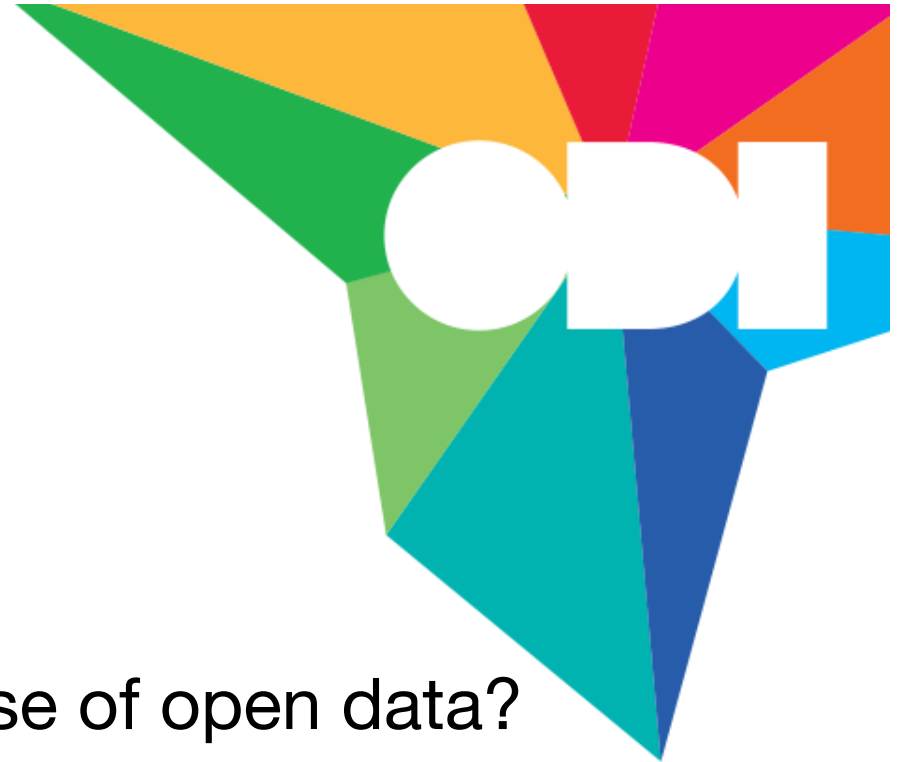




<http://training.theodi.org/InPractice>

David Tarrant · @davetaz

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



Agenda - Today

The characteristics of data

Data discovery patterns

*** Lunch ***

Data publication platforms

Quick big data break

Practical publication hands-on



