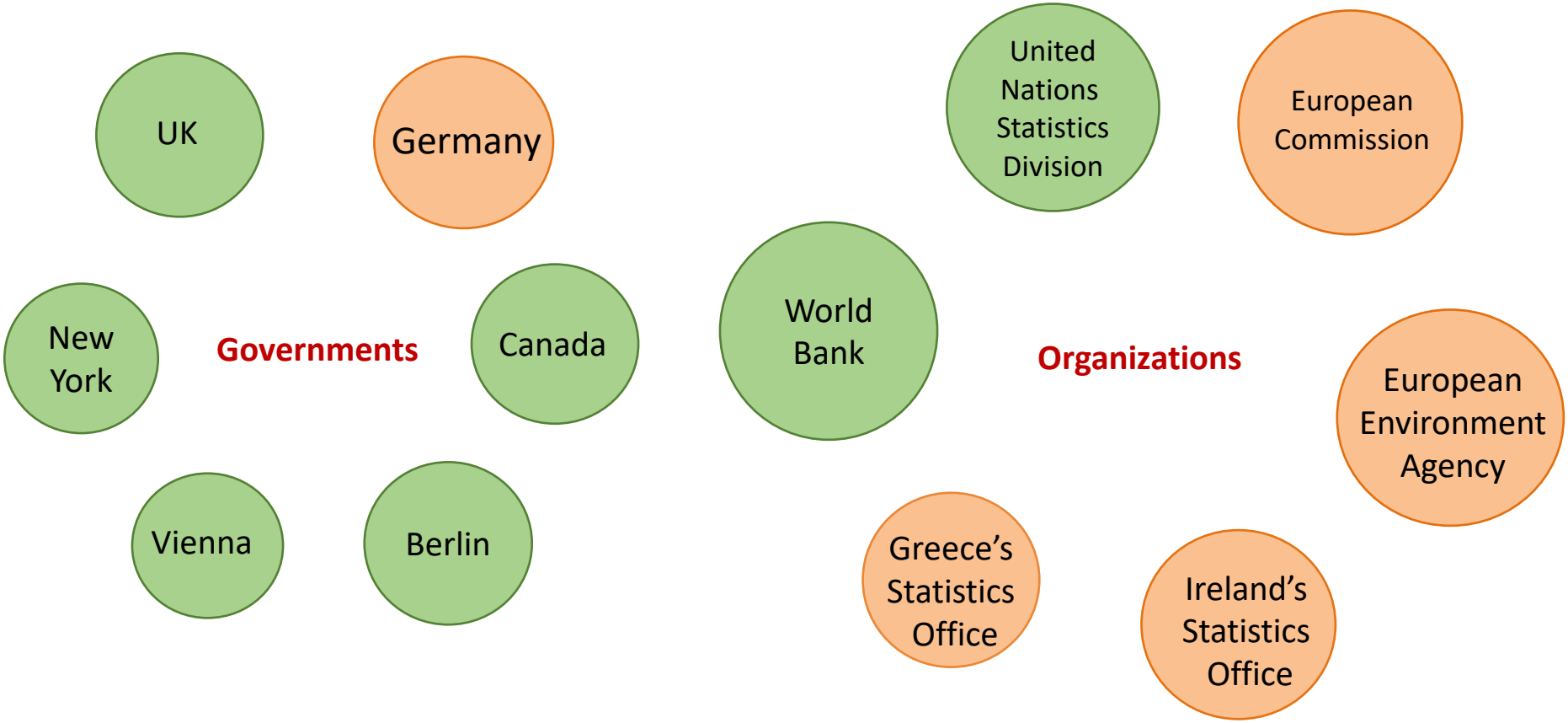
ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Semantic Exploration and Integration of Statistical Data



Statistical data is published by
various governments and organizations!

UK

Germany

United Nations Statistics Division

European Commission

New York

Governments

Canada

World Bank

Organizations

European Environment Agency

Vienna

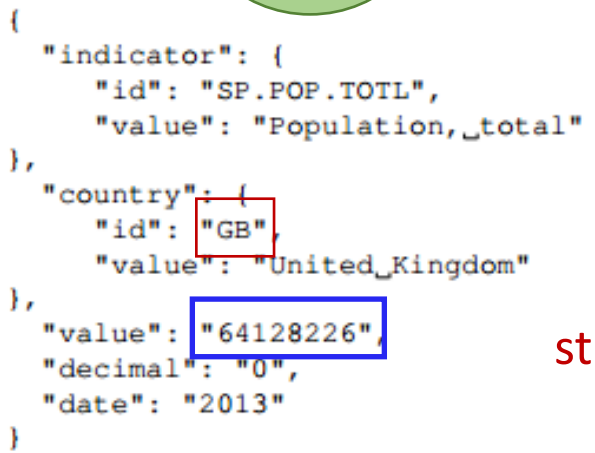
Berlin

Greece's Statistics Office

Ireland's Statistics Office

What is the population of the UK?

World Bank



```
{
  "indicator": {
    "id": "SP.POP.TOTL",
    "value": "Population, total"
},
  "country": {
    "id": "GB",
    "value": "United Kingdom"
},
  "value": "64128226",
  "decimal": "0",
  "date": "2013"
}
```

JSON

UK

Mid-Year	Mid-Year Population (millions)	Annual Percentage Change
2011	63.3	0.84
2012	63.7	0.66
2013	64.1	0.63

CSV

Heterogeneity of formats, encodings, scales, structures, access mechanisms

European Environment Agency

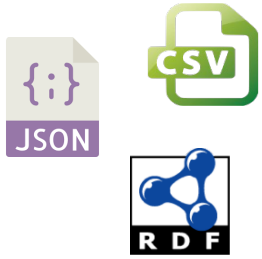
How does it fit together?

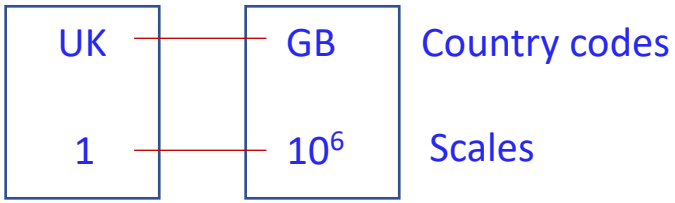
sdmxd: freq	sdmxd: timePeriod	sdmxd: refArea	sd-mxd:age	sdmxd: sex	sdmxm: obsValue
sdmx-code:freqA	2013	geo:UK	ag:TOTAL	sdmxcode: sex-T	63,905,297
sdmx-code:freqA	2013	geo:UK	ag:Y_LT15	sdmxcode: sex-T	11,260,549
sdmx-code:freqA	2013	geo:UK	ag:Y15–64	sdmxcode: sex-F	20,917,257

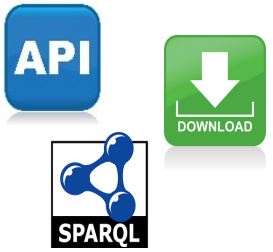
RDF

Motivation & Challenges

- Motivation: facilitate exploration and integration of heterogeneous statistical data sets
- Main challenges
  - Syntactic heterogeneity
  - Semantic heterogeneity
  - Access heterogeneity







Research Gaps

- Statistical data **integration**
  "The integration of data cubes is still an unexploited research topic"
  Karamanou et al., ***Linked Data Cubes: Research results so far***

- Statistical data **exploration**
  - Focusing on visualization of individual data sets [1-7]
  - Identification of relatable data set is an open problem

Research Questions

How can users be enabled to explore and integrate heterogenous statistical data sources?

RQ1. How can we address statistical data heterogeneity in terms of formats?

RQ2. How can we establish interconnections between statistical data sets?

RQ3. How can we provide uniform access to heterogenous statistical data sets?

Syntactic Heterogeneity

RQ1. How can we address statistical data heterogeneity in terms of formats?

- RDF: standard for data representation [9, 10]

	In advance transformation	Query time transformation
Mechanism	Transforms data into RDF then stores data in endpoints [11-15]	Transforms data into RDF when requested by users
Options	Mapping languages [11, 12], ETL tools [13- 15]	Mapping languages
Advantages	Manages data sets easily	Provides up-to-date data Requires small volume for mapping storage
Disadvantages	Increases data volume Faces with out-of-date data	Increases query answering time

- Approach
  - Uses mapping language for data transformation at query time
  - Creates a cache to reduce query answering time

Mapping Languages

	CSV	Spreadsheet	JSON	XML	Database
XLWrap [16]	✓				
M^2 [17]		✓			
XSPARQL [18]				✓	
D2RMap [19]					✓
R2RML [20]					✓
M2RML [21]					✓
RML [22]	✓		✓	✓	

W3C

An overview of existing mapping languages

- Choose RML for data transformation
- Two extensions for the RML's processor
  - Support for spreadsheet formats
  - Support for variables in mappings

RML mapping for the World Bank

http://statspace.linkedwidgets.org/mapping/wb.ttl&indicator=NY.GDP.PCAP.CD&refArea=AT

```
<#Variables>
  rmlx:defaultValue
    [rmlx:varName "indicator"; rr:constant "SP.POP.TOTL"],
    [rmlx:varName "refArea"; rr:constant "all"].

<#Observation>
  rml:logicalSource [
    rml:source
        "http://api.worldbank.org/countries/{refArea}/indicators/{indicator}?
         format=json&page=1&per_page=15000";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$[1].*"
  ];
  rr:subjectMap[
    rr:class qb:Observation;
    rr:template "http://statspace.linkedwidgets.org/dataset/WorldBank-
                 {indicator.id}/Obs-{country.id}-{date}";
    rr:termType rr:IRI
  ];
  rr:predicateObjectMap [
    rr:predicate sdmxd:refArea;
    rr:objectMap [
       rr:template
           "http://statspace.linkedwidgets.org/codelist/cl_area/{country.value}";
       rr:termType rr:IRI
    ]
];
```

AT

NY.GDP.PCAP.CD

Description of the input data set

Template to coin subjects

(Subject, Predicate, Object)

Template to coin objects

Now, data is syntactically homogenous (RDF)

UK

World Bank

European Environment Agency

Heterogeneity of URIs and scales

uk:country	uk:year	uk:value	uk:unit
uk:UnitedKingdom	uk:2013	64.1	uk:Million

wb:country	wb:year	wb:value	wb:unit
wb:GB	wb:2013	64128226	wb:AbsoluteScale

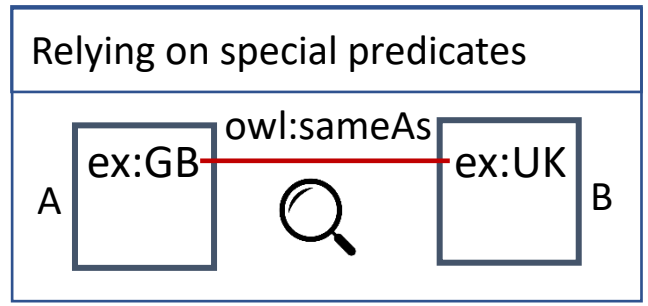
sdmxd: freq	sdmxd: timePeriod	sdmxd: refArea	sd-mxd:age	sdmxd: sex	sdmxm: obsValue
sdmx-code:freqA	2013	geo:UK	ag:TOTAL	sdmxcode: sex-T	63,905,297
sdmx-code:freqA	2013	geo:UK	ag:Y_LT15	sdmxcode: sex-T	11,260,549
sdmx-code:freqA	2013	geo:UK	ag:Y15–64	sdmxcode: sex-F	20,917,257

Semantic Heterogeneity

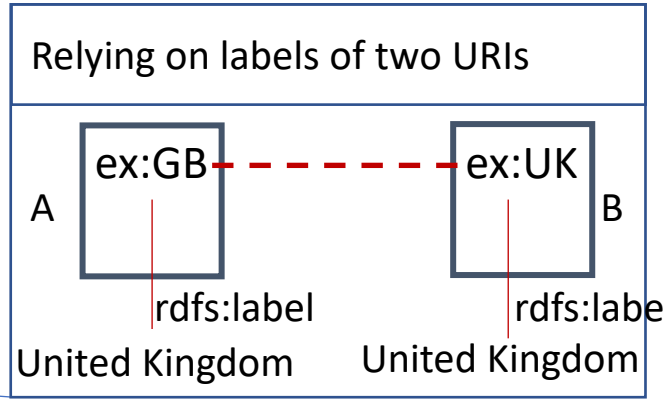
RQ2. How can we establish interconnections between statistical data sets?

