Publier des données avec Datalift

Laurent Bihanic, Jérôme Euzenat

http://www.datalift.org

September 6, 2013

Prérequis

1. Publier des données en RDF
2. Le projet Datalift

Un exemple de bout en bout

3. La plateforme Datalift: architecture et technologie
4. Publier étapes par étapes
   - Importer des données
   - Sémantiser ces données
   - Choix de license
   - Lier les données publiées

Download

- Allez sur le site `http://datalift.org`
- Puis sur download
- Téléchargez la version correspondant à votre architecture
- Décompressez
- Doublie-cliquez

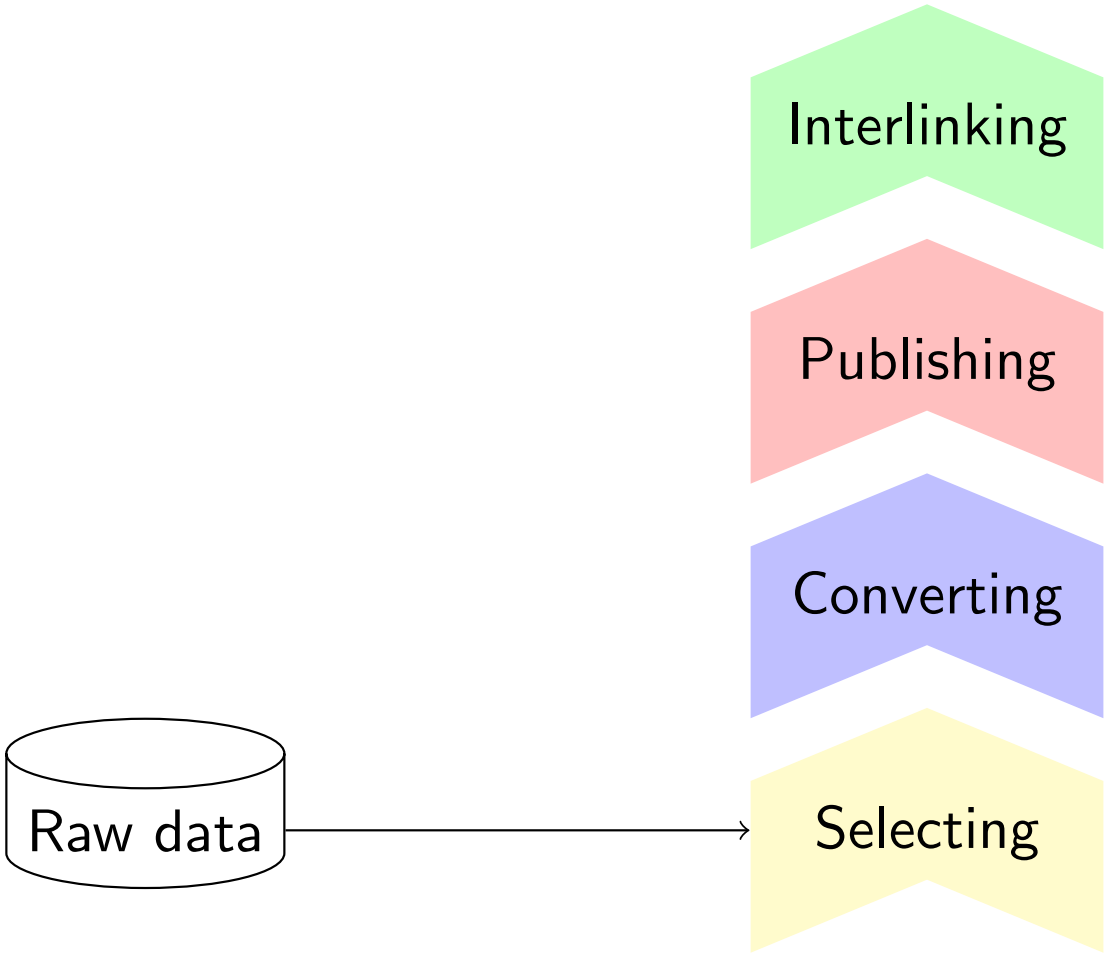
Outline

Datalift

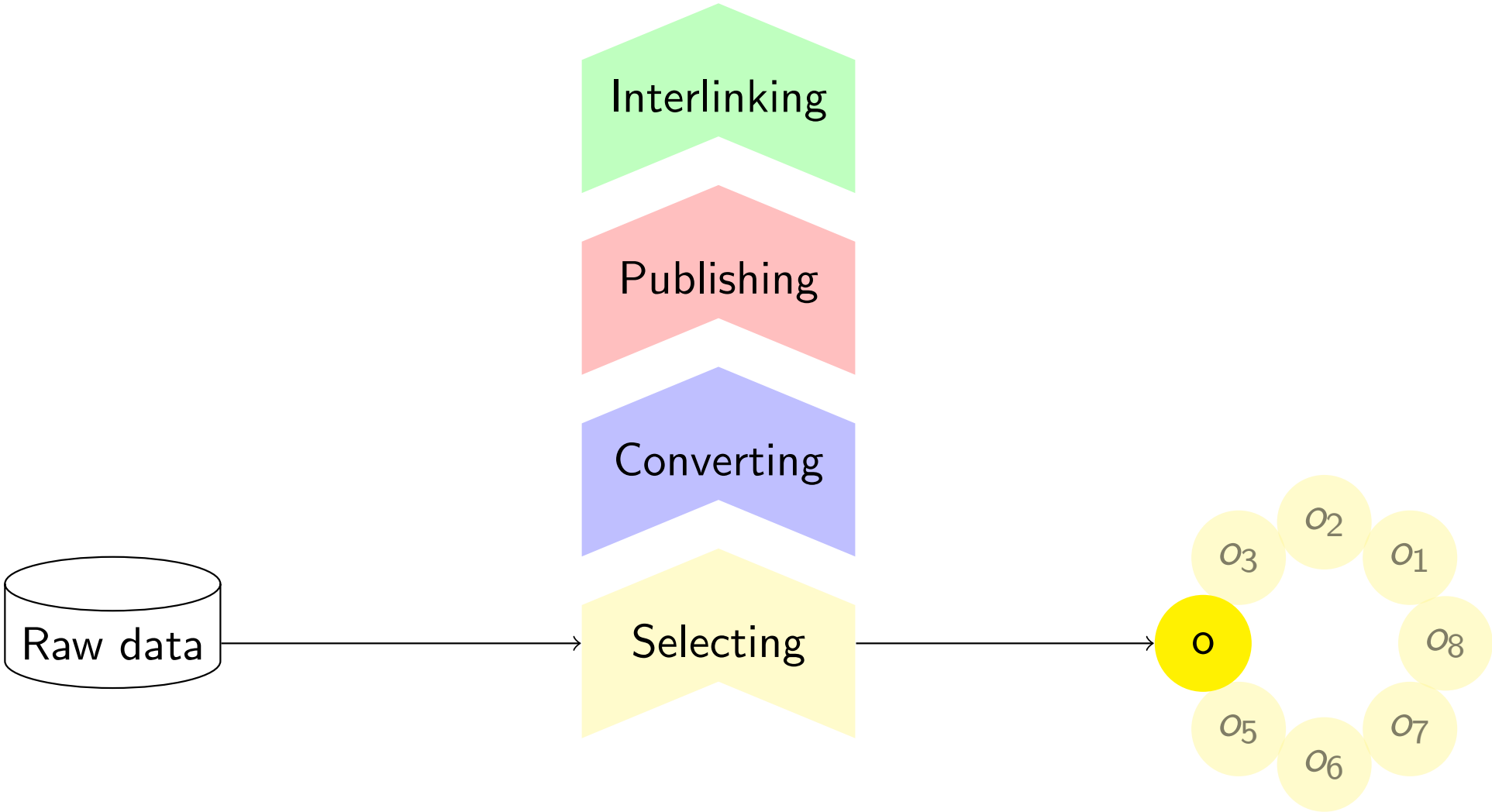
- French ANR project 2010-2013
- Research: INRIA, LIRMM, Eurecom
- Technology companies: ATOS, Mondeca
- Data providers: IGN, INSEE
- FING
- Goal: producing the data elevator turning legacy data sets in properly linked datasets

The Datalift platform

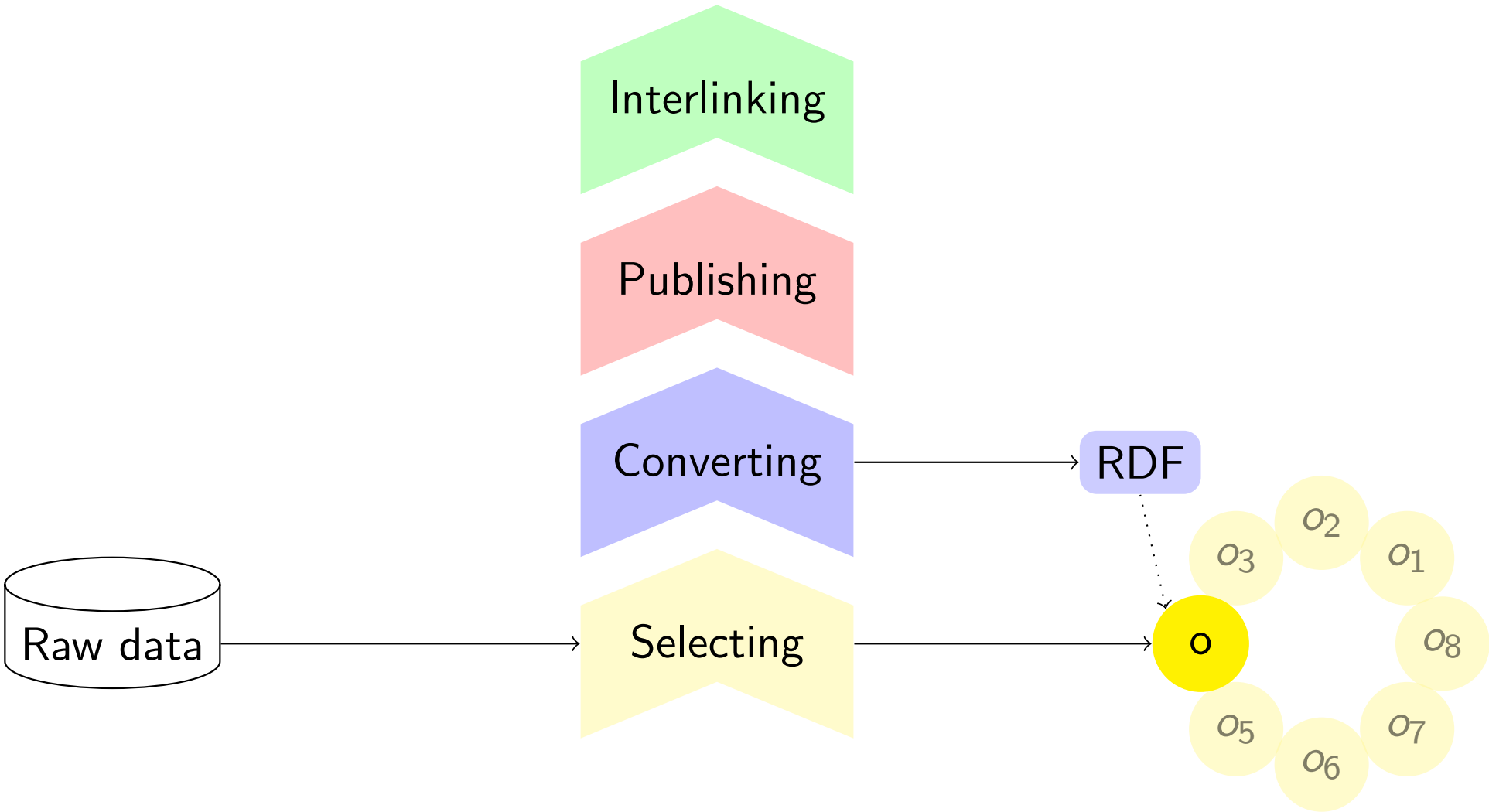
The Datalift platform



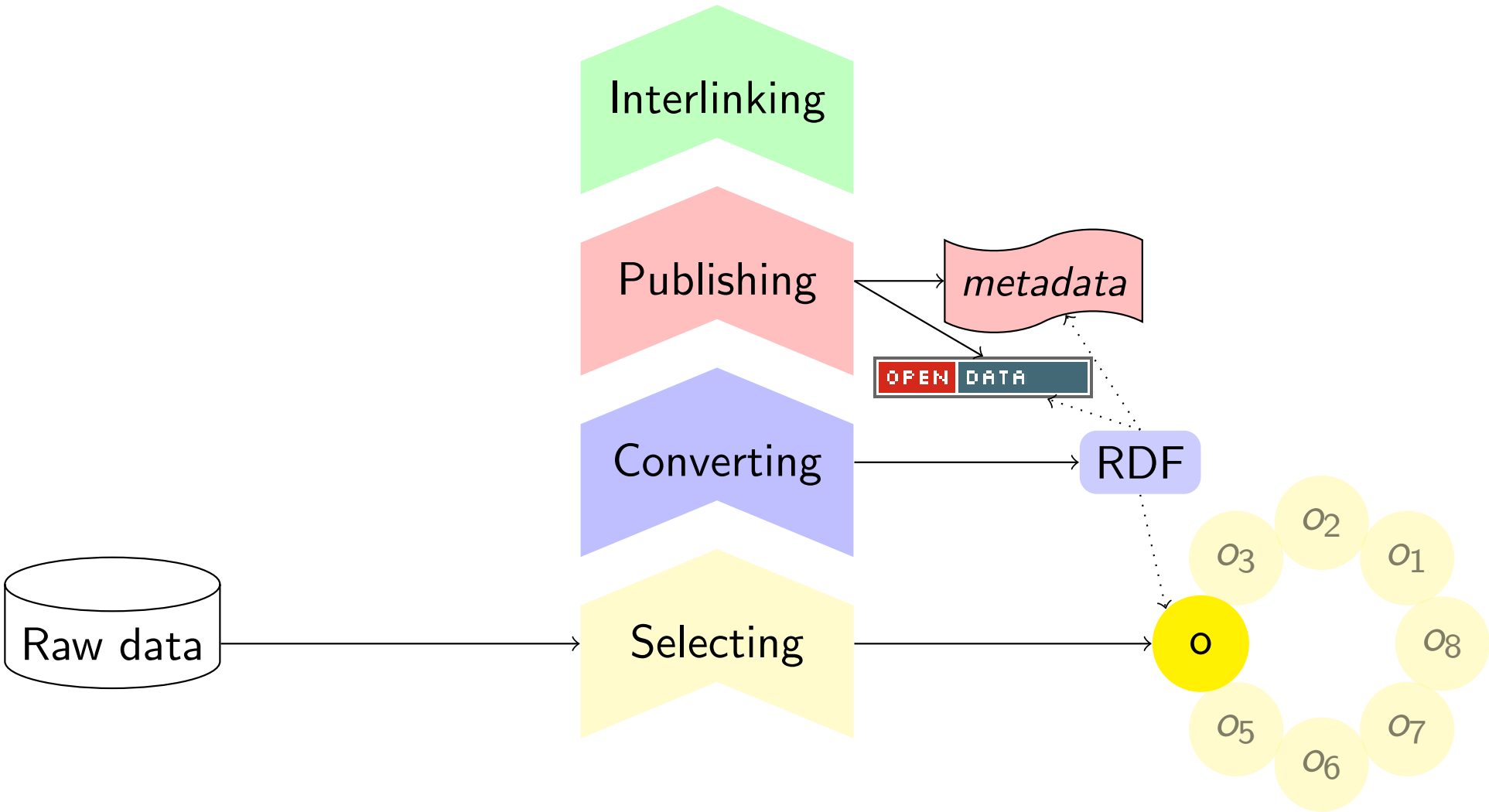
The Datalift platform



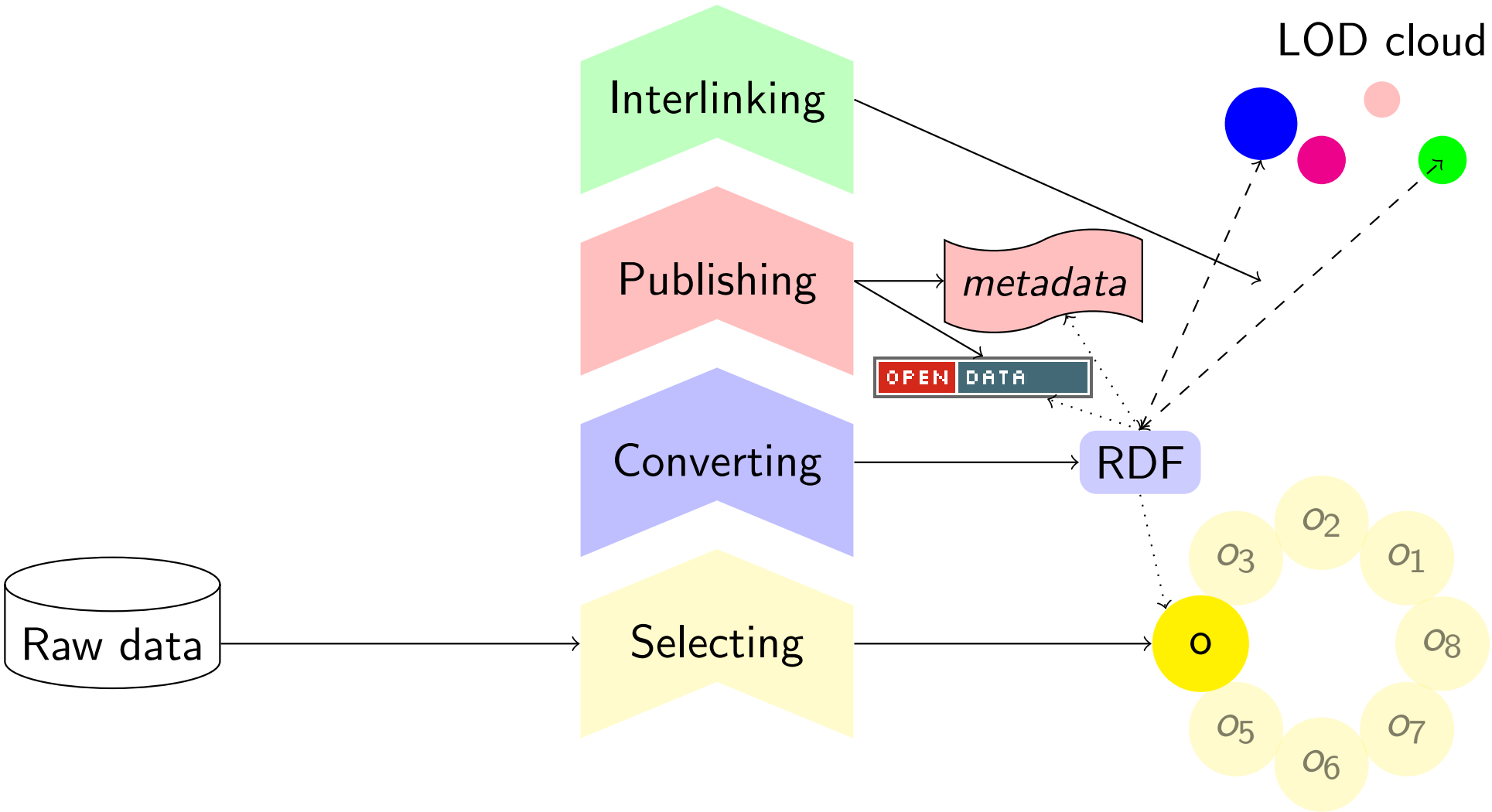
The Datalift platform



The Datalift platform



The Datalift platform



Datalift output

Plate-forme Datalift L'ascenseur pour les données;

LOV L'annuaire des ontologies du lod-cloud (`http://lov.okfn.org`)

Sites de l'INSEE et de l'IGN (`http://data.insee.fr`, `http://data.ign.fr`)

Outline

Interlinking insee data to Eurostat NUTS

This is a typical example of interlinking:

- two datasets with different breadth and depth;
- with internal identifiers.

