Representing Verifiable Statistical Computations as linked data

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

Hania Farham
The Web Foundation
London
hania@webfoundation.org

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

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

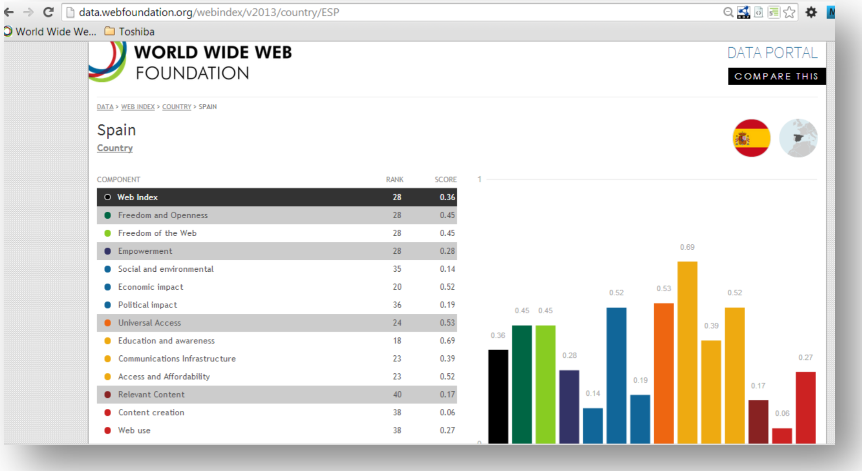
- Developed by the Web Foundation

Web page:

- http://thewebindex.org

Linked data portal:

- http://data.webfoundation.org/webindex/2013



data.webfoundation.org/webindex/v2013/country/ESP

DATA > WEB INDEX > COUNTRY > SPAIN

Spain

Country

COMPONENT	RANK	SCORE
○ Web Index	28	0.36
● Freedom and Openness	28	0.45
● Freedom of the Web	28	0.45
● Empowerment	28	0.28
● Social and environmental	35	0.14
● Economic impact	20	0.52
● Political impact	36	0.19
● Universal Access	24	0.53
● Education and awareness	18	0.69
● Communications Infrastructure	23	0.39
● Access and Affordability	23	0.52
● Relevant Content	40	0.17
● Content creation	38	0.06
● Web use	38	0.27

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.