Recap session

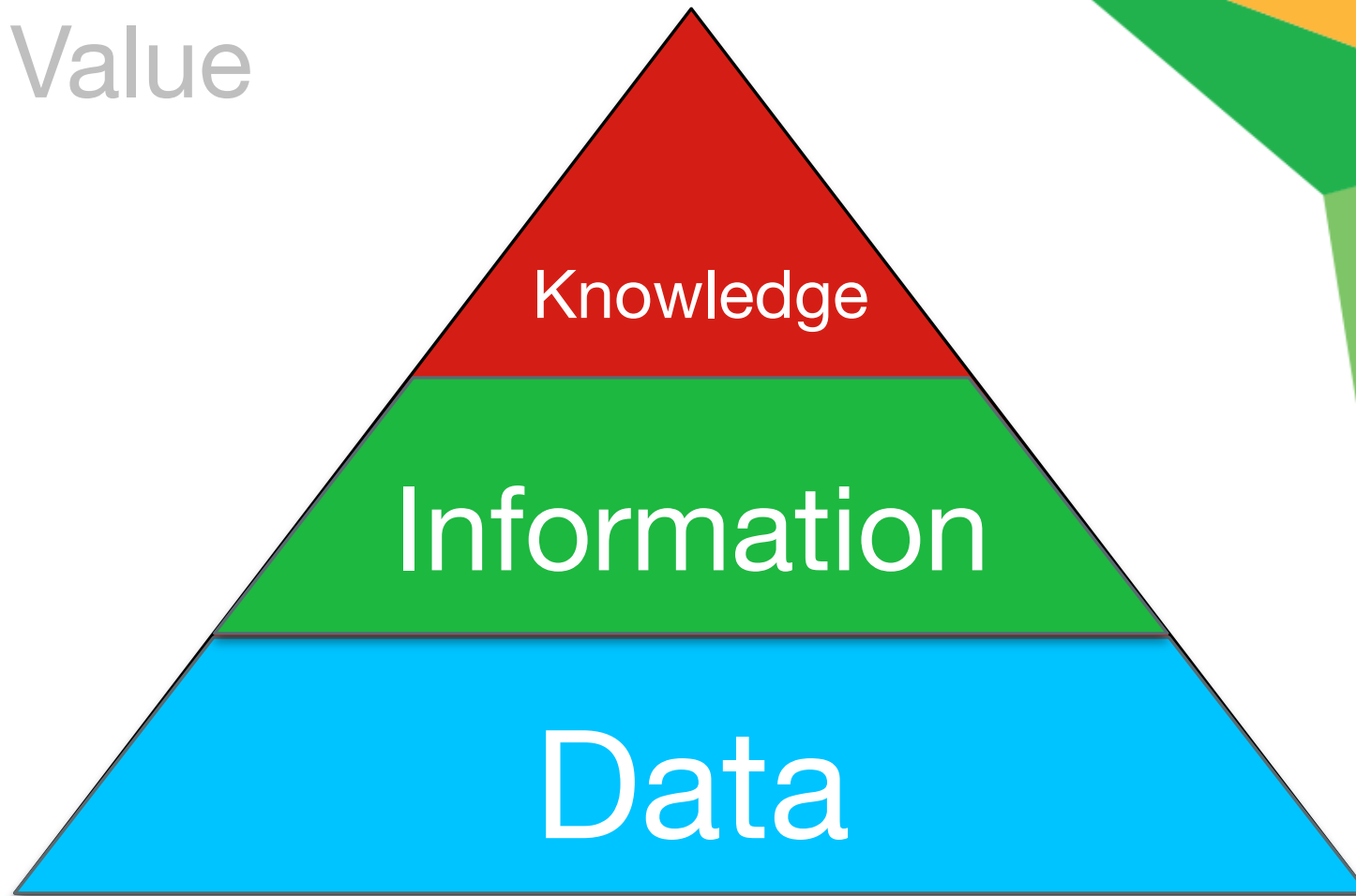


Exercise

What is Data?



Value



Exercise

What is Open Data?



Option A

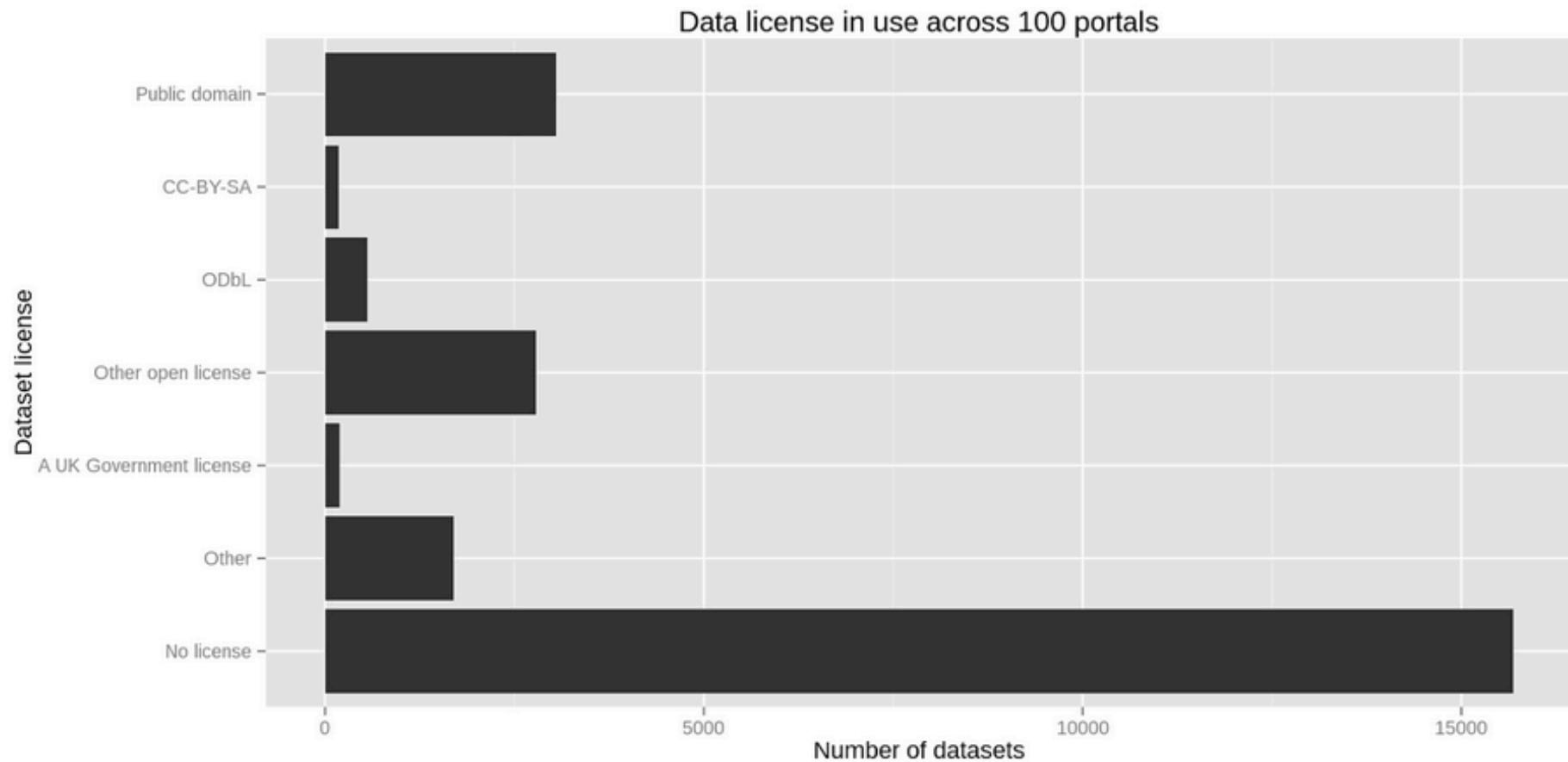
Open data is data that is made
available by **organisations**,
businesses and **individuals** for
anyone to **access**, **use** and **share**.

