gesis
Leibniz Institute for the Social Sciences

Object Property Matching utilizing the Overlap between Imported Ontologies

Ben Zapilko Brigitte Mathiak
GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

Member of the Leibniz Association

Background

```
ex1:Obs1 a   qb:Observation;
             ex1:date   „2004“;
             ex1:value   „5367561“;
             ex1:ages   codelists:cl/ages#20-29;
             ex1:gender   codelists:cl/gender#Sex-M;
             ex1:geo   countries:countries#BE.
```

Open Government Data, e.g. statistical data, is often used for research.

Background

```
ex1:Obs1 a   qb:Observation;
         ex1:date   „2004“;
         ex1:value   „5367561“;
         ex1:ages   codelists:cl/ages#20-29;
         ex1:gender   codelists:cl/gender#Sex-M;
         ex1:geo   countries:countries#BE.
```

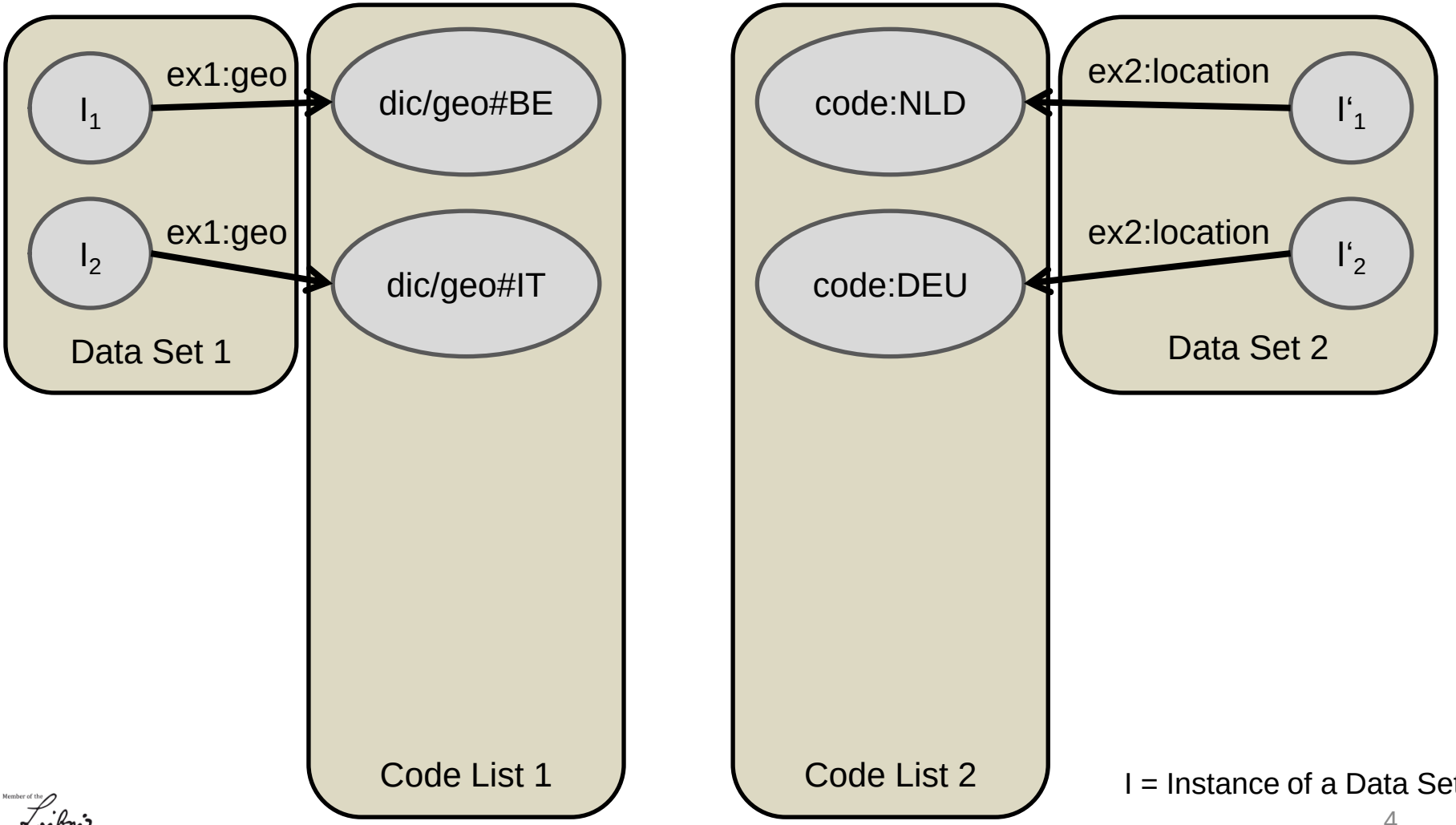
Open Government Data, e.g. statistical data, is often used for research.

```
countries:countries#BE a owl:Class;
rdfs:label „Belgium“.

countries:countries#NL a owl:Class;
rdfs:label „The Netherlands“.

countries:countries#LU a owl:Class;
rdfs:label „Luxembourg“.
```

```
codelists:cl/ages#20-29 a owl:Class;
rdfs:label „From 20 to 29 years“.

codelists:cl/ages#30-39 a owl:Class;
rdfs:label „From 30 to 39 years“.

codelists:cl/gender#Sex-F a owl:Class;
rdfs:label „Female“.

codelists:cl/gender#Sex-M a owl:Class;
rdfs:label „Male“.
```