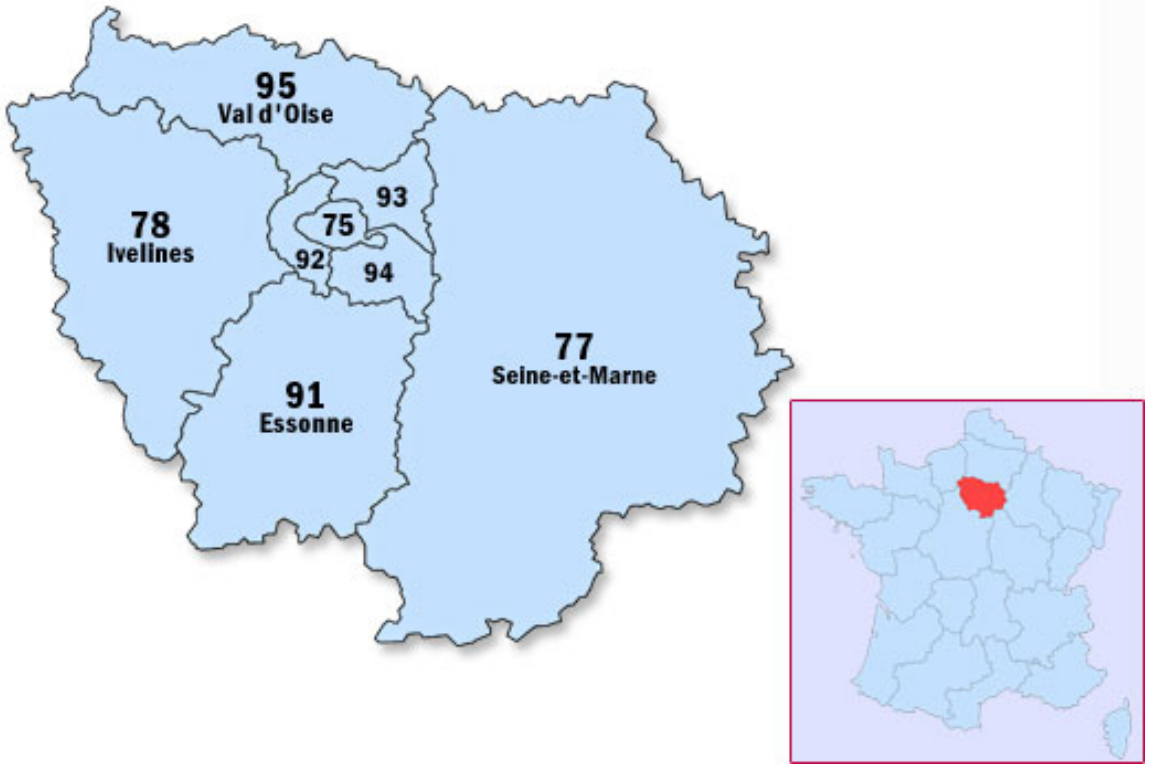
Example: INSEE dataset

Région table:

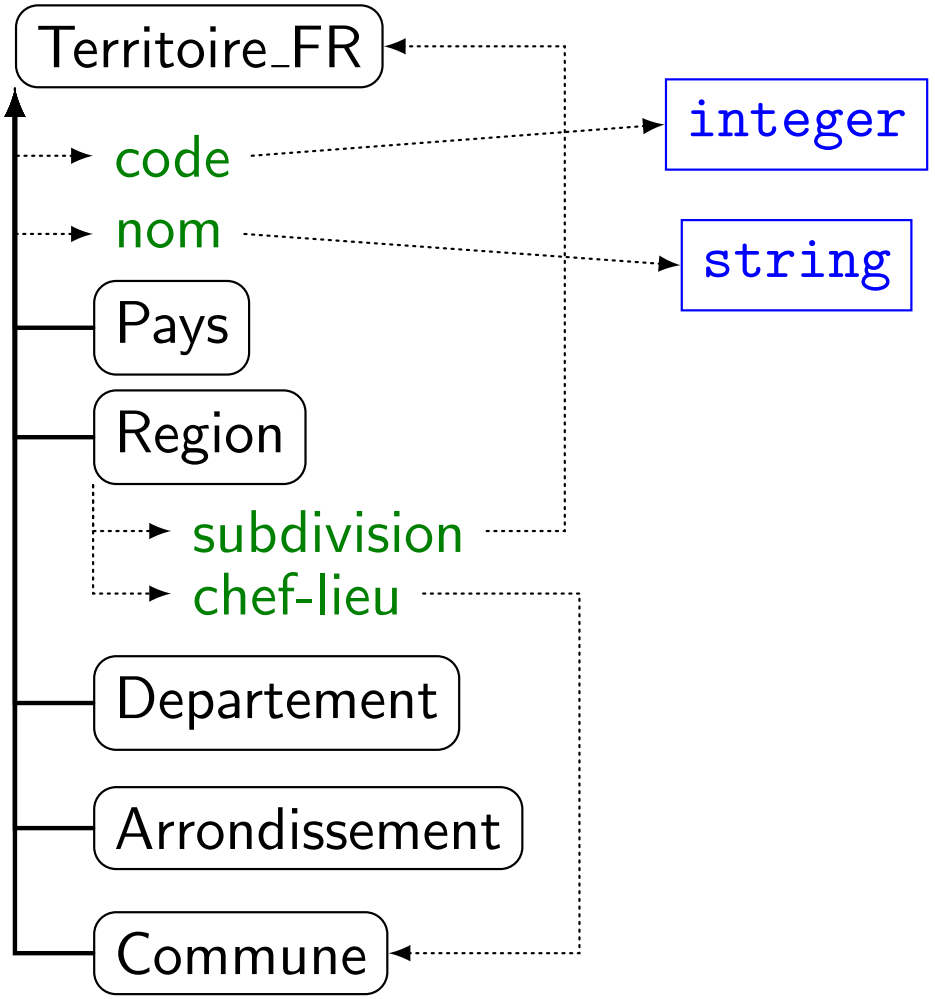
code	nom	chef-lieu
11	Île-de-France	75056
21	Champagne-Ardenne	51108
22	Picardie	80021

Sous-région table:

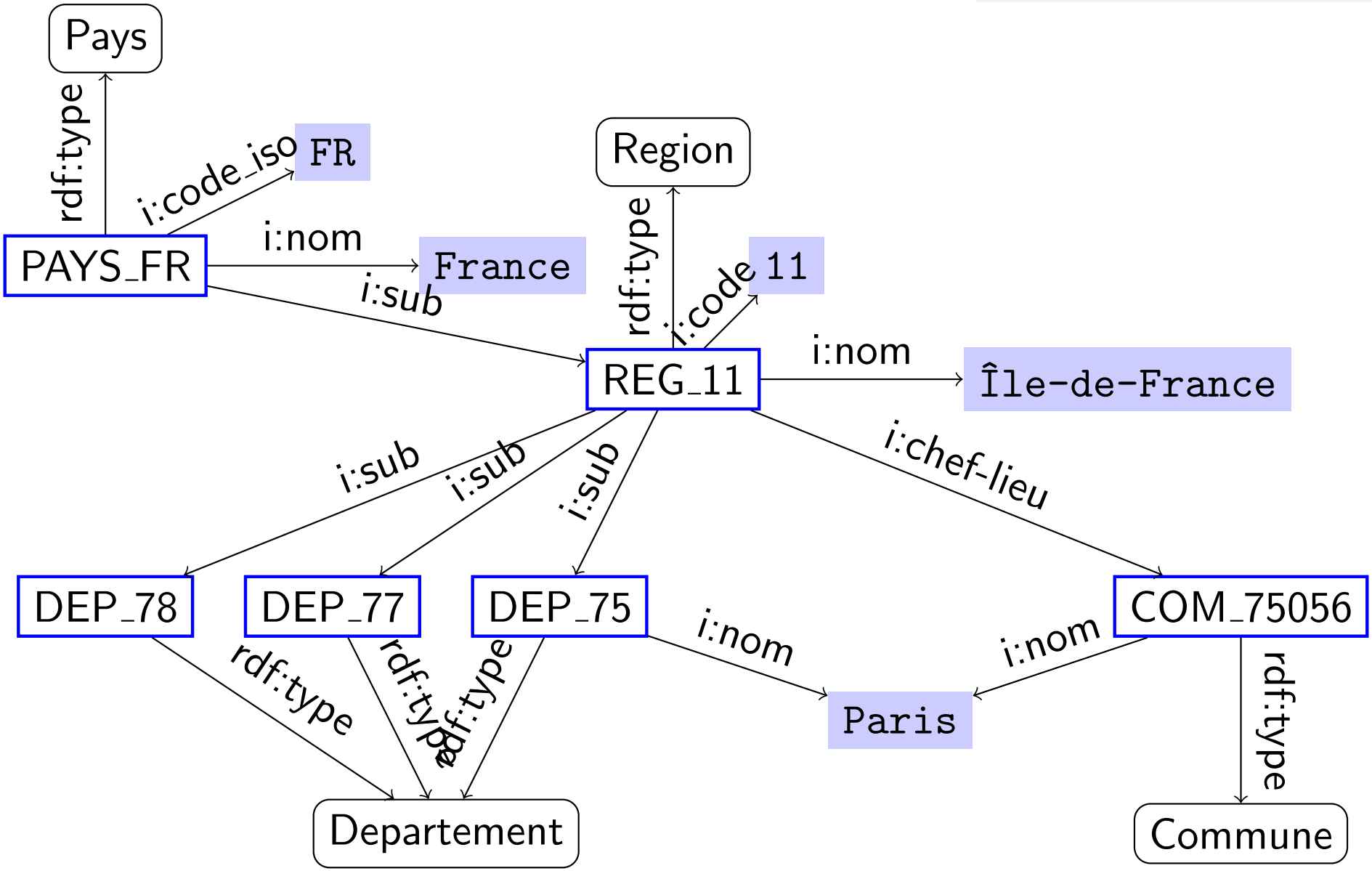
région	département
11	75
11	77
11	78
11	91
11	92
11	93



Example: Administrative ontology



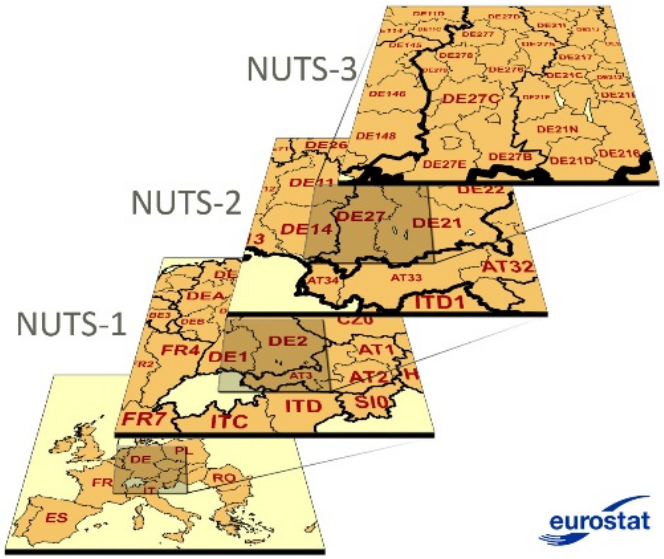
Example: RDF


Pays
rdf:type
i:code_iso
FR
Region
rdf:type
i:code
11
i:nom
PAYS_FR
France
i:sub
REG_11
i:nom
Île-de-France
i:sub
i:sub
i:sub
i:chef-lieu
DEP_78
DEP_77
DEP_75
COM_75056
rdf:type
rdf:type
rdf:type
i:nom
i:nom
rdf:type
Paris
Departement
Commune

Example: NUTS dataset

NUTSRegion table:

level	code	name	hasParentRegion
0	FR	FRANCE	
1	FR1	ÎLE DE FRANCE	FR
2	FR10	Île de France	FR1
3	FR101	Paris	FR10
3	FR104	Essonne	FR10



Example: Linking INSEE and NUTS

NUTS: Nomenclature of territorial units for statistics

#INSEE	INSEE name	NUTS Level	#NUTS
1	Pays	0	34
		1	142
26	Région	2	344
100	Département	3	1488
342	Arrondissement		
4036	Canton	4	
52422	Commune	5	

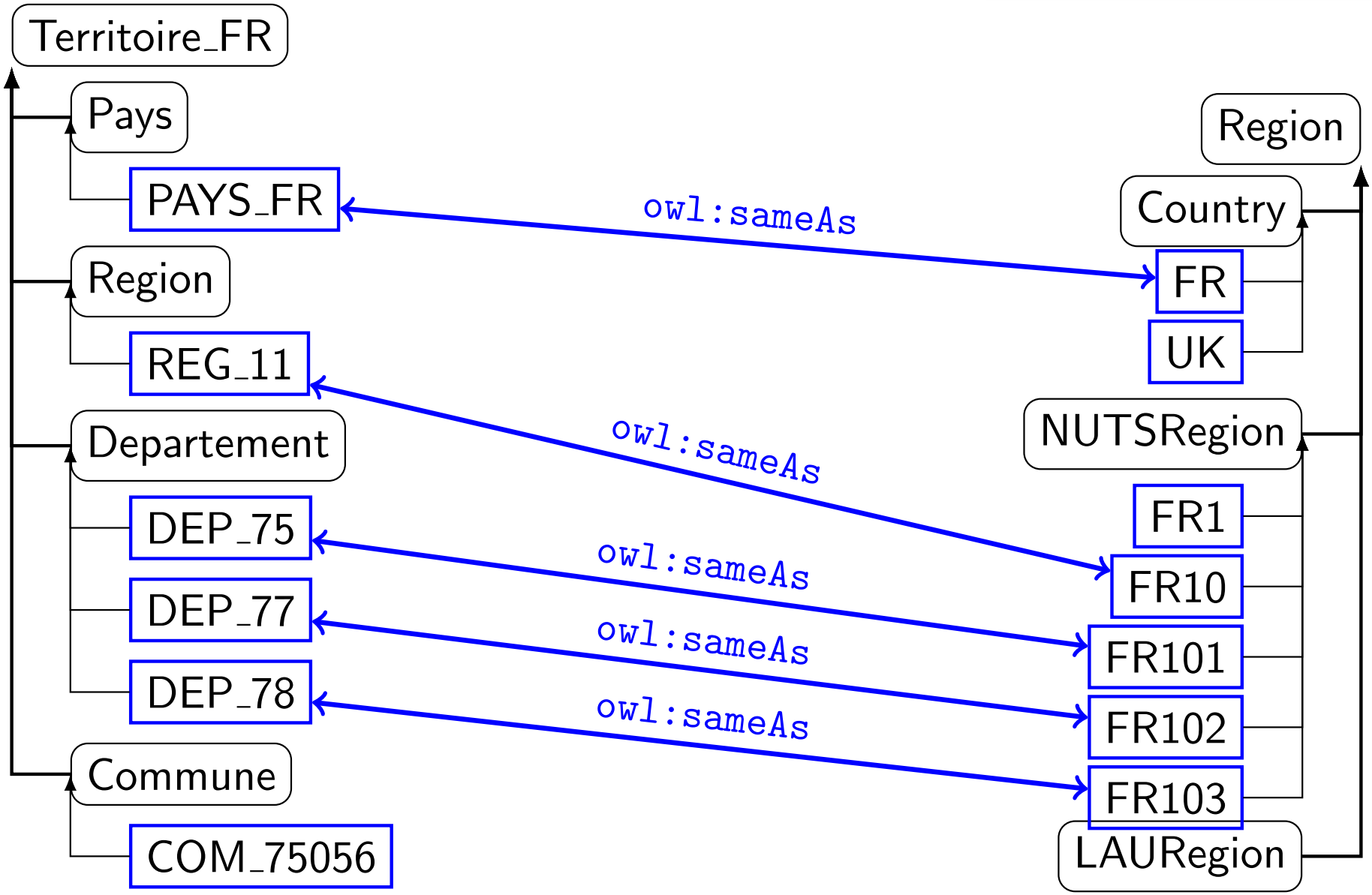
Example: Linking INSEE and NUTS

NUTS: Nomenclature of territorial units for statistics

#INSEE	INSEE name	NUTS Level	#NUTS
1	Pays	0	34
		1	142
26	Région	2	344
100	Département	3	1488
342	Arrondissement		
4036	Canton	4	
52422	Commune	5	

Å vs. Saint-Rémy-en-Bouzemont-Saint-Genest-et-Isson or Montbonnot Saint-Martin

Example: Linking INSEE and NUTS



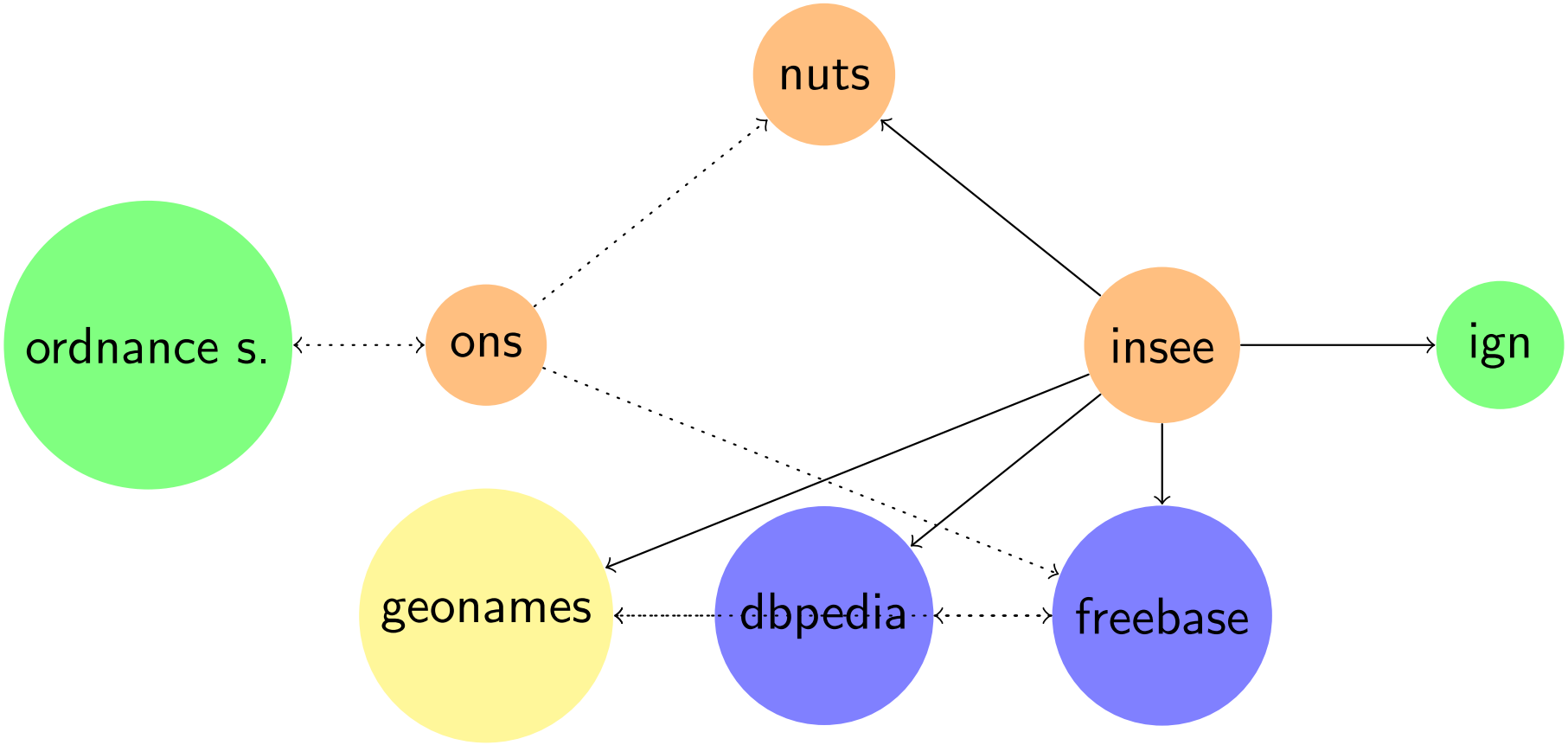
Example: Linksets

Specific data sets containing URIs.

```
<http://www.example.org/linkset/INSEE-NUTS>
 a void:Linkset ;
 void:target <http://rdf.insee.fr/geo/regions-2011.rdf>;
 void:target <http://nuts.psi.enakting.org/id/>;
```

```
insee:PAYS_FR owl:sameAs nuts:FR
insee:REG_11 owl:sameAs nuts:FR10
insee:DEP_75 owl:sameAs nuts:FR101
insee:DEP_77 owl:sameAs nuts:FR102
insee:DEP_78 owl:sameAs nuts:FR103
```

Example: interesting sets



Outline

Example Silk

- Déclarer les sources de données (DataSource);
- Circonscrire les entités à comparer (Source/TargetDataset);
- Décrire comment elles seront comparées (LinkageRule):
  - Sélectionner les propriétés à comparer à l'aide de chemins (Input);
  - Calculer une distance entre eux (Compare+seuil);
  - Aggréger les différentes comparaisons (Aggregate);
- Sélectionner les paires d'entités à lier (Filter);
- Engendrer les liens (Output+seuil).

