Open Data Workshop

Data Economy-BDA Initiative,
Innovation Capital
Multimedia Development Corporation

MDEC Driving Transformation

DIGITAL MALAYSIA National Digital Economy Initiative

MSC MALAYSIA National ICT Initiative

Open Data- Introduction & Principles

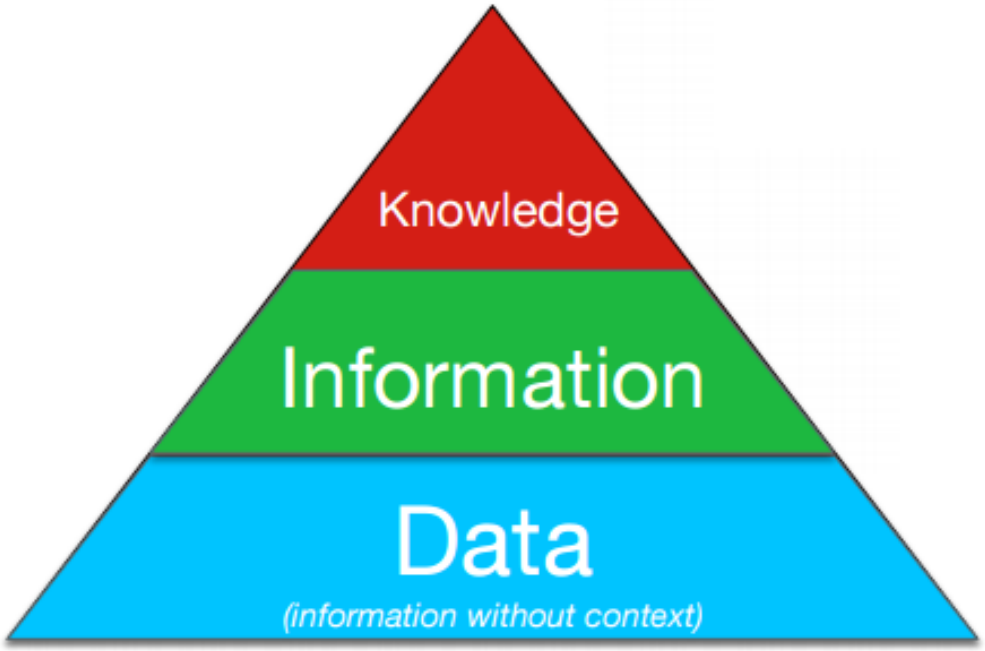
Data Economy-BDA Initiative,
Innovation Capital
Multimedia Development Corporation

MDEC Driving Transformation

DIGITAL MALAYSIA National Digital Economy Initiative

MSC MALAYSIA National ICT Initiative

WHAT IS DATA?



- A collection of facts, information and statistics that can be analyse to develop new knowledge
- The lowest level of abstraction from which information and then knowledge are derived
- Data can be collected through various method

" Information is not a form of knowledge"- Albert Einstein

MDEC Driving Transformation

DIGITAL MALAYSIA National Digital Economy Initiative

MSC MALAYSIA National ICT Initiative

OBJECTIVE

- Provide good foundation in the principles of open data and explore the good example of data usage
- Know-how of open data, datasets, data publication platform and data visualisation
- Best practices of open data
- Law, principles and open data licensing
- Encourage adoption of open data in public sector

OPEN DATA

What is Open Data?

OPEN DATA (2)

"Open data is data that can be freely used, reused and redistributed by anyone "

Source: Open Data Handbook. Org

"Open data is data that is published in an open format, is machine readable and is published under a license that allows for free reuse"

Source: Data.gov.uk

"Open data is data that is made available by organizations, businesses and individuals for anyone to access, use and share"

Source: The ODI UK

OPEN DATA (3)

The various definition of open data:

http://thegovlab.org/open-data-whats-in-a-name/

TYPE OF PERSONAL DATA

Open personal data

Data about people not a person

Available to anyone

Has been anonymised

e.g. number of people attending event, gender split, age ranges. (bigger numbers are better!)

Available personal data

Data about a person

Available to the person only!

Often known as MiData

e.g. credit scores, energy and other consumption data.

Personal data

Data about a person which is neither open nor available.

Might belong to you or be collected by a company.

TYPE OF PERSONAL DATA (2)

- Open Data: Bring transparency and open peer review
- Big Data: Bring evidences
- Personal Data: Makes it relevant

SO, WHY OPEN DATA?

OPEN DATA IMPACT

$125 Billion

Worldwide market for data driven and analytics services by 2015

Source: IDC & IIA

RM5.5 Billion

Converted potential productivity saving in Malaysia from **traffic efficiency** via Open Data e.g. creation of open innovation in public transport apps

Source: MDeC Analysis

42%

Companies invested in data analytics or planning to do so by 2015

Source: Gartner

£27M /Month

Potentially savings from NHS in England last year from healthcare datasets shared in UK