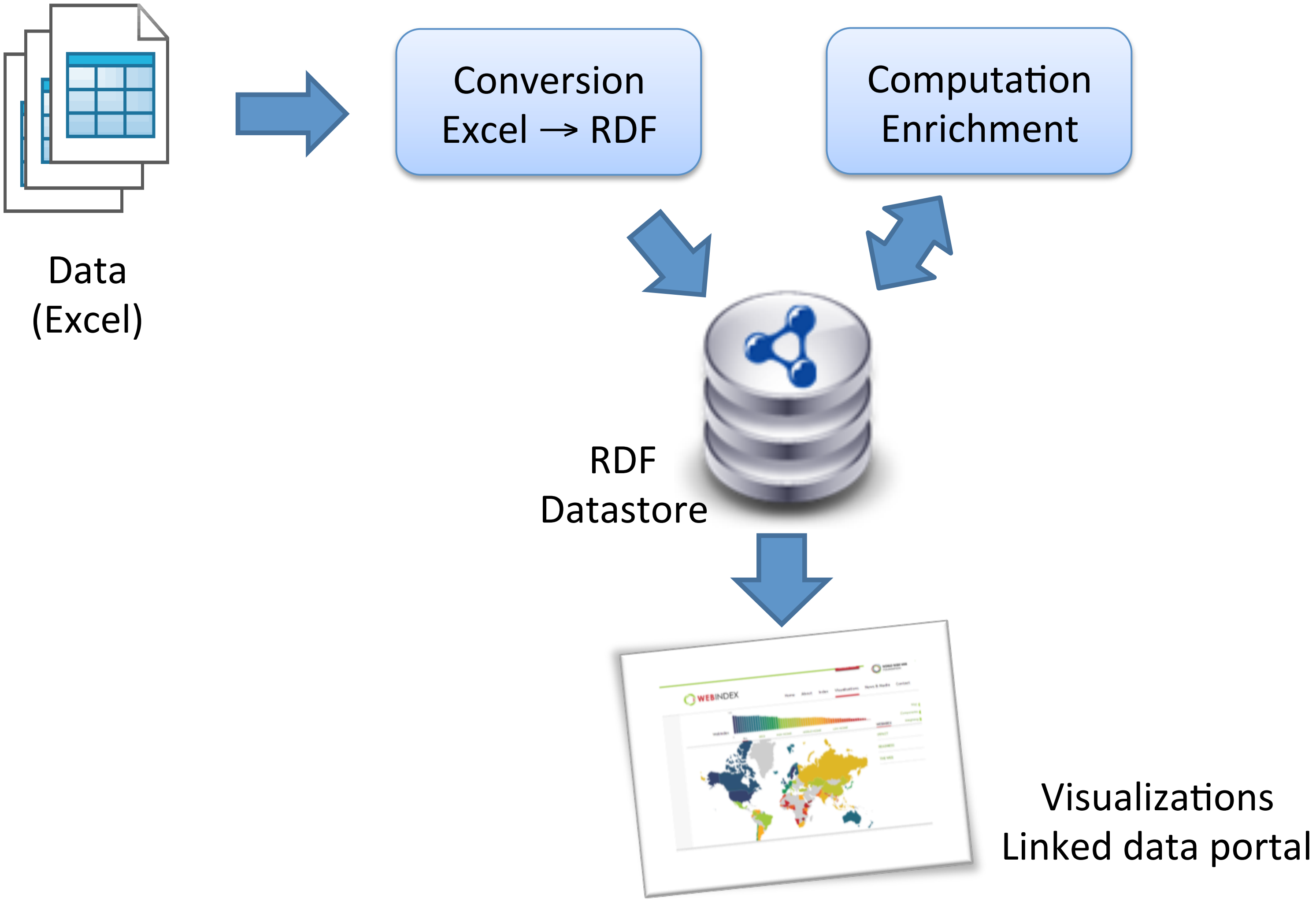
Different versions

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

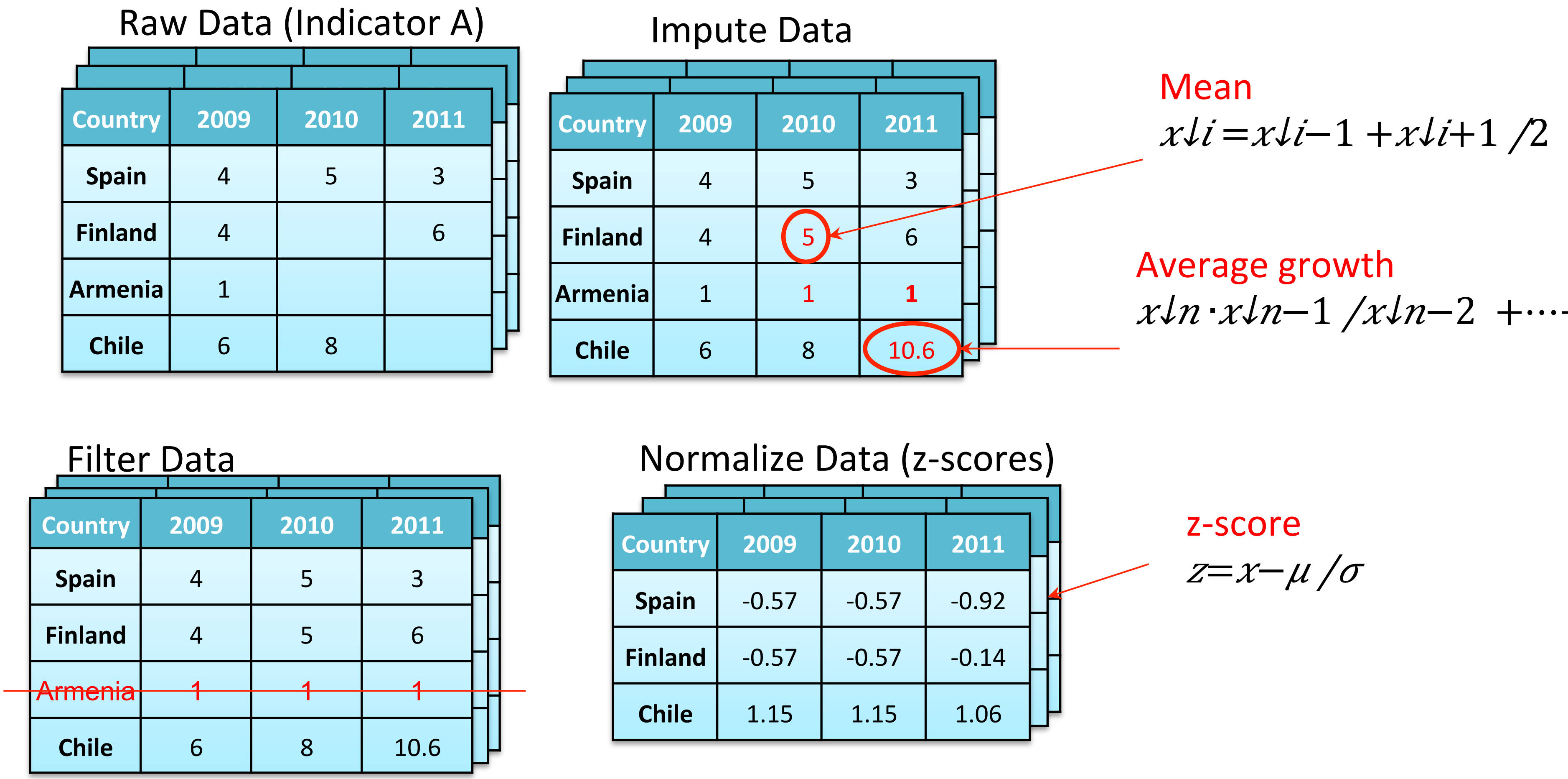
Computation
Enrichment

RDF
Datastore

Visualizations
Linked data portal

Computation process (1)

Simplified with one indicator, 3 years and 4 countries



Raw Data (Indicator A)

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

Impute Data

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

Mean

$x\downarrow i = x\downarrow i-1 + x\downarrow i+1 /2$

Average growth

$x\downarrow n \cdot x\downarrow n-1 / x\downarrow n-2 + \cdots$

Filter Data

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