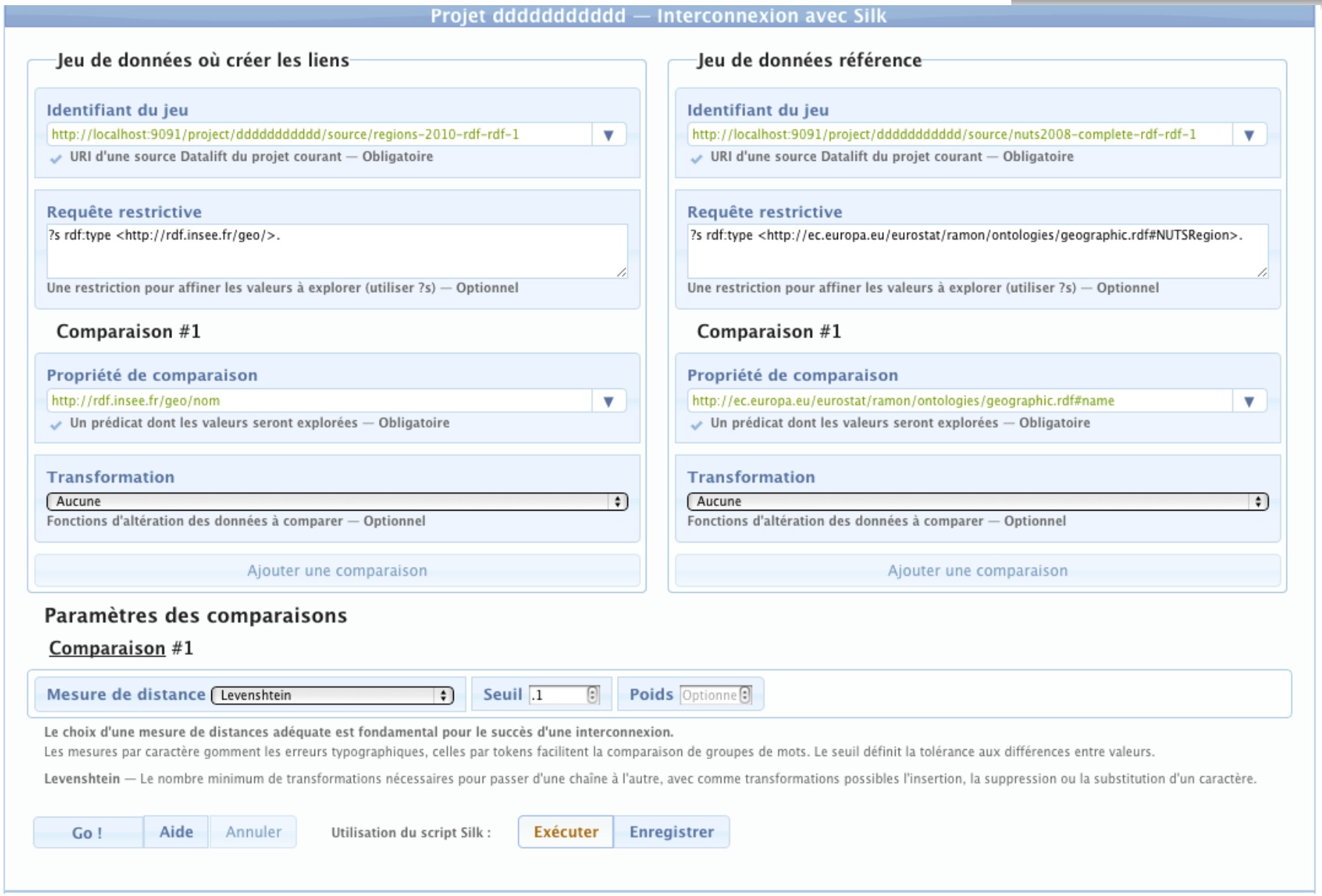
Silk script

Consider a linking script between INSEE and NUTS:

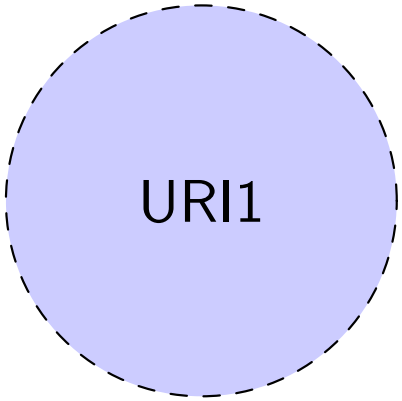
```
<Silk>
  <Prefix id="nuts"
      namespace="http://ec.europa.eu/.../geographic.rdf#" />
  <Prefix id="insee"
      namespace="http://rdf.insee.fr/geo/" />

  <DataSource id="nuts2008"
    type="sparqlEndpoint">
    <Param name="endpointURI"
     value="http://localhost:9091/.../internal"/>
    <Param name="graph"
     value="http://localhost:9091/.../nuts2008-complete-1"/>
  </DataSource>
  <DataSource id="insee2010"
    type="sparqlEndpoint">
     <Param name="endpointURI"
     value="http://localhost:9091/.../internal"/>
    <Param name="graph"
     value="http://localhost:9091/.../source/regions-2010-1"/>
  </DataSource>

  <Thresholds accept="0.9" verify="0.7" />
  <Outputs>
     <Output type="sparul">
      <Param name="graphUri"
          value="http://localhost:9091/.../source/insee-nuts-silk"/>
      <Param name="uri"
          value="http://localhost:9091/.../lifted/"/>
      <Param name="parameter" value="update"/>
    </Output>
```

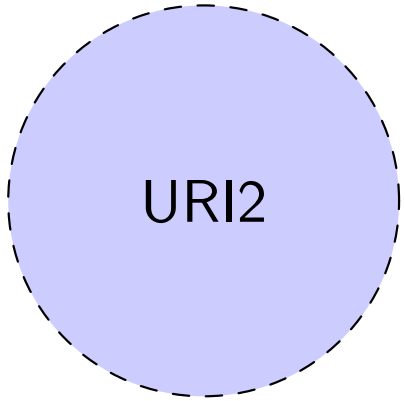
```
<Interlinks>
  <Interlink id="linkingNUTS">
    <LinkType>owl:sameAs</LinkType>

    <SourceDataset dataSource="nuts2008" var="s">
      <RestrictTo>?s rdf:type nuts:NUTSRegion.
        ?s nuts:level 2.
      </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource="insee2010" var="ss">
      <RestrictTo>?ss rdf:type insee:Region</RestrictTo
    </TargetDataset>

    <LinkageRule>
      <Aggregate type="max">
        <Compare metric="levenshteinDistance"
            threshold=".2">
          <Input path="?s/nuts:name"/>
          <Input path="?ss/insee:nom"/>
        </Compare>
      </Aggregate>
    </LinkageRule>

  </Interlinks>
 </Interlink>
</Silk>
```