Source: ODI UK

OPEN DATA BENEFITS

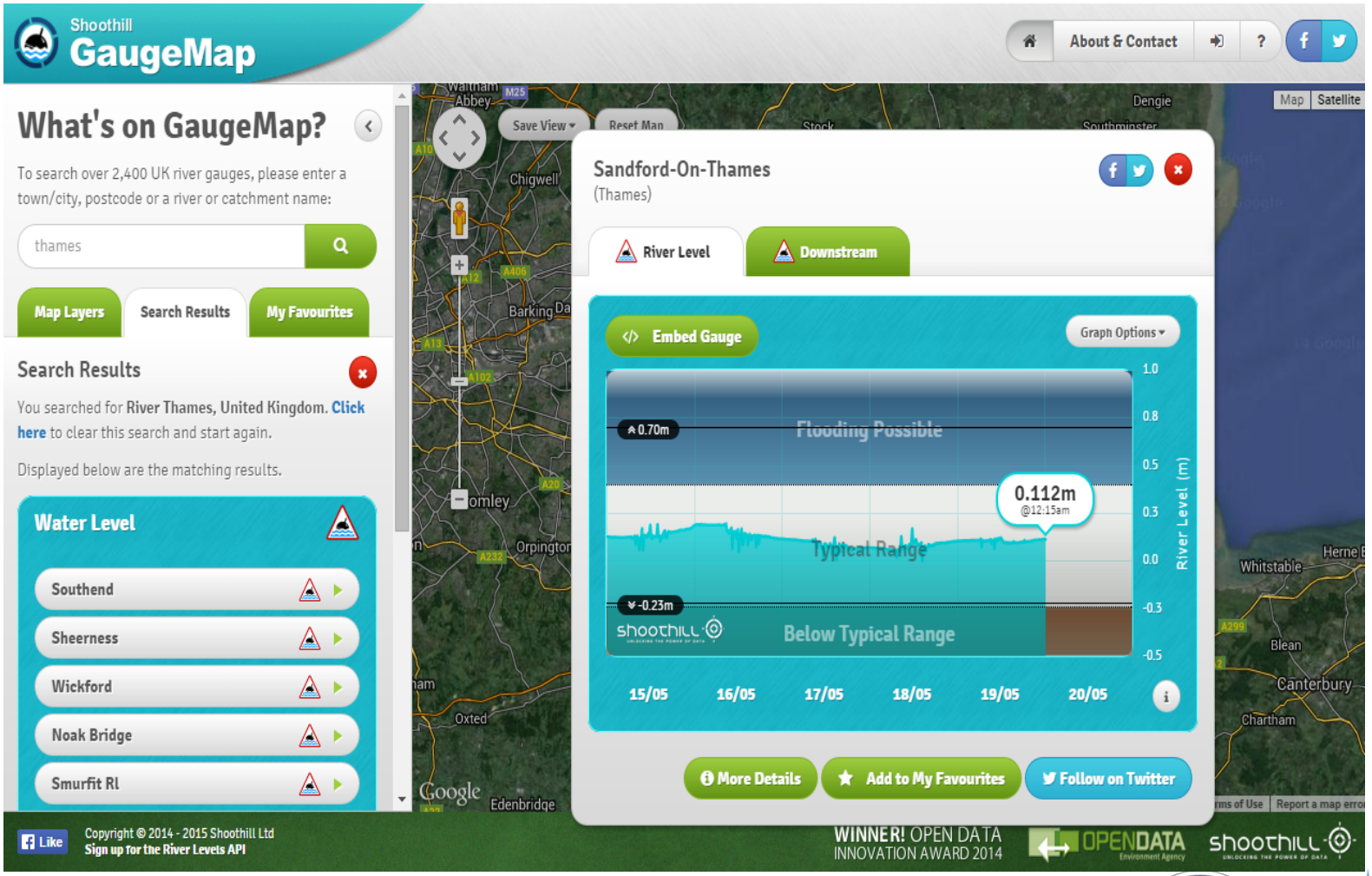
- Enabling economy – open innovation, creating jobs
- Reduce cost – data drive understanding of conventional system and explore value added to the current system
- Transparency – driving towards transparent culture in government and businesses
- Improve services – via innovative apps, productivity gain

OPEN DATA BENEFITS (2)

- Efficient data sharing – between G2G, G2B, G2C
- More high impact solutions can be created for benefits of community and rakyat

