Converting WHO's Global Health Observatory Data to RDF

Amrapali Zaveri

AKSW, Institut für Informatik

Outline

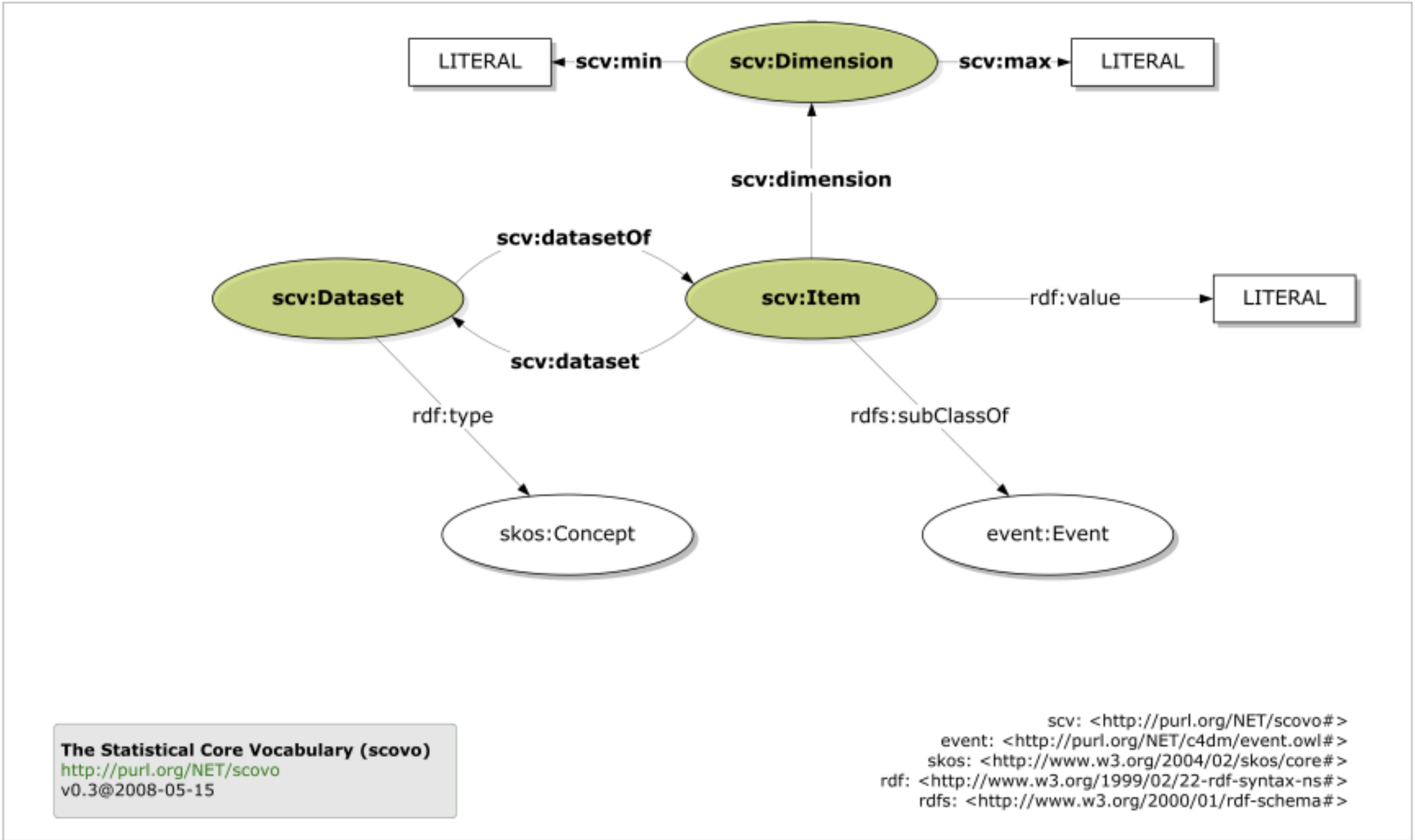
- Background
- What is SCOVO?
- Semi-automated approach
- OntoWiki's CSV Import plug-in
- SCOVOfied WHO data
- Challenges and Future Work
- References

Background

- Biomedical statistical data
  - Published as Excel sheets
- Advantage
  - Readable by humans
- Disadvantages
  - Cannot be queried efficiently
  - Difficult to integrate with other data (in different formats)
- Our approach
  - Converting data into a single data model - RDF
  - Using SCOVO (Statistical Core Vocabulary)*
    - designed particularly to represent multidimensional statistical data using RDF.

**Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. Scovo: Using statistics on the web of data. In ESWC, pages 708–722, 2009.*

What is SCOVO?



The Statistical Core Vocabulary (scovo)
http://purl.org/NET/scovo
v0.3@2008-05-15

scv: <http://purl.org/NET/scovo#>
event: <http://purl.org/NET/c4dm/event.owl#>
skos: <http://www.w3.org/2004/02/skos/core#>
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>

