### Use of Open Data in Hong Kong

Sammy Fung

sammy.hk

Incu-Lab ICE in StartMeUpHK - Open Data Initiative Gathering 2013/12/04

# We want a better life with public data.

# We want a easier way to access the public data.

## Agenda

- What is Open Data?
- Use of Open Source Software in web crawling.
- Starting new Open Source project hk0weather to create Open Weather Data.

## Sammy Fung

- Software Developer
  - to use and develop open source sofware.
  - Perl → PHP → Python.
  - interests on Data Mining / Web Crawling.
  - own a startup of web and mobile technology.

## Sammy Fung

- 15+ years in Open Source Communities.
  - Founding Chairman, Hong Kong Linux User Group.
  - Founding Chairman, Open Source Hong Kong.
  - Member, GNOME Asia committee.
  - Mozilla Representative
  - Member, program committee at COSCUP
    - Conference for Open Source Coders, Users and Developers.
    - Largest open source conference in Taiwan.

# What is Open Data?

## Open Data

Three Laws of Open Government Data by David Eaves.

- 1.If it can't be spidered or indexed, it doesn't exist.
- 2.If it isn't available in open and machine readable format, it can't engage.
- 3.If a legal framework doesn't allow it to be repurposed, it doesn't empower.

http://eaves.ca/2009/09/30/three-law-of-open-government-data/

# Open Data

- Tim Berners-Lee, the inventor of the Web.
  - 5stardata.info
  - 5 star deployment scheme of Open Data.

# \* One Star - Open Data

- 1.make your stuff available on the Web (whatever format) under an open license.
- 2.make it available as structured data (e.g., Excel instead of image scan of a table)
- 3.use non-proprietary formats (e.g., CSV instead of Excel)
- 4.use URIs to denote things, so that people can point at your stuff.
- 5. link your data to other data to provide context.

# \*\* Two Star - Open Data

- 1.make your stuff available on the Web (whatever format) under an open license.
- 2.make it available as structured data (e.g., Excel instead of image scan of a table)
- 3.use non-proprietary formats (e.g., CSV instead of Excel)
- 4.use URIs to denote things, so that people can point at your stuff.
- 5. link your data to other data to provide context.

# \*\*\* Three Star - Open Data

- 1.make your stuff available on the Web (whatever format) under an open license.
- 2.make it available as structured data (e.g., Excel instead of image scan of a table)
- 3.use non-proprietary formats (e.g., CSV instead of Excel)
- 4.use URIs to denote things, so that people can point at your stuff.
- 5. link your data to other data to provide context.