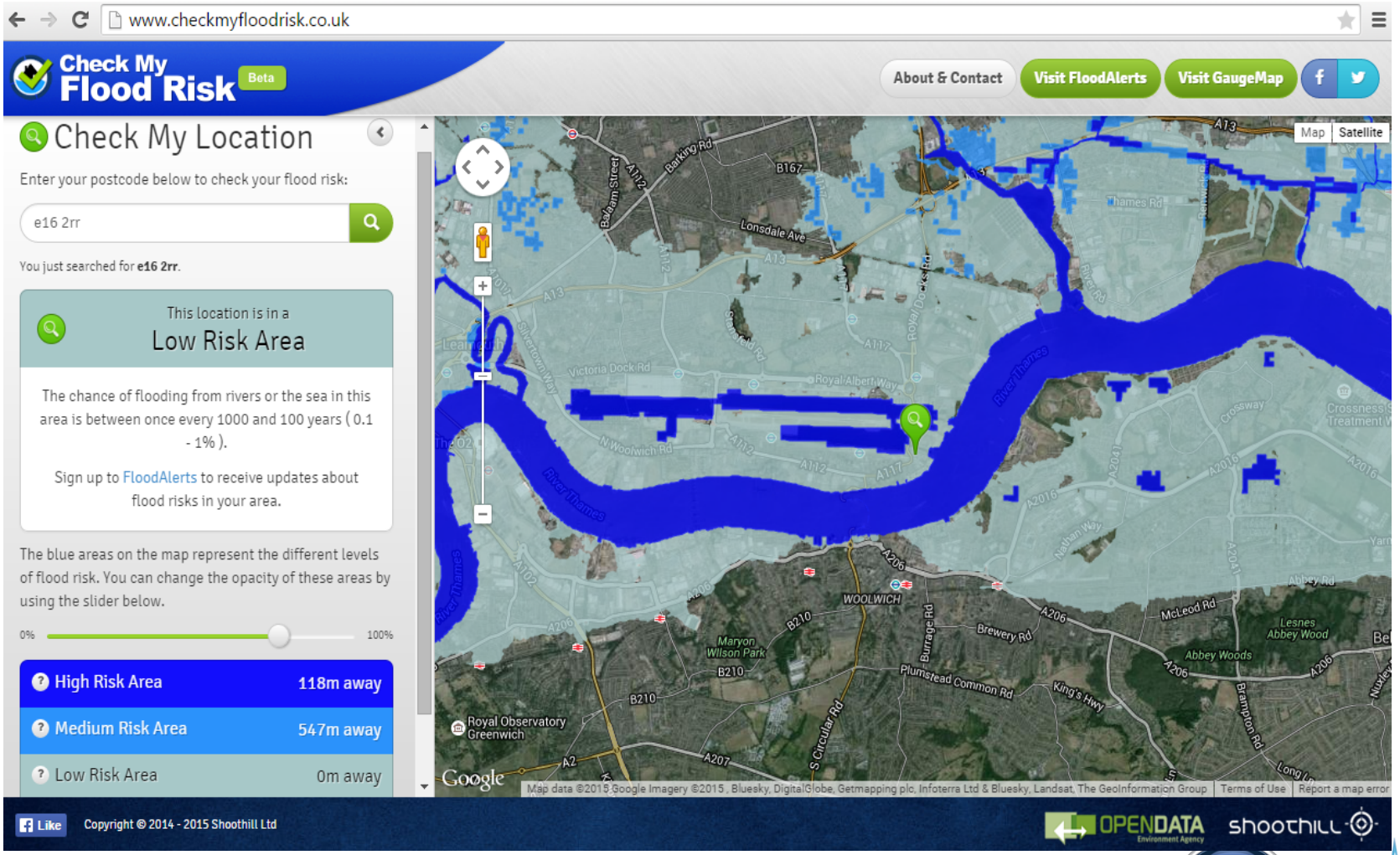
EXAMPLE OPEN DATA APPS

- Gaugemap (Measure river depth)



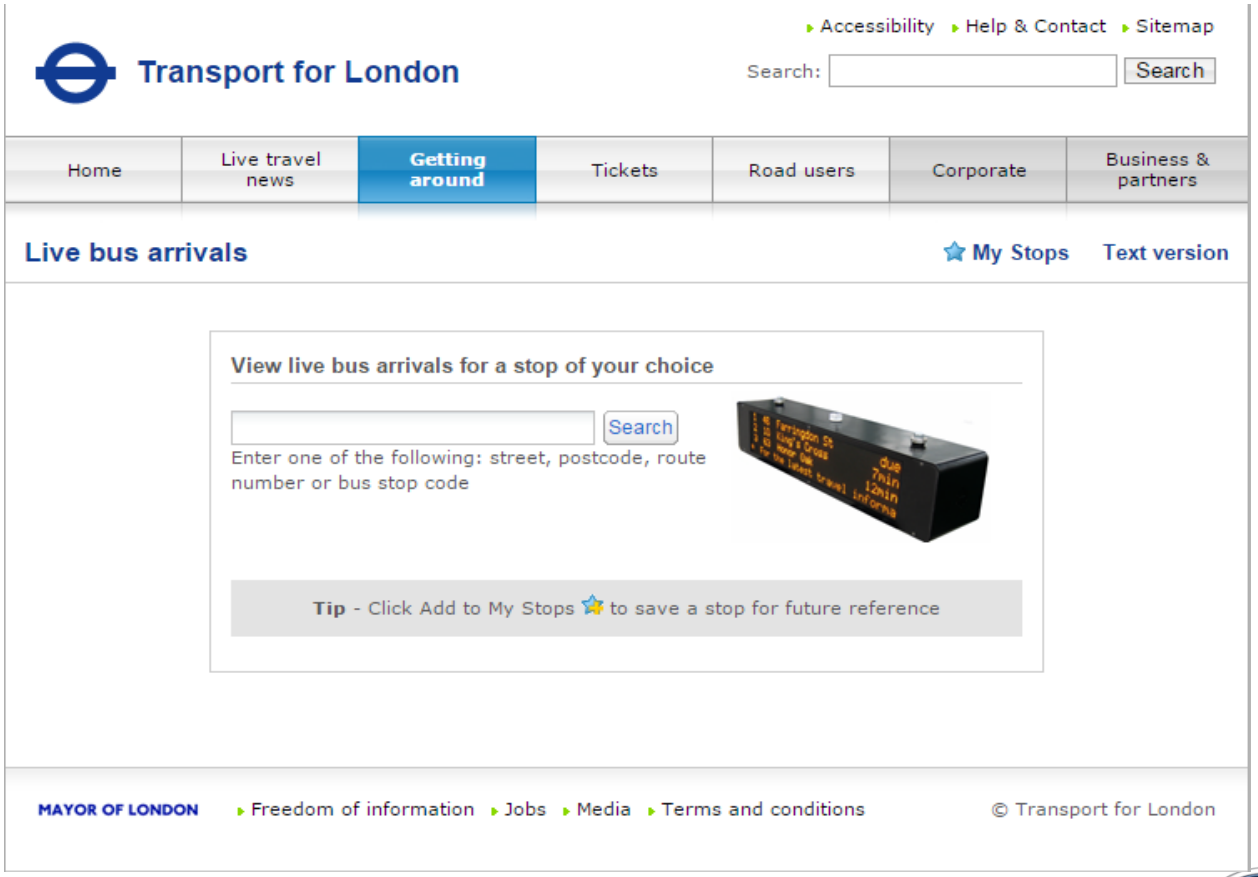
EXAMPLE OPEN DATA APPS (2)

- checkMyFloodRisk (Flood alert apps)