- 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									✓	

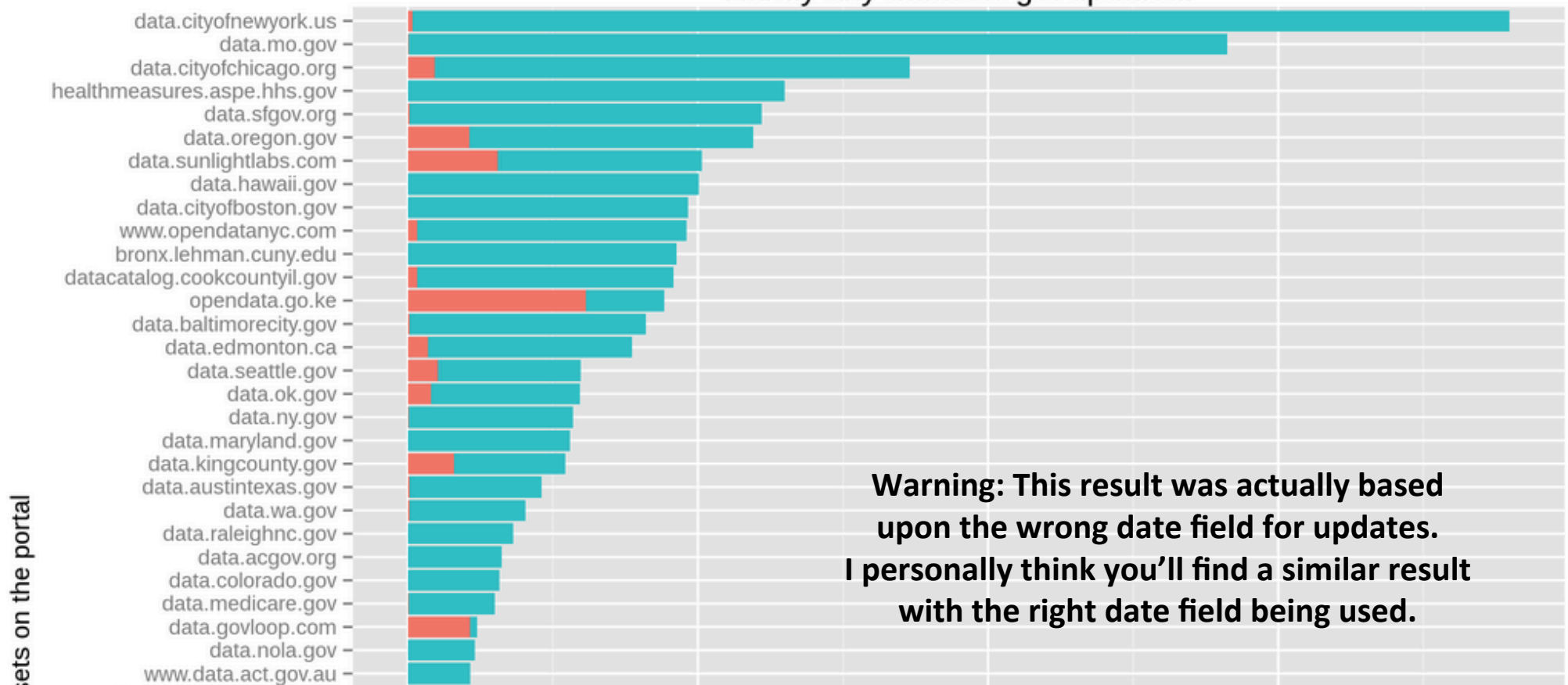
Open data is hardly ever appropriately licensed.



