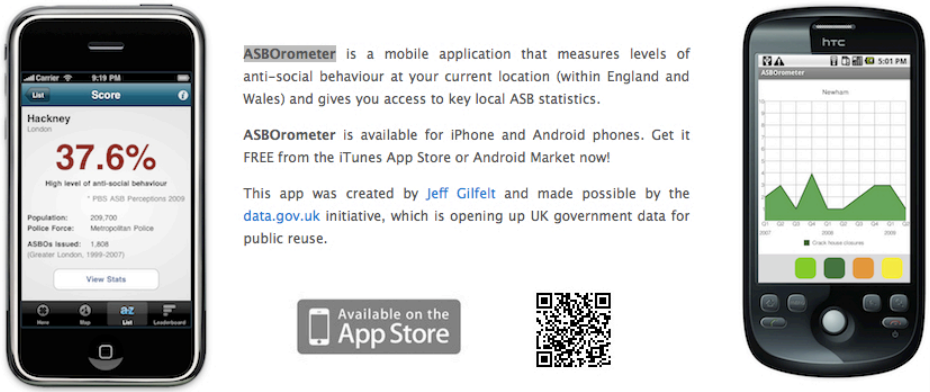
- Sir Tim Berners-Lee
- Professor Nigel Shadbolt
- Number 10
- Digital Engagement
- Digigov (COI)
- Digital Britain
- Jeni Tennison

Have you used our data to create an application?
Let us know...
Share your app

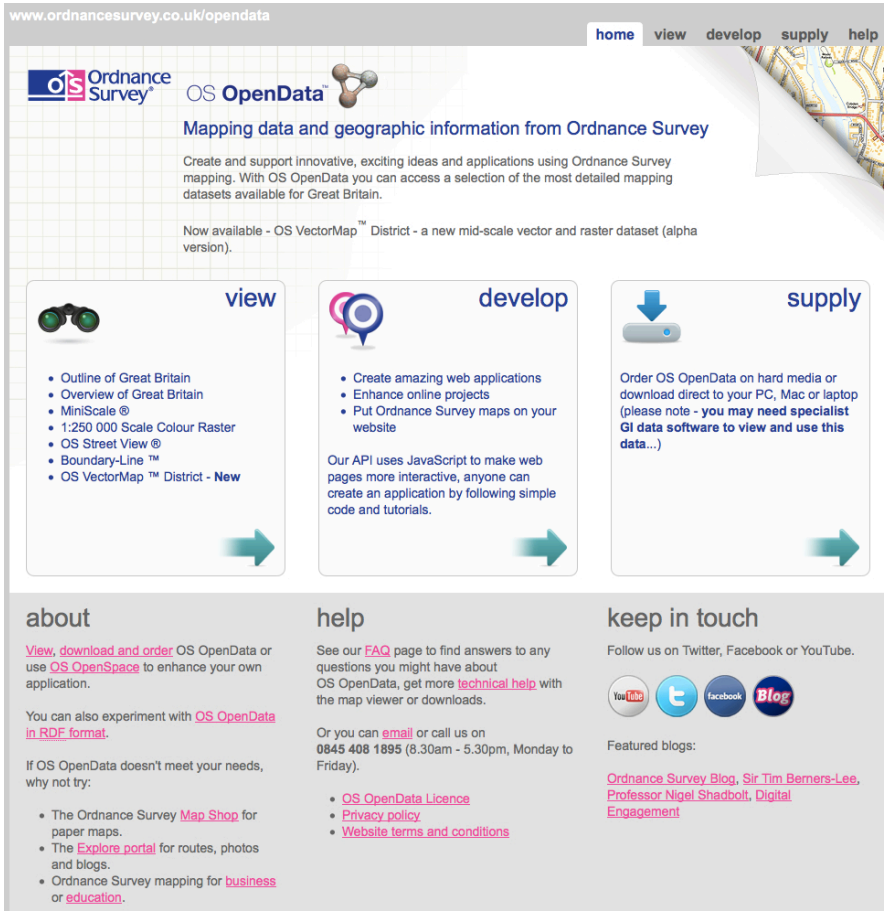
Have you got a good idea for the data?
We can help you find a developer....
Share your idea

