Challenges on Developing Tools for Exploiting Linked Open Data Cubes

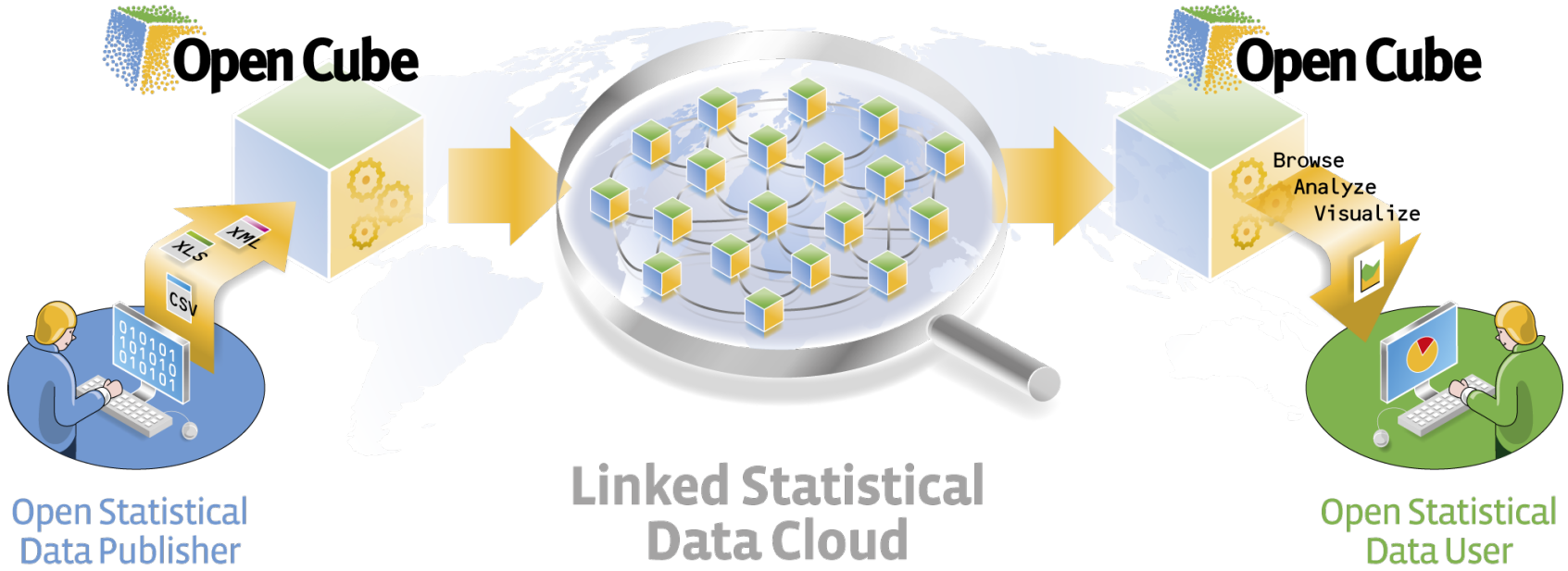
Kalampokis, Roberts, Karamanou, Tambouris, Tarabanis, Hermans

Bill Roberts

@billroberts

http://swirrl.com


Swirrl


Open Cube
XML
XLS
CSV
Open Statistical
Data Publisher
Linked Statistical
Data Cloud
Open Cube
Browse
Analyze
Visualize
Open Statistical
Data User

Pilot partners: government statistics publishers

- Central Statistics Office, Ireland
- Department for Communities and Local Government, UK
- Statistics Office of Flanders, Belgium

http://www.opencube-project.eu

http://www.opencube-toolkit.eu

Publishing Platforms

Some examples of PublishMyData in use

- http://opendatacommunities.org
- http://statisticsbeta.com
- http://data.hampshirehub.net
- http://data.surreycc.gov.uk
- http://gmdatastore.org.uk

Challenges

- Understanding the 'shape' of the data
- Selecting the slice(s) you want
- Viewing it easily
- Exporting it easily
- Accessing via API (with/without SPARQL)
- Combining data together from different datasets, different publishers
- Knowing whether and how to aggregate it

