USC Viterbi
School of Engineering

Finding Concept Coverings in Aligning Ontologies of Linked Data

Rahul Parundekar, Craig A. Knoblock and Jose-Luis Ambite
{parundek,knoblock,ambite}@usc.edu
University of Southern California

USC

USC Viterbi
School of Engineering

INTRODUCTION

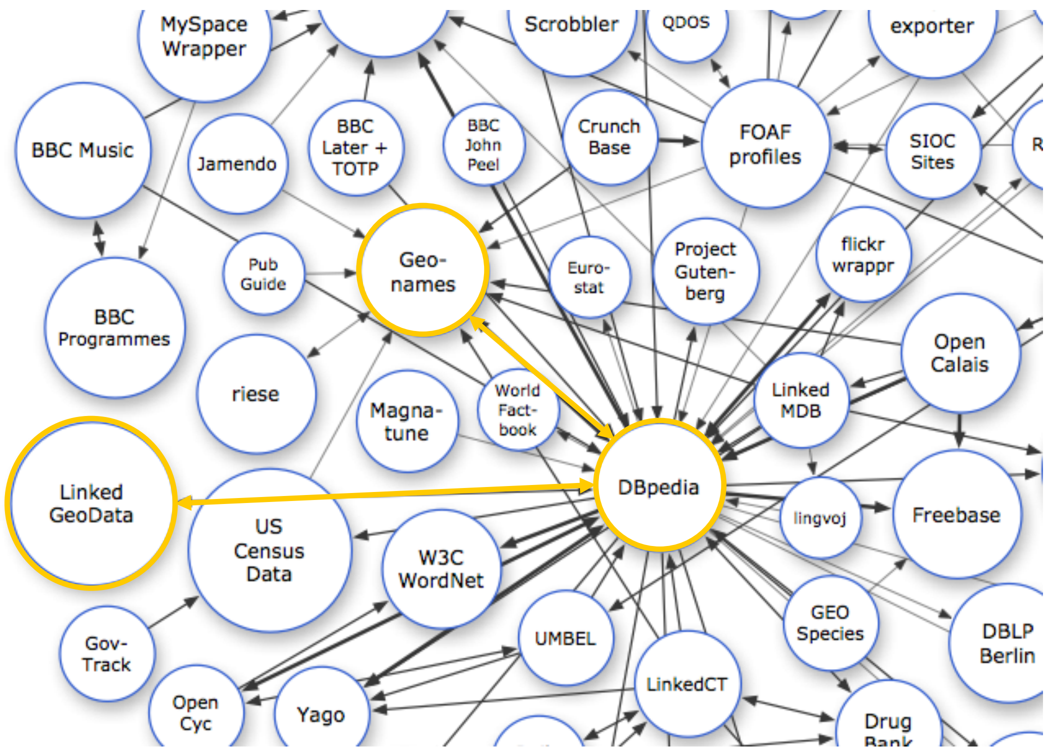
USC

Web of Linked Data

USC Viterbi School of Engineering

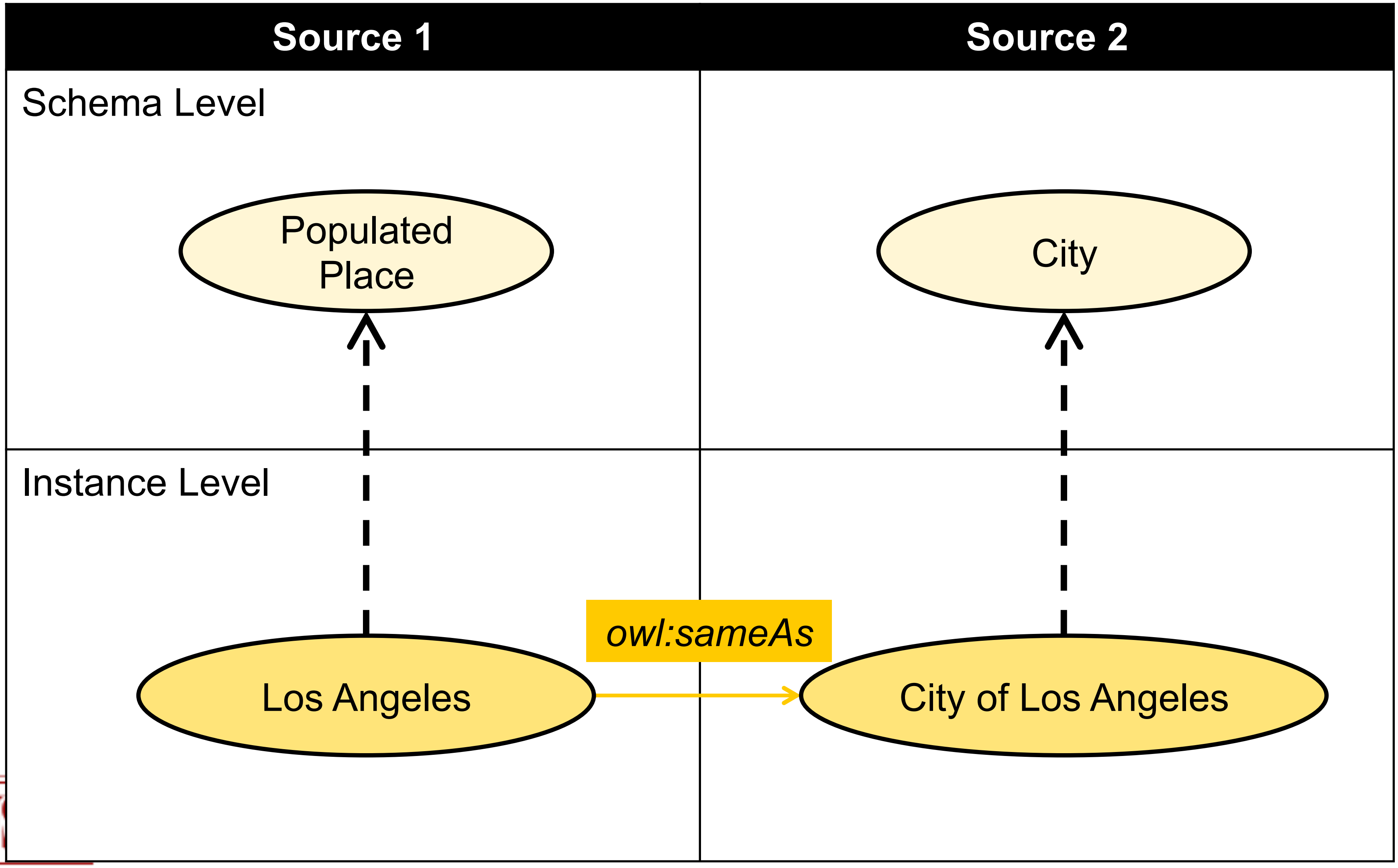
- Different sources with different schemas
- Equivalent instances in the different domains connected with *owl:sameAs*