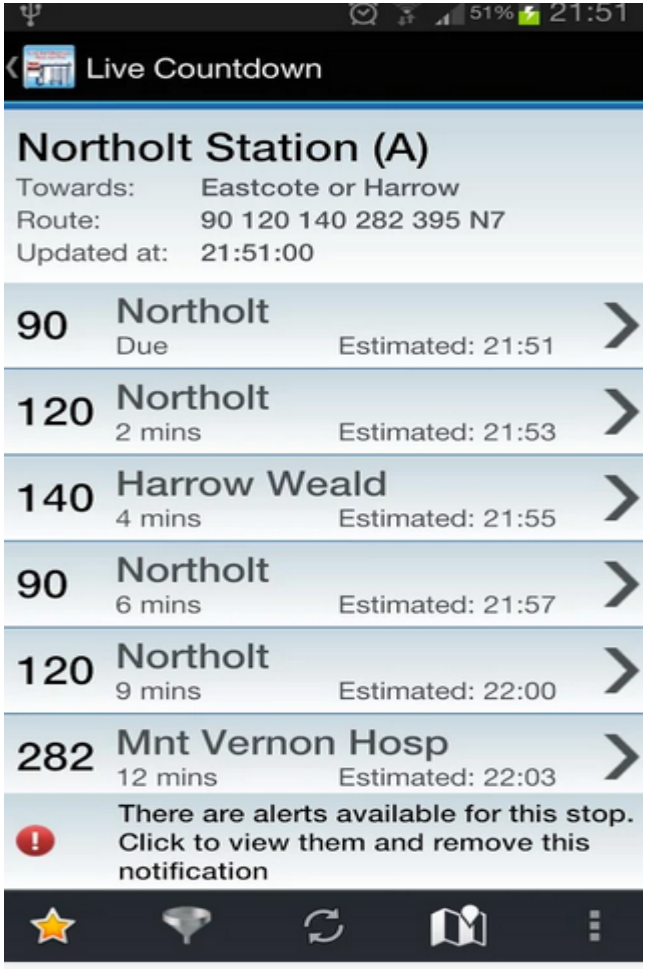
EVOLUTION OF APPS FROM OPEN DATA

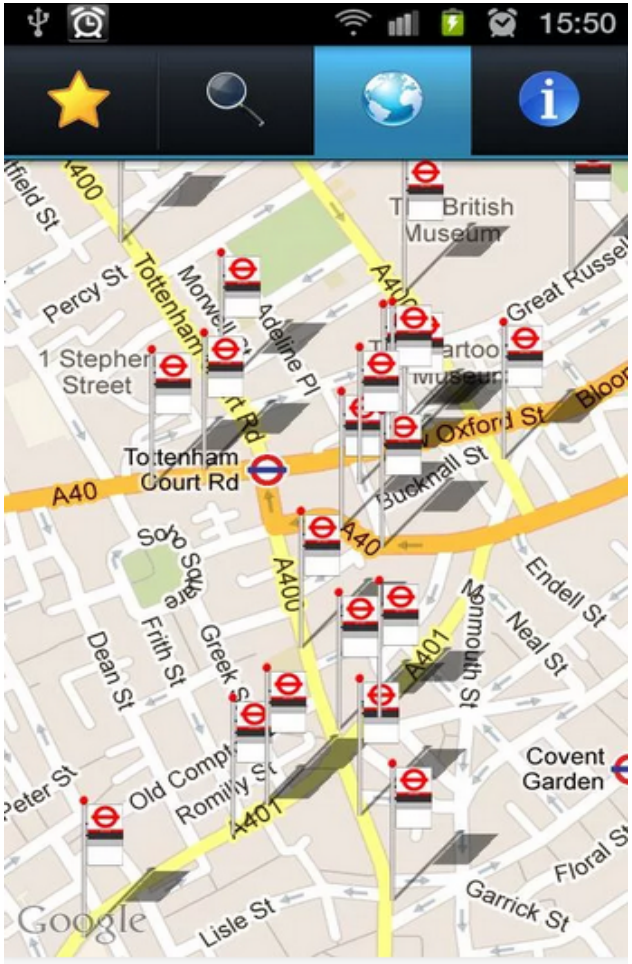
- For example, check TFL websites (http://countdown.tfl.gov.uk/)



EVOLUTION OF APPS FROM OPEN DATA

- Now check this apps (Live London Bus Tracker) , what is the similarity? What data that this app used?





MDEC Driving Transformation

DIGITAL MALAYSIA National Digital Economy Initiative

MSC MALAYSIA National ICT Initiative

DISCUSSION

- Why are you opening the data? What kind of data that can be open? Give examples
- What type of application from the data you open?
- How can it bring benefits?

Open Data – Law and Licensing

Data Economy-BDA Initiative,
Innovation Capital
Multimedia Development Corporation

OBJECTIVE

- Understand what is Open Data License
- Increase confidence in using Open Data License
- Understand *Garis Panduan Data Terbuka Sektor Awam*

LAW AND LICENSE

- Why Open Data License?

Provides clarity and set out how and what user and re-user are permitted to do with your datasets

- What is Open Data License?

Open Data License identify what is Open Data and who can use it, what is permitted & what is the restriction of using the data.

LAW AND LICENSE (2)

- Malaysia in the midst of strengthen Open Data License on Data.gov.my
- Example taken from UK Open Data License

http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

LAW AND LICENSE (3)

Using Information under this licence

Use of copyright and database right material expressly made available under this licence (the 'Information') indicates your acceptance of the terms and conditions below.

The Licensor grants you a worldwide, royalty-free, perpetual, non-exclusive licence to use the Information subject to the conditions below.

This licence does not affect your freedom under fair dealing or fair use or any other copyright or database right exceptions and limitations.

You are free to:

- copy, publish, distribute and transmit the Information;
- adapt the Information;
- exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.

You must, where you do any of the above:

- acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence;

LAW AND LICENSE (4)

Exemptions

This licence does not cover:

- personal data in the Information;
- information that has neither been published nor disclosed under information access legislation (including the Freedom of Information Acts for the UK and Scotland) by or with the consent of the Information Provider;
- departmental or public sector organisation logos, crests and the Royal Arms except where they form an integral part of a document or dataset;
- military insignia;
- third party rights the Information Provider is not authorised to license;
- other intellectual property rights, including patents, trade marks, and design rights; and
- identity documents such as the British Passport

LAW AND LICENSE (5)

Be careful!

