

Towards Easy Matching Between Statistical Linked Data: Dimension Patterns


Hideto Sato and Wen Wen

First International Workshop on Semantic Statistics
(SemStats 2013)
22 October 2013, Sydney

Introduction

- For matching statistical data from **different sources**, **upper concepts** and **schema-level links** are important.
- Three Problems
 - (1) **A small number of upper concepts** are available.
 - (2) Certain patterns of dimension description **prevent some schema-level links**.
 - (3) Usage of **external codes** is hard to find in a schema-level.
- This paper focuses on (2) and (3), and propose **patterns of dimension description** to improve them.

Trial Matching

- Italian Immigration Statistics
⇒ the numbers of immigrants to Italy
by birth country by year
 - World Bank Statistics
⇒ the total population
by country by year
- 
- Integrated Statistics
Percentage of Immigrants to Italy
by country by year

Italian Immigration Statistics

World Bank Statistics

DataSet

istat:dataset-DCIS_POPSTRCIT

dataset:world-development-indicators

qb:DataSet

istat:dsd-DCIS_POPSTRCIT

d-indicators:structure

qb:DataSet

DataSetDefinition

(Empty Node)

(Empty Node)

DimensionProperty

istat:dimension-paesi

sdmx-dimension:refArea or sdmx-dimension:visArea

country-dimension

rdfs:subPropertyOf

rdfs:subPropertyOf

Code Class

istat:code-range-paesi

sdmx-code:Area

country-code-class

rdfs:subClassOf

rdfs:subClassOf

rdfs:range

qb:codeList

istat:code-paesi

http://sws.geonames.org/ontology#Feature/

classification:country

skos:hasTopConcept

rdf:type

rdf:type

rdf:type

skos:hasTopConcept

Code

istat:code-paesi-al

http://sws.geonames.org/783754/

classification:country/AL

skos:exactMatch

owl:sameAs

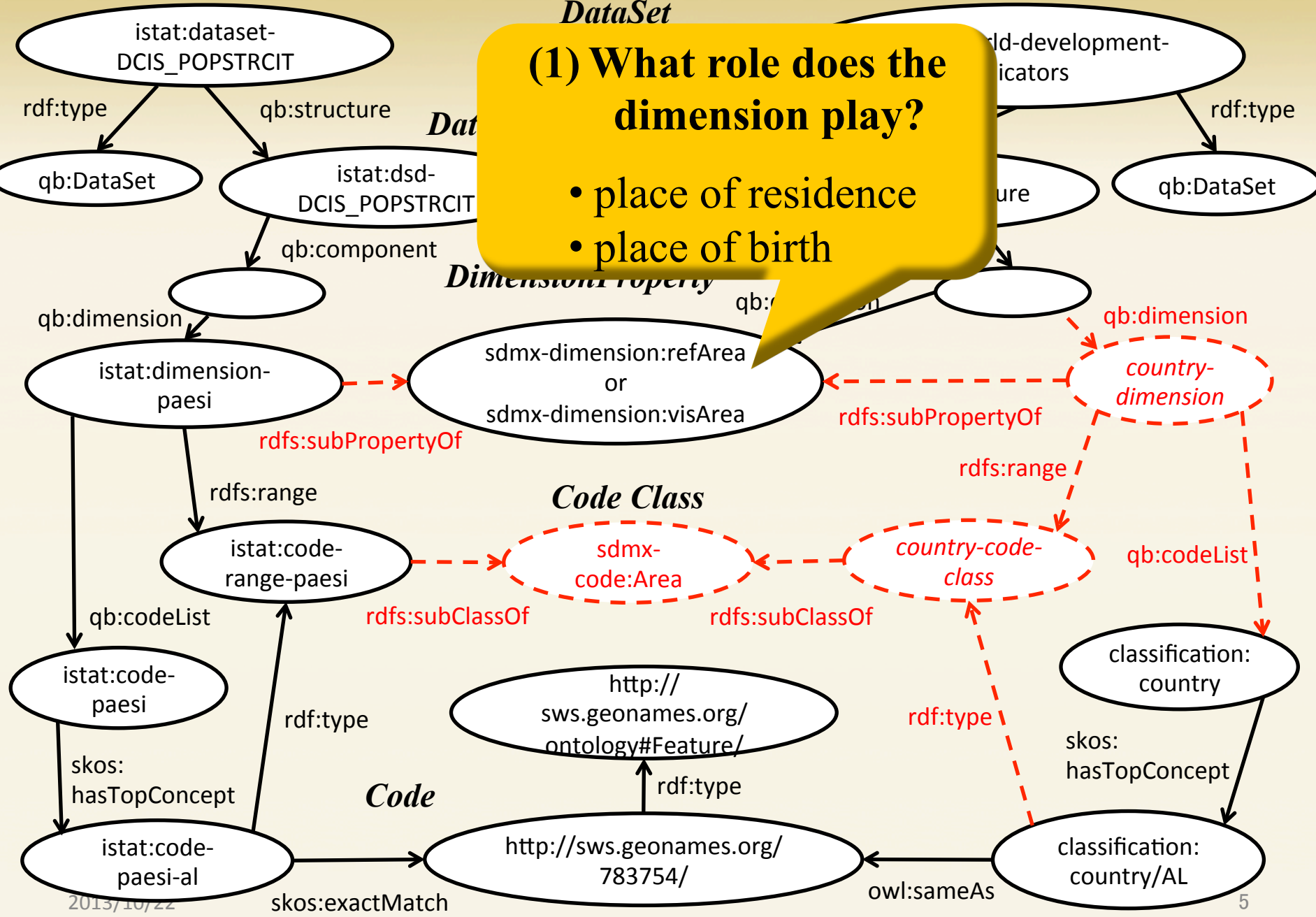
Italian Immigration Statistics

World Bank Statistics

DataSet

(1) What role does the dimension play?