- Basics for data integration: identify equivalent entities

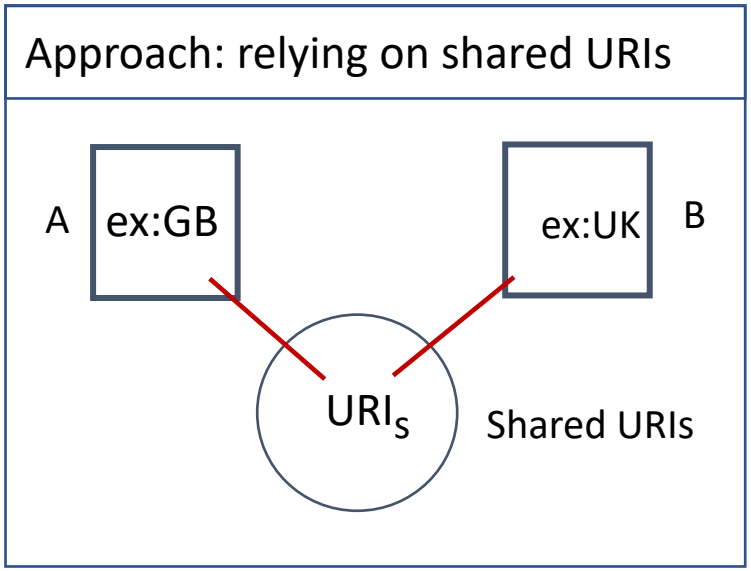
establish direct relationships



Used in coreference resolution services [24, 25]



Used in research on statistical data integration [11, 12, 14]



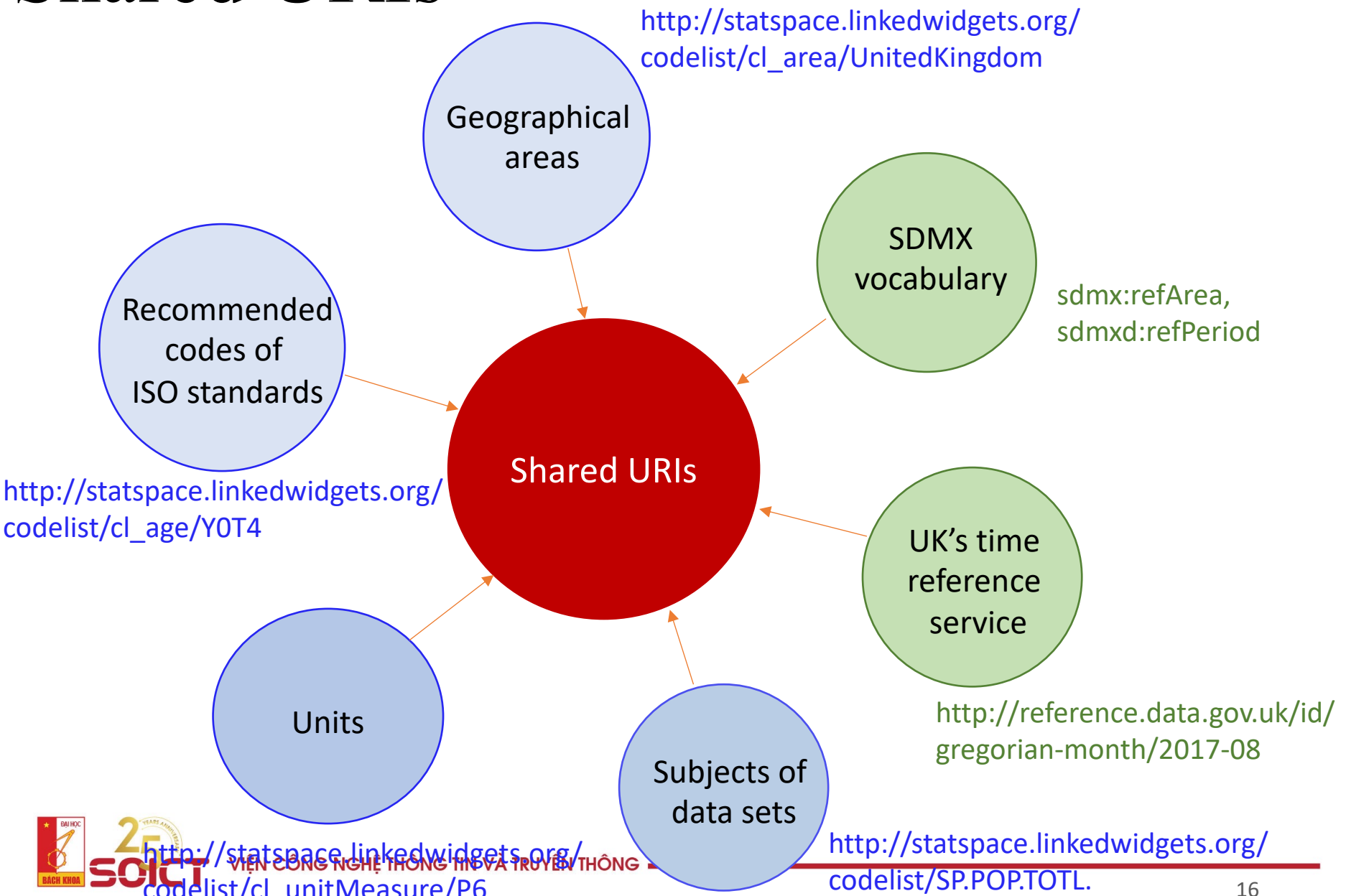
A set of shared URIs for linking facilitates query rewriting and result rewriting

Shared URIs – Example

- Geographical areas
  - Pattern: *Country/Administrative Area Level 1/.../Administrative Area Level n*
  - Example: Austria/Vienna/Vienna/Floridsdorf

- Temporal values
  - Patterns: rely on the UK time reference service
  - Example:
    - Year: http://reference.data.gov.uk/id/gregorian-year/{year}
    - Month: http://reference.data.gov.uk/id/gregorian-month/{month}

Shared URIs


http://statspace.linkedwidgets.org/
codelist/cl_area/UnitedKingdom
Geographical areas
SDMX vocabulary
sdmx:refArea,
sdmxd:refPeriod
Recommended codes of ISO standards
http://statspace.linkedwidgets.org/
codelist/cl_age/Y0T4
Shared URIs
UK's time reference service
http://reference.data.gov.uk/id/
gregorian-month/2017-08
Units
Subjects of data sets
http://statspace.linkedwidgets.org/
codelist/cl_unitMeasure/P6
http://statspace.linkedwidgets.org/
codelist/SP.POP.TOTL.

Spatial Dimension Mapping Algorithm

- Spatial dimension: maps different URIs, such as *ex:country, ex:area,* etc. to *sdmxd:refArea*

- Spatial values: uses Google's geocoding service for mapping discovery

Vienna Austria

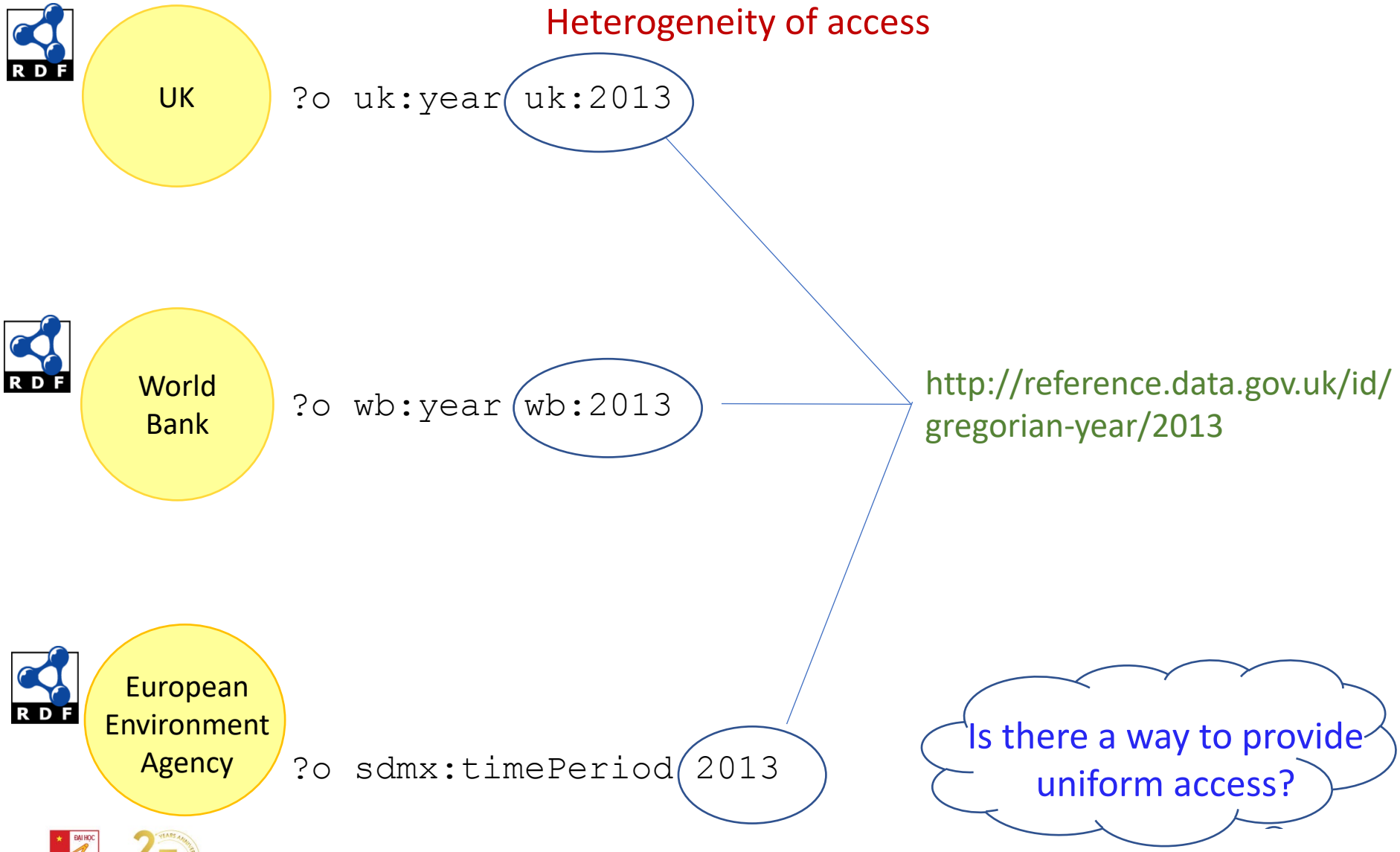
Wien Österreich

Vienne Autriche

Google's geocoding service


```
"results" : [
  {
    "address_components" : [
      {
        "long_name" : "Vienna",
        "short_name" : "Vienna",
        "types" : [ "locality", "political" ]
      },
      {
        "long_name" : "Vienna",
        "short_name" : "Vienna",
        "types": ["administrative_area_level_1", "political"]
      },
      {
        "long_name" : "Austria",
        "short_name" : "AT",
        "types" : [ "country", "political" ]
      }
    ],
    "formatted_address" : "Vienna, Austria",
    "geometry" : {
       "location" : {
          "lat" : 48.2081743,
          "lng" : 16.3738189
       }
    }
  }
]
```

Now, RDF data sets are linked!