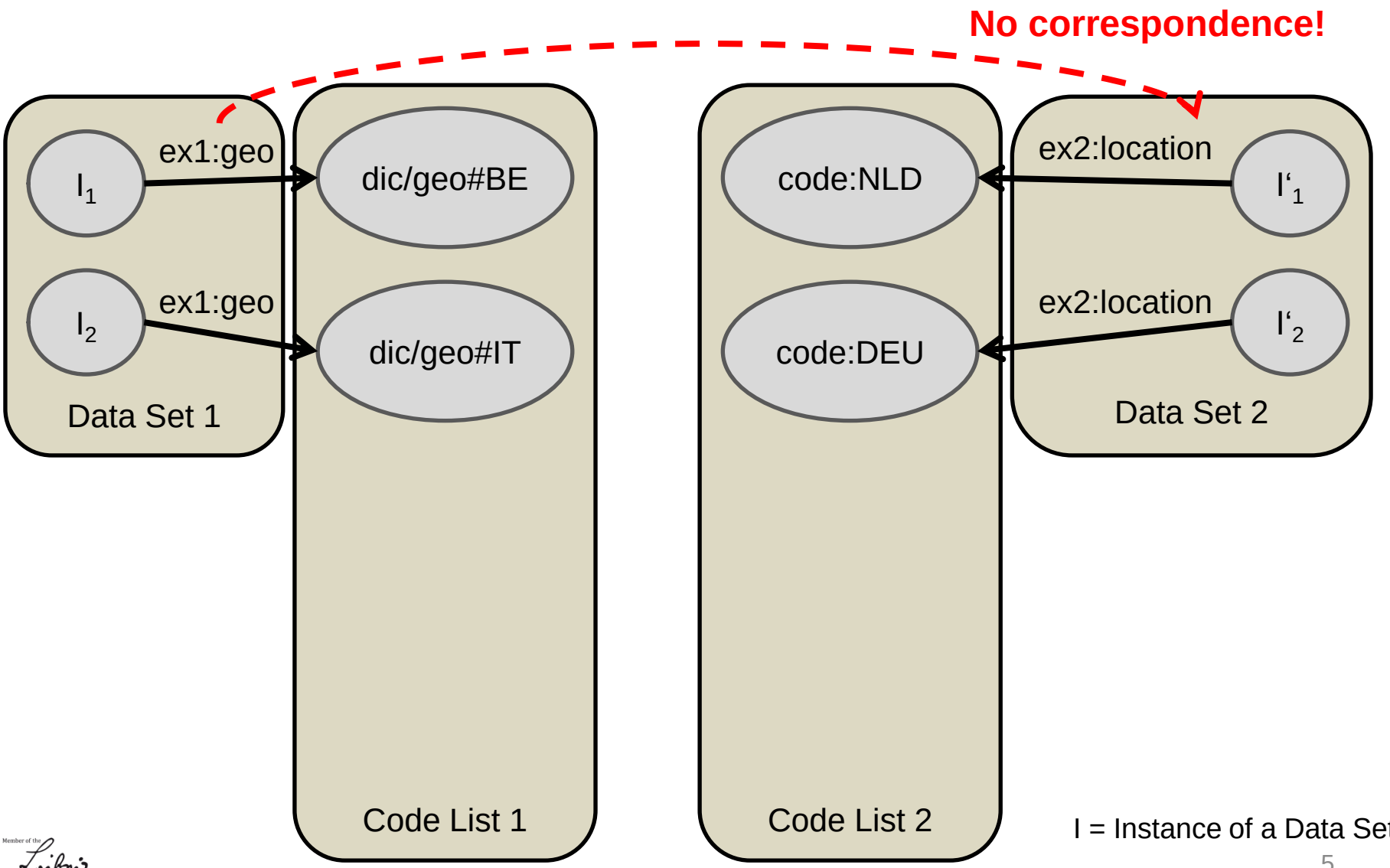
Result of analysis of 140 data sets of DataHub, PlanetData Wiki and OAEI