- place of residence
- place of birth



Italian Immigration Statistics

World Bank Statistics

DataSet

DataSet DataStructureDefinition

(2) What type of code does the dimension use ?

- Countries
- Domestic Administrative Areas
- River Basins, and so on.

Code Class

sdmx-code:Area

country-code-class

country-dimension

classification:country

skos:hasTopConcept

classification:country/AL

<http://sws.geonames.org/783754/>

<http://sws.geonames.org/ontology#Feature/>

Code

istat:dataset-DCIS_POPSTRCIT

qb:DataSet

istat:dsd-DCIS_POPSTRCIT

istat:dimension-paesi

istat:code-range-paesi

istat:code-paesi

istat:code-paesi-al

d-indicators:structure

dataset:world-development-indicators

qb:DataSet

classification:country

classification:country/AL

Italian Immigration Statistics

World Bank Statistics

DataSet

DataSet

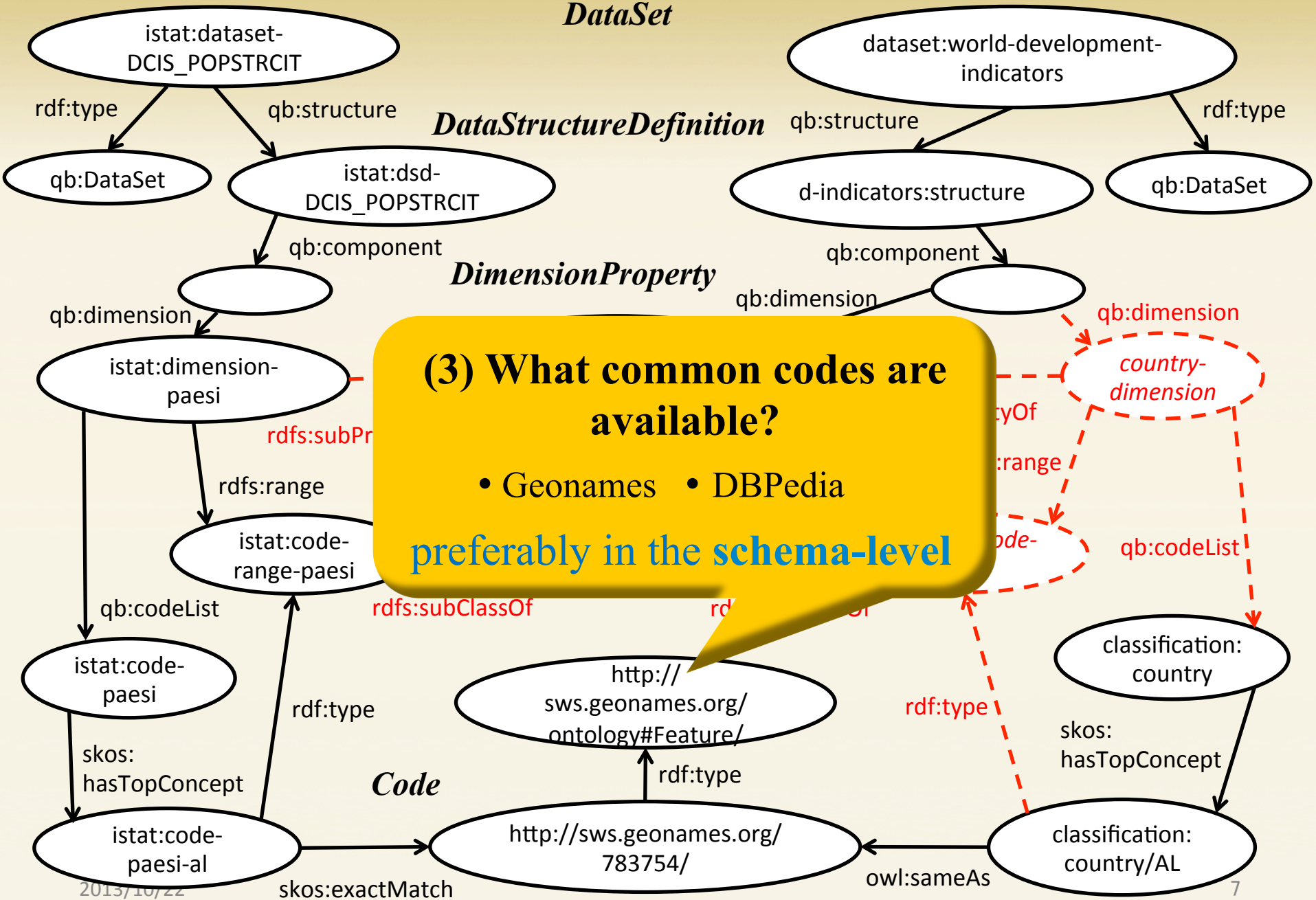
DimensionProperty

(3) What common codes are available?