- Personal Data
- Information that been accessed unlawfully
- Third party IP rights i.e. datasets that contain other party rights, map, images, excerpts of text

GARIS PANDUAN DATA TERBUKA SEKTOR AWAM

- What?

Basically a guideline of classifying, identifying and publishing open data across agencies and public sector.

Enabling public sector to understand open data principal

GARIS PANDUAN DATA TERBUKA SEKTOR AWAM (2)

- Imply the definition and principal of open data, terms of open data usage, checklist of open datasets and implementation council of open data in public sector

http://data.gov.my/folders/others/Garis_PanduanBeta_Ver_1.0_01042015.pdf

Open Data – Best Practices and Ensuring Data Quality

Data Economy-BDA Initiative,
Innovation Capital
Multimedia Development Corporation

BEST PRACTICES

Open Data definition according to Data.gov.my

"Open data is data that is published in an open format, is machine readable and is published under a license that allows for free reuse."

BEST PRACTICES (2)

Definition of Open Format:

- Data must be available and accessible.
- Reuse and redistribute: data must be permitted to be re-use and redistributed including intermixing with other datasets.
- Universal participations: data must be able to universally reuse, share and redistribute without any discrimination e.g. “for non-commercial purpose” that prevent commercialization of the open datasets.

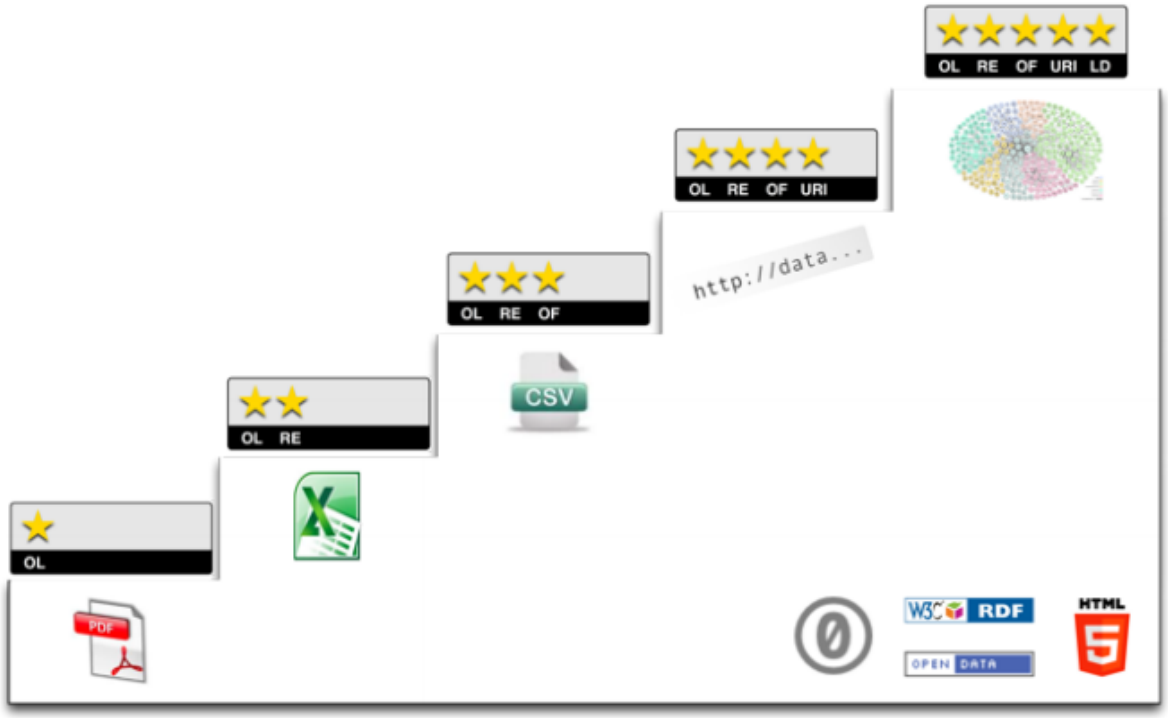
BEST PRACTICES (3)

Definition of Machine Readable:

- Data or metadata which is in a format that is understood by machine

- Examples: Comma separated values (CSV), Excel (XLS), XML, HTML etc.

BEST PRACTICES (4)



★ Available on the web (whatever format) *but with an open licence, to be Open Data*

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ use URIs to denote things and when people look up these URIs, provide useful information and/or data.

★★★★★ All the above, plus: Link your data to other people's data to provide context

Source:http://5stardata.info/

BEST PRACTICES (5)

4 aspects to look for opening up data:

- Legal
- Technical
- Practical
- Social

BEST PRACTICES (6)

Adding aspects context:

Who Collected it? Owns it? Publishes it? Is the Audience?	**Where** Was it collected? Is it used? Is it described? Is it located?
What Is it (title/description)? Type of data is it? Type of objects?	**When** Collected? Published? Updated? Due next update?

BEST PRACTICES (7)

Legal

- ✓ Right to publish
- ✓ Data licensed
- ✓ Content licensed
- ✓ Clear privacy statement
- ✓ Source of data documented
- ✓ Audited anonymisation

BEST PRACTICES (8)

Technical

- ✓ Data hosted online
- ✓ Type of data refined
- ✓ Machine readable format
- ✓ Clear technical documentation
- ✓ Data can be verified

BEST PRACTICES (9)

Practical

- ✓ Quality controlled
- ✓ Listed in collection
- ✓ Useable period described
- ✓ Discoverable from main homepage
- ✓ Referenced from publication or application

BEST PRACTICES (10)

Social

- ✓ Support for improving/fixing
- ✓ Email support
- ✓ Discussion group/forum
- ✓ Social media channel
- ✓ Supported community
- ✓ Tools and guide available to work with data

BEST PRACTICES (10)

Our objectives:

- ✓ At least 3 star data ranking
- ✓ Enhancement on data portal
- ✓ Balanced supply and demand of open datasets
- ✓ Continuous support from community

Q&A

Thank you!