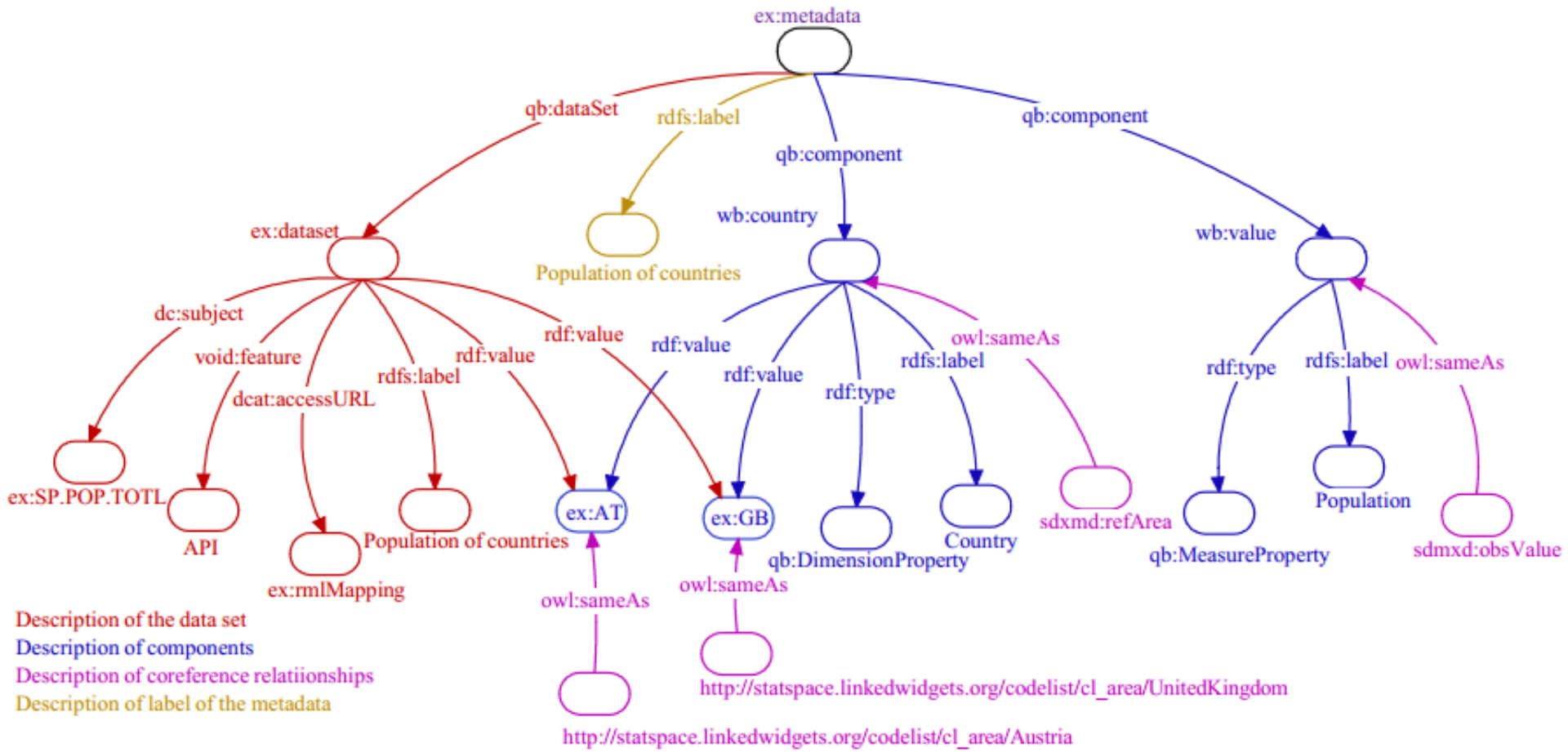


Uniform Access

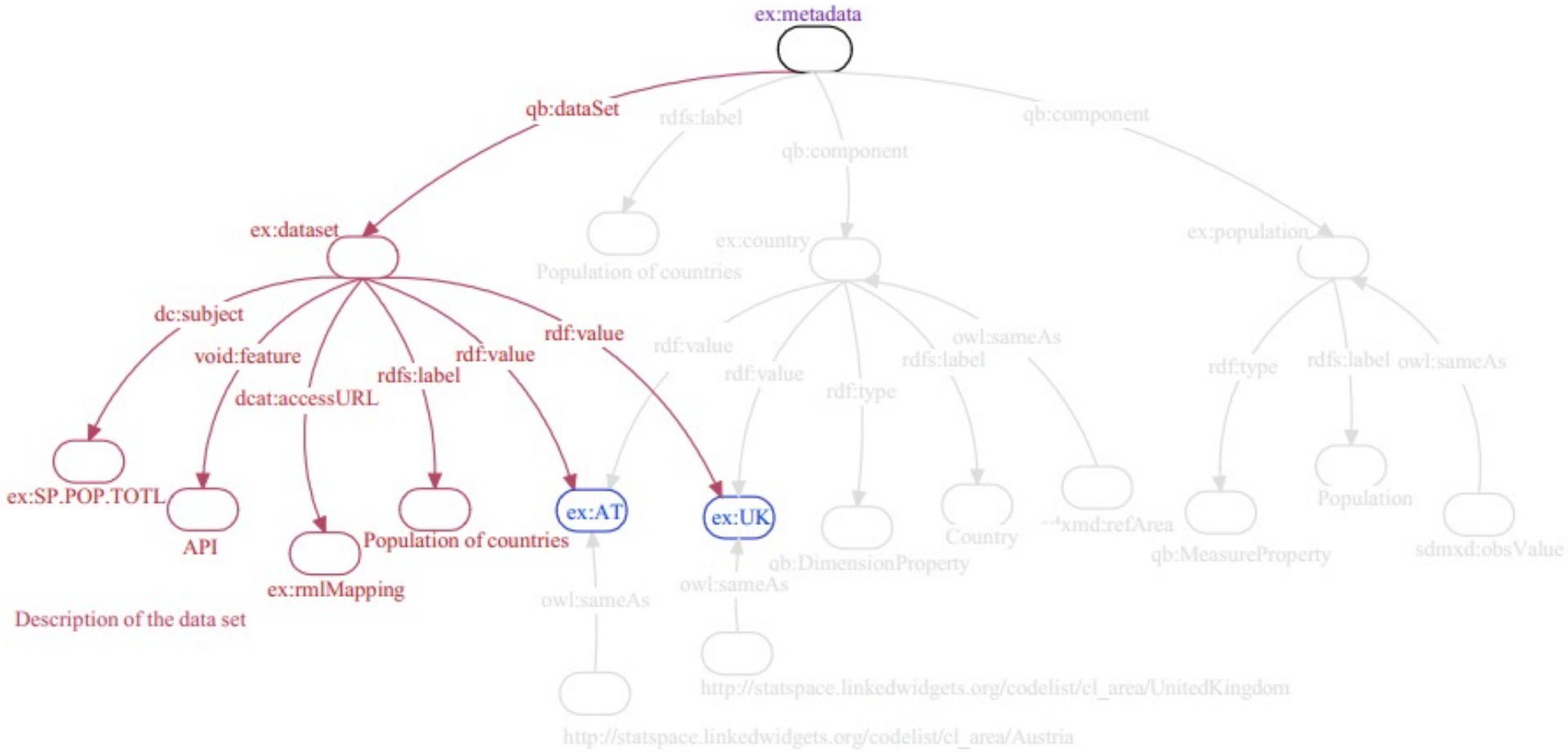
RQ3. How can we provide uniform access to heterogeneous statistical data sets?

- No related work on statistical data domain!
- Approach
  - Metadata description for each statistical data set
    - Information about data structure, access mechanism, and equivalent relationships
    - Standardized conceptual layer over each data set
  - Mediator
    - Query rewriting and result rewriting
    - Single point of access

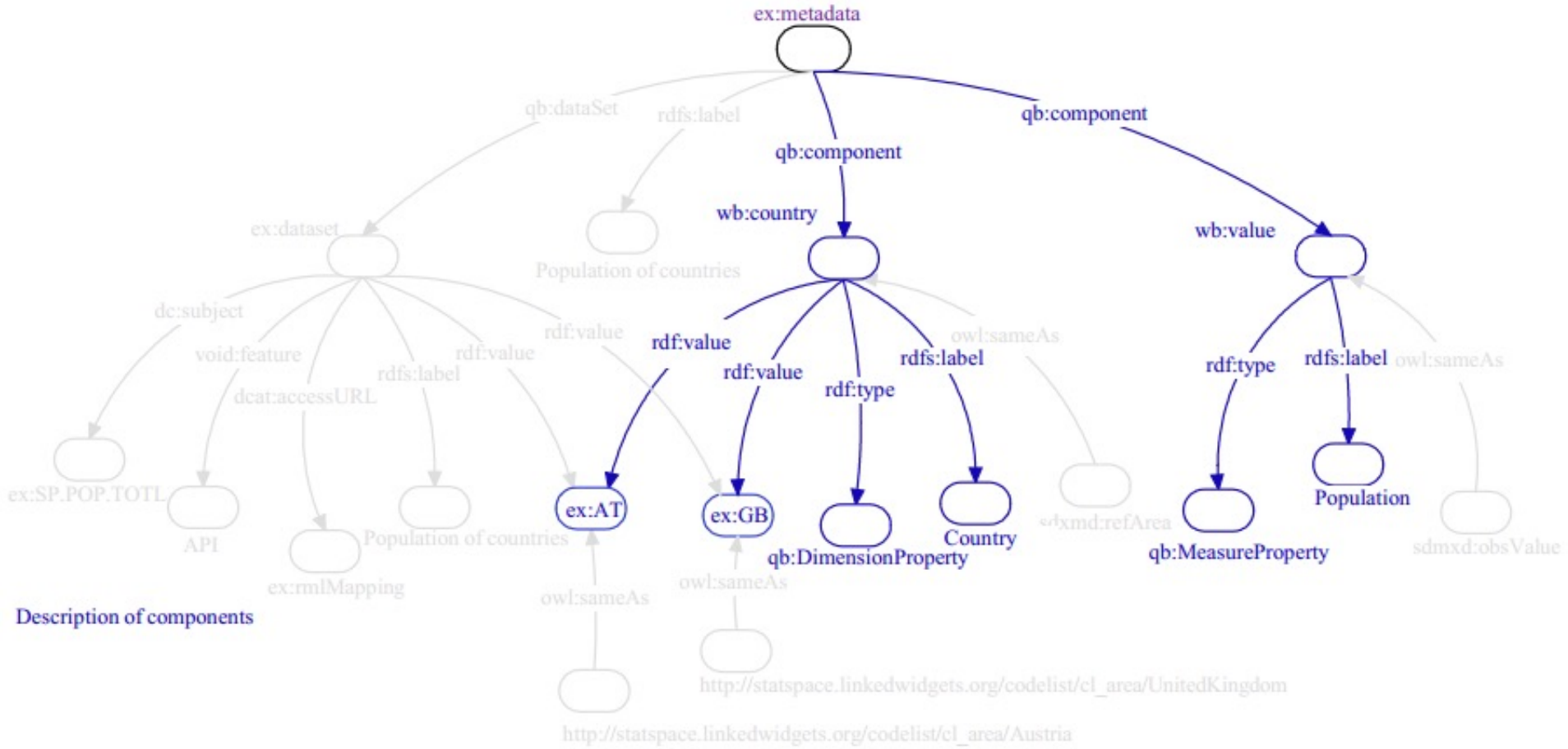
Metadata Description - Example



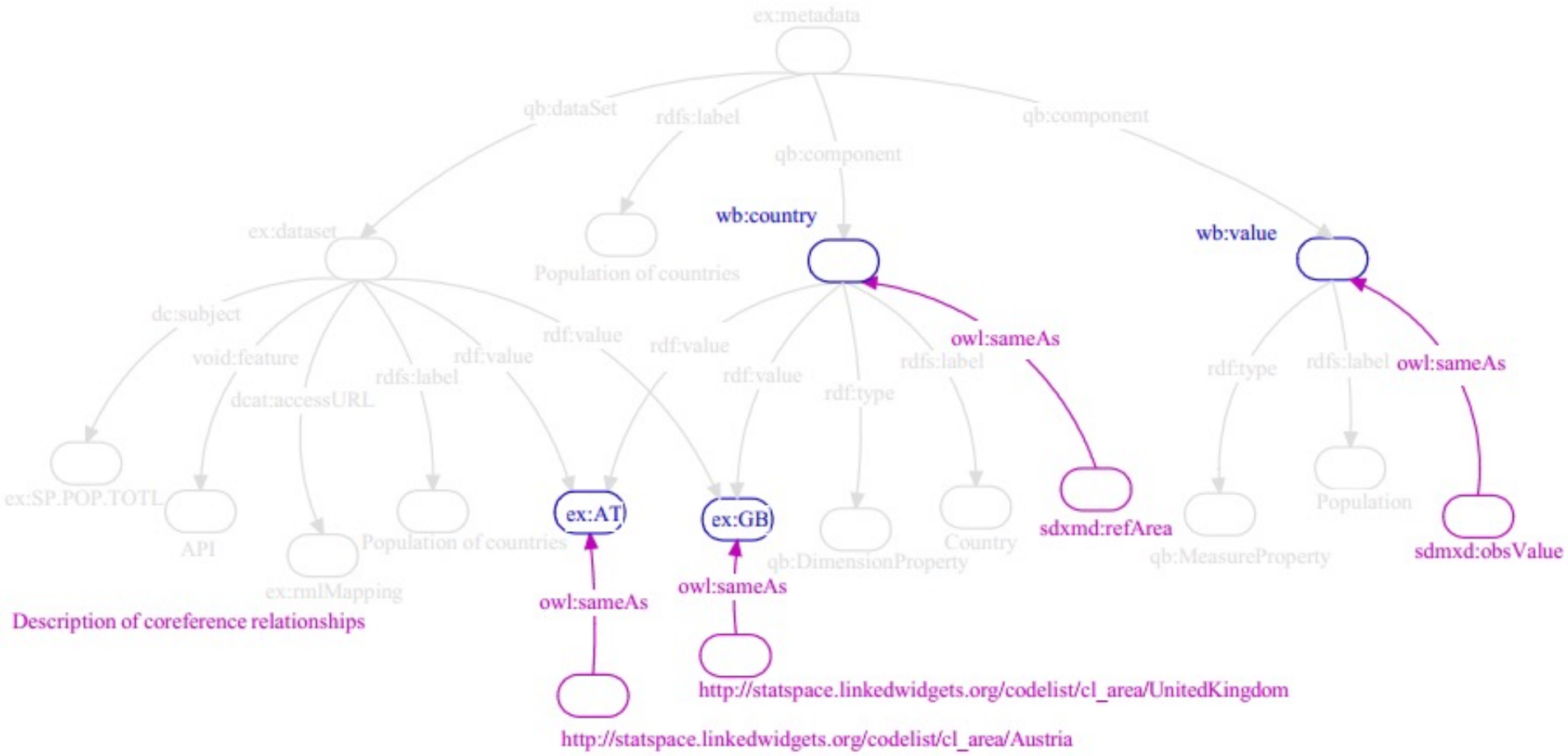
Metadata Description – Description of the data set