- Geonames
- DBPedia

preferably in the **schema-level**

Code



Matching Data from Different Sources

The following questions are important for each dimension. As for an **area dimension**,

For Dimension Properties

What role does the dimension play?

- Place of Birth
- Place of Residence

For Code Class (Range of Dimension)

What type of code does the dimension use?

- Countries
- Domestic Administrative Areas
- River Basins

For Code Values

What common codes are available?

- Geonames
- DBPedia

Matching Data from Different Sources

The following questions are in
dimension. As for an **area di**

Upper Concepts

For Dimension Properties

What role does the dimension play?

- Place of Birth
- Place of Residence

For Code Class (Range of Dimension)

What type of code does the dimension use?

- Countries
- Domestic Administrative Areas

For Code Values

What common codes are available?

- Geonames
- DBpedia

Schema-Level Description

QB and Upper Concepts

QB : The RDF Data Cube Vocabulary

QB provides a bridge to **upper concepts** by referring to the **SDMX-RDF vocabulary**.

Upper Concepts and SDMX-RDF

Upper concept

Upper resource in SDMX-RDF

Dimension Property

Place of Birth → sdmx-dimension:visArea

Place of Residence → sdmx-dimension:refArea

Code Class (Range of Dimension)

Area → sdmx-code:Area

Country → (not defined)

Domestic Area → (not defined)

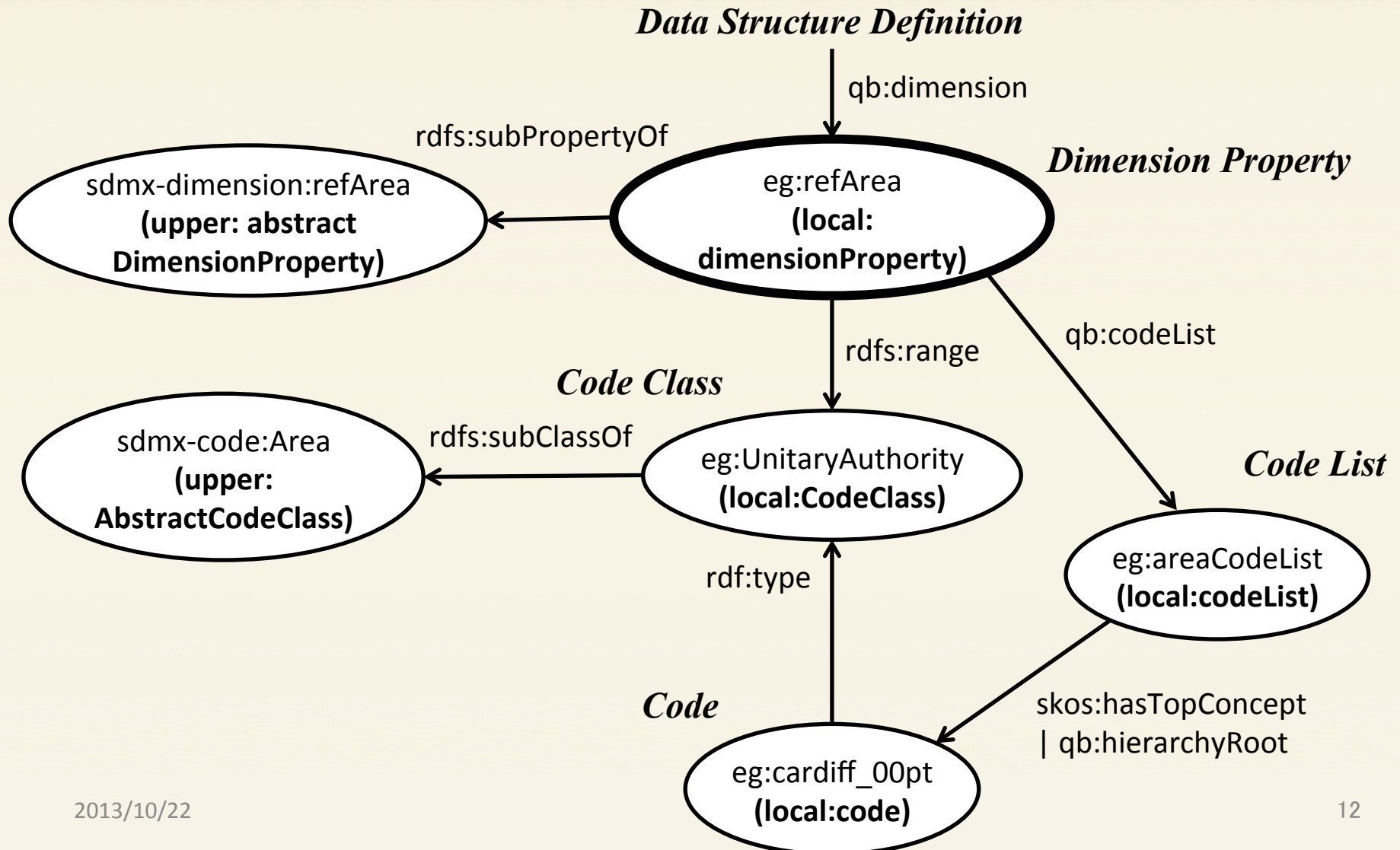
River Basin → (not defined)

(sdmx-dimension:visArea has been removed in the current version of SDMX-RDF.)