Normalize 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-score

$z = x - \mu / \sigma$

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

Computation Process (2)

Simplified with one indicator, 3 years and 4 countries

Normalize 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

Adjust 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	...

$x\downarrow i = x\downarrow i + \delta$

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/

WebIndex data model

Model based on RDF Data Cube

Main entity = Observation

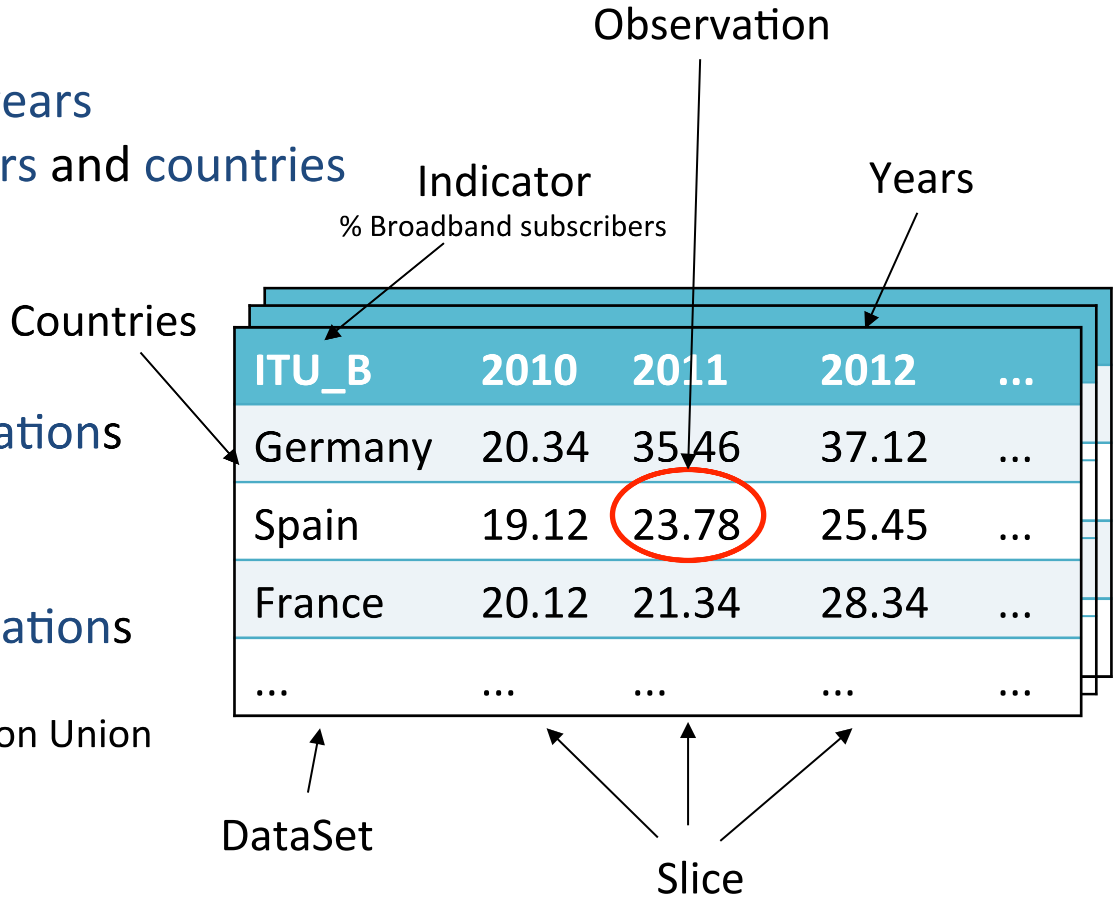
- Observations have values by years
- Observations refer to indicators and countries

