Interlinking the web of data: challenges and solutions

Work in collaboration with François Scharffe and others

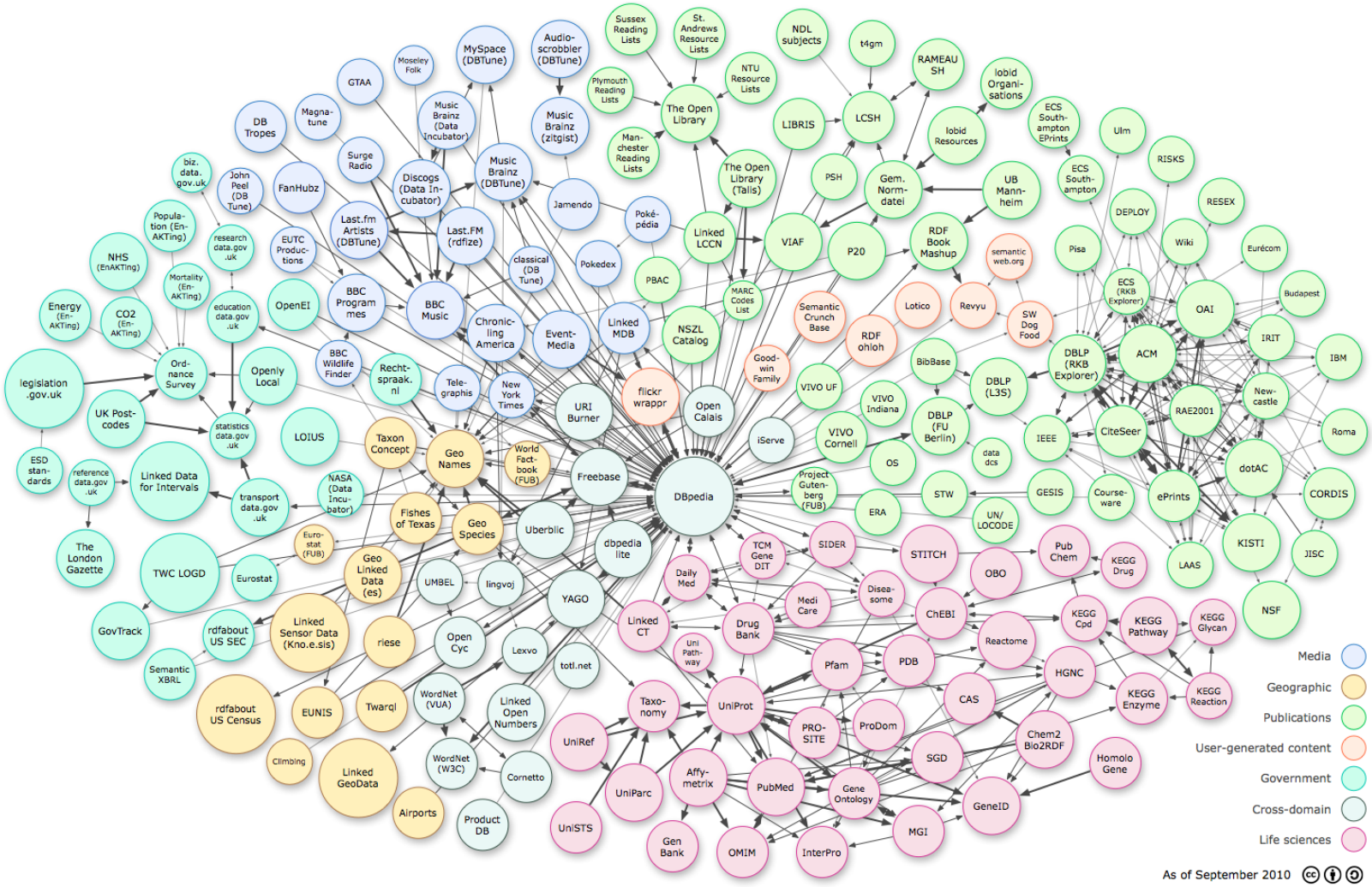
Jérôme Euzenat

June 8, 2011

Outline

Outline

Linked (open) data cloud



Publishing data on the web

Four publication principles

1. Use URIs for identifying resources
2. Use dereferencable URIs
3. When a URI is dereferenced, a description of the identified resource is returned
4. Published datasets must be interconnected to other datasets

If possible using semantic web technologies: URI, HTTP, RDF, OWL

What for?

- If you are an authority publishing data (gov), this allows you to put your data in the context other authority data and the others to reuse your data;
- If you are a link producer, i.e., someone who knows how to add value by cross-linking data, you have standard references to work with;
- If you are an application developer, e.g., seevl.net, you can take advantage of this data always on the web.

and

- the data is constantly up-to-date;
- the data can be freely linked;
- Data is browsable.