Dimension Description in QB

Upper

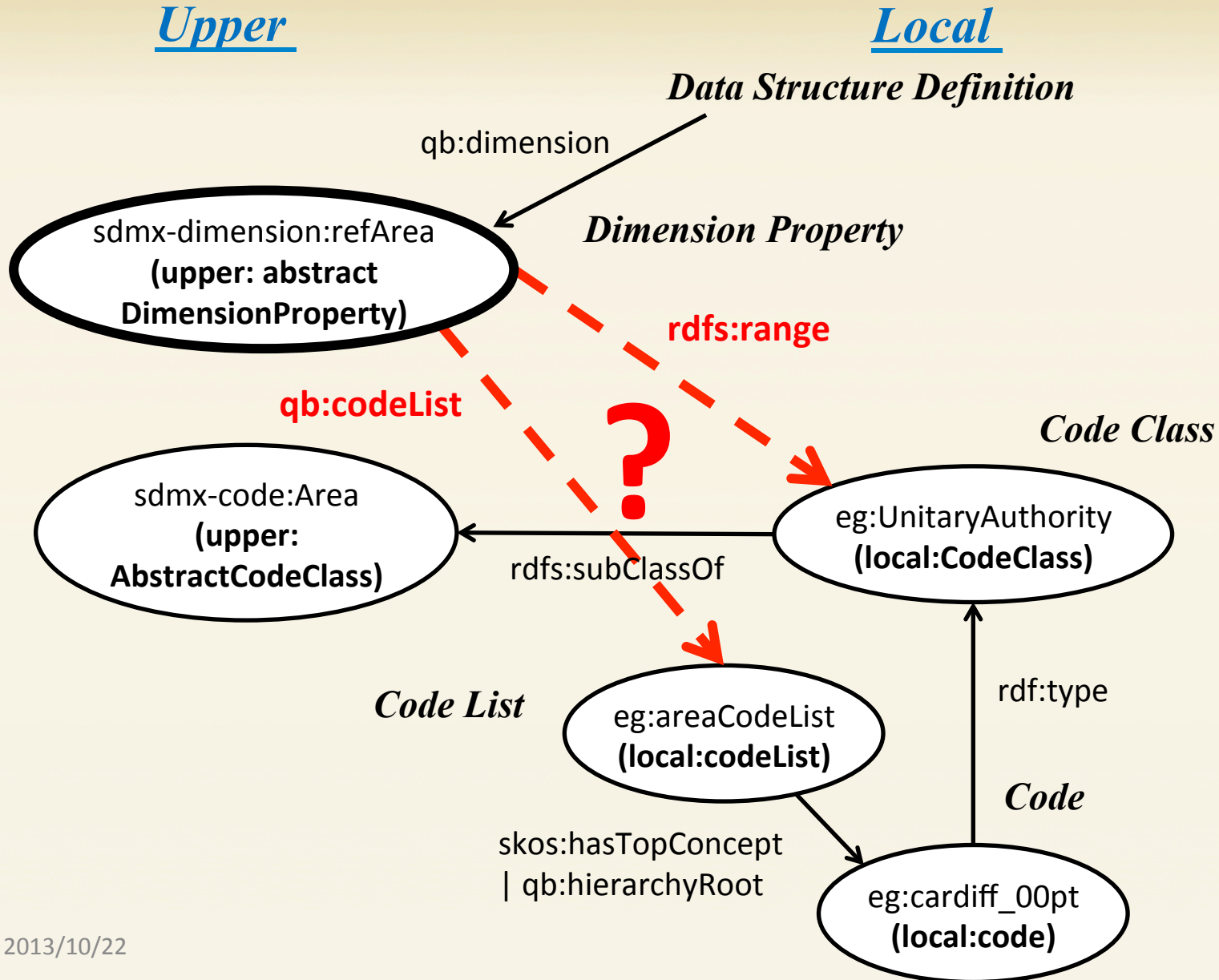
Local



Anti-Patterns

- **Two Anti-Patterns** prevent describing **schema-level links** properly.
 - Direct use of **an abstract upper resource**
 - Direct use of **an external code class**

