

# Linking and Building Ontologies of Linked Data

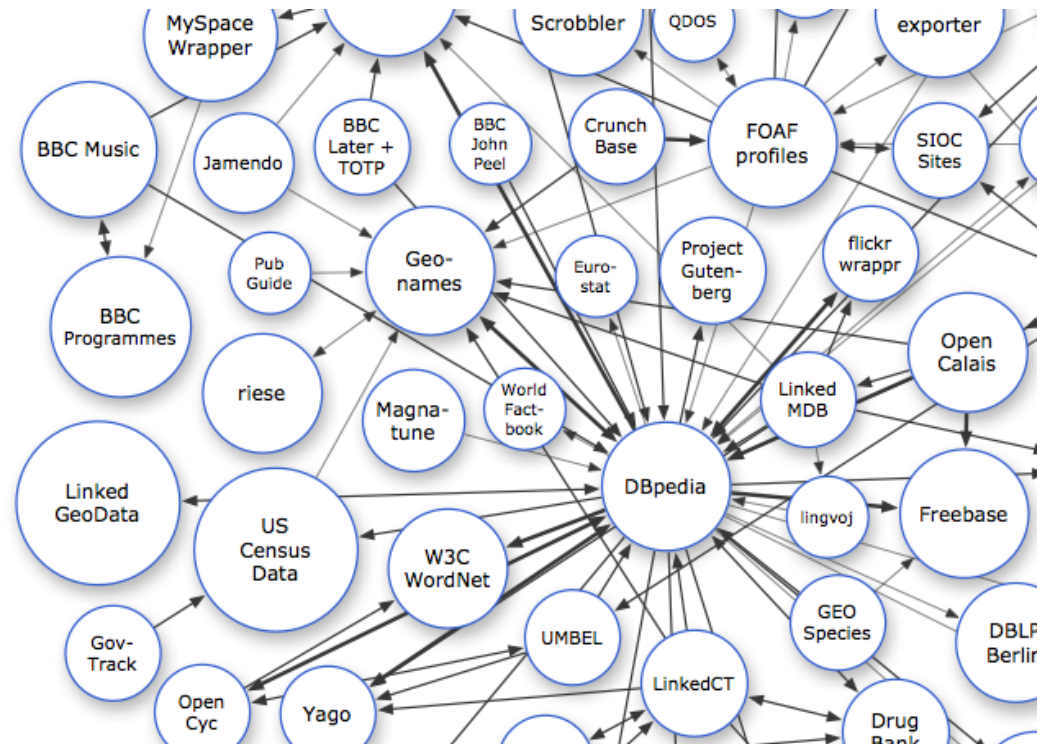
**Rahul Parundekar, Craig A. Knoblock and Jose-Luis Ambite**

{parundek,knoblock,ambite}@isi.edu

**University of Southern California**

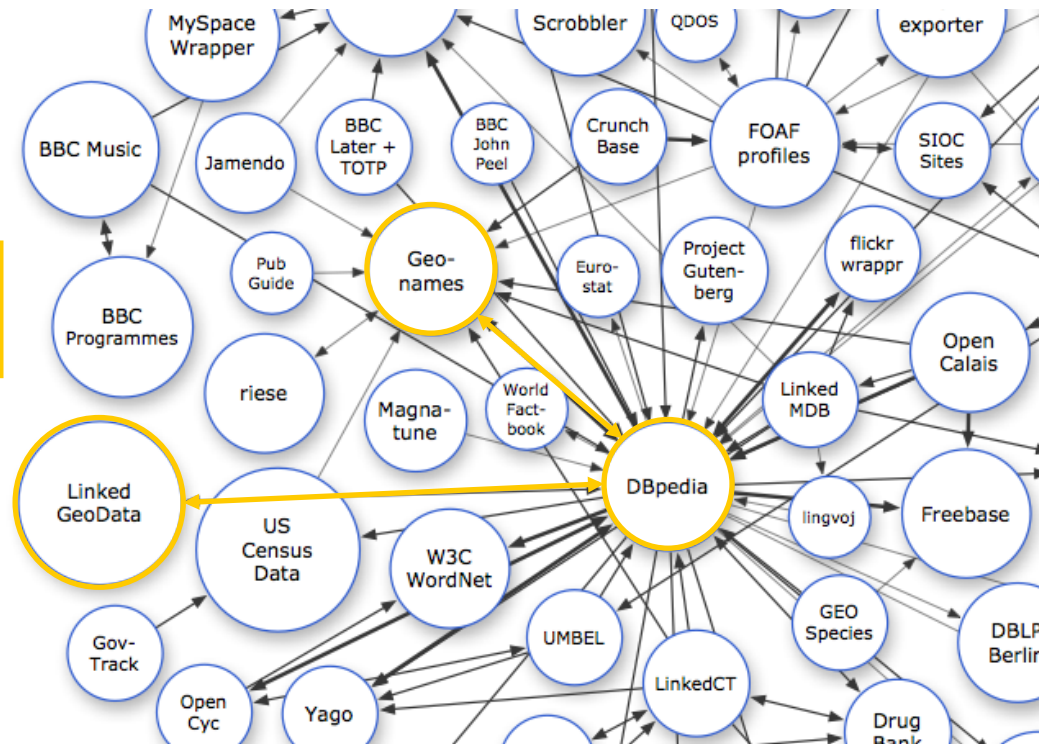
# Web of Linked Data

- Vast collection of interlinked information
- Different sources with different schemas

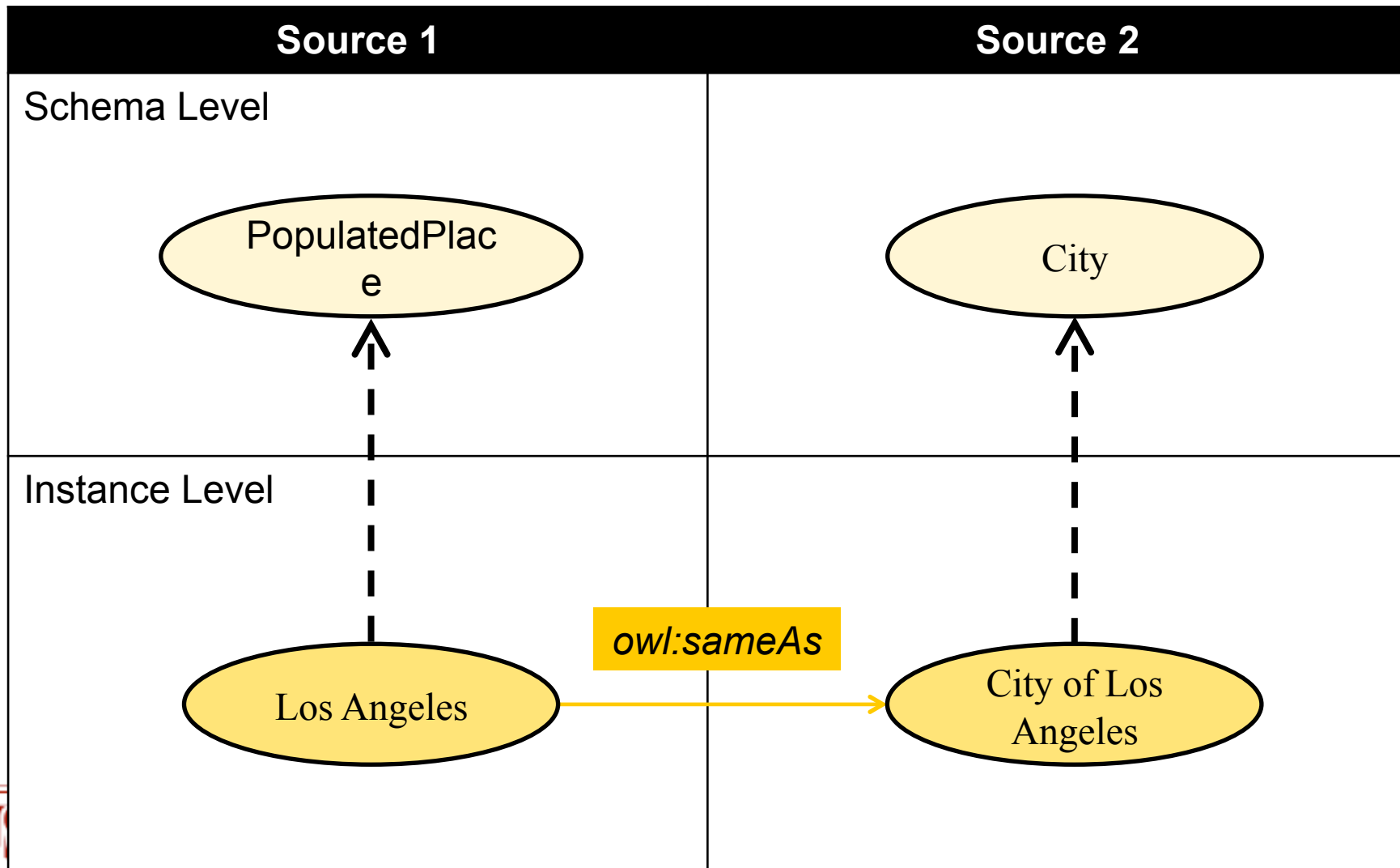


- Interlinked instances in the various domains
- Equivalent instances linked with *owl:sameAs*