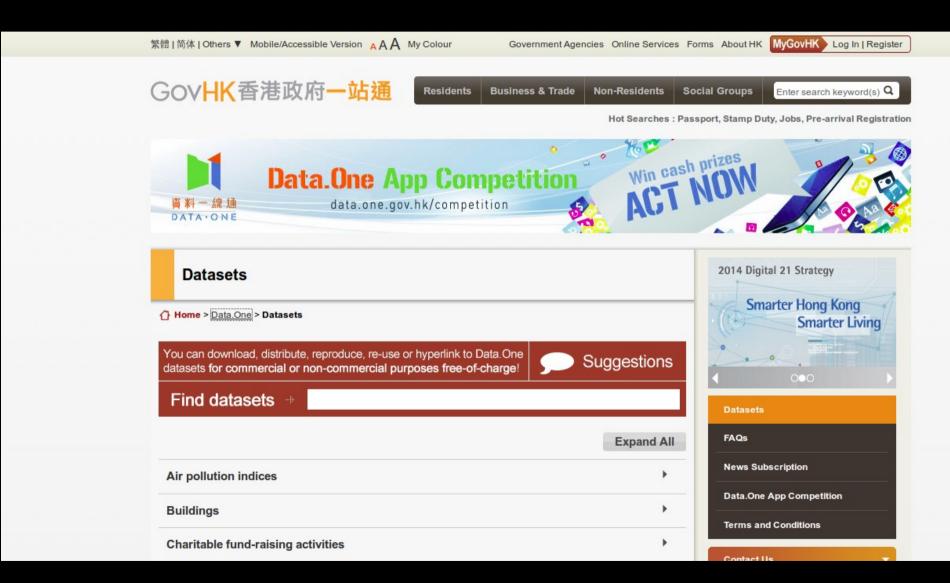
# \*\*\*\* Four Star - Open Data

- 1.make your stuff available on the Web (whatever format) under an open license.
- 2.make it available as structured data (e.g., Excel instead of image scan of a table)
- 3.use non-proprietary formats (e.g., CSV instead of Excel)
- 4.use URIs to denote things, so that people can point at your stuff.
- 5. link your data to other data to provide context.

# \*\*\*\*\* Five Star - Open Data

- 1.make your stuff available on the Web (whatever format) under an open license.
- 2.make it available as structured data (e.g., Excel instead of image scan of a table)
- 3.use non-proprietary formats (e.g., CSV instead of Excel)
- 4.use URIs to denote things, so that people can point at your stuff.
- 5.link your data to other data to provide context.

# Open Data in Hong Kong



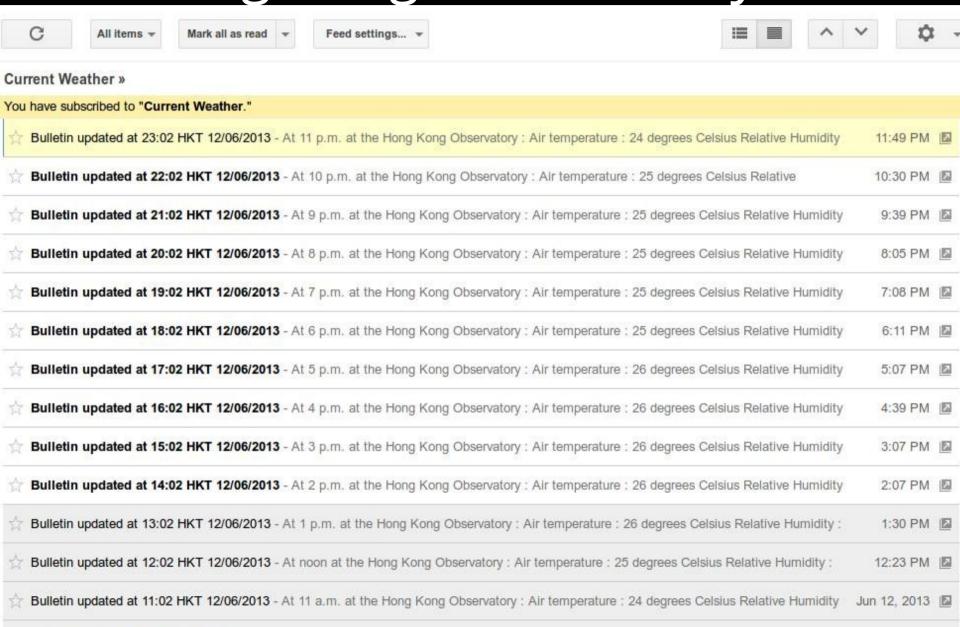
# Open Data in Hong Kong

- Data.One
  - http://www.gov.hk/en/theme/psi
  - released on 2011/3/31.
  - First App Competition on Data.One
    - Call for Submission now till 2014/02/28.