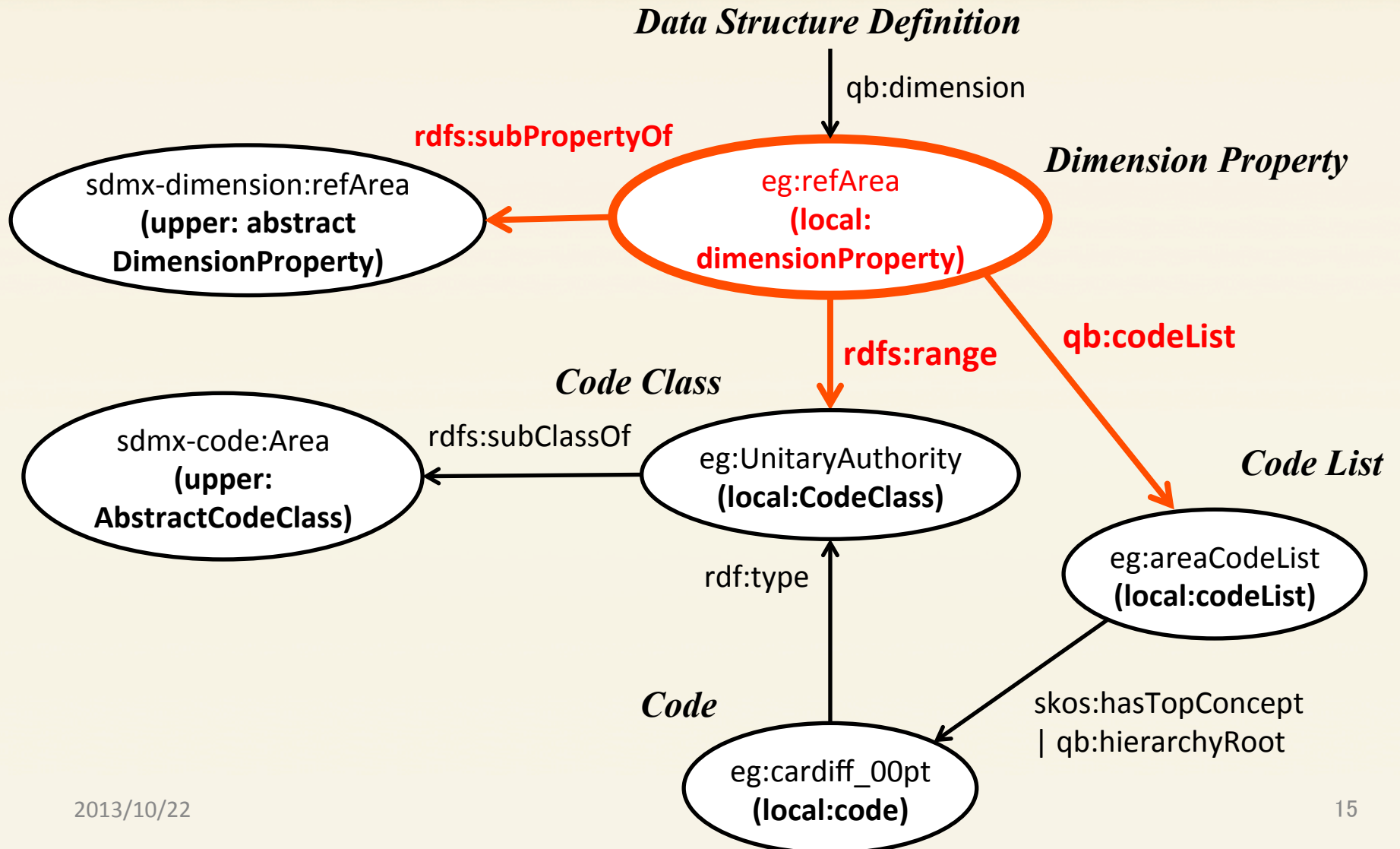
Anti-Pattern: Direct Use of an Upper Resource