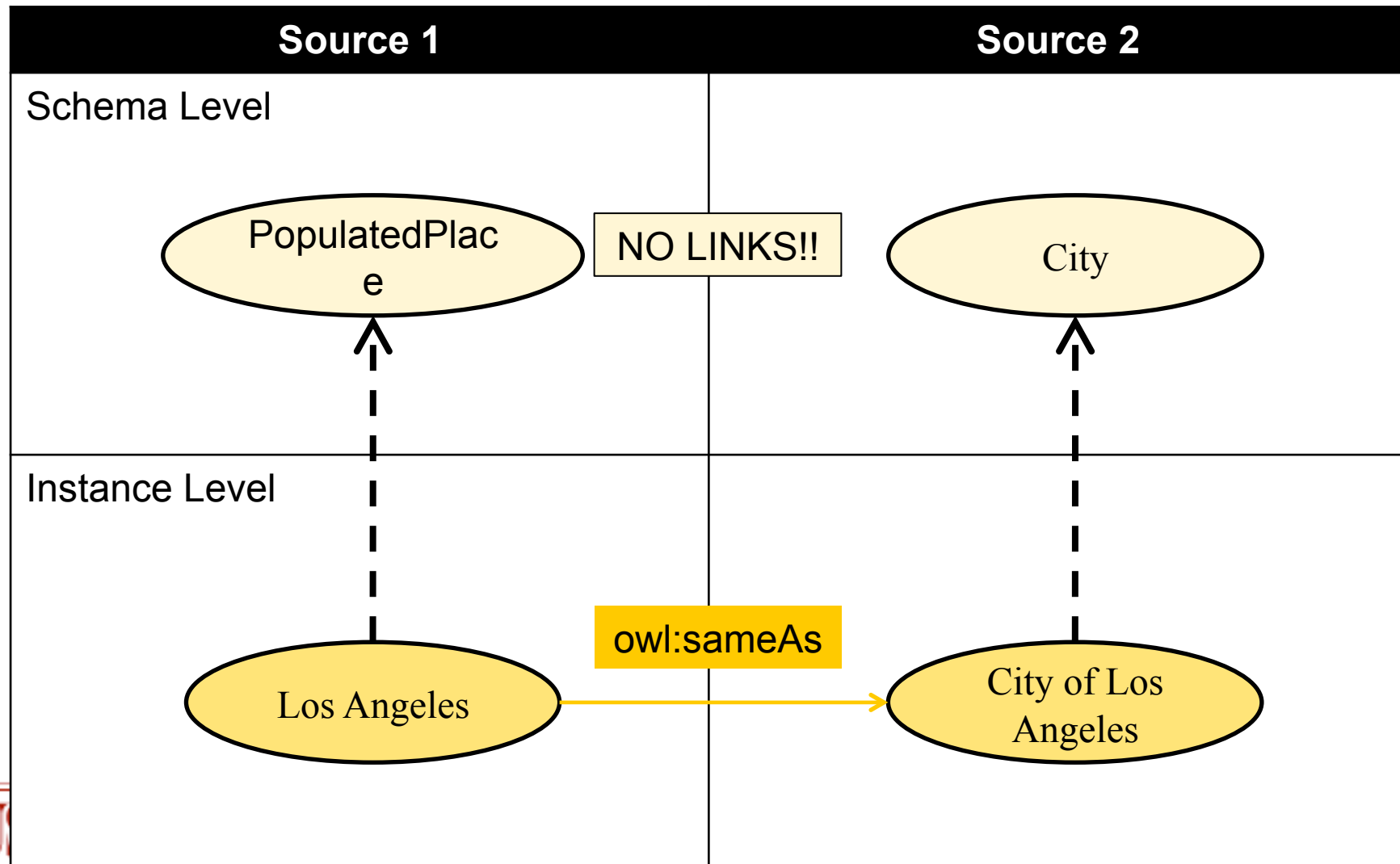
Geospatial  
Domain



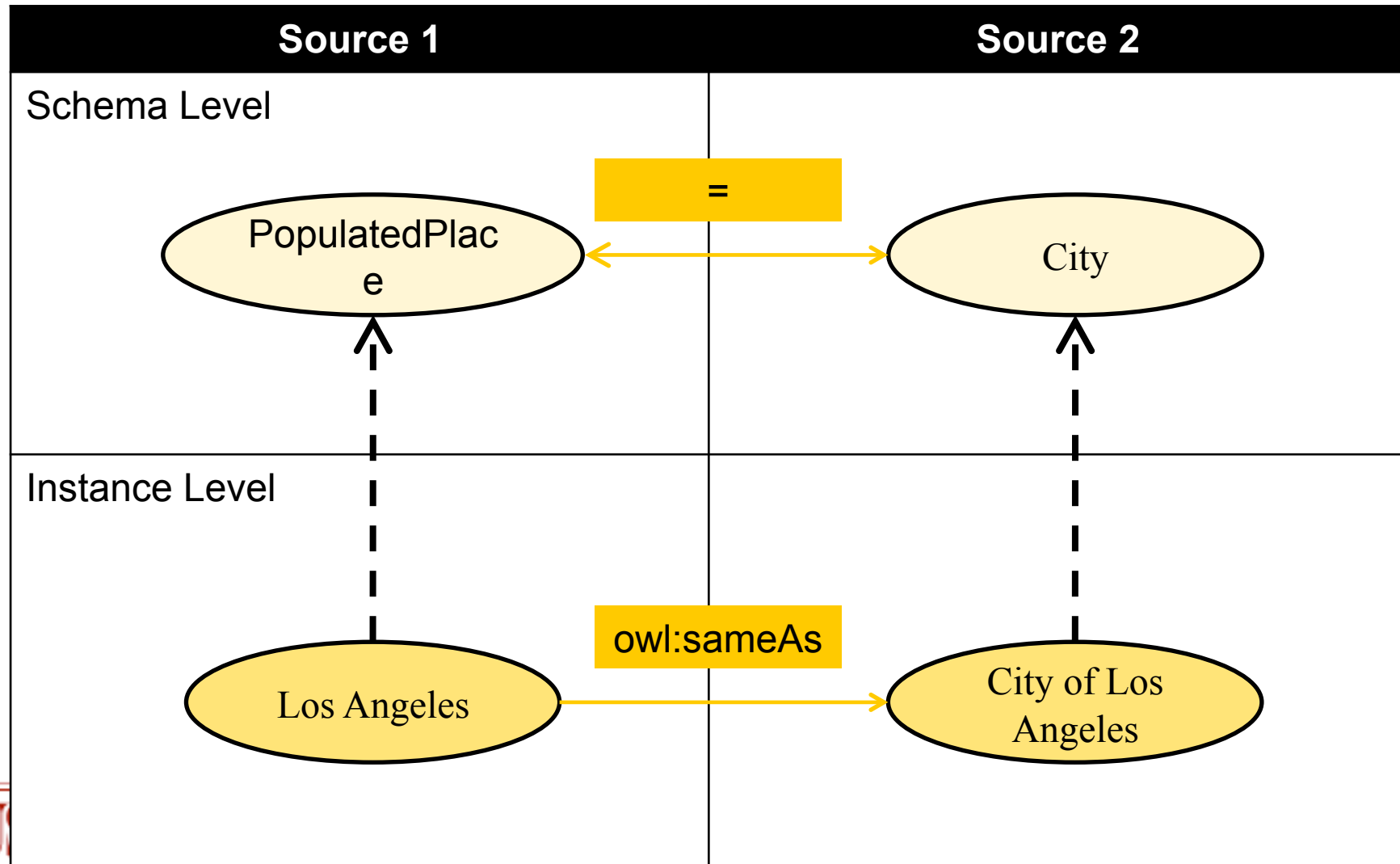
# Interlinked Instances



# Disjoint Schemas



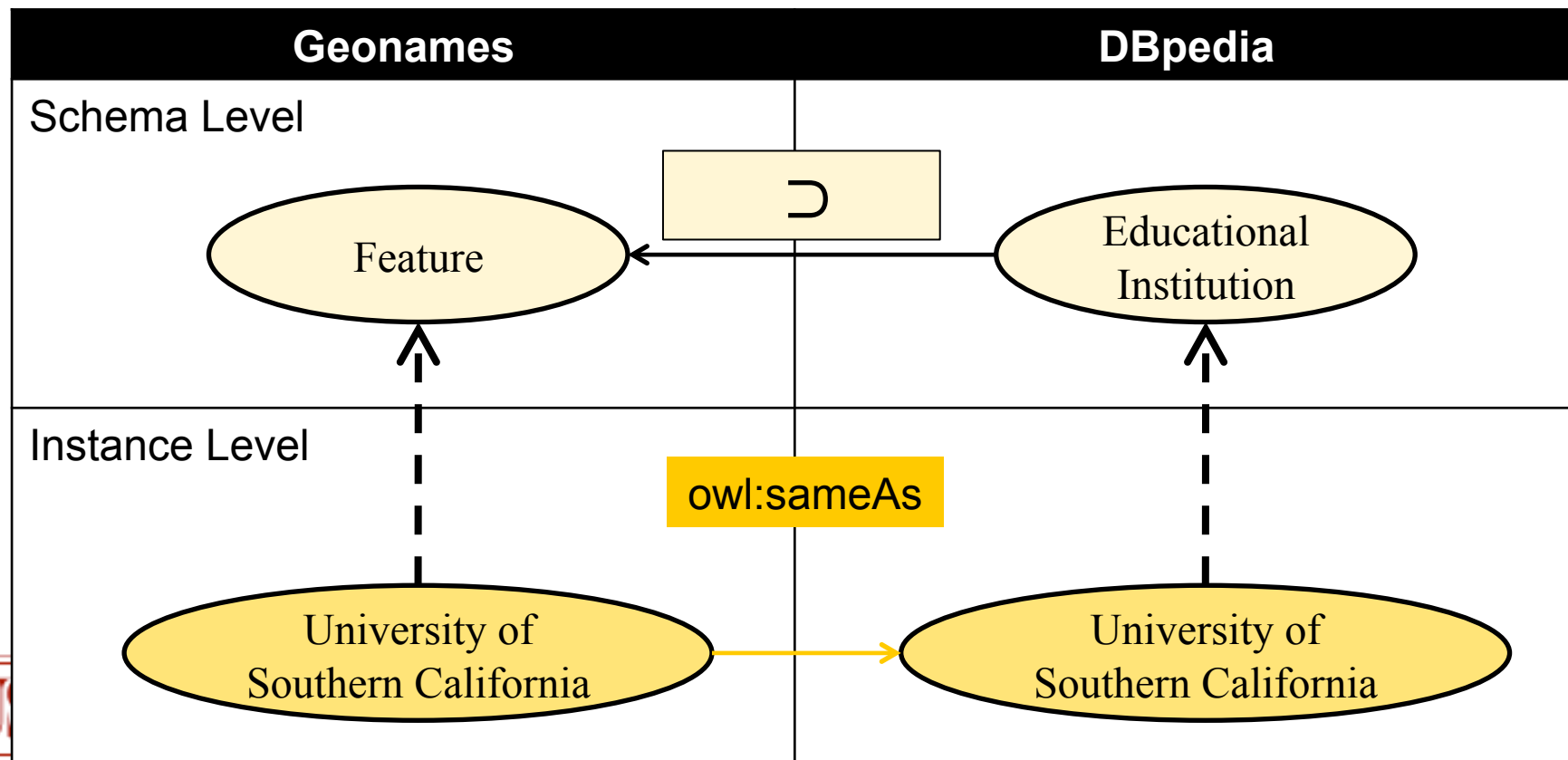
# Objective 1: Find Schema Alignments



- **Ontologies can be highly specialized**
  - e.g. DBpedia has classes for *Educational Institutions*, *Bridges*, *Airports*, etc.
- **But some can be rudimentary**
  - e.g. in Geonames all instances only belong to a single class – ‘Feature’
  - Derived from RDBMS schemas from which Linked Data was generated

