Validating statistical index data represented in RDF using SPARQL queries

Jose Emilio Labra Gayo   Jose María Álvarez Rodríguez

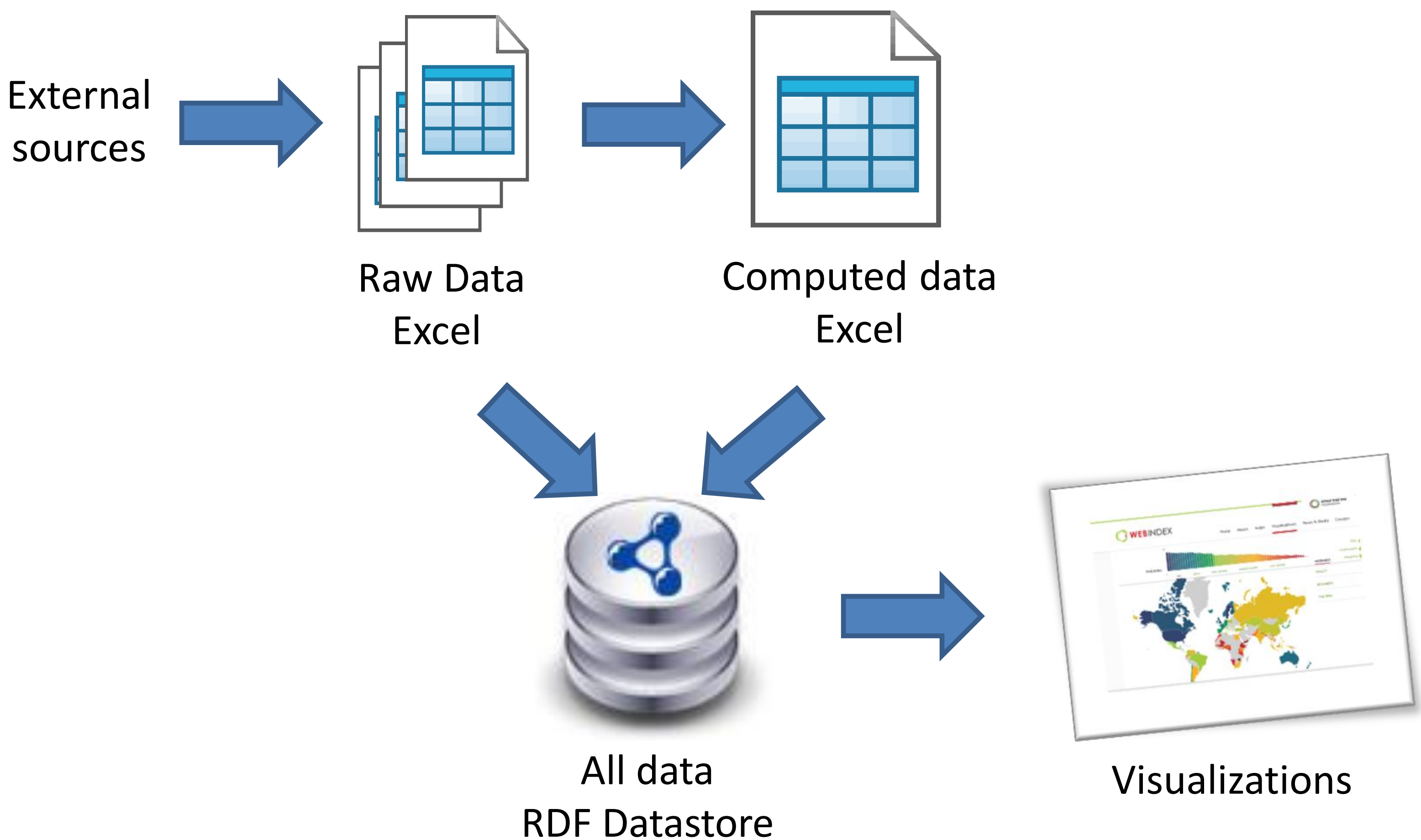
WESO Research Group
University of Oviedo, Spain

Motivation

The WebIndex Project

- Measure impact of the Web in different countries
- First publication: 2012, 2 web sites:
- http://thewebindex.org
  - Visualizations
- http://data.webfoundation.org
  - Data portal

WebIndex Workflow



Technical details

- Index made from
  - 61 countries (100 planned for 2013)
  - 85 indicators:
    - 51 Primary (questionnaires)
    - 34 Secondary (external sources)
- RDF data
  - Modeled on top of RDF Data Cube
  - More than 1 million triples
  - Linked data: DBPedia, Organizations, etc.

WebIndex computation process (1)

Simplified with one indicator, 3 years and 4 countries

Raw Data

Country	2009	2010	2011
Spain	4	5	3
Finland	4		6
Armenia	1		
Chile	6	8	

Imputed Data

Country	2009	2010	2011
Spain	4	5	3
Finland	4	5	6
Armenia	1	1	1
Chile	6	8	10.6

$$x_i = \frac{x_{i-1} + x_{i+1}}{2}$$

$$x_n \cdot \frac{\frac{x_{n-1}}{x_{n-2}} + \cdots + \frac{x_2}{x_1}}{n-1}$$

Filtered Data (Indicator A)

Country	2009	2010	2011
Spain	4	5	3
Finland	4	5	6
~~Armenia~~	~~1~~	~~1~~	~~1~~
Chile	6	8	10.6

Normalized Data (z-scores)

Country	2009	2010	2011
Spain	-0.57	-0.57	-0.92
Finland	-0.57	-0.57	-0.14
Chile	1.15	1.15	1.06

$$z = \frac{x - \mu}{\sigma}$$

More details can be found here: http://thewebindex.org/about/methodology/computation/

WebIndex computation Process (2)

Simplified with one indicator, 3 years and 4 countries

Normalized Data (z-scores)

Country	2009	2010	2011
Spain	-0.57	-0.57	-0.92
Finland	-0.57	-0.57	-0.14
Chile	1.15	1.15	1.06

Adjusted data

Country	A	B	C	D	...
Spain	8	7	9.1	7.1	...
Finland	7	8	7.1	8	...
Chile	8	9	7.6	6	...