Open Data – Data Validation, Publication & Visualization

Data Economy-BDA Initiative,
Innovation Capital
Multimedia Development Corporation

DISCOVERING OPEN DATA

- The best practical to publish data, is to think as a data consumer itself
- There is no point to publish data that not reliable and can't be use
- Few questions we need to ask:
  - Do we understand what is the dataset?
  - Do we understand about the data?
  - Does it clean?
  - Am I able to access the data?
  - Is the data too granular? Too generic? Etc.

DISCOVERING OPEN DATA (2)

Please open/fork Workshop Github for exercise:

http://github.com/chlorofell/OpenDataWorkshop

DISCOVERING OPEN DATA (3)

Exercise 1:

To understand what is dataset and best format of datasets

Key Points:

- Differentiate both datasets
- Analyse best practice between both datasets
- Improvement on datasets

DATA CLEANING AND VALIDATION

What would you do next? Once you identify the datasets that can be open?

- Verify that the datasets according to Open Data Principles – anonymous, not too aggregated etc.
- Perform data cleaning- granular as possible, meaningful, no duplication etc.
- Format change to machine readable CSV, XLS, XML etc.

DATA CLEANING AND VALIDATION (2)

Once we identify the dataset, we need to ensure that dataset is free from error. A problem dataset can contain few errors such as:

1. Multiple representation
2. Duplicate records
3. Summation records – summation formula etc.
4. Redundant Data
5. Mix use of numerical scale
6. Spelling errors
7. Date Validation – British and America Date

DATA CLEANING AND VALIDATION (2)

There are few methods of how to perform data cleaning and validation:

1. Manual method – self-identifying
2. Using data cleaning tools :
   - Open Refine (openrefine.org)
   - Data Cleaner (datacleaner.org)
   - Data Wrangler (http://vis.stanford.edu/wrangler/)

DATA CLEANING AND VALIDATION (3)

Exercise 2:

To perform data cleaning using Open Refine

Key Points:

- Identify errors on datasets
- Perform data cleaning
- Understand to data cleaning tools

http://github.com/chlorofell/OpenDataWorkshop

DATA PUBLICATIONS

- Data publication platform is which all the catalogues of datasets resides and published, for public access

- Data can be publish using open source platform or using traditional websites

- Examples: data.gov.my, data.gov.uk hosted on self-build platform

DATA PUBLICATIONS (2)

- 2 types of Data Publication platform:

✓Integrated platform

✓Specialist platform

DATA PUBLICATIONS (3)

- Integrated Platform

- Build on top of organisation main website
- No separation from other authoritative data
- No new platform to learn
- Easy to search for data

- Specialist platform

- Open Data focus platform
- Clear workflow of publishing data
- Virtualization tools
- Ease to setup and maintain

DATA PUBLICATIONS (4)

- Best practices of publishing data:
- Separate from the main websites
- Design to publish data, not fulfilling other organisations' goal
- Host the data, control and monitor the input and output log
- Provide data services – tools and visualization in one portal
- Best if data can be share as in API e.g. transport API

DATA PUBLICATIONS (5)

- Example of Data Publication Specialist:

1. CKAN (http://demo.ckan.org/)

2. Open Data Soft (https://www.opendatasoft.com)

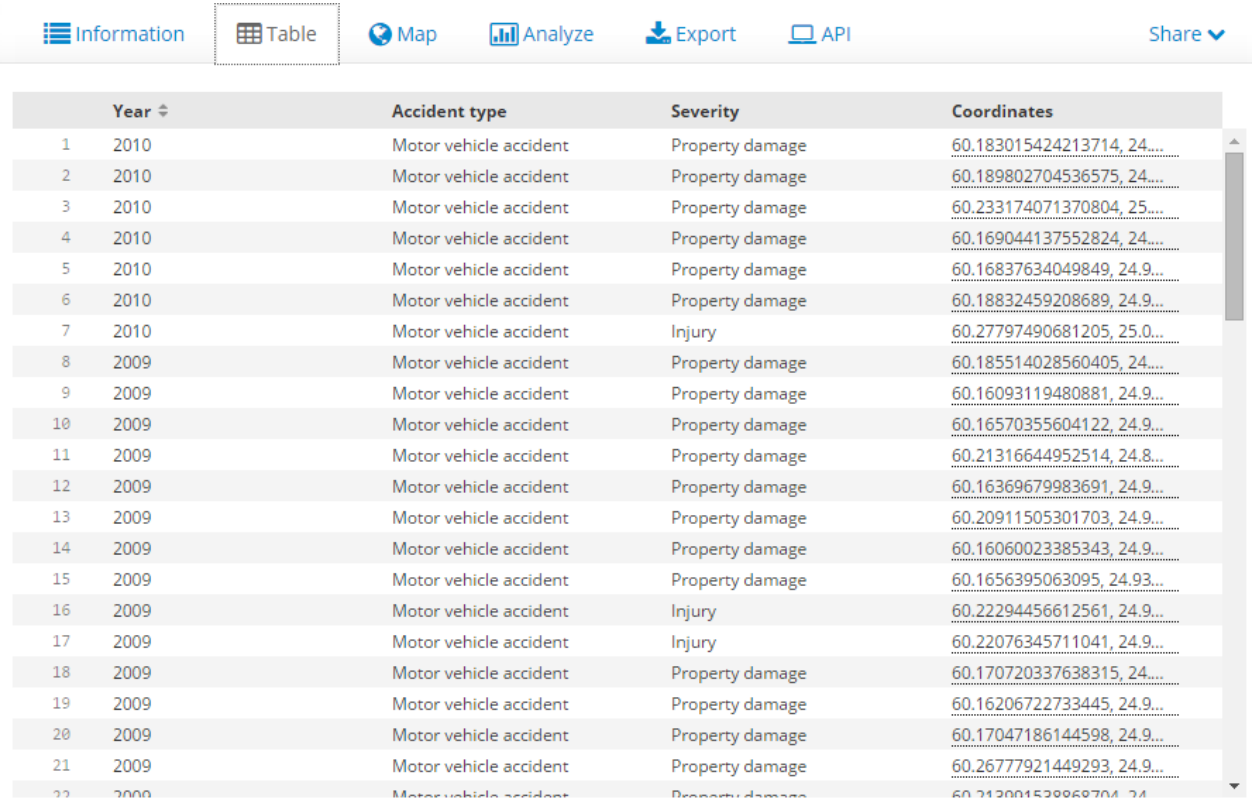
3. Socrata (www.socrata.com)