Matching



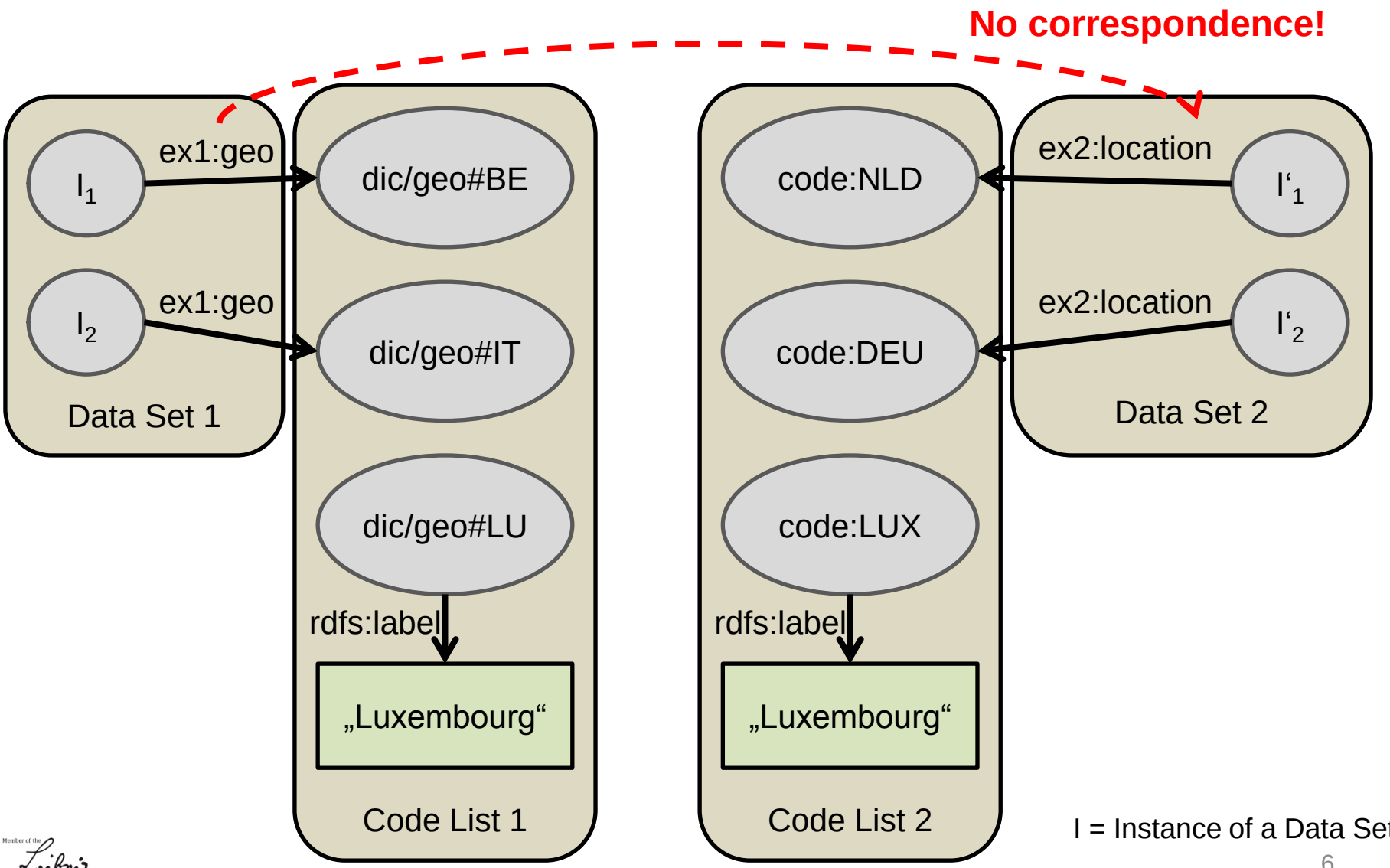
I = Instance of a Data Set

Matching



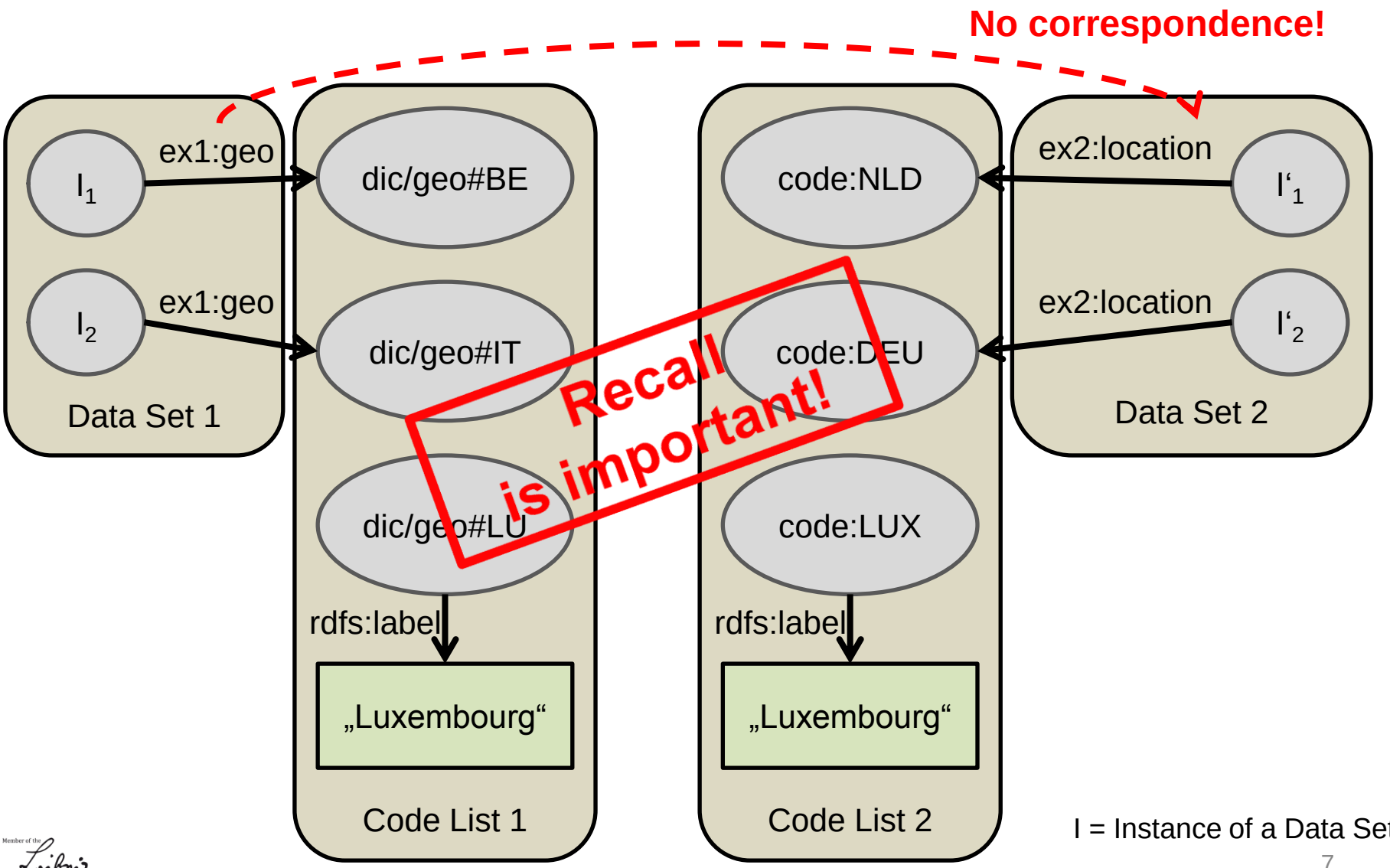
I = Instance of a Data Set

Matching



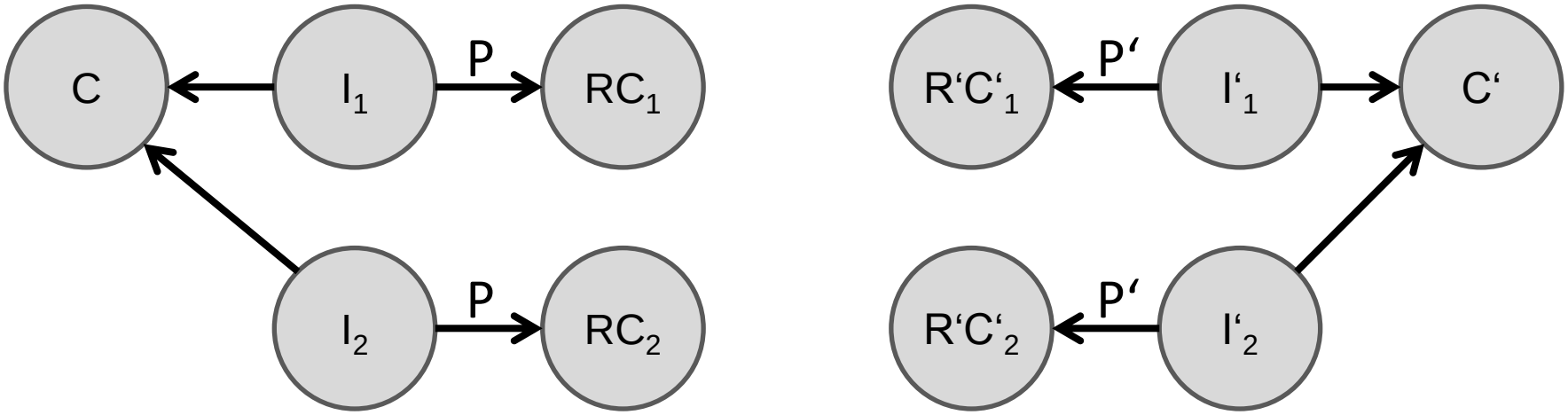
I = Instance of a Data Set

Matching



I = Instance of a Data Set

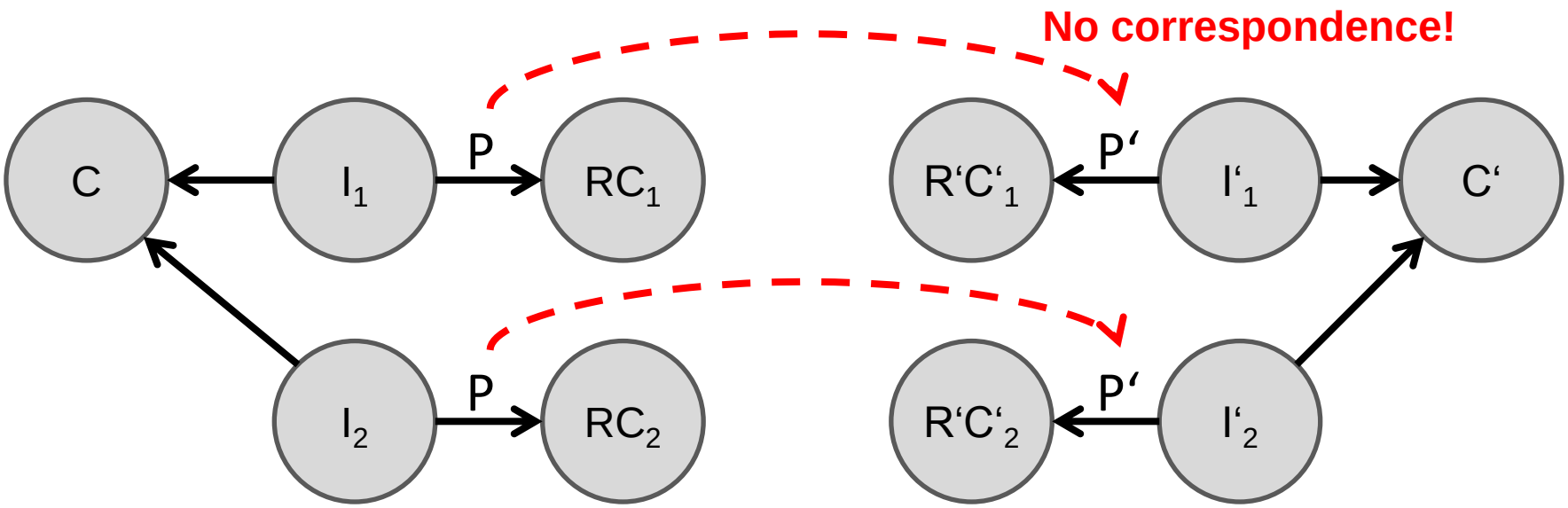
Problem



- Correspondences between object properties may be missed when the linked classes are from imported ontologies
- Matching systems are not trained against this problem, because there are no such data sets in current, established evaluation campaigns