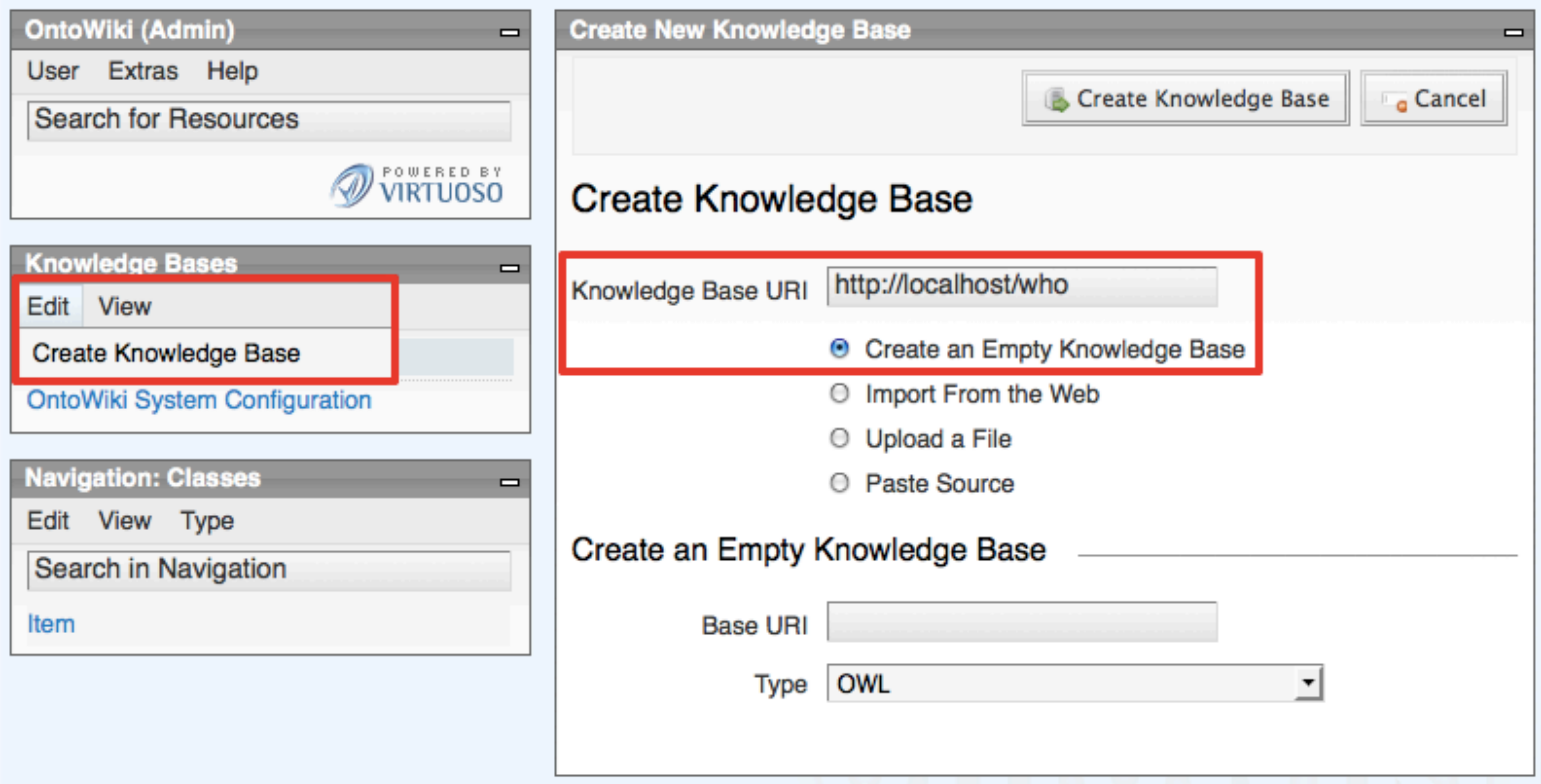
Fig.: The Statistical Core Vocabulary (scovo)

Semi-automated approach

- Transforming CSV to RDF in a fully automated way is not feasible.
    - Dimensions may often be encoded in heading or label of a sheet
- Our semi-automatic approach:
    - As a plug-in in OntoWiki#
        - a semantic collaboration platform developed by the AKSW research group.
    - A CSV file is converted into RDF using SCOVO

# Sören Auer, Sebastian Tramp (geb. Dietzold), Jens Lehmann, and Thomas Riechert: OntoWiki: A Tool for Social Semantic Collaboration In: Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge CKC 2007 at the 16th