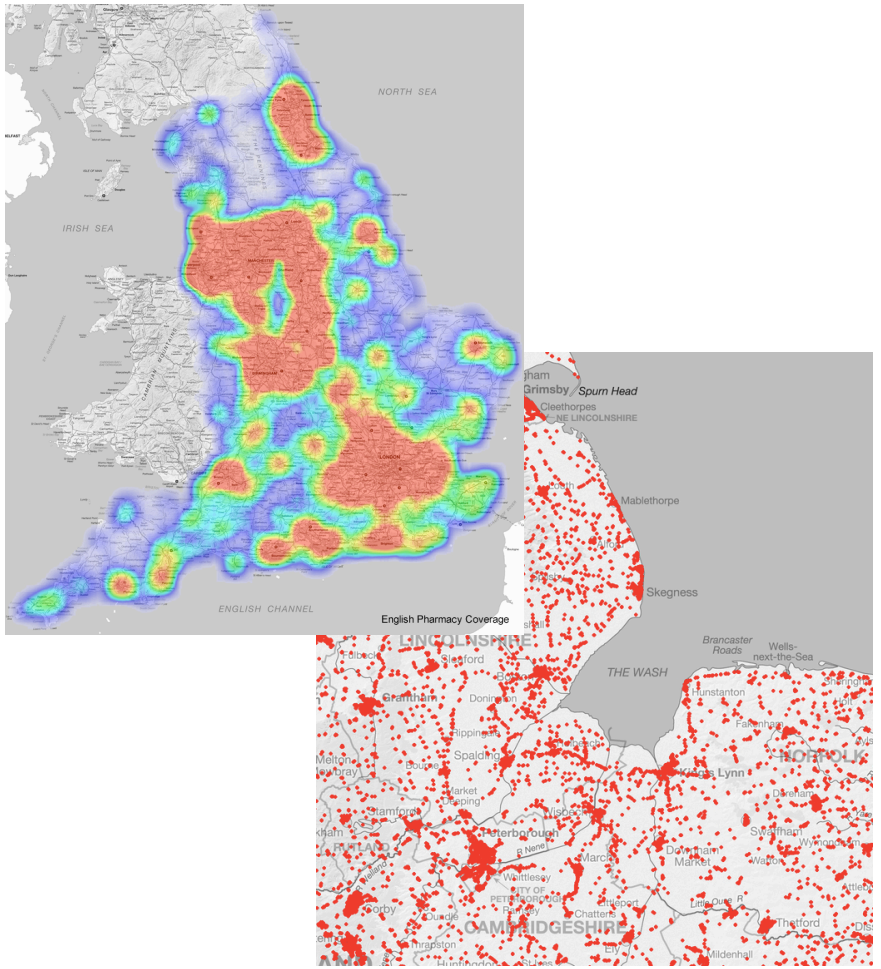
Do you want more data?
Request new data

Early examples of OGD - ASBOrometer



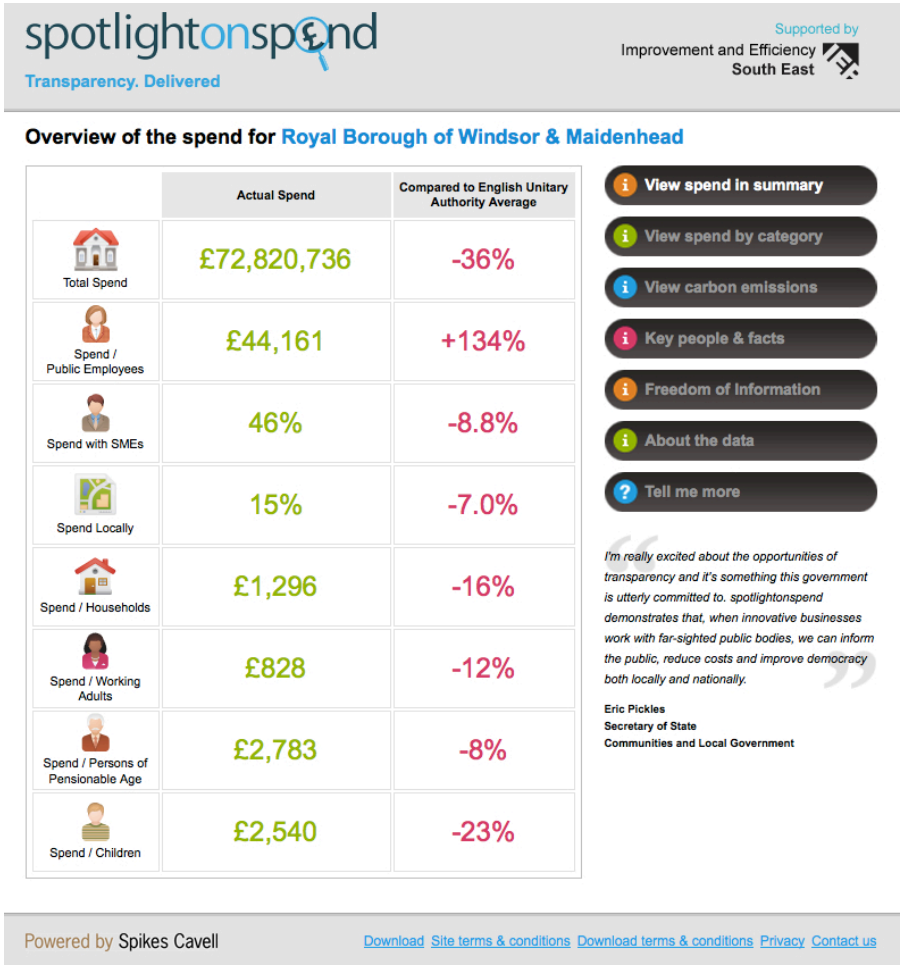
OGD – location, location, location…





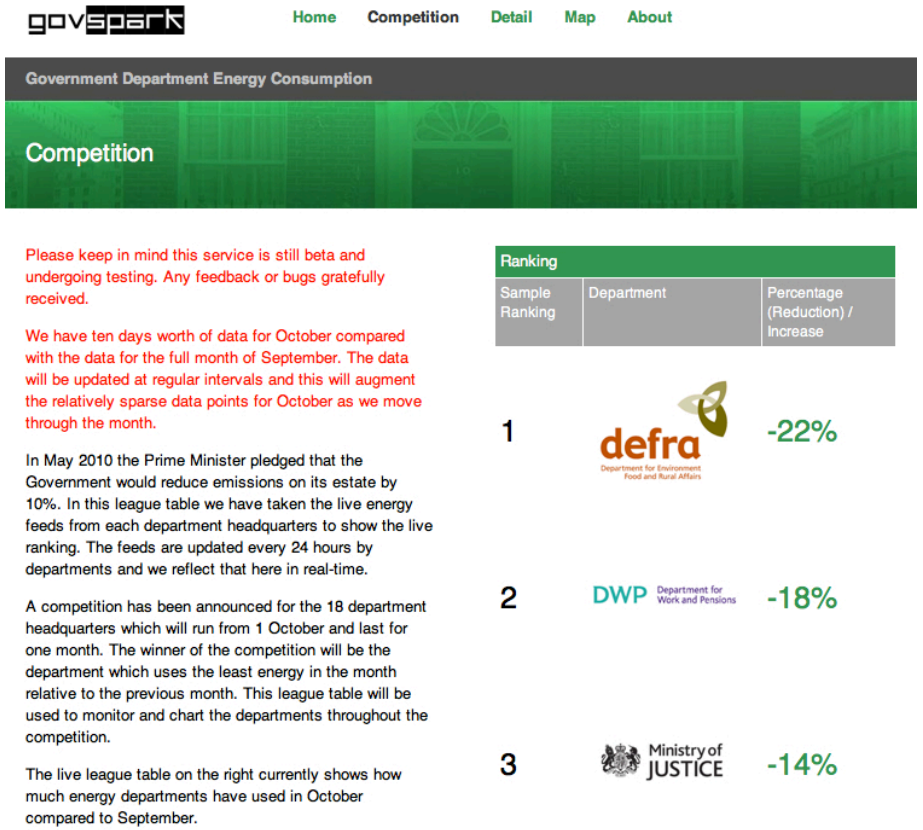
Data to decisions - spending...

New items of Local government spending over £500 – council by council from Jan 2011



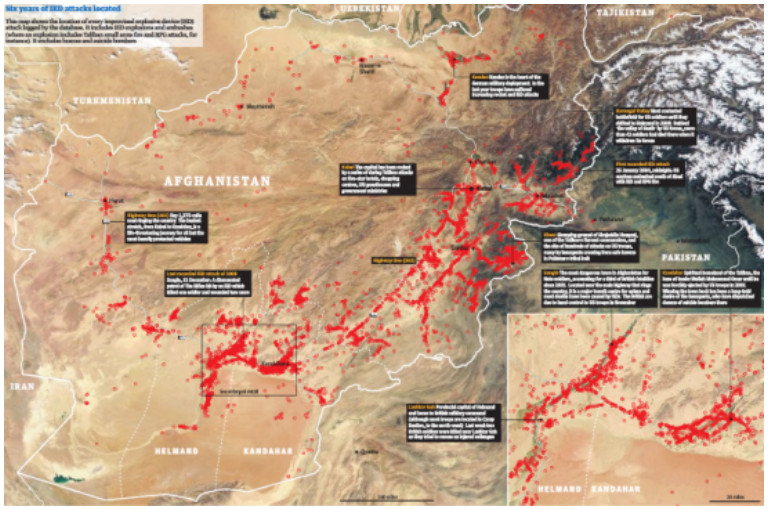
Data to decisions - behaviour...

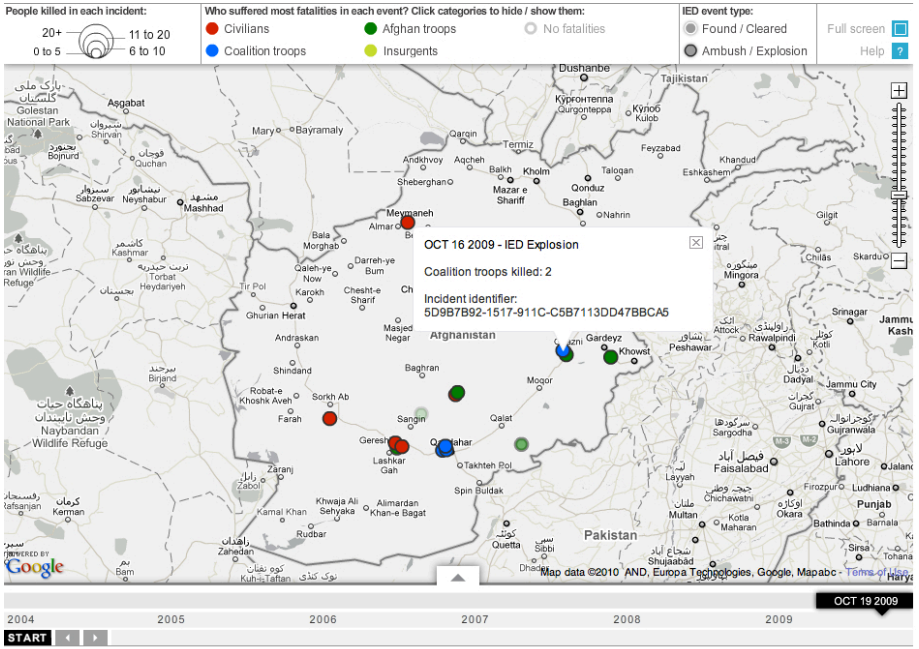
Real time energy data is changing behaviours – the decisions people make



Data to decisions - behaviour...

IEDS from Afghanistan War Logs – raw data available from Guardian



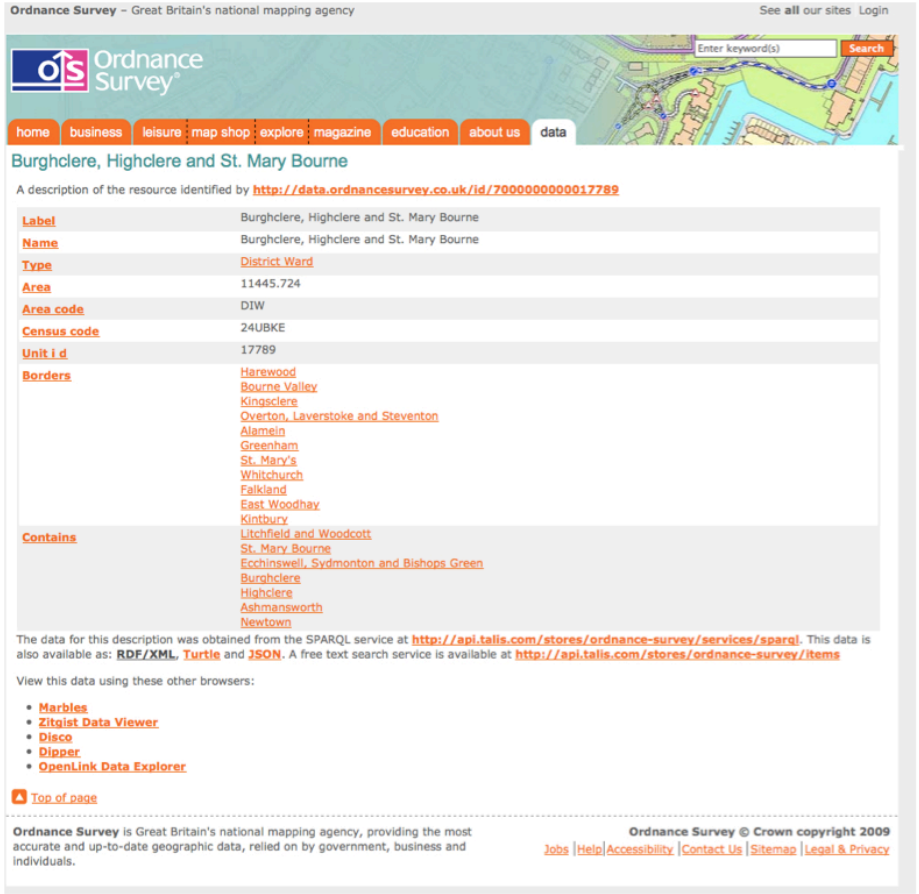


UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

Open Standards

Why 5★ Linked Data?

- National digital infrastructure being built
- URIs for schools, roads, bus stops, post codes, admin boundaries...
- Some of the data links across and connects other data together
- Key data link points exist



Ordnance Survey – Great Britain's national mapping agency

See **all** our sites Login

Ordnance Survey

Enter keyword(s) Search

home | business | leisure | map shop | explore | magazine | education | about us | data

Burghclere, Highclere and St. Mary Bourne

A description of the resource identified by http://data.ordnancesurvey.co.uk/id/7000000000017789

Label	Burghclere, Highclere and St. Mary Bourne
Name	Burghclere, Highclere and St. Mary Bourne
Type	District Ward
Area	11445.724
Area code	DIW
Census code	24UBKE
Unit i d	17789
Borders	Harewood Bourne Valley Kingsclere Overton, Laverstoke and Steventon Alamein Greenham St. Mary's Whitchurch Falkland East Woodhay Kintbury
Contains	Litchfield and Woodcott St. Mary Bourne Ecchinswell, Sydmonton and Bishops Green Burghclere Highclere Ashmansworth Newtown

The data for this description was obtained from the SPARQL service at http://api.talis.com/stores/ordnance-survey/services/sparql. This data is also available as: RDF/XML, Turtle and JSON. A free text search service is available at http://api.talis.com/stores/ordnance-survey/items

View this data using these other browsers:

- Marbles
- Zitgist Data Viewer
- Disco
- Dipper
- OpenLink Data Explorer

Top of page

Ordnance Survey is Great Britain's national mapping agency, providing the most accurate and up-to-date geographic data, relied on by government, business and individuals.