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
  - ...



Observation

Indicator

% Broadband subscribers

Years

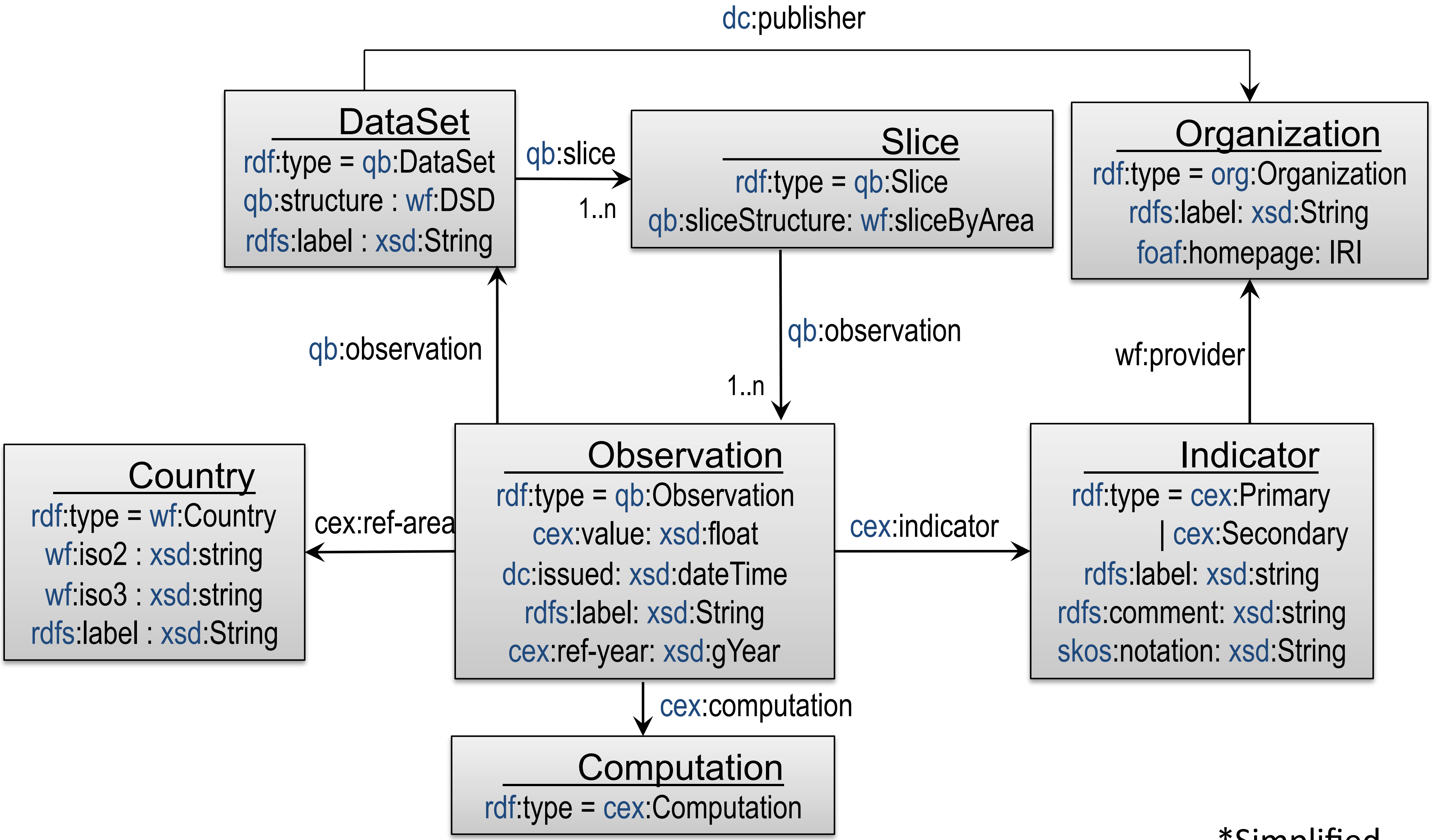
Countries

ITU_B	2010	2011	2012	...