1. Create Knowledge Base



Create Knowledge Base

Knowledge Base URI: http://localhost/who

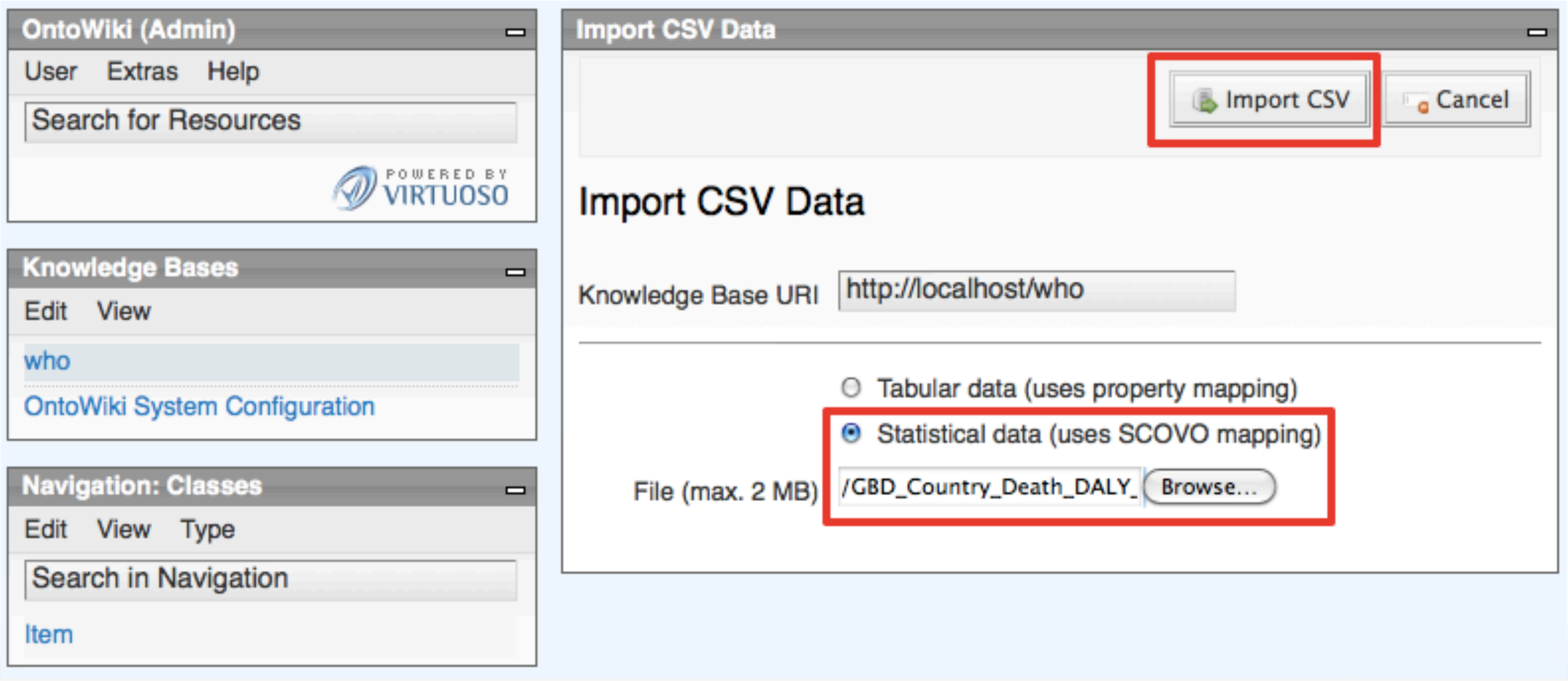
● Create an Empty Knowledge Base
○ Import From the Web
○ Upload a File
○ Paste Source

Create an Empty Knowledge Base

Base URI:

Type: OWL

2. Import a CSV file



Import CSV Data

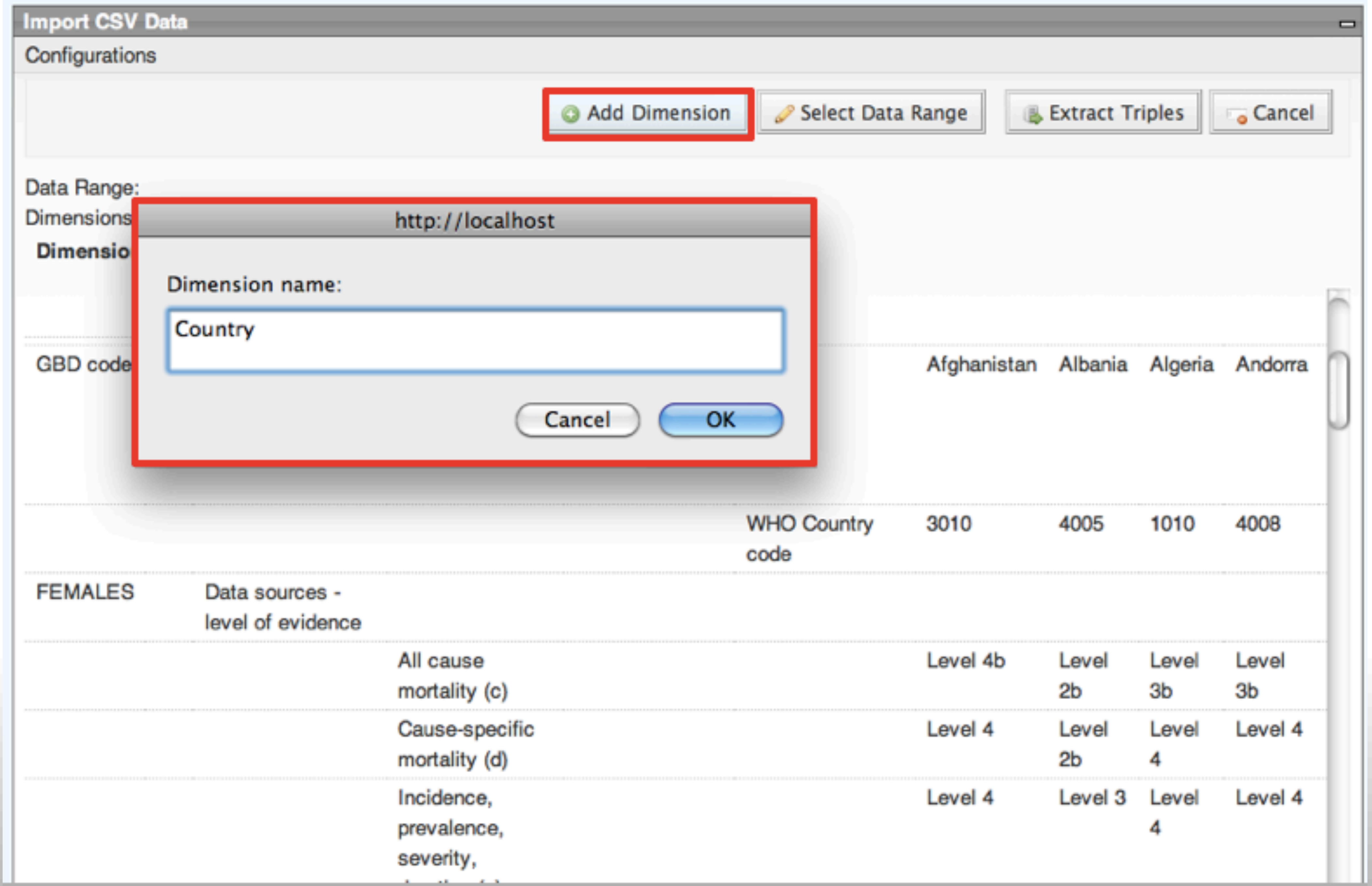
Knowledge Base URI: http://localhost/who

○ Tabular data (uses property mapping)

● Statistical data (uses SCOVO mapping)

File (max. 2 MB): /GBD_Country_Death_DALY_ [Browse...]