## Weather Information in Hong Kong

- Hong Kong Observatory
  - Hourly Hong Kong Weather Report
  - Regional Weather in Hong Kong (10 min updates)
  - Weather Forecast and Weekly Weather Forecast
  - Typhoon Report and Forecast

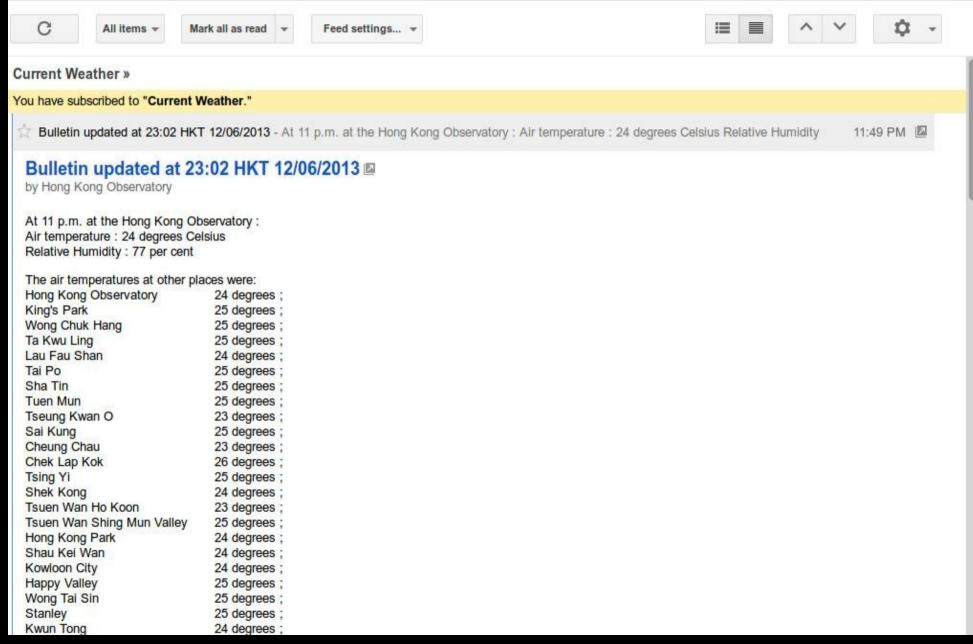
# Hong Kong Observatory RSS



Bulletin updated at 10:02 HKT 12/06/2013 - At 10 a.m. at the Hong Kong Observatory: Air temperature: 24 degrees Celsius Relative Humidity

Jun 12 2013

# Hong Kong Observatory RSS



#### Weather at Data.One

- I posted a blog 'Progress of Open Government Data in Hong Kong' on 2013/01/17.
- Weather at Data.One provides 7 dataset URLs, returns RSS (XML) format (Eng/TChi/SChi)
  - One word: Useless.
  - Data.One dataset (RSS) is completely different with HKO own paid service (XML).

#### Weather at Data.One

- Example Current local weather report:
- Plain text report in RSS.
- Difference to quote report content:
  - Website: a pair of HTML tags, eg. <PRE>....</PRE>.
  - Data.One: a pair of RSS description tags,
     <description>....</description>.
- Other weather data is missing, eg. Regional temperture updates per each 12 mins.

#### Weather at Data.One

- Weather at Data. One is 'report' but not 'data'.
- Weather RSS is already released by HKO before launch of Data.One.
- Technically, json/xml format is better readable by computer programs.

- In November 2013, 43 datasets are available.
  - JSON/XML = 18
  - RSS = 10
  - -XLS = 6
  - -CSV = 4
  - JPG/PNG = 3
  - HTML/MDB = 2

- JSON/XML (18 datasets)
  - Air Pollution.
    - Past 24-hour Air Pollution Index from stations.
  - Approved Charitable Fund-raising Activities
  - Restaurant and Food Licences.
  - Details of facility locations.
  - Reward Notices from Police Force.
  - Marine Traffic (Arrival/Departure).
  - Traffic Speed and special news.
  - EventHK information.

- RSS (10 datasets)
  - Weather Information (7 datasets)
  - Beach Water Quality (1 datasets)
  - Current Air Pollution Index range and forecase (2 datasets)

- JPG/PNG (3 datasets)
  - Exhibition gallery of government building projects.
  - Speed map panels.
  - Traffic snapshot images.

