Early analysis and debugging of linked open data cubes

Enrico Daga[1] Mathieu d'Aquin[1] Aldo Gangemi[2] Enrico Motta[1]

[1]**KMi - The Open University**
{enrico.daga,mathieu.daquin,enrico.motta}@open.ac.uk
[2]**ISTC-CNR - Italian National Research Council** aldo.grangemi@open.ac.uk

Second International Workshop on Semantic Statistics (SemStats)
ISWC 2014, Riva del Garda - Trentino, Italy

Summary

- *Linked Data Cubes* can be complex for both publishers and consumers
- We published the Linked Data version of **Unistats** (LinkedUp!)
- Vital: a tool to support the **early analysis** and **debugging** of linked data cubes

The Unistats Dataset

`https://www.hesa.ac.uk/unistats-dataset`

Published by UK the Higher Education Statistical Agency (HESA)
Statistics about universities: courses, subjects, employment outcomes, accommodation information, fees ...

- Collected from UK universities
- Accessible using a **Web API**
- Downloadable as a single **XML** or CSV
- Documentated on the HESA web site (HTML or PDF)

We are a university, and we might reuse this data as well ...

Example: the Employment dataset

... it shouldn't be hard, right?

```
<INSTITUTION>
[...]
 <KISCOURSE>
[...]
  <EMPLOYMENT>
    <EMPSBJ>090</EMPSBJ>
    <WORKSTUDY>95</WORKSTUDY>
    <STUDY>25</STUDY>
    <ASSUNEMP>5</ASSUNEMP>
    <BOTH>5</BOTH>
    <NOAVAIL>5</NOAVAIL>
    <WORK>60</WORK>
    <EMPPOP>25</EMPPOP>
    <EMPAGG>24</EMPAGG>
  </EMPLOYMENT>
  <EMPLOYMENT>
[...]
```

Dimensions:

- INSTITUTION
- KISCOURSE
- EMPSBJ

Measures:

- WORKSTUDY
- STUDY
- ASSUNEMP
- BOTH
- NOAVAIL
- WORK

Attributes:

- EMPPOP
- EMPAGG

The analyst needs to build familiarity with the **syntactic** *format of the data but especially with its* **semantics**, *jumping to the documentation on the web site.*

Benefits of linked data

- help to make sense of the information (everything is a **link**...);
- schema and documentation are **embedded** in the data;
- data can be queried (SPARQL) for **task-oriented** tailored views;
- we can make **links to other data**, also from third parties;
- **new dimensions** exploiting links and paths
  eg: postcodes derived from institutions

The RDF Data Cube Vocabulary

A vocabulary to publish multi-dimensional data in RDF!

```
[] a qb:Observation ;
   qb:dataSet ex:stats_OU
   ex:organization ou:the_open_university ;
   ex:inUK false ;
   ex:amount "13000"^^xsd:int ;
   ex:rounded ex:down ;
   ex:year "2013"^^xsd:gYear .

ex:organization a qb:DimensionProperty .
ex:inUK  a qb:DimensionProperty .
ex:year  a qb:DimensionProperty .
ex:amount  a qb:MeasureProperty .
ex:rounded  a qb:AttributeProperty .
```

Observations are grouped in **datasets**. An observation has:

- **dimensions** identify the observation (eg: year, subject, institution, job type)
- **measures** contain the observed values (eg: amount, working students)
- **attributes** qualify the values (eg: provisional, round off, unit)

Schema is specified

A **data structure definition** specifies components - *dimensions*, *measures* or *attributes*.

```
dset:employment a qb:DataSet;
    rdfs:label "Employment"@en ;
    skos:prefLabel "Employment" ;
    rdfs:comment "Contains data relating to the employment outcomes of students"@en;
    qb:structure dset:employmentStructure ;
    rdfs:seeAlso <http://www.hesa.ac.uk/includes/C13061_resources/
      Unistats_checkdoc_definitions.pdf?v=1.12> ;
    rdfs:isDefinedBy kis:
    .

dset:employmentStructure a qb:DataStructureDefinition;
    rdfs:label "Employment (structure)"@en ;
    skos:prefLabel "Employment (structure)" ;
    rdfs:comment "Information relating to the employment outcomes of students"@en ;
     qb:component [
    a qb:ComponentSpecification ;
    qb:dimension  dcterm:subject ;
    skos:prefLabel "Specification dimension: subject" ];
[...]
     rdfs:seeAlso <http://www.hesa.ac.uk/includes/C13061_resources/
      Unistats_checkdoc_definitions.pdf?v=1.12> ;
rdfs:isDefinedBy kis:
     .
```