Interface d'interconnexion

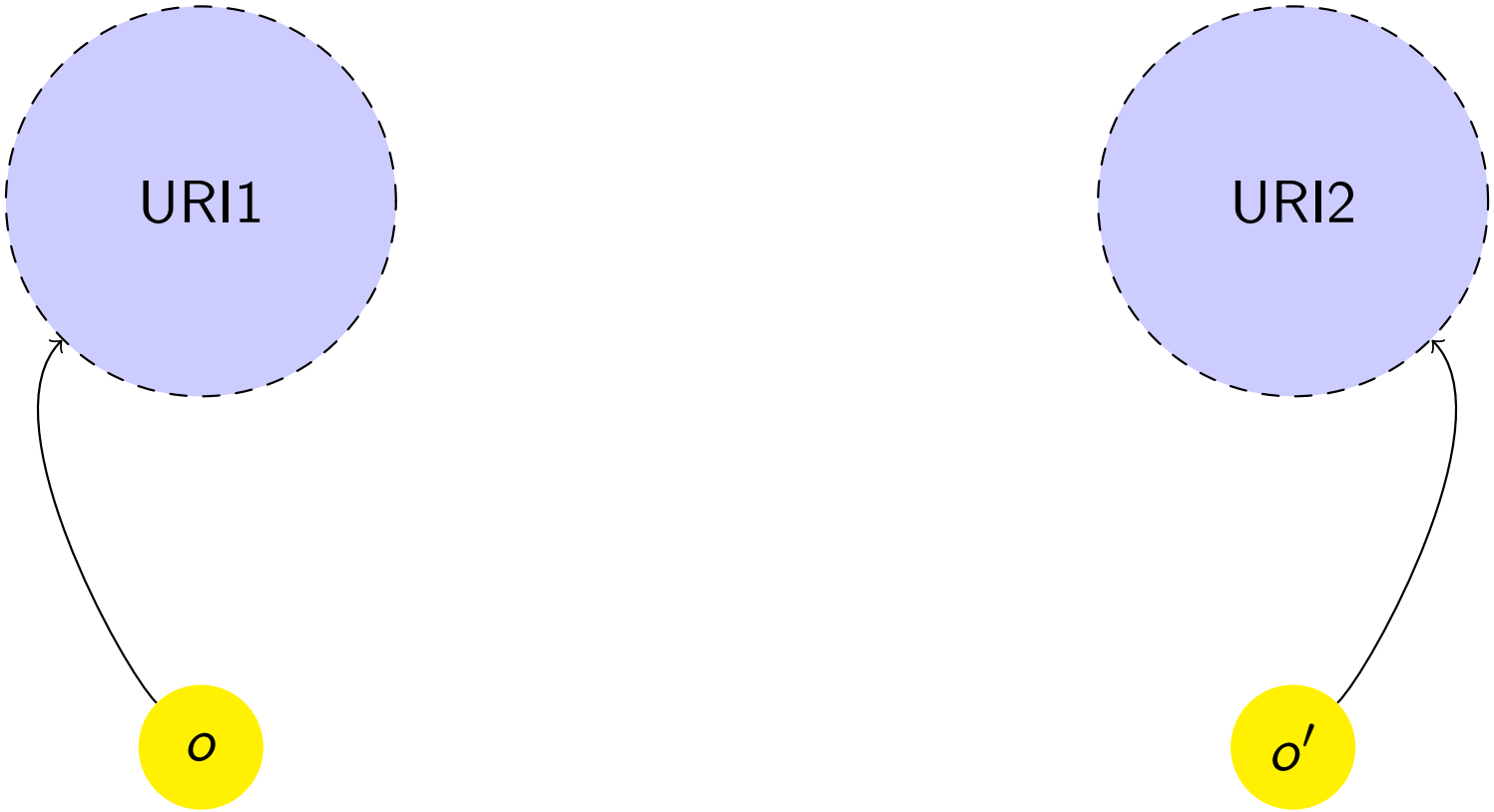


The data interlinking problem

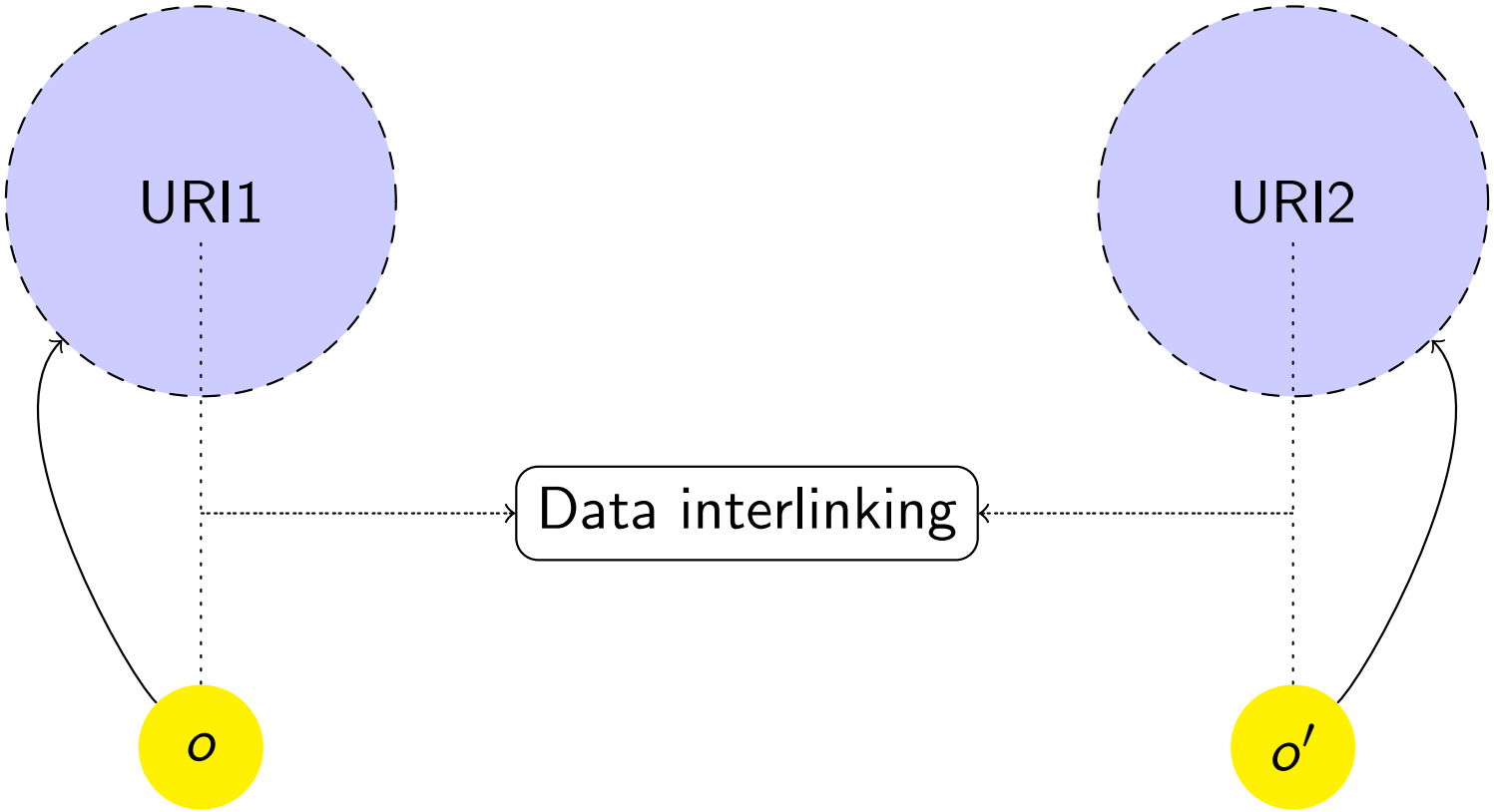




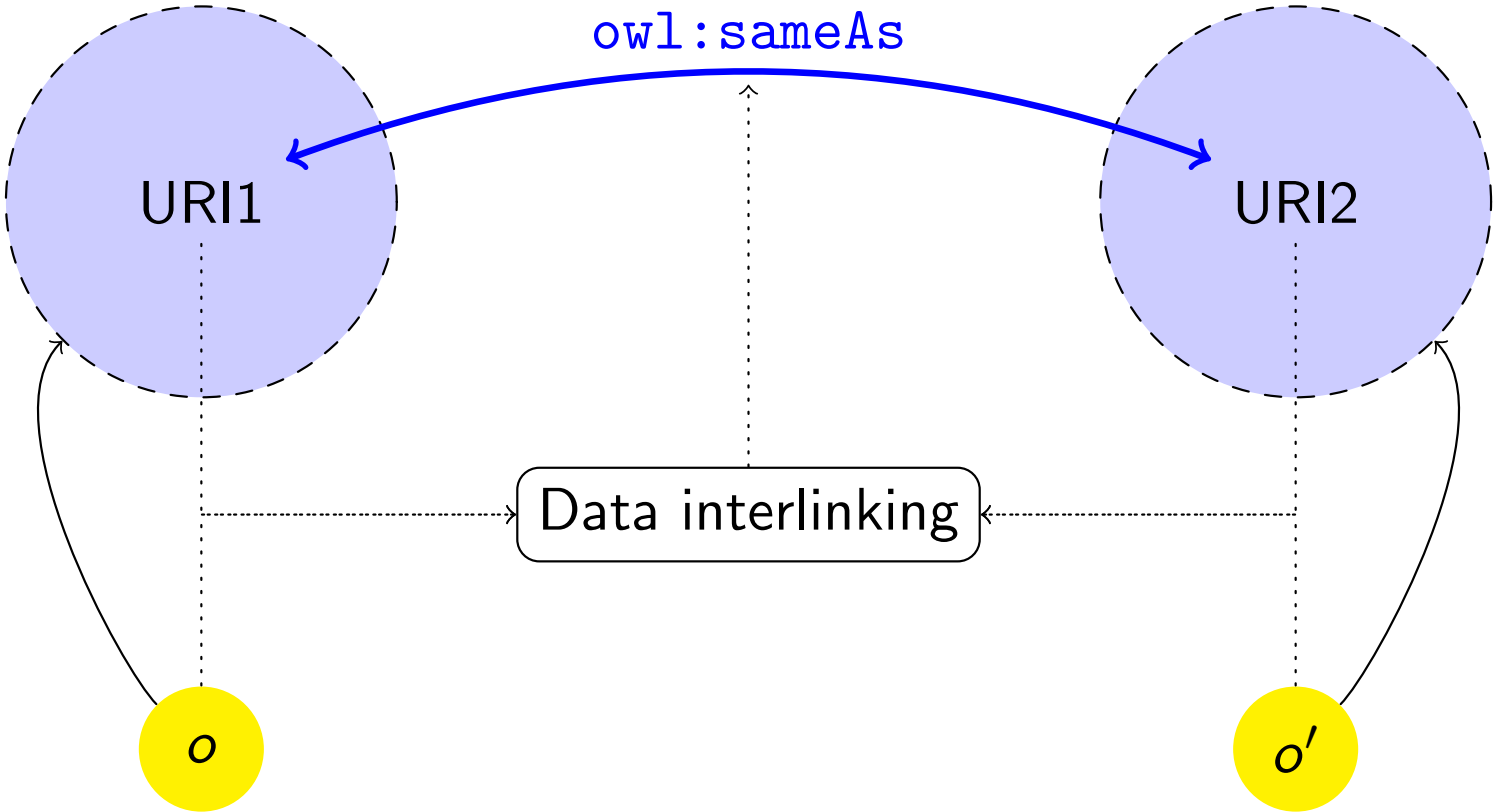
The data interlinking problem



The data interlinking problem



The data interlinking problem

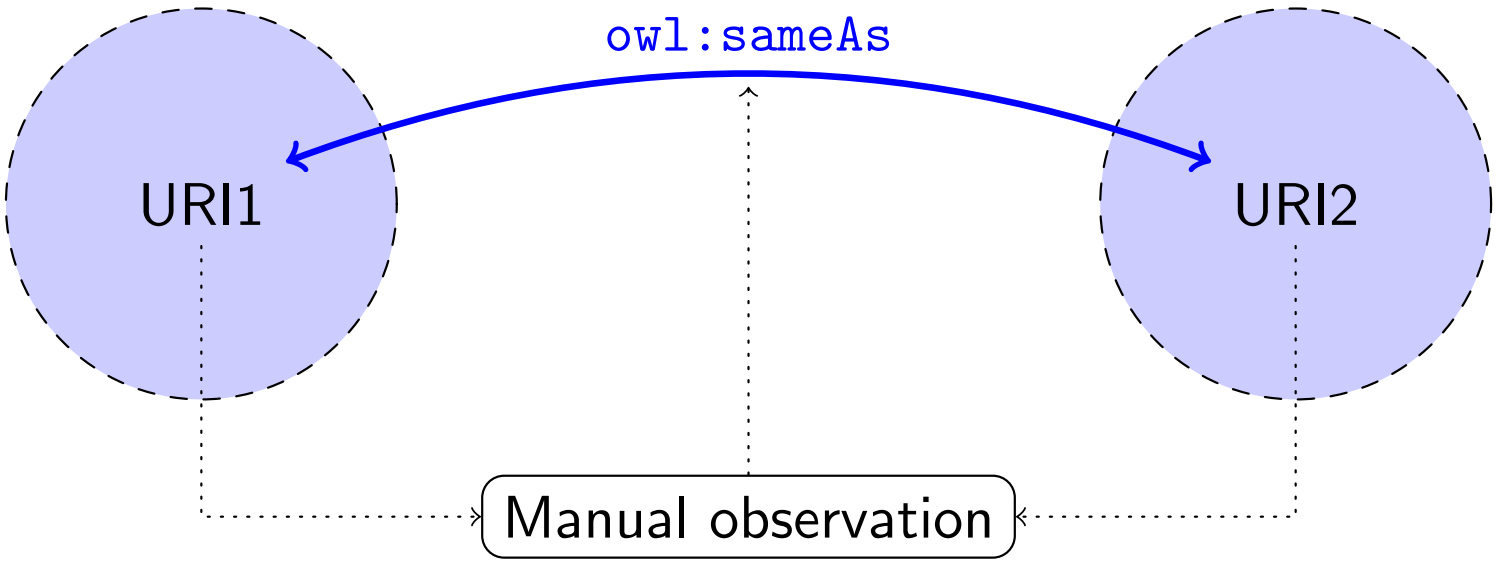


Classification

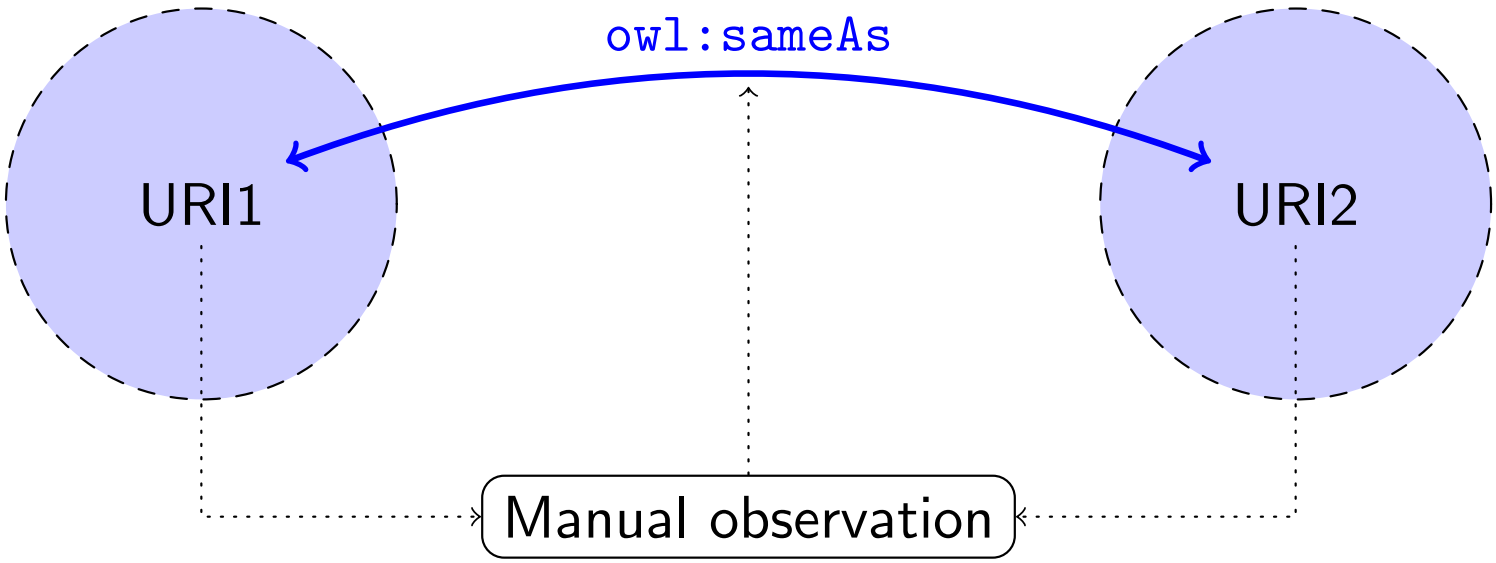
Data interlinking techniques may be based on:

- Data ID (URIs);
- Data keys
- Data content
- External relations (links)
- Common ontologies
- Ontology alignments

Manual resource matching

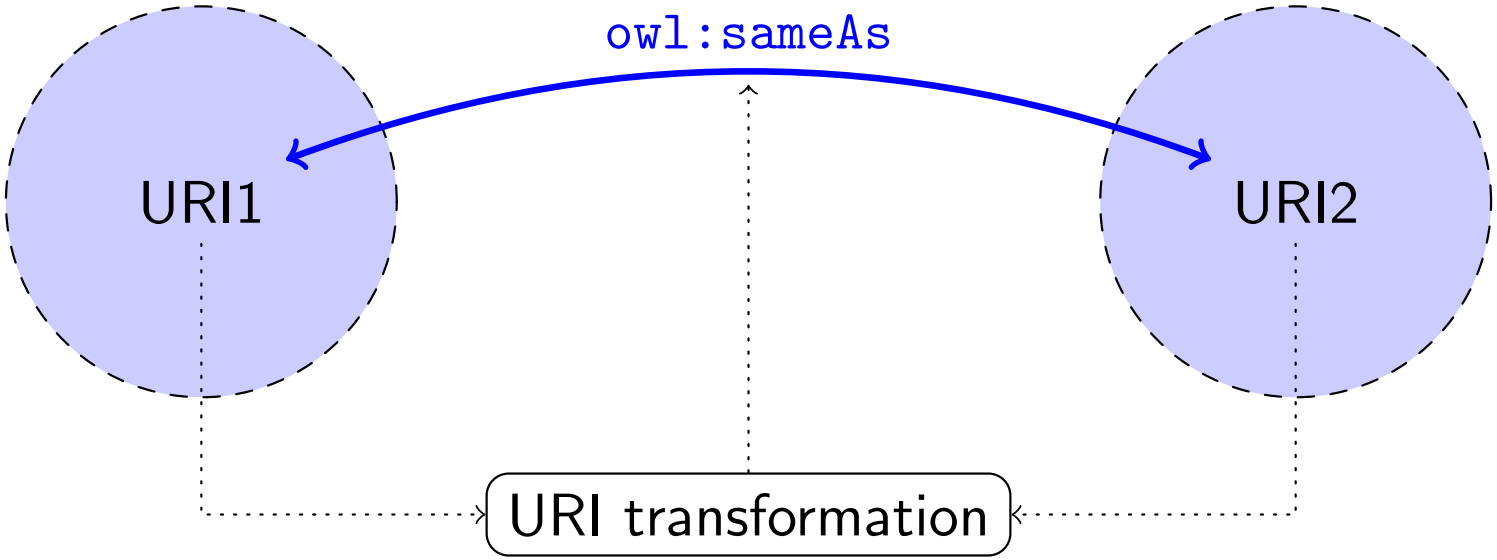


Manual resource matching

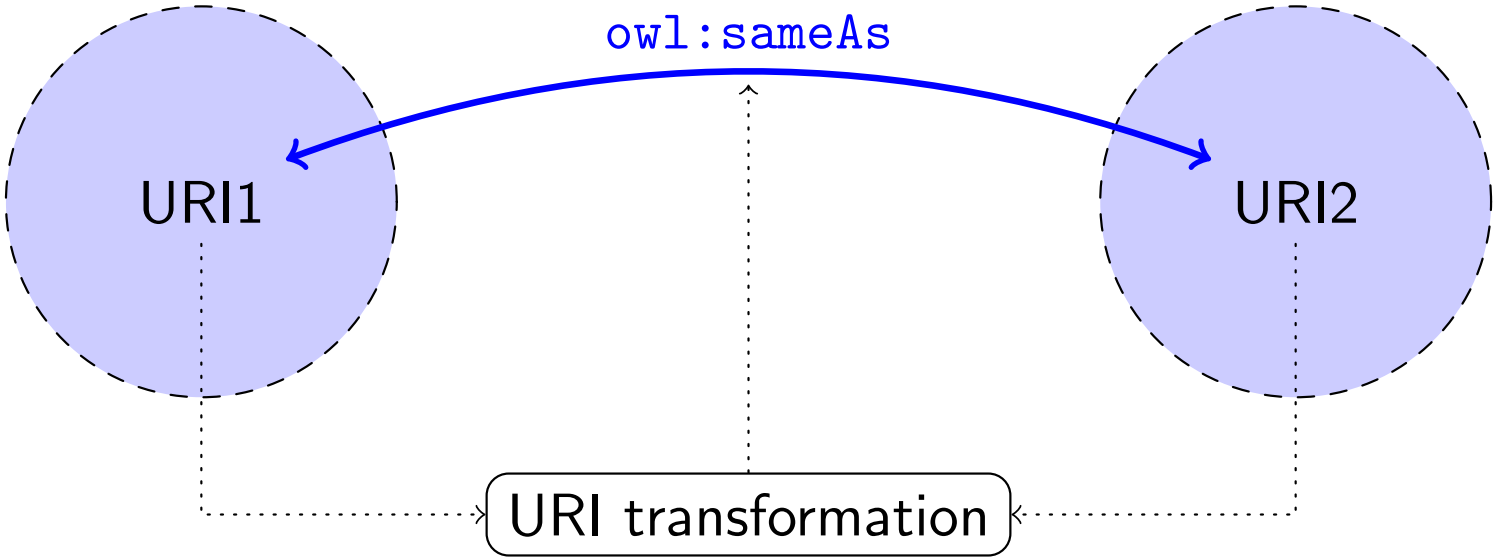


This does not scale.

URI matching

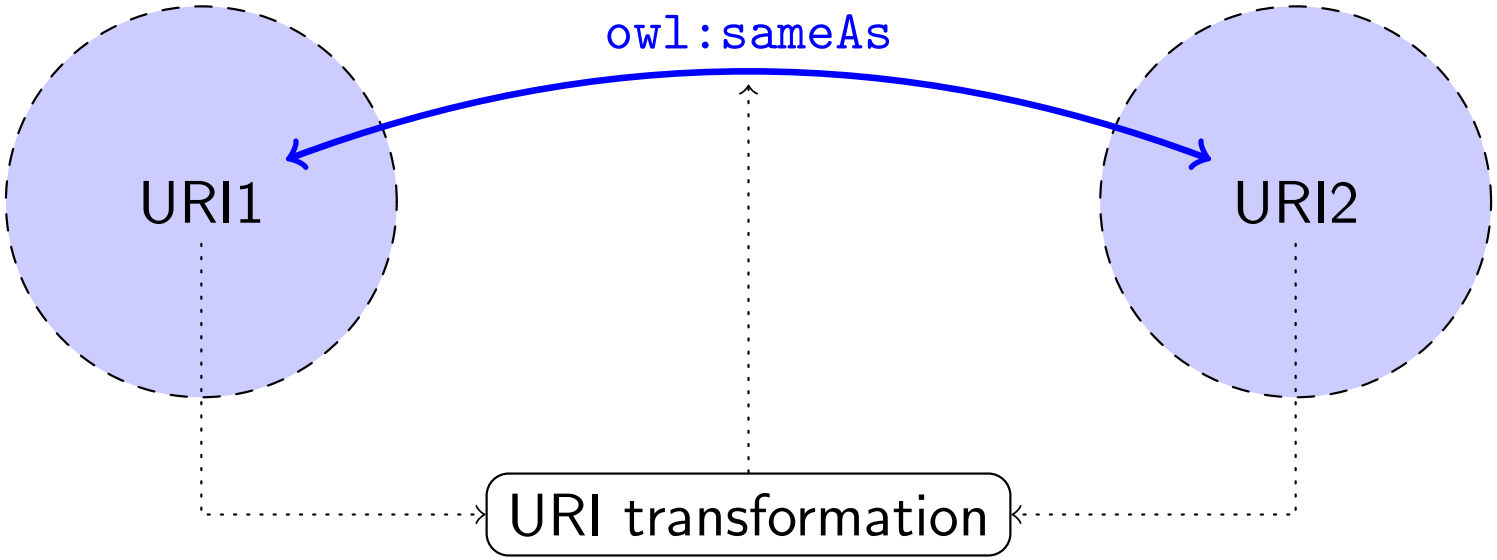


URI matching



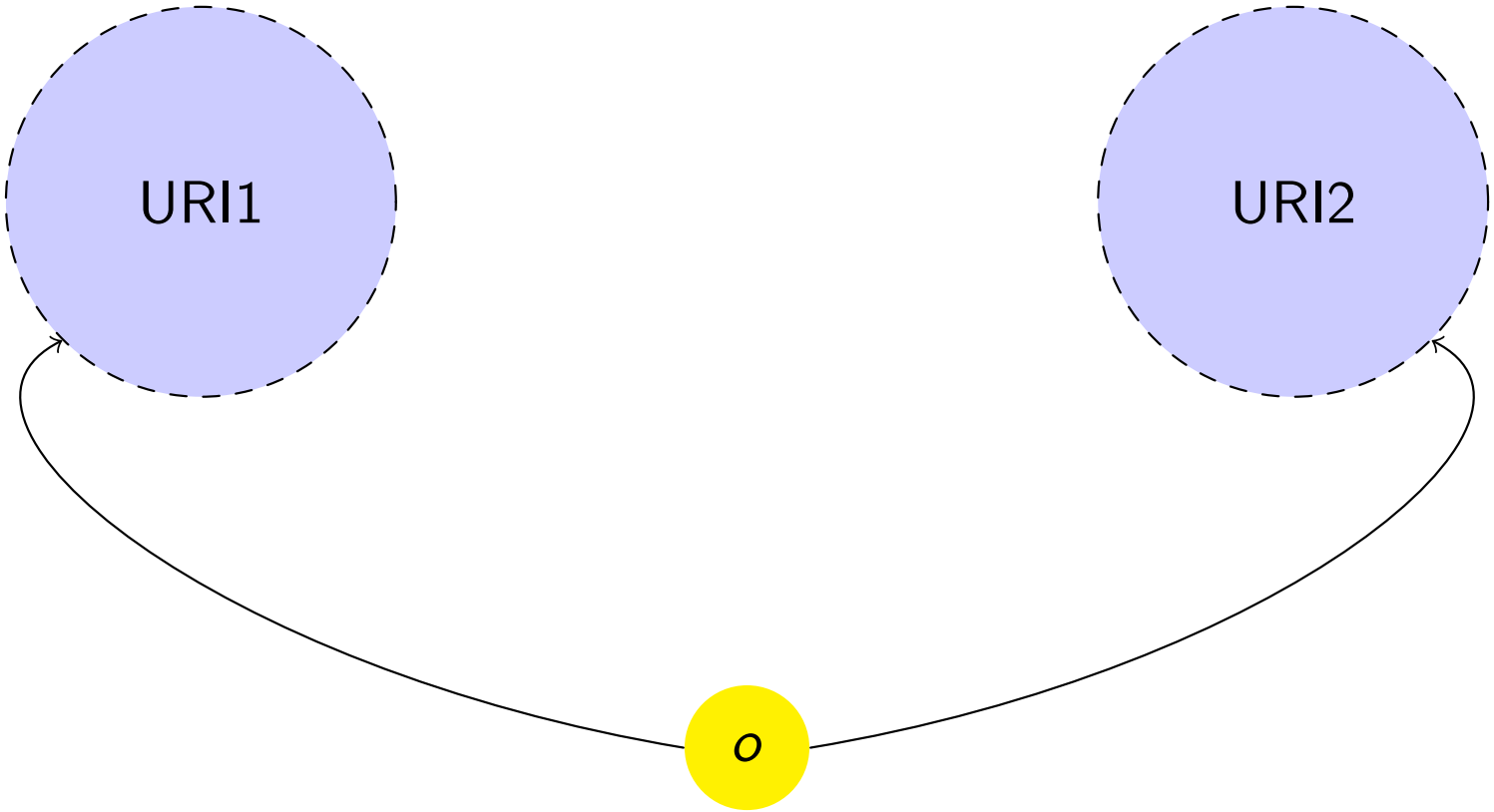

```
http://dbpedia.org/resource/Johann_Sebastian_Bach owl:sameAs
    http://www.lastfm.fr/music/Johann+Sebastian+Bach
```

URI matching

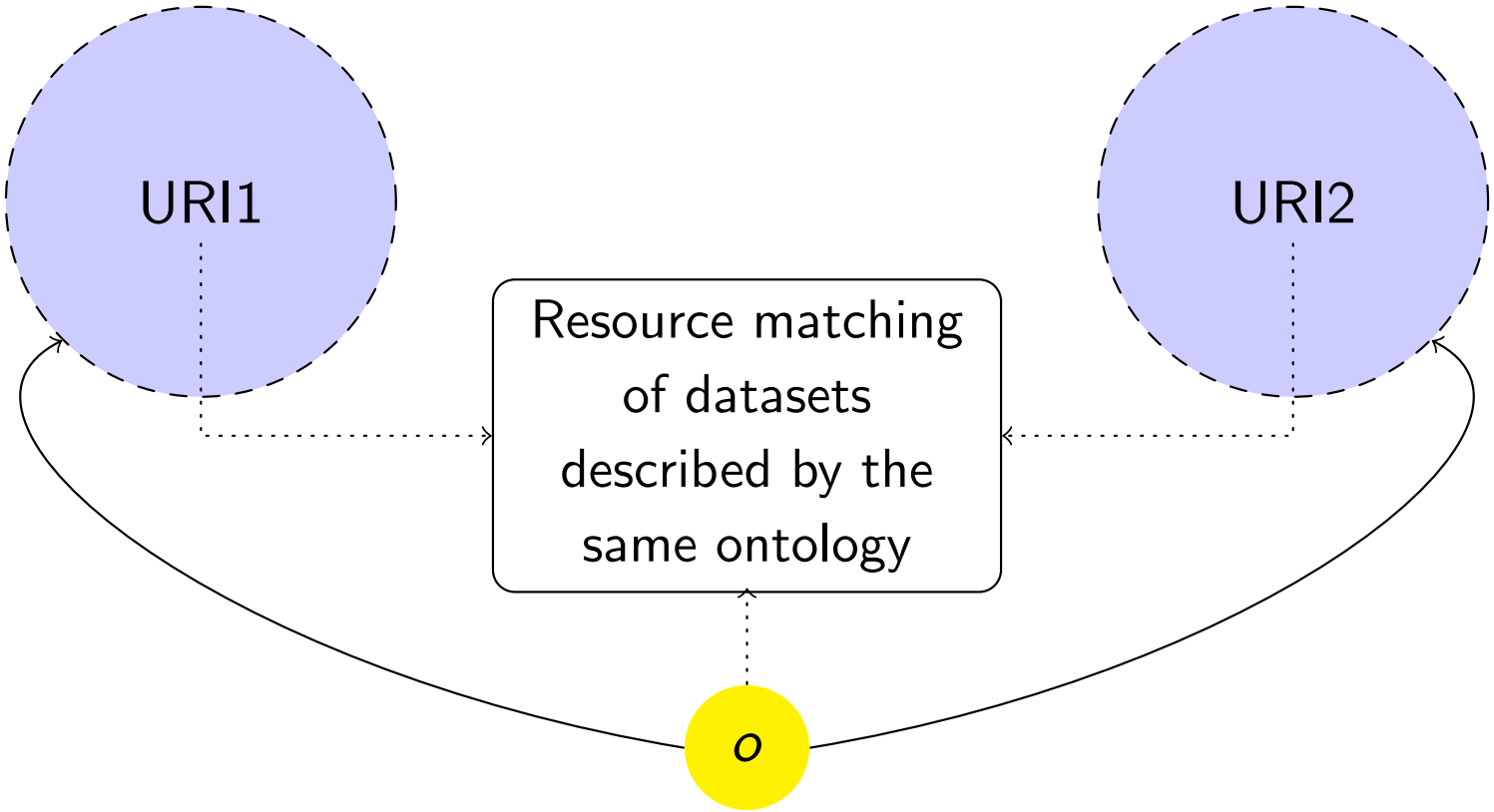



```
http://dbpedia.org/resource/Johann_Sebastian_Bach owl:sameAs
    http://www.lastfm.fr/music/Johann+Sebastian+Bach

http://rdf.insee.fr/geo/regions-2011.rdf#REG_11 ?
    http://ec.europa.eu/eurostat/ramon/rdfdata/nuts2008/FR10
```