The Pattern for Using a Local Code Class

Upper

Local



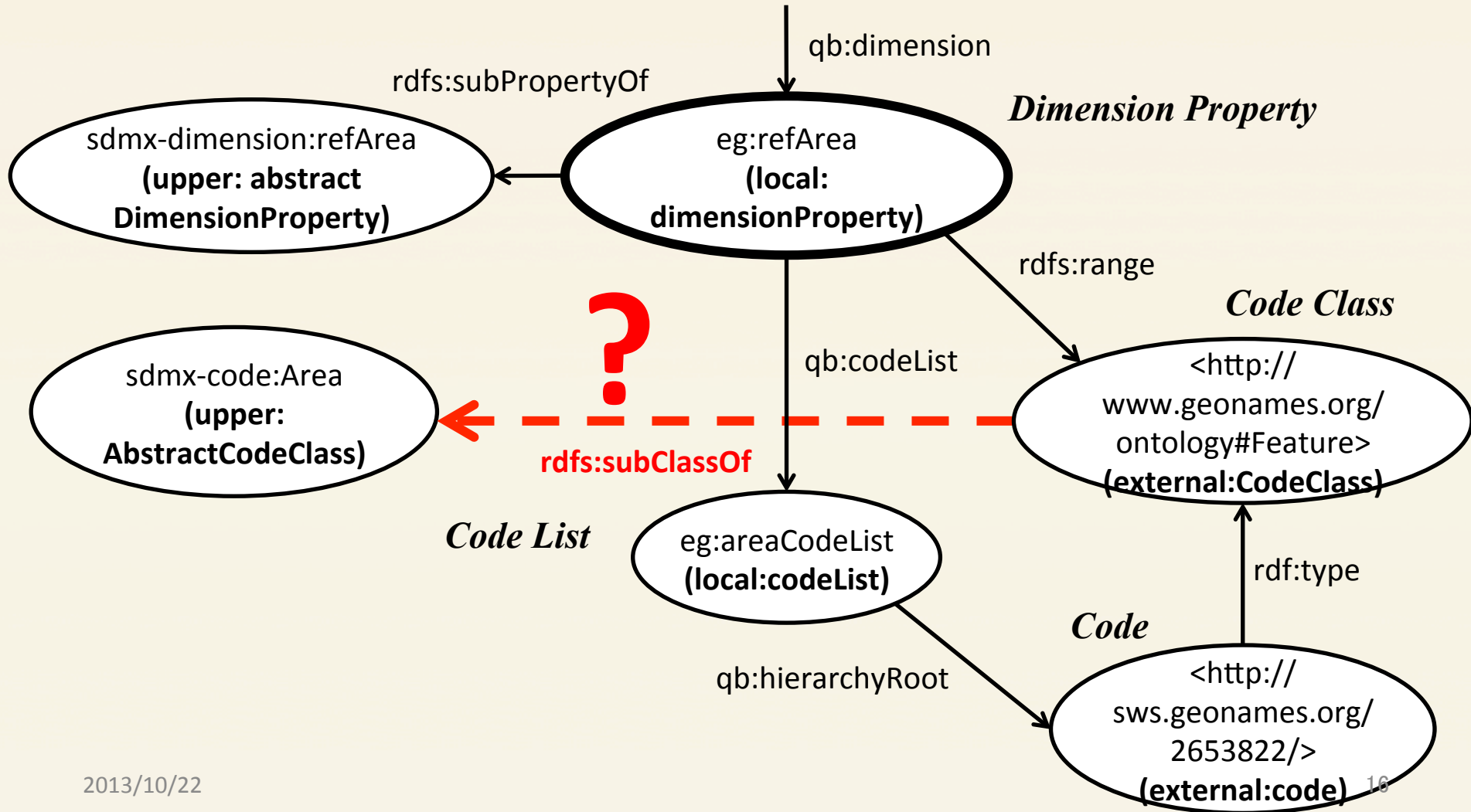
Anti-Pattern: Direct Use of an External Code Class