Producing RDF Data Cube data

- Grafter: http://grafter.org
- JSONstat2qb
- R2RML

Challenges

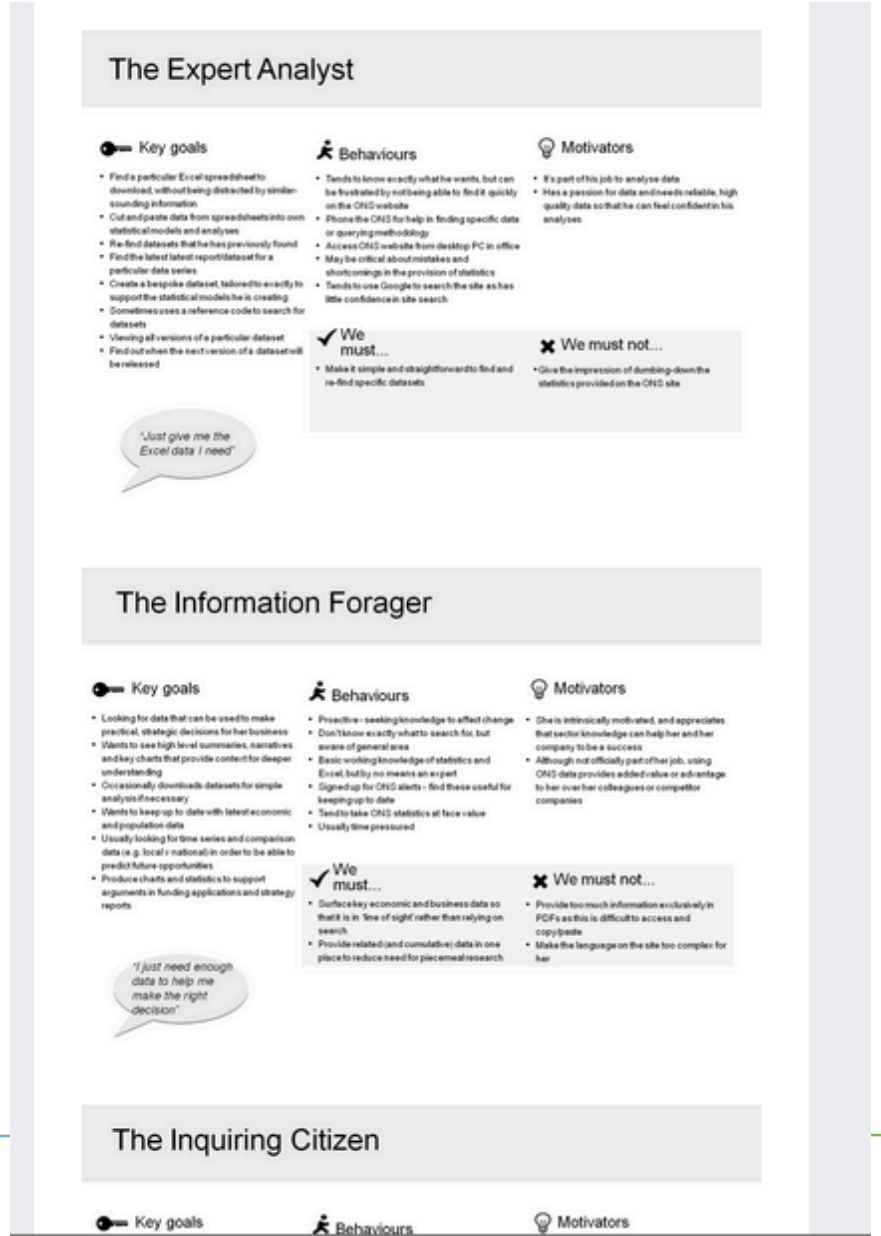
- Design issues: user interfaces and user experience

- Standardisation issues: describing the data to maximise interoperability

What kind of users?