DATA VISUALIZATION

- Data visualization transform data into more meaningful insight for analysis
- It helps to improve the understanding of the data
- Use data as the source to tell the stories and making impact i.e. about the trends, historical data, maps/cartograph etc.

DATA VISUALIZATION (2)

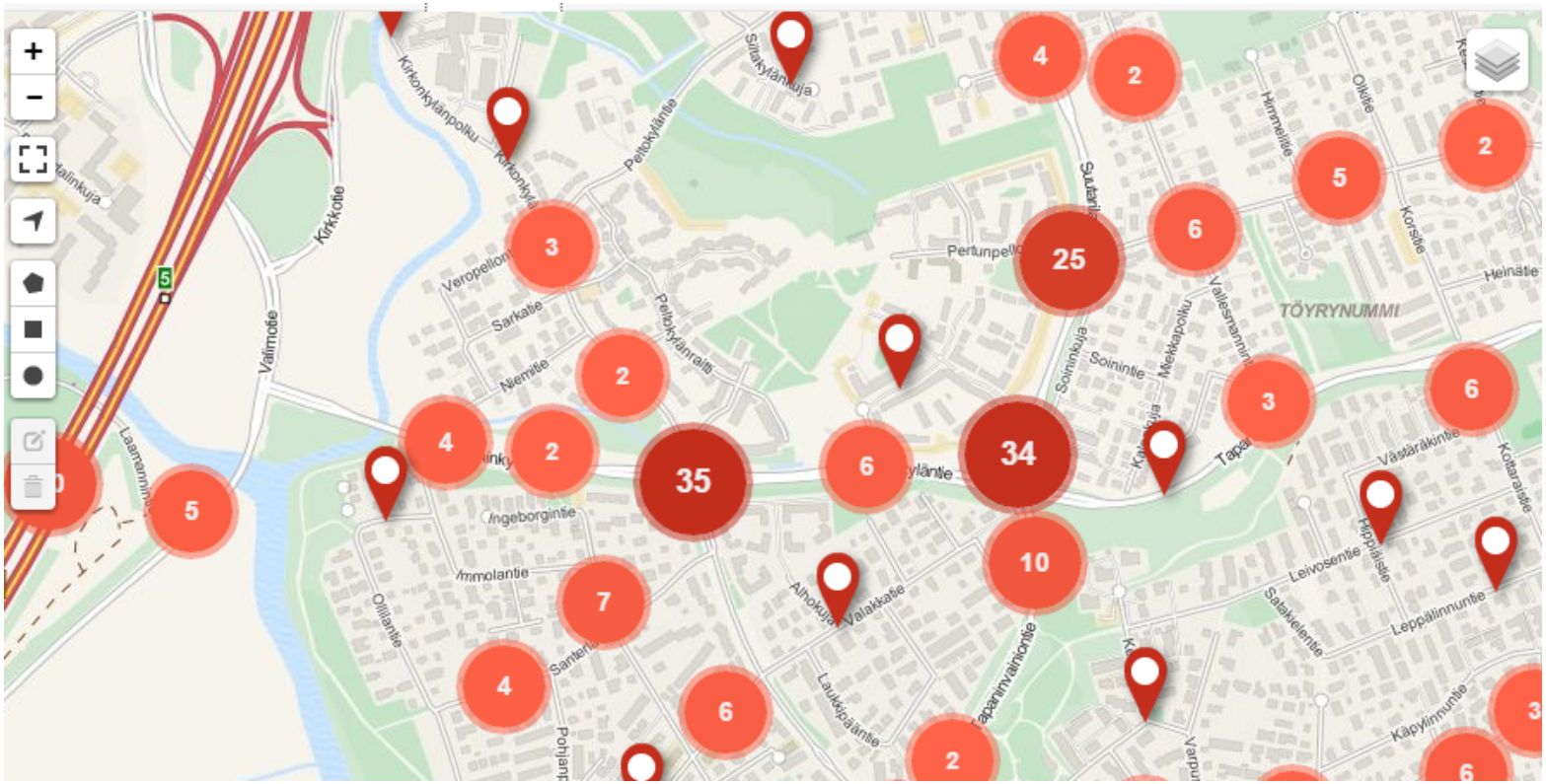
- For example, using OpenDataSoft platform to transform Accident datasets into user-friendly visualization for better insight