Example: Geospatial Domain

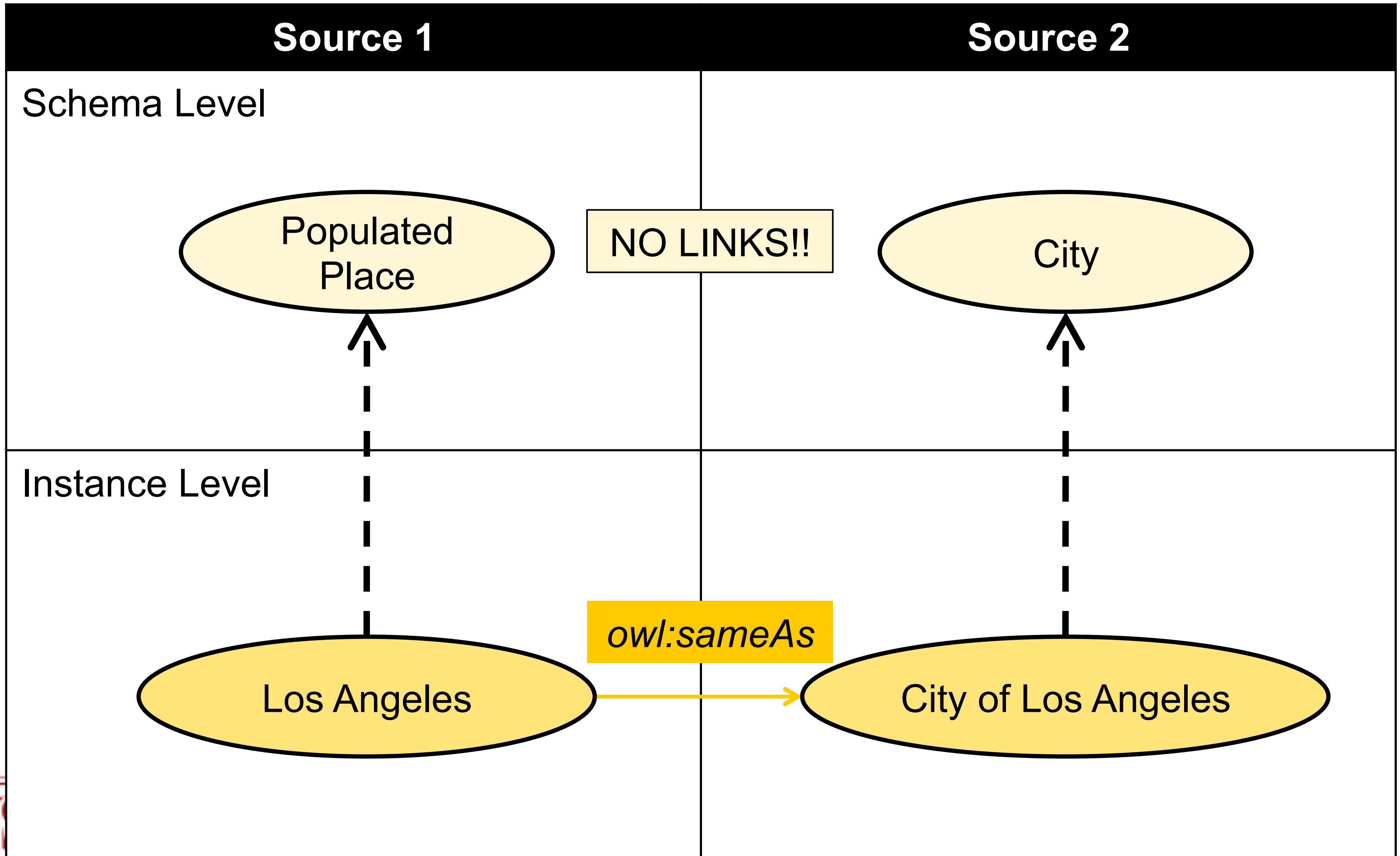


USC

Interlinked instances…

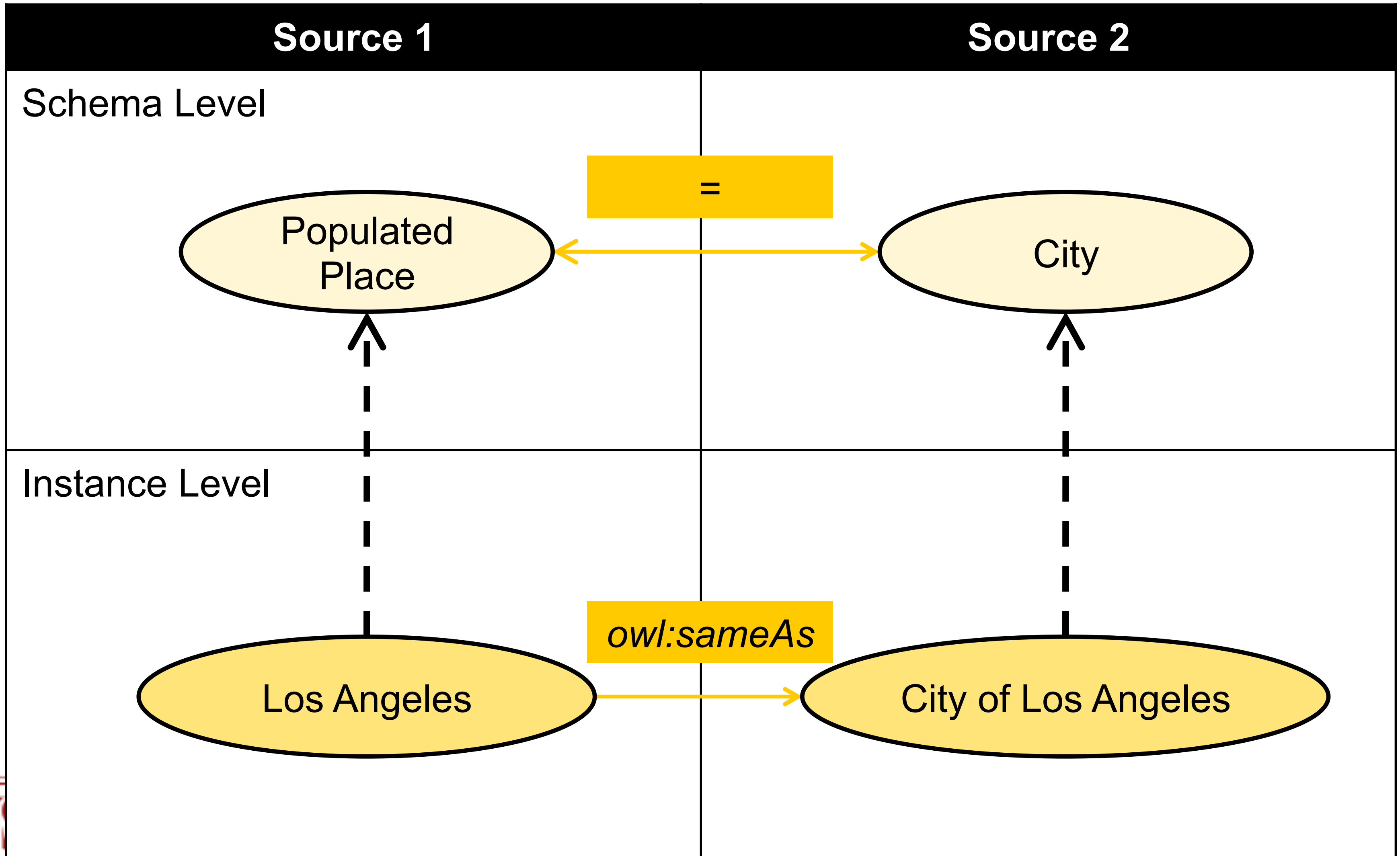


…with disjoint schemas



USC Viterbi School of Engineering

Can we find schema alignments?



USC Viterbi School of Engineering

Previous Work @ ISWC 2010

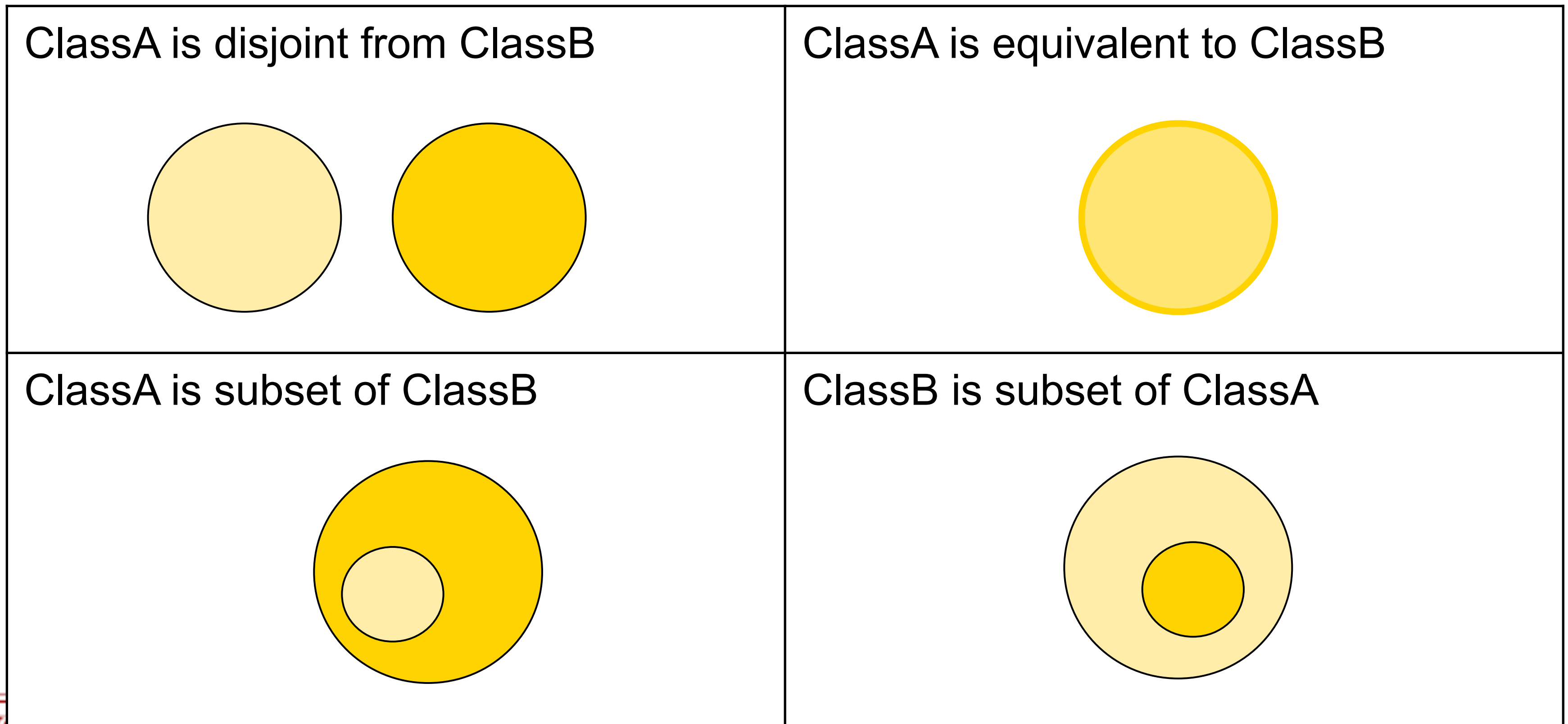
Linking and Building Ontologies of Linked Data

USC

Extensional Approach to Ontology Alignment

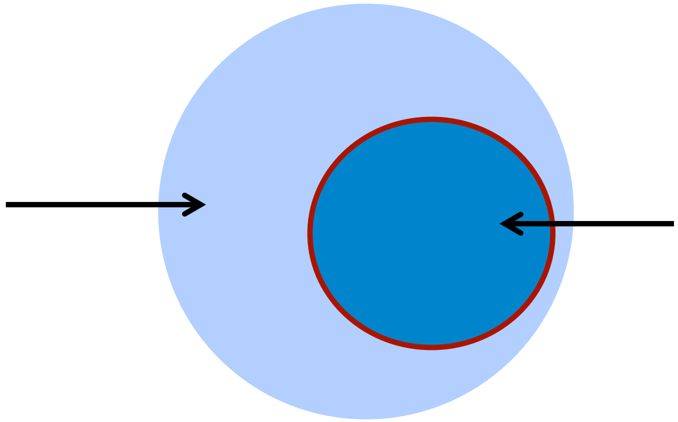
Represents set of instances belonging to ClassA

Represents set of instances belonging to ClassB



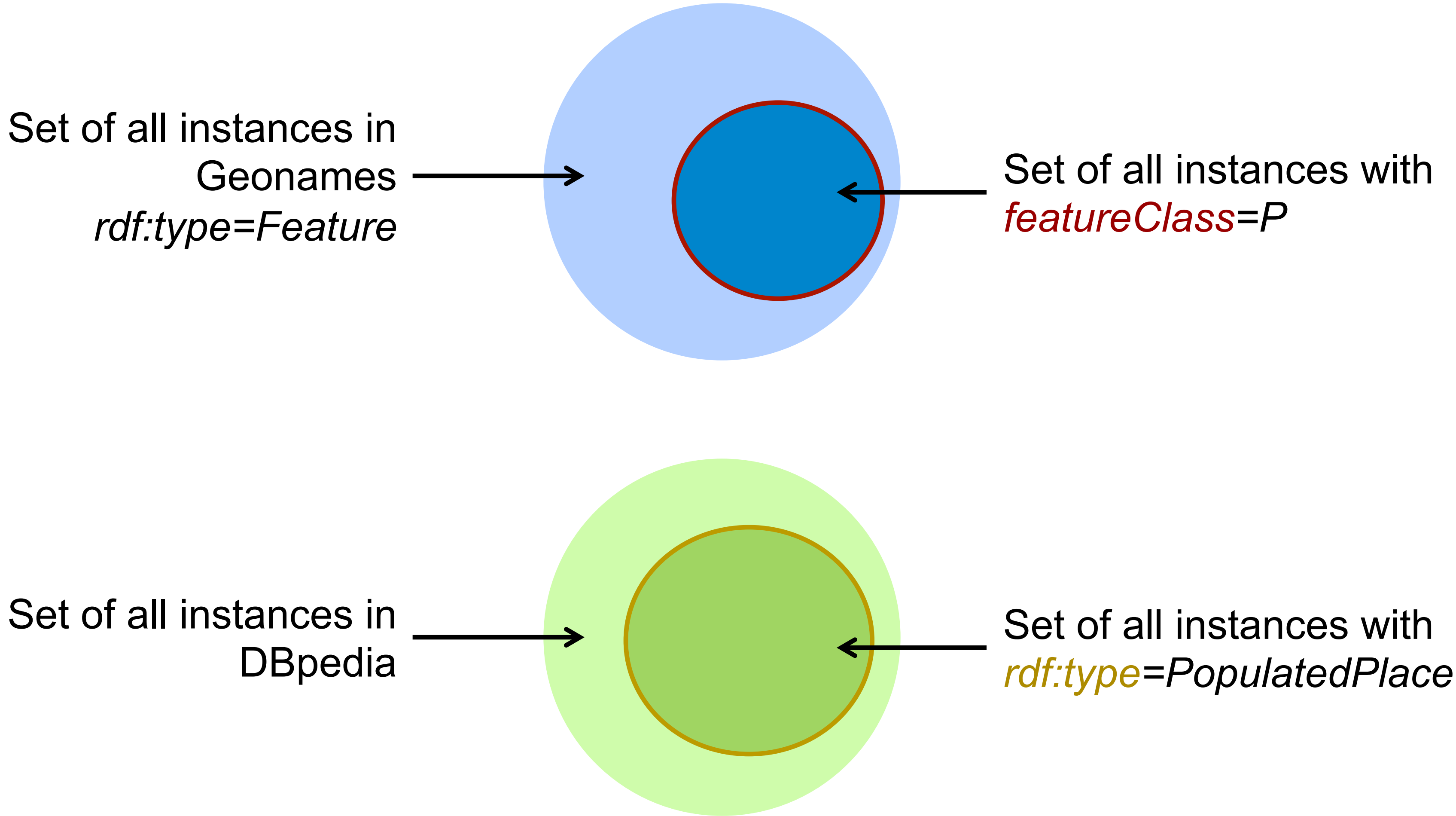
Classes are created extensionally by adding value restrictions on properties