C = Class of an Ontology
P = Object Property
I = Instance of Class C
RC = Class of Imported Ontology

Problem

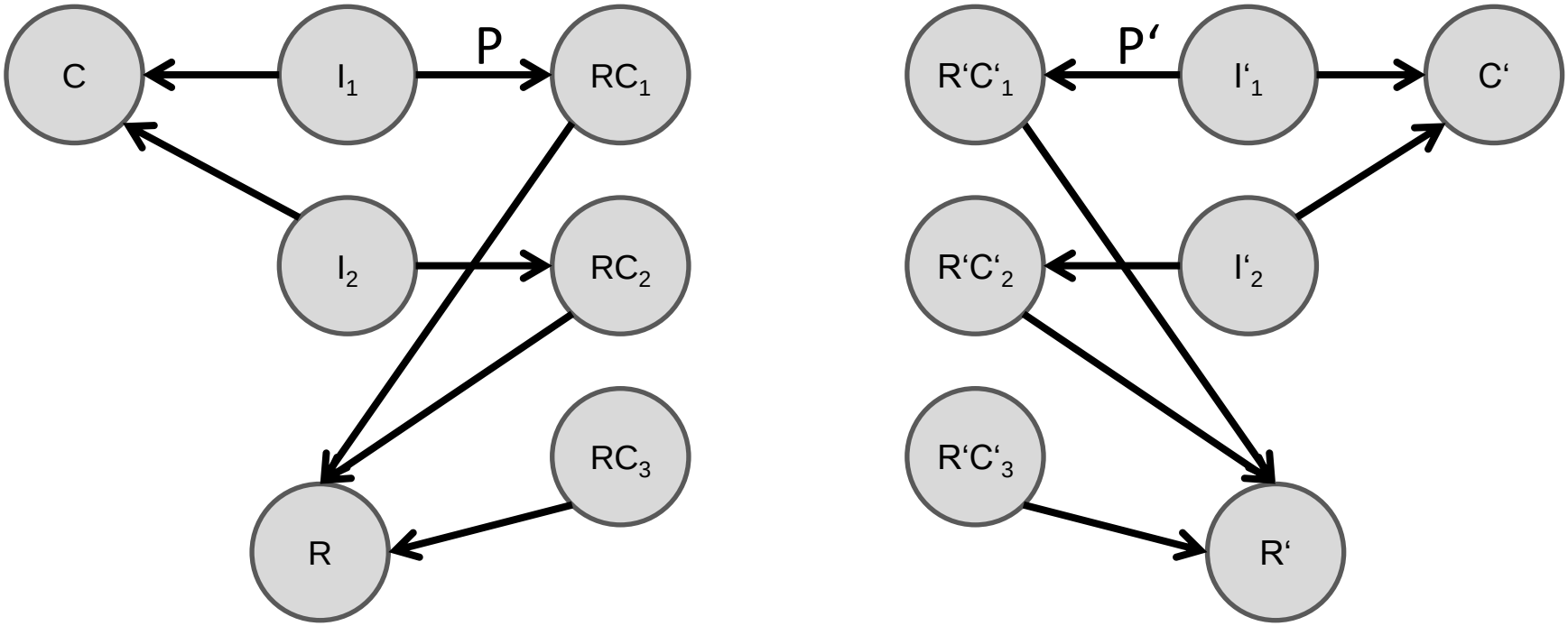


- Correspondences between object properties may be missed when the linked classes are from imported ontologies
- Matching systems are not trained against this problem, because there are no such data sets in current, established evaluation campaigns

C = Class of an Ontology
P = Object Property
I = Instance of Class C
RC = Class of Imported Ontology

Approach

Utilizing the overlap between the imported ontologies to generate a correspondence between their linking object properties

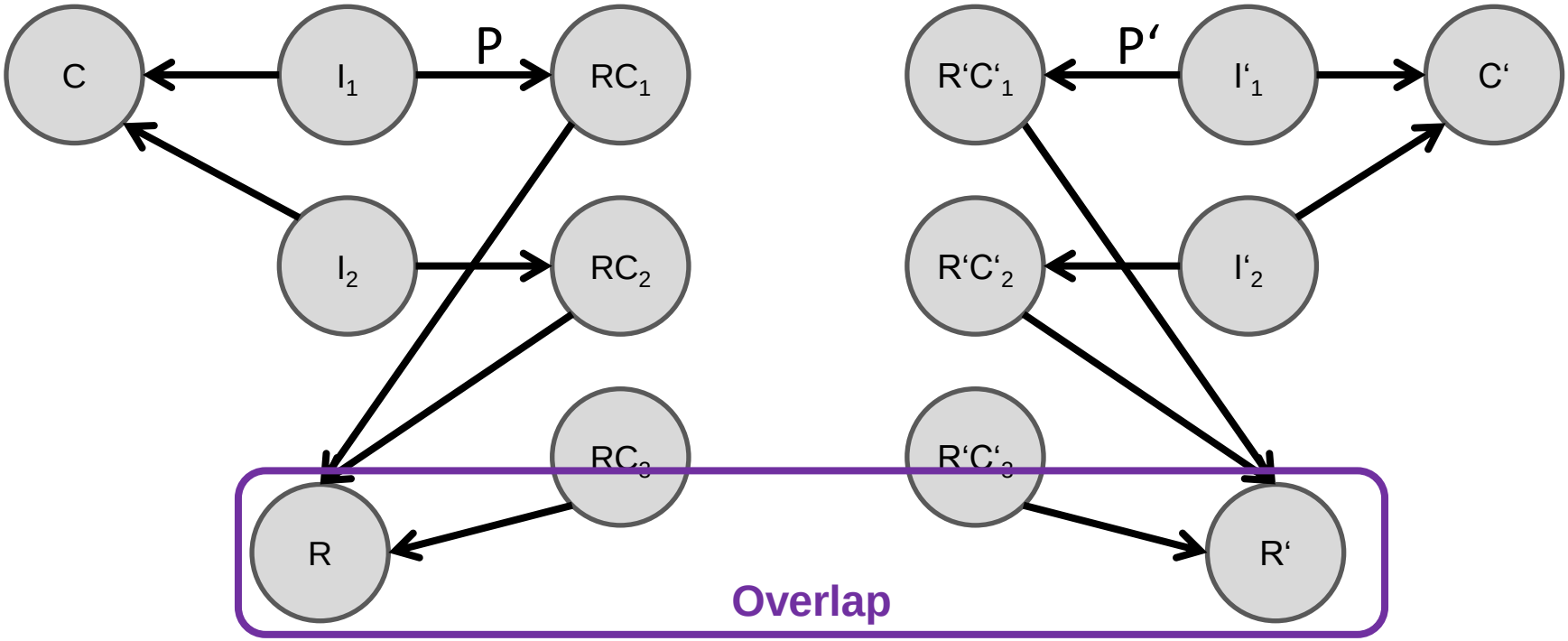


C = Class of an Ontology
P = Object Property
R = Imported Ontology

I = Instance of Class C
RC = Class of Imported Ontology

Approach

Utilizing the overlap between the imported ontologies to generate a correspondence between their linking object properties