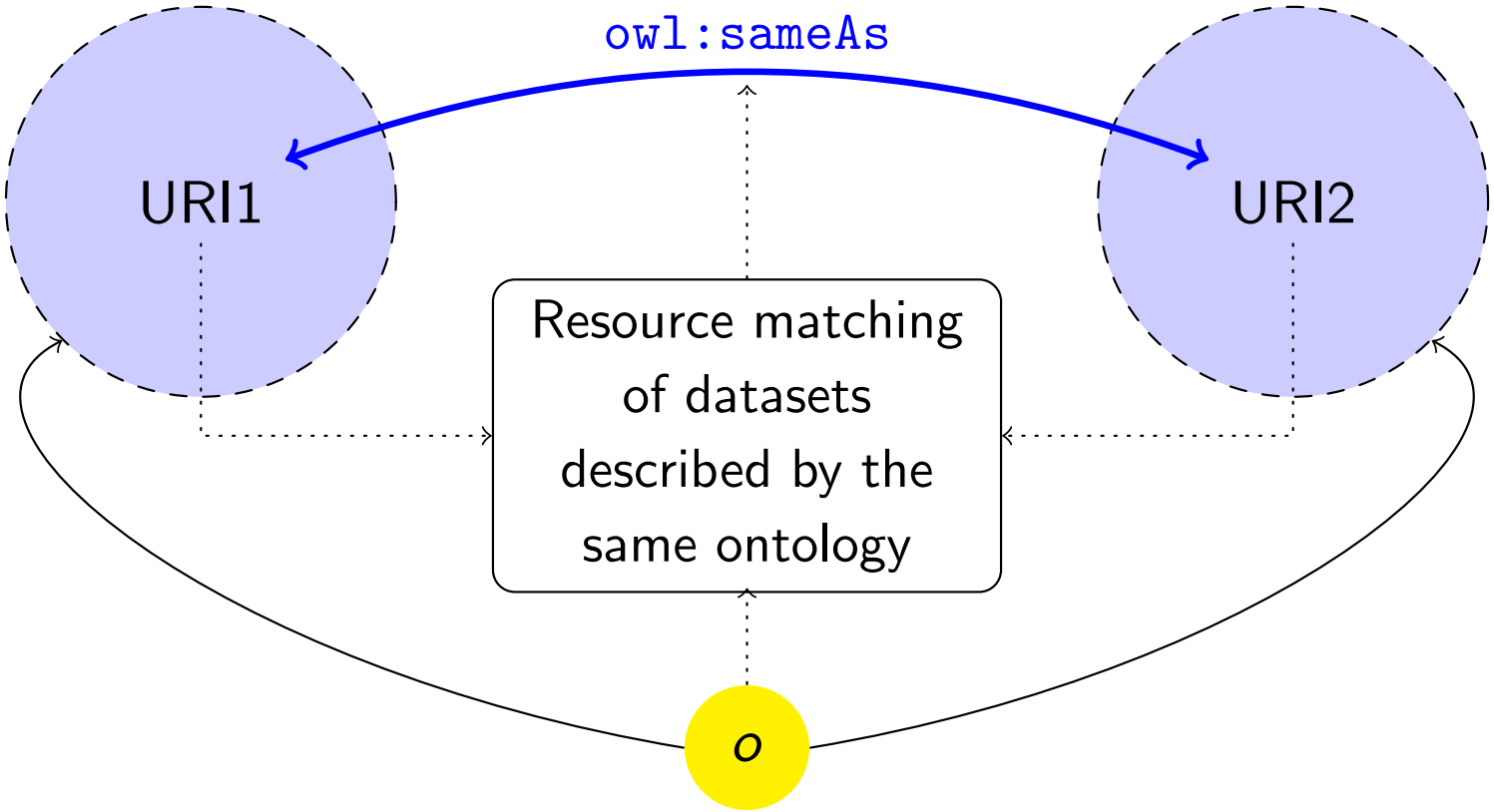
Data matching through a common ontology



Data matching through a common ontology



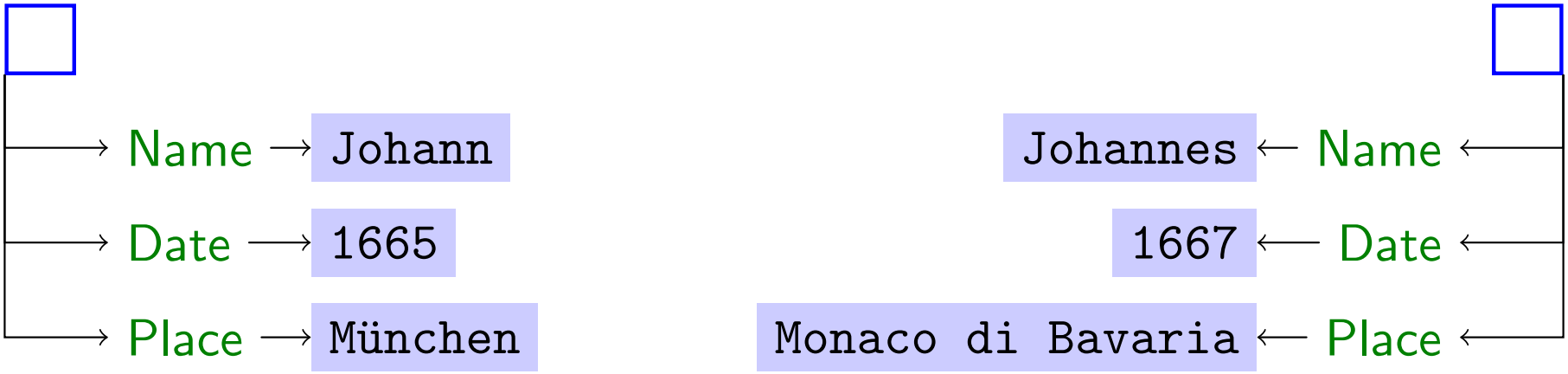
Data matching through a common ontology



Matching with a common ontology

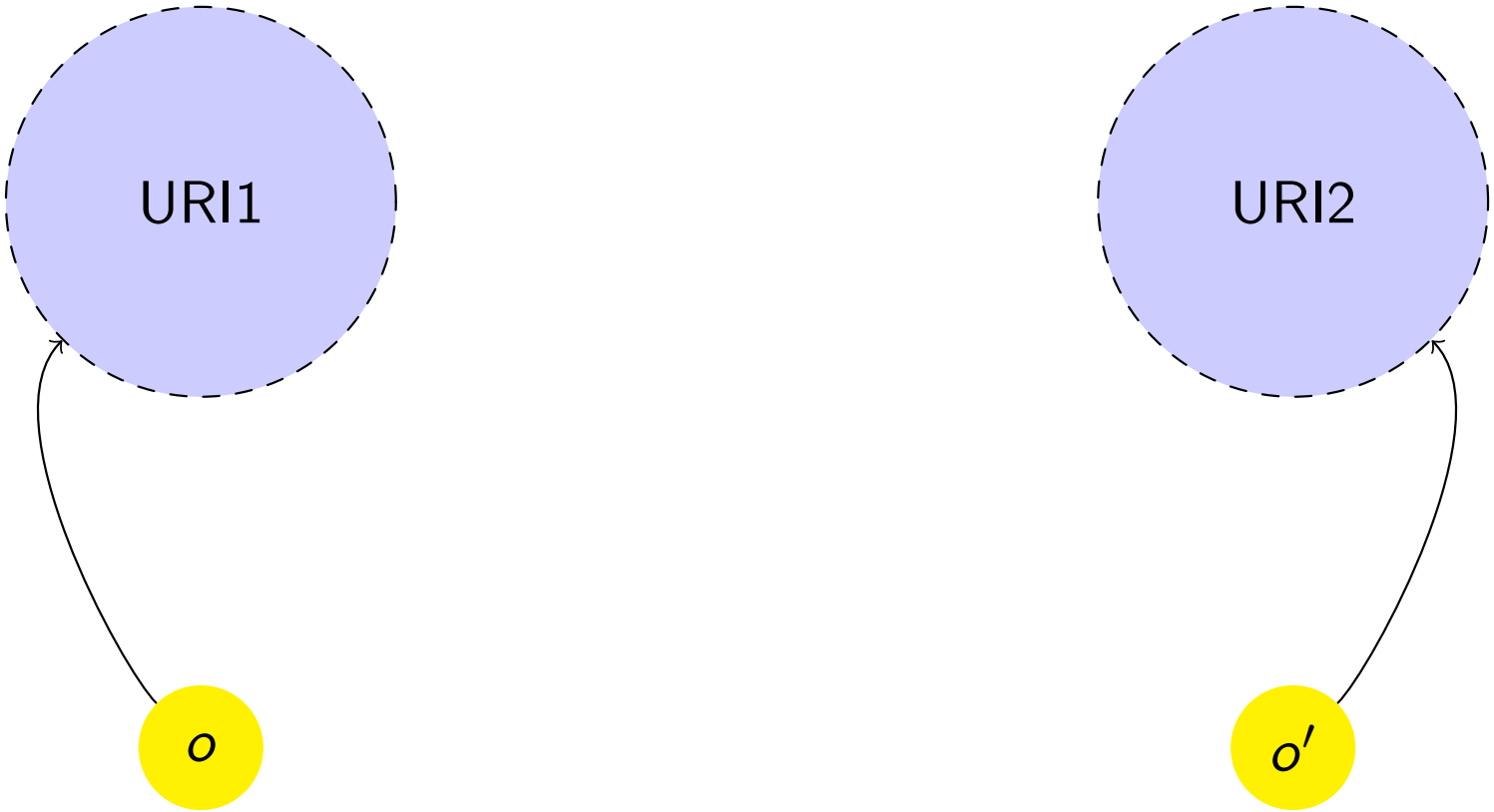
+ Focus the search: only match instances of the same class;
– Not sufficient: it remains to identify corresponding entities
  + If keys are defined (OWL 2), this is done;
  + At least we know which properties to compare;
  – Inferring secondary keys may be useful;
  – Correcting discrepancies: record linkage.

Record linkage

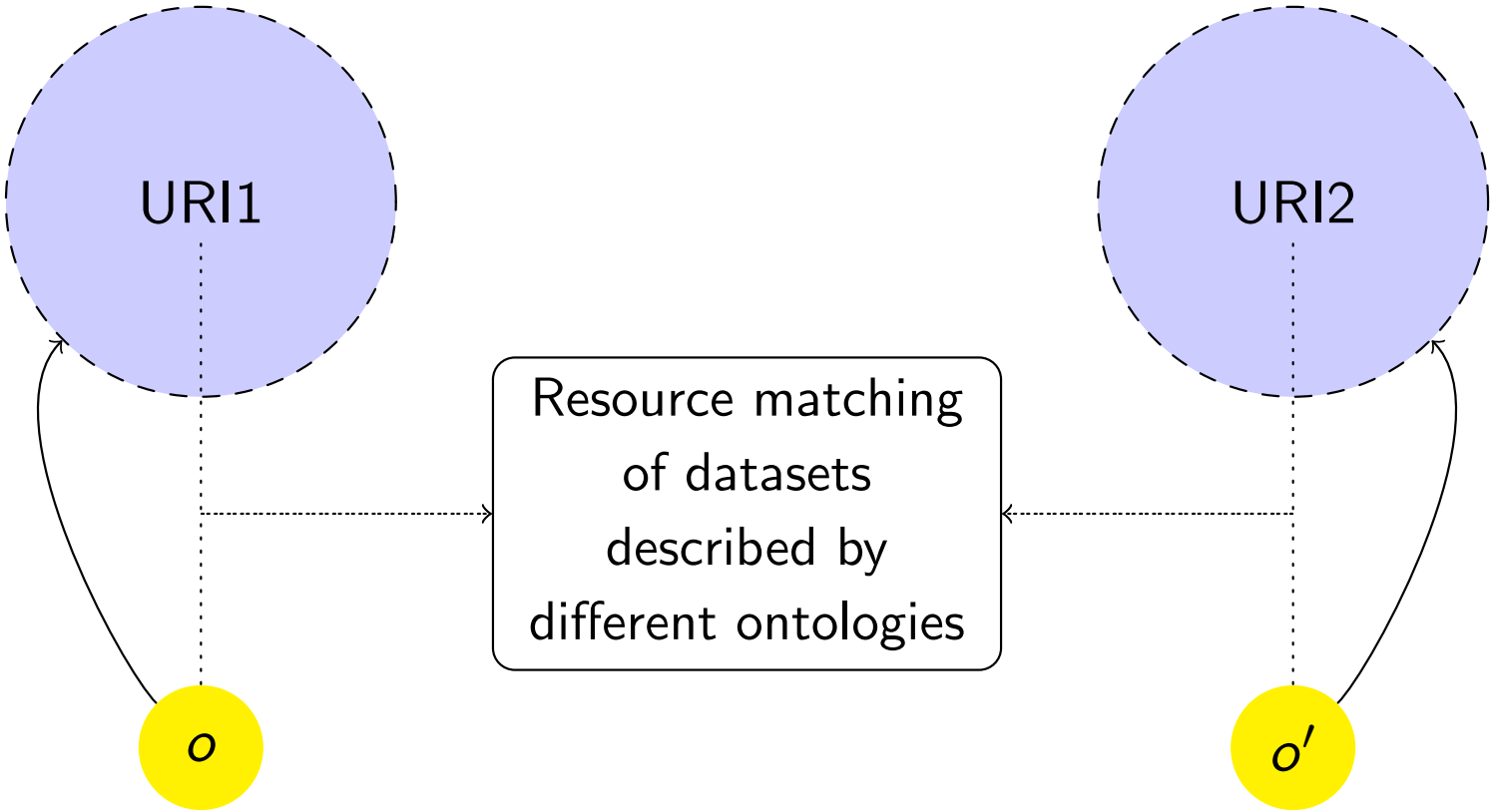


Having a common ontology does not solve the problem.

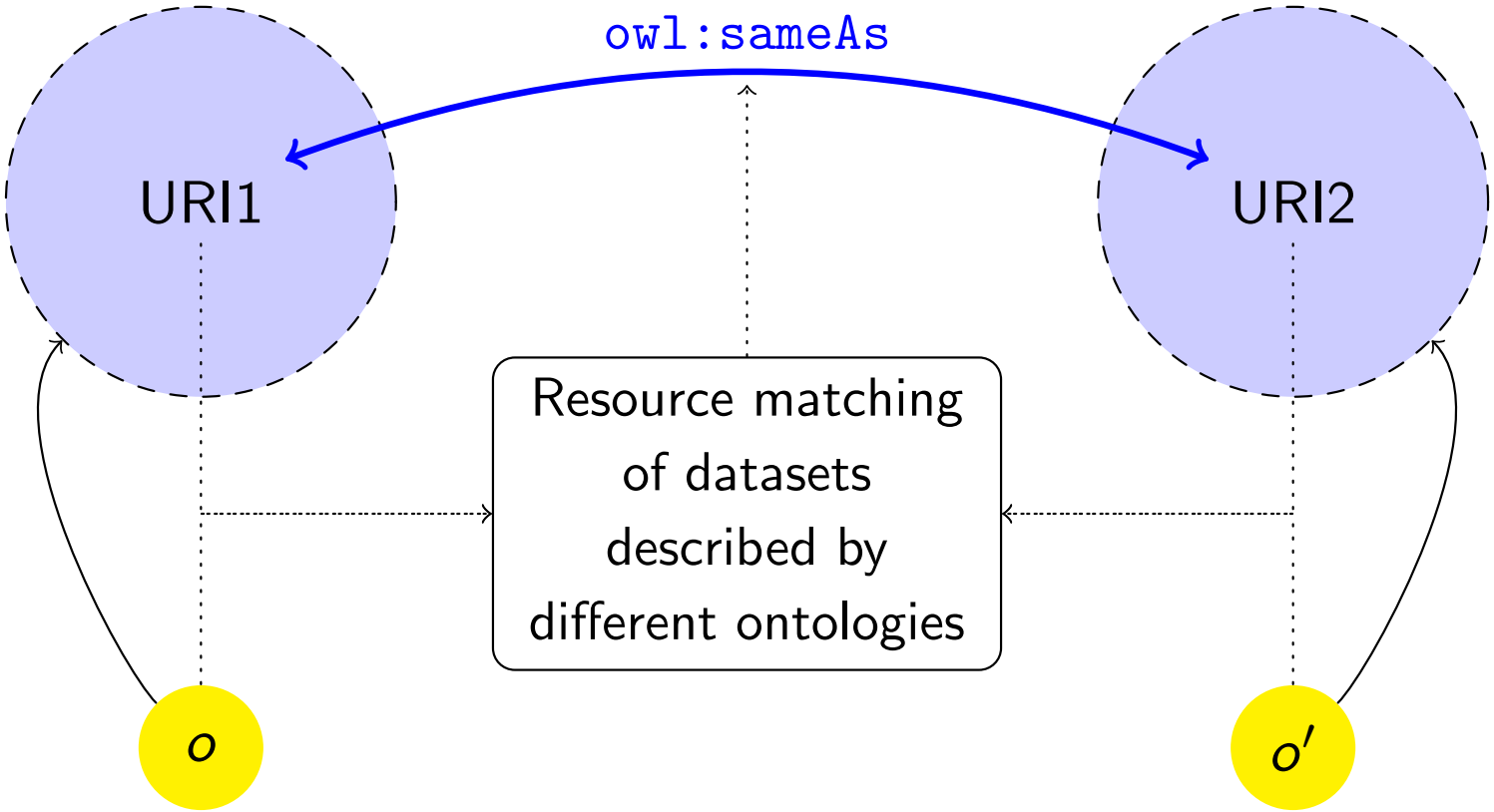
Data matching with different ontologies (implicit alignment)



Data matching with different ontologies (implicit alignment)



Data matching with different ontologies (implicit alignment)



Independent key computation

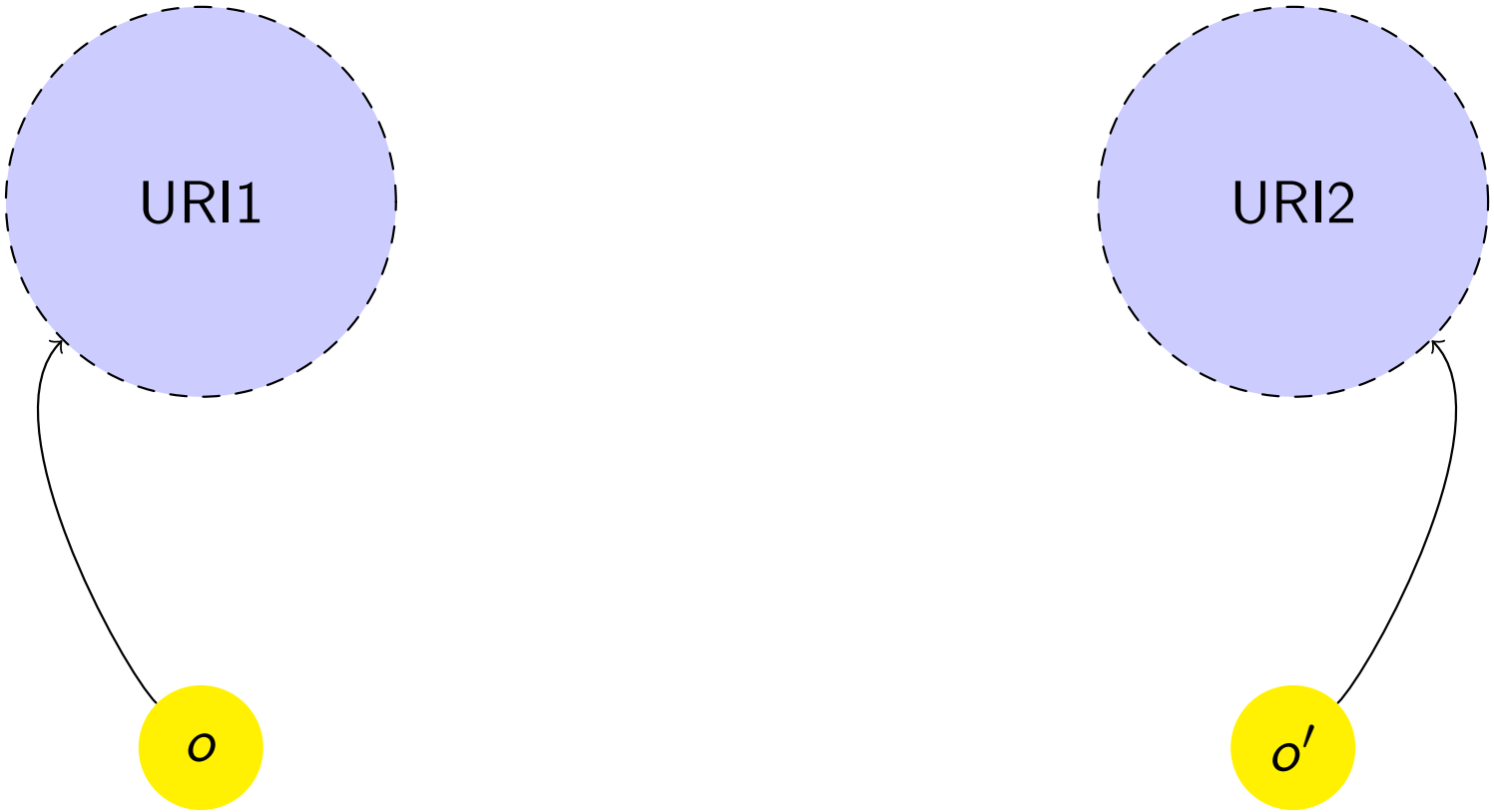
- Different span requires different key (France is not a key for INSEE);
- Differences in schema and depths makes difference in what is a key ("Paris" is both a department name (DEP_75) and a municipality name (COM_75056) for INSEE while the region name may be a key for NUTS)
- Keys are often meaningless: they are data-independent and database-dependent, hence they cannot be used for matching entities (REG_11 vs. FR_10).