# Traditional Alignments

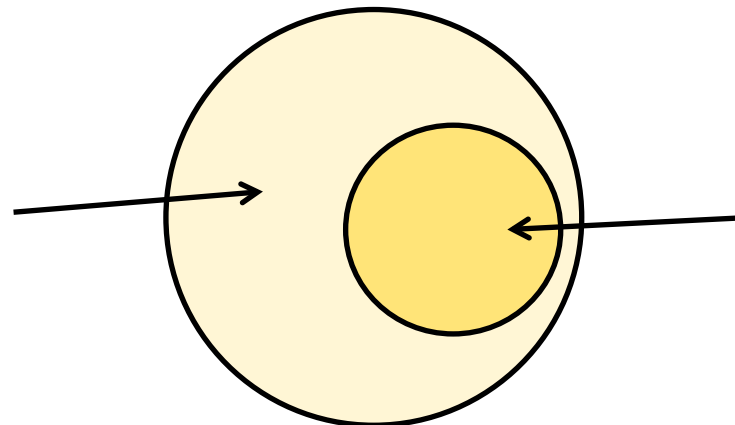
- There might not exist exact equivalences between classes in two sources
- Only subset relations possible





- A specialized class can be created by restricting the value of one or more properties
- The following Venn diagram explains a restriction class in Geonames with a restriction on the value of the *featureCode* property as 'S.SCH'

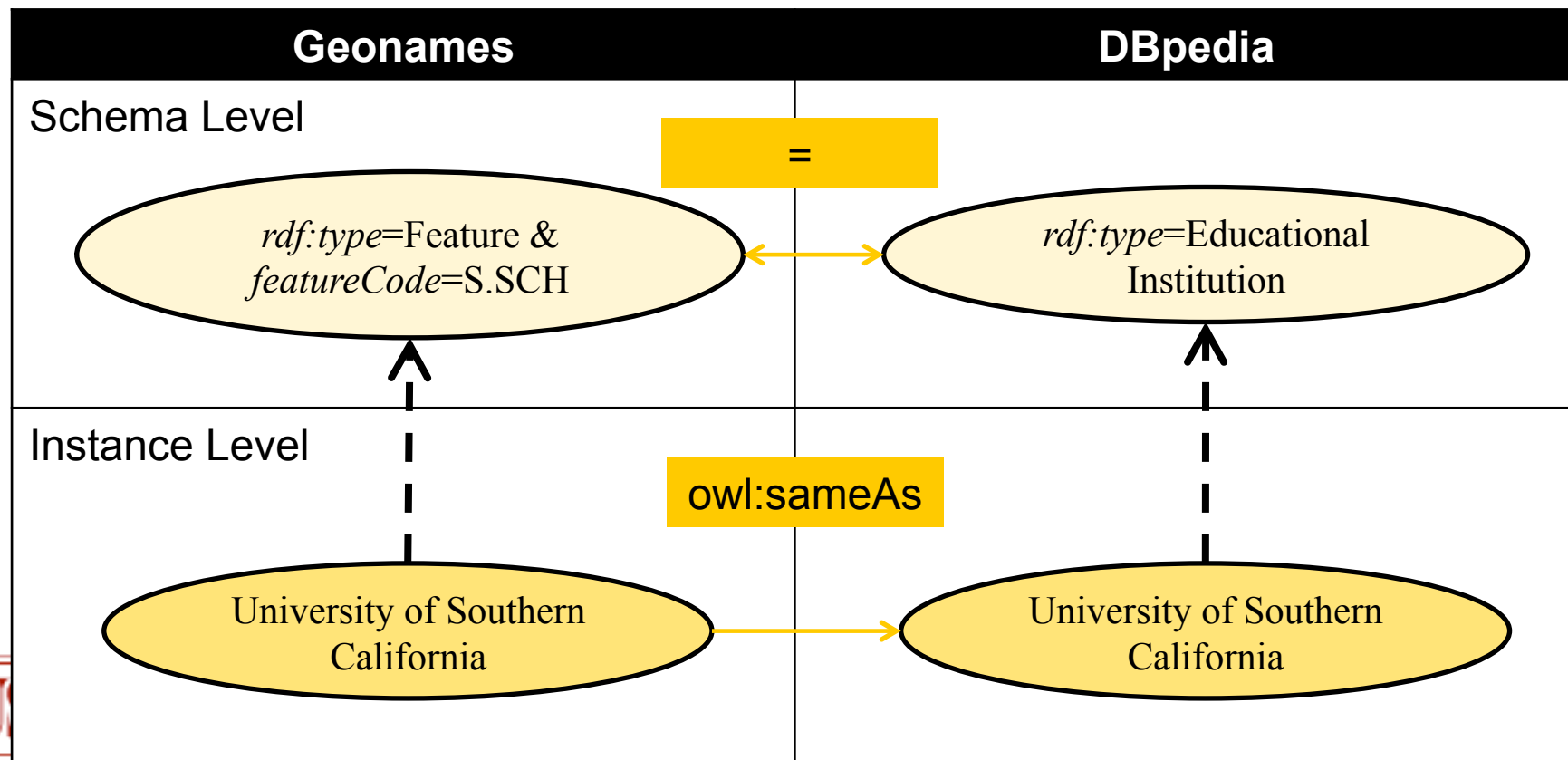
Set of all instances in  
Original Class -  
*rdf:type=Feature*