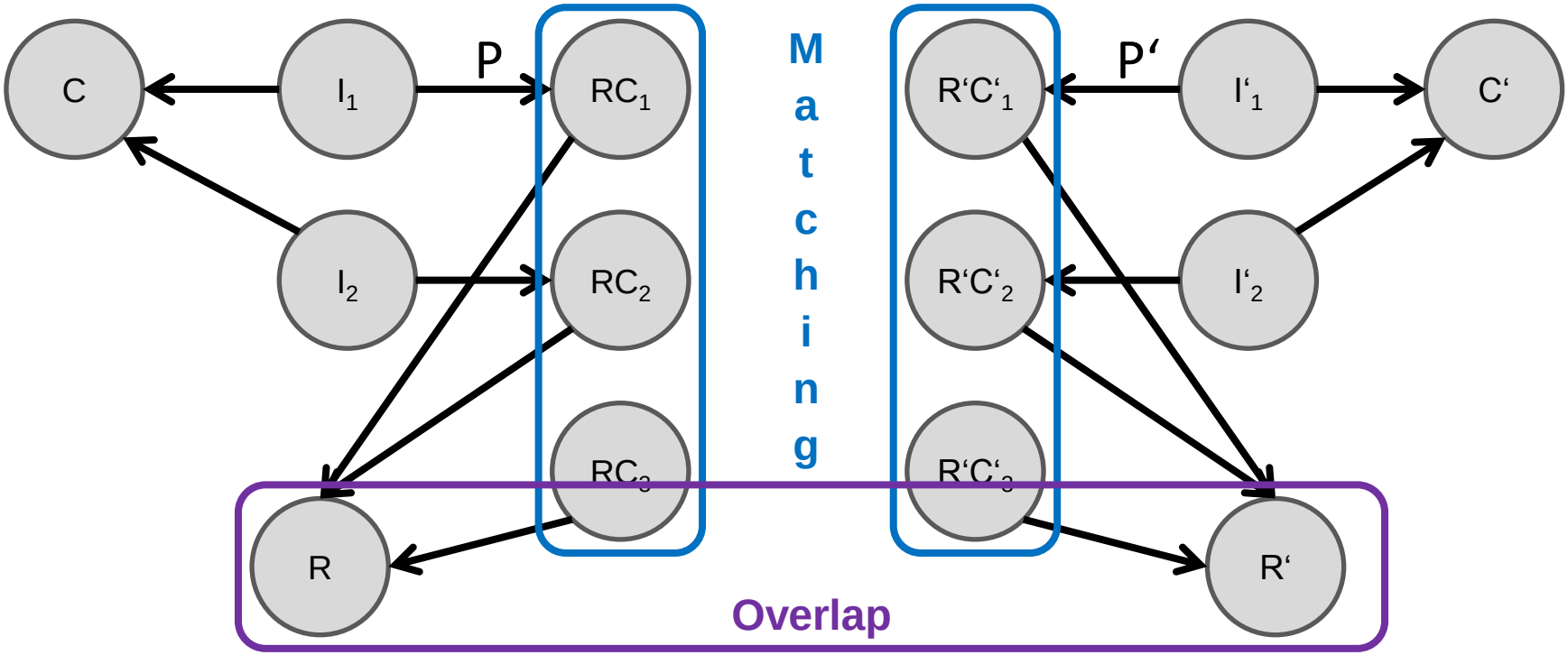


C = Class of an Ontology
P = Object Property
R = Imported Ontology

I = Instance of Class C
RC = Class of Imported Ontology

Approach

Utilizing the overlap between the imported ontologies to generate a correspondence between their linking object properties