Independent key computation

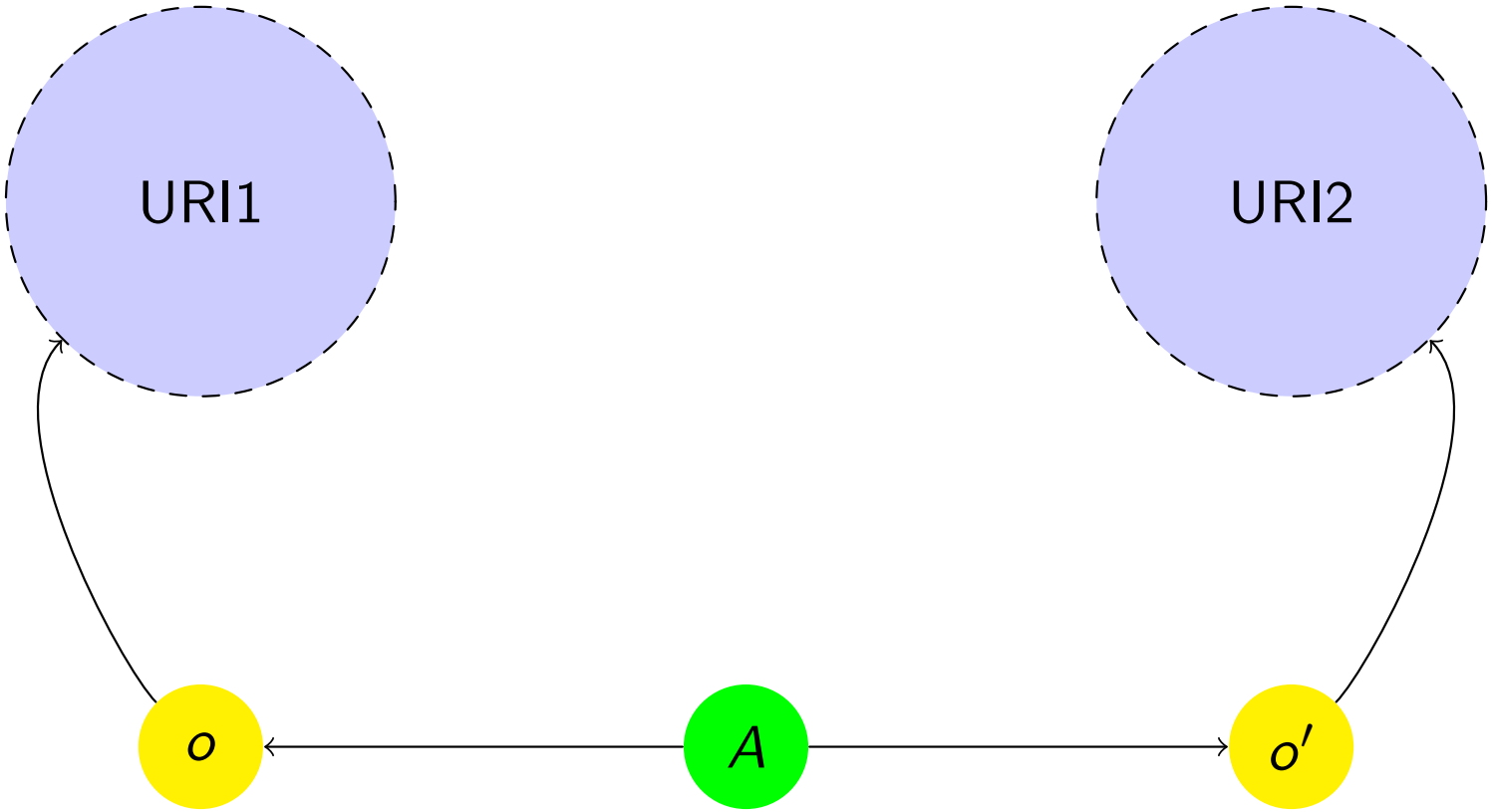
- Different span requires different key (France is not a key for INSEE);
- Differences in schema and depths makes difference in what is a key ("Paris" is both a department name (DEP_75) and a municipality name (COM_75056) for INSEE while the region name may be a key for NUTS)
- Keys are often meaningless: they are data-independent and database-dependent, hence they cannot be used for matching entities (REG_11 vs. FR_10).

- rdf:type and insee:nom are keys for INSEE (Region);
- nuts:level and nuts:name are keys for NUTS (NUTSRegion);
- insee:nom corresponds to nuts:name; there exists a correspondence between rdf:type in INSEE and nuts:level.

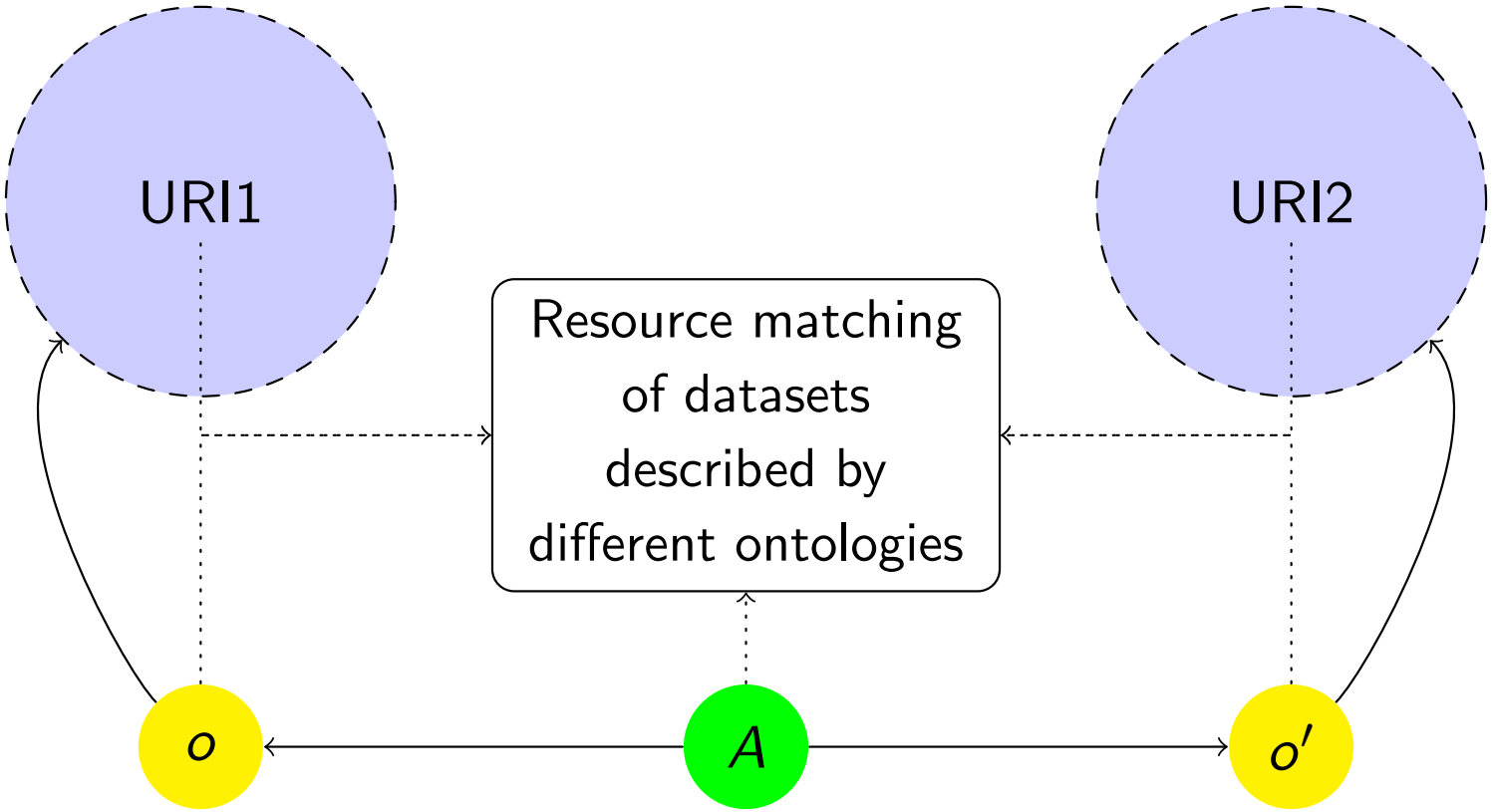
Data matching with different ontologies (explicit alignment)



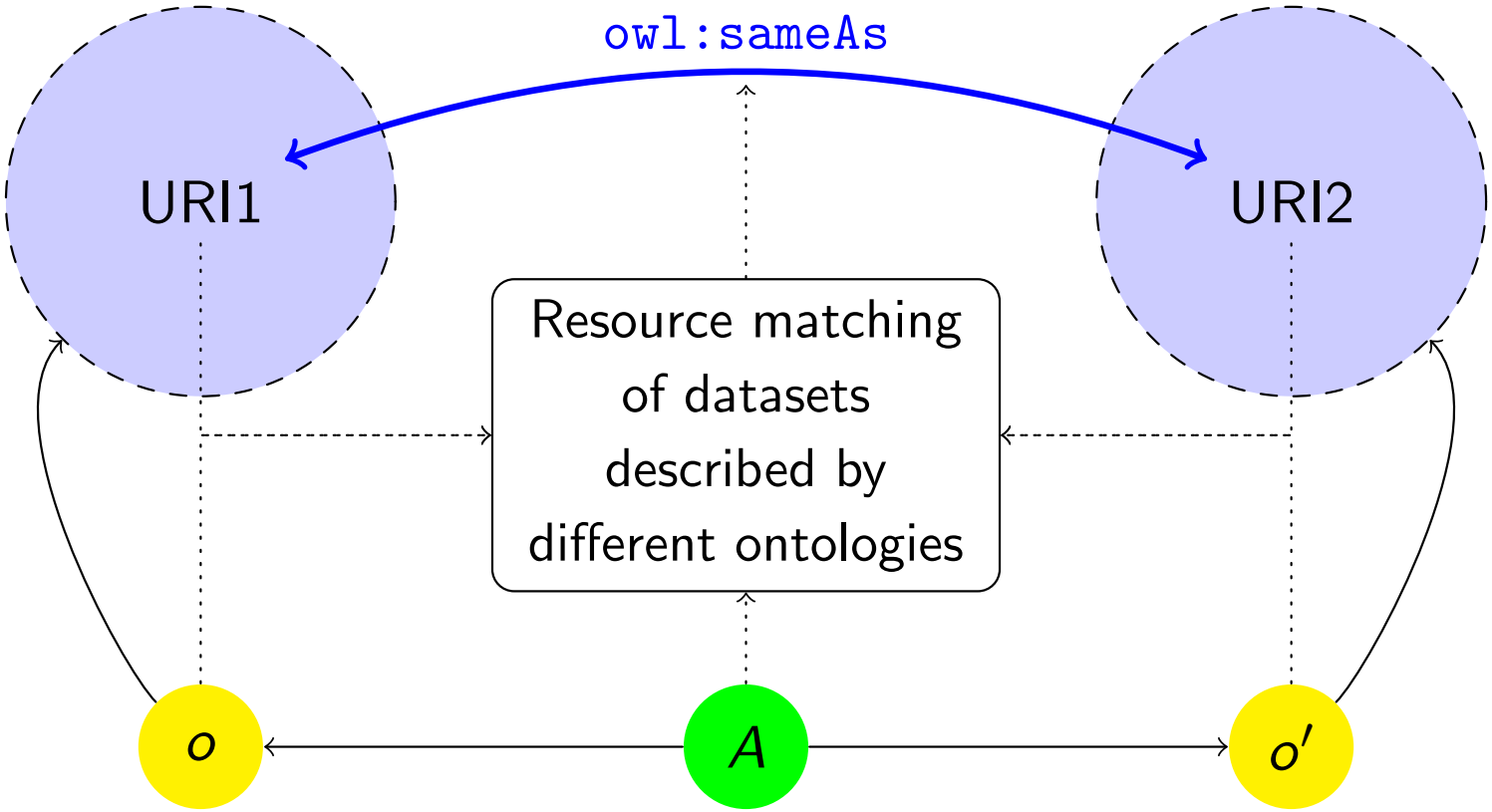
Data matching with different ontologies (explicit alignment)



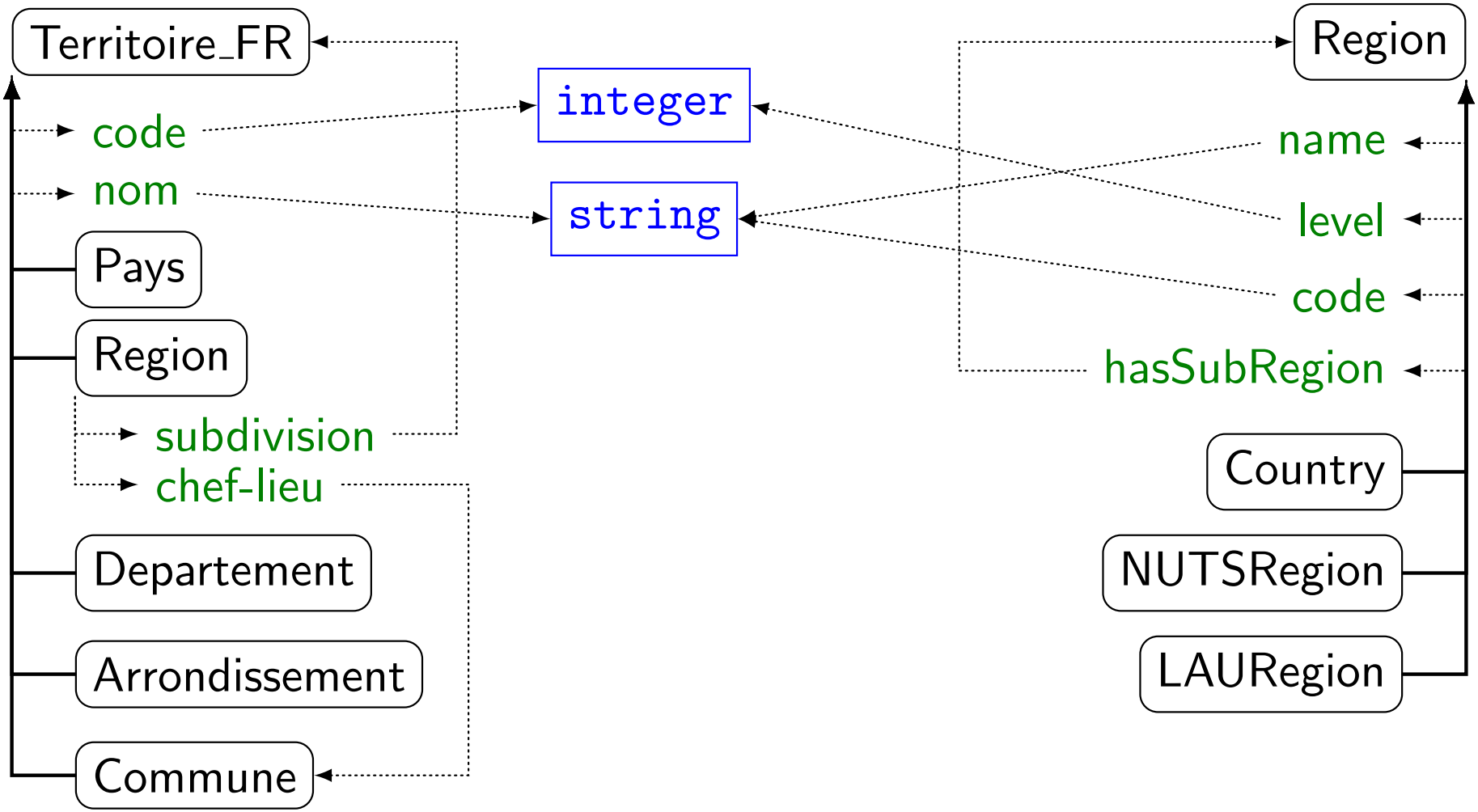
Data matching with different ontologies (explicit alignment)



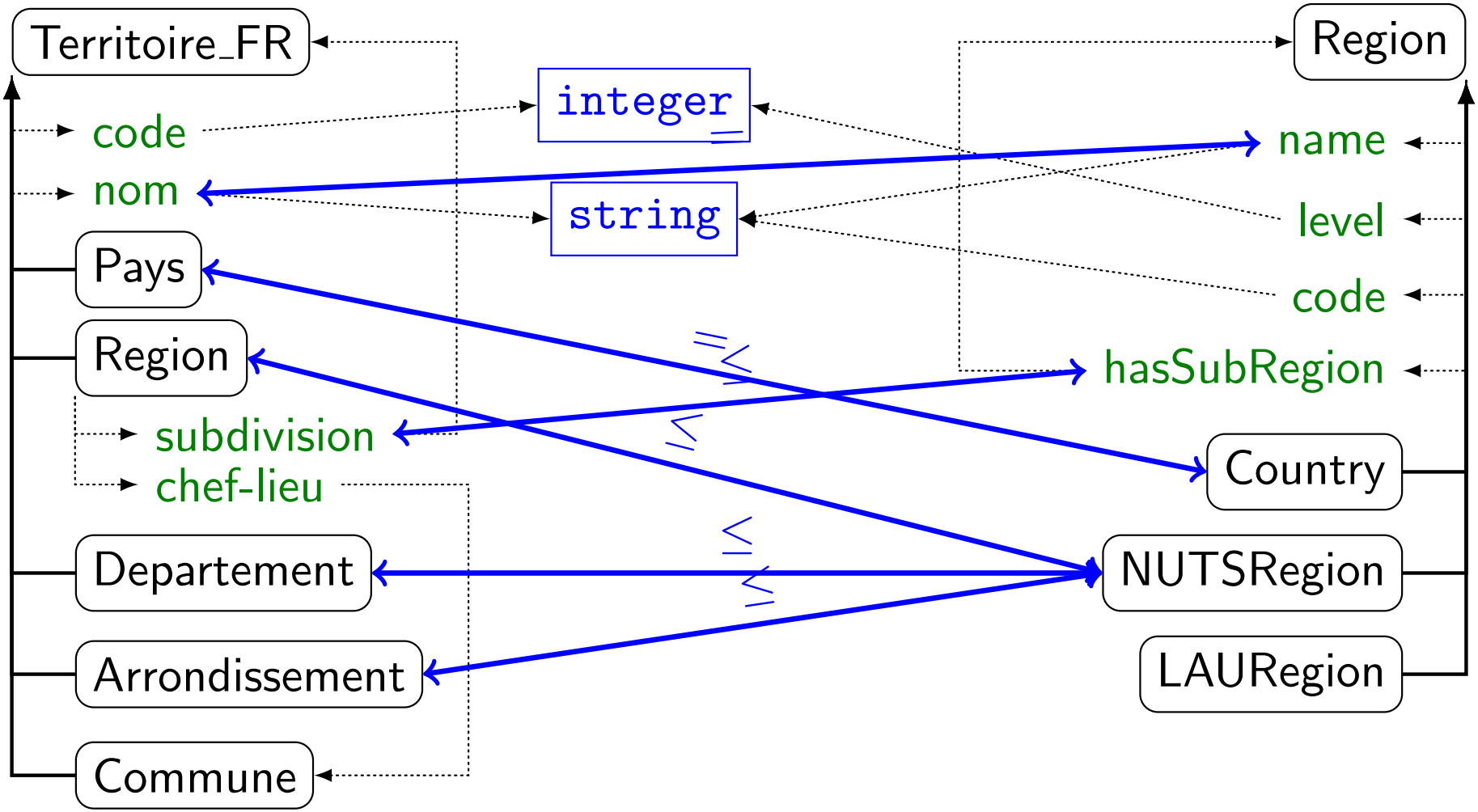
Data matching with different ontologies (explicit alignment)