Germany	20.34	35.46	37.12	...
Spain	19.12	23.78	25.45	...
France	20.12	21.34	28.34	...
...	...	...	...	...

DataSet

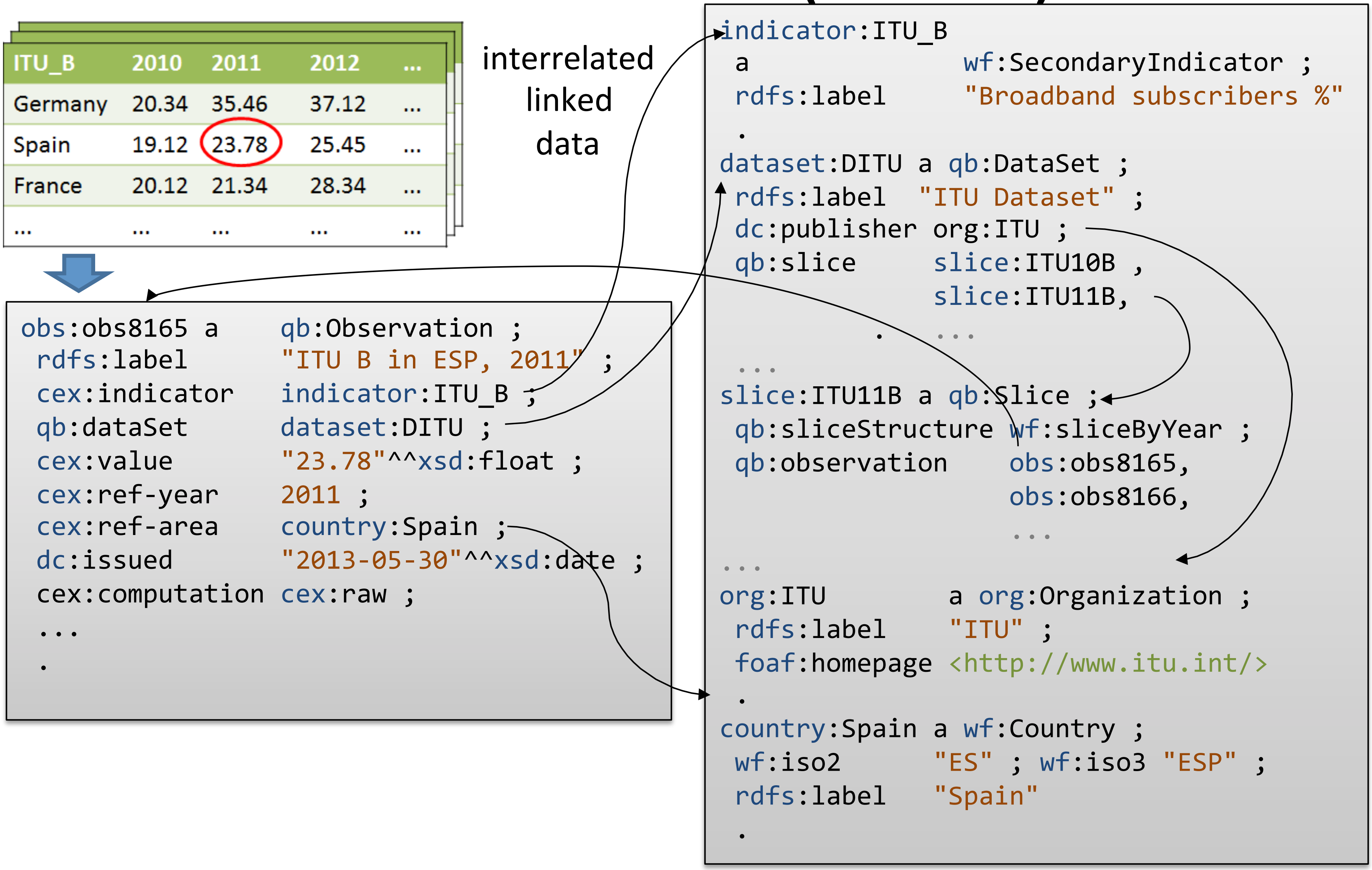
Slice

Data model*



*Simplified

Excel → RDF (Turtle)