Set of all instances in Geonames *rdf:type=Feature*

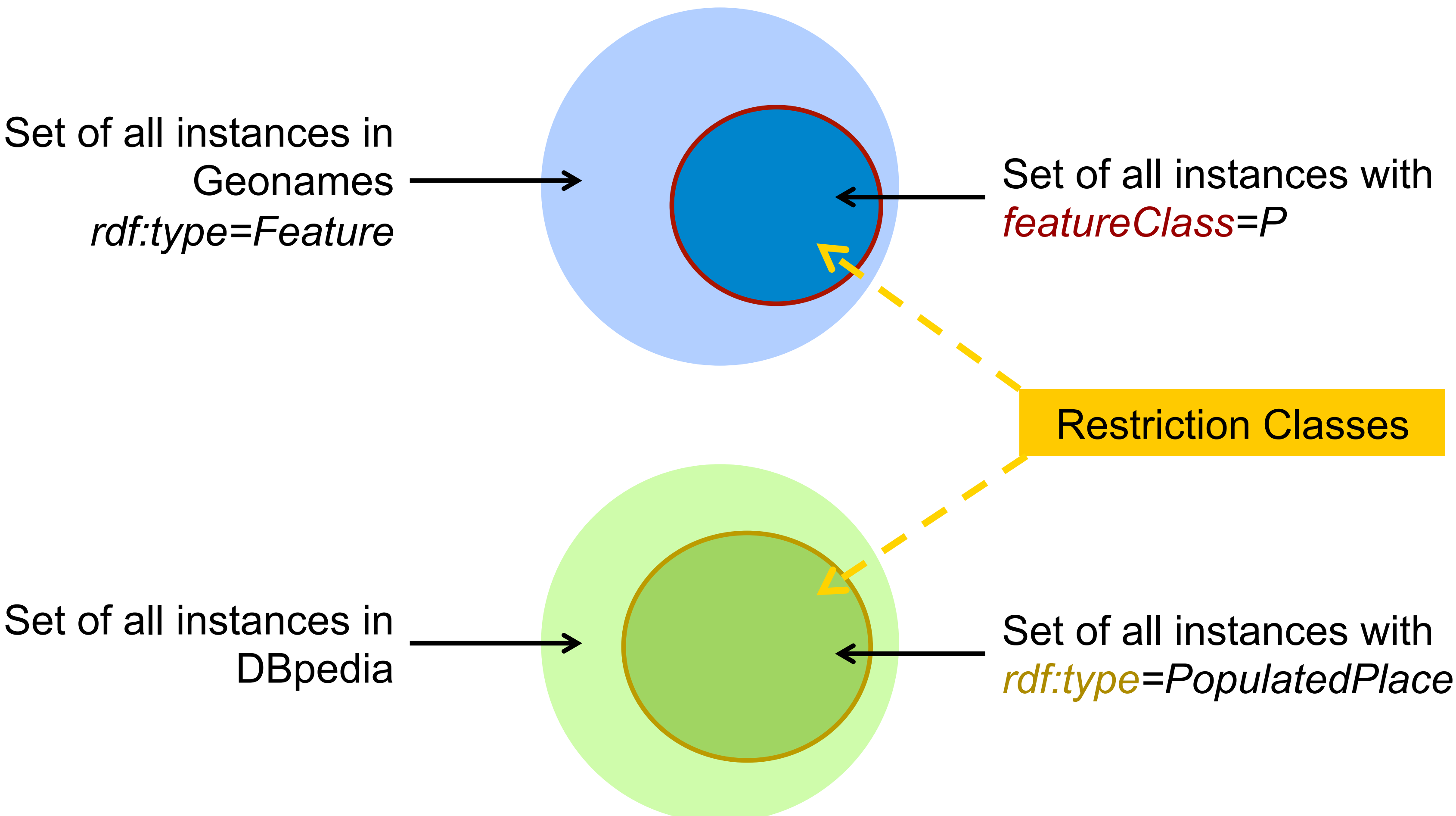


Set of all instances with *featureClass=P*

Classes are created extensionally by adding value restrictions on properties

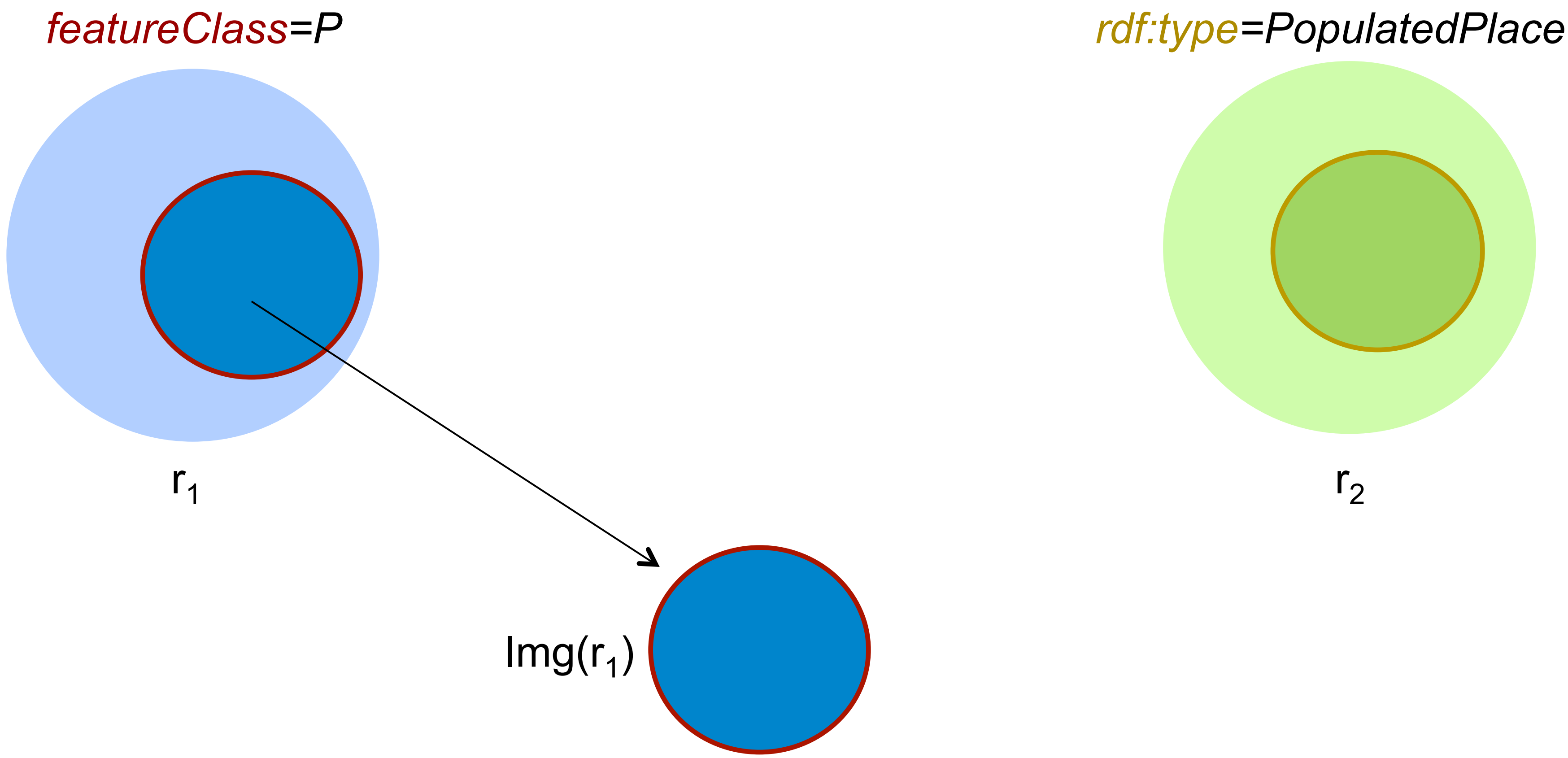


Classes are created extensionally by adding value restrictions on properties



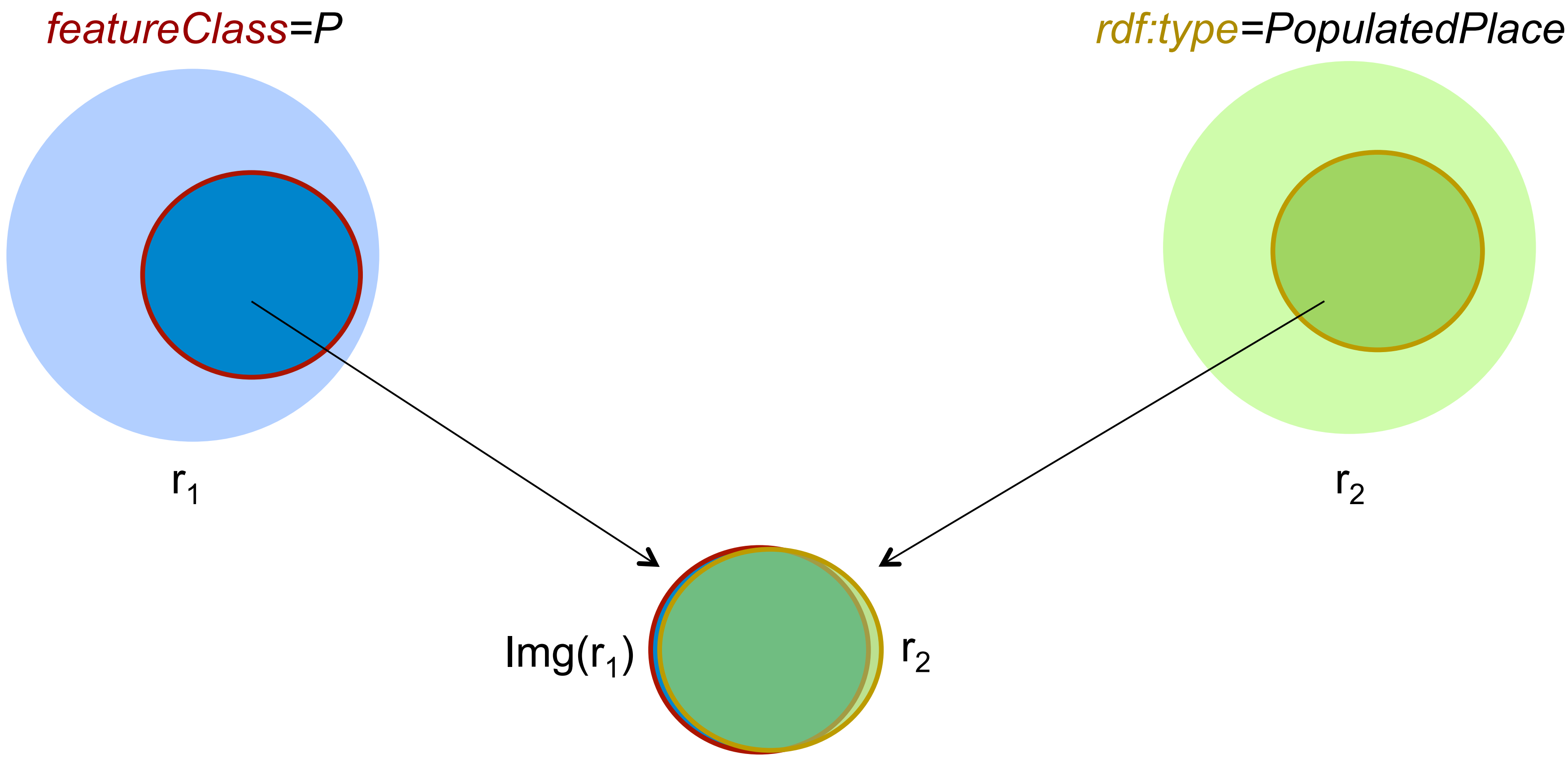
Aligning Restriction Classes Using Extensional Approach

Aligning Restriction Classes Using Extensional Approach



Set of instances from DBpedia that r_1 is linked to

Aligning Restriction Classes Using Extensional Approach



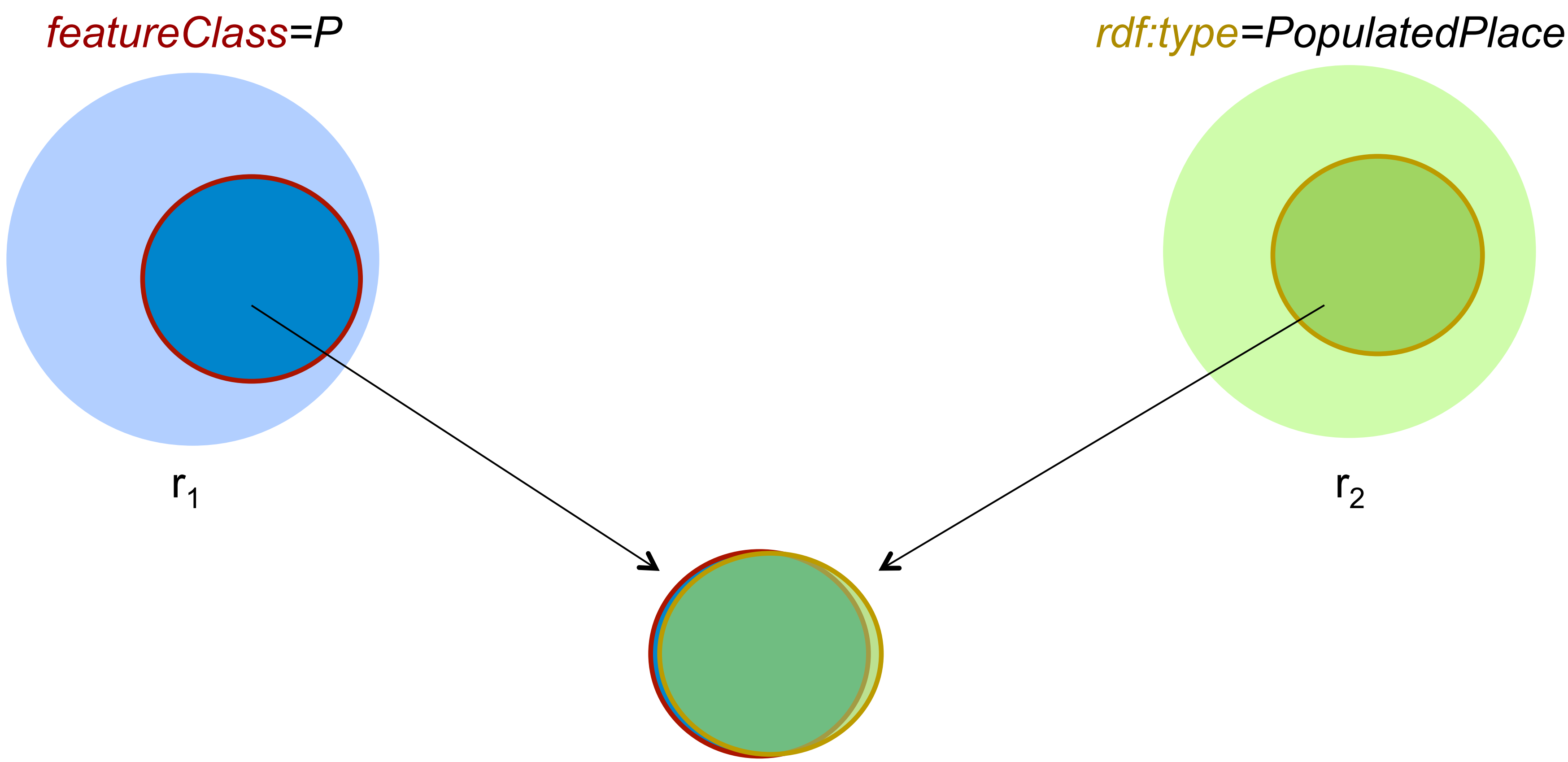
Extensionally, when are two classes equal?

Represents set of instances belonging to ClassA

Represents set of instances belonging to ClassB

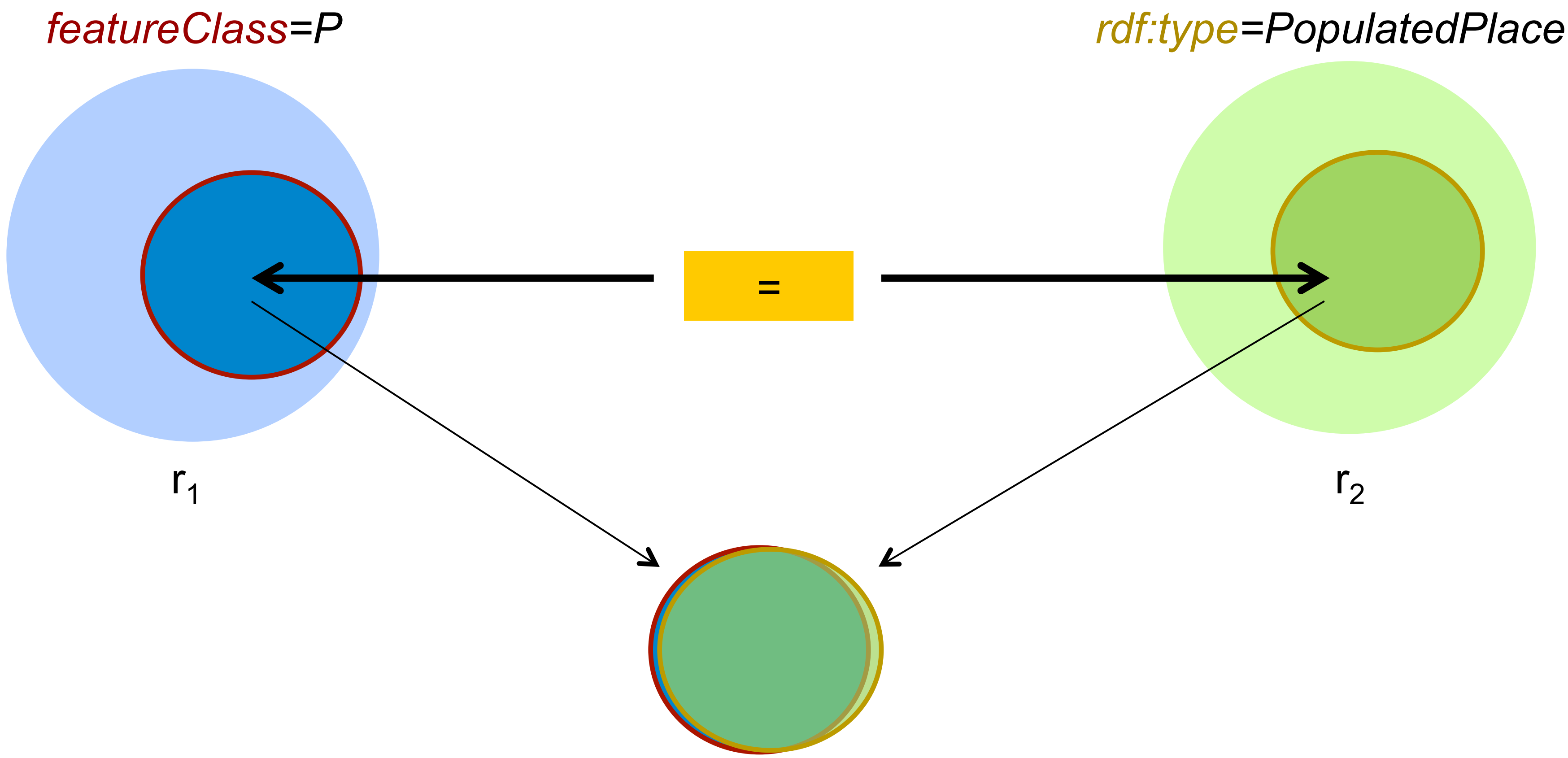
$$\frac{|ClassA \cap ClassB|}{|ClassA|} = \frac{|ClassA \cap ClassB|}{|ClassB|} = 1$$

Aligning Restriction Classes Using Extensional Approach



$$\frac{|\,Img(r_1) \cap r_2|}{|\,Img(r_1)\,|} > 0.9 \qquad \frac{|\,Img(r_1) \cap r_2|}{|r_2|} > 0.9$$

Aligning Restriction Classes Using Extensional Approach



$$\frac{|\,Img(r_1) \cap r_2|}{|\,Img(r_1)\,|} > 0.9 \qquad \frac{|\,Img(r_1) \cap r_2|}{|r_2|} > 0.9$$

Alignments Found in the ISWC'10 Paper

- Algorithm was able to
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology
- Alignments based on the actual data
  - reflects the semantics of the sources in practice
- 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	-	-

USC Viterbi School of Engineering

Alignments Found in the ISWC'10 Paper

- Algorithm was able to
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology
- Alignments based on the actual data
  - reflects the semantics of the sources in practice
- 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	-	-

Can we use the subset relations to find more meaningful alignments?