#### CSV

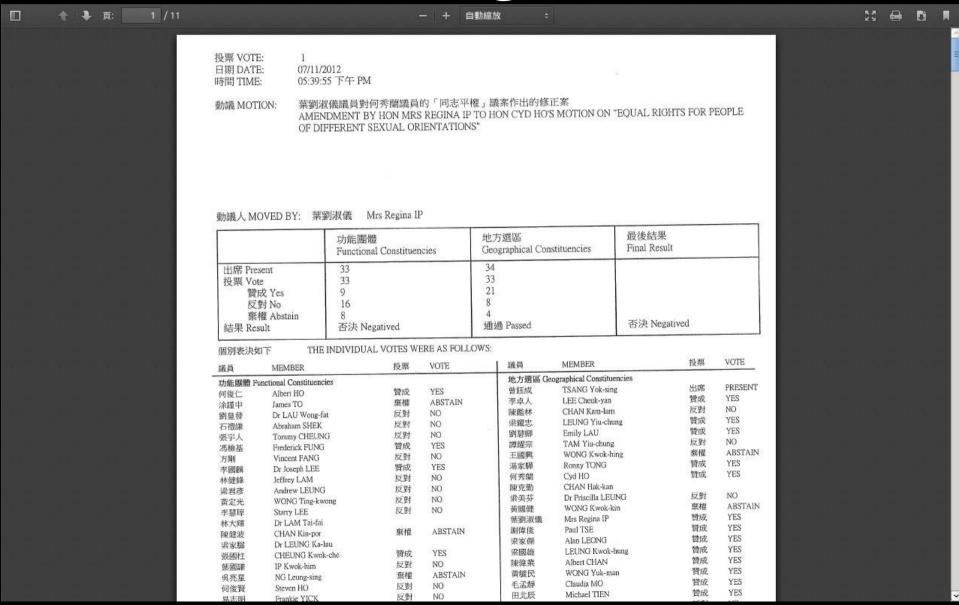
- Past Record of Air Pollution Index
- Locations of Public Facility and GovWifi
- Marine Shipping directory of HK

#### • HTML

- HTML version of Marine Traffic.
- XLS, MDB
  - 2011 Population Census.
  - Property Market Statistics.
  - Monthly Digested Stats and Registers of Auth Persons from Building Dept.
  - Routes and fares of public transport.

- Many departments does not release their useful data, and release current information available on their website.
  - Few of them keep available open data in their own.
- Most of them does not understand what is 'real' open data.
  - Data insteads of Information.
  - Open data format insteads of proprietary data format.
  - Useful of data.
- Some departments should manage their open data in better data structure.

# Legco Meeting Minutes and Voting Results



# Legco Meeting Minutes and Voting Results



# Legco Meeting Minutes and Voting Results

- In October 2013, LegCo start to publish voting results of House Committe in XML.
- It is <u>not</u> a part of Data. One project.
- My open source software on LegCo vote result XML:
  - http://github.com/smamyfung/legcovotes

# Open Data is important to citizens.

# User of Open Source Software in web crawling

# Web Scraping

- a computer software technique of extracting information from websites. (Wikipedia)
- for business, hobbies, research purposes.

## Web Scraping

- Look for right URLs to scrap.
- Look for right content from webpages.
- Saving data into data store.
- When to run the web scraping program?

# Use of Open Source Software in Web Crawling

- Use Open Source Tools to collect useful and meaningful machine-readable data.
- Doesn't need to wait provider to release data in machine-readable format.

### Open Source Tools

- Python programming lanugage
- with Regular Expression library
- Scrapy web crawling framework

### Why python + scrapy?

- python: my current favourite programming language for few years.
- scrapy: web crawling framework written in Python.