Ordnance Survey © Crown copyright 2009

Jobs | Help | Accessibility | Contact Us | Sitemap | Legal & Privacy

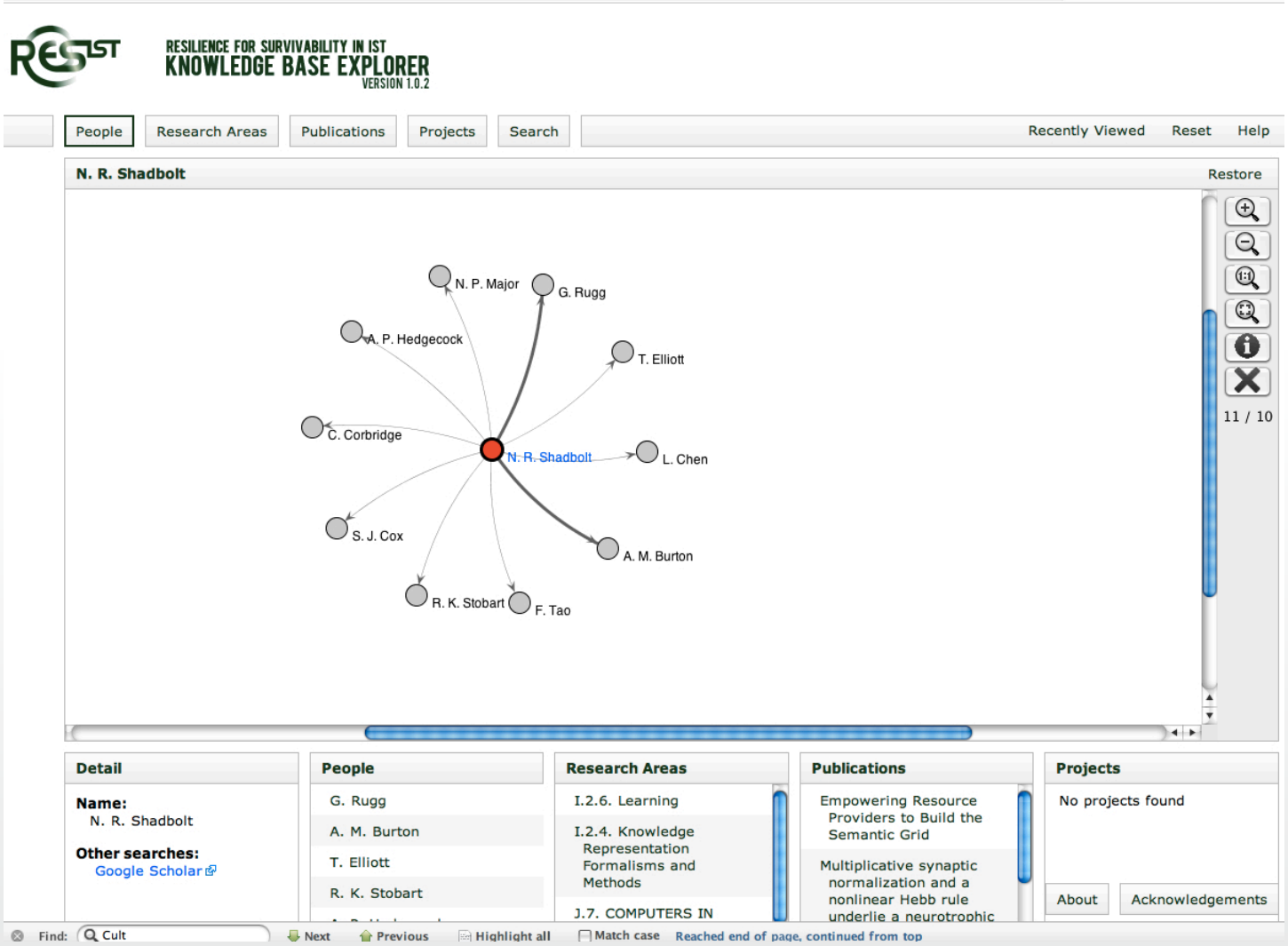
Why 5★ Linked Data?

- National digital infrastructure being built
- URIs for schools, roads, bus stops, post codes, admin boundaries...
- Some of the data links across and connects other data together
- Key data link points exist

http://data.ordnancesurvey.co.uk/doc/7000000000017789.rdf

Litchfield and Woodcott	label	Litchfield and Woodcott
Burghclere, Highclere and St. Mary Bourne	Area	11445.724
	Area Code	DIW
	Census Code	24UBKE
	Unit ID	17789
	borders	Greenham
		Falkland
		Kintbury
		Overton, Laverstoke and Steventon
		Harewood
		St. Mary's
		Whitchurch
		Kingsclere
		East Woodhay
		Bourne Valley
		Alamein
	contains	St. Mary Bourne
		Ecchinswell, Sydmonton and Bishops Green
		Newtown
		Burghclere
		Highclere
		Litchfield and Woodcott
		Ashmansworth
	type	District Ward
	label	Burghclere, Highclere and St. Mary Bourne
	name	Burghclere, Highclere and St. Mary Bourne
Highclere	label	Highclere
Harewood	label	Harewood
Alamein	label	Alamein
Ecchinswell, Sydmonton and Bishops Green	label	Ecchinswell, Sydmonton and Bishops Green
Burghclere	label	Burghclere
Greenham	label	Greenham
St. Mary Bourne	label	St. Mary Bourne
Newtown	label	Newtown
Bourne Valley	label	Bourne Valley
District Ward	label	District Ward
East Woodhay	label	East Woodhay
Falkland	label	Falkland
Kingsclere	label	Kingsclere
Kintbury	label	Kintbury
St. Mary's	label	St. Mary's
Whitchurch	label	Whitchurch
Ashmansworth	label	Ashmansworth
Overton, Laverstoke and Steventon	label	Overton, Laverstoke and Steventon
http://data.ordnancesurvey.co.uk/doc/7000000000017789	Has Format	http://data.ordnancesurvey.co.uk/doc/7000000000017789.html
		http://data.ordnancesurvey.co.uk/doc/7000000000017789.json
		http://data.ordnancesurvey.co.uk/doc/7000000000017789.rdf
		http://data.ordnancesurvey.co.uk/doc/7000000000017789.ttl
	Title	Linked Data for Burghclere, Highclere and St. Mary Bourne
	type	http://purl.org/dc/dcmitype/Text
		Document
	primary topic	Burghclere, Highclere and St. Mary Bourne
http://data.ordnancesurvey.co.uk/doc/7000000000017789.rdf	Format	application/rdf+xml
	Is Format Of	http://data.ordnancesurvey.co.uk/doc/7000000000017789
	Title	Linked Data in RDF/XML format for Burghclere, Highclere and St. Mary Bourne
	type	http://purl.org/dc/dcmitype/Text
		Document
	primary topic	Burghclere, Highclere and St. Mary Bourne
http://data.ordnancesurvey.co.uk/doc/7000000000017789.html	Format	text/html

RKBExplorer: The Power of Linked Data



Linked Data - Technical Principles

The four micro principles of the Semantic Web

1. All entities of interest, such as information resources, real-world objects, and vocabulary terms should be identified by URI references.
2. URI references should be dereferenceable, meaning that an application can look up a URI over the HTTP protocol and retrieve RDF data about the identified resource.
3. Data should be provided using the RDF/XML syntax
4. Data should be interlinked with other data.

The Semantic Web Revisited

Nigel Shadbolt and Wendy Hall, *University of Southampton*
Tim Berners-Lee, *Massachusetts Institute of Technology*

Nigel Shadbolt is a professor of artificial intelligence in the School of Electronics and Computer Science at Southampton University. Contact him at nrs@ecs.soton.ac.uk.

Tim Berners-Lee is the director of the World Wide Web Consortium, a senior researcher at the Massachusetts Institute of Technology's Computer Science and Artificial Intelligence Laboratory, and a professor of computer science in the Department of Electronics and Computer Science at Southampton University. Contact him at timbl@w3.org.

Wendy Hall is a professor of computer science in the School of Electronics and Computer Science at Southampton University. Contact her at wh@ecs.soton.ac.uk.

Principle 1: URIs

- Uniform Resource identifier
- A Web based identifier
- You can dereference them – get something back
- Information about objects, relations, properties and values

http://rdf.ecs.soton.ac.uk/person/2686

http://rdf.ecs.soton.ac.uk/project/464

http://rdf.ecs.soton.ac.uk/publication/11065

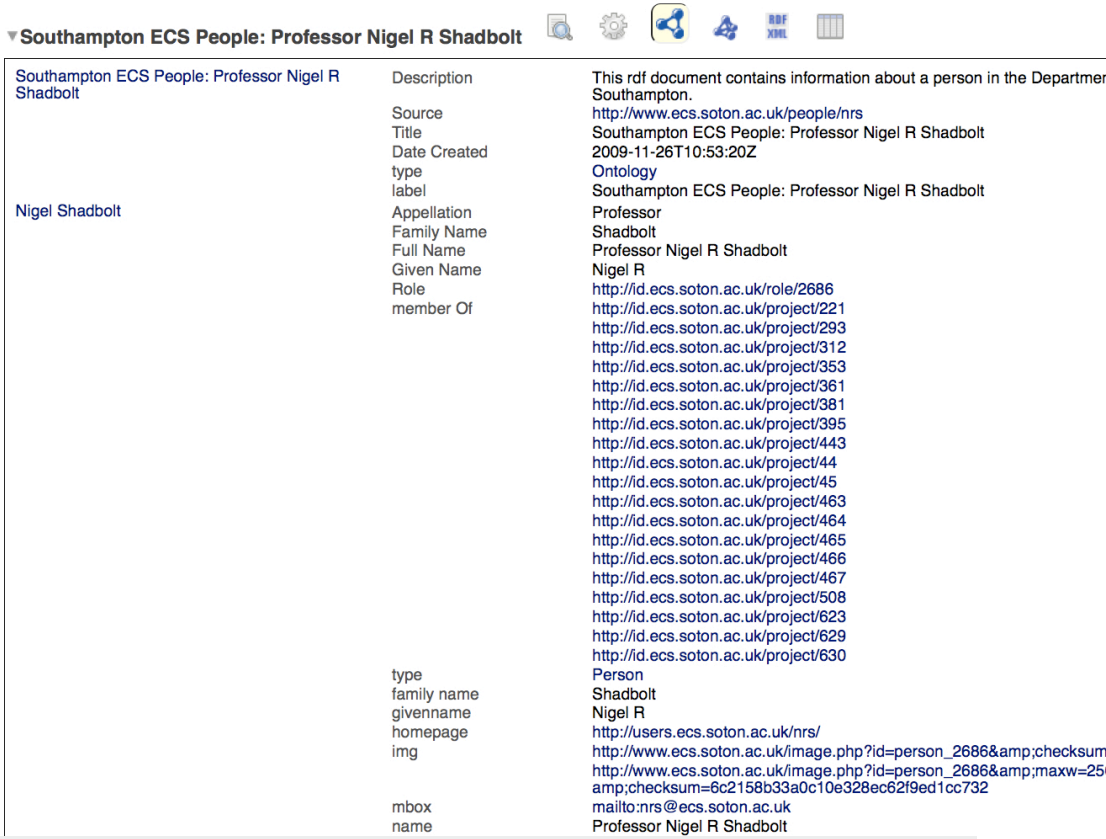
http://education.data.gov.uk/doc/school/120805

http://southampton.rkbexplorer.com/id/person-02686

http://dbpedia.org/resource/Nigel_Shadbolt

Principle 2: Dereferencable

- Uniform Resource identifier
- A Web based identifier
- You can dereference them – get something back
- Information about objects, relations, properties and values

