Challenges on Developing Tools for Exploiting Linked Open Data Cubes

Kalampokis, Roberts, Karamanou, Tambouris, Tarabanis, Hermans
Bill Roberts

@billroberts

http://swirrl.com
Linked Statistical Data Cloud

Open Statistical Data Publisher

Open Cube

Browse Analyze Visualize

Open Cube

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

Information Workbench

PUBLISH MY DATA
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
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
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference area</td>
<td><a href="http://statistics.gov.scot/df/statistical-dimensions/refArea">http://statistics.gov.scot/df/statistical-dimensions/refArea</a> (8475 geographies)</td>
</tr>
<tr>
<td>Age</td>
<td><a href="http://statistics.gov.scot/df/dimension/age">http://statistics.gov.scot/df/dimension/age</a> 65 And Over, All</td>
</tr>
<tr>
<td>Gender</td>
<td><a href="http://statistics.gov.scot/df/dimension/gender">http://statistics.gov.scot/df/dimension/gender</a> All, Female, Male</td>
</tr>
<tr>
<td>Admission Type</td>
<td><a href="http://statistics.gov.scot/df/dimension/admissionType">http://statistics.gov.scot/df/dimension/admissionType</a> Accidents, Cancer, Cerebrovascular Disease (CVD), Coronary Heart Disease (CHD), Disease Of The Digestive System (DDS), Elective (Planned), Emergency, Respiratory</td>
</tr>
</tbody>
</table>

MEASURE: what each cell shows

Hospital Admissions (Count(Admissions))

DIMENSIONS

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>show as table columns All, Emergency</td>
</tr>
<tr>
<td>Gender</td>
<td>show as table columns Male</td>
</tr>
</tbody>
</table>

Table:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scotland</td>
<td>236296</td>
<td>232098</td>
<td>122888</td>
<td>233877</td>
<td>243076</td>
<td>244118</td>
<td>268468</td>
<td>157376</td>
<td>251242</td>
<td>260157</td>
<td>259508</td>
</tr>
</tbody>
</table>
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
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 containing

IC-1. Unique DataSet

Every gb:Observation has exactly one associated gb:DataSet.

ASQ {

  { # Check observation has a data set
    ?obs a gb:Observation .
    FILTER NOT EXISTS { ?obs gb:dataset ?dataset1 . }
  } UNION {
    # Check has just one data set
    ?obs a gb:Observation ;
    FILTER (?dataset1 != ?dataset2)
  }

} } 

IC-2. Unique DSD

Every gb:DataSet has exactly one associated gb:DataStructureDefinition.

ASQ {

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

} } 

IC-3. DSD includes measure

Every gb: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

Copyright © 2015 W3C® (MIT, ERCIM, Keio, Beihang). W3C liability, trademark and document use rules apply.
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