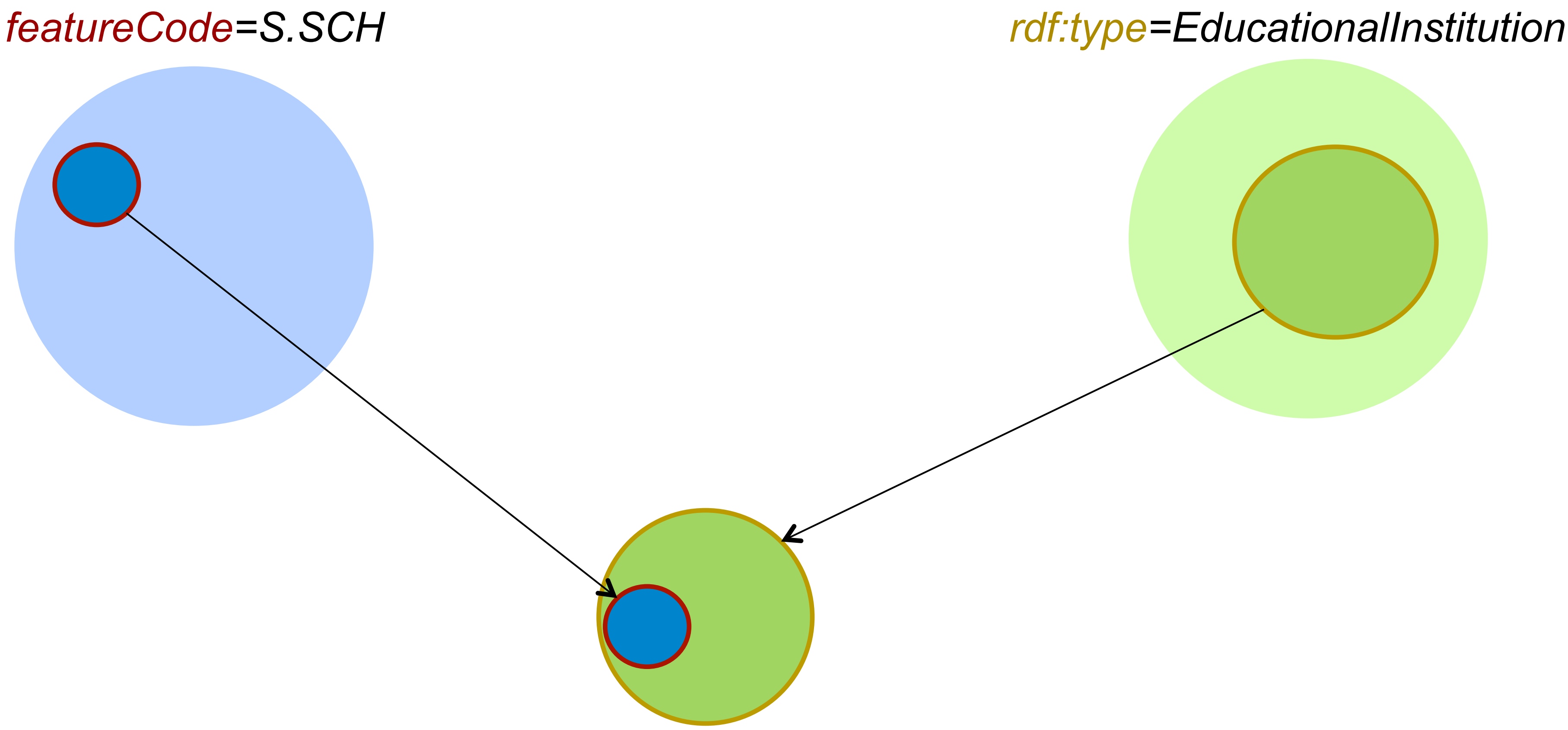
Know@LOD Workshop – ESWC 2012

FINDING CONCEPT COVERINGS IN ALIGNING ONTOLOGIES OF LINKED DATA

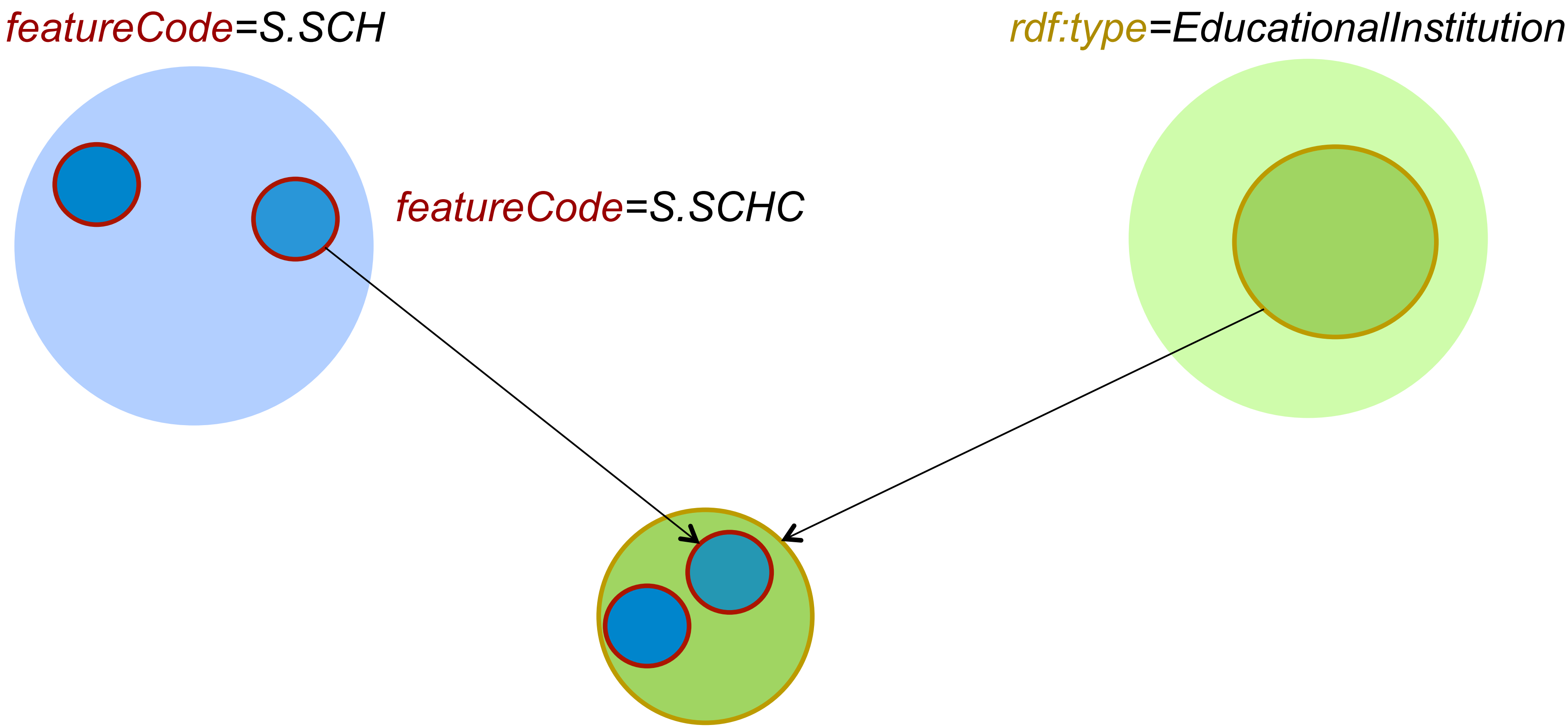
Is there a pattern in the subset relations?

Let’s look at 3 of the subset relations we found…

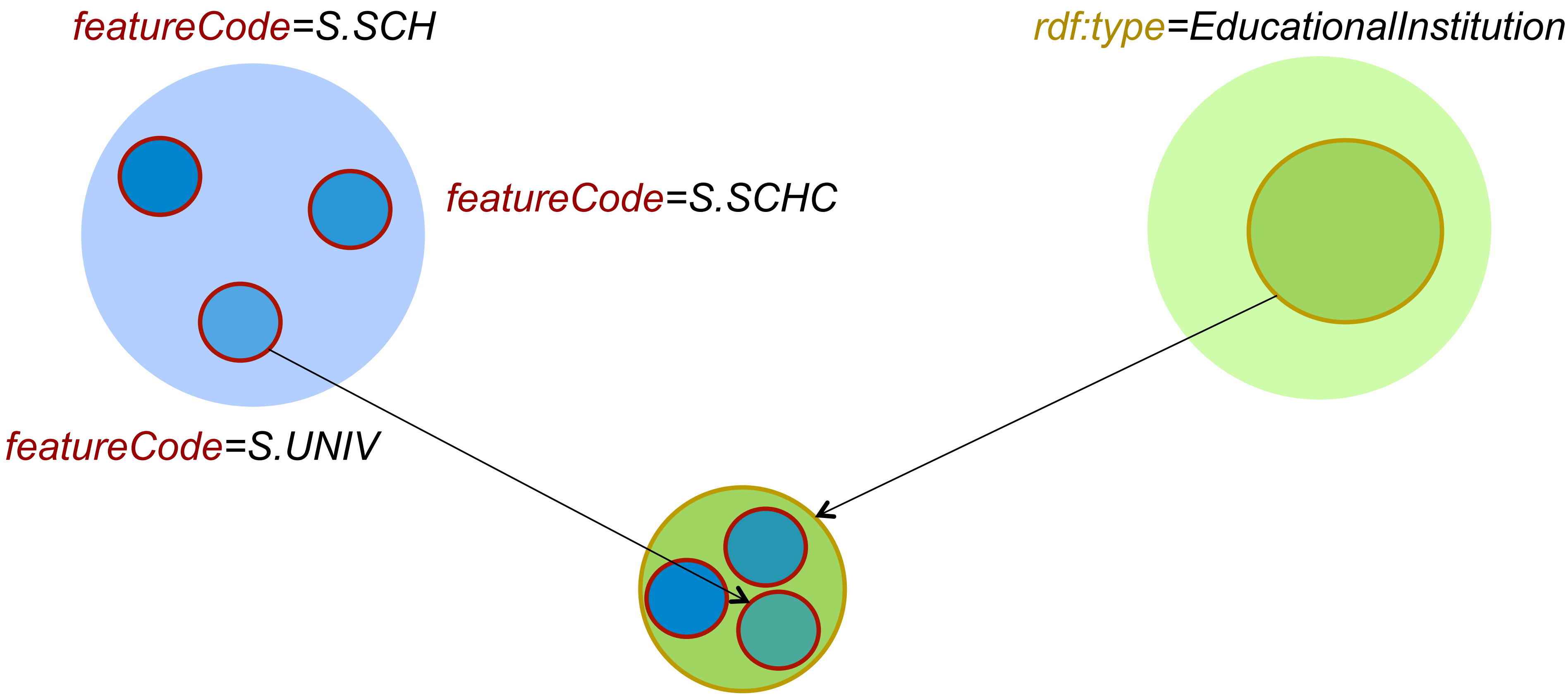
1) Schools in *Geonames* are Educational Institutions in *DBpedia*