Metadata Description – Description of components

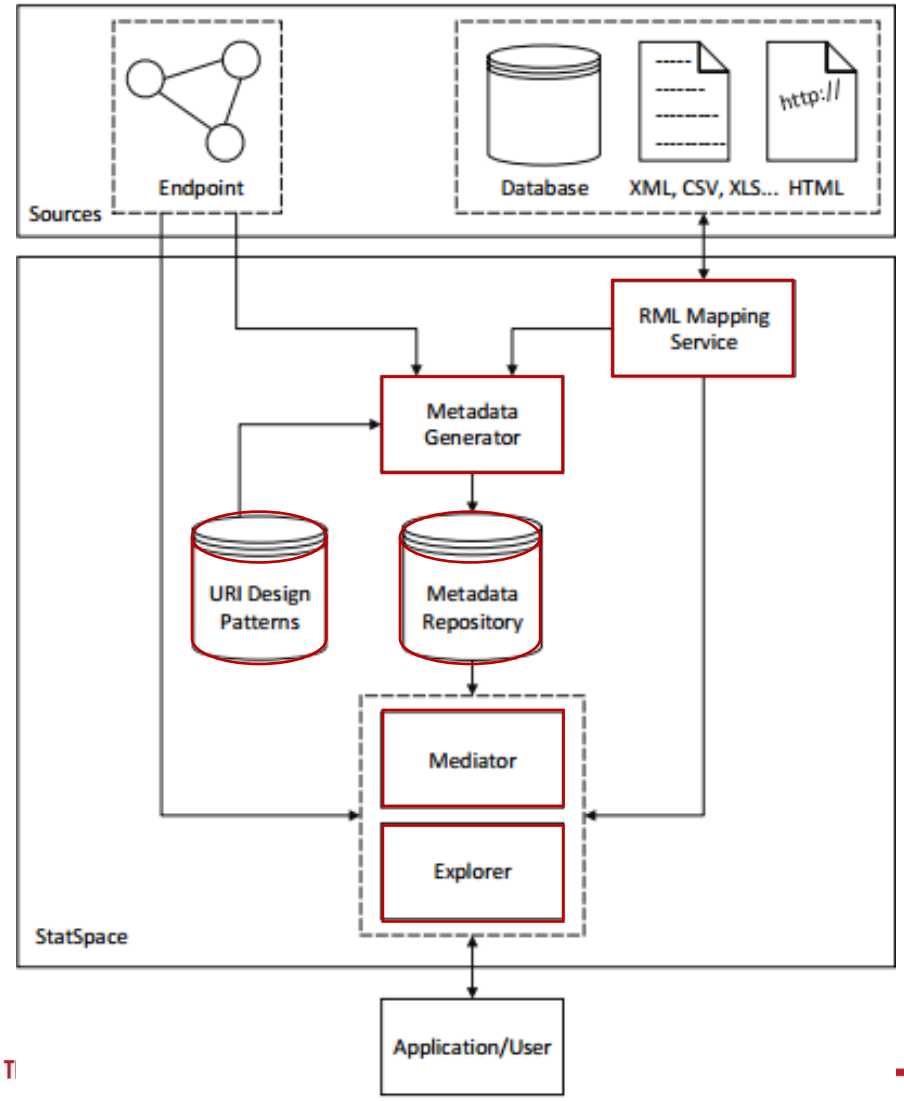


Metadata Description – Description of coreference relationships



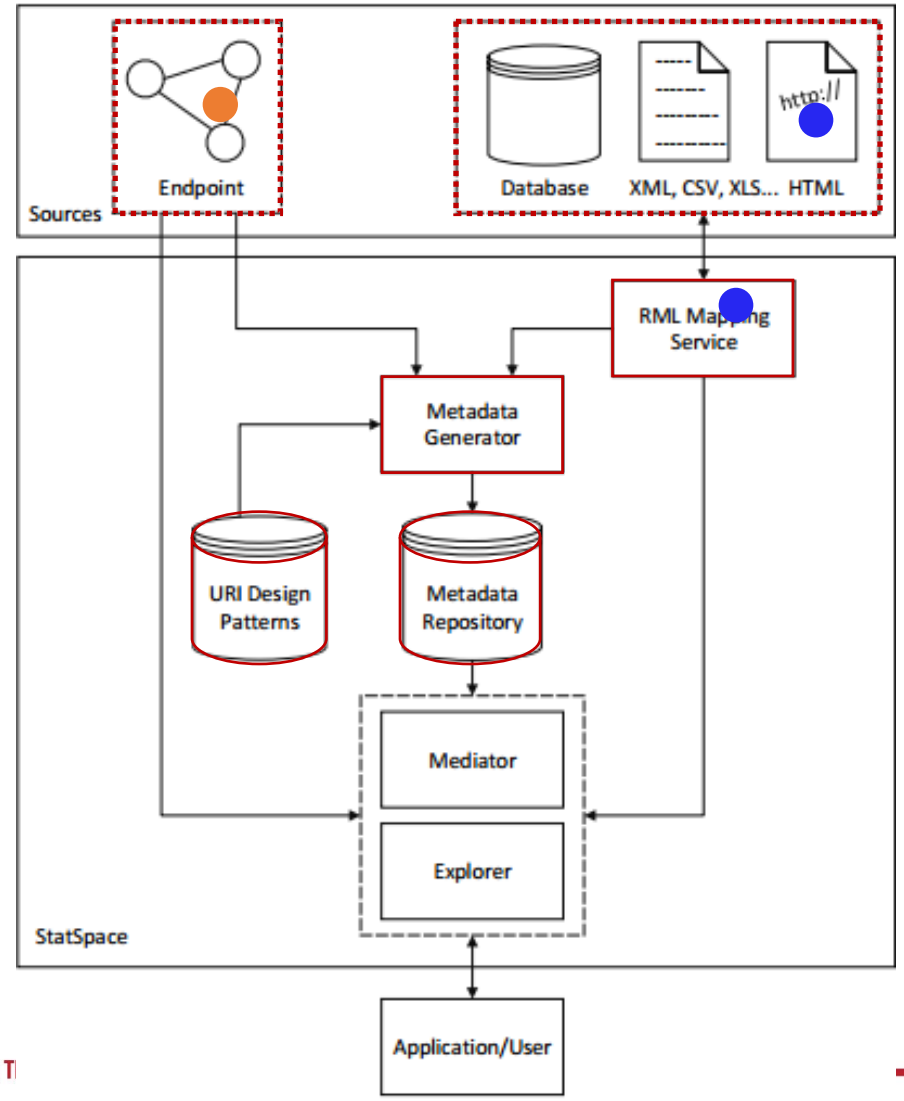
Architecture for Statistical Data Exploration and Integration

StatSpace Architecture



StatSpace Architecture

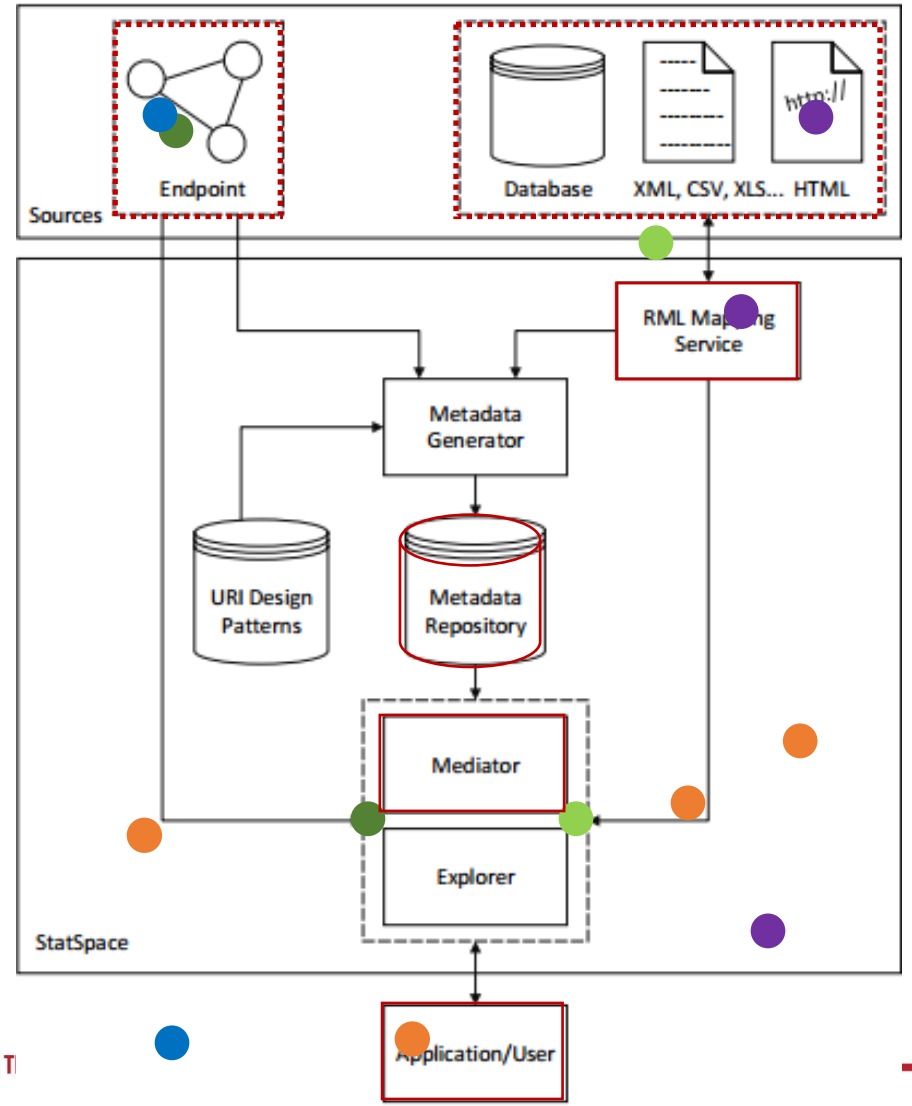
Metadata repository building



StatSpace Architecture

Mediator –

Cross-data set SPARQL querying



Mediator

Query Analysis

Query Rewriting

Result Rewriting

Result Integration

```
PREFIX qb:      <http://purl.org/linked-data/cube#>
PREFIX sdmxd:   <http://purl.org/linked-data/sdmx/2009/dimension#>
PREFIX sdmxm    <http://purl.org/linked-data/sdmx/2009/measure#>
PREFIX sdmxa:   <http://purl.org/linked-data/sdmx/2009/attribute#>
PREFIX dc:      <http://purl.org/dc/terms/>

SELECT * WHERE {
?ds dc:subject
<http://statspace.linkedwidgets.org/codelist/cl_subject/SP.POP.TOTL>.
        ?o qb:dataSet ?ds.
        ?o sdmxm:obsValue  ?obsValue.
        ?o sdmxd:refPeriod ?refPeriod.
        ?o sdmxd:refArea   ?refArea.
        ?o sdmxa:unitMeasure ?unit.
Filter(?refArea=<
http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom>)}
```

Mediator

Query Analysis

Query Rewriting

Result Rewriting

Result Integration

```
SELECT * WHERE {
    ?o qb:dataSet
<http://rdfdata.eionet.europa.eu/eurostat/data/demo_pjanbroad>.
    ?o sdmxm:obsValue ?obsValue.
    ?o sdmxd:timePeriod ?refPeriod.
    ?o sdmxd:refArea ?refArea.
    ?o sdmxd:freq <http://purl.org/linked-data/sdmx/2009/code#freq-A>.
    ?o sdmxd:age <http://dd.eionet.europa.eu/vocabulary/eurostat/age/TOTAL>.
    ?o sdmxd:sex  <http://purl.org/linked-data/sdmx/2009/code#sex-T>.
FILTER(?refArea= <http://dd.eionet.europa.eu/vocabulary/eurostat/geo/UK>) }
```

EEA data set query

```
http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&indicator=SP.POP.TOTL&refArea=GB
```

WB data set query

```
http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/uk7.ttl
```

UK data set query

Mediator

Query Analysis

Query Rewriting

Result Rewriting

Result Integration

- Rewrites each result based on co-reference relationships

```
http://dd.eionet.europa.eu/vocabulary/eurostat/geo/UK  =>
http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom
http://dd.eionet.europa.eu/vocabulary/worldbank/country/GB =>
http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom
```

- Consolidates different scales in results
  - EEA and WB data sets: absolute scale
  - UK data set: million scale => multiply one million
- Integrates results

Mediator

Query Analysis

Query Rewriting

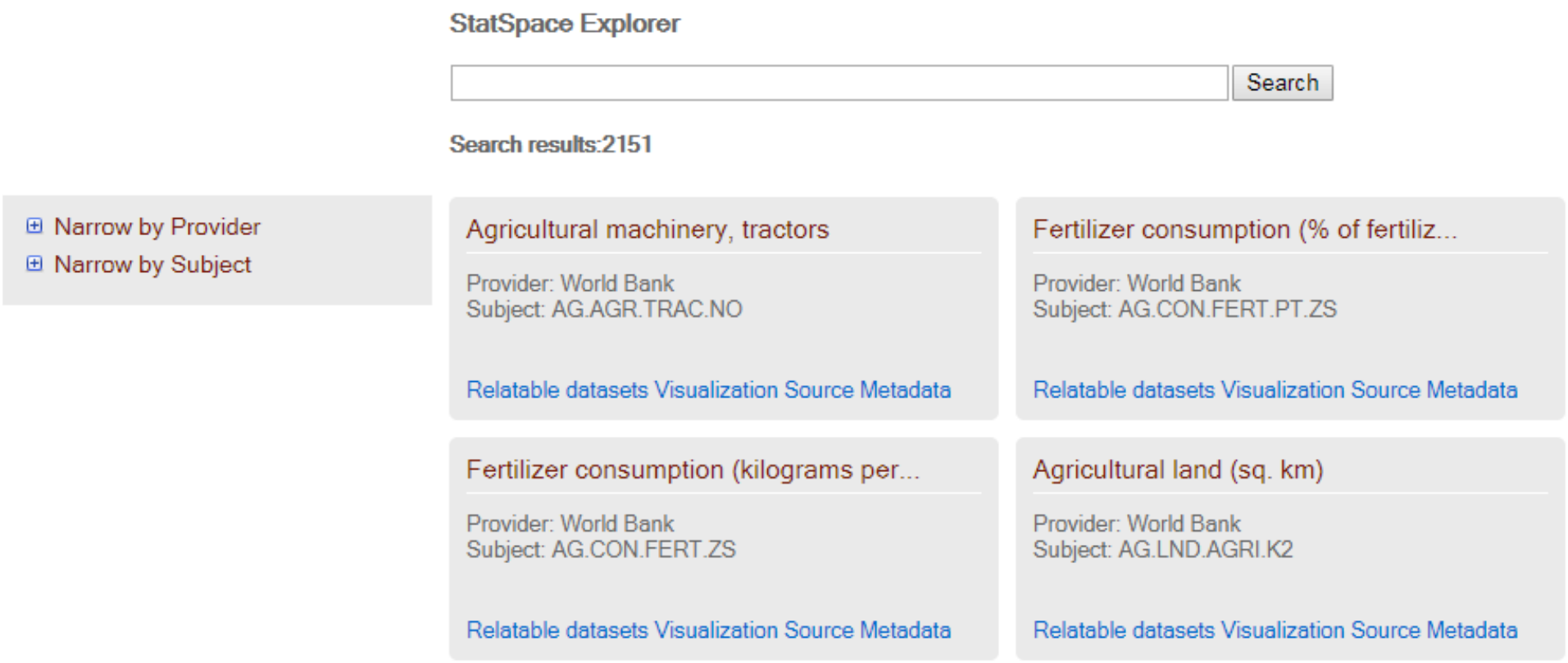
Result Rewriting

Result Integration

Number of matching datasets: 7

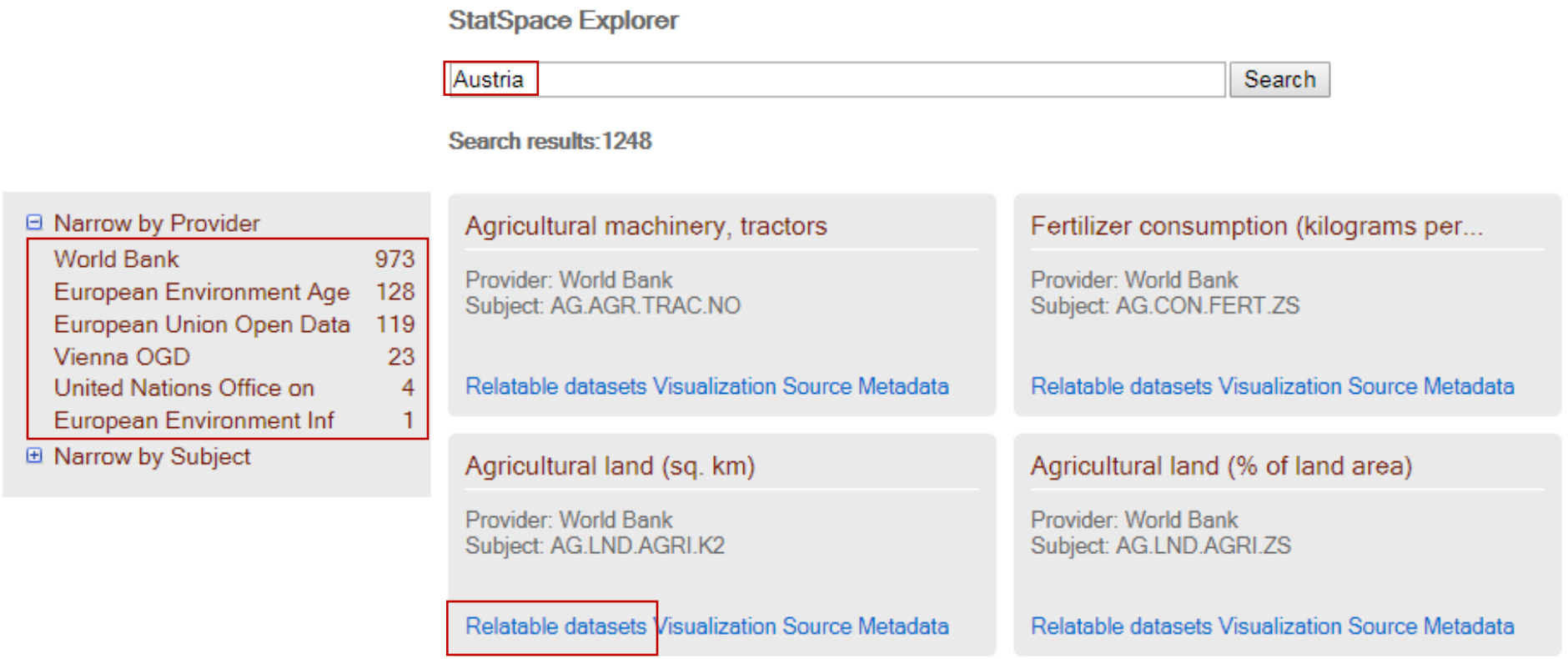
http://rdfdata.eionet.europa.eu/worldbank/dataset/wdi/SP.POP.TOTL	http://reference.data.gov.uk/id/gregorian-year/2010	62231336
http://rdfdata.eionet.europa.eu/eurostat/data/tps00001	http://reference.data.gov.uk/id/gregorian-year/2010	62510197
http://rdfdata.eionet.europa.eu/eurostat/data/demo_pjanbroad	http://reference.data.gov.uk/id/gregorian-year/2010	62510197
http://rdfdata.eionet.europa.eu/who/data/WHS9_86	http://reference.data.gov.uk/id/gregorian-year/2010	62066000
http://statspace.linkedwidgets.org/dataset/ONS-Population-change	http://reference.data.gov.uk/id/gregorian-year/2010	62800000
http://statspace.linkedwidgets.org/dataset/WorldBank-SP.POP.TOTL	http://reference.data.gov.uk/id/gregorian-year/2010	62766365
http://statspace.linkedwidgets.org/dataset/ONS-Population-1851-2014	http://reference.data.gov.uk/id/gregorian-year/2010	62759456