Set of all instances in  
Restricted Class -  
*rdf:type=Feature &  
featureCode=S.SCH*

## Objective 2: Find Alignments Between Restriction Classes

- Find and model specialized descriptions of classes

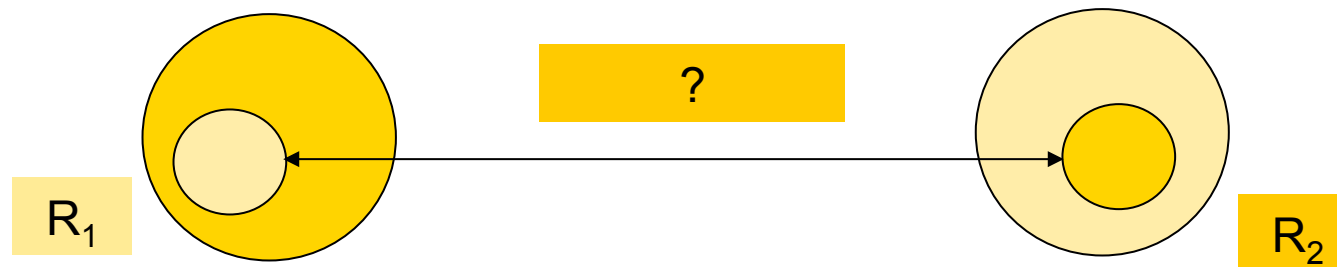


- **Geospatial**
  - Dbpedia
  - LinkedGeoData
  - Geonames
- **Zoology**
  - Geospecies
  - Dbpedia
- **Genetics (Bio2RDF)**
  - GeneID
  - MGI

- Aligning Restriction Classes

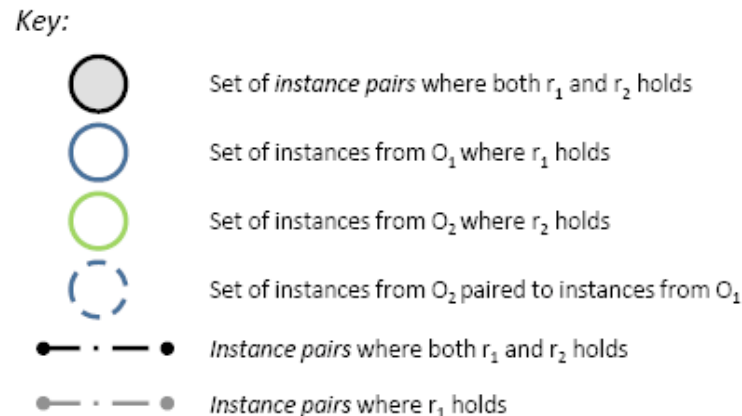
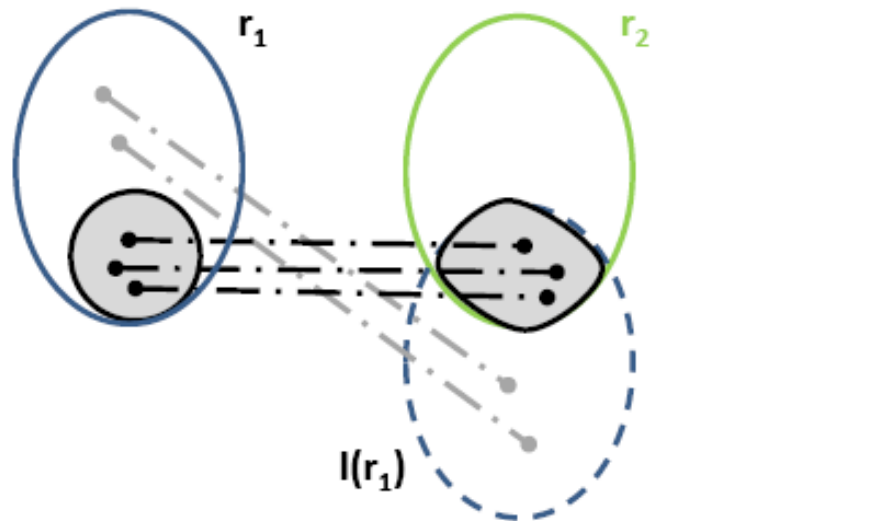







- Aligning Restriction Classes



- Find relation between the two restriction classes
  - Equivalent
  - Subset

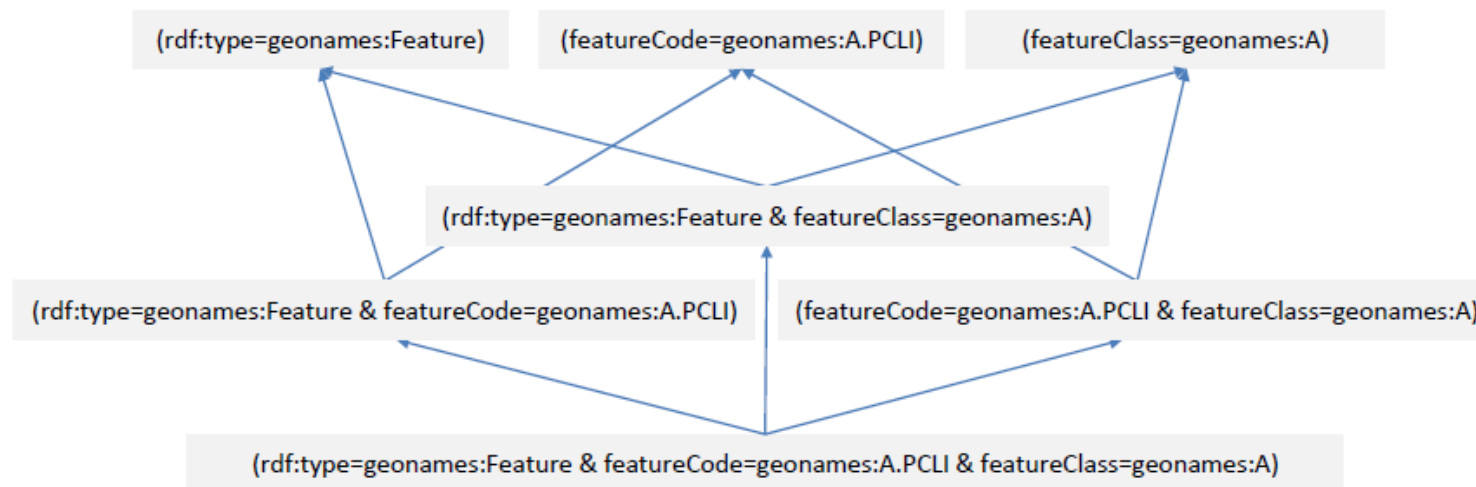
# Extensional Approach to Ontology Alignment



Set Representation	Relation
	Disjoint
	$r_1 \subset r_2$
	$r_2 \subset r_1$
	$r_1 = r_2$
	Not enough support