ITU_B	2010	2011	2012	...
Germany	20.34	35.46	37.12	...
Spain	19.12	23.78	25.45	...
France	20.12	21.34	28.34	...
...	...	...	...	...

interrelated
linked
data

```
obs:obs8165 a    qb:Observation ;
 rdfs:label      "ITU B in ESP, 2011" ;
 cex:indicator   indicator:ITU_B ;
 qb:dataSet      dataset:DITU ;
 cex:value       "23.78"^^xsd:float ;
 cex:ref-year    2011 ;
 cex:ref-area    country:Spain ;
 dc:issued       "2013-05-30"^^xsd:date ;
 cex:computation cex:raw ;
 ...
 .
```

```
indicator:ITU_B
 a              wf:SecondaryIndicator ;
 rdfs:label     "Broadband subscribers %"
 .
dataset:DITU a qb:DataSet ;
 rdfs:label  "ITU Dataset" ;
 dc:publisher org:ITU ;
 qb:slice      slice:ITU10B ,
               slice:ITU11B,
            .  ...
 ...
slice:ITU11B a qb:Slice ;
 qb:sliceStructure wf:sliceByYear ;
 qb:observation    obs:obs8165,
                   obs:obs8166,
                   ...
...
org:ITU          a org:Organization ;
 rdfs:label     "ITU" ;
 foaf:homepage <http://www.itu.int/>
 .
country:Spain a wf:Country ;
 wf:iso2      "ES" ; wf:iso3 "ESP" ;
 rdfs:label   "Spain"
 .
```

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

Scala

wiCompute

http://www.github.com/weso/wiCompute

Computation representation

Computex Vocabulary

- Describes statistical computation procedures
- Compatible with RDF Data Cube

Some terms:

cex:Concept	Entities that are beind indexed
cex:Indicator	Dimension whose values add information to the index
cex:Computation	Represents the different computation types It can be: cex:Raw, cex:Mean, cex:Increment, cex:Copy, cex:Z-Score, cex:Ranking, cex:AverageGrowth, cex:WeightedMean
cex:WeightSchema	Weight schema for a list of indicators

Example of a computed observation

```
obs:c39049  a    qb:Observation ;
 rdfs:label      "ITU B in ESP, 2011, Normalized" ;
 cex:indicator   indicator:ITU_B ;
 qb:dataSet      dataset:computed366 ;
 cex:value       "0.859"^^xsd:double ;
 cex:ref-year    2011 ;
 cex:ref-area    country:Spain ;
 cex:computation wi-comp:comp39050 ;
...
 .
```

Normalization using z-score

$$z = x - \mu / \sigma \qquad = 23.78 - 12.816 / 12.766 = 0$$

✓ ok!

```
wi-comp:39050  a cex:Normalize ;
 cex:stdDesv      "12.766"^^xsd:double ;
 cex:mean         "12.816"^^xsd:double ;
 cex:slice       wi-slice:sliceITUB_2011 ;
 cex:observation obs:obs8165 ;
.
```

```
obs:obs8165       a qb:Observation ;
 cex:value        "23.78"^^xsd:double ;
 ...
.
```

```
wi-slice:sliceITU_B_2011 a    qb:Slice ;
 qb:observation  obs:8471,
                 obs:8434, ...;
.
```

URI of computed observation:
http://data.webfoundation.org/webindex/v2013/observation/computed_2011_1386752461095_39049

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 .
   ?obs cex:value ?value1 .
   ?obs cex:value ?value2 .
 FILTER ( ?value1 != ?value2 )
}
```

Detects if one observation has more than 1 value

SPARQL validation

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 {
        ?comp cex:observation ?obs1 .
        ?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