RDF can be self-descriptive

The data can be extended to also include well-formed documentation:

- resources to have at least one `rdfs:label` with language information specified;
- resources to have an `rdfs:comment`;
- resources to have a single `skos:prefLabel`;
- schema elements to have `rdfs:isDefinedBy` links to the vocabulary specification;
- resources to include links to external documentation using `rdfs:seeAlso`.

Datasets

The LinkedUp Project published Unistats as Linked Open Data!
See: `http://www.linkedup-project.eu`.
SPARQL: `http://data.linkedu.eu/kis/query`

Name	Observ.
Course Stages	82642
Continuation	30115
Entry Qualifications	30013
National Student Survey Results	29381
Employment	29300
Tariffs	27732
Common Jobs	26849
Job Types	26308
Degree Classes	22548
Salaries	22430
National Student Survey NHS Results	389

7.144.510 triples in total with 327.707 `qb:Observations` in 11 `qb:DataSets`

Challenges

For the publisher - *debugging*:

- Consistency between observations and the data structure definition
  SPARQL queries are defined in the QB specification
- Coherence between RDF model and original model
  How to detect remodelling issues?
- Documentation aligned with the origin and exposed within the data
  How to assess it's quality?

For the consumer - *early analysis*:

- Data *fitness for use* needs to be evaluated by users, but building the right query requires a-priori knowledge of the data

Question

How feasible is it to use self-descriptive RDF to give an overview of the data in a way that would facilitate debugging and exploration of statistical linked open data?

Objective

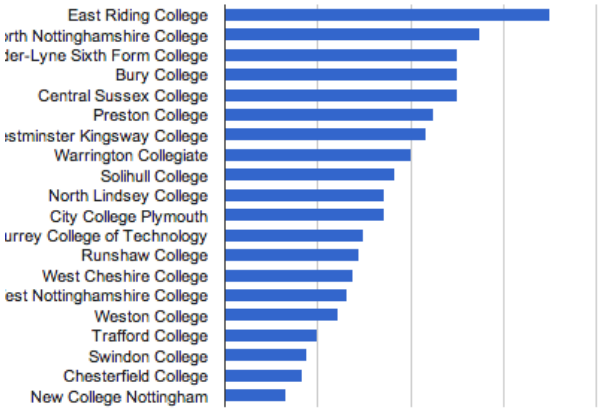
Support for data publisher and consumer:

- A methodology and a tool to debug linked data cubes and assess possible data quality issues.
- A suitable procedure in order to obtain representative samples of the data, for early analysis and evaluation

by relying only on the SPARQL endpoint

Methodology 1/3

Bar charts are simple and intuitive!



However a basic bar chart is only made up of two axis!
It can show one dimension (the **categories** observed) and one measure (the **values** compared).

Methodology 2/3

- Following the data structure definition **we can generate views automatically**, by picking a single dimension and a single measure
- We **aggregate** the measure values by applying a SPARQL aggregation function:
  - if the values datatype is numeric, we compute the *average* (AVG);
  - otherwise, we apply the `COUNT` function to the set of `DISTINCT` values, thus obtaining the number of different values for all observations under the dimension - a *diversity* score.