Upper

Local

External

Data Structure Definition

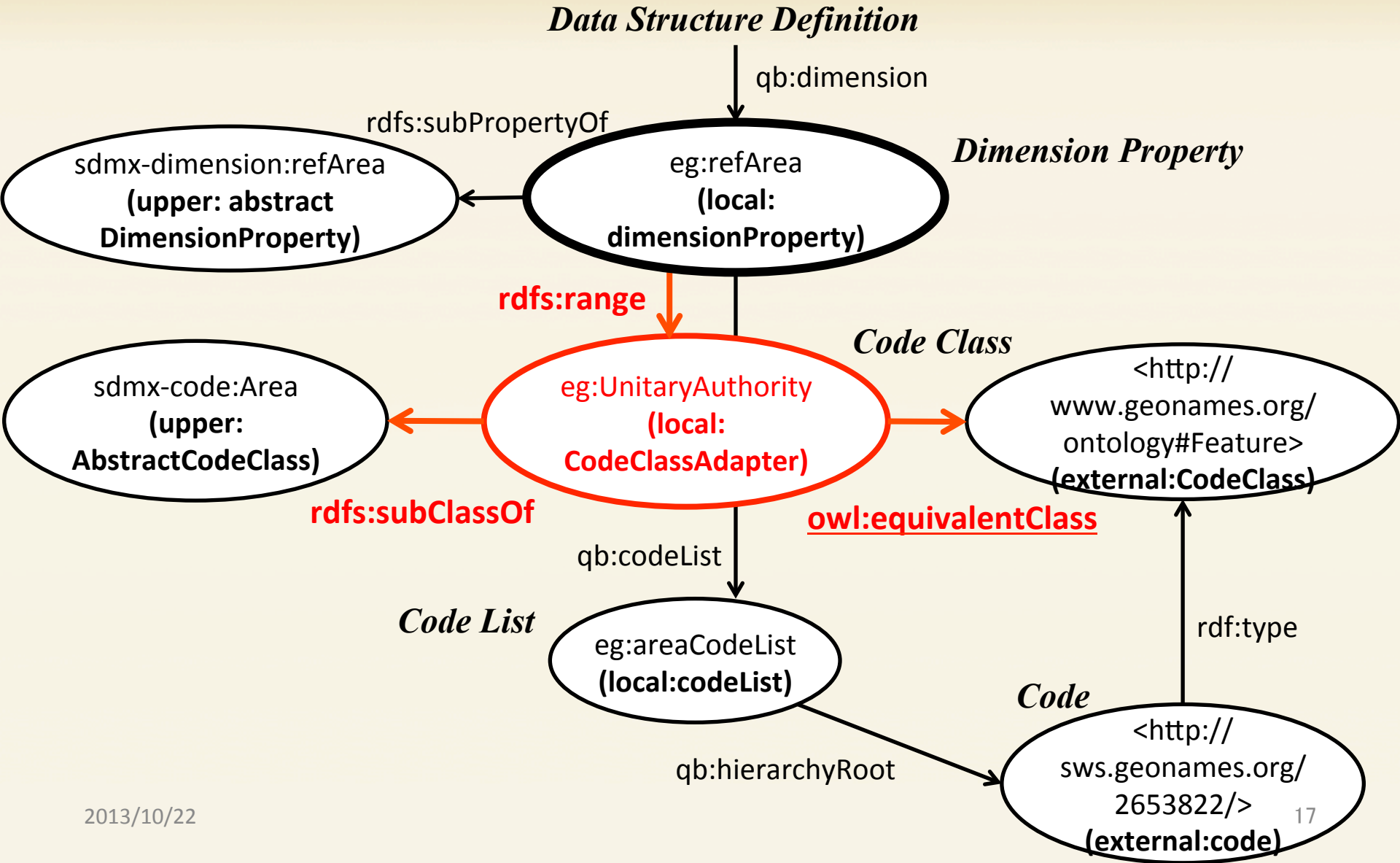


The Pattern for Using an External Code Class

Upper

Local

External

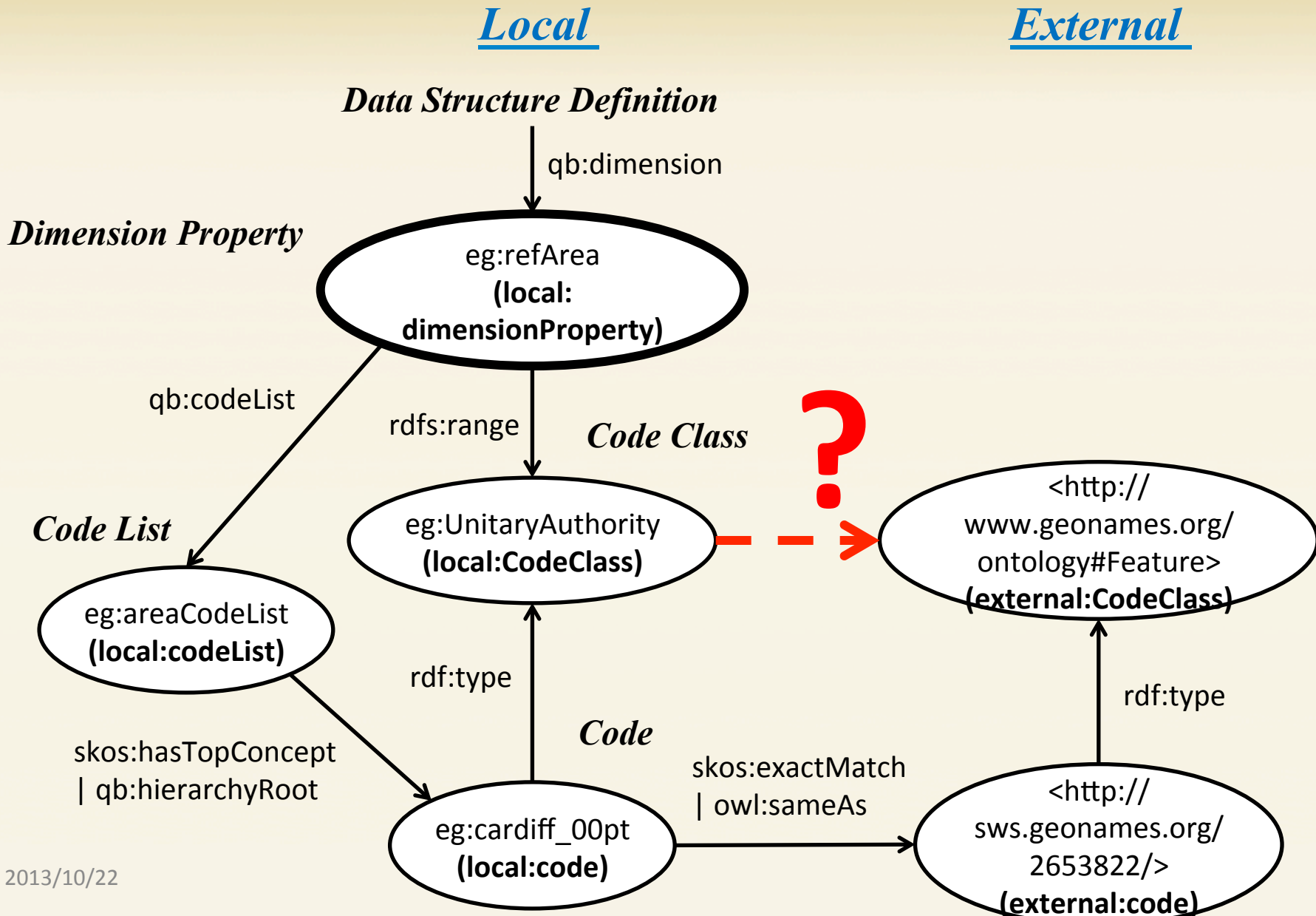


Alternate Code Class

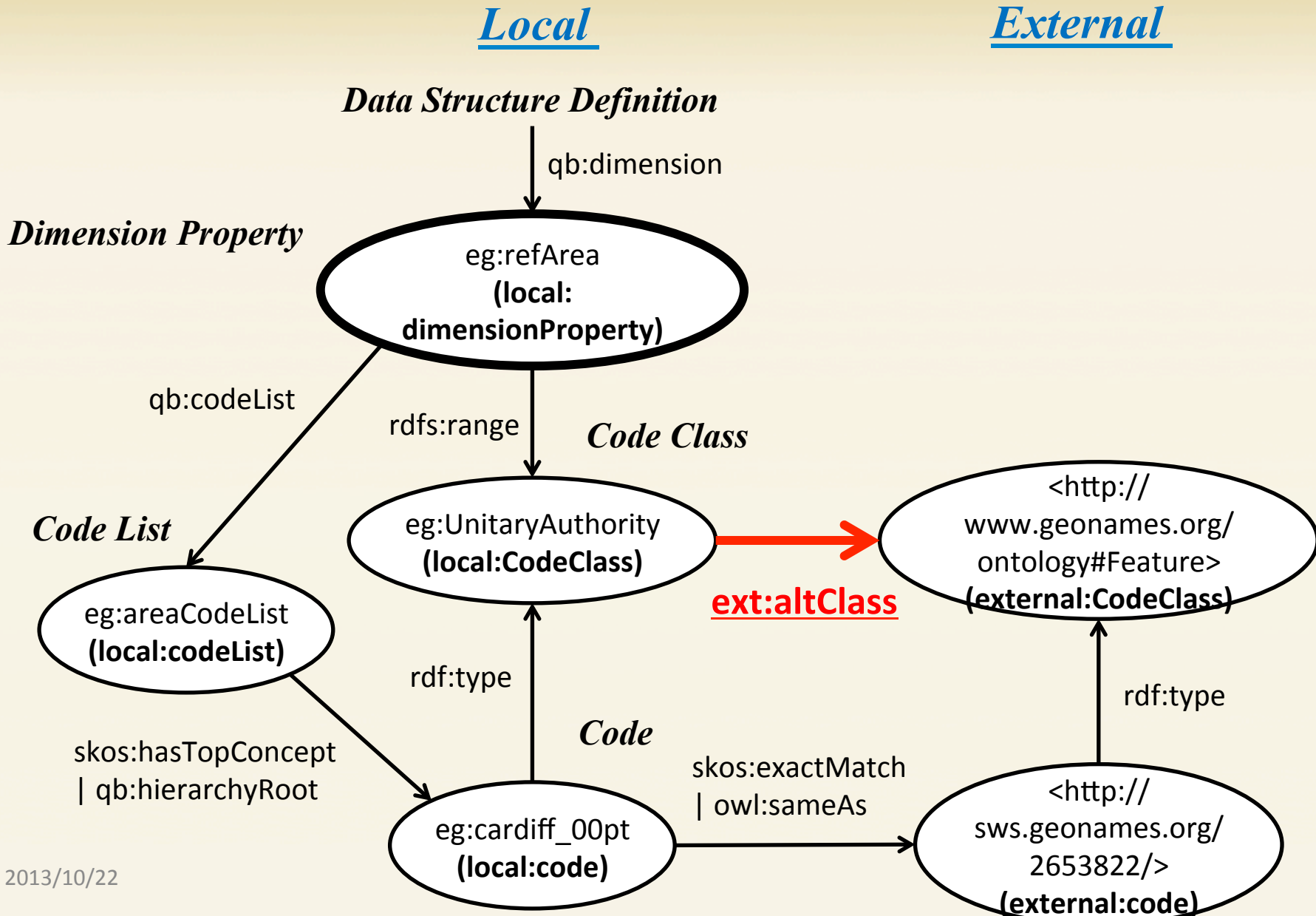
When using both **local and external code classes**, it is difficult to find whether an external code class is employed or not.

We need a **schema-level description** for an **alternate code class**.

Using Local and External Code Classes



Proposal of an additional link (ext:altClass)



From Our Survey

	Area Dimension	Time Dimension
Direct Use of an Upper Resource	3/12	3/12
Direct Use of an External Code Class	2/12	8/12
Use of Alternate Code Classes	10/12	1/12

The counts are DSDs (Data Structure Definitions) found in the endpoints listed at http://www.w3.org/2011/gld/wiki/Data_Cube_Implementations.

Conclusion

- We introduced **dimension patterns** for describing **schema-level links** including **references to upper resources** and **alternate class links**.
- These will extract the QB's power of description to its full extent.
- However, only **a few upper resources** are available now. Therefore, the part of the patterns concerning to upper concepts are **preparatory** for the future.
- We think that it is an urgent task **to enrich upper resources** suitable for statistical data.