3. Define dimensions



Data Range:

Dimensions

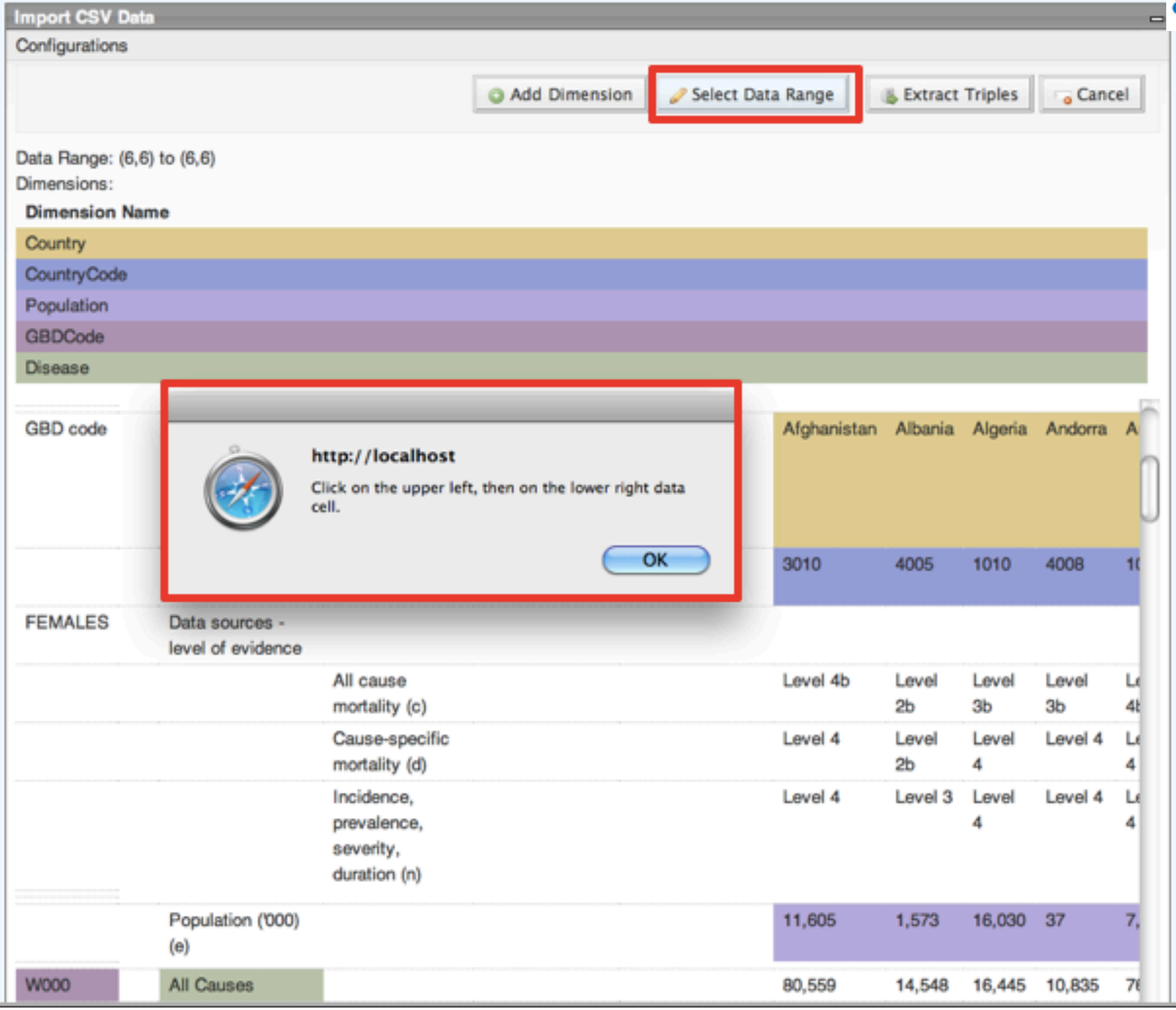
Dimensio

Dimension name:

Country

GBD code				Afghanistan	Albania	Algeria	Andorra
			WHO Country code	3010	4005	1010	4008
FEMALES	Data sources - level of evidence						
		All cause mortality (c)		Level 4b	Level 2b	Level 3b	Level 3b
		Cause-specific mortality (d)		Level 4	Level 2b	Level 4	Level 4
		Incidence, prevalence, severity,		Level 4	Level 3	Level 4	Level 4

4. Define data range



Import CSV Data

Configurations

Add Dimension | Select Data Range | Extract Triples | Cancel

Data Range: (6,6) to (6,6)

Dimensions:

Dimension Name
Country
CountryCode
Population
GBDCode
Disease

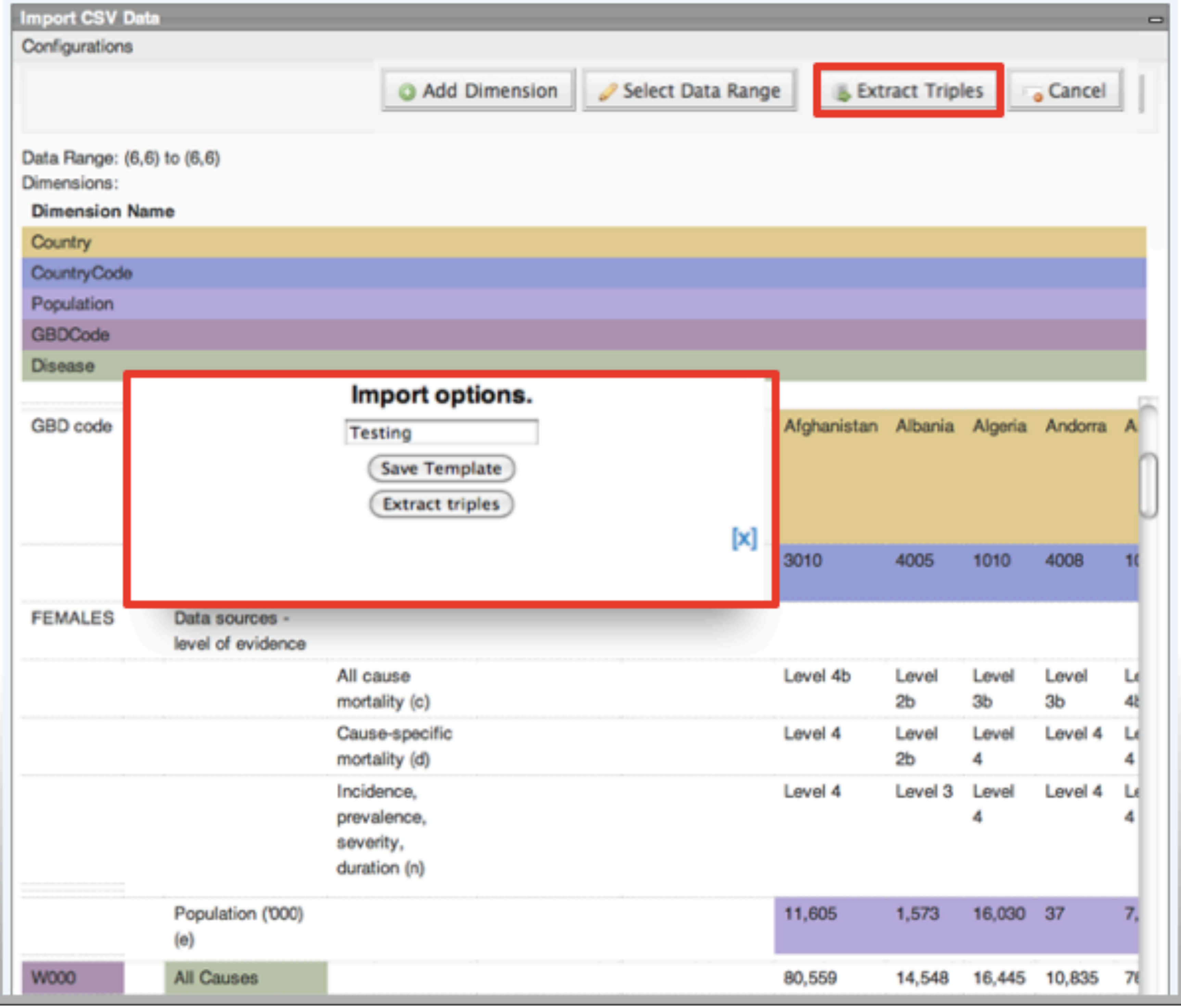
http://localhost

Click on the upper left, then on the lower right data cell.

OK

GBD code			Afghanistan	Albania	Algeria	Andorra	A
			3010	4005	1010	4008	1
FEMALES	Data sources - level of evidence						
		All cause mortality (c)	Level 4b	Level 2b	Level 3b	Level 3b	Le 4b
		Cause-specific mortality (d)	Level 4	Level 2b	Level 4	Level 4	Le 4
		Incidence, prevalence, severity, duration (n)	Level 4	Level 3	Level 4	Level 4	Le 4
	Population ('000) (e)		11,605	1,573	16,030	37	7,
W000	All Causes		80,559	14,548	16,445	10,835	7

5. Save template, extract triples



Import CSV Data

Configurations

Data Range: (6,6) to (6,6)

Dimensions:

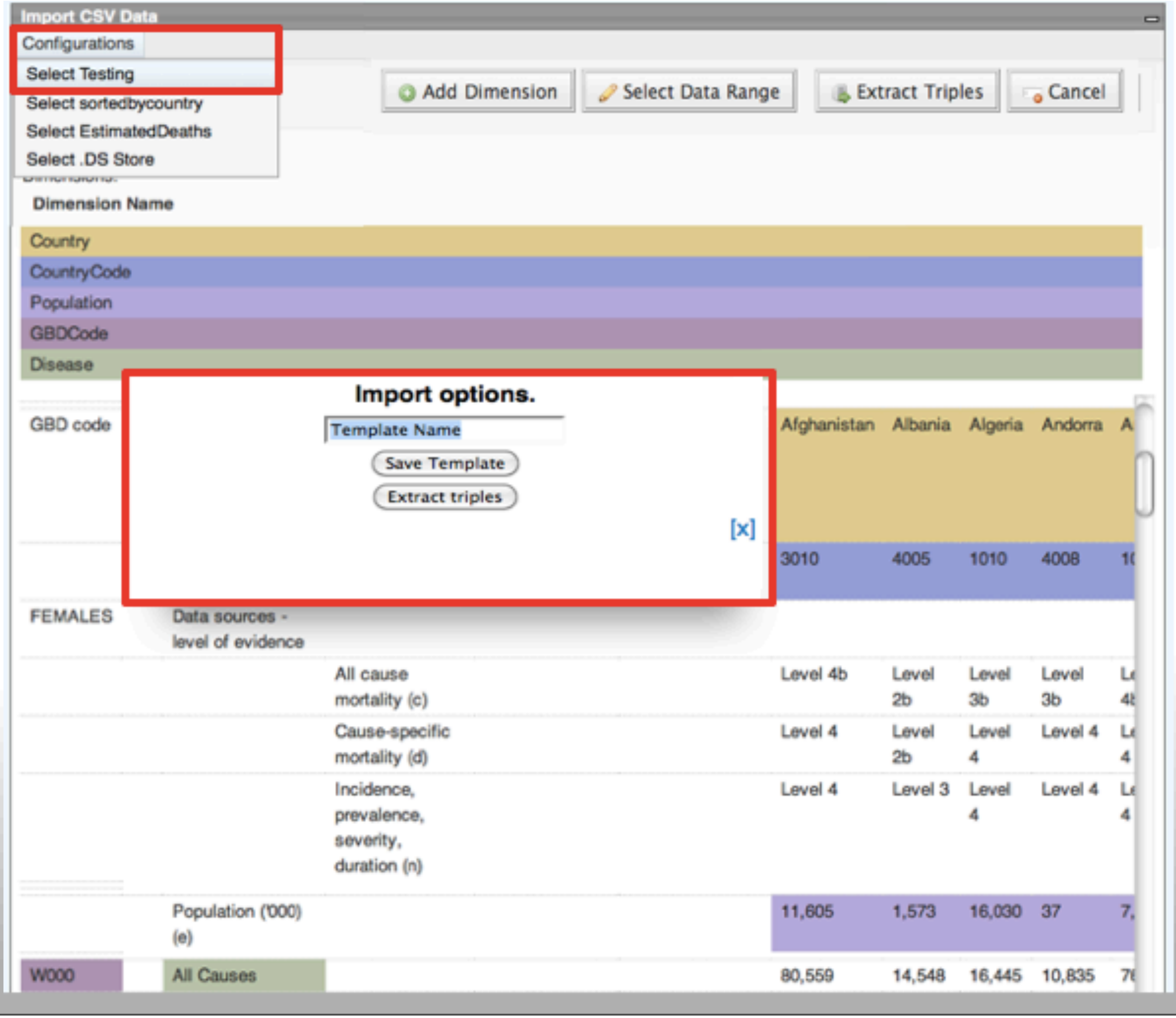
Dimension Name
Country
CountryCode
Population
GBDCode
Disease

Import options.

Testing

GBD code			Afghanistan	Albania	Algeria	Andorra	A
			3010	4005	1010	4008	10
FEMALES	Data sources - level of evidence						
		All cause mortality (c)	Level 4b	Level 2b	Level 3b	Level 3b	Le 4b
		Cause-specific mortality (d)	Level 4	Level 2b	Level 4	Level 4	Le 4
		Incidence, prevalence, severity, duration (n)	Level 4	Level 3	Level 4	Level 4	Le 4
	Population ('000) (e)		11,605	1,573	16,030	37	7,
W000	All Causes		80,559	14,548	16,445	10,835	7

6. Re-use template for similar files



Dimensions:

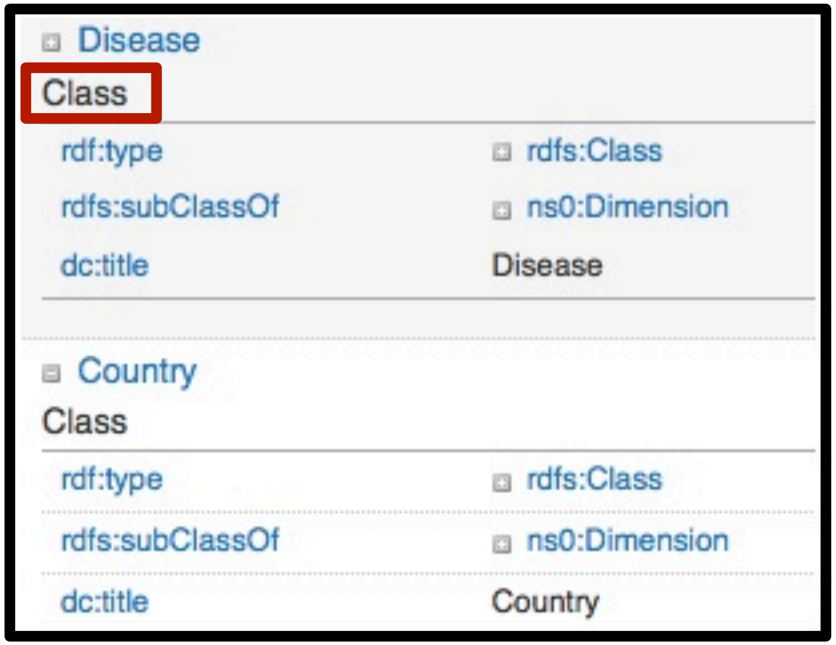
Dimension Name

- Country
- CountryCode
- Population
- GBDCode
- Disease

GBD code			Afghanistan	Albania	Algeria	Andorra	A
			3010	4005	1010	4008	10
FEMALES	Data sources - level of evidence						
		All cause mortality (c)	Level 4b	Level 2b	Level 3b	Level 3b	Le 4b
		Cause-specific mortality (d)	Level 4	Level 2b	Level 4	Level 4	Le 4
		Incidence, prevalence, severity, duration (n)	Level 4	Level 3	Level 4	Level 4	Le 4
	Population ('000) (e)		11,605	1,573	16,030	37	7,
W000	All Causes		80,559	14,548	16,445	10,835	76