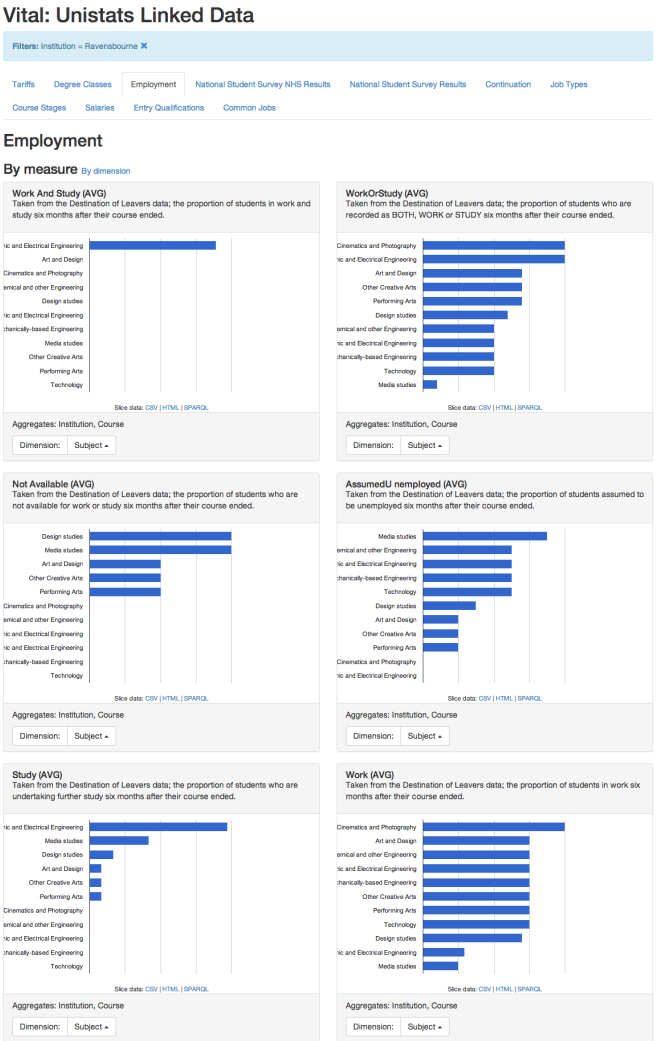
Methodology 3/3

The Employment Dataset
3 dimensions: Institution, Course and Subject
6 measures: Study, Work, Work and Study, Work or Study Assumed Unemployed, Not available.
= 18 diagrams are generated.
The view Employment: Institution + Work:

```
SELECT (?dimension as ?cat) (AVG ( ?value ) as ?nb) ?label
WHERE {
[] a qb:Observation ;
   qb:dataSet <http://data.linkedu.eu/kis/dataset/employment> ;
   <http://data.linkedu.eu/kis/ontology/institution> ?dimension ;
   <http://data.linkedu.eu/kis/ontology/work> ?value .
   optional { ?dimension skos:prefLabel ?label } .
}
GROUP BY ?dimension ?label
ORDER BY DESC(?nb) ?label
LIMIT 20
```

Following this approach our system generates 230 diagrams.

http://data.open.ac.uk/demo/vital



Debugging 1/2

The debugging methodology is a continous iteration of the following steps:

1. browse the diagrams and identify a suspicious graphs (eg: an empty diagram)
2. inspect the SPARQL query and the result sets
3. identify the error (eg: a wrong data type leading to a void result set)
4. perform the fix on the data remodelling tool

Debugging 2/2

The issues found can be grouped in the following categories:

1. insufficient/wrong documentation
2. wrong or unspecified data types
3. syntax errors in URIs
4. inconsistency between the data structure specifications and the observations.

With this tool we have been able to discover them very easily, and to fix the data remodelling procedure accordingly.

Conclusions

The Vital approach takes into account with a simple tool the basic needs of data publishers and early adopters:

- allows the exploration of data sets, dimensions, measures;
- supports the exploration of the documentation attached to the data;
- supports the publisher on detecting remodelling issues;
- it can contribute to the design of more relevant SPARQL queries from initial drafts;
- applying the tool to other data sources is straight forward, requiring minor configuration (basically, pointing to the SPARQL endpoint).

Open issues

- Consideration of other Data Cube components to debug and evaluate, such as Code Lists.
- We only use two simple aggregation functions: AVG for numeric values and COUNT for string values.
- The semantics of the measure, that may suggest specific aggregations.
- We do not consider attributes. Attributes can then affect the way measures need to be processed for aggregations.

  However our objective was to support the analyist in gaining an early understainding of the core capabilities of a statistical linked open dataset.

Thank you!

Enrico Daga
enrico.daga@open.ac.uk