Explorer



Web-based application for non-expert users

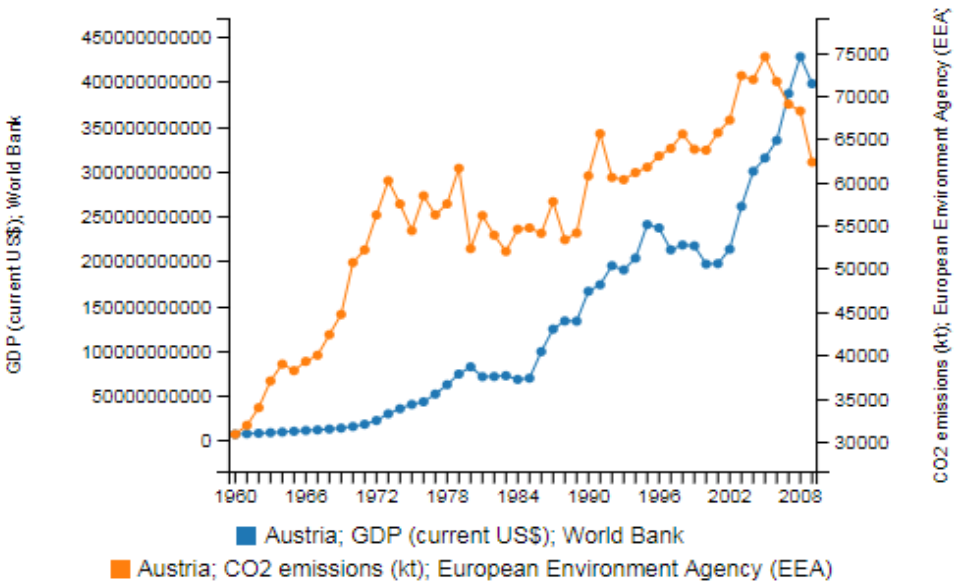
Finding all statistical data sets about an area

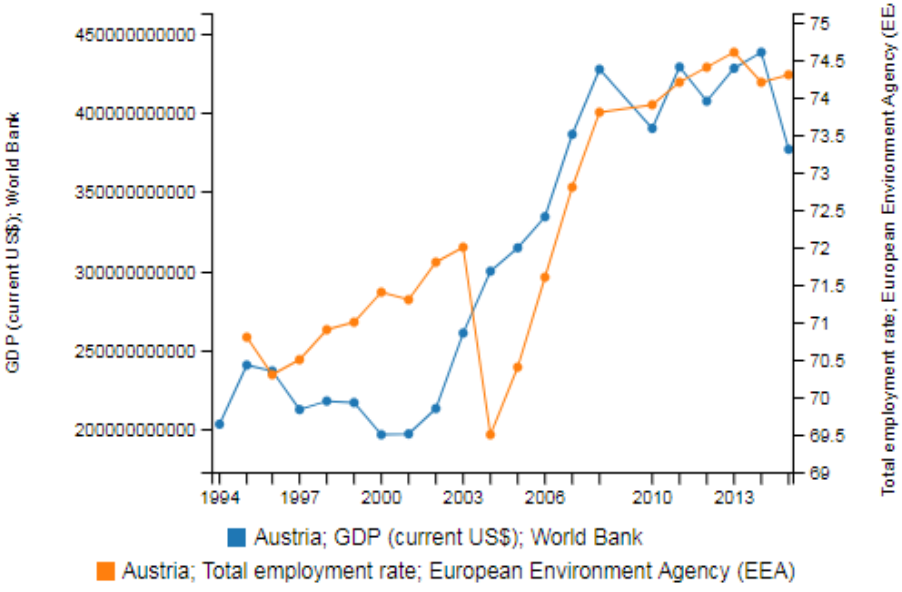


Identifying relatable data sets for a selected data set

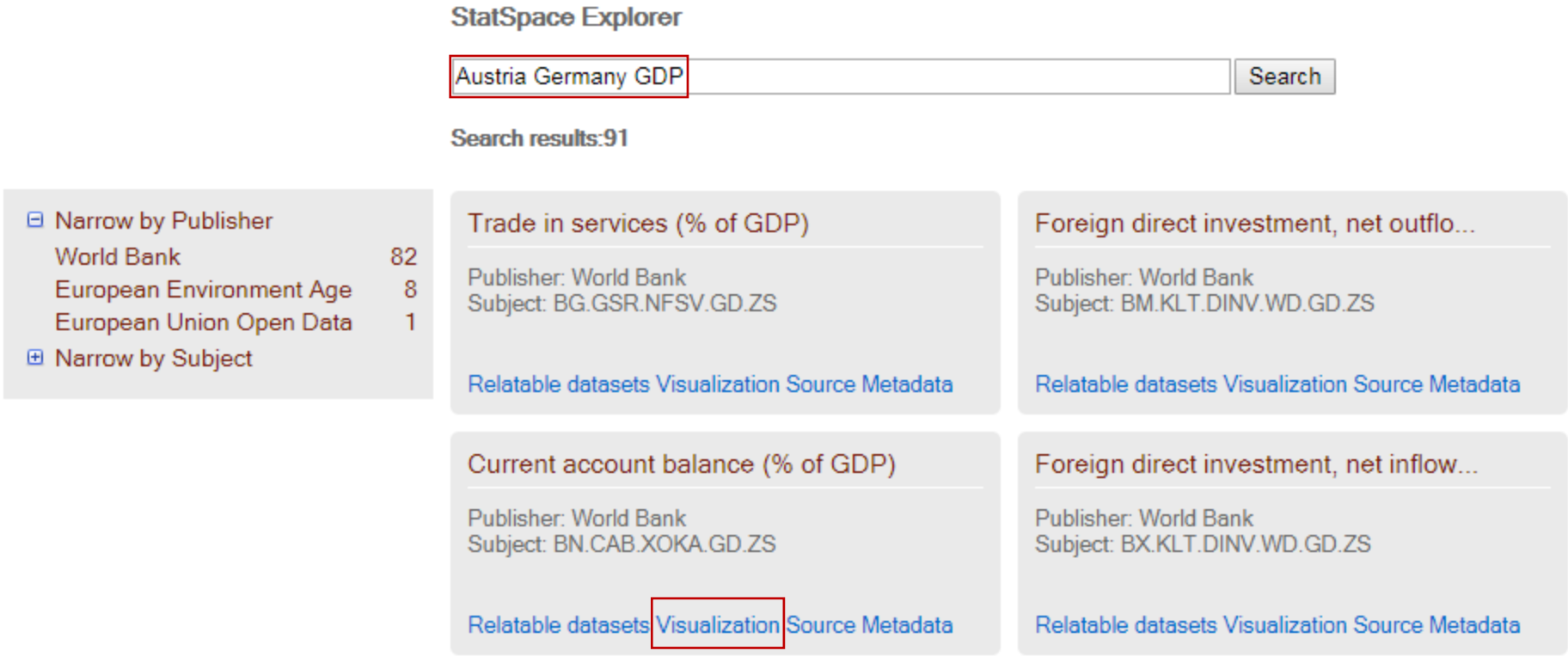
Relatable data sets

Data visualization between multiple indicators

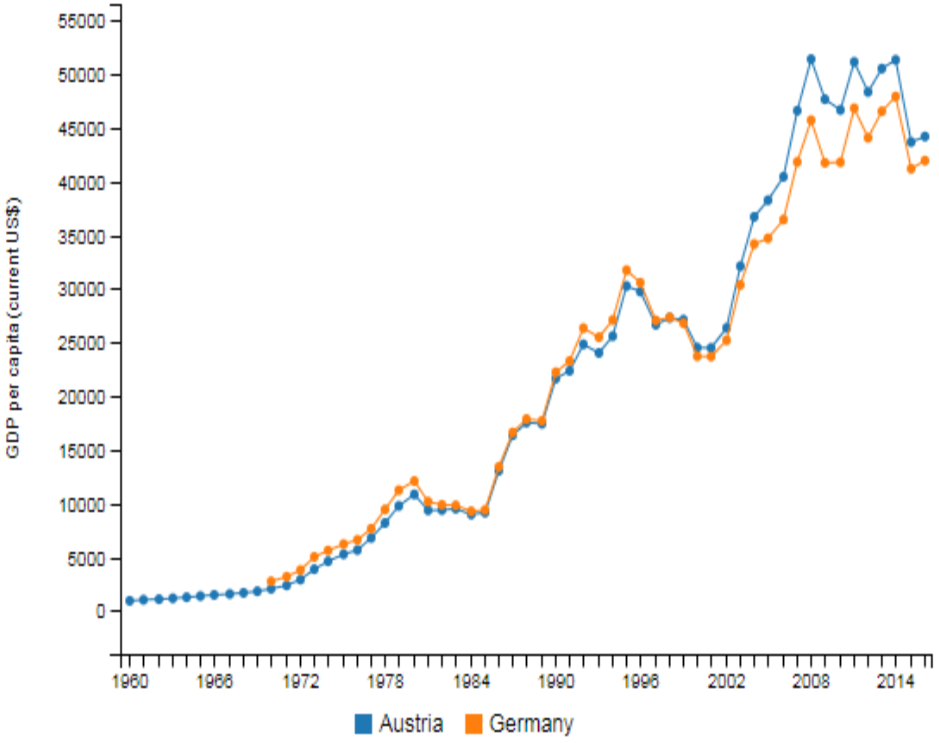


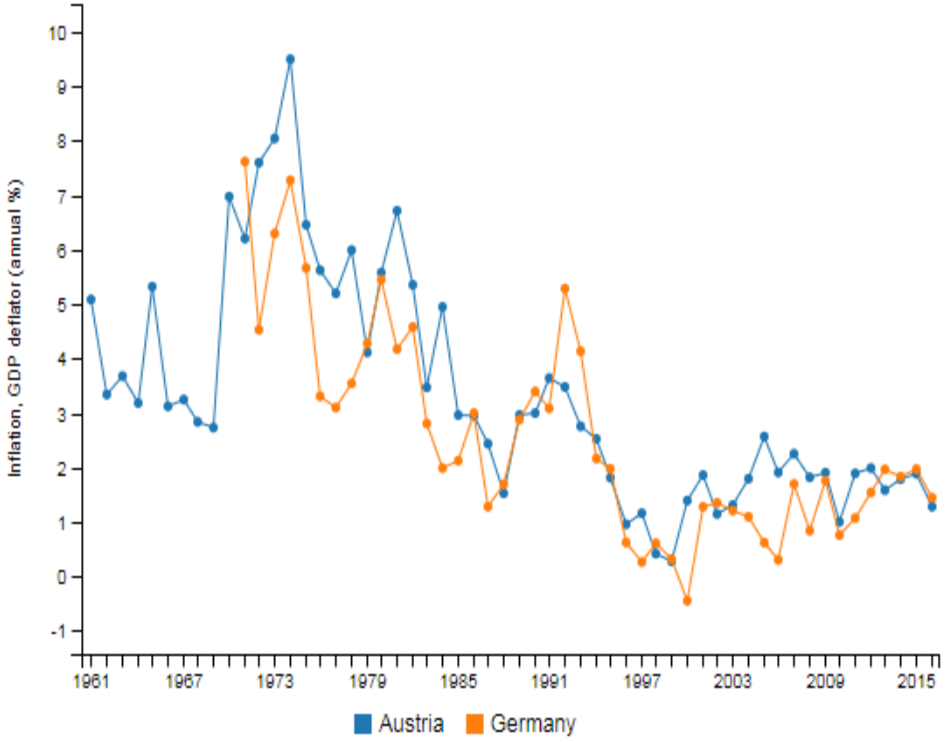


Comparing statistical data about different areas



Comparing statistical data about different areas





Evaluation

Evaluation – Metadata Repository

- 601 descriptions of RDF data sets
- 1,459 descriptions of raw data sets

Source	Data format	|Data sets used|	|Metadata descriptions|	Metadata size (KB)
EUODP	RDF	151	151	826
EEA	RDF	147	147	3,331
CSO	RDF	61	61	413
ScotStat	RDF	23	23	428
ODC	RDF	3	3	268
VOGD	RDF	39	214	2,385
ONS	XLS	2	8	103
World Bank	JSON	1,451	1,451	22,045

Sources and numbers of data sets covered

Evaluation – Metadata Generator

- Goal: evaluates the correctness of the mappings
- Settings
  - 25 mappings for spatial dimension and 25 mappings for temporal dimension between the URIs of the European Environment Agency data and shared URIs
  - Five experts
- Result
  - All experts agree the correctness of 48 mappings
  - One mapping: lacks necessary information for evaluation
    2007^http://www.w3.org/2001/XMLSchema#int and
    http://reference.data.gov.uk/id/gregorian-year/2007
  - One wrong mapping: due to error from Google's service

    "Twente, Overijssel" => "Hof van Twente, Overijseel"

    http://dd.eionet.europa.eu/vocabulary/eurostat/geo/NL213 (label: Twente) and
    http://statspace.linkedwidgets.org/codelist/cl_area/Netherlands/Overijssel/HofvanTwente

Evaluation - A comparison of research on statistical data integration

	Capadisli et al. [14]	Sabou et al. [11, 12]	Kämpgen et al. [26, 27]	Statspace
Original format	SDMX-XML	database	SPARQL	raw formats, SPARQL
Approach	warehousing	warehousing	warehousing	virtual integration
Requirements of data structure	same structure	–	same structure	relatable structures
Scale conversion	–	–	✓	✓
Uniform access	✗	✗	✗	✓

RQ1

RQ2

RQ3

Conclusions and Future Work

Conclusions

How can users be enabled to explore and integrate heterogenous statistical data sources?