7. View resources

OntoWiki



Disease

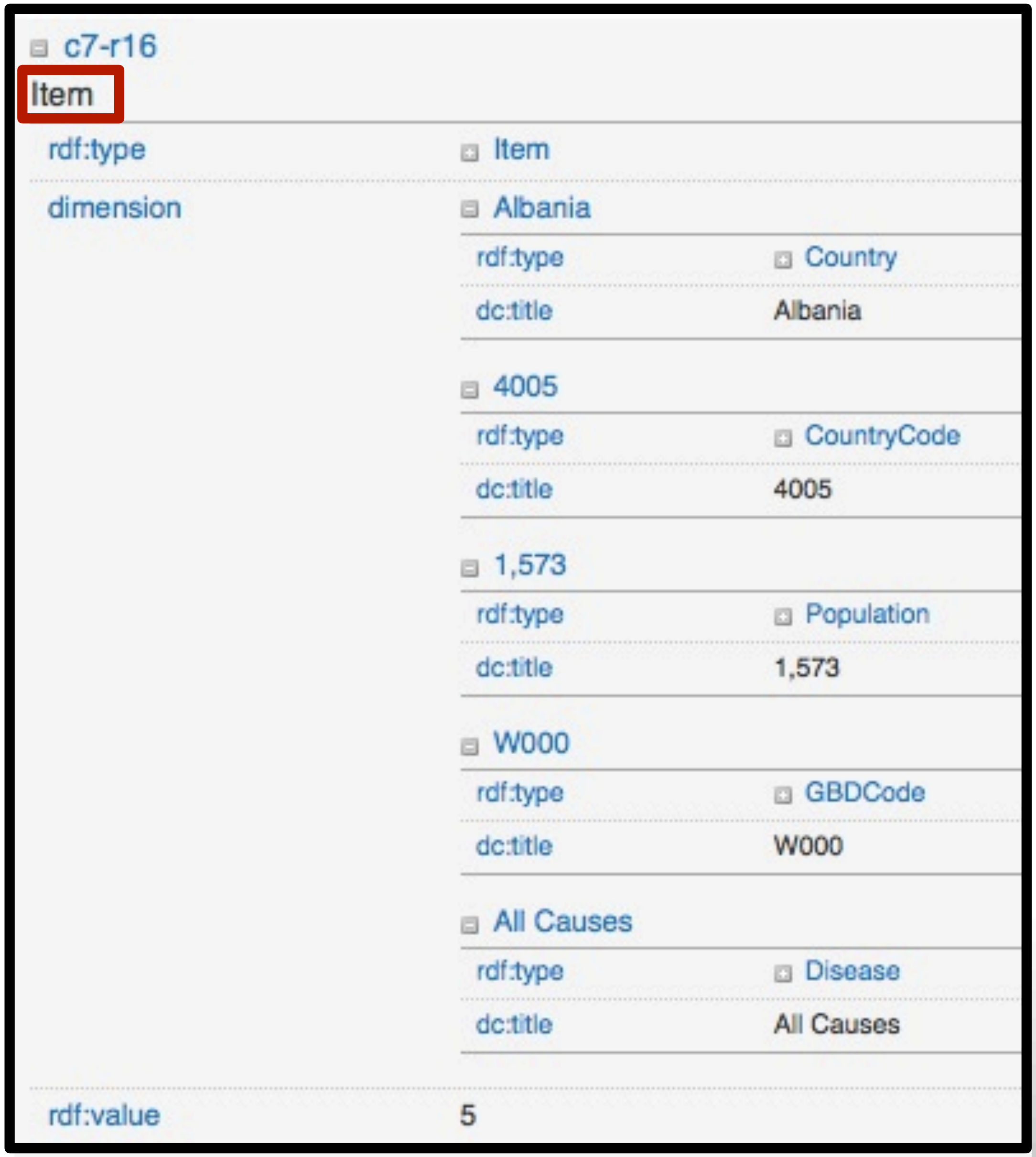
Class

rdf:type	rdfs:Class
rdfs:subClassOf	ns0:Dimension
dc:title	Disease

Country

Class

rdf:type	rdfs:Class
rdfs:subClassOf	ns0:Dimension
dc:title	Country



c7-r16

Item

rdf:type	Item		
dimension	Albania		
		rdf:type	Country
		dc:title	Albania
	4005		
		rdf:type	CountryCode
		dc:title	4005
	1,573		
		rdf:type	Population
		dc:title	1,573
	W000		
		rdf:type	GBDCode
		dc:title	W000
	All Causes		
		rdf:type	Disease
		dc:title	All Causes
rdf:value	5		