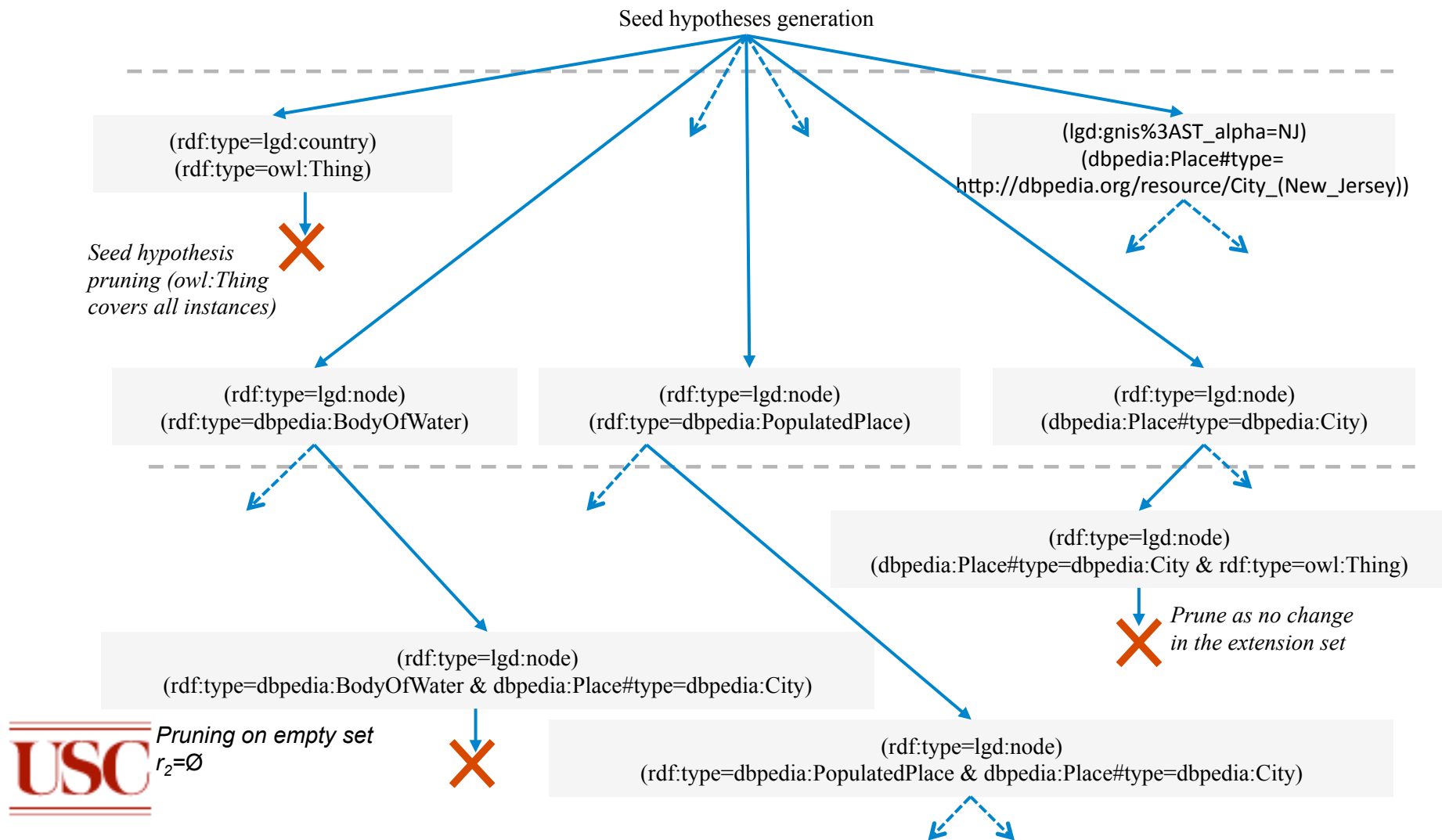
# Lattice of Restriction Classes

- Instances belonging to a restriction class also belong to parent restriction class
  - e.g. restrictions from Geonames below



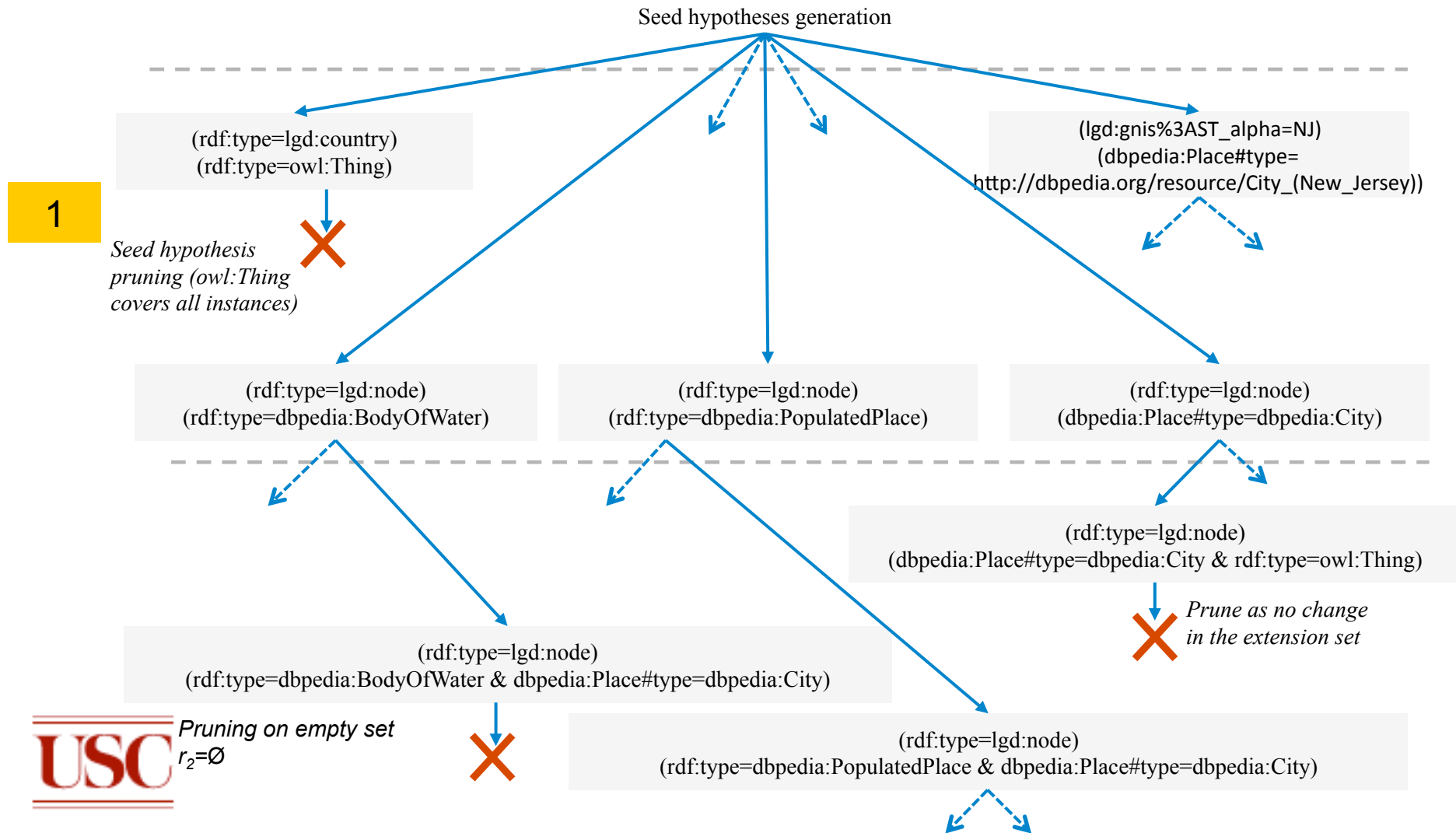
- This also results in a hierarchy in the alignments, which our algorithm exploits