- Architecture for statistical data exploration and integration
  - Up-to-date data to users
  - Uniform access to heterogeneous data sets
- Linked statistical data space: 1,800+ heterogeneous data sets from eight providers, http://statspace.linkedwidgets.org/

Research Questions

- RQ1. How can we address statistical data heterogeneity in terms of formats? RDF, mapping language
- RQ2. How can we establish interconnections between statistical data sets? shared URIs, mapping algorithms
- RQ3. How can we provide uniform access to heterogenous statistical data sets? metadata descriptions and mediator

References

1. Maali et al. A dynamic faceted browser for data cube statistical data, Workshop on Using Open Data, 2012
2. Salas et al., Publishing Statistical Data on the Web, Conference on Semantic Computing, 2012
3. Klímek et al., Payola: Collaborative Linked Data Analysis and Visualization Framework, *ESWC 2013 Satellite Events, 2013*
4. Ermilov et al., Linked Open Data Statistics: Collection and Exploitation, *Conference on Knowledge Engineering and Semantic Web, 2013*
5. Hoefler et al., Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints, LDOW, 2014
6. Kalampokis et al., Exploiting linked data cubes with opencube toolkit, ISWC Posters & Demos Track, 2014
7. Kämpgen et al., OLAP4LD: A Framework for Building Analysis Applications over Governmental Statistics, ESWC 2014 Satellite Events, 2014
8. Bayerl et al., Linked Data Integration based on the RDF Data Cube Vocabulary, *Conference on Web Intelligence, Mining and Semantics,* 2015

References

9. RDF Working Group, Resource Description Framework (RDF), https://www.w3.org/RDF/, 2014
10. Haase et al., A Comparison of RDF Query Languages, ISWC, 2004
11. Sabou et al., Tourmislod: A tourism linked data set, Semantic Web, 2013
12. Sabou et al., Linked Data for Cross Domain Decision-Making in Tourism, Conference on Information and Communication Technologies in Tourism, 2015
13. Kalampokis et al., Creating and utilizing linked open statistical data for the development of advanced analytics services, Semantics Statistics, 2014
14. Capadisli et al. Towards Linked Statistical Data Analysis, SemStat, 2013
15. Trinh et al., Linked Widgets: An Approach to Exploit Open Government Data, DATA conference, 2014

References

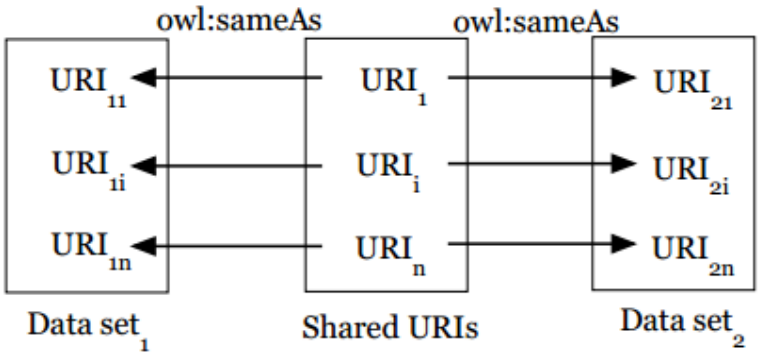
16. Langegger et al., XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL, ISWC, 2009
17. Connor et al., Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL, ISWC, 2010
18. Bischof et al., Mapping between RDF and XML with XSPARQL. *Journal on Data Semantics, 2012*
19. Bizer et al., D2R map-a database to RDF mapping language, *WWW – Poster Track*. 2003
20. Das et al., R2RML, https://www.w3.org/TR/r2rml/, 2012
21. Ghasemi et al., M2RML: Multidimensional to RDF Mapping Language, *Workshop on Database and Expert Systems Applications,* 2014
22. Dimou et al., RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data, LDOW, 2014
23. RML Processor. https://github.com/RMLio/RML-Processor

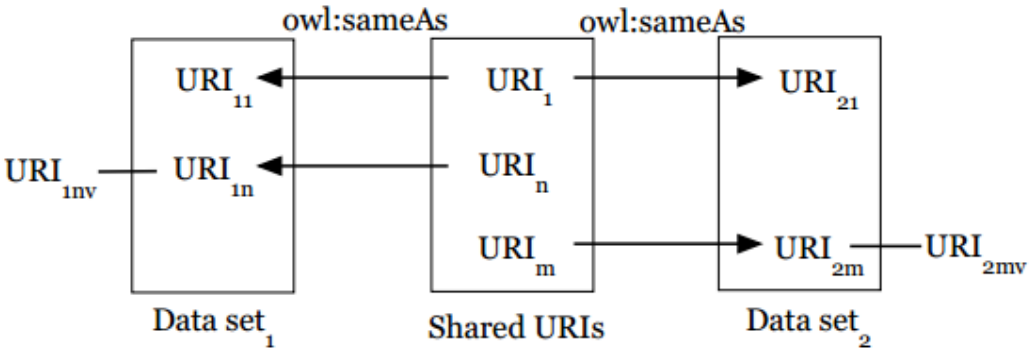
References

24. Hugh Glaser et al., Managing co-reference on the semantic web, LODW, 2009
25. Kai Schlegel et al., Balloon fusion: SPARQL rewriting based on unified co-reference information, *Conference on Data Engineering,* 2014
26. Kämpgen et al., Querying the Global Cube: Integration of Multidimensional Datasets from the Web, Conference on Knowledge Engineering and Knowledge Management, 2014
27. Benedikt Kämpgen, Flexible Integration and Effcient Analysis of Multidimensional Datasets from the Web, PhD Thesis, 2015
28. Mutlu et al., Automated Visualization Support for Linked Research Data, I-SEMANTICS Posters and Demos, 2013
29. Kotu et al., Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. Morgan Kaufmann, 2014
30. Saed Sayad. Data Exploration, http://chem-eng.utoronto.ca/~datamining/Presentations/Data_Exploration.pdf, 2010

BACKUP SLIDES

Relatable data structures



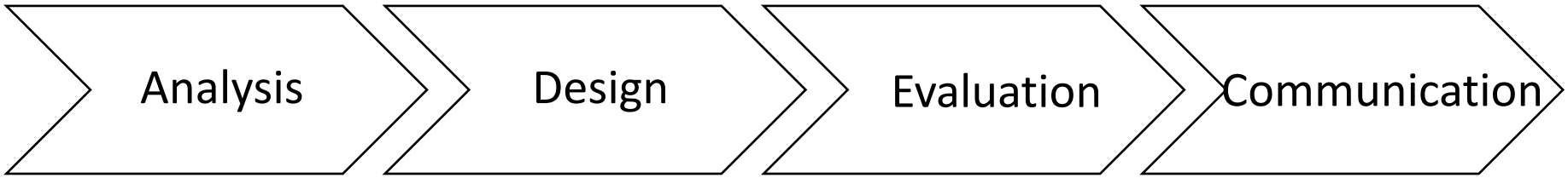


uk:country	uk:year	uk:value	uk:unit
uk:UnitedKingdom	uk:2013	64.1	uk:Million

wb:country	wb:year	wb:value	wb:unit
wb:GB	wb:2013	64128226	wb:AbsoluteScale

sdmxd: freq	sdmxd: timePeriod	sdmxd: refArea	sd-mxd:age	sdmxd: sex	sdmxm: obsValue
sdmx-code:freqA	2013	geo:UK	ag:TOTAL	sdmxcode: sex-T	63,905,297
sdmx-code:freqA	2013	geo:UK	ag:Y_LT15	sdmxcode: sex-T	11,260,549
sdmx-code:freqA	2013	geo:UK	ag:Y15–64	sdmxcode: sex-F	20,917,257

Design Science Research Methodology



Peffers et al., ***A design science research methodology for information systems research***, Journal of management information systems, 2007

Shared URIs – Overview

- URIs for Components
  - 11 URIs for dimensions, e.g., sdmxd:refArea, sdmxd:sex, sdmxd:freq, sdmxd:occupation, sdmx:civilStatus, etc.
  - 1 URI for measure (generic measure): sdmxm:obsValue
  - 1 URI for attribute (unit of measure): sdmxa:unitMeasure
- URIs for Subjects
  - *Topic.General Subject.Specific Subject.Extensions* (e.g. *SP.POP.TOTL.FE*)
  - 55 codes for Topics, 198 codes for General topics, 629 codes for Specific Subjects, 147 codes for Extensions
- URIs for Units: 43 URIs, e.g., http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/P6, refers to people in million (10^6)

Shared URIs – Code lists of Components

Code list	URI design patterns	Number of URIs	Hierarchy Support
CL_Area	1	Unlimited	Yes
CL_Period	11	Unlimited	Yes
CL_Economic_Activity	1	996	Yes
CL_Age	1	209	Yes
CL_Education_Level	1	9	No
CL_COICOP	1	230	Yes
CL_COFOP	1	188	Yes
CL_COPP	1	51	Yes
CL_COPNI	1	65	Yes
CL_Occupation	1	619	Yes
CL_Currency	1	180	No
CL_Civil_Status	1	8	No
CL_Frequency	1	9	No
CL_Sex	1	5	Yes
CL_Unit_Measure	1	43	No

Characteristics of code lists

Shared URIs – Code lists of Components

The *expenditure dimension* consists of four code lists, i.e.,

- classification of individual consumption by purpose (COICOP): food, clothing, etc.
- classification of functions of government (COFOG): defense, health, etc.
- classification of purposes of non-profit institutions serving households (COPNI): R&D for health, R&D for education, etc.
- classification of outlays of producers by purpose (COPP): sales promotion, maintenance, etc.

Evaluation – Shared URIs

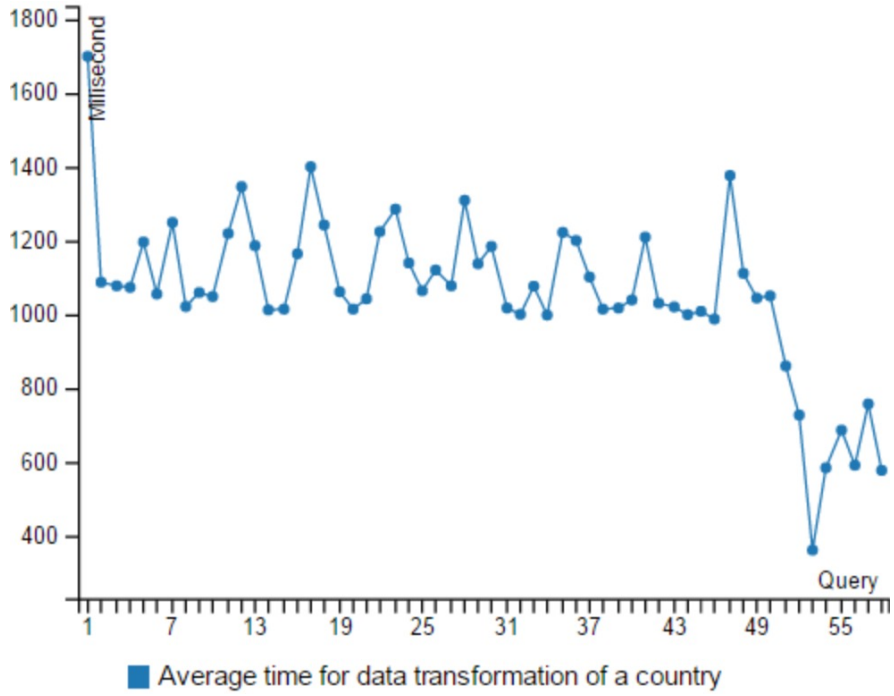
Source	|Area|	|Time|	|Age|	|Sex|	|Frequency|	|Occupation|
EUODP	31	12	0	0	0	0
EEA	3,114	609	132	0	3	0
ScotStat	16,718	167	83	3	0	0
CSO	4,806	1	40	3	0	11
ODC	33,286	167	0	0	4	0
VOGD	2,542	823	113	3	0	0
World Bank	304	57	–	–	–	–
URI Design Patterns	Unlimited	Unlimited	209	5	9	619

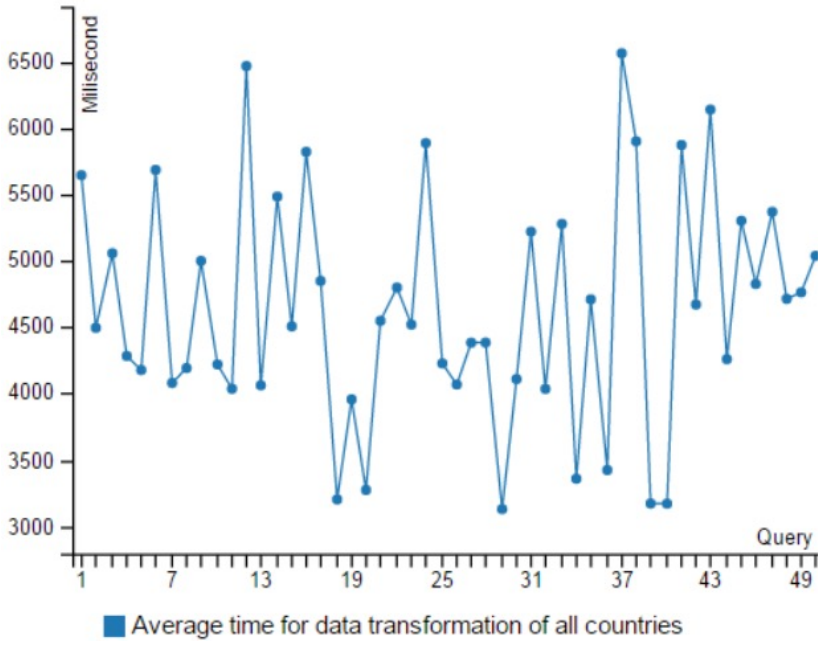
A comparison of size of code lists for six popular dimensions