$x_i = x_i + \delta$

Group indicators

Country	Readiness	Impact	Web	Composite
Spain	5.7	3.5	5.1	4.5
Finland	5.5	3.9	7.1	4.9
Chile	6.7	4.5	7.6	5.1

Rankings

Country	Readiness	Impact	Web	Composite
Spain	2	3	3	**3**
Finland	3	2	2	**2**
Chile	1	1	1	**1**

More details can be found here: http://thewebindex.org/about/methodology/computation/

Example of data representation in RDF

Example:

```
obs:obsM23 a qb:Observation ;
cex:computation
  [ a cex:Z-Score ;
    cex:observation obs:obsA23 ;
    cex:slice slice:sliceA09 ;
  ] ;
cex:value -0.57 ;
cex:md5-checksum "2917835203..." ;
cex:indicator indicator:A ;
cex:concept country:ESP ;
qb:dataSet dataset:A-Normalized ;
# ... other declarations omitted for brevity
```

It declares that the value of this observation was obtained as z-score of `obs:obsA23` over `slice:sliceA09`

Each observation follows the RDF Data Cube vocabulary extented with metadata about how it was obtained

Vocabulary of statistical computations: Computex

Can be seen as a RDF Data Cube specialization

Available at: http://purl.org/weso/computex

Some terms:

```
cex:Concept
cex:Indicator
cex:Computation
cex:WeightSchema
qb:Observation
...
```

Validation process

- Last year (2012)
  - Shape/template based validation
  - MD5 checksum of each observation
- This year (2013)
  - SPARQL based validation
  - 3 levels of validation
    - RDF Data Cube
    - Shapes of data
    - Computation process
  - Ultimate goal: automatically compute the index

Validation approach

We used SPARQL CONSTRUCT queries instead of ASK

IF (no error) THEN empty model

ELSE RDF graph with error information

```
CONSTRUCT {
 [ a cex:Error ;
  cex:errorParam
          [cex:name "obs"; cex:value ?obs ] ,
          [cex:name "value1"; cex:value ?value1 ] ,
          [cex:name "value2"; cex:value ?value2 ] ;
   cex:msg "Observation has two different values" . ]
} WHERE {
   ?obs a qb:Observation .
   ?obs cex:value ?value1 .
   ?obs cex:value ?value2 .
   FILTER ( ?value1 != ?value2 )
}
```

CONSTRUCT queries facilitate debugging

SPARQL queries
RDF Data Cube

RDF Data Cube integrity constraints can easily be converted from ASK to CONSTRUCT queries

```
CONSTRUCT {
  [ a cex:Error ;
    cex:errorParam [cex:name "dim"; cex:value ?dim ] ;
    cex:msg "Every Dimension must have a declared range" .
  ]
}  WHERE  {
  ?dim a qb:DimensionProperty .
  FILTER NOT EXISTS { ?dim rdfs:range []  }
}
```

SPARQL expressivity

SPARQL can express complex validation patterns.

Example: Mean

```
CONSTRUCT {
    [ a cex:Error ;
      cex:errorParam # ...omitted
      cex:msg "Mean value does not match" ] .
} WHERE {
    ?obs a qb:Observation ;
    cex:computation ?comp ;
    cex:value ?val .
    ?comp a cex:Mean .
    { SELECT (AVG(?value) as ?mean) ?comp WHERE {
          ?comp cex:observation ?obs1 .
          ?obs1 cex:value ?value ;
      } GROUP BY ?comp }
    FILTER( abs(?mean - ?val) > 0.0001) }}
```

Limitations of SPARQL expressivity

Some built-in functions are not standardized

Example: z-score employs standard deviation. It requires built-in function: `sqrt`

- Available in some SPARQL implementations

Limits of SPARQL expressivity

Ranking of values (2 approaches)

- Using GROUP_CONCAT
- Check a value against all the other values

Limits of SPARQL expressivity

Handling series with RDF Collections

Average growth: $\dfrac{\frac{v_n}{v_{n-1}} + \cdots + \frac{v_2}{v_1}}{n-1}$

```
CONSTRUCT {   # ... omitted for brevity
} WHERE {
?obs cex:computation
           [a cex:AverageGrowth; cex:observations ?ls] ;
           cex:value ?val .
           ?ls rdf:first [cex:value ?v1 ] .
   { SELECT ( SUM(?v_n / ?v_n1)/COUNT(*) as ?meanGrowth)
    WHERE {
      ?ls rdf:rest* [ rdf:first [ cex:value ?v_n ] ;
                      rdf:rest [ rdf:first [ cex:value ?v_n1 ]]] .
    }}  FILTER (abs(?meanGrowth * ?v1 - ?val) > 0.001)
}
```

Computex Validation tool

Available at:

http://herokuapp.computex.com

Work in progress

- Validates both RDF Data Cube and Computex datasets
- Generates error messages and EARL report
- Selection of validation profile
- Source code available:

  http://github.com/weso/computex

Conclusions

WebIndex = use case for RDF Validation

RDF Validation using SPARQL queries seems promising

Challenges:

- Expressivity limits of SPARQL
- Complexity of some queries

Future work

Generic validation and computation (expansion)

- RDF Data Cube = Profile
- Computex = Profile derived from RDF Data Cube
- User defined profiles
- Other profiles?