# Exploration of Hypotheses Search Space (LinkedGeoData with DBpedia)

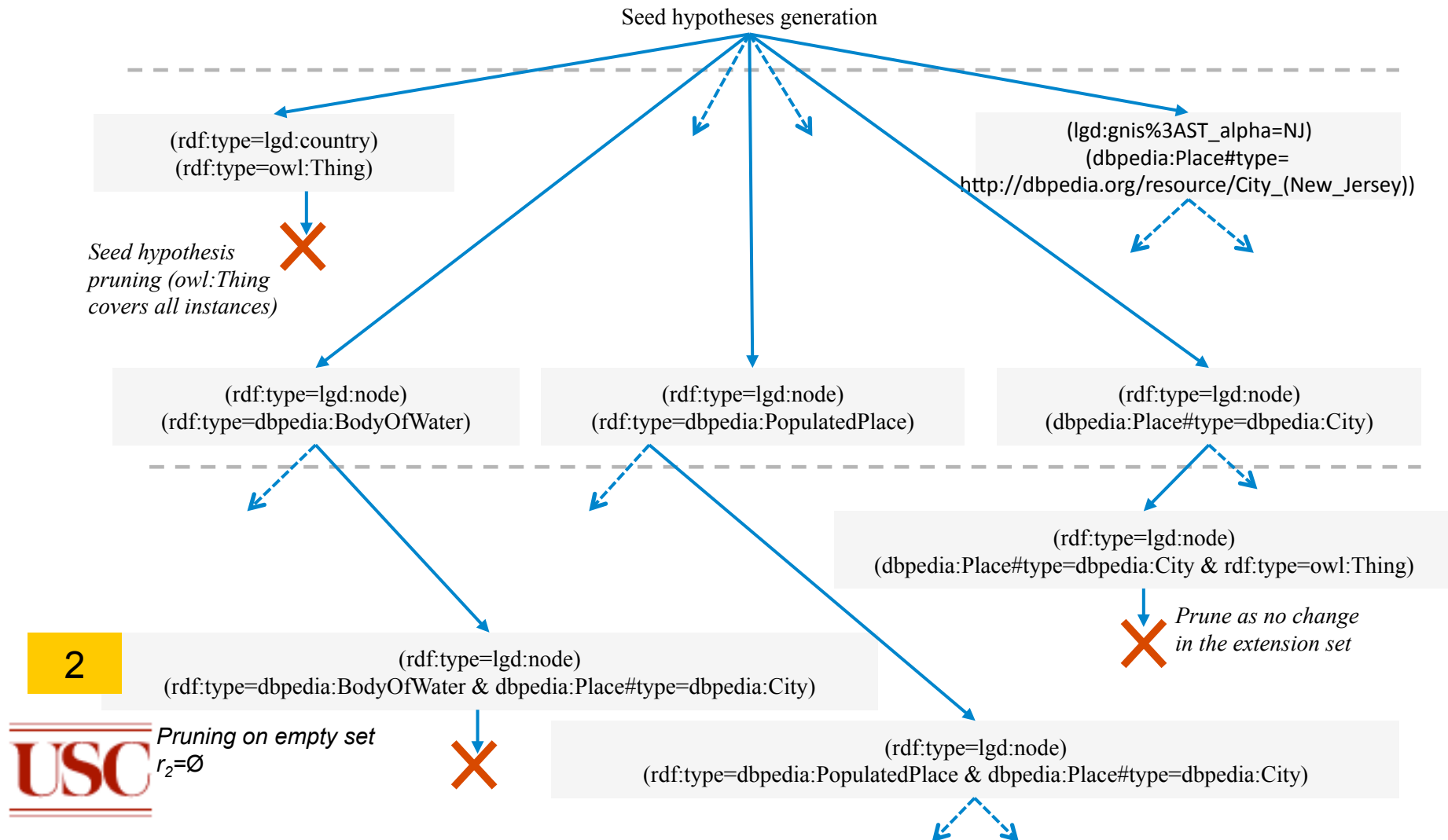




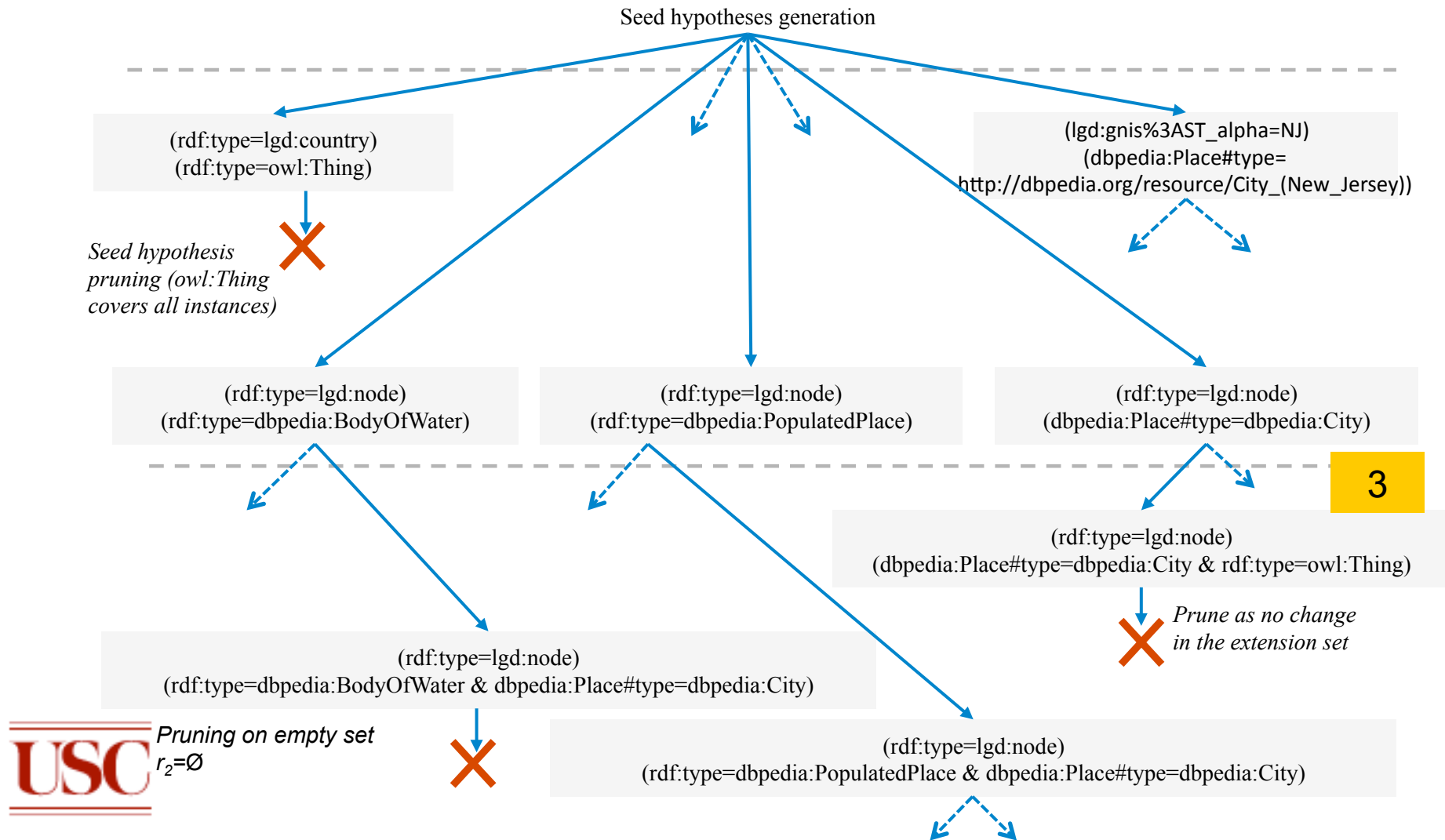
# 1. Prune seed hypothesis if either restriction covers all instances in that source