- Analysts and researchers
- Information seekers
- Developers of visualisations and applications

http://digitalpublishing.ons.gov.uk/2014/04/02/the-persona-touch/



The Expert Analyst

Key goals

- Find a particular Excel spreadsheet to download, without being distracted by similar-sounding information
- Cut and paste data from spreadsheets into own statistical models and analyses
- Re-find datasets that he has previously found
- Find the latest latest report/dataset for a particular data series
- Create a bespoke dataset, tailored to exactly to support the statistical models he is creating
- Sometimes uses a reference code to search for datasets
- Viewing all versions of a particular dataset
- Find out when the next version of a dataset will be released

Behaviours

- Tends to know exactly what he wants, but can be frustrated by not being able to find it quickly on the ONS website
- Phone the ONS for help in finding specific data or querying methodology
- Access ONS website from desktop PC in office
- May be critical about mistakes and shortcomings in the provision of statistics
- Tends to use Google to search the site as has little confidence in site search

Motivators

- It's part of his job to analyse data
- Has a passion for data and needs reliable, high quality data so that he can feel confident in his analyses

We must...

- Make it simple and straightforward to find and re-find specific datasets

We must not...

- Give the impression of dumbing-down the statistics provided on the ONS site

"Just give me the Excel data I need"

The Information Forager

Key goals

- Looking for data that can be used to make practical, strategic decisions for her business
- Wants to see high level summaries, narratives and key charts that provide context for deeper understanding
- Occasionally downloads datasets for simple analysis if necessary
- Wants to keep up to date with latest economic and population data
- Usually looking for time series and comparison data (e.g. local v national) in order to be able to predict future opportunities
- Produce charts and statistics to support arguments in funding applications and strategy reports

Behaviours

- Proactive - seeking knowledge to affect change
- Don't know exactly what to search for, but aware of general area
- Basic working knowledge of statistics and Excel, but by no means an expert
- Signed up for ONS alerts - find these useful for keeping up to date
- Tend to take ONS statistics at face value
- Usually time pressured

Motivators

- She is intrinsically motivated, and appreciates that sector knowledge can help her and her company to be a success
- Although not officially part of her job, using ONS data provides added value or advantage to her over her colleagues or competitor companies

We must...

- Surface key economic and business data so that it is in 'line of sight' rather than relying on search
- Provide related (and cumulative) data in one place to reduce need for piecemeal research

We must not...

- Provide too much information exclusively in PDFs as this is difficult to access and copy/paste
- Make the language on the site too complex for her

"I just need enough data to help me make the right decision"

The Inquiring Citizen

Key goals

Behaviours

Motivators

```
<http://statistics.gov.scot/data/economic-activity-b                    /full-time-employment/year/2006/S12000034/people-working-full-time/count> a
<http://purl.org/linked-data/cube#Observation> ;
        <http://statistics.gov.scot/def/statistical-dim                 http://statistics.gov.scot/id/statistical-geography/S12000034> ;
        <http://statistics.gov.scot/def/statistical-dime                > <http://reference.data.gov.uk/id/year/2006> ;
        <http://purl.org/linked-data/cube#measureType> <h               ics.gov.scot/def/measure-properties/count> ;
        <http://statistics.gov.scot/def/measure-properties              00"^^<http://www.w3.org/2001/XMLSchema#integer> ;
        <http://purl.org/linked-data/sdmx/2009/attribute#               http://statistics.gov.scot/def/concept/measure-units/people-working-full-time> ;
        <http://purl.org/linked-data/cube#dataSet> <htt                 v.scot/data/economic-activity-benefits-and-tax-credits/full-time-employment> .
```

Understanding the shape of the data

- Possibly lots of dimensions
- Possibly long lists of possible values
- Possibly 'sparse' cubes
- Ensuring good performance of tools even with large data collections

Dimension	Values
Reference area http://statistics.gov.scot/def/statistical-dimensions/refArea	(8475 geographies)
Reference period http://statistics.gov.scot/def/statistical-dimensions/refPeriod	2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012.
Age http://statistics.gov.scot/def/dimension/age	65 And Over All
Gender http://statistics.gov.scot/def/dimension/gender	All Female Male
Admission Type http://statistics.gov.scot/def/dimension/admissionType	Accidents Cancer Cerebrovascular Disease (CVD) Coronary Heart Disease (CHD) Disease Of The Digestive System (DDS) Elective (Planned) Emergency Respiratory