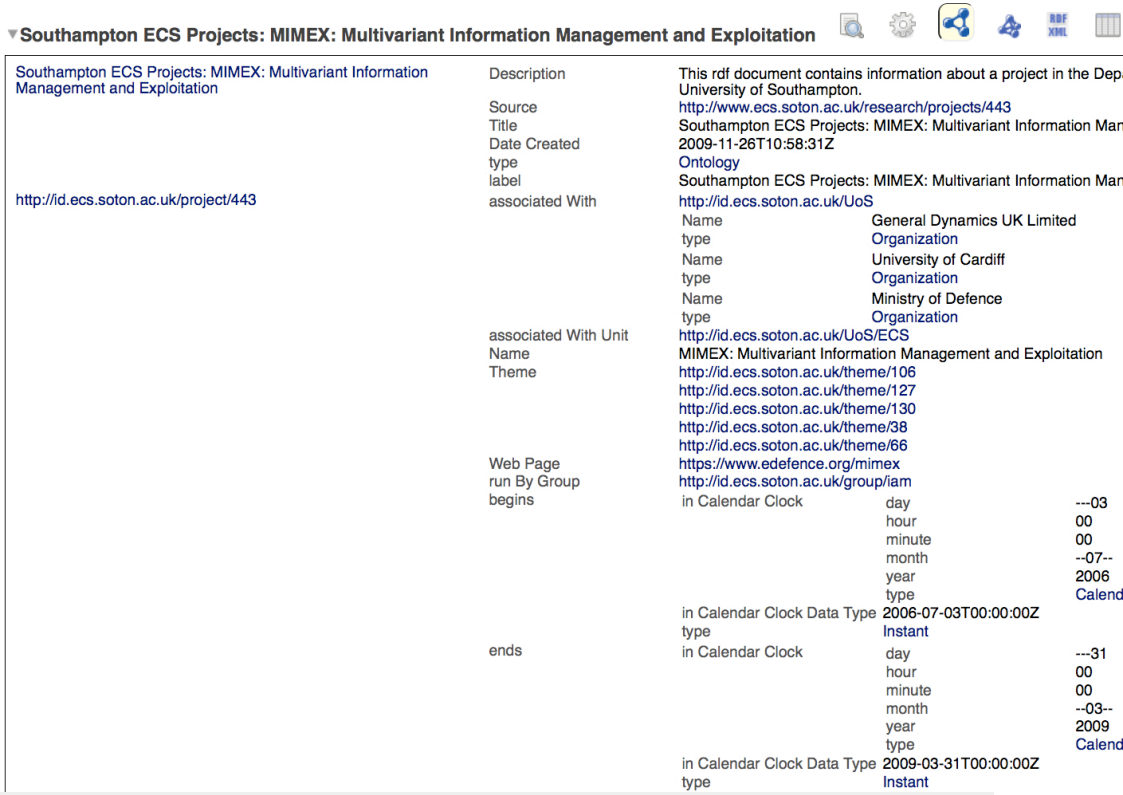


Southampton ECS People: Professor Nigel R Shadbolt

Southampton ECS People: Professor Nigel R Shadbolt	Description	This rdf document contains information about a person in the Departmer Southampton.
	Source	http://www.ecs.soton.ac.uk/people/nrs
	Title	Southampton ECS People: Professor Nigel R Shadbolt
	Date Created	2009-11-26T10:53:20Z
	type	Ontology
	label	Southampton ECS People: Professor Nigel R Shadbolt
Nigel Shadbolt	Appellation	Professor
	Family Name	Shadbolt
	Full Name	Professor Nigel R Shadbolt
	Given Name	Nigel R
	Role	http://id.ecs.soton.ac.uk/role/2686
	member Of	http://id.ecs.soton.ac.uk/project/221
		http://id.ecs.soton.ac.uk/project/293
		http://id.ecs.soton.ac.uk/project/312
		http://id.ecs.soton.ac.uk/project/353
		http://id.ecs.soton.ac.uk/project/361
		http://id.ecs.soton.ac.uk/project/381
		http://id.ecs.soton.ac.uk/project/395
		http://id.ecs.soton.ac.uk/project/443
		http://id.ecs.soton.ac.uk/project/44
		http://id.ecs.soton.ac.uk/project/45
		http://id.ecs.soton.ac.uk/project/463
		http://id.ecs.soton.ac.uk/project/464
		http://id.ecs.soton.ac.uk/project/465
		http://id.ecs.soton.ac.uk/project/466
		http://id.ecs.soton.ac.uk/project/467
		http://id.ecs.soton.ac.uk/project/508
		http://id.ecs.soton.ac.uk/project/623
		http://id.ecs.soton.ac.uk/project/629
		http://id.ecs.soton.ac.uk/project/630
	type	Person
	family name	Shadbolt
	givenname	Nigel R
	homepage	http://users.ecs.soton.ac.uk/nrs/
	img	http://www.ecs.soton.ac.uk/image.php?id=person_2686&checksum
		http://www.ecs.soton.ac.uk/image.php?id=person_2686&maxw=25 amp;checksum=6c2158b33a0c10e328ec62f9ed1cc732
	mbox	mailto:nrs@ecs.soton.ac.uk
	name	Professor Nigel R Shadbolt

Principle 3: RDF back

- Uniform Resource identifier
- A Web based identifier
- You can dereference them – get something back
- Web Knowledge Representation called RDF
- Information about objects, relations, properties and values



Southampton ECS Projects: MIMEX: Multivariant Information Management and Exploitation

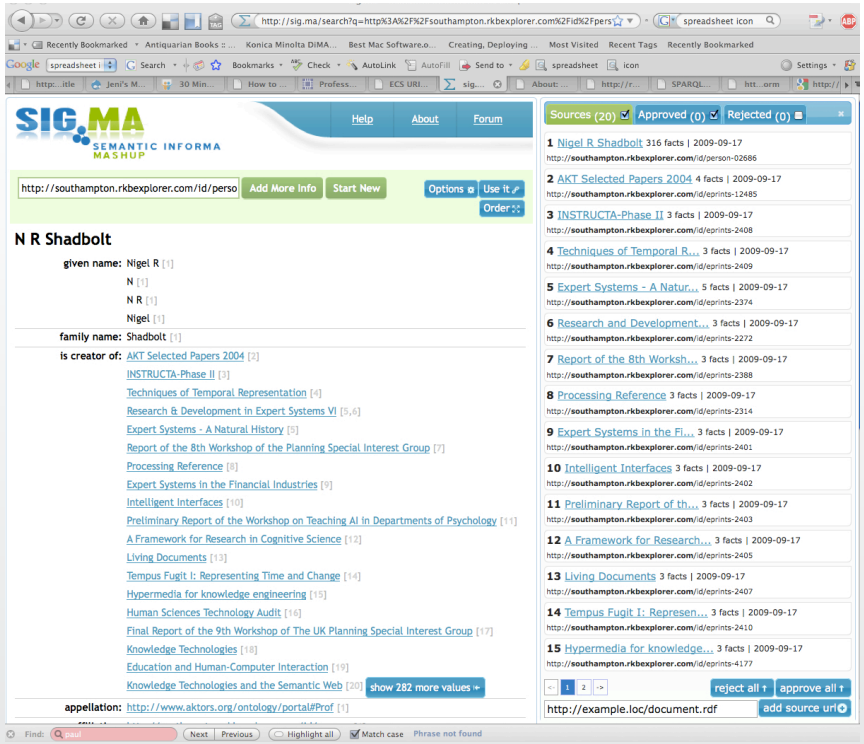
Southampton ECS Projects: MIMEX: Multivariant Information Management and Exploitation	Description	This rdf document contains information about a project in the Dep University of Southampton.	
	Source	http://www.ecs.soton.ac.uk/research/projects/443	
	Title	Southampton ECS Projects: MIMEX: Multivariant Information Mar	
	Date Created	2009-11-26T10:58:31Z	
	type	Ontology	
	label	Southampton ECS Projects: MIMEX: Multivariant Information Mar	
http://id.ecs.soton.ac.uk/project/443	associated With	http://id.ecs.soton.ac.uk/UoS	
		Name	General Dynamics UK Limited
		type	Organization
		Name	University of Cardiff
		type	Organization
		Name	Ministry of Defence
		type	Organization
	associated With Unit	http://id.ecs.soton.ac.uk/UoS/ECS	
	Name	MIMEX: Multivariant Information Management and Exploitation	
	Theme	http://id.ecs.soton.ac.uk/theme/106	
		http://id.ecs.soton.ac.uk/theme/127	
		http://id.ecs.soton.ac.uk/theme/130	
		http://id.ecs.soton.ac.uk/theme/38	
		http://id.ecs.soton.ac.uk/theme/66	
	Web Page	https://www.edefence.org/mimex	
	run By Group	http://id.ecs.soton.ac.uk/group/iam	
	begins	in Calendar Clock	day ---03
			hour 00
			minute 00
			month --07--
			year 2006
			type Calend
		in Calendar Clock Data Type	2006-07-03T00:00:00Z
		type	Instant
	ends	in Calendar Clock	day ---31
			hour 00
			minute 00
			month --03--
			year 2009
			type Calend
		in Calendar Clock Data Type	2009-03-31T00:00:00Z
		type	Instant