## 2. Number of instance pairs supporting hypothesis must be above a threshold

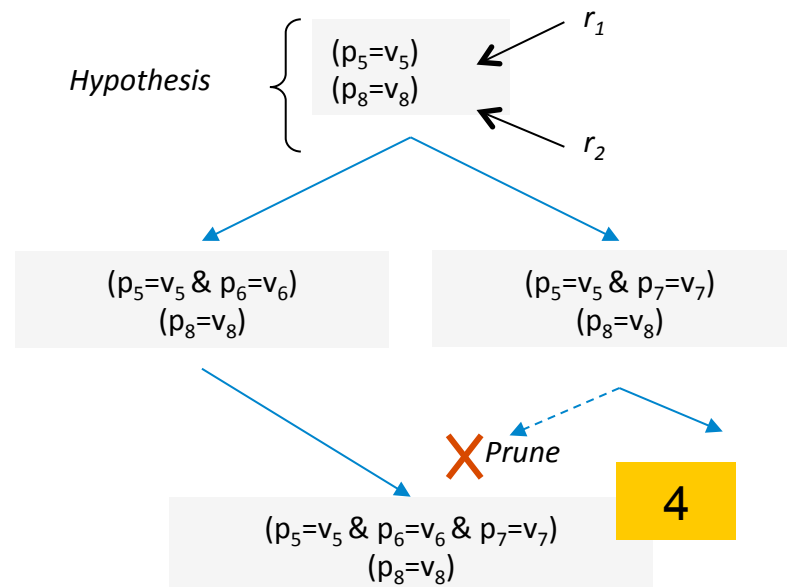


# 3. Prune if the added constraint does not change the extension

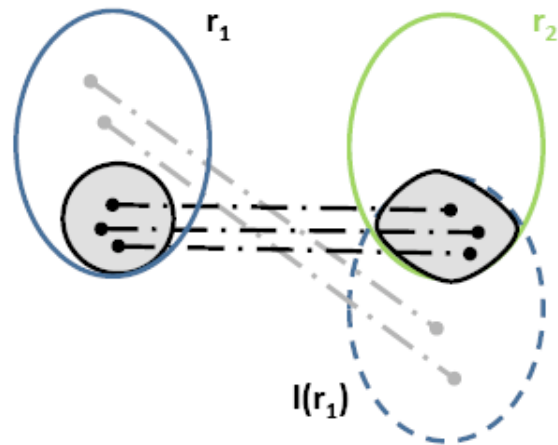


## 4. Lexicographic ordering

Lexicographic ordering provides a systematic search by pruning hypotheses with reverse order



# Relaxed Scoring

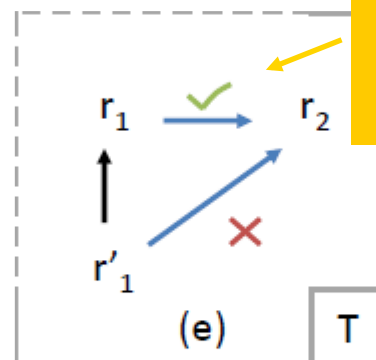


Set Representation	Relation	$P = \frac{ I(r_1) \cap r_2 }{ r_2 }$	$R = \frac{ I(r_1) \cap r_2 }{ r_1 }$	$P'$	$R'$
	Disjoint	= 0	= 0	$\leq 0.01$	$\leq 0.01$
	$r_1 \subset r_2$	< 1	= 1	$> 0.01$	$\geq 0.90$
	$r_2 \subset r_1$	= 1	< 1	$\geq 0.90$	$> 0.01$
	$r_1 = r_2$	= 1	= 1	$\geq 0.90$	$\geq 0.90$
	Not enough support	$0 < P < 1$	$0 < R < 1$	$0.01 < P' < 0.90$	$0.01 < R' < 0.90$

- Compensates for missing, inconsistent in the data

# Post-processing: Removing Implied Alignments

GEONAMES restriction	DBPEDIA restriction
geonames:featureCode=geonames:S.SCH	rdf:type=dbpedia:EducationalInstitution
geonames:featureCode=geonames:S.SCH & geonames:inCountry=geonames:US	rdf:type=dbpedia:EducationalInstitution



Keep the simpler definition  
&  
Remove the implied definition

Key:

$r_i \rightarrow r_j$  : Subset relations ( $r_i \subset r_j$ )  
found by the algorithm.

$r_i \cdots \rightarrow r_j$  : Implied subset relations.

$r'_i \rightarrow r_j$  : Subset relation by construction.

T: Transitivity in subset relations.

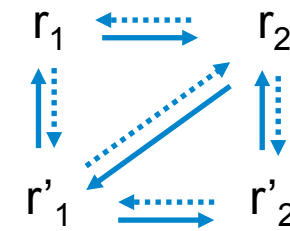
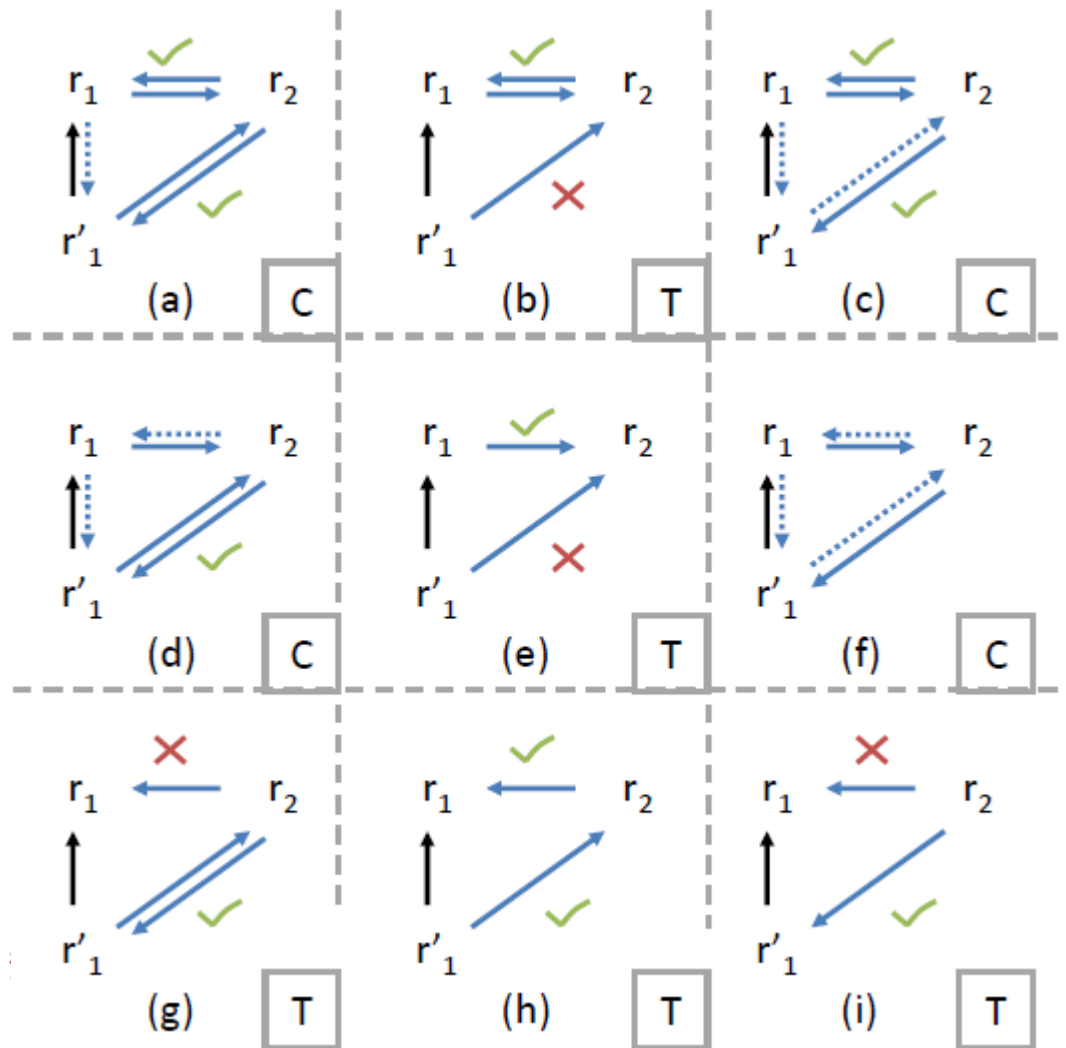
One relation can be eliminated.

C: Cycle in subset relations. Hence,  
all classes are equivalent.

✗ : Relation eliminated by the rule.

✓ : Relation retained by the rule.

# Removing Implied Alignments



Cascading

Key:

$r_i \rightarrow r_j$  : Subset relations ( $r_i \subset r_j$ ) found by the algorithm.

$r_i \cdots \rightarrow r_j$  : Implied subset relations.

$r'_i \rightarrow r_j$  : Subset relation by construction.

T: Transitivity in subset relations.

One relation can be eliminated.

C: Cycle in subset relations. Hence, all classes are equivalent.

X : Relation eliminated by the rule.

✓ : Relation retained by the rule.