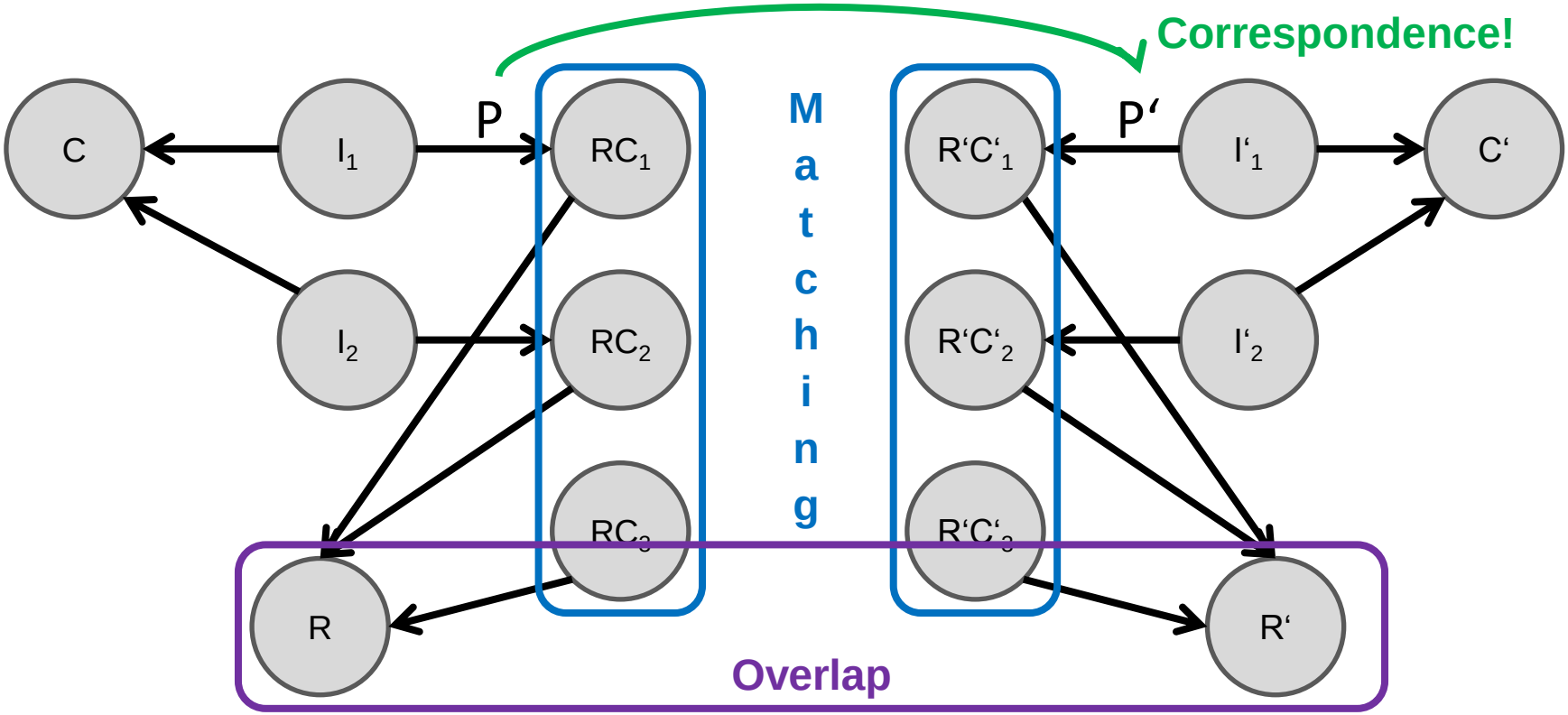


C = Class of an Ontology
P = Object Property
R = Imported Ontology

I = Instance of Class C
RC = Class of Imported Ontology

Approach

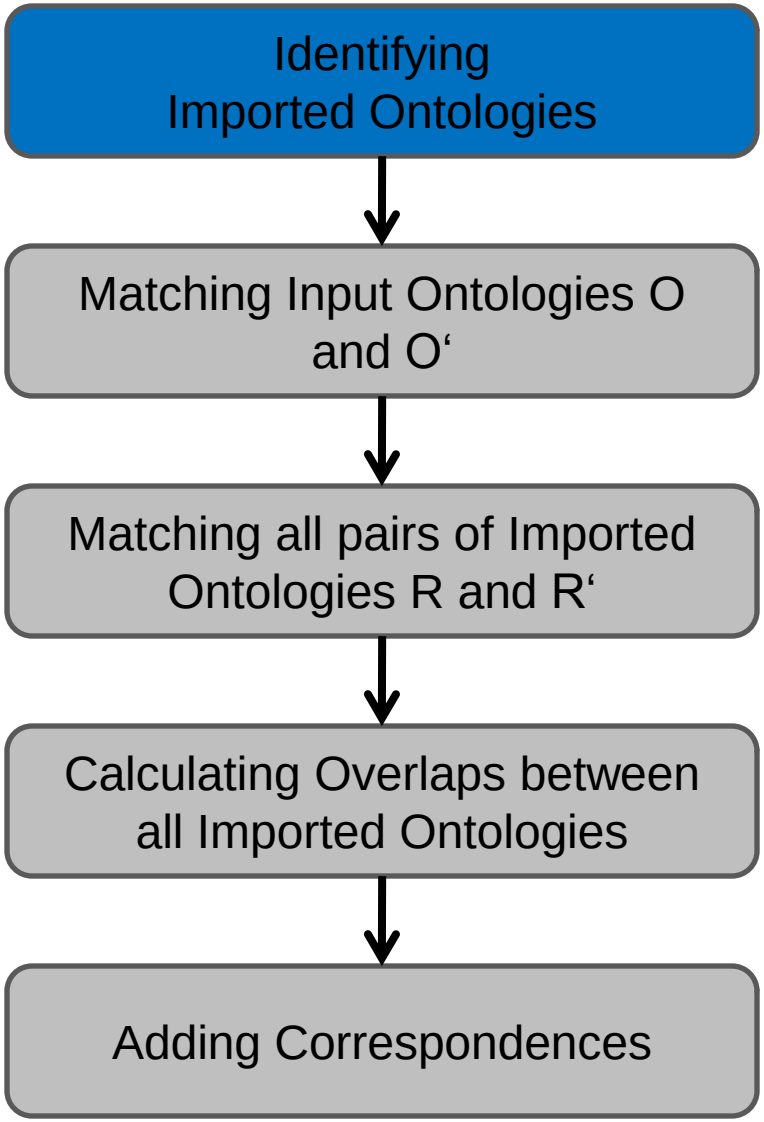
Utilizing the overlap between the imported ontologies to generate a correspondence between their linking object properties



C = Class of an Ontology
P = Object Property
R = Imported Ontology

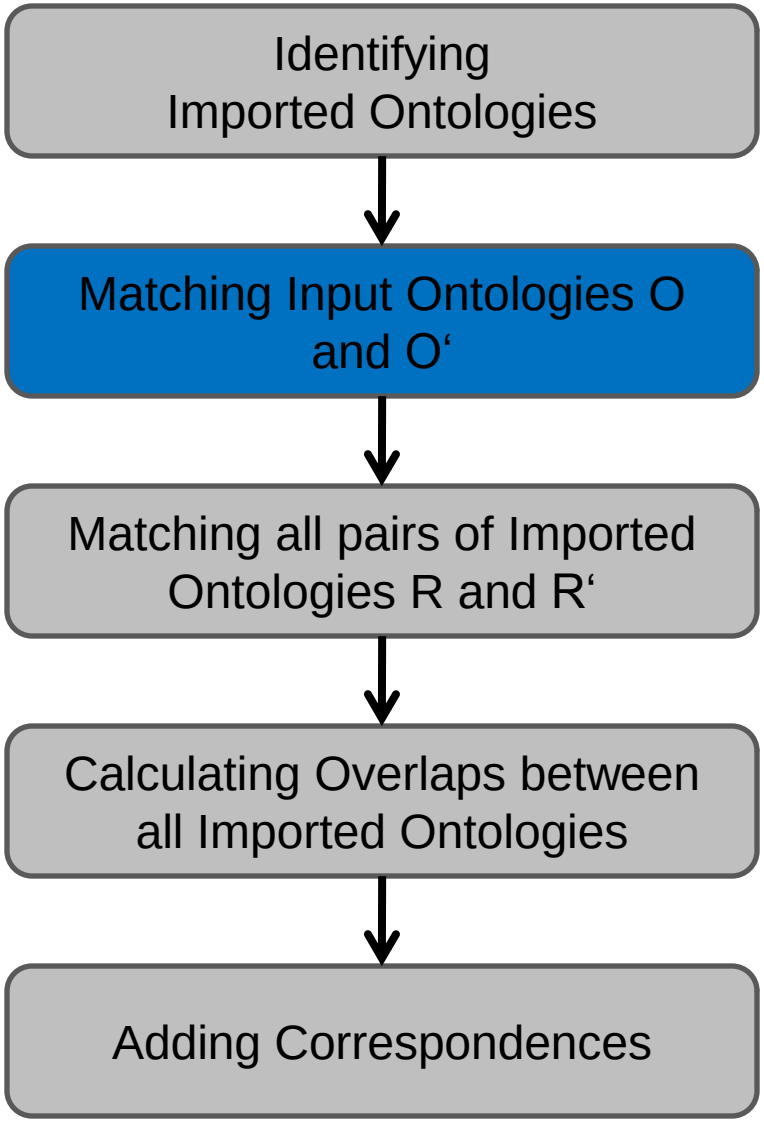
I = Instance of Class C
RC = Class of Imported Ontology