INSEE and NUTS: ontology alignment



INSEE and NUTS: ontology alignment



Ontology alignment

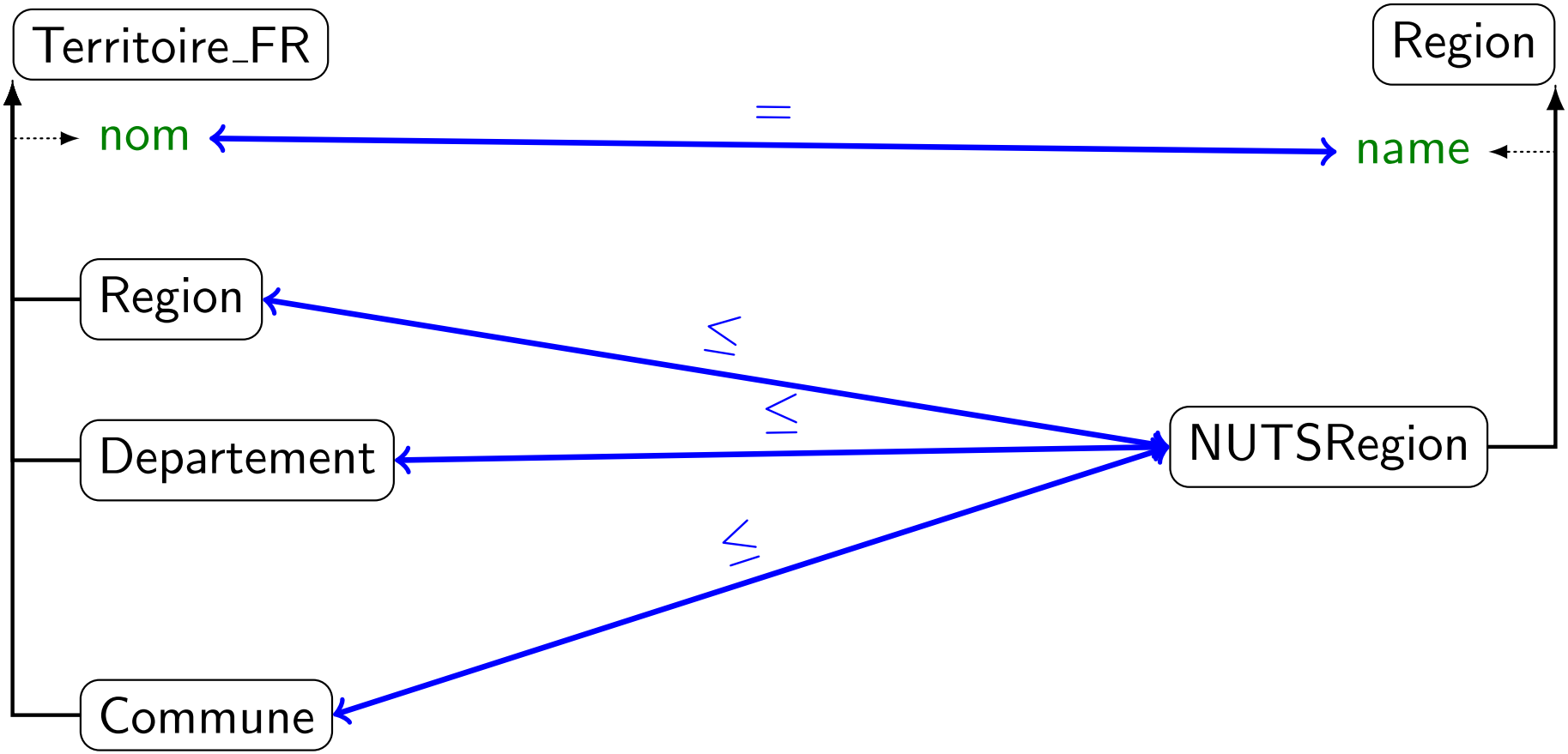
+ Ontology alignment restore the benefits of common ontology: it helps focussing the search;
– It is not exact science! (but alignments may be available);
+ Ontology alignment and data linking reinforce each others.

A simple algorithm

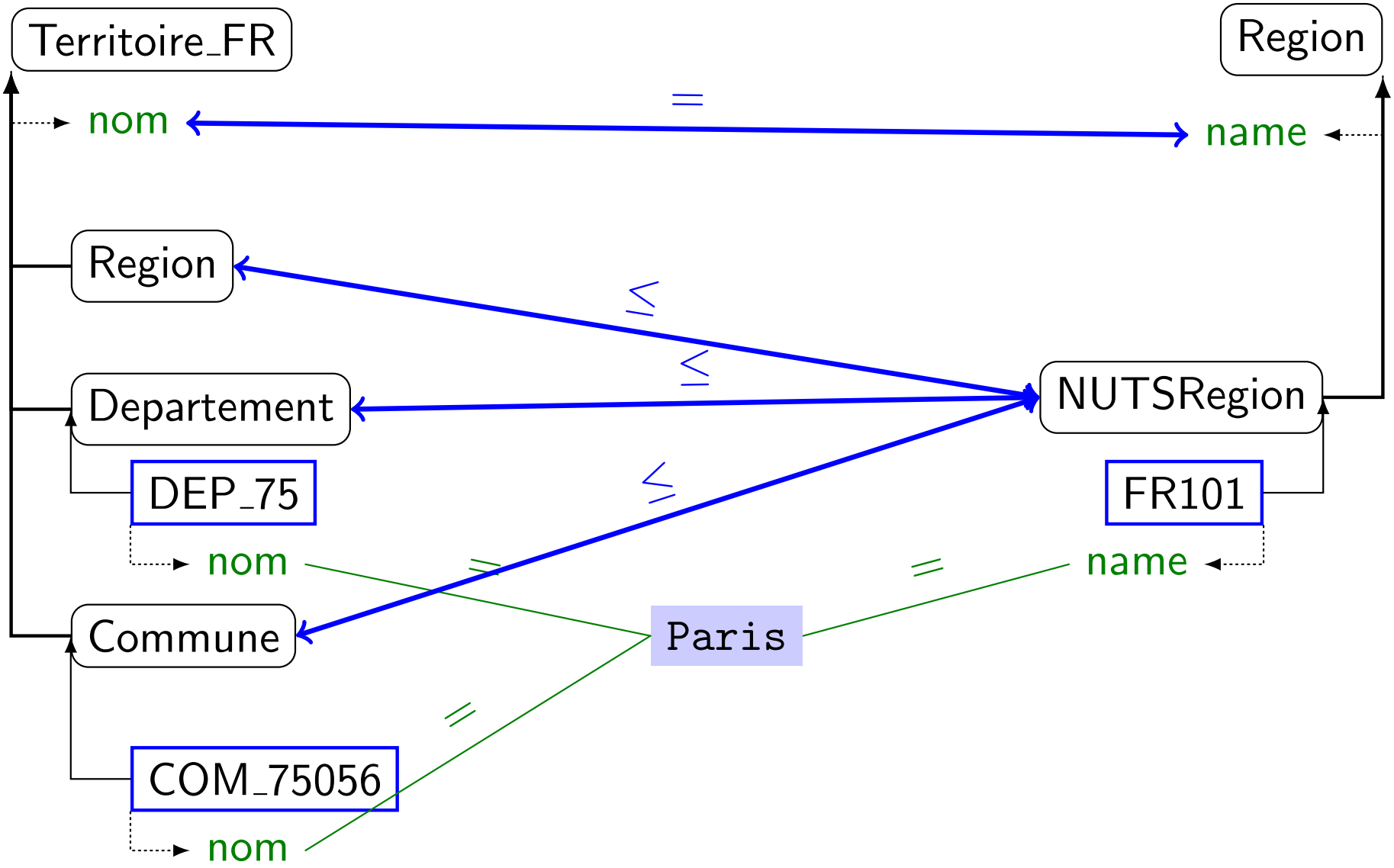
- Find matching concepts [concept matching];
- For each of them, determine matching properties based on the similarity between their values in both datasets [property matching];
- From them find property combinations identifying corresponding entities [key extraction];
- Link corresponding entities [link generation].

For instance, nom/Region$_{INSEE}$ $\subseteq$ name/NUTSRegion$_{NUTS}$ and moreover they are unambiguous.

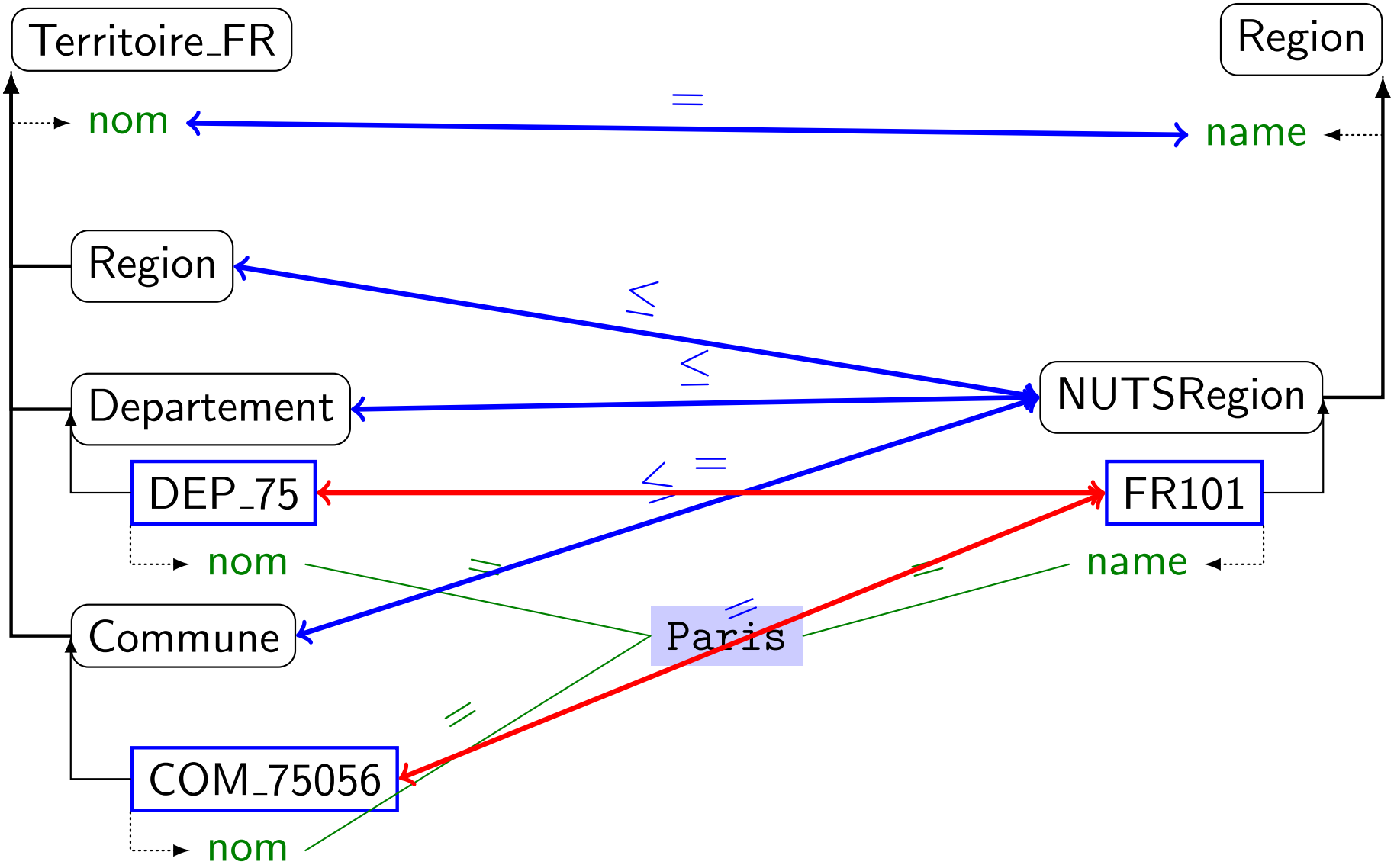
Simple alignments are not sufficient



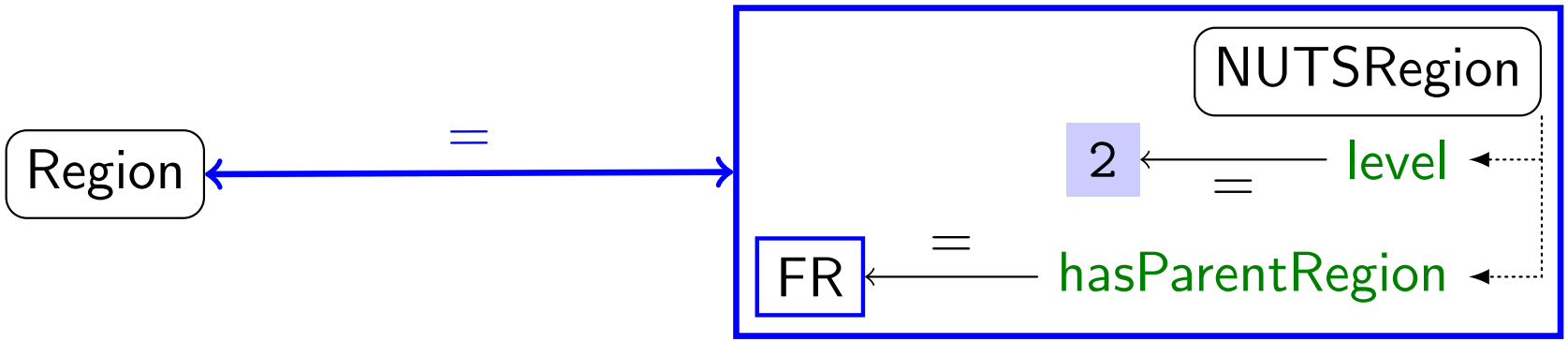
Simple alignments are not sufficient



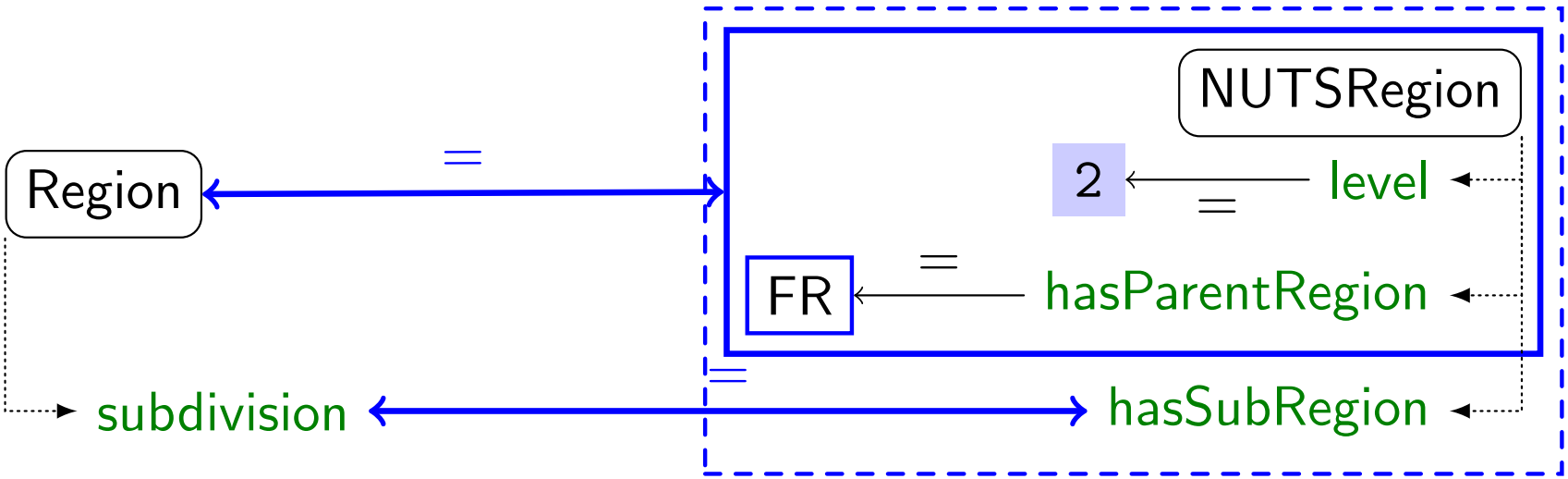
Simple alignments are not sufficient



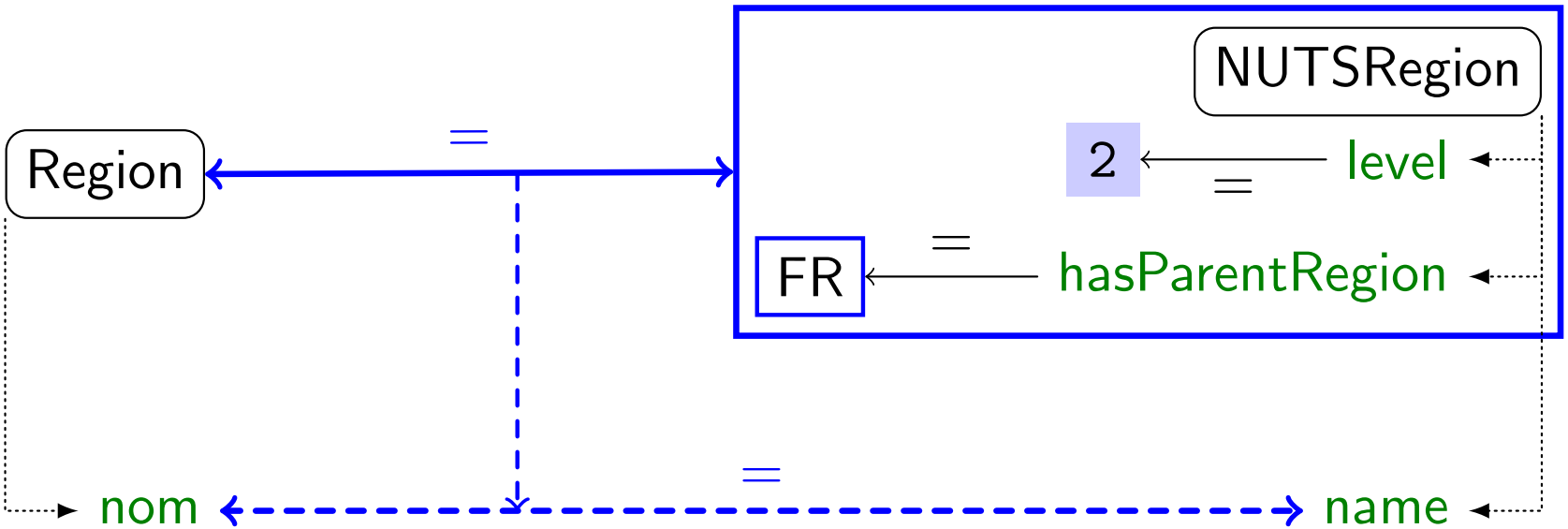
Expressive alignments are necessary



Expressive alignments are necessary



Expressive alignments are necessary



Query generation

```
SELECT ?r
PREFIX insee: <http://rdf.insee.fr/ontologie-geo-2006.rdf#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
FROM <http://rdf.insee.fr/geo/regions-2011.rdf>
WHERE {
    ?r rdf:type insee:Region .
}
```

```
SELECT ?n
PREFIX nuts: <http://ec.europa.eu/eurostat/ramon/ontologies/geographi
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
FROM <http://ec.europa.eu/eurostat/ramon/rdfdata/nuts2008/>
WHERE {
    ?n rdf:type nuts:NUTSRegion .
    ?n nuts:level 2^^xsd:int .
    ?n nuts:hasParentRegion nuts:FR1 .
}
```

Query generation

```
CONSTRUCT { ?r owl:sameAs ?n . }
PREFIX insee: <http://rdf.insee.fr/ontologie-geo-2006.rdf#>
PREFIX nuts: <http://ec.europa.eu/eurostat/ramon/ontologies/geographi
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
FROM <http://rdf.insee.fr/geo/regions-2011.rdf>
FROM <http://ec.europa.eu/eurostat/ramon/rdfdata/nuts2008/>
WHERE {
    ?r rdf:type insee:Region .
    ?r insee:nom ?l .
    ?n rdf:type nuts:NUTSRegion .
    ?n nuts:name ?l .
    ?n nuts:level 2^^xsd:int .
    ?n nuts:hasParentRegion nuts:FR1 .
}
```

What does this mean?