#	<b>LINKEDGEODATA restriction</b>	<b>DBPEDIA restriction</b>	<b>Relation</b>
1	rdf:type=lgd:node	rdf:type=owl:Thing	$r_1 = r_2$
2	rdf:type=lgd:aerodrome	rdf:type=dbpedia:Airport	$r_1 = r_2$
3	rdf:type=lgd:island	rdf:type=dbpedia:Island	$r_1 = r_2$
4	lgd:gnis_%3AST_alpha=NJ	dbpedia:Place#type= <a href="http://dbpedia.org/resource/City_(New_Jersey)">http://dbpedia.org/resource/City_(New_Jersey)</a>	$r_1 = r_2$
5	rdf:type=lgd:village	rdf:type=dbpedia:PopulatedPlace	$r_1 \subset r_2$
#	<b>GEONAMES restriction</b>	<b>DBPEDIA restriction</b>	<b>Relation</b>
6	geonames:featureClass=geonames:P	rdf:type=dbpedia:PopulatedPlace	$r_1 = r_2$
7	geonames:featureClass=geonames:H	rdf:type=dbpedia:BodyOfWater	$r_1 = r_2$
8	geonames:parentFeature= <a href="http://sws.geonames.org/3174618/">http://sws.geonames.org/3174618/</a>	dbpedia:City_region= <a href="http://dbpedia.org/resource/Lombardy">http://dbpedia.org/resource/Lombardy</a>	$r_1 = r_2$
9	geonames:featureCode=geonames:S.SCH	rdf:type=dbpedia:EducationalInstitution	$r_1 = r_2$
10	geonames:featureCode=geonames:S.SCH & geonames:inCountry=geonames:US	rdf:type=dbpedia:EducationalInstitution	$r_1 = r_2$
11	geonames:featureCode=geonames:T.MT	rdf:type=dbpedia:Mountain	$r_1 \subset r_2$



#	GEOSPECIES restriction	DBPEDIA restriction	Relation
12	geospecies:inKingdom=http://lod.geospecies.org/kingdoms/Aa	rdf:type=dbpedia:Animal	$r_1 = r_2$
13	geospecies:hasOrderName=Lepidoptera	dbpedia:order=http://dbpedia.org/resource/Lepidoptera	$r_1 = r_2$
14	geospecies:hasOrderName=Lepidoptera	dbpedia:kingdom=http://dbpedia.org/resource/Animal & dbpedia:order=http://dbpedia.org/resource/Lepidoptera	$r_1 = r_2$
15	geospecies:hasGenusName=Falco	dbpedia:genus=http://dbpedia.org/resource/Falcon	$r_1 = r_2$
16	geospecies:hasOrderName=Primates	dbpedia:order=http://dbpedia.org/resource/Primates	$r_2 \subset r_1$

#	GEOSPECIES restriction	GEOSPECIES restriction	Relation
20	geospecies:hasKingdomName=Animalia	geospecies:inKingdom=http://lod.geospecies.org/kingdoms/Aa	$r_1 = r_2$
21	geospecies:hasClassName=Insecta	geospecies:inClass= http://lod.geospecies.org/bioclasses/aQado	$r_1 \subset r_2$
22	geospecies:inFamily= http://lod.geospecies.org/families/amTJ9	geospecies:hasSubfamilyName=Sigmodontinae	$r_2 \subset r_1$

#	<i>MGI restriction</i>	<i>GENEID restriction</i>	<b>Relation</b>
17	bio2rdf:subType=Pseudogene	bio2rdf:subType=pseudo	$r_1 = r_2$
18	bio2rdf:subType=Pseudogene & mgi:genomeStart=17	geneid:chromosome=17 & bio2rdf:subType=pseudo	$r_1 = r_2$
19	bio2rdf:chromosomePosition=-1.00 & mgi:genomeStart=4	geneid:chromosome=4 & bio2rdf:subType=pseudo	$r_2 \subset r_1$

## Results: Alignments Found

- Equivalences, Subset alignments before and after removing implied alignments

Source 1 ( $O_1$ )	Source 2 ( $O_2$ )	$\#(r_1 = r_2)$ total	$\#(r_1 = r_2)$ best matches	$\#(r_1 \subset r_2)$ before	$\#(r_1 \subset r_2)$ after	$\#(r_2 \subset r_1)$ before	$\#(r_2 \subset r_1)$ after
LinkedGeoData	DBpedia	158	152	2528	1837	1804	1627
Geonames	DBpedia	31	19	809	400	1384	1247
Geospecies	DBpedia	509	420	9112	2294	6098	4455
MGI	GeneID	10	9	2031	1869	3594	2070
Geospecies	Geospecies	94	88	1550	1201	-	-

Linking and Building Ontologies of Linked Data

Rahul Parundekar, Craig A. Knoblock and José-Luis Ambite  
University of Southern California,  
Information Sciences Institute  
4676 Admiralty Way, Marina del Rey, CA 90292  
{parundek,knoblock,ambite@isi.edu}

This page provides the dataset used in the paper on [Linking and Building Ontologies of Linked Data](#).

- This dataset is organized as follows:
  - The 5 source pairs, discussed in the paper, each have a compressed file containing the *instance pairs* input to the algorithm and the *alignments* generated by the algorithm from it.
  - Each compressed file contains 3 *Comma separated variables (\*.csv)* files containing the *instance pairs*, *alignments before and after post processing*. There are two data sources that are being aligned *source1* and *source2* (which may be the same), in each of these files.
  - The *instancepairs\_source1\_source2* file:
    - Each row in the file represents an *instance pair* which is a join of the flattened property-value pairs of the instances from each source (see the paper), where the join is on the property that asserts instance equivalence.
    - The first line in the csv file lists the properties in that source. It is of the form  $uri_1, property1\_source1, \dots, propertyn\_source1, uri_2, property1\_source2, \dots, propertym\_source2$ .
    - The other lines in the file contain a URI of the instance from the first source, the values of the properties under each of its columns (? if no value exists) and a similar vector for the URI of the second source.
    - Preprocessing has already been performed on these instances.
  - The *alignments\_source1\_source2* file:
    - Each row in the file represents an alignment generated by the algorithm along with the stats that support that hypothesis.
    - The first line in the csv file contains the column headings.
      - The columns in this file are
        - Restriction class from Ontology 1: (R1)* property-value pairs representing restriction class from Source/Ontology 1
        - Restriction class from Ontology 2: (R2)* property-value pairs representing restriction class from Source/Ontology 2
        - $|img(R1) \text{ int } R2| / |img(R1)|$ : support score for the alignment from the first source. (See *R* from paper in Fig. 5 Metrics)
        - $|img(R1) \text{ int } R2| / |R2|$ : support score for the alignment from the second source. (See *P* from paper in Fig. 5 Metrics)
        - Relation*: Equivalent, R1 subset R2 or R2 subset R1
        - Size of Intersection*
        - Size of Restriction 1*
        - Size of Restriction 2*
      - These alignments were produced by the algorithm described in the paper.
      - Important Note: Alignments still have pending post-processing
  - The *results\_source1\_source2* file:
    - Each row in the file represents an alignment generated by the algorithm after post-processing along with the stats.
    - The columns in the file are similar to the *alignments\_source1\_source2* file.
      - These columns are:
        - Restriction class from Ontology 1: (R1)* property-value pairs representing restriction class from Source/Ontology 1
        - Restriction class from Ontology 2: (R2)* property-value pairs representing restriction class from Source/Ontology 2
        - $|img(R1) \text{ int } R2| / |img(R1)|$ : support score for the alignment from the first source. (See *R* from paper in Fig. 5 Metrics)
        - $|img(R1) \text{ int } R2| / |R2|$ : support score for the alignment from the second source. (See *P* from paper in Fig. 5 Metrics)
        - Relation*: Equivalent, Similar, R1 subset R2 or R2 subset R1

- **Euzenat et al. – Ontology Matching**
  - Terminological
  - Structural
  - Semantic
- **FCA-Merge, Duckham et al.**
  - Use extensional techniques
- **GLUE**
  - Uses an extensional technique after performing machine learning operations

- Our algorithm generates alignments, consisting of conjunctions of restriction classes
  - Extensional approach on Linked Data
  - Use of restriction classes
- **Alignments based on the actual data**
  - We determine the relationships based on the data
  - Schemas of linked sources can be readily modeled and used
- **Algorithm also able to**
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology

- How to actually understand these alignments

#	MGI restriction	GENEID restriction	Relation
17	bio2rdf:subType=Pseudogene	bio2rdf:subType=pseudo	$r_1 = r_2$
18	bio2rdf:subType=Pseudogene & mgi:genomeStart=17	geneid:chromosome=17 & bio2rdf:subType=pseudo	$r_1 = r_2$
19	bio2rdf:chromosomePosition=-1.00 & mgi:genomeStart=4	geneid:chromosome=4 & bio2rdf:subType=pseudo	$r_2 \subset r_1$

- Scalability
  - Pre-processing of the sources
  - Faster alignment processing