Evaluation – RML Mapping Service

- Test environment
  - Windows 7 Enterprise 64 bit, Ram - 8.00 GB of DDR3, Processor - Intel Core i5-3470, CPU@3.20 GHz
  - Internet connection speed: around 800 Mbps for download and 900 Mbps for upload
- Number of queries
  - 50 queries to transform data of a single country from the WB
  - 8 queries to transform data from the UK data source
  - 50 queries to transform data of all countries from the WB

Evaluation – RML Mapping Service





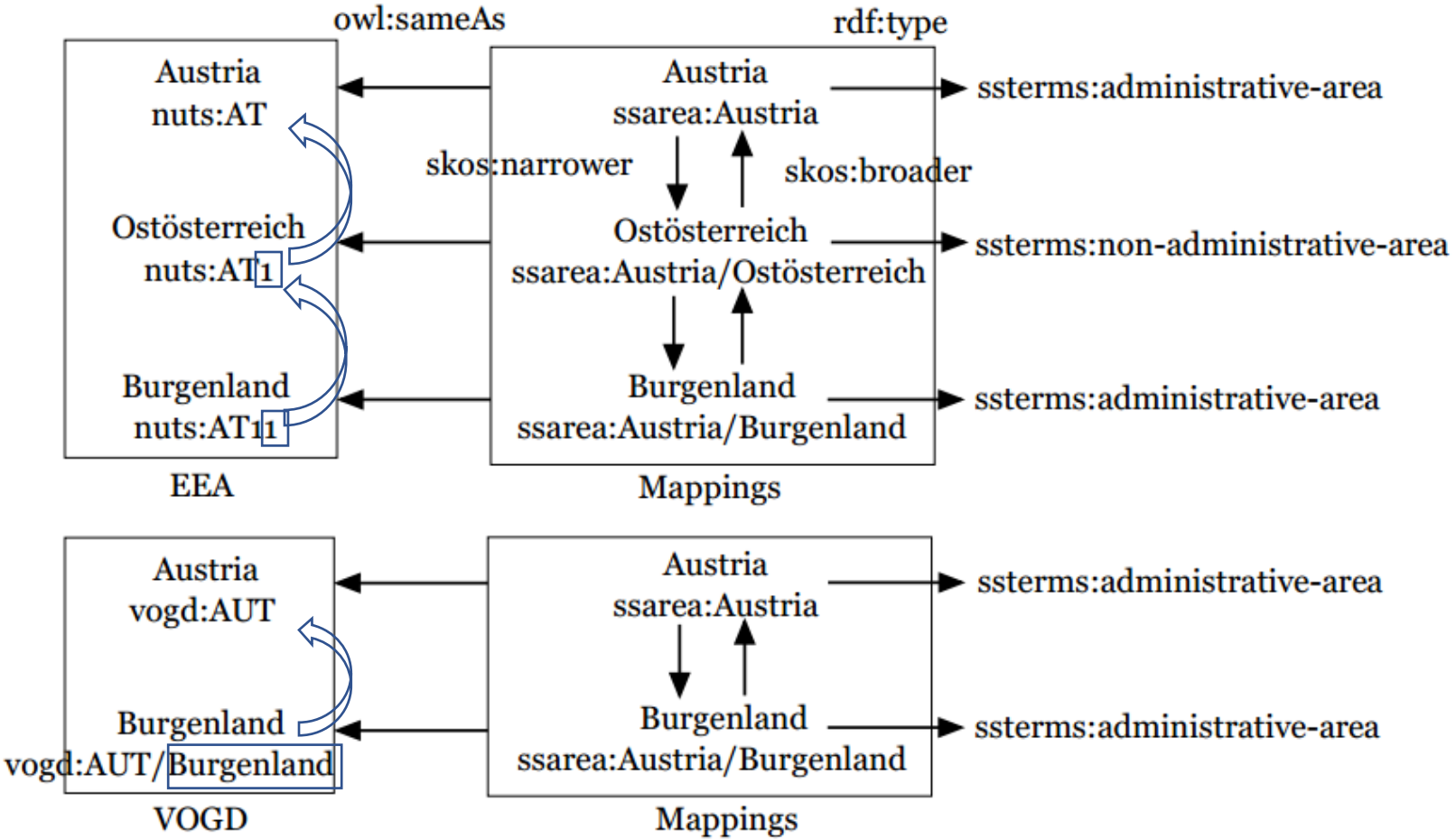
Limitations

- Data collection methodology is not modelled
- Territorial changes and differences in fiscal years are not considered
- Do not have participation of data publishers/experts to valid mappings between URIs used in data sources and shared URIs
- Mediator's implementation needs to be improved

Spatial Dimension Mapping Algorithm

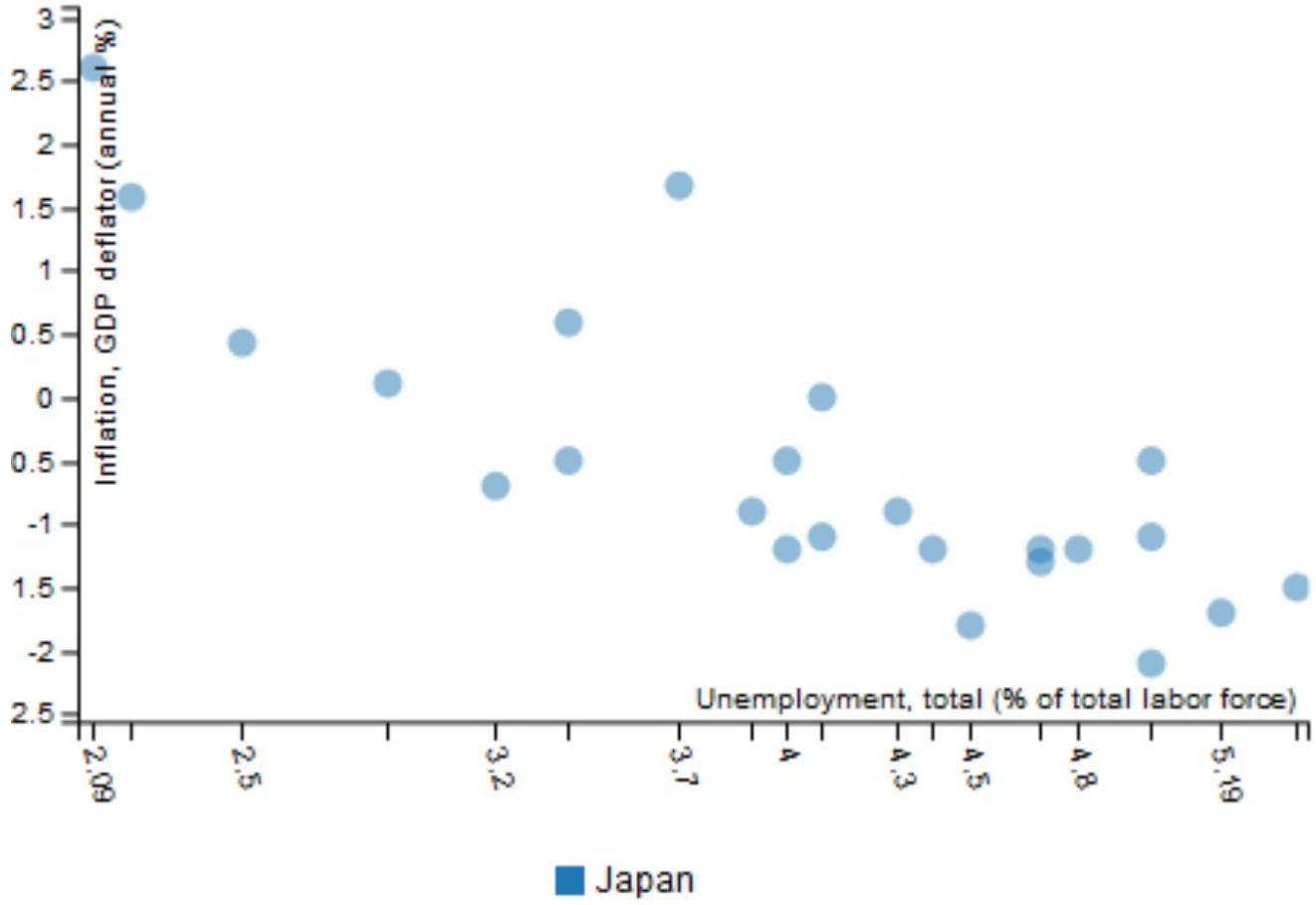
- Ambiguity: *Vienna – Austria*, *Vienna – USA*
  - Orders input areas in the ascending order of geographical levels
  - Adds label of the broader area (e.g., Austria) to the queries of its narrower areas (e.g., Vienna)
- Direction-based regions: Ostösterreich - Eastern Austria
  - Assigns URI based on URI of its broader area and its label
  - Sets type "non-administrative area"

Spatial Dimension Mapping Algorithm - Example



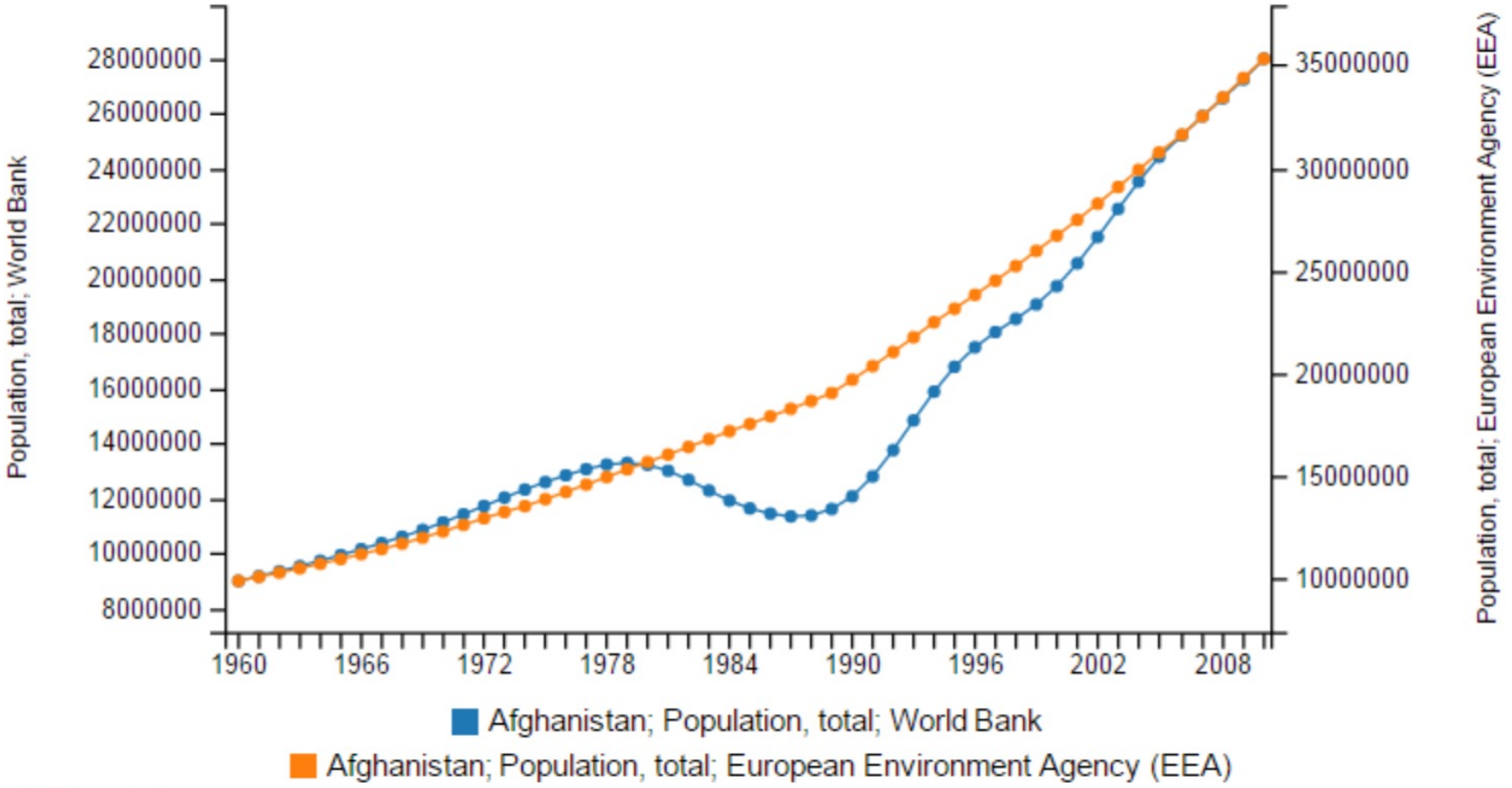
nuts: http://dd.eionet.europa.eu/vocabulary/common/nuts/
vogd: http://ogd.ifs.tuwien.ac.at/vienna/geo/
ssarea: http://statspace.linkedwidgets.org/codelist/cl_area/