	Year	Accident type	Severity	Coordinates
1	2010	Motor vehicle accident	Property damage	60.183015424213714, 24....
2	2010	Motor vehicle accident	Property damage	60.189802704536575, 24....
3	2010	Motor vehicle accident	Property damage	60.233174071370804, 25....
4	2010	Motor vehicle accident	Property damage	60.169044137552824, 24....
5	2010	Motor vehicle accident	Property damage	60.16837634049849, 24.9...
6	2010	Motor vehicle accident	Property damage	60.18832459208689, 24.9...
7	2010	Motor vehicle accident	Injury	60.27797490681205, 25.0...
8	2009	Motor vehicle accident	Property damage	60.185514028560405, 24....
9	2009	Motor vehicle accident	Property damage	60.16093119480881, 24.9...
10	2009	Motor vehicle accident	Property damage	60.16570355604122, 24.9...
11	2009	Motor vehicle accident	Property damage	60.21316644952514, 24.8...
12	2009	Motor vehicle accident	Property damage	60.16369679983691, 24.9...
13	2009	Motor vehicle accident	Property damage	60.20911505301703, 24.9...
14	2009	Motor vehicle accident	Property damage	60.16060023385343, 24.9...
15	2009	Motor vehicle accident	Property damage	60.1656395063095, 24.93...
16	2009	Motor vehicle accident	Injury	60.22294456612561, 24.9...
17	2009	Motor vehicle accident	Injury	60.22076345711041, 24.9...
18	2009	Motor vehicle accident	Property damage	60.170720337638315, 24....
19	2009	Motor vehicle accident	Property damage	60.16206722733445, 24.9...
20	2009	Motor vehicle accident	Property damage	60.17047186144598, 24.9...
21	2009	Motor vehicle accident	Property damage	60.26777921449293, 24.9...
22	2009	Motor vehicle accident	Property damage	60.213091528868704, 24....

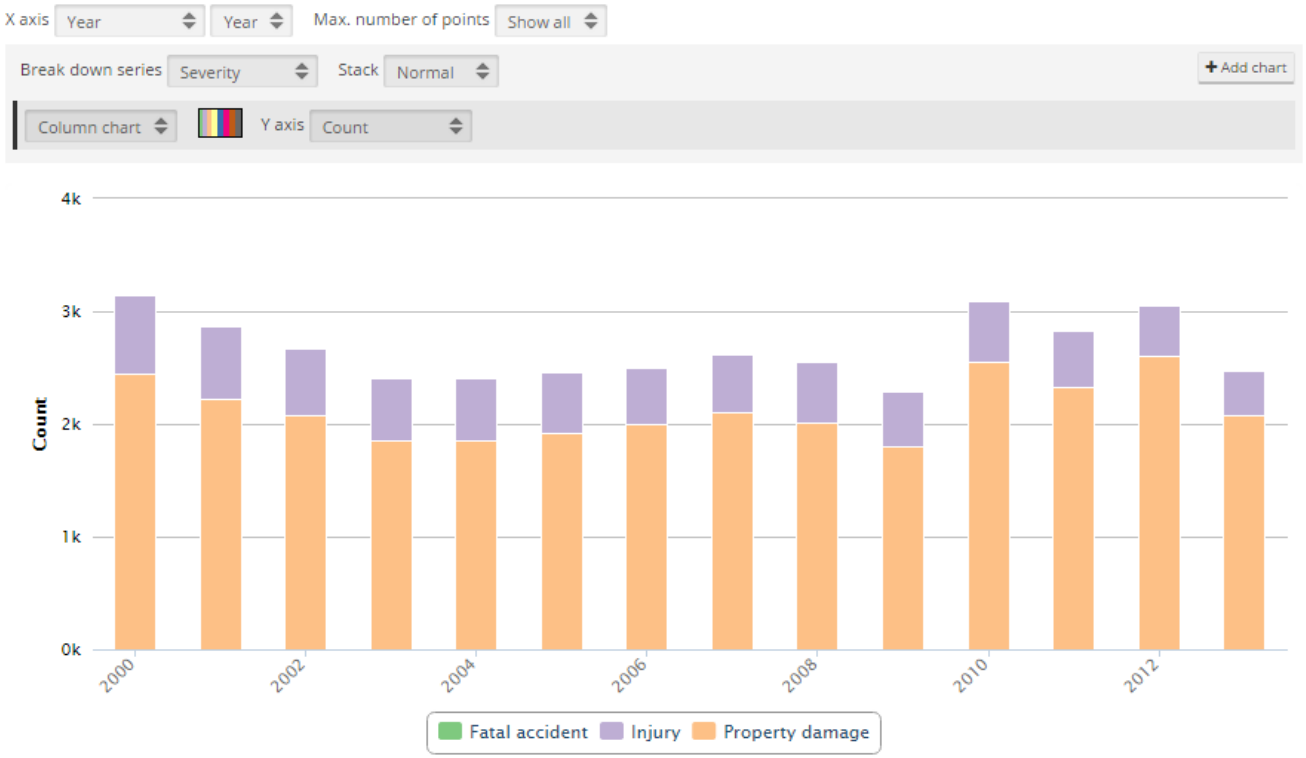
DATA VISUALIZATION (3)

- Visualization transforms the datasets to cartograph for analysis



DATA VISUALIZATION (4)

- Visualization transforms the datasets to Bar Chart for analysis



DATA VISUALIZATION (5)

- Other data-driven visualization tools:

-Plot.ly (www.plot.ly)

-CartoDB (www.cartodb.com)

-D3 (http://d3js.org) – Using Java Script

DATA VISUALIZATION (6)

Exercise 3:

Learn translating dataset to a meaningful visualization and analysis insights

Key Points:

- Identify use-cases, datasets and create story from data
- Perform data cleaning and validation using ScraperWiki
- Use plot.ly to visualize dataset

http://github.com/chlorofell/OpenDataWorkshop

DATA VISUALIZATION (7)

Exercise 3:

Publish Visualization Output at:

http://chlorofell.github.io/OpenDataWorkshop

Thank you

"Data is a means, not an end."

.