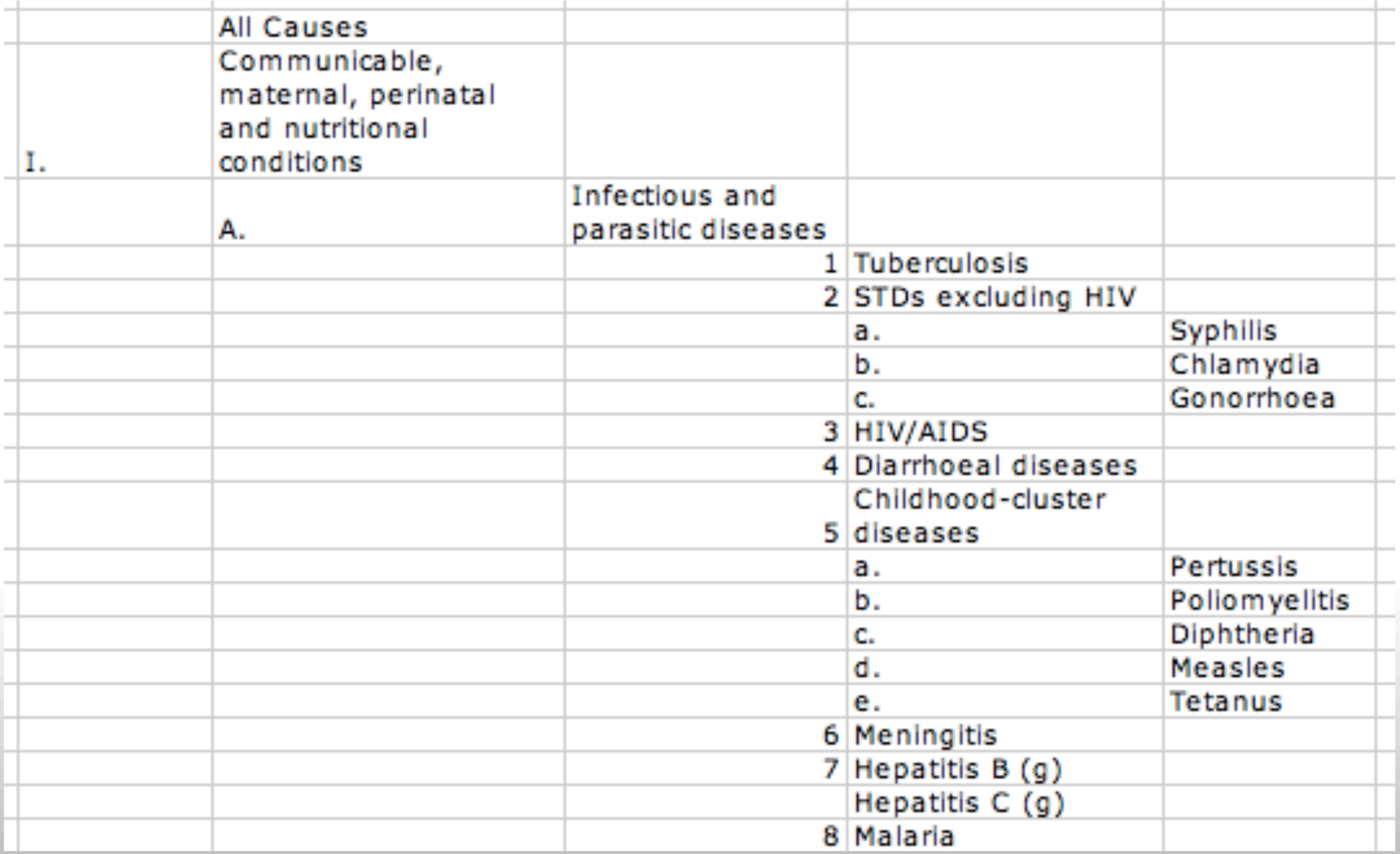
SCOVOfied WHO’s Global Health Observatory Data

```
prefix ex:<http://example.org/who-data>
prefix scv:<http://purl.org/NET/scovo>

ex:Country           rdfs:subClassOf     scv:Dimension;
                     rdf:type            rdfs:Class;
                     dc:title            "Country".
ex:Disease           rdfs:subClassOf     scv:Dimension;
                     rdf:type            rdfs:Class;
                     dc:title            "Disease".
ex:CountryCode       rdfs:subClassOf     scv:Dimension;
                     rdf:type            rdfs:Class;
                     dc:title            "CountryCode".

ex: Afghanistan      rdf:type            ex:Country;
                     dc:title            "Afghanistan" .
ex:Tuberculosis      rdf:type            ex:Disease;
                     dc:title            "Tuberculosis" .
ex:3010              rdf:type            ex:CountryCode;
                     dc:title            “3010”.

ex:c1-r6             rdf:type            scv:Item;
                     rdf:value           127;
                     scv:dimension       ex:Afghanistan;
                     scv:dimension       ex:Tuberculosis .
                     scv:dimension       ex:3010
```

After converting a file containing 5 dimensions and 22384 statistical data items, an RDF model containing

Challenges and Future Work

- There may be some Excel sheets that contain taxonomies only readable by humans.



	All Causes			
I.	Communicable, maternal, perinatal and nutritional conditions			
	A.	Infectious and parasitic diseases		
		1	Tuberculosis	
		2	STDs excluding HIV	
			a.	Syphilis
			b.	Chlamydia
			c.	Gonorrhoea
		3	HIV/AIDS	
		4	Diarrhoeal diseases	
		5	Childhood-cluster diseases	
			a.	Pertussis
			b.	Poliomyelitis
			c.	Diphtheria
			d.	Measles
			e.	Tetanus
		6	Meningitis	
		7	Hepatitis B (g)	
			Hepatitis C (g)	
		8	Malaria	

Future Work

- Converting other WHO datasets
  - ○ WHO Global InfoBase Online
  - ○ Global Health Atlas
  - ○ Regional Statistics
- Evolution patterns$
  - Facilitate seamless evolution of knowledge-base

$ Christoph Rieß, Norman Heino, Sebastian Tramp (geb. Dietzold), and Sören Auer: EvoPat -- Pattern-Based Evolution and Refactoring of RDF Knowledge Bases. In: Proceedings of the 9th International Semantic Web Conference ISWC2010