Use Case – Scatter plot for Correlation Mining

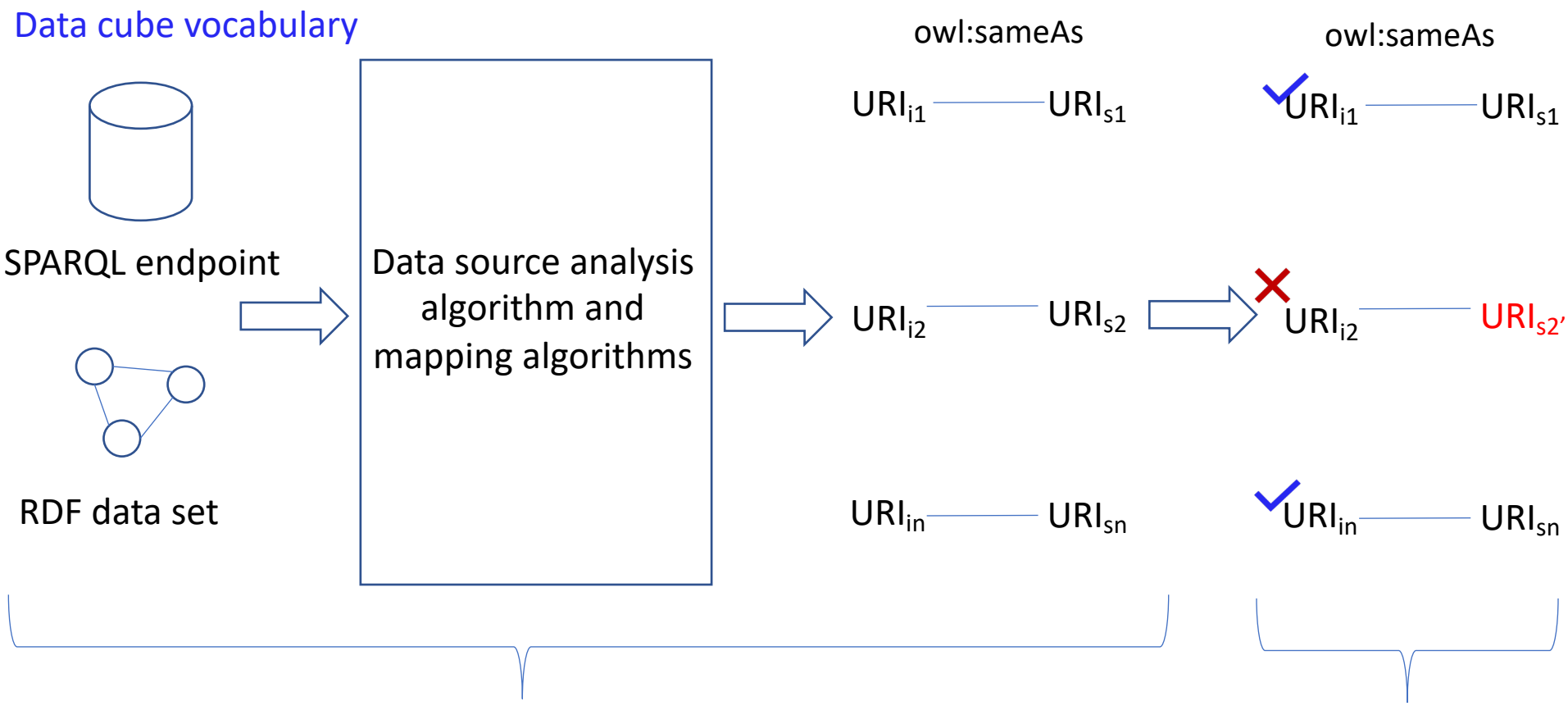


Relationship between Inflation and Unemployment indicators for Japan

Use Case – Showing Out-of-date Data



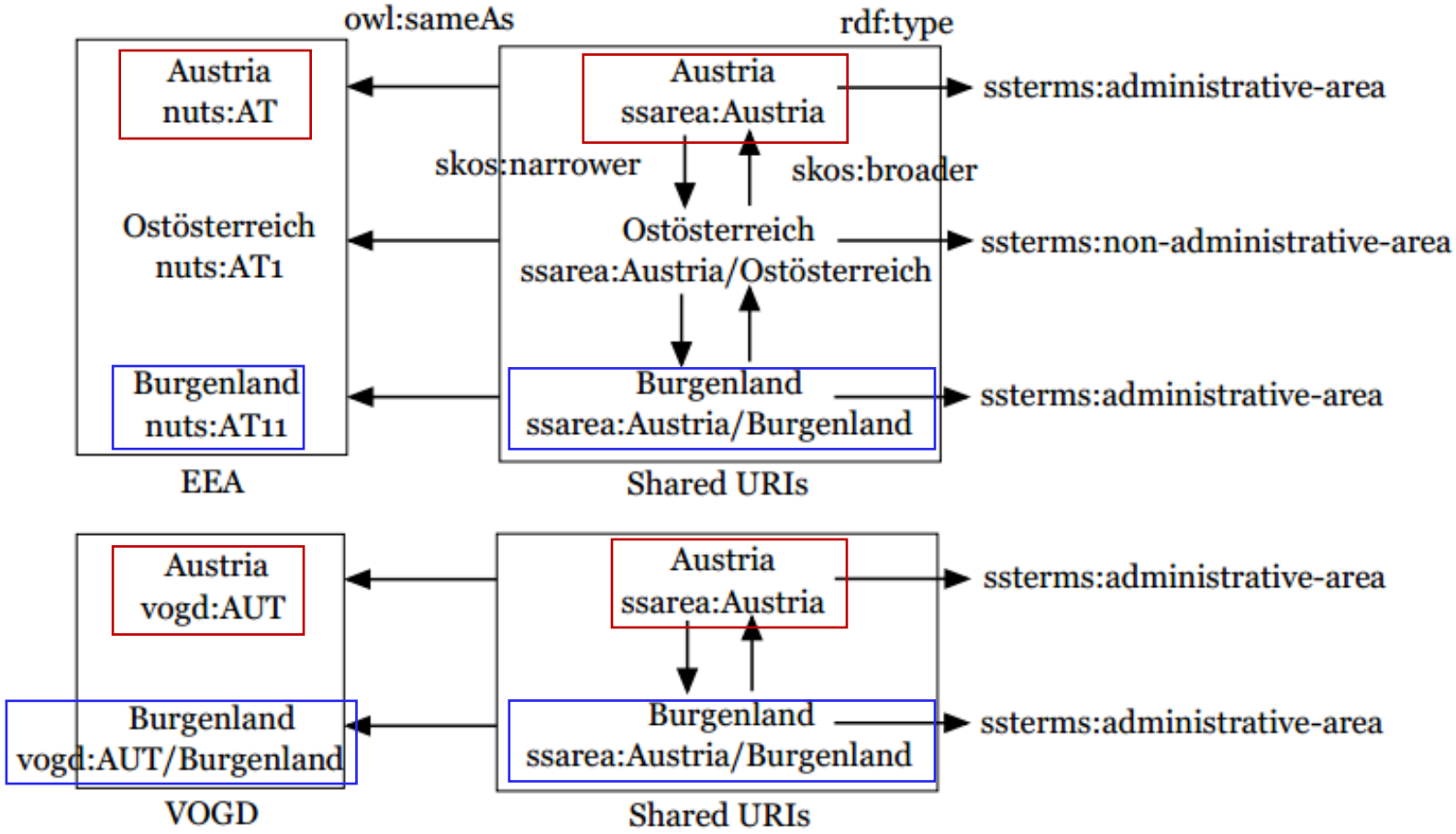
Mapping Generation



Mapping Generation

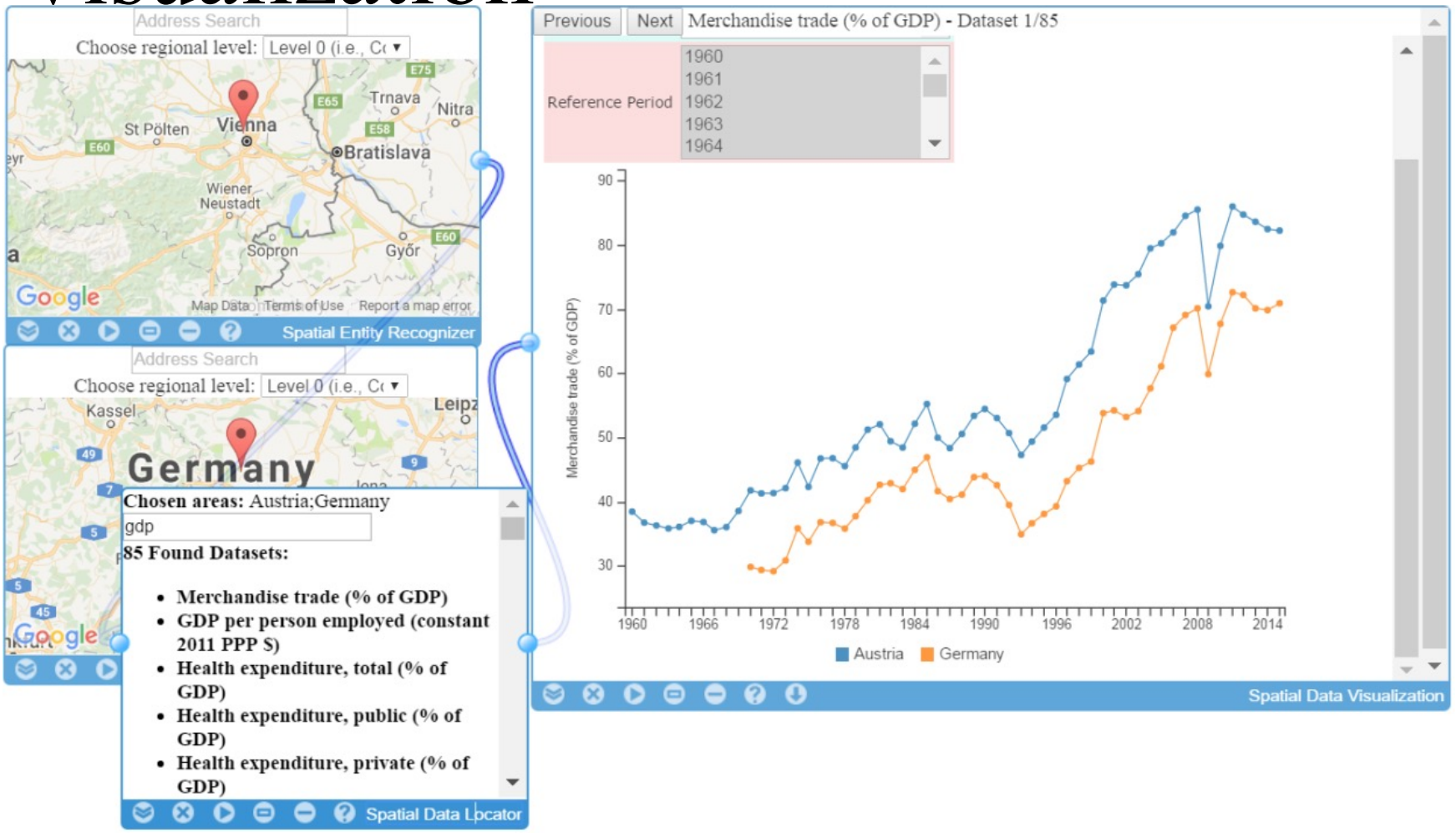
- Data source analysis algorithm
  - Input: a SPARQL endpoint or an RDF data set following Data cube vocabulary
  - Output: A list of data sets, each contains label, components, and their values
- Mapping algorithms
  - Eleven mapping algorithms for eleven dimensions
  - One mapping algorithm for the unit attribute
  - Mapping algorithm for spatial dimension: relies on Google's geocoding service
  - Mapping algorithms for other components: rely on patterns and keywords

Spatial Dimension Mapping Algorithm - Example



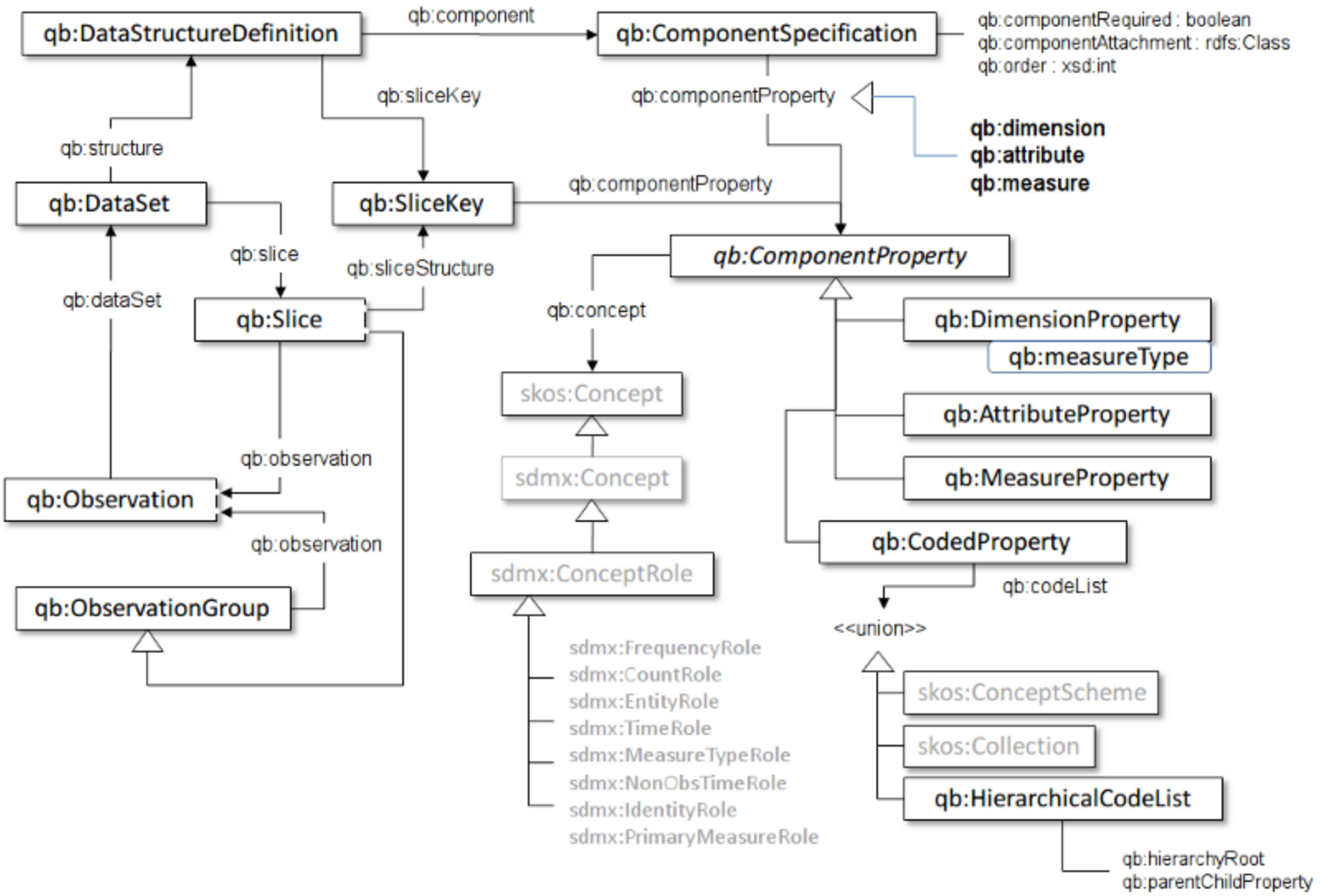
nuts: http://dd.eionet.europa.eu/vocabulary/common/nuts/
vogd: http://ogd.ifs.tuwien.ac.at/vienna/geo/
ssarea: http://statspace.linkedwidgets.org/codelist/cl_area/

Use Case – Spatial Data Visualization



Statistical data comparison of different areas

Data Cube Vocabulary



Data Cube Vocabulary - Example