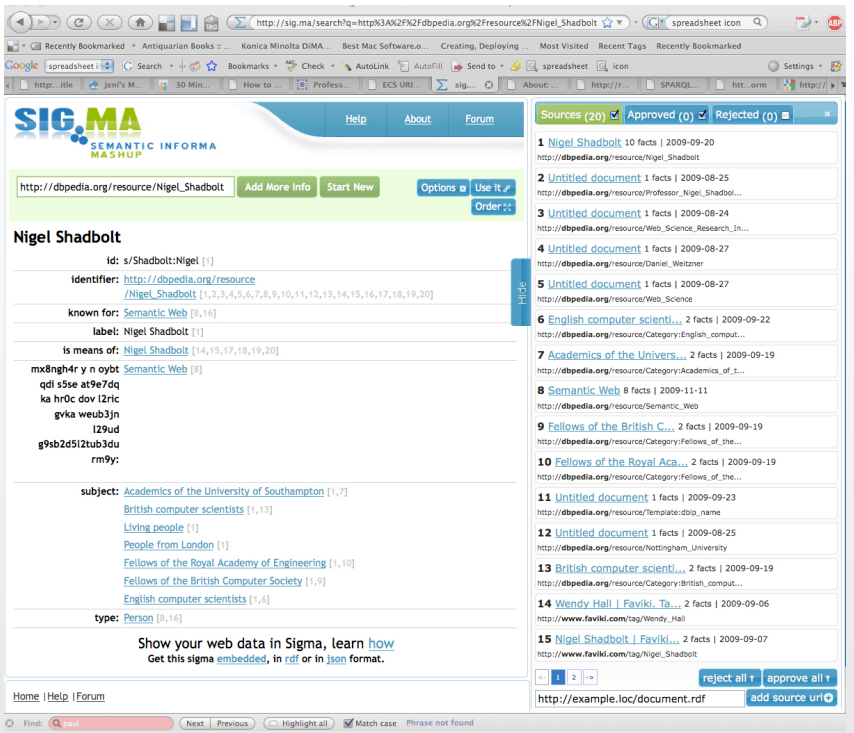
Principle 4: Link data

http://rdf.ecs.soton.ac.uk/person/2686 SameAs

http://southampton.rkbexplorer.com/id/person-02686 SameAs

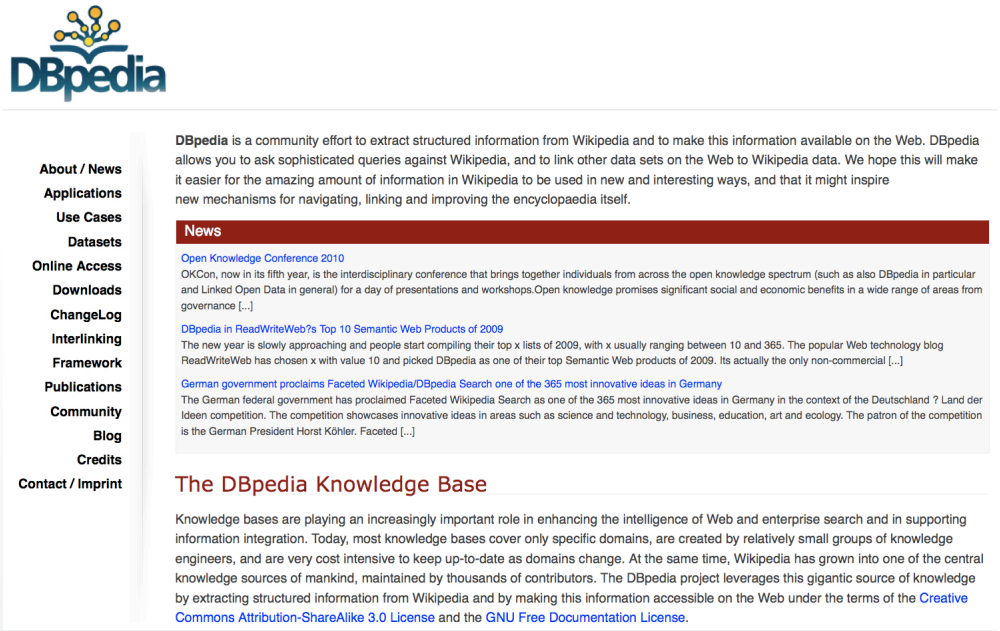
http://dbpedia.org/resource/Nigel_Shadbolt





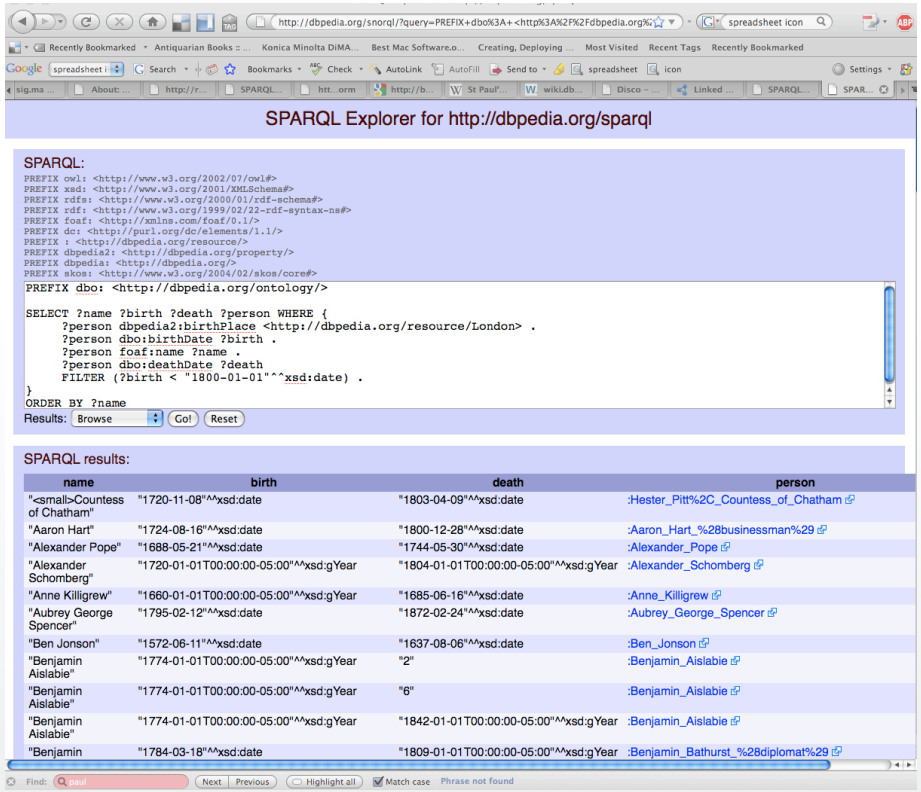
Wikipedia as Data: DBpedia

- 2.9 million things
- > 282,000 persons
- 339,000 places
- 88,000 music albums
- 44,000 films
- 15,000 video games
- 119,000 organizations
- 130,000 species
- 4400 diseases.

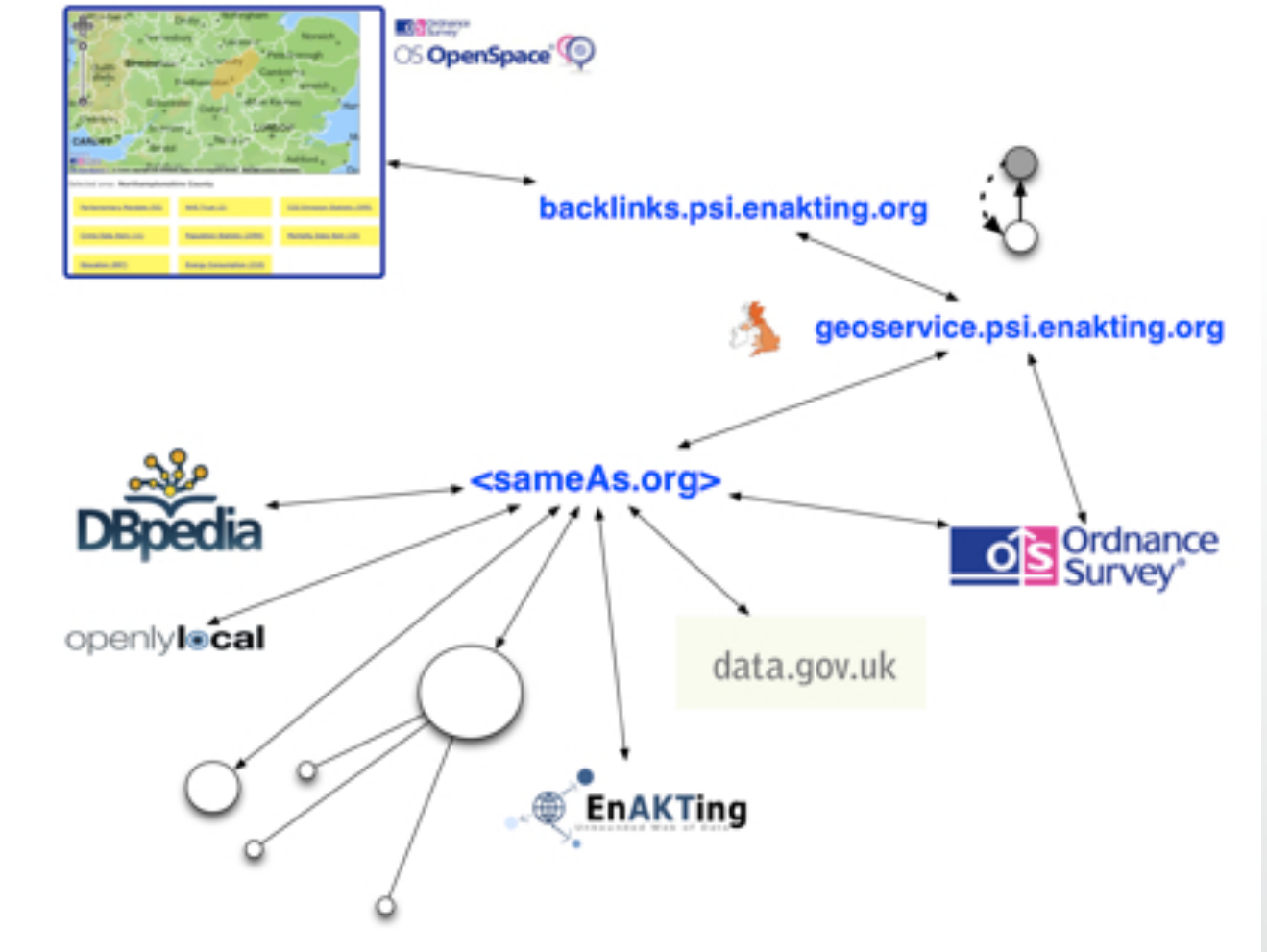


SPARQL

- A data access language for RDF on the Web
- Queries across diverse data sources
- SPARQL contains capabilities for querying required and optional patterns



Joining up the dots…
http://map.psi.enakting.org/how



Joining up the dots – UK Linked Data
http://myarea.psi.enakting.org/

How good is my area ?

feedback ... please email us

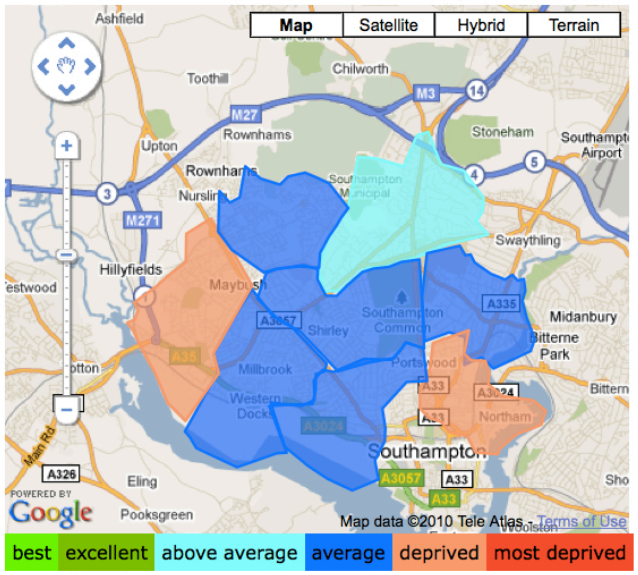
Sources:
statistics.data.gov.uk, data.ordnancesurvey.co.uk,
IMD Database, Boundary data from OS Open Data

This is a **BETA** app developed in a EnAKTing hack day. It is built on top of Semantic Web technologies and Linked Data.