SPREADSHEET VIEW

MEASURE *ie: what each cell shows*

Hospital Admissions: Count (Admissions)

DIMENSIONS

Reference period
- show as *table columns*
- is 2002

Age
- show as *table columns*
- **is** All

Gender
- show as *table columns*
- **is** Male

Admission Type
- show as *table columns*
- **is** Emergency

Reference Area ↓	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Country											
Scotland	236296	232698	232888	233077	243076	264180	260468	257376	254242	260157	259590

Play nicely with other tools

- CSV and Excel downloads of extracts
- R
- Visualisation libraries: d3.js, leaflet.js, Google maps/charts, Tableau…

Ordering and hierarchy of code lists

- ui:sortPriority - other options?
- skos:broader and skos:narrower are not enough
- ➔ XKOS
  - Levels
  - Knowing whether a hierarchy is exhaustive and or exclusive
  - Hierarchies change over time – e.g. administrative geography

Please select a cube to visualize on map: Cube employment

Map Type: Choropleth map

Language: en

Dimensions:

- [x] The period of time or point in time to which the measured observation refers.
- [x] Age group

(Select at most 1 levels)

- [x] Age group Level1
- [ ] Age group Level2
- [ ] Age group Level3
- [ ] Age group Level4

- [ ] The state of being male or female.

Measures:

- [x] Employment rate

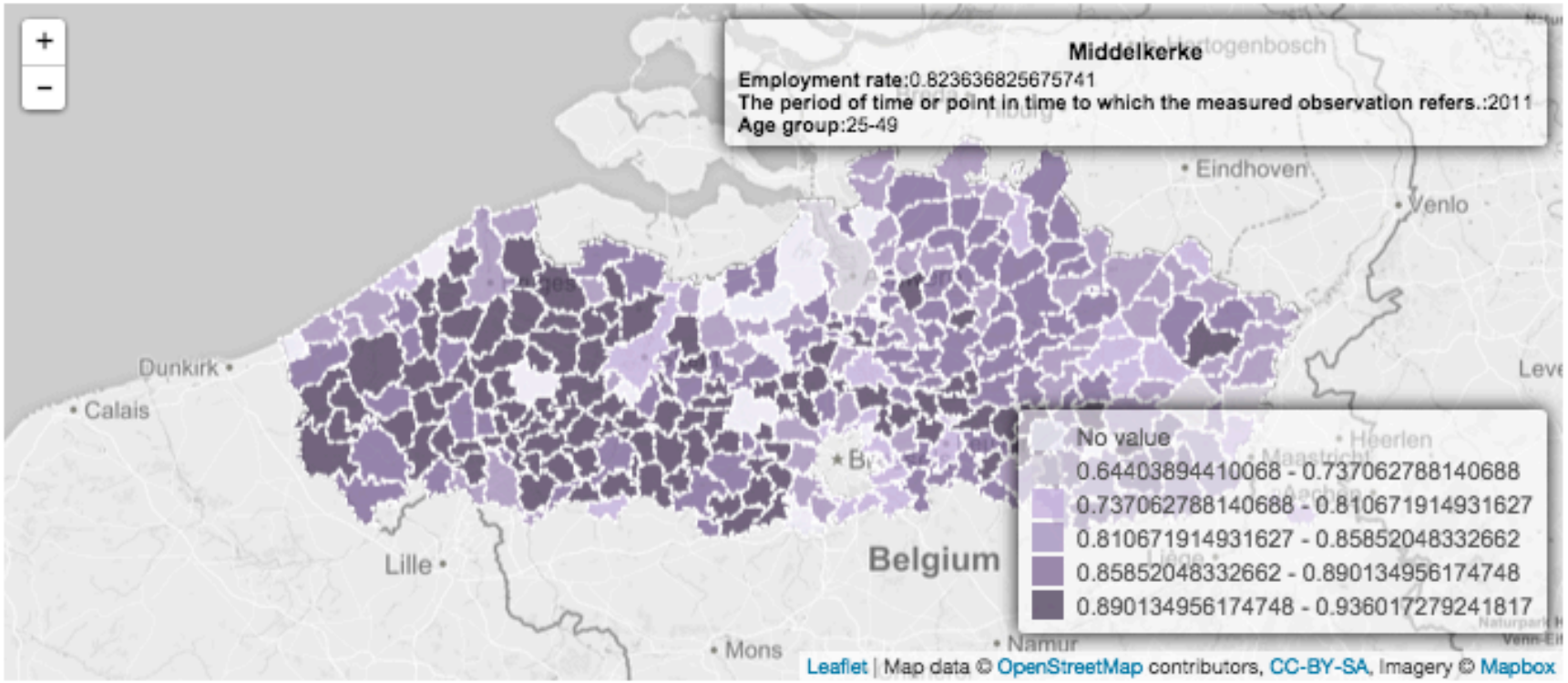
Filter:

The period of time or point in time to which the measured observation refers-: 2011

Age group: 25-49

Geography granularity

- [ ] Region
- [ ] Province
- [ ] District
- [x] Municipality



Aggregation

- Metadata to indicate 'aggregatability'
- XKOS to describe hierarchies
  - Combine the statistical data with external reference and structural data
- Ratios – link to numerator and denominator observations

Improving interoperability of data cubes

- A denser network of interlinks
- Better discovery of re-usable code lists and ontologies
- Auto-processing of equivalent concepts
- Different approaches to measure properties:
  - numberOfPeopleWithDementiaInLondon
  - numberOfPeopleWithDementia (plus refArea = London)
  - numberOfPeople (refArea=London, condition=dementia)
  - number (unitMeasure=People, refArea=London, condition=dementia)
  - obsValue

- Is there a shortlist of standard measures that would be useful?

What's missing from the RDF Data Cube vocabulary?

- Several choices for representing measures
- Aggregation
- Hierarchical code lists

- Choices/patterns for 'where to put the semantics' – measure, unit, dimensions
- Recommend use of XKOS?
- Metadata for aggregatability (including ratios)

IC-0. Datatype consistency

The RDF graph must be consistent under RDF D-entailment [RDF-MT] using a datatype map containin

IC-1. Unique DataSet

Every `qb:Observation` has exactly one associated `qb:DataSet`.

```
ASK {
  {
    # Check observation has a data set
    ?obs a qb:Observation .
    FILTER NOT EXISTS { ?obs qb:dataSet ?dataset1 . }
  } UNION {
    # Check has just one data set
    ?obs a qb:Observation ;
       qb:dataSet ?dataset1, ?dataset2 .
    FILTER (?dataset1 != ?dataset2)
  }
}
```

IC-2. Unique DSD

Every `qb:DataSet` has exactly one associated `qb:DataStructureDefinition`.

```
ASK {
  {
    # Check dataset has a dsd
    ?dataset a qb:DataSet .
    FILTER NOT EXISTS { ?dataset qb:structure ?dsd . }
  } UNION {
    # Check has just one dsd
    ?dataset a qb:DataSet ;
       qb:structure ?dsd1, ?dsd2 .
    FILTER (?dsd1 != ?dsd2)
  }
}
```

IC-3. DSD includes measure

Every `qb:DataStructureDefinition` must include at least one declared measure.

Shapes Constraint Language (SHACL)

W3C First Public Working Draft 08 October 2015

This version:
http://www.w3.org/TR/2015/WD-shacl-20151008/
Latest published version:
http://www.w3.org/TR/shacl/
Latest editor's draft:
http://w3c.github.io/data-shapes/shacl/
Editors:
Holger Knublauch, TopQuadrant, Inc.
Arthur Ryman, Invited Expert

Conclusion

- We still love Linked Data and RDF Data Cube!
- We've persuaded some statisticians to love it too
- Understand the audience and design for them
- Opportunities for improved standardisation and guidance

Thanks!

bill@swirrl.com

ekal@uom.gr