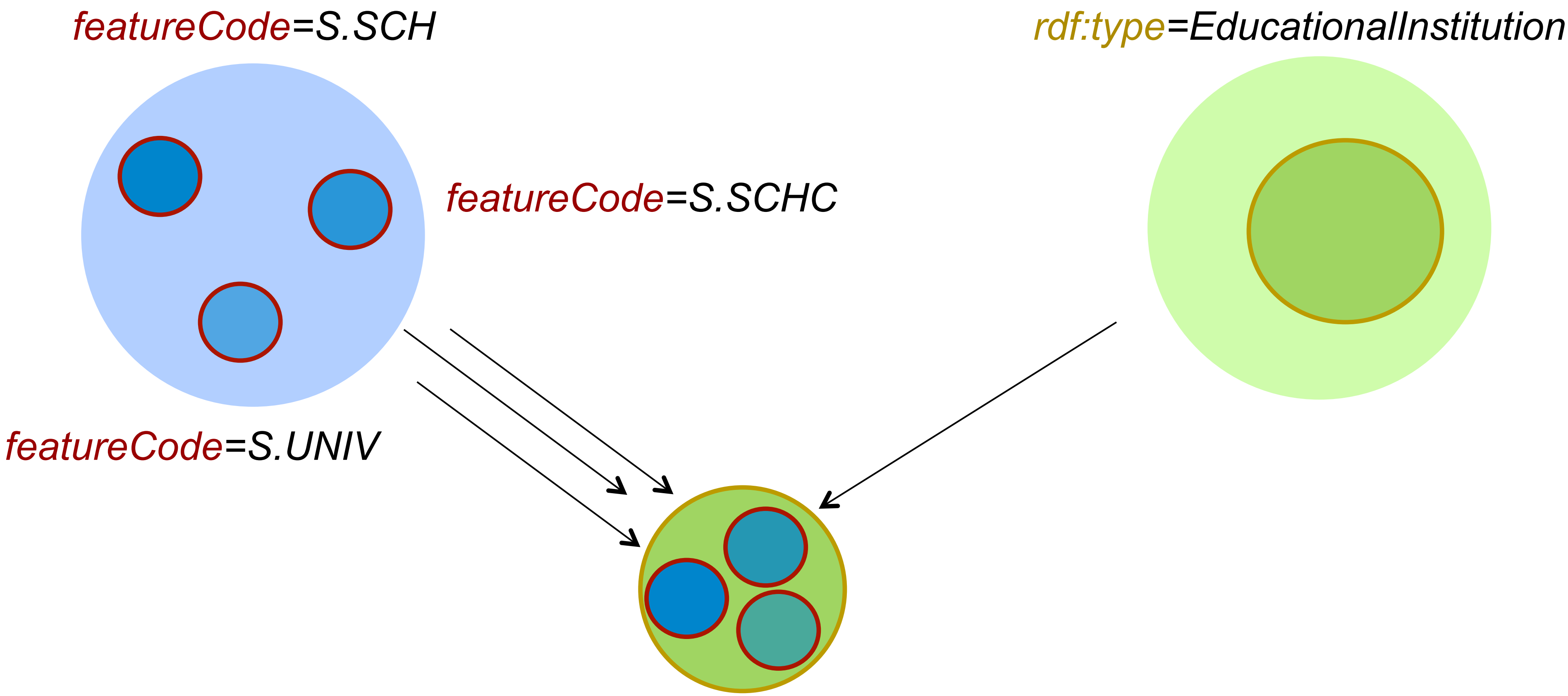
2) Colleges in *Geonames* are Educational Institutions in *DBpedia*



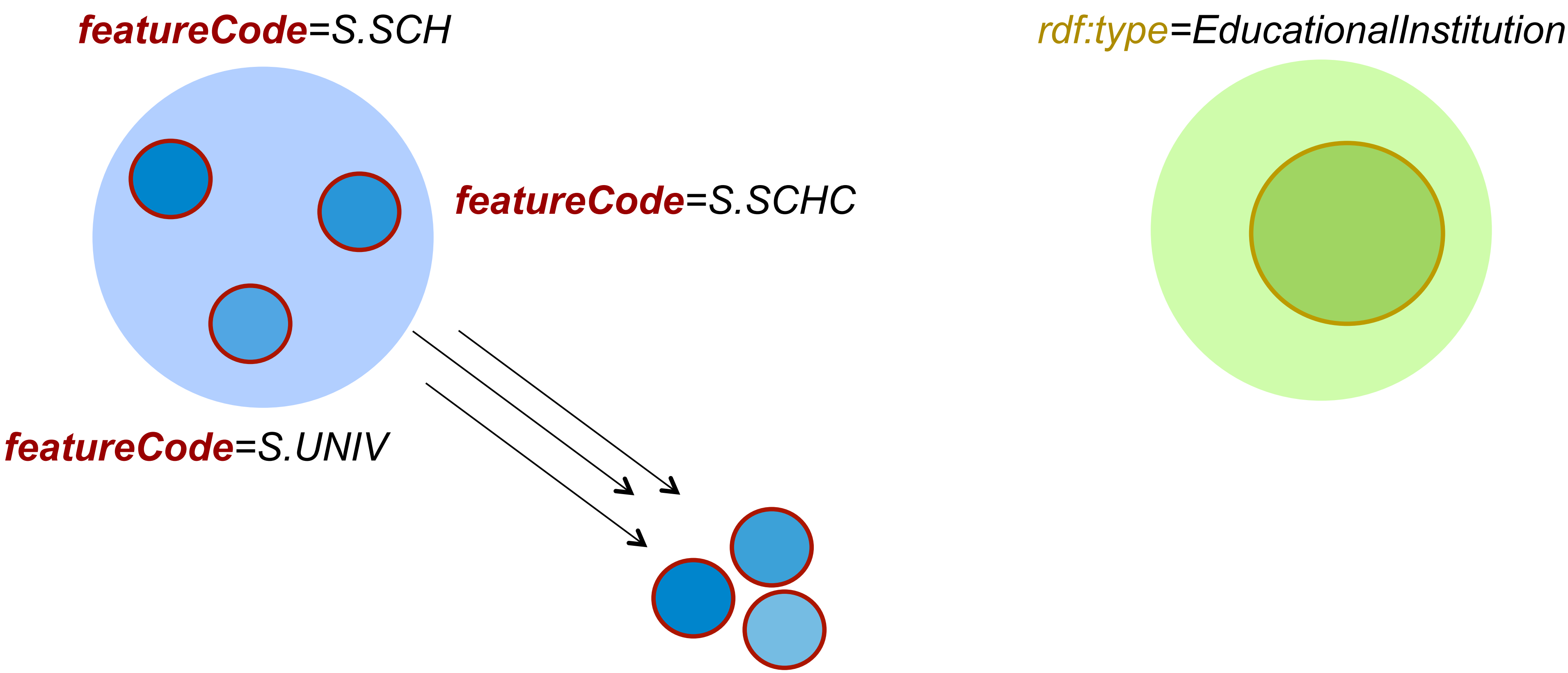
3) Universities in *Geonames* are Educational Institutions in *DBpedia*



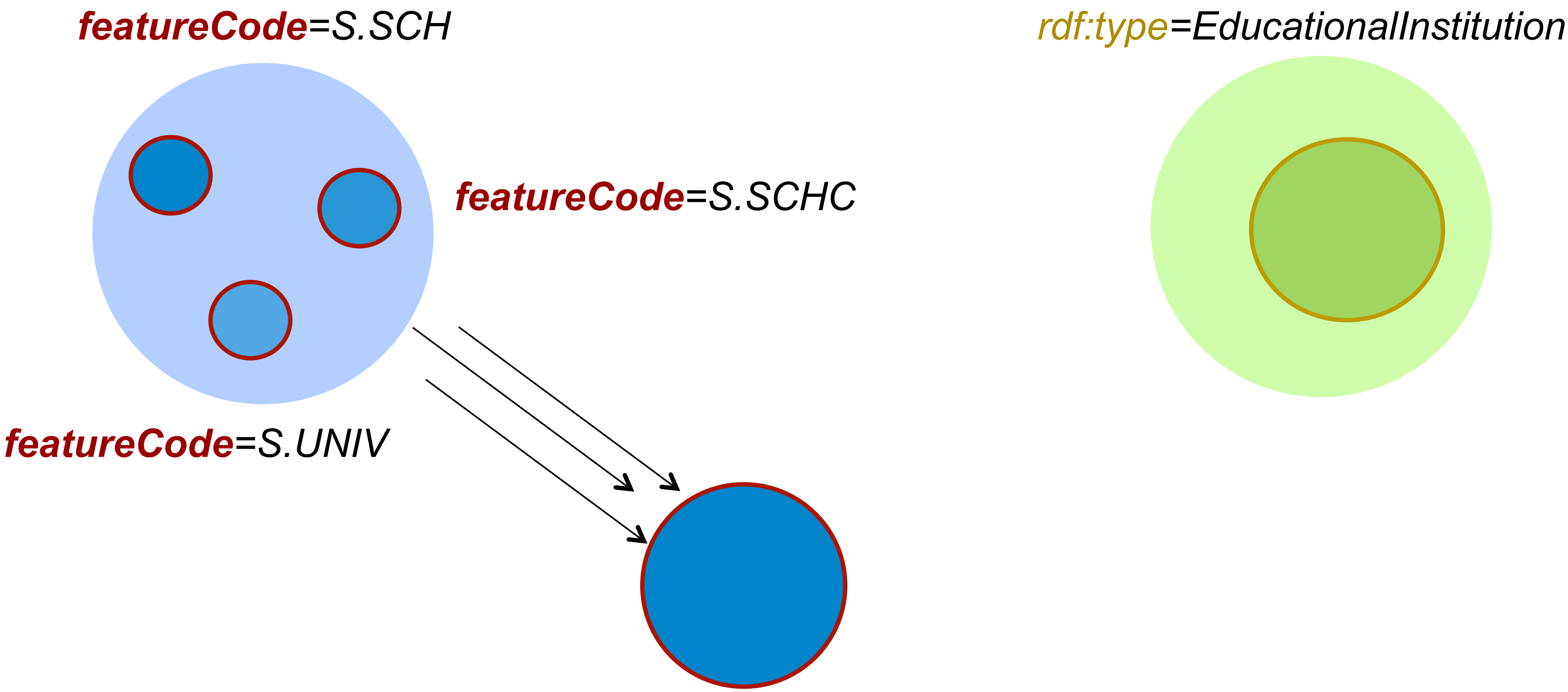
Taken by themselves, the subset relations are not useful