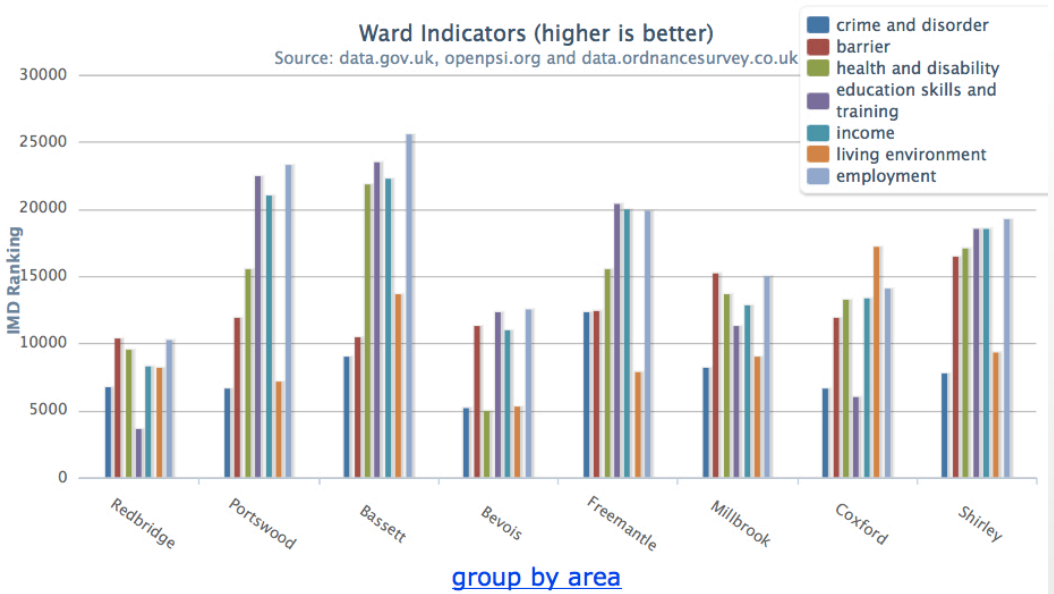
It displays a region and its surroundings showing how good are those areas based on the IMD database ranking.

shirley find

introduce an area name i.e: Portswood, Chilworth, Camden, ...



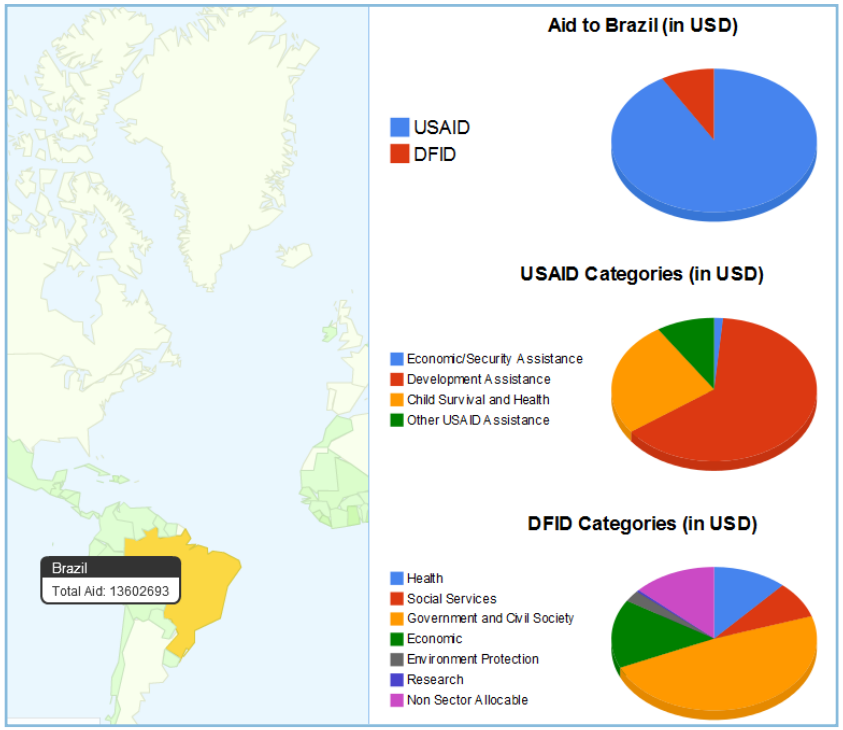
Area: Bevois

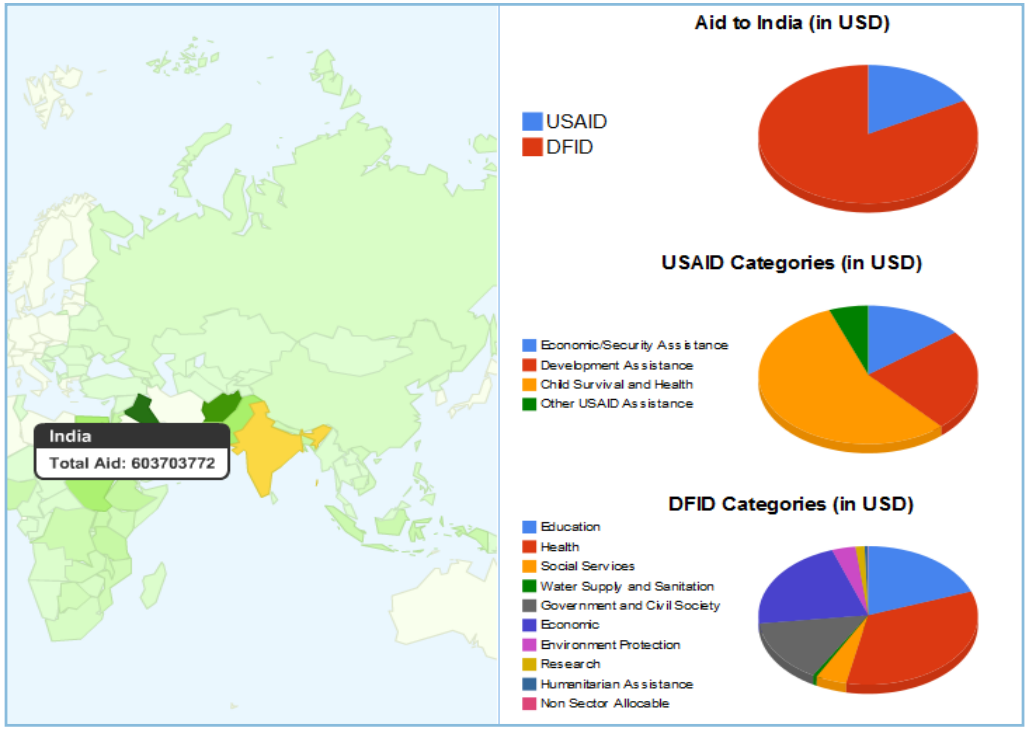


group by area

US and UK Foreign AID FY2007:

Integrating data from two countries





Data Sources:

USAID FROM THE AMERICAN PEOPLE

UKaid from the Department for International Development

	AID	Major aids from US	Major aids from UK
Brazil	US >UK	Development Assistance	Gov & civil society, Economic
India	UK > US	Child Survival and Health	Health, Economic

[Spatial Mashup] Data.gov (USAID) + Data.gov.uk (DFID)

Being extended to allow user annotations

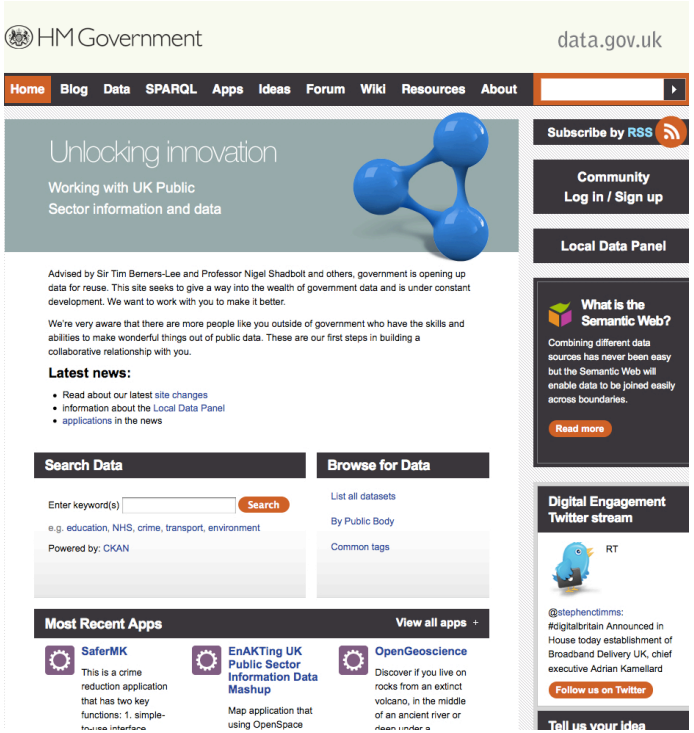
Created by James Michaelis, RPI, http://data-gov.tw.rpi.edu/demo/linked/aidviz-1554-10030.html