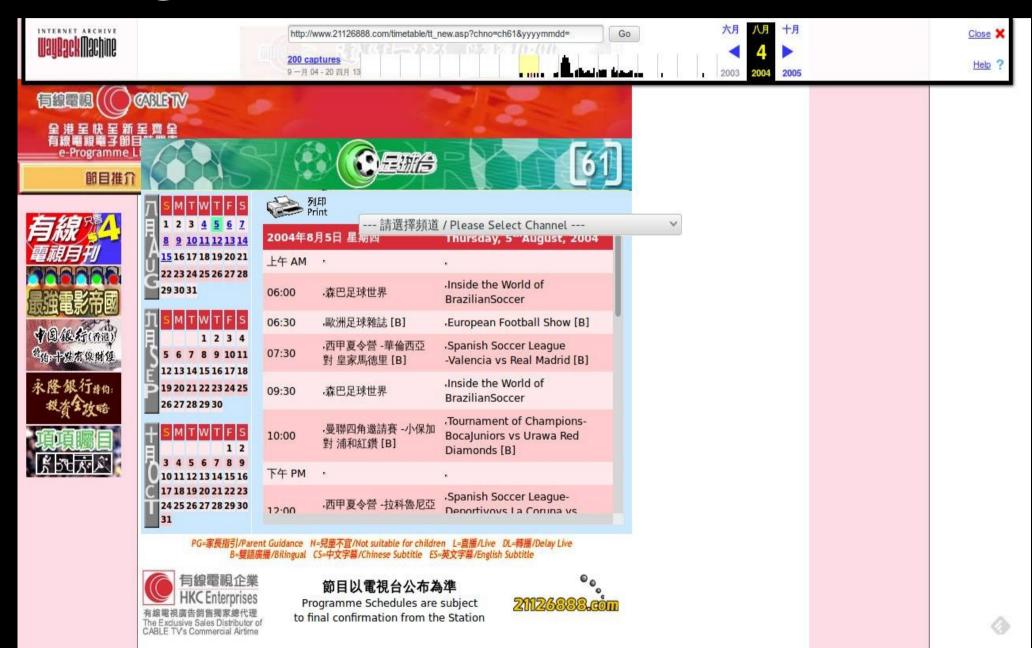
### What is Scrapy?

- An open source web scraping framework for Python.
- Scrapy is a fast high-level screen scraping and web crawling framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.

### Scrapy Features

- define data you want to scrapy
- write spider to extract data
- Built-in: selecting and extracting data from HTML and XML
- Built-in: JSON, CSV, XML output
- Interactive shell console
- Built-in: web service, telnet console, logging
- Others

### Programme List of Paid TVs in 2004



### Programme List of Paid TVs in 2004

- I want to know live football match was showing on which channel.
- Paid TV web site = M\$ + IIS + ASP + Flash
- Slow...... Very Slow..... Extremely Slow!
- Couldn't connect at any peak hours!
- Wrote my first web crawler in PHP in 2004.

### Public Transportation in 2006-2010

- Kowloon Motor Bus (KMB)
  - No map view for a bus route
- Public Transportation Enquiry System (PTES)
  - Exteremly Poor, Ugly (or much worse) map UI on PTES.

## HK Observatory and Joint Typhoon Warning Center

- Any typhoon is coming to Hong Kong? And When will it come?
- No easy data exchange format.
- No RSS nor ATOM.
- We aren't check websites everyday.

### My Products

- WeatherHK ← ← ←
- TCTrack

### WeatherHK

- http://twitter.com/weatherhk
- hourly current weather report
- weather forecast report
- tropical signal warning



乎23至27度。吹和緩偏南風,明晚轉吹偏東風。

### WeatherHK

- Backend: Python + Scrapy + Database + Twitter + NNTP.....
- Frontend: Twitter + Newsgroup



### WeatherHK

- http://twitter.com/weatherhk
- Interview by MetroPop in 2009.



### My Products

- WeatherHK
- TCTrack ← ← ←

### TCTrack

http://sammy.hk/projects/tctrack/tctrack.php

Plot TC current and forecast tracks over

Google Map.

- Source:
  - JTWC
  - HKO



### TCTrack

- http://sammy.hk/projects/tctrack/tctrack.php
- Probably first tctrack map in HK using GoogleMap
- Use of GMap: TCTrack -> Weather Underground Hong Kong -> HKO



### TCTrack

- http://twitter.com/tctrack
- Tweet JTWC updates for Northwest Pacific.