We use the common ***featureCode*** property as a hint…

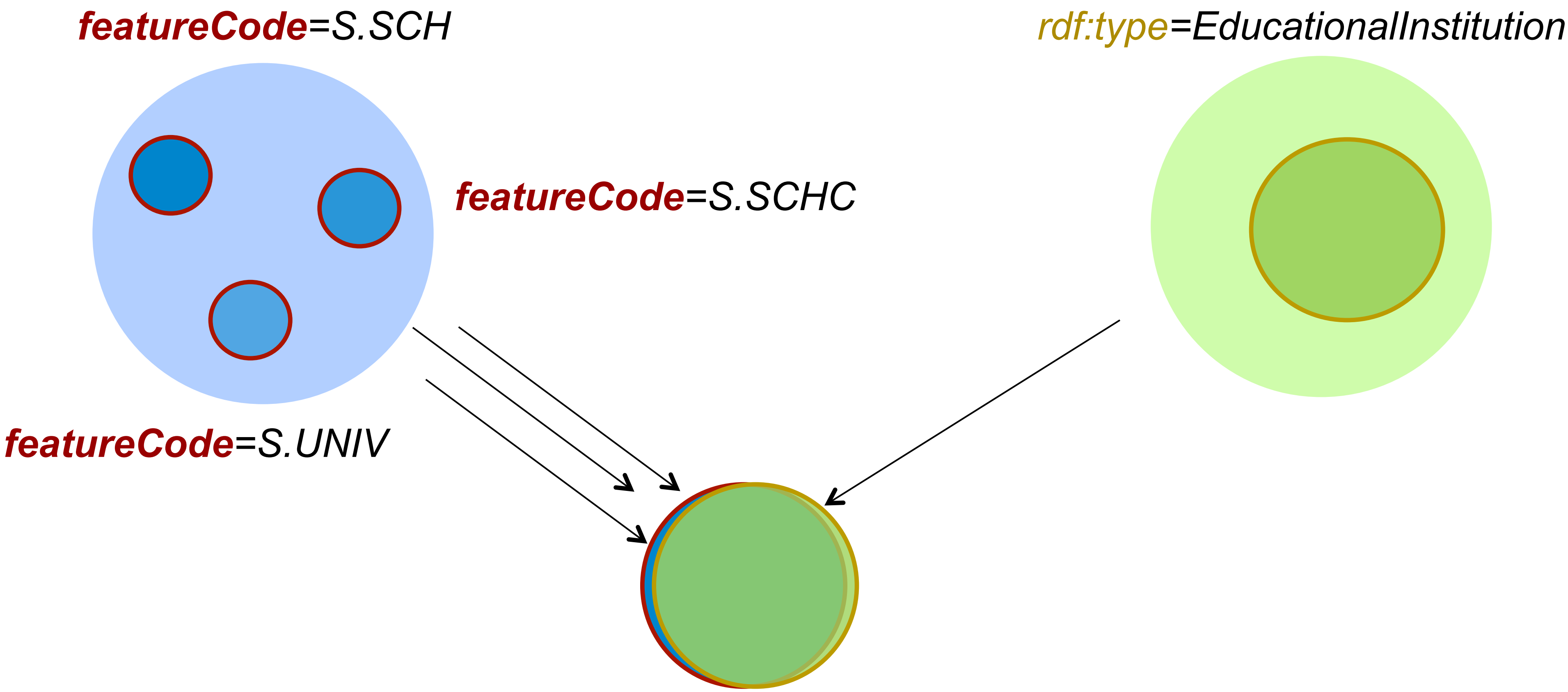


…to form a ***Union*** of Restriction Classes



featureCode*=S.SCH* ∪ ***featureCode****=S.SCHC* ∪ ***featureCode****=S.UNIV*

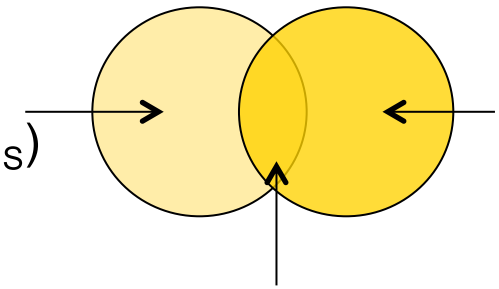
Contribution 1: Find Union Alignments



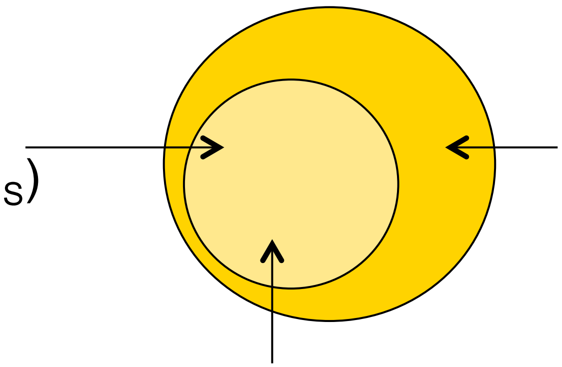
featureCode=*S.SCH* ∪ ***featureCode***=*S.SCHC* ∪ ***featureCode***=*S.UNIV*

Finding Union Alignments: Approach

- For all alignments found in the ISWC2010 paper marked as subsets
  1. We group all subset alignments according to the common larger restriction class
  2. We form a *union concept* such that all restriction classes
     - have the same property
     - have a single *property-value pair* each
  3. We then try to match the *union concept* to the larger class
  4. This forms a hypothesis *Union Alignment*



Finding Union Alignments: Scoring



$$\frac{|U_A|}{|U_S|} = 1$$ since by definition, all smaller classes are subsets

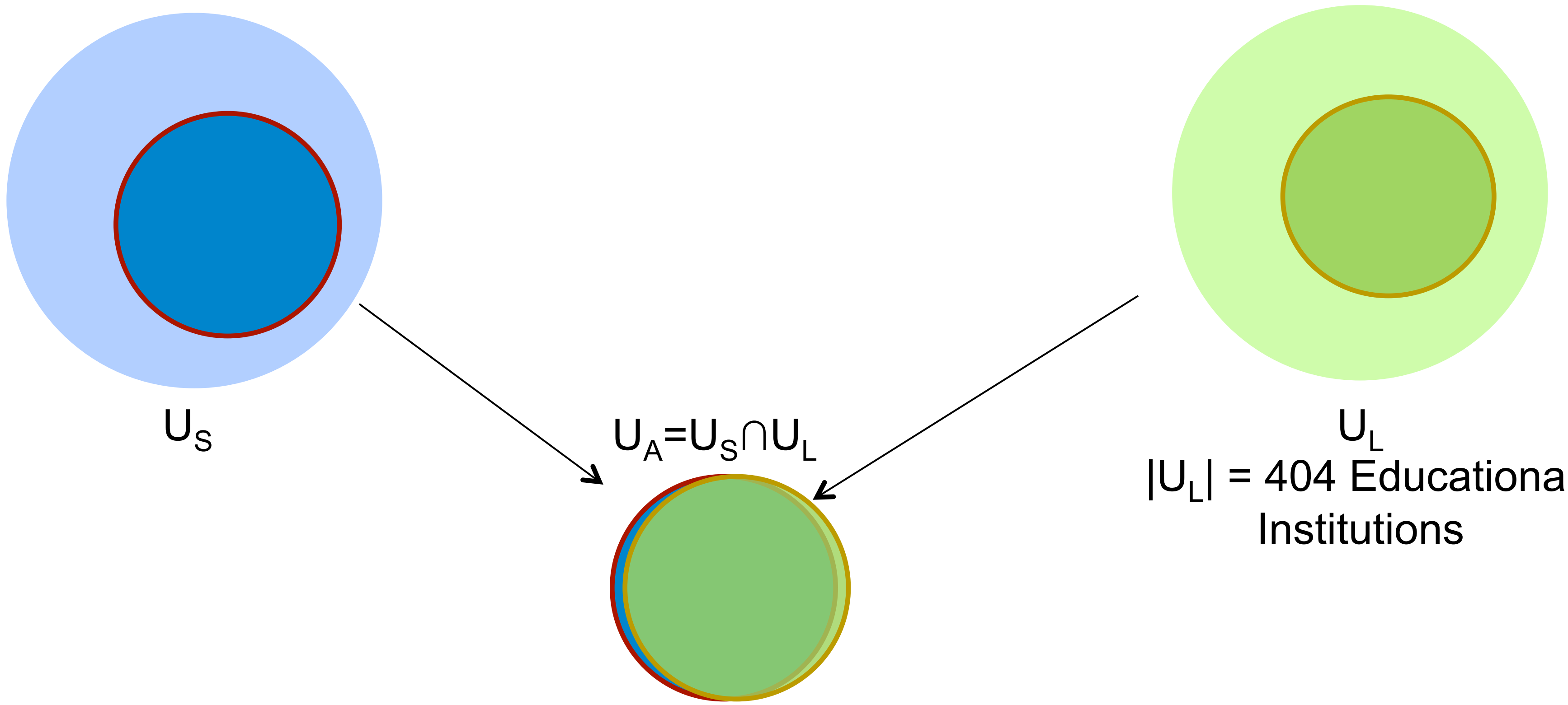
So, if $\frac{|U_A|}{|U_L|} = 1$, then the larger class U_L is equivalent to U_S

Practically, we use a relaxed subset assumption: $\frac{|U_A|}{|U_S|}, \frac{|U_A|}{|U_L|} > 0.9$

Contribution 1: Find Union Alignments

featureCode=*{S.SCH, S.SCHC, S.UNIV}*

rdf:type=*EducationalInstitution*



$$\frac{|U_A|}{|U_S|} > 0.9$$

$$\frac{|U_A|}{|U_L|} = \frac{396}{404} = 0.98 > 0.9$$

WHAT ARE THE OTHER 8 EDUCATIONAL INSTITUTIONS?