UNIVERSITY OF
Southampton
School of Electronics
and Computer Science

Open Source

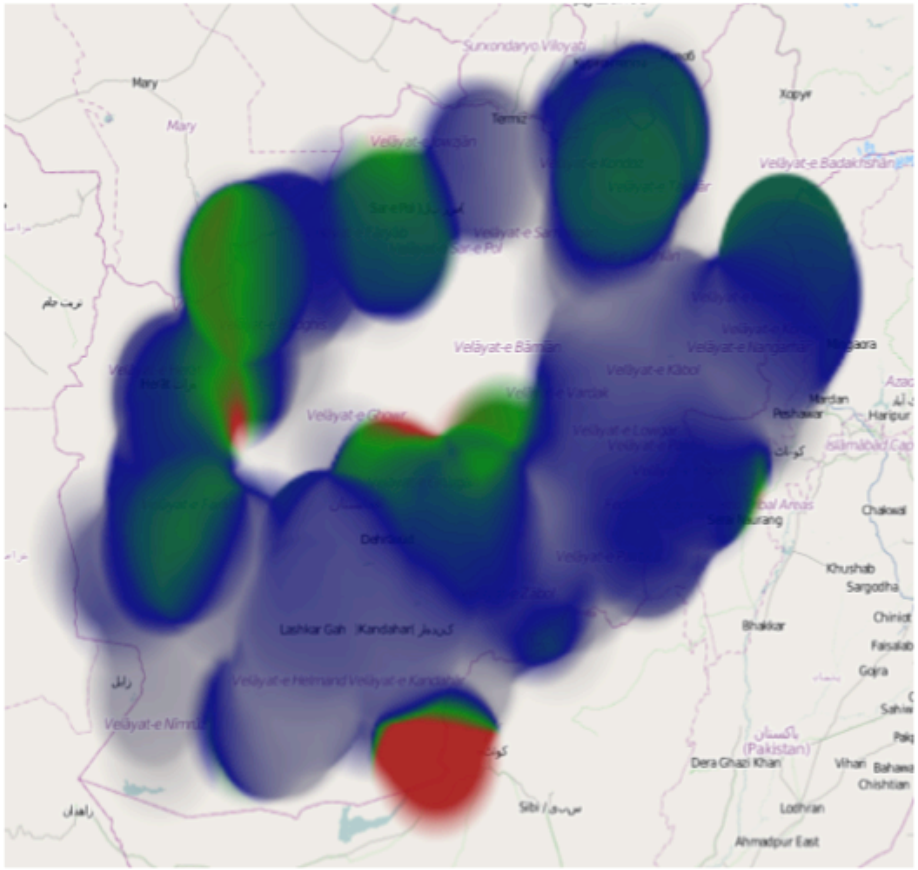
data.gov.uk

- Drupal and Media Wiki
- Cheap
- Robust
- Resilient
- Maintained



IED Data

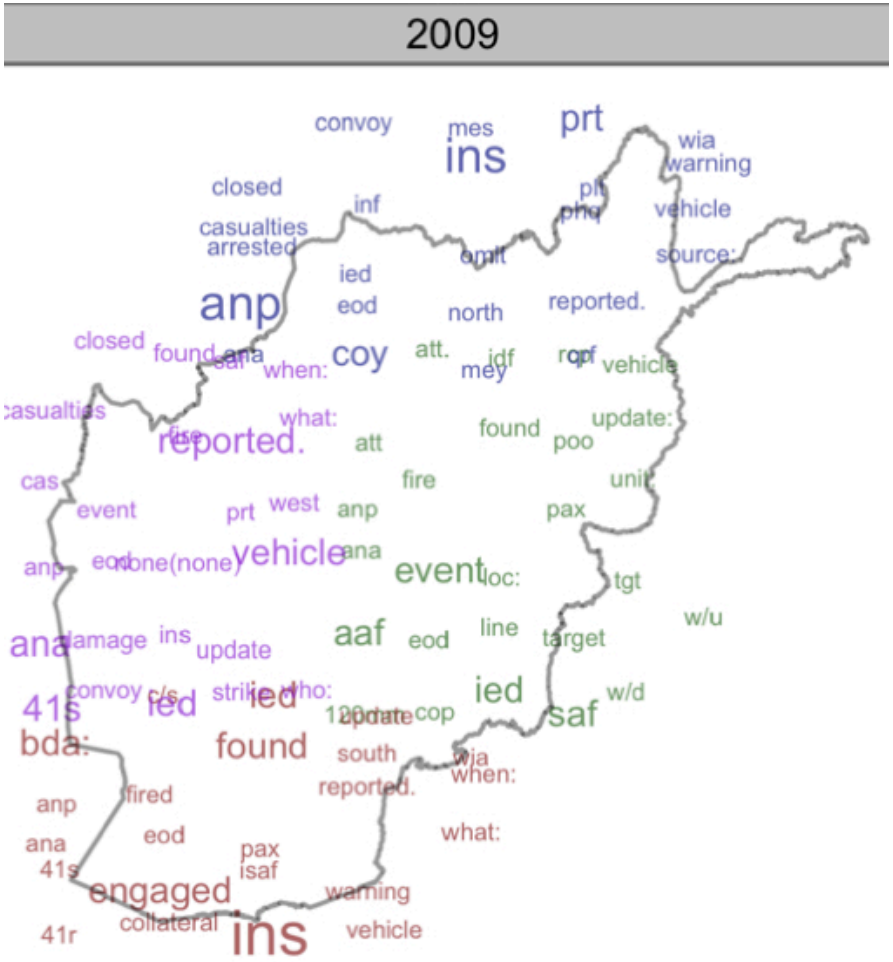
http://www.wired.com/dangerroom/2010/08/open-source-wikileaked-docs-illustrated-afghan-meltdown/



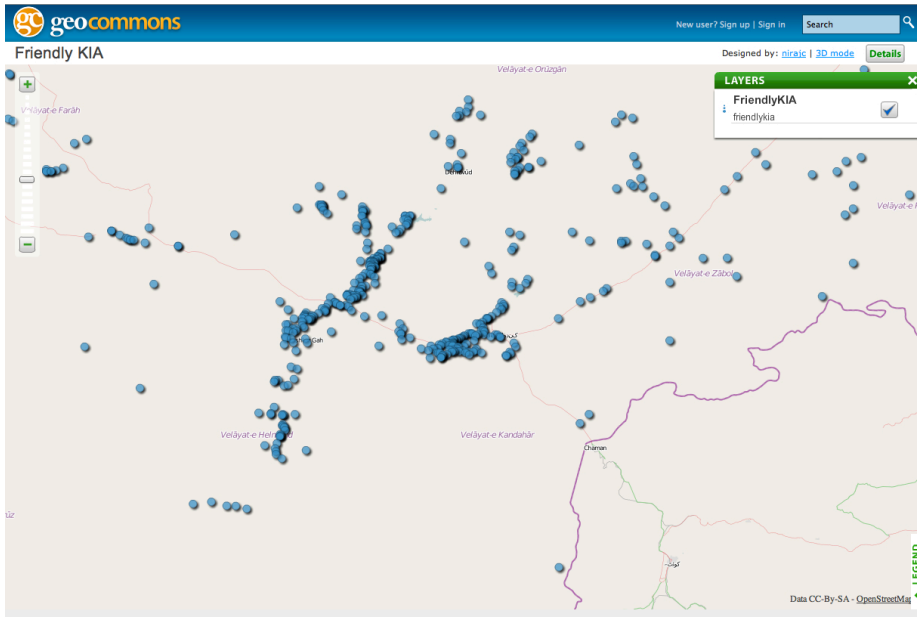
IED Data

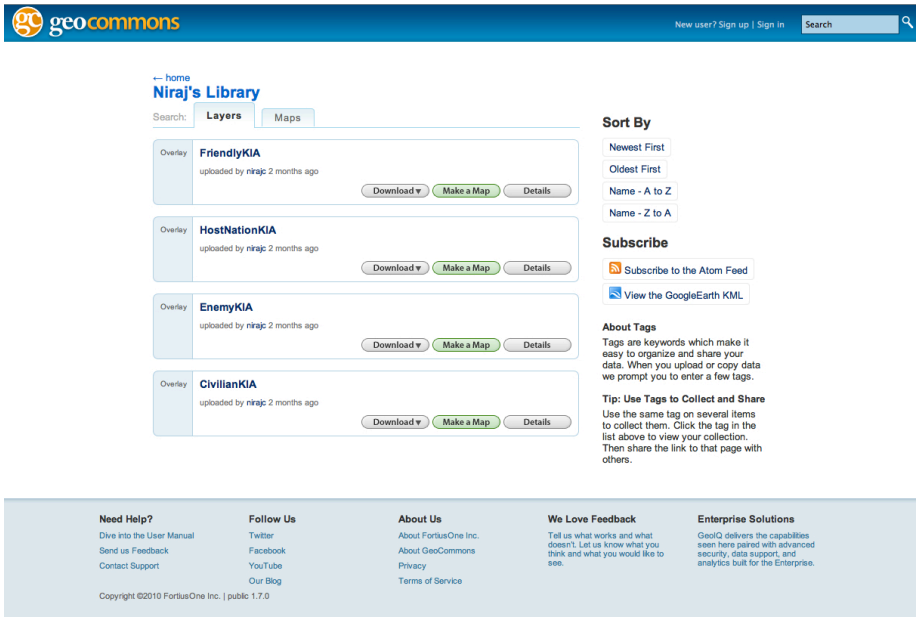
100644 | 227 lines (200 sloc) | 9.271 kb raw | blame | histor

```
# File-Name:       wikileaks_tm.R
# Date:            2010-08-11
# Author:          Drew Conway
# Email:           drew.conway@nyu.edu
# Purpose:         Perform basic text mining operations on WL data
# Data Used:       see load_data.R
# Packages Used:
# Output File:
# Data Output:
# Machine:         Drew Conway's MacBook Pro

# Copyright (c) 2010, under the Simplified BSD License.
# For more information on FreeBSD see: http://www.opensource.org/licenses/bsd-license.php
# All rights reserved.

source("load_data.R")
source("utils.R")

## LDA topic models
library(topicmodels)

# Create corpus from Summary column for all regions
create.lda<-function(afg.df, num.topics=3, num.words=10, stemming=FALSE, stopwords = TRUE, minWordLength = 3) {
    # Create nested list of Summary topic models for all Year/Region pairs
    ## WARNING: with above dtm.control settings this takes a VERY LONG TIME
    ## To speed the process set stemming=FALSE and stopwords=FALSE, however,
    ## results will be different/worse.
    afg.terms<-list()
    regions<-levels(as.factor(afg.df$Region))
    years<-levels(as.factor(afg.df$Year))
    dtm.control<-list(stemming = stemming, stopwords = stopwords, minWordLength = minWordLength)
    for(y in years){
        y.sub<-subset(afg.df,afg.df$Year==y)
        for(r in regions){
            r.sub<-subset(y.sub,y.sub$Region==r)
            yr.corp<-Corpus(VectorSource(r.sub$Summary))
            dtm<-DocumentTermMatrix(yr.corp,control=dtm.control)
            dtm<-removeSparseTerms(dtm,0.95)
            afg.terms[[y]][[r]]<-get_terms(LDA(dtm,control = list(alpha = 0.05), k = num.words),num.topics)
        }
        names(afg.terms[[y]])<-regions
    }
    names(afg.terms)<-years
    return(afg.terms)
}

## Only run this portion if you have not yet generated LDA
afg.terms<-create.lda(afg,stopwords=c(stopwords("english"),"report","reported","reports","report)","(delayed","(s//rel","ro
```