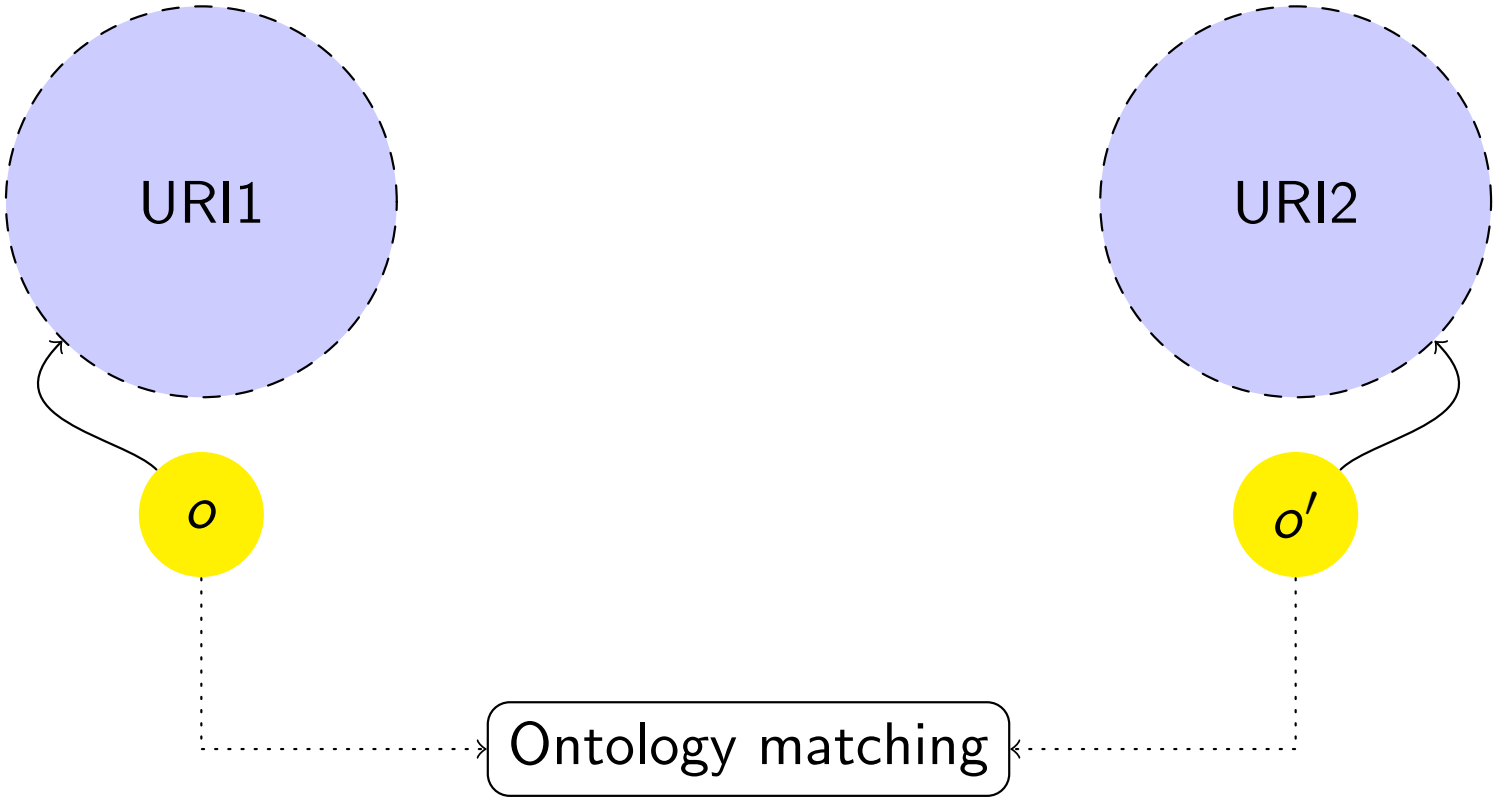
- Ontology alignments are schema-level expression of correspondences;
- They are useful for focussing the search;
- Expressive alignments are necessary;
- They can be turned into SPARQL-based link generators.

but it is also necessary to express instance level constraints:

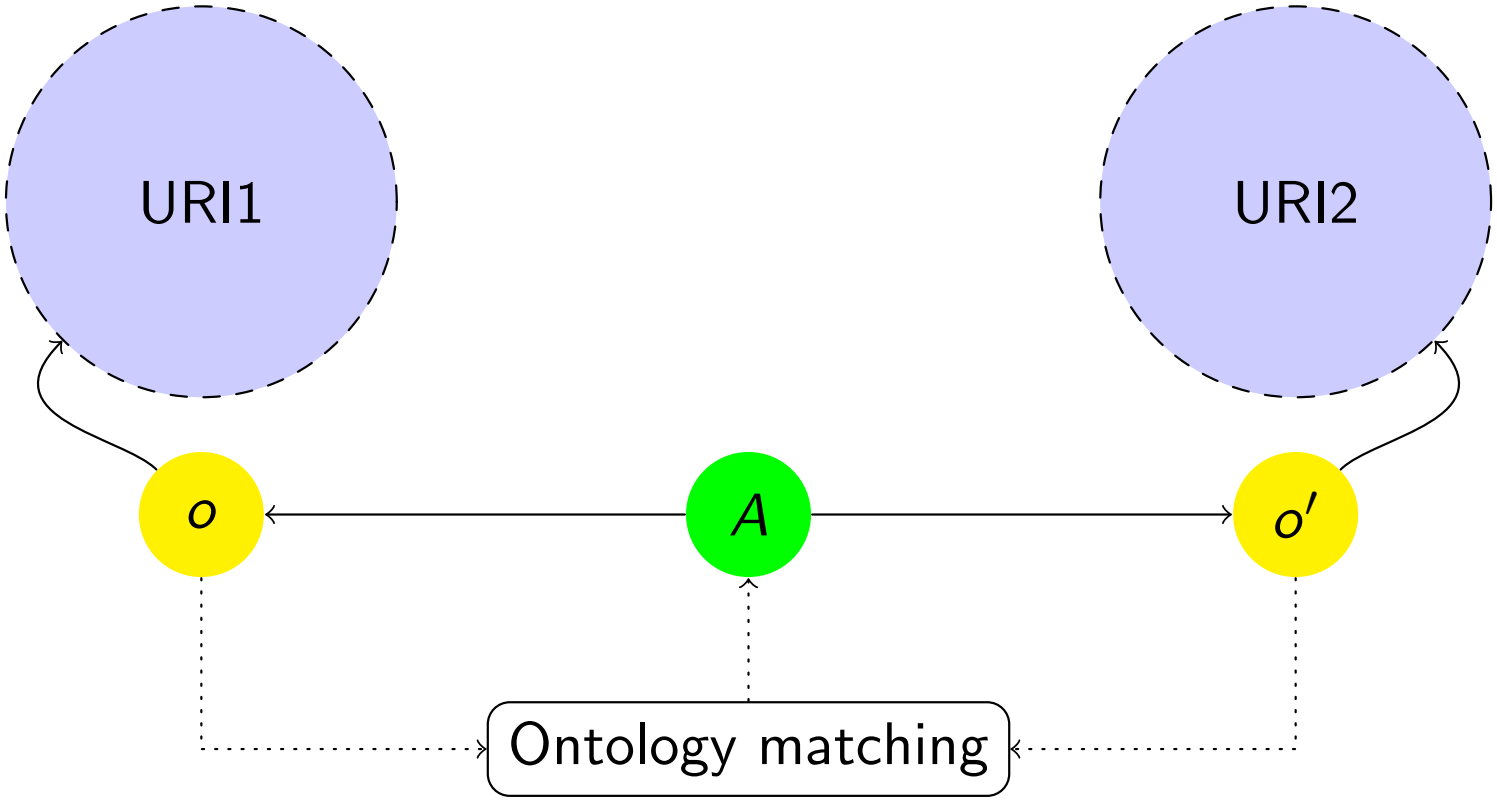
- for converting data (e.g., mph vs. m/s);
- for expressing matching constraint on data (e.g., similarity).

General framework

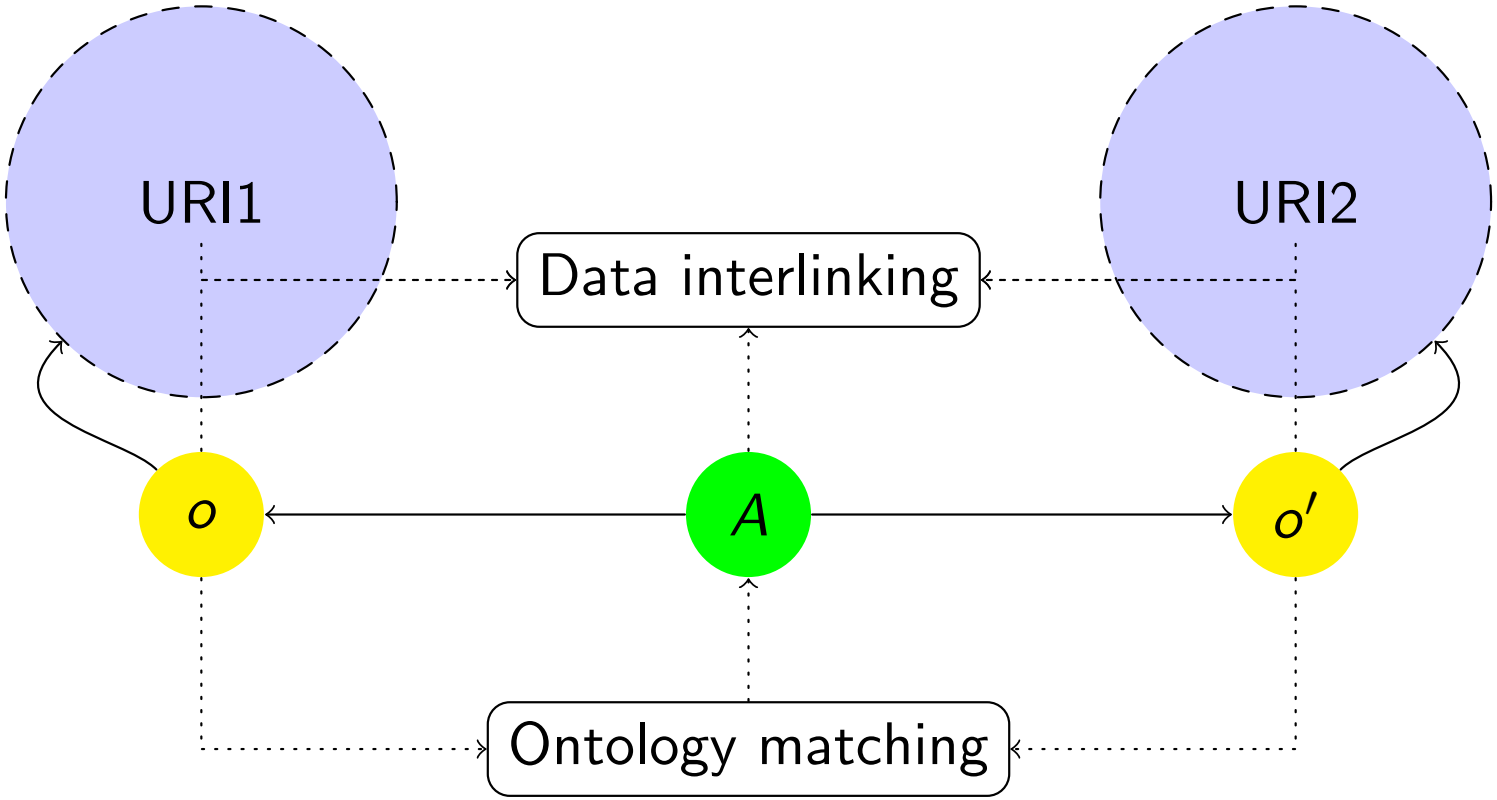
General framework



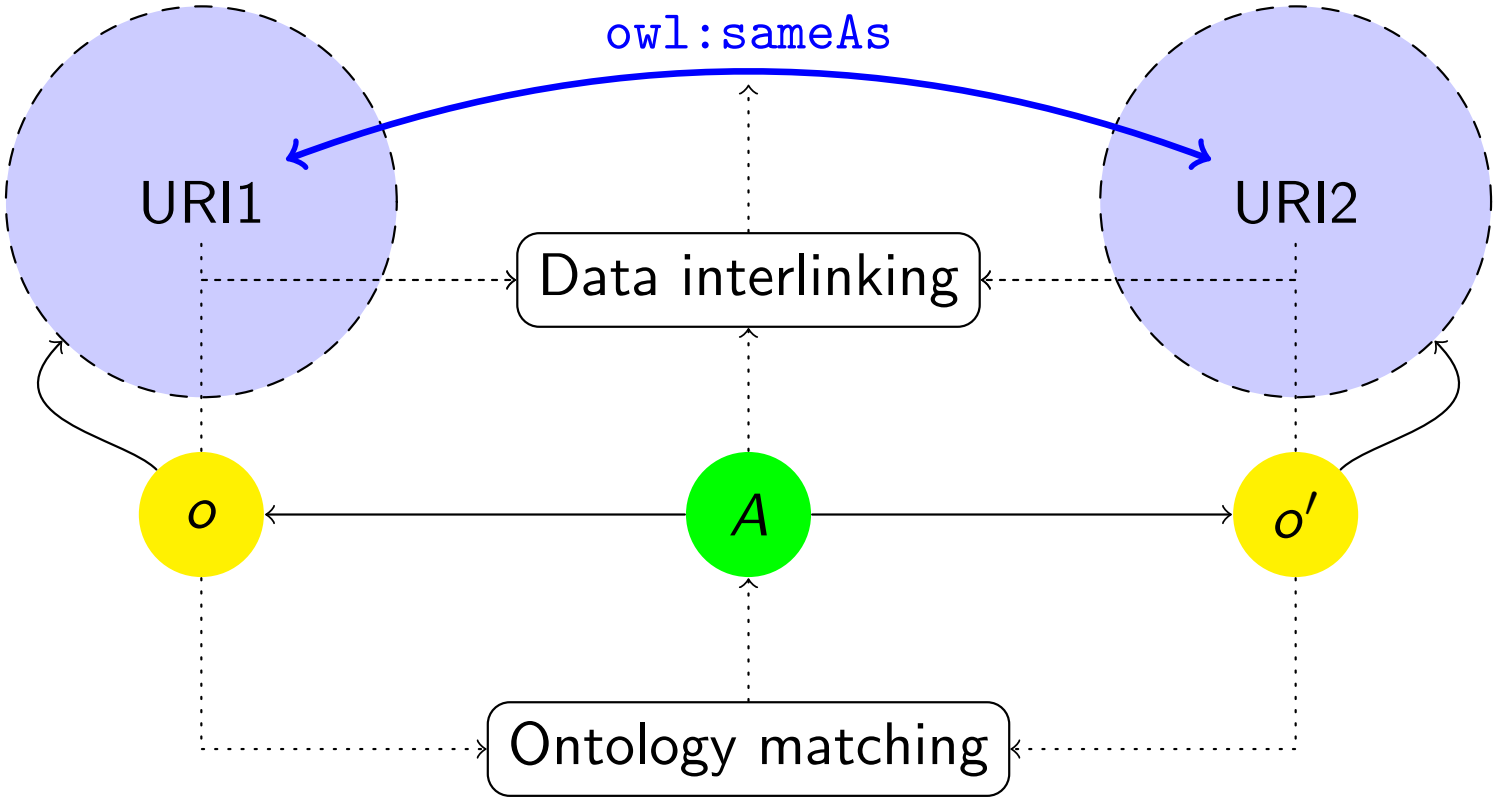
General framework



General framework



General framework



Application

Consider a linking script between DBPedia and Geonames:

```
<Silk>
  <Prefix id="rdfs" namespace=
          "http://www.w3.org/2000/01/rdf-schema#" />
  <Prefix id="dbpedia" namespace=
          "http://dbpedia.org/ontology/" />
  <Prefix id="gn" namespace=
          "http://www.geonames.org/ontology#" />

  <DataSource id="dbpedia">
    <EndpointURI>http://demo_sparql_server1/sparql
    </EndpointURI>
    <Graph>http://dbpedia.org</Graph>
  </DataSource>

  <DataSource id="geonames">
    <EndpointURI>http://demo_sparql_server2/sparql
    </EndpointURI>
    <Graph>http://sws.geonames.org/</Graph>
  </DataSource>

  <Thresholds accept="0.9" verify="0.7" />
  <Output acceptedLinks="accepted_links.n3"
      verifyLinks="verify_links.n3"
      mode="truncate" />
```

```
<Interlink id="cities">
    <LinkType>owl:sameAs</LinkType>
    <SourceDataset dataSource="dbpedia" var="a">
      <RestrictTo>
         ?a rdf:type dbpedia:City
      </RestrictTo>
    </SourceDataset>
    <TargetDataset dataSource="geonames" var="b">
      <RestrictTo>
        ?b rdf:type gn:P
      </RestrictTo>
    </TargetDataset>
    <LinkCondition>
      <AVG>
        <Compare metric="jaroSimilarity">
          <Param name="str1" path="?a/rdfs:label" />
          <Param name="str2" path="?b/gn:name" />
        </Compare>
        <Compare metric="numSimilarity">
          <Param name="num1"
                 path="?a/dbpedia:populationTotal" />
          <Param name="num2" path="?b/gn:population" />
        </Compare>
      </AVG>
    </LinkCondition>
 </Interlink>
</Silk>
```

Application

Here is the alignment implicitly contained in this linking script.

```
:dbp-geo a align:Alignment;
   align:onto1 <http://dbpedia.org/ontology/>;
   align:onto2 <http://www.geonames.org/ontology#>;
   align:map [ :map1 a align:Cell;
     align:entity1 dbpedia:City;
     align:entity2 gn:P;
     align:relation align:subsumedBy.
   ];
   align:map [ :map2 a align:Cell;
     align:entity1 dbpedia:populationTotal;
     align:entity2 gn:population;
     align:relation align:equivalent.
   ];
   align:map [ :map3 a align:Cell;
     align:entity1 rdfs:label;
     align:entity2 gn:name;
     align:relation align:equivalent.
   ].
```

```
  align:map [ :map2 a align:Cell;
      align:entity1 [ a align:Property;
        edoal:and dbpedia:populationTotal.
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain dbpedia:City.
];
      align:entity2 [ a align:Property;
        edoal:and gn:population;
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain gn:P. ];
      align:relation align:equivalent.
    ];
    align:map [ :map2 a align:Cell;
      align:entity1 [ a align:Property;
        edoal:and rdfs:label.
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain dbpedia:City.
];
      align:entity2 [ a align:Property;
        edoal:and gn:name;
edoal:and [ a edoal:PropertyDomainRestriction;
  edoal:domain gn:P. ];
      align:relation align:equivalent.
    ].
```

Application

It is then possible to simplify the linking script and keep a declarative link specification.

```
<UseAlignment rdf:resource="#dbp-geo" />

<Interlink id="cities">
  <LinkType>owl:sameAs</LinkType>
  <LinkCell rdf:resource="#map1" />
  <LinkCondition>
    <AVG>
      <Compare metric="jaroSimilarity">
        <CellParam rdf:resource="#map2" />
      </Compare>
      <Compare metric="numSimilarity">
        <CellParam rdf:resource="#map3" />
      </Compare>
    </AVG>
  </LinkCondition>

  <Thresholds accept="0.9" verify="0.7" />
  <Output acceptedLinks="accepted_links.n3"
    verifyLinks="verify_links.n3"
    mode="truncate" />
</Interlink>
```

Outline

Conclusions

- A large part of linked data added value is in links;
- They may not be easy to find;
- Many techniques are available for automating interlinking;
- Having a general framework may help integrating them.

Jerome.Euzenat@inria.fr

http://exmo.inrialpes.fr