Contribution 2: Find Outliers / Discrepancies

- We are also able to point out where the instances that disagree with the alignment lie
- These instances were not part of the alignment because
  - Their restriction class was not a subset (P’<0.9)
  - Some of these instances are
    - Linked Incorrectly with *owl:sameAs*
    - Assigned wrong value during RDF generation*
    - Common in both sets (could be debatable)
    - Did not have a minimum support size of 2 instances (set with 1 instance cannot be relied on)
- Outliers help in understanding discrepancies in the Linked Data

*See Country of http://dbpedia.org/page/Skegness

What are the other 8 Educational Institutions?

- 1 with *featureCode*=S.HSP (Hostpitals)
  - There are 31 instances with S.HSP because of which Hospitals are not subsets
- 3 with *featureCode*=S.BLDG (Buildings)
- 1 with *featureCode*=S.EST (Establishment)
- 1 with *featureCode*=S.LIBR (Library)
- 1 with *featureCode*=S.MUS (Museum)

- 1 doesn't have a *featureCode* property

RESULTS

Results: *Geonames-DBpedia*

#	$\{r_1\}$	$p_2 \in \{v_2\}$	$R'_U = \frac{\|U_A\|}{\|U_L\|}$	$\|U_A\|$	$\|U_L\|$	*Outliers*	*# Explained Instances*
DBpedia **(larger) -** ***GeoNames*** **(smaller)**							
1	*{rdf:type = dbpedia:EducationalInstitution}*	*geonames:featureCode* ∈ {S.SCH, S.SCHC, S.UNIV}	0.9801	396	404	S.BLDG (3/122), S.EST (1/13), S.LIBR (1/7), S.HSP (1/31), S.MUS (1/43)	403
	As described in Section 4, Schools, Colleges and Universities in *GeoNames*make Educational Institutions in *DBpedia*						
2	*{dbpedia:country = dbpedia:Spain}*	*geonames:countryCode = ES*	0.9997	3917	3918	IT (1/7635)	3918
	The concepts for the country Spain are equal in both sources. The only outlier has it's country as Italy, an erroneous assertion.						
3	*dbpedia:region = dbpedia:Basse-Normandie*	*geonames:parentADM2* ∈ {geonames:2989247, geonames:2996268, geonames:3029094}	1.0	754	754		754
	We confirm the hierarchical nature of administrative divisions with alignments between administrative units at two different levels.						
4	*{rdf:type = dbpedia:Airport}*	*geonames:featureCode* ∈ {S.AIRB, S.AIRP}	0.9924	1981	1996	S.AIRF (9/22), S.FRMT (1/5), S.SCH (1/404), S.STNB (2/5) S.STNM (1/36), T.HLL (1/61)	1996
	In alignmening airports, an airfield should have been an an airport. However, there was not enough instance support.						
GeoNames **(larger) -** ***DBpedia*** **(smaller)**							
5	*{geonames:countryCode = NL}*	*dbpedia:country* ∈ {dbpedia:The_Netherlands, dbpedia:Flag_of_the _Netherlands.svg, dbpedia:Netherlands}	0.9802	1939	1978	dbpedia:Kingdom_of _the_Netherlands	1940
	The Alignment for Netherlands should have been as straightforward as #2. However we have possible alias names, such as *The Netherlands* and *Kingdom of Netherlands*, as well a possible linkage error to *Flag of the Netherlands.svg*						
6	*{geonames:countryCode = JO}*	*dbpedia:country* ∈ {dbpedia:Jordan, dbpedia:Flag_of_Jordan.svg}	0.95	19	20		20
	The error pattern in #5 seems to repeat systematically, as can be seen from this alignment for the coutry of Jordan.						

Results: *LinkedGeoData-DBpedia*

#	$\{r_1\}$	$p_2 \in \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
***DBpedia* (larger) - *LinkedGeoData* (smaller)**							
7	{*dbpedia:bundesland = Saarland*}	*lgd:OpenGeoDBLicensePlate-*Number ∈ { HOM, IGB, MZG, NK, SB, SLS, VK, WND}	0.93	46	49		46
	Our algorithm also produces interesting alignments between different properties. In this case, we find 8 of the 10 license plates in the state of Saarland						
8	{*rdf:type, dbpedia:EducationalInstitution*}	*rdf:type* ∈ {lgd:Amenity, lgd:K2543, lgd:School, lgd:University, lgd:WaterTower}	0.9901	2609	2610		2609
	Educational Institutions in *DBpedia* can be explained with classes in *LinkedGeoData*. An example of an incorrent alignment, a water tower has been linked to as an educational institution.						
***LinkedGeoData* (larger) - *DBpedia* (smaller)**							
9	{*lgd:gnisST_alpha = NJ*}	*dbpedia:subdivisionName* ∈ {Atlantic, Burlington, {Cape May, Hudson, Hunterdon, Monmoth, New Jersey, Ocean, Passaic}	1.0	214	214		214
	Due to missing instance alignments, this *union alignment* incorrectly claims that the state of New Jersey is composed of 9 counties while actually it has 21.						
10	{*rdf:type = lgd:Waterway*}	*rdf:type* ∈ dbpedia:River dbpedia:Stream}	0.97	33	34	dbpedia:Place(1/94989)	34
	Waterways in *LinkedGeoData* as equal to the union of streams and rivers from *DBpedia*						

Results: *Geospecies-DBpedia*

#	$\{r_1\}$	$p_2 \in \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
***DBpedia* (larger) - *Geospecies* (smaller)**							
11	{*rdf:type* = *dbpedia:Amphibian*} *dbpedia:Amphibian* }	*geospecies:hasOrderName* ∈ {Anura, Caudata, Gymnophionia}	0.99	90	91	Testudines (1/7)	91
	Species from *Geospecies* with the order names Anura, Caudata & Gymnophionia are all Amphibians We also find inconsistancies due to misaligned instances, e.g. one Turtle (Testidune) was classified as amphibian.						
12	{*rdf:type* = *dbpedia:Salamander*}	{*geospecies:hasOrderName* = Caudata}	0.94	16	17	Testudines (1/7)	17
	Upon further inspection of #11, we find that the culprit is a Salamander						
***Geospecies* (larger) - *DBpedia* (smaller)**							
13	{*rdf:type* = *dbpedia:Plant*}	{*geospecies:inKingdom* = geospecies:kingdoms/Ab}	0.99	1874	1876	geospecies:kingdoms/Ac(1/8)	1875
	The Kingdom Plantae, from both sources, almost matches perfectly. The only inconsistant instance happens to be a fungus.						
14	{*geospecies:inOrder* = *geospecies:orders/jtSaY*}	*dbpedia:ordo* ∈ {dbpedia:Carnivora, dbpedia:Carnivore}	0.99	247	247		247
	Inconsistancies in the object values can also be seen - Carnivores from *Geospecies* are aligned with both : Carnivora & Carnivore.						
15	{*geospecies:hasOrderName* = *Chiroptera*}	*dbpedia:ordo* ∈ {Chiroptera@en, dbpedia:Bat}	1	111	111		111
	We can detect that species with order Chiroptera correctly belong to the order of Bats. Unfortunatey, due to values of the property being the literal "Chiropta@en", the alignment is not clean.						

Results: *GeneID-MGI*

#	$\{r_1\}$	$p_2 \in \{v_2\}$	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
***GeneID* (larger) - *MGI* (smaller)**							
16	{*bio2rdf:subType* = *pseudo*}	{*bio2rdf:subType* = Pseudogene}	0.93	5919	6317	Gene (318/24692)	6237
	Due to the absence of a clear hierarchy, we found only a few hierarchical relations. For example, alignments of the classes Pseudogenes.						
17	{*bio2rdf:xTaxon* = *taxon:10090*}	*bio2rdf:subType* ∈ {Complex Cluster/Region, DNA Segment, Gene, Pseudogene}	1	30993	30993		30993
	The Mus Musculus (house mouse) taxonomy is completely composed of complex clusters, DNA segments, Genes and Pseudogenes .						
***MGI* (larger) - *GeneID* (smaller)**							
18	{*bio2rdf:subType* = *Pseudogene*}	*bio2rdf:subType* = *pseudo*	0.94	5919	6297	other (4/230) protein-coding (351/39999) unknown(23/570)	6297
	Inconsistancies are also evident as the values pseudo and Pseudogene are used to denote the same thing.						
19	{*mgi:genomeStart* = *1*}	*geneid:location* ∈ {1, 1 0.0 cM, 1 1.0 cM, 1 10.4 cM, ...}	0.98	1697	1735	"" (37/1048) 5 (1/52)	1735
20	{*mgi:genomeStart* = *X*}	*geneid:location* ∈ {X, X 0.5 cM, X 0.8 cM, X 1.0 cM, ...}	0.99	1748	1758	"" (10/1048)	1758
	We find interesting alignments like #19 & #20 , which align the genome start position in *MGI* with the location in *GeneID* As can be seen, the values of the locations (distances in centimorgans) in *GeneID* contain genome start value as a prefix. Inconsistancies are also seen, e.g. in #19 a gene that starts with 5 is misaligned and in #20, where the value is an empty string.						

USC Viterbi School of Engineering

Results: Summary

We find a total of 7069 Union Alignments that cover 77966 subset relations for a compression of 90%

Source1	Source2	Union Alignments 12 (Subset Alignments 12)	Union Alignments 21 (Subset Alignments 21)	Total union alignments
GeoNames	*DBpedia*	434 (2197)	318 (7942)	752
LinkedGeoData	*DBpedia*	2746 (12572)	3097 (48345)	5843
Geospecies	*DBpedia*	191 (1226)	255 (2569)	446
GeneID	*MGI*	6 (29)	22 (3086)	28

Results also available at

http://www.isi.edu/integration/data/UnionAlignments

USC

Related Work

- BLOOMS, BLOOMS+ ([4][5] in paper)
  - Linked Open Data ontologies aligned with ‘Proton’
  - Constructs a forest of concepts and computes structural similarity
  - Geonames – Proton has “poor performance” because of small number and vague classes in Geonames
- Volker et al. ([8] in paper)
  - Statistical schema induction
  - Mines associativity rules from intermediate *‘transaction datasets’*
  - Develops OWL2 Axioms
- AgreementMaker [2]
  - Similarity Metrics on labels of classes

Conclusion and Future Work

- Conclusion
  - We were able to find *Union Alignments* in the Geospatial, Biological Classification & Genetics Domain
    - Find alignments where no direct equivalence was evident
    - Introduced a disjunction operator to restriction classes
  - We were able to find *Outliers*
    - Help identify inconsistencies in the data
- Future work
  - Experimental comparison with other approaches
  - Preliminary findings suggest patterns in properties like *geonames:countryCode* and *dbpedia:country*

USC Viterbi School of Engineering

Any questions?

THANK YOU

USC