Representing Verifiable Statistical Computations as linked data

**Jose Emilio Labra Gayo**  
WESO Research group  
University of Oviedo  
Spain  
labra@uniovi.es

**Juan Castro Fernández**  
WESO Research Group  
University of Oviedo  
Spain  
juan@weso.es

**Hania Farham**  
The Web Foundation  
London  
hania@webfoundation.org

**Jose Emilio Labra Gayo**  
WESO Research group  
University of Oviedo  
Spain  
labra@uniovi.es

**Juan Castro Fernández**  
WESO Research Group  
University of Oviedo  
Spain  
juan@weso.es

**Hania Farham**  
The Web Foundation  
London  
hania@webfoundation.org

**Jose María Álvarez Rodríguez**  
Dept. Computer Science  
Carlos III University  
Spain  
josemaria.alvarez@uc3m.es
This talk in one slide

Describe the WebIndex Project

- Represents an statistical index
- Data Model based
- Computation and validation process
- Visualization
Web Index

Measure WWW's contribution to development and human rights by country

Developed by the Web Foundation

Web page:

http://thewebindex.org

Linked data portal:

http://data.webfoundation.org/webindex/2013
Technical details

Index made from

81 countries, 5 years (2007-12)

116 indicators:

84 Primary (questionnaires)
32 Secondary (external sources)

Linked data portal

Modeled on top of RDF Data Cube
Linked data: DBPedia, Organizations, etc.
Different versions

2012. Visualizations & linked data portal
   RDF representation based on RDF Data Cube
   Internal validation
   No representation of computations

2013. Include data about computations
   Goal: External agents can verify data & computations

2014. Currently in development
Webindex workflow

Data (Excel) → Conversion Excel → RDF

Computation Enrichment

RDF Datastore

Visualizations
Linked data portal
Computation process (1)

Simplified with one indicator, 3 years and 4 countries

Raw Data (Indicator A)

<table>
<thead>
<tr>
<th>Country</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Finland</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Armenia</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Chile</td>
<td>6</td>
<td>8</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Impute Data

<table>
<thead>
<tr>
<th>Country</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Finland</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Armenia</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chile</td>
<td>6</td>
<td>8</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Filter Data

<table>
<thead>
<tr>
<th>Country</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Finland</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Armenia</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chile</td>
<td>6</td>
<td>8</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Normalize Data (z-scores)

<table>
<thead>
<tr>
<th>Country</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>-0.57</td>
<td>-0.57</td>
<td>-0.92</td>
</tr>
<tr>
<td>Finland</td>
<td>-0.57</td>
<td>-0.57</td>
<td>-0.14</td>
</tr>
<tr>
<td>Armenia</td>
<td>1.15</td>
<td>1.15</td>
<td>1.06</td>
</tr>
<tr>
<td>Chile</td>
<td>1.15</td>
<td>1.15</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Mean \( x_{i} = \frac{x_{i-1} + x_{i+1}}{2} \)

Average growth \( x_{n} \cdot \frac{x_{n-1}}{x_{n-2}} + \cdots + \frac{x_{1}}{x_{1}} \)

\( z \)-score \( z = \frac{x - \mu}{\sigma} \)

More details can be found here: http://thewebindex.org/about/methodology/computation/
Simplified with one indicator, 3 years and 4 countries

### Normalize Data (z-scores)

<table>
<thead>
<tr>
<th>Country</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>-0.57</td>
<td>-0.57</td>
<td>-0.92</td>
</tr>
<tr>
<td>Finland</td>
<td>-0.57</td>
<td>-0.57</td>
<td>-0.14</td>
</tr>
<tr>
<td>Chile</td>
<td>1.15</td>
<td>1.15</td>
<td>1.06</td>
</tr>
</tbody>
</table>

### Adjust data

\[ x_{\downarrow i} = x_{\downarrow i} + \delta \]

<table>
<thead>
<tr>
<th>Country</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>8</td>
<td>7</td>
<td>9.1</td>
<td>7.1</td>
<td>...</td>
</tr>
<tr>
<td>Finland</td>
<td>7</td>
<td>8</td>
<td>7.1</td>
<td>8</td>
<td>...</td>
</tr>
<tr>
<td>Chile</td>
<td>8</td>
<td>9</td>
<td>7.6</td>
<td>6</td>
<td>...</td>
</tr>
</tbody>
</table>

### Group indicators

<table>
<thead>
<tr>
<th>Country</th>
<th>Readiness</th>
<th>Impact</th>
<th>Web</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>5.7</td>
<td>3.5</td>
<td>5.1</td>
<td>4.5</td>
</tr>
<tr>
<td>Finland</td>
<td>5.5</td>
<td>3.9</td>
<td>7.1</td>
<td>4.9</td>
</tr>
<tr>
<td>Chile</td>
<td>6.7</td>
<td>4.5</td>
<td>7.6</td>
<td>5.1</td>
</tr>
</tbody>
</table>