Source: Thomas Levine

Source: Thomas Levine

Hardly any datasets get updated.

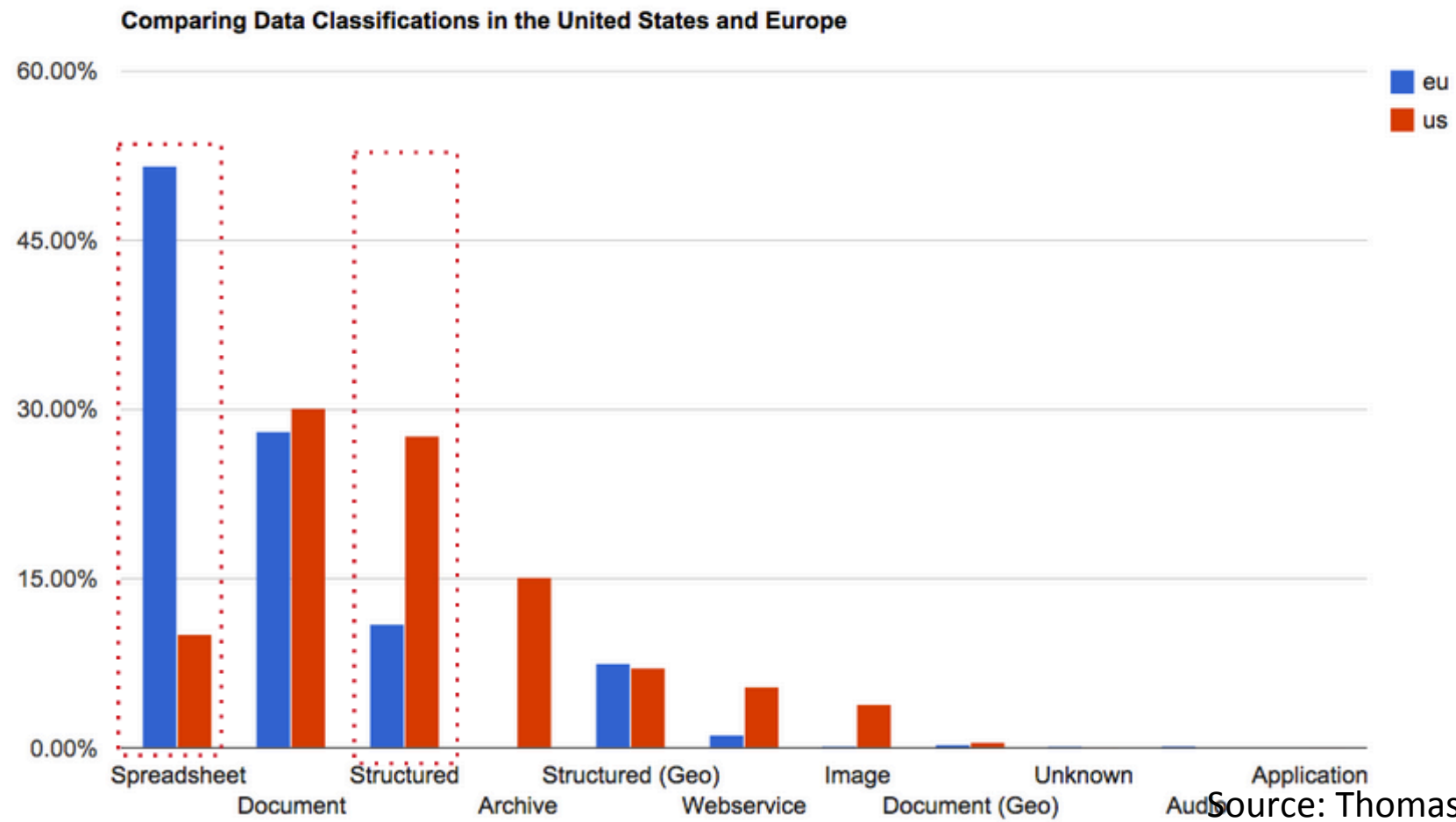


Warning: This result was actually based upon the wrong date field for updates. I personally think you'll find a similar result with the right date field being used.