Approach



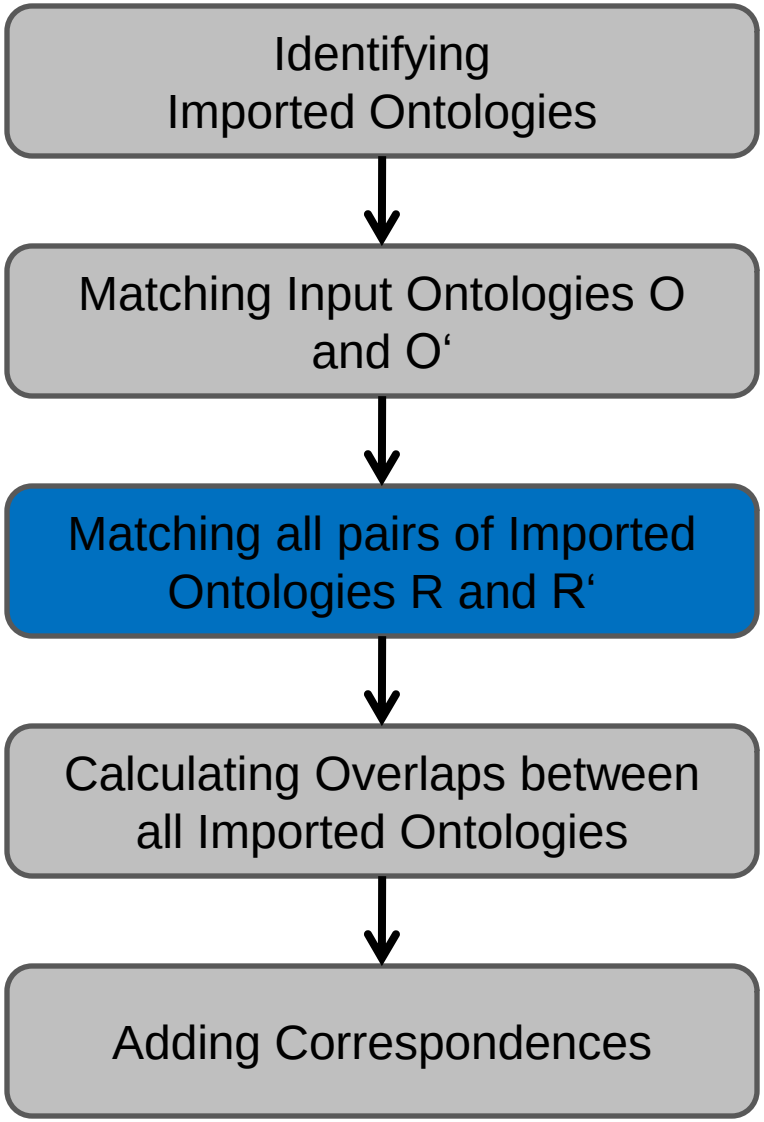
Grouping all linked instances (namespaces, URIs)

Approach



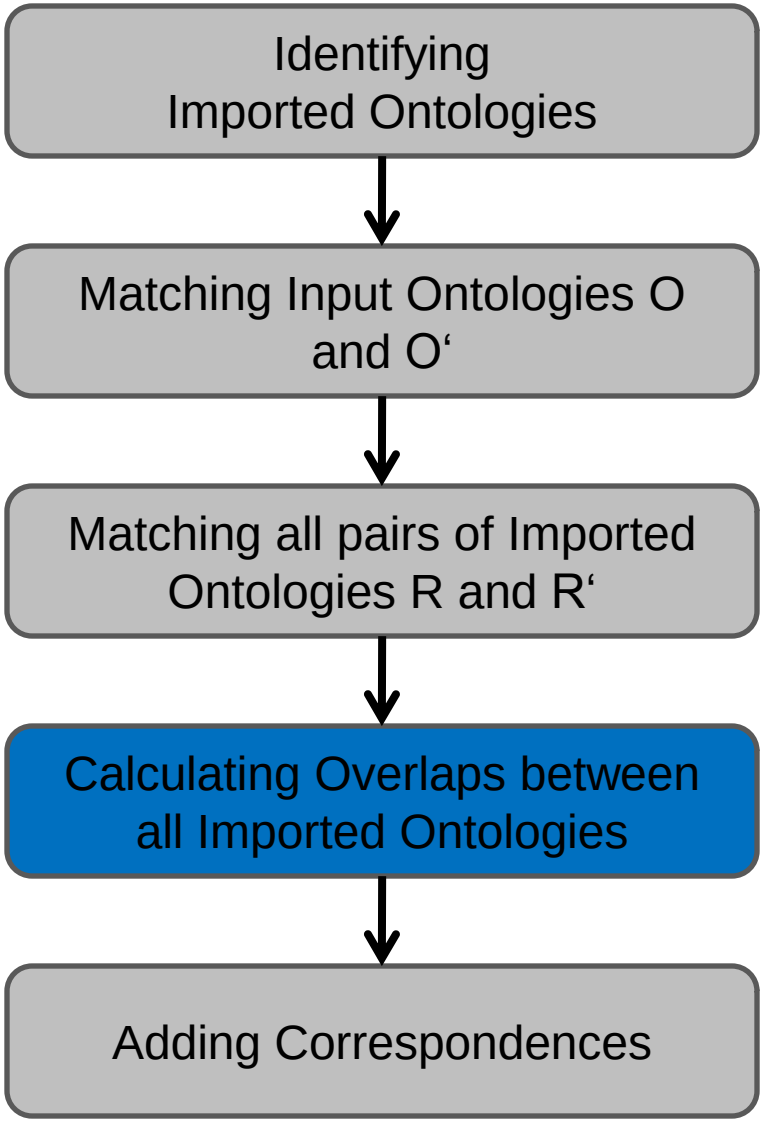
Instance-based matcher is used as black box