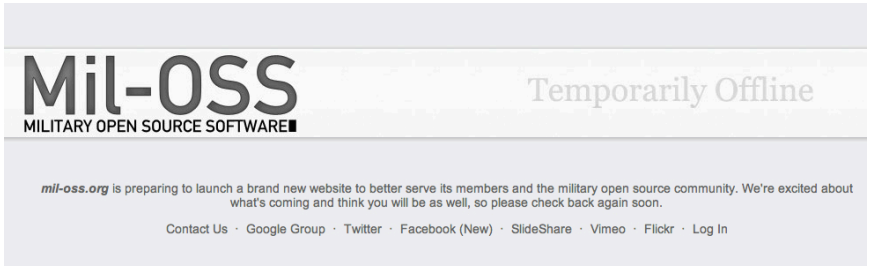
Geo Commons


geocommons
New user? Sign up | Sign in
Search
Friendly KIA
Designed by: niraic | 3D mode
Details
LAYERS
FriendlyKIA
friendlykia
Velāyat-e Orūzgān
Velāyat-e Farāh
Velāyat-e Zābol
Velāyat-e Kandahār
Chaman
Data CC-By-SA - OpenStreetMap


geocommons
New user? Sign up | Sign in
Search
← home
Niraj's Library
Search:
Layers
Maps
Overlay
FriendlyKIA
uploaded by niraic 2 months ago
Download
Make a Map
Details
HostNationKIA
uploaded by niraic 2 months ago
EnemyKIA
uploaded by niraic 2 months ago
CivilianKIA
uploaded by niraic 2 months ago
Sort By
Newest First
Oldest First
Name - A to Z
Name - Z to A
Subscribe
Subscribe to the Atom Feed
View the GoogleEarth KML
About Tags
Tags are keywords which make it easy to organize and share your data. When you upload or copy data we prompt you to enter a few tags.
Tip: Use Tags to Collect and Share
Use the same tag on several items to collect them. Click the tag in the list above to view your collection. Then share the link to that page with others.
Need Help?
Dive into the User Manual
Send us Feedback
Contact Support
Follow Us
Twitter
Facebook
YouTube
Our Blog
About Us
About FortiusOne Inc.
About GeoCommons
Privacy
Terms of Service
We Love Feedback
Tell us what works and what doesn't. Let us know what you think and what you would like to see.
Enterprise Solutions
GeoIQ delivers the capabilities seen here paired with advanced security, data support, and analytics built for the Enterprise.
Copyright ©2010 FortiusOne Inc. | public 1.7.0

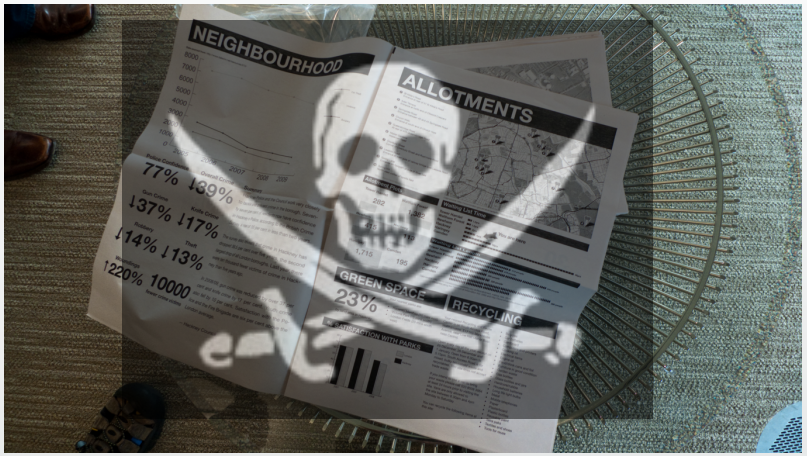
Emerging DoD Interest

- http://www.forge.mil/



UNIVERSITY OF Southampton
School of Electronics and Computer Science

Open Licence



Open Licencing

- OPEN SOURCE AND THESE UNITED STATES
- Major Justin Seiferth

  http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA398898

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Open Minds

Challenges

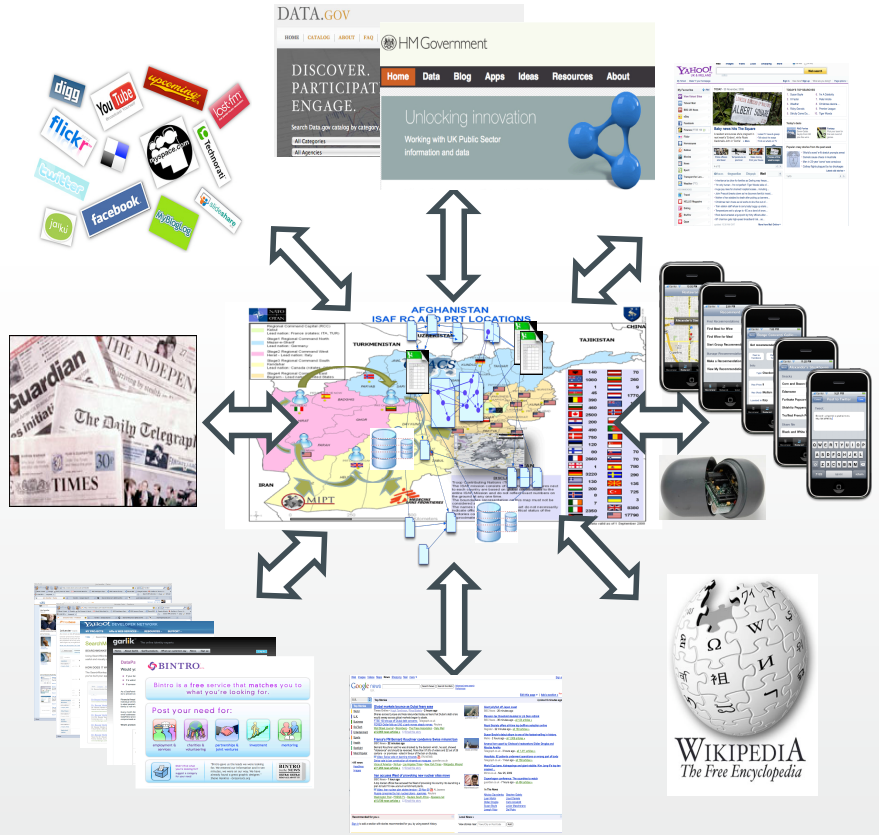
- Cultural & Organisational
  - Data hugging
  - We have released it - sort of…
  - Licence impediments
- Worries about
  - Confidentiality
  - How it will be Interpreted
  - QoI
  - New IT procurement
- Capability
  - What assets have we got
  - Who can do this work
  - How to share best practice
- Policy
  - Adopting and extending the principles
  - Consistency in what is published and its format
  - Range of exceptions do need to be appreciated
  - Confidentiality, accuracy, liability and reputation

Web of Linked Data

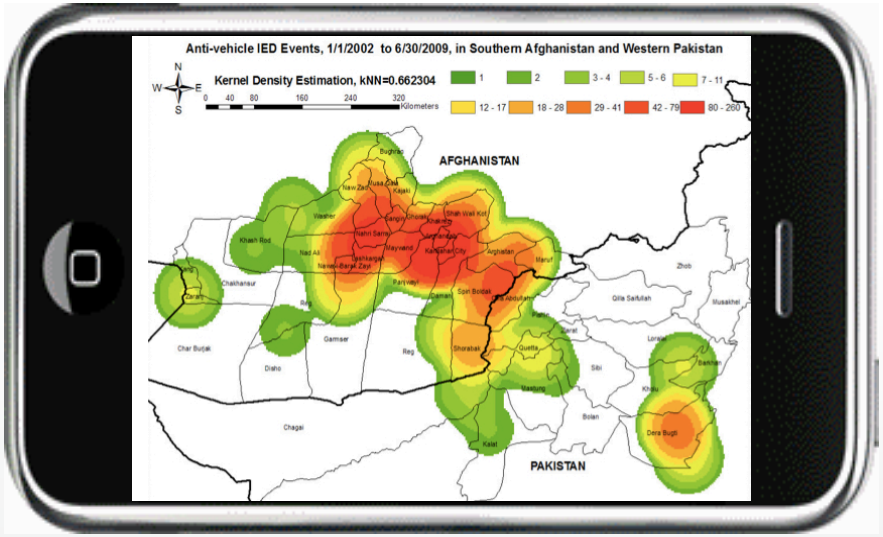
- The Web of Linked Data is next
- Can change
  - Government
  - Business
  - Cities
  - Universities
  - :
- Exploitation of
  - open data
  - open standards
- Can apply to data intranets as well as extranets

Why does this matter to you

- Supports information interoperability
- Sum is greater than the parts
- Exploitation of
  - open data
  - open standards
  - open source
  - open licence
- Can apply to data intranets as well as extranets
- Military and intelligence contexts are classic examples where there is heterogeneous information that could be shared, linked, mapped and enriched using this approach



Challenge – this in 2 days?



Challenge – this in 2 days?

Thank you

- Colleagues with me to also answer questions!
- Any further information or follow ups regarding research please email

  nrs@ecs.soton.ac.uk
- Any further information or follow ups regarding application development or consultancy please email
- tim.organ@seme4.com

www.seme4.com