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.

What best practice guidelines and tools will help us achieve this phase 2 goal?



Guidelines



5-Stars

5-Stars



 <http://5stardata.info/>



<http://data...>



OPEN DATA



Open Data Certificate



<http://certificates.theodi.org>

Introducing Open Refine

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

Google Refine: government IT contracts

Mass edit 2350 cells in column Type of Contract Undo

Facet / Filter Undo / Redo x

5200 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Freebase +

Type of Contract

591 choices Sort by: name count cluster

Agreement 30
HITSS Task Order 32
CPAF 39
T&M w/ FFP: Time & Materials w/
Firm Fixed Price mix 29
Time and Material 37
Firm Fixed 28
Term 25
Firm Fixed Price 34
Labor Hours 21
Seed 31
FFP LDE: Firm Fixed Price Level

	Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	End Date	Total value of Contract	Contract Awarded
1.	1838	ADAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/2008	04/01/2008	06/30/2011	1.952	yes
2.	1940	8842 SOFTWARE DISTRIBUTION INCORPORATED	Rarely Service Desk Maintenance	04/01/2008	04/01/2008	03/01/2010	0.891	yes
3.	1941	GOVCONNECTION INCORPORATED	Class Shield	05/01/2009	05/01/2009	04/30/2011	0.307	yes
4.	1942	ITS CORPORATION	Time & Materials	12/01/2008	01/01/2009	12/31/2011	20	yes
5.	7430	ISNET INTERNATIONAL CORPORATION		05/04/2003	05/05/2008	07/03/2009	0.040157	yes
6.	1945			01/26/2009	01/26/2010	06/30/2010	0.738	yes
7.	1946	IT FEDERAL SALES LIMITED LIABILITY COMPANY		10/01/2009	10/01/2009	09/30/2010	0.349	yes
8.	1947		Firm Fixed Price	08/09/2003	10/01/2008	08/30/2010	0.884	yes
9.	1948		Firm Fixed Price	11/05/2009	11/05/2009	05/30/2010	0.362	yes
10.	1949	REDHAWK IT SOLUTIONS LLC	Firm Fixed Price	01/22/2009	01/01/2010	12/31/2010	0.913	yes

0:00 / 6:48

YouTube

<http://openrefine.org>



Session 1

The characteristics of 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



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 = Infrequent, 2 = Frequent, 3 = Live

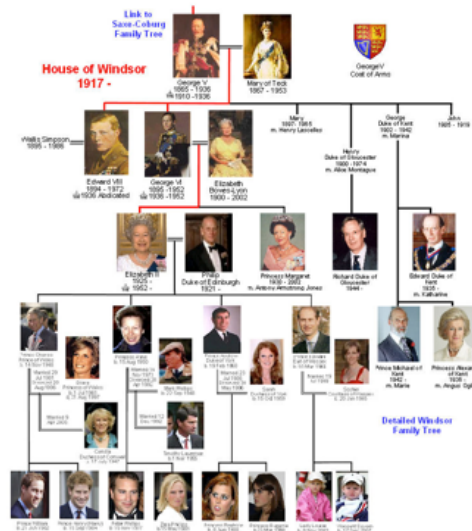


Data Representations

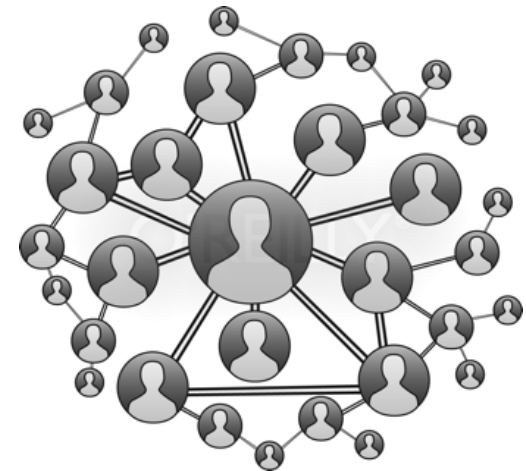
Tabular

Region	Production			Y
Country	Level 2	Production (thousand MT)	Change from last year	Change from 5 year average
Brazil		57289	-4.05%	+2.66%
	Mato Grosso	18,008	0.90%	6.17%
	Paraná	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	-7.09%	-1.97%
	Minas Gerais	2,667	5.12%	2.41%
	Bahia	2,512	-8.50%	4.84%
	São Paulo	1,382	-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%
	Pernambuco	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





Thank-you