### Rankings

<table>
<thead>
<tr>
<th>Country</th>
<th>Readiness</th>
<th>Impact</th>
<th>Web</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Finland</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Chile</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

More details can be found here: [http://thewebindex.org/about/methodology/computation/](http://thewebindex.org/about/methodology/computation/)
WebIndex data model

Model based on RDF Data Cube
Main entity = Observation
Observations have values by years
Observations refer to indicators and countries

DataSets are published by Organizations
Datasets contain several slices
Slices group observations

Indicators are provided by Organizations
Examples
ITU = International Telecommunication Union
UN = United Nations
WB = World bank
...

<table>
<thead>
<tr>
<th>Countries</th>
<th>Indicator</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>% Broadband subscribers</td>
<td>2010</td>
</tr>
<tr>
<td>Spain</td>
<td>19.12</td>
<td>23.78</td>
</tr>
<tr>
<td>France</td>
<td>20.12</td>
<td>21.34</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

...
Excel → RDF (Turtle)
Computation process

1. First computation
   Statistics experts using Excel

2. Second computation (WESO team)
   1st. approach: SPARQL Update queries
      Can reuse the validation queries
      Declarative approach
      Problem: Efficiency & debugging
   2nd. approach: Special purpose program
      Performs computations and adds metadata

http://www.github.com/weso/wiCompute
**Computation representation**

**Computex Vocabulary**

Describes statistical computation procedures

Compatible with RDF Data Cube

**Some terms:**

<table>
<thead>
<tr>
<th><strong>cex:Concept</strong></th>
<th>Entities that are being indexed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cex:Indicator</strong></td>
<td>Dimension whose values add information to the index</td>
</tr>
<tr>
<td><strong>cex:Computation</strong></td>
<td>Represents the different computation types</td>
</tr>
<tr>
<td></td>
<td>It can be:</td>
</tr>
<tr>
<td></td>
<td>cex:Raw, cex:Mean, cex:Increment, cex:Copy,</td>
</tr>
<tr>
<td></td>
<td>cex:Z-Score, cex:Ranking, cex:AverageGrowth,</td>
</tr>
<tr>
<td></td>
<td>cex:WeightedMean</td>
</tr>
<tr>
<td><strong>cex:WeightSchema</strong></td>
<td>Weight schema for a list of indicators</td>
</tr>
</tbody>
</table>
Example of a computed observation

Normalization using z-score

\[ z = \frac{x - \mu}{\sigma} \]

\[ = \frac{23.78 - 12.816}{12.766} \approx 0.859 \]

\[ \text{z-score} \]

URI of computed observation:
Verifying linked data contents

Once the linked data has been published
How can an external agent verify it?

2 approaches:

- SPARQL Queries
- Shape expressions
SPARQL validation

CONSTRUCT queries like:

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

Detects if one observation has more than 1 value
More advanced queries like:

```
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 {
    ?obs1 cex:value ?value ;
  } GROUP BY ?comp }
} FILTER (abs(?mean - ?val) > 0.0001)
```

Detects if an observation whose computation is declared as the mean is really the mean.
Shape Expressions validation

Shape expressions declare the shape of RDF data

- Human readable and machine processable

Shape Expressions for team communication

- Developers know which triples must generate/consume

```xml
<Observation> {
    rdf:type (qb:Observation),
    cex:value xsd:float ?,
    dc:issued xsd:dateTime,
    rdfs:label xsd:string ?,
    qb:dataSet @<DataSet>,
    cex:ref-area @<Country>,
    cex:indicator @<Indicator>,
    cex:ref-year xsd:gYear,
    cex:computation @<Computation>
}
```
Protection

Protection tool: Wesby, Inspired by Pubby

Enables easy customization by templates

Different templates are chosen based on rdf:type

Data load on demand

SPARQL queries

Responsive design and mobile friendly
Visualization

Example: Template for Observations
http://data.webfoundation.org/webindex/v2013/observation/obs8003
Visualization

Example: Template for Countries
http://data.webfoundation.org/webindex/v2013/country/ESP
Conclusions

WebIndex:

Linked data portal (medium size ≈ 3,5 mill triples)
It adds data about computation
Computations represented as linked data
We explored some possibilities for validation
SPARQL validation: very expressive, declarative
Shape Expressions: more readable
Visualization by templates
Future work

Computex vocabulary was a first attempt
  Further work to employ it in similar projects

Visualization of computations
  Define wesby templates to visualize computations

Question: Was it worth the effort?
  Producer/consumers balance
  We produced data that can be externally verified
  However, we still don't have consumers who need it
End of presentation

More info:
WESO Research group
http://www.weso.es