# Releases information to citizens in a better presentation.

### Starting new Open Source project hk0weather to create Open Weather Data.

## Starting new Open Source projects to create Open Data

- Develop a open source project.
- Release data in standard machine-readable data format.

- https://github.com/sammyfung/hk0weather
- Open Source Hong Kong Weather Project.
- convert to JSON data from HKO webpages.
- python + scrapy
- 1st version: from current weather report, extracting temperture and humidity from 20+ weather stations, export in json format.

- https://github.com/sammyfung/hk0weather
- \$ virtualenv hk0weatherenv
- \$ source hk0weatherenv/bin/activate
- \$ pip install scrapy
- \$ git clone https://github.com/sammyfung/hk0weather.git
- \$ cd hk0weather
- \$ scrapy crawl currwx -t json -o testresult

- Python
  - import re
- Scrapy
  - web crawling framework written in Python.
  - HtmlXPathSelector.
  - built-in JSON, CSV, XML output.

```
[{"humidity": 80, "station": "hko", "temperture": 17, "time": 1360785720},
{"station": "kingspark", "temperture": 16, "time": 1360785720},
{"station": "wongchukhang", "temperture": 17, "time": 1360785720},
{"station": "takwuling", "temperture": 16, "time": 1360785720},
{"station": "laufaushan", "temperture": 15, "time": 1360785720},
{"station": "taipo", "temperture": 16, "time": 1360785720},
{"station": "shatin", "temperture": 17, "time": 1360785720},
{"station": "tuenmun", "temperture": 17, "time": 1360785720},
{"station": "tseungkwano", "temperture": 16, "time": 1360785720},
{"station": "saikung", "temperture": 16, "time": 1360785720},
{"station": "cheungchau", "temperture": 17, "time": 1360785720},
{"station": "cheungchau", "temperture": 17, "time": 1360785720}.
{"station": "tsingyi", "temperture": 17, "time": 1360785720},
{"station": "shekkong", "temperture": 15, "time": 1360785720}
{"station": "tsuenwanhokoon", "temperture": 15, "time": 1360785720},
{"station": "tsuenwanshingmunvalley", "temperture": 17, "time": 1360785720},
{"station": "hongkongpark", "temperture": 17, "time": 1360785720},
{"station": "shaukeiwan", "temperture": 16, "time": 1360785720},
{"station": "kowlooncity", "temperture": 16, "time": 1360785720},
{"station": "happyvalley", "temperture": 18, "time": 1360785720}
{"station": "wongtaisin", "temperture": 17, "time": 1360785720},
{"station": "stanley", "temperture": 16, "time": 1360785720},
{"station": "kwuntong", "temperture": 15, "time": 1360785720},
{"station": "shamshuipo", "temperture": 17, "time": 1360785720}]
```

### Items.py

```
class Hk0WeatherItem(Item):
    time = Field()
    station = Field()
    temperture = Field()
    humidity = Field()
```

### Currwx.py

```
start_urls = (
'http://www.weather.gov.hk/wxinfo/currwx/curr
entc.htm',
)
```

### Currwx.py

```
def parse(self, response):
    laststation = "
    temperture = int()
    stations = []
    hxs = HtmlXPathSelector(response)
    report = hxs.select('//div[@id="ming"]')
```

### libhk0

```
class hk0:
 stations = [
  (u'天文台', 'hko'),
  (u' 京 士 柏 ', 'kingspark'),
  (u' 黃 竹 坑 ', 'wongchukhang'),
  (u' 打鼓嶺 ', 'takwuling'),
  (u'流浮山', 'laufaushan'),
```

### libhk0

```
class hk0:
    def gettime(self, report):
    ...
    def hk0current(self, report):
    ...
```

### Agenda

- What is Open Data?
- Use of Open Source Software in web crawling.
- Starting new Open Source project hk0weather to create Open Weather Data.

# We want a easier way to access the public data.

# We want a better life with public data.

### Thank You!

sammy.hk