Approach



Detected correspondences are kept separately for each pair

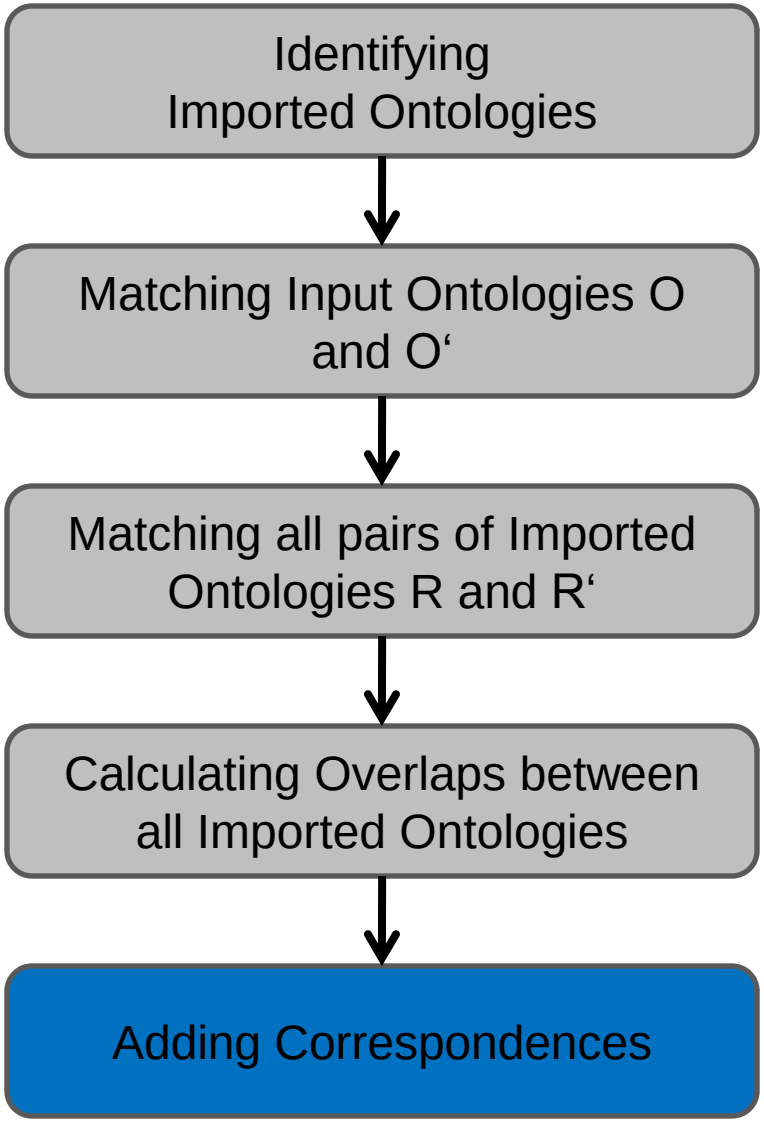
Approach



Pairwise overlap calculation based on correspondences using Jaccard Coefficient

$$JC = \frac{|R_n \bigcap R'_m|}{|R_n \bigcup R'_m|}$$

Approach



Overlap between imported ontologies indicates a correspondence between linking object properties

Evaluation Setup

- 2 Scenarios
  - Benchmark for Statistical Data (50 instances, 6 object properties, 10 variants)
  - Real-world data
    - Eurostat (16783 instances, 7 object properties)
    - OECD (5343 instances, 8 object properties)
- Reference Alignments have been created by 3 domain experts
- Measuring Precision, Recall, F-measure
- Tools
    - FALCON-AO
    - AgreementMaker

Benchmark for Statistical Data

- No suitable benchmark exists
- Benchmark oriented on established benchmarks (e.g. OAEI, IIMB)
  - One seed ontology
  - Randomly populated with instances
  - Generation of variations

Tests	Variation
001	Duplicate of seed ontology
010 – 011	Names of object properties
020 – 024	Labels, class names, URIs, namespaces of imported classes and ontologies
030 - 031	No overlap

Example class of the seed ontology

```
STATABOX:Entry11        a  STATTBOX:DataEntry,
                           owl:NamedIndividual;
          STATTBOX:date  „1981/08/02“^^xsd:integer;
          STATTBOX:obsValue  „886“^^xsd:integer;
          STATTBOX:ageGroup  ages:20-29;
          STATTBOX:gender  sex:Sex-M;
          STATTBOX:geo  countriesISO:DE
          STATTBOX:maritalStatus  concepts:CL_MAR_TOTAL;
          STATTBOX:occupation  indic_1:Occup_value3;
          STATTBOX:satisfaction  indic_2:Sat_value4.
```

```
ages:20-29 a owl:Class;
rdfs:label     „From 20 to 29 years“.

sex:Sex-M a owl:Class;
rdfs:label     „Male“.

countriesISO:DE a owl:Class;
rdfs:label     „Germany“.

concepts:CL_MAR_TOTAL a owl:Class;
rdfs:label     „Total“.

indic_1:Occup_value3 a owl:Class;
rdfs:label     „Unemployed“.

indic_2:Sat_value4 a owl:Class;
rdfs:label     „Very dissatisfied“.
```

Results

Approach	State of the Art (SotA)						Object Property Matching					
System	Agreement Maker			FALCON-AO			Agreement Maker			FALCON-AO		
Test	P	R	F	P	R	F	P	R	F	P	R	F
001	1	1	1	1	1	1	1	1	1	1	1	1
010-010	1	.45	.61	1	.34	.46	1	.89	.94	1	.78	.87
020-024	1	.42	.59	1	.29	.41	1	.85	.92	1	.67	.79
030-031	1	.45	.61	1	.34	.46	1	.45	.61	1	.34	.46
Real World Data	1	.40	.70	1	.40	.70	.83	1	.92	.45	1	.73

- Improvement on recall up to 2.5 times
- Loss in precision occurs when matching the imported ontologies

P = Precision
R = Recall
F = F-measure

Variants of Jaccard Coefficient

Jaccard Coefficient can get very unbalancy

- Considering the minimum # of all entities

$$JC_{min} = \frac{|R_n \bigcap R'_m|}{|min(|R_n|, |R'_m|)|}$$

- Considering the # of classes that are linked to

$$JC_{res} = \frac{|R_n \bigcap R'_m|}{|R_{n-Linked} \bigcup R'_{m-Linked}|}$$

- Considering the minimum # of classes that are linked to

$$JC_{min+res} = \frac{|R_n \bigcap R'_m|}{|min(|R_{n-Linked}|, |R'_{m-Linked}|)|}$$

Calculated Overlap Scores

Can the loss in precision be tackled?

Detected Correspondences (AM)	JC	JC_{min}	JC_{res}	$JC_{min+res}$	SotA
Correct Correspondences					
geo = LOCATION	.002	1	.688	1	0
indic_bt = VAR	.132	.909	1	1	0
nace_r2 = ISIC3	.006	.979	.959	1	0
obs_status = OBS_STATUS	.750	1	X	X	.969
timeformat = TIME_FORMAT	.571	1	1	1	.872
False Positives					
geo = ISIC3	.004	.354	.369	1	0

JC = Jaccard Coefficient
JC_{res} = Resource Jaccard
SotA = State of the Art
AM = AgreementMaker

JC_{min} = Minimum Jaccard
$JC_{min+res}$ = Minimum+Resource Jaccard
X = no calculation possible (Divide by Zero)

Calculated Overlap Scores

Detected Correspondences (FALCON-AO)	JC	JC_{min}	JC_{res}	$JC_{min+res}$	SotA
Correct Correspondences					
geo = LOCATION	.002	.909	.588	.909	0
indic_bt = VAR	.012	.090	.083	.333	0
nace_r2 = ISIC3	.007	.188	.103	.191	0
obs_status = OBS_STATUS	.647	.917	X	X	1
timeformat = TIME_FORMAT	.571	1	1	1	1
False Positives					
geo = ISIC3	.004	.354	.340	1	0
nace_r2 = VAR	.001	.090	.017	.100	0
nace_r2 = OBS_STATUS	.012	.938	.306	.306	0
freq = VAR	.053	.111	.1	.100	0
freq = OBS_STATUS	.389	.778	X	X	0
freq = TIME_FOTRMAT	.300	.750	1	1	0

Conclusion

- Instance-based object property matching approach
- Utilizing overlap between imported ontologies
  - Including classes that are not linked to
- Improvement on recall up to 2.5 times
  - Loss in precision may occur
  - Choice of similarity measure seems to be less relevant than choice of matching system
  - Larger evaluation may bring more insight
- Any instance-based matching system can be applied

Future Work

- Object Property Matching
  - Larger evaluation
  - Experimenting with additional similarity measures
- Benchmark for Statistical Data
  - More object properties including diverse classifications and code lists
  - More variants
  - Applicable for other purposes

Thank You!

Contact:
Ben Zapilko
ben.zapilko@gesis.org
GESIS – Leibniz Institute for the Social Sciences,
Cologne, Germany

Benchmark for Statistical Data:
http://code.google.com/p/matching-statistics/