```
<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>
}
```

Documentation http://weso.github.io/wiDoc

Visualization

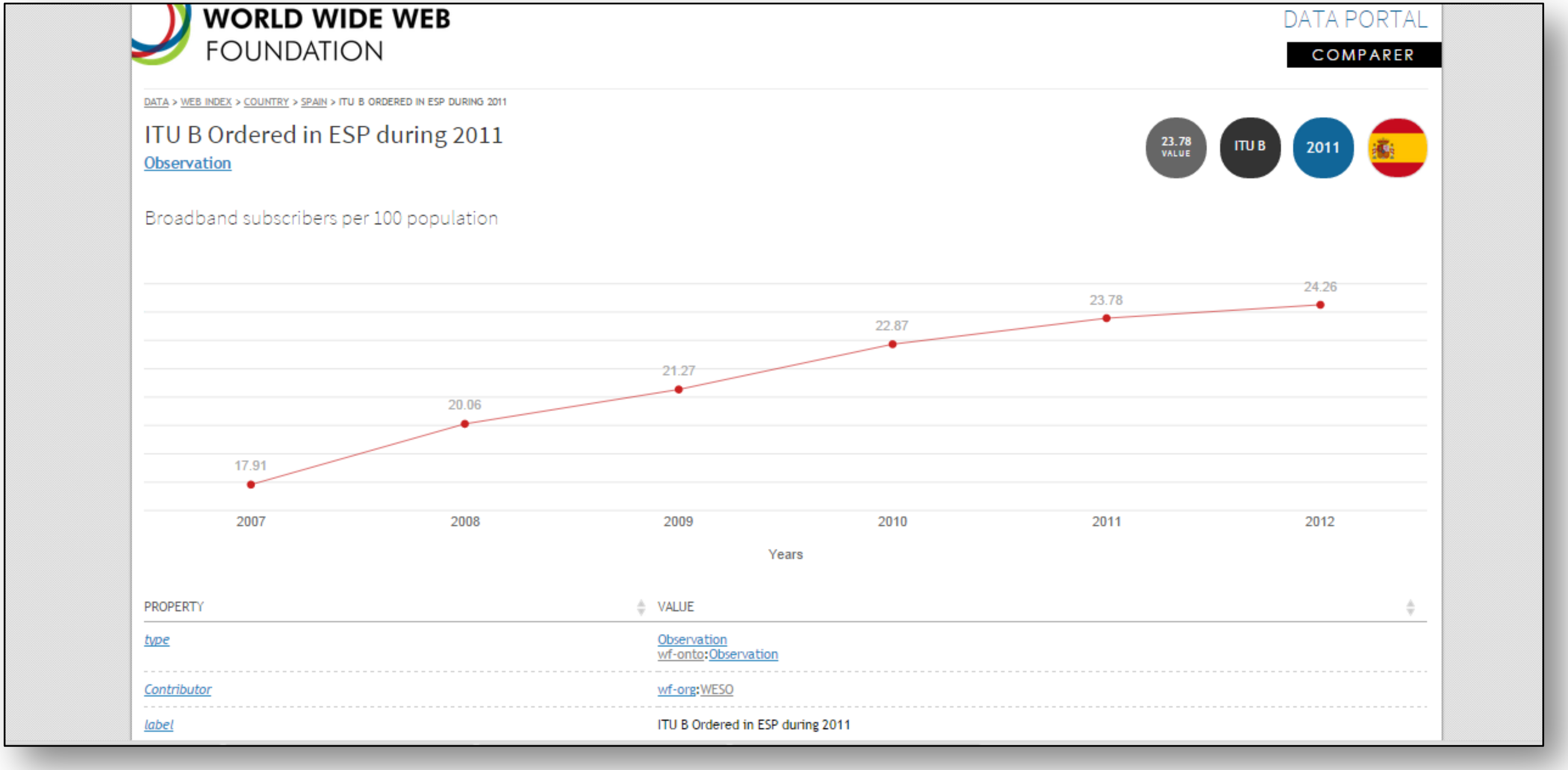
Wesby
Linked Open Data Viewer

- Visualization 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

Wesby
Linked Open Data Viewer

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



WORLD WIDE WEB FOUNDATION

DATA PORTAL

COMPARER

DATA > WEB INDEX > COUNTRY > SPAIN > ITU B ORDERED IN ESP DURING 2011

ITU B Ordered in ESP during 2011

Observation

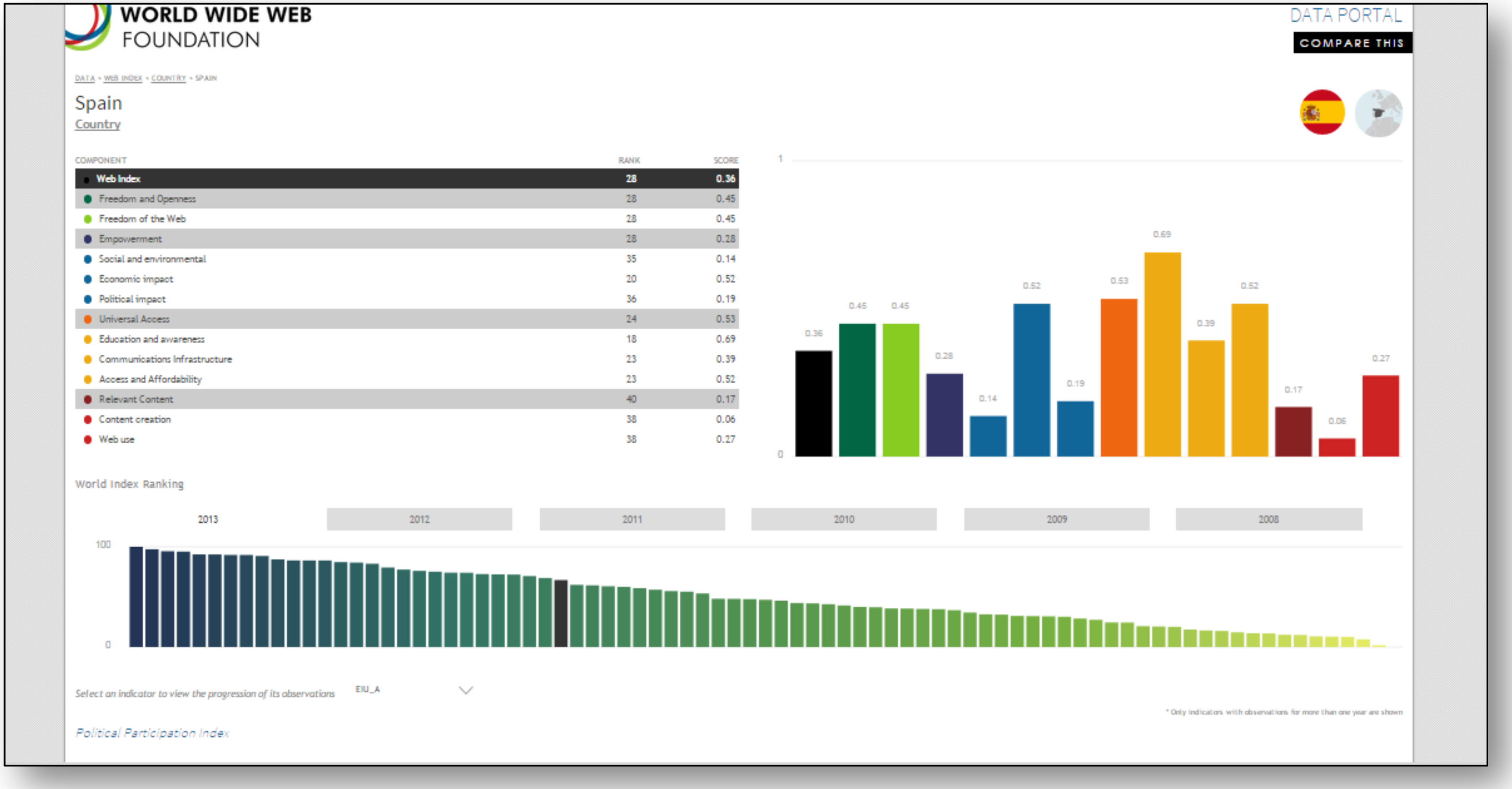
Broadband subscribers per 100 population

PROPERTY	VALUE
type	Observation wf-onto:Observation
Contributor	wf-org:WESO
label	ITU B Ordered in ESP during 2011

Visualization

Wesby
Linked Open Data Viewer

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



Spain
Country

COMPONENT	RANK	SCORE
● Web Index	28	0.36
● Freedom and Openness	28	0.45
● Freedom of the Web	28	0.45
● Empowerment	28	0.28
● Social and environmental	35	0.14
● Economic impact	20	0.52
● Political impact	36	0.19
● Universal Access	24	0.53
● Education and awareness	18	0.69
● Communications Infrastructure	23	0.39
● Access and Affordability	23	0.52
● Relevant Content	40	0.17
● Content creation	38	0.06
● Web use	38	0.27

World Index Ranking

Select an indicator to view the progression of its observations EIU_A

* Only indicators with observations for more than one year are shown

Political Participation Index

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