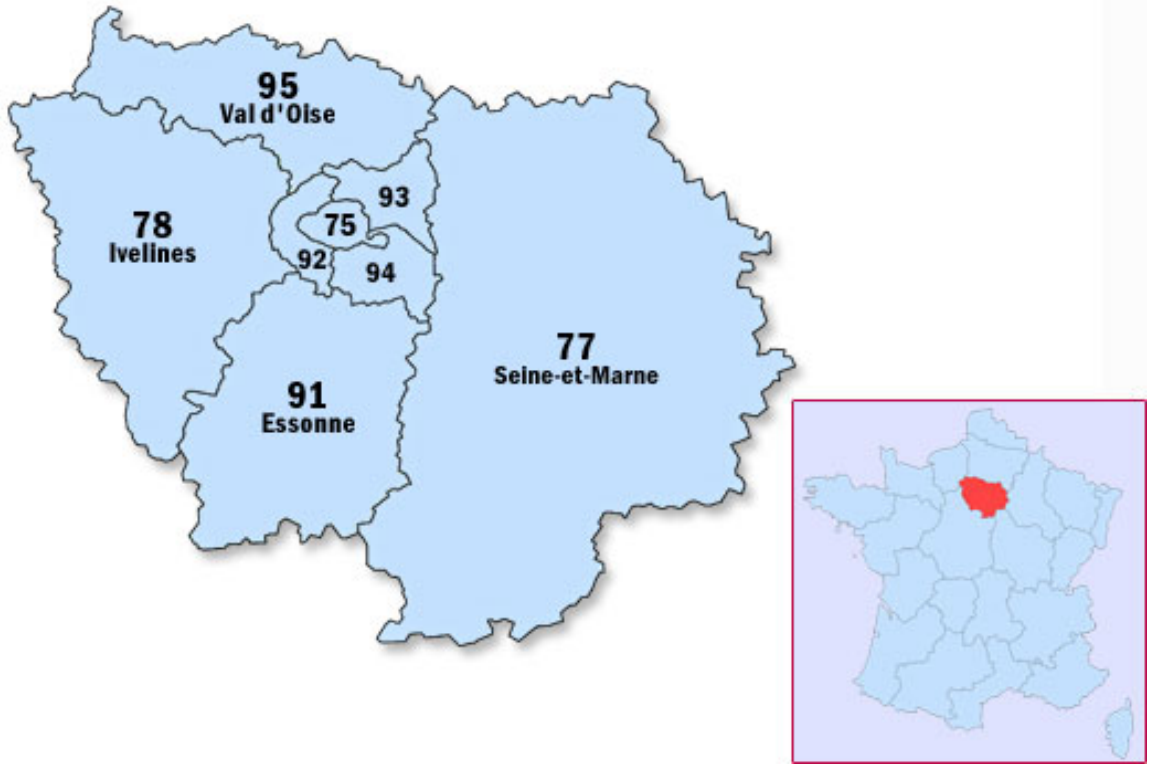
Example: INSEE dataset

Région table:

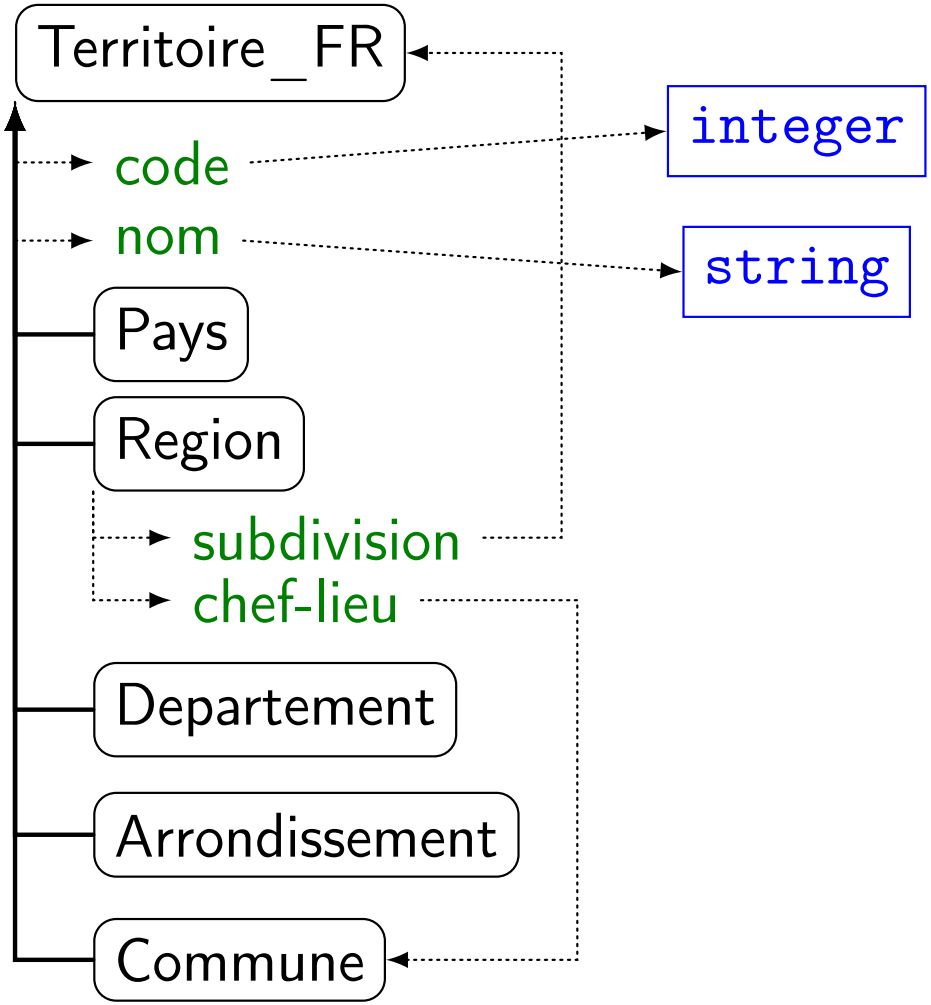
code	nom	chef-lieu
11	Île-de-France	75056
21	Champagne-Ardenne	51108
22	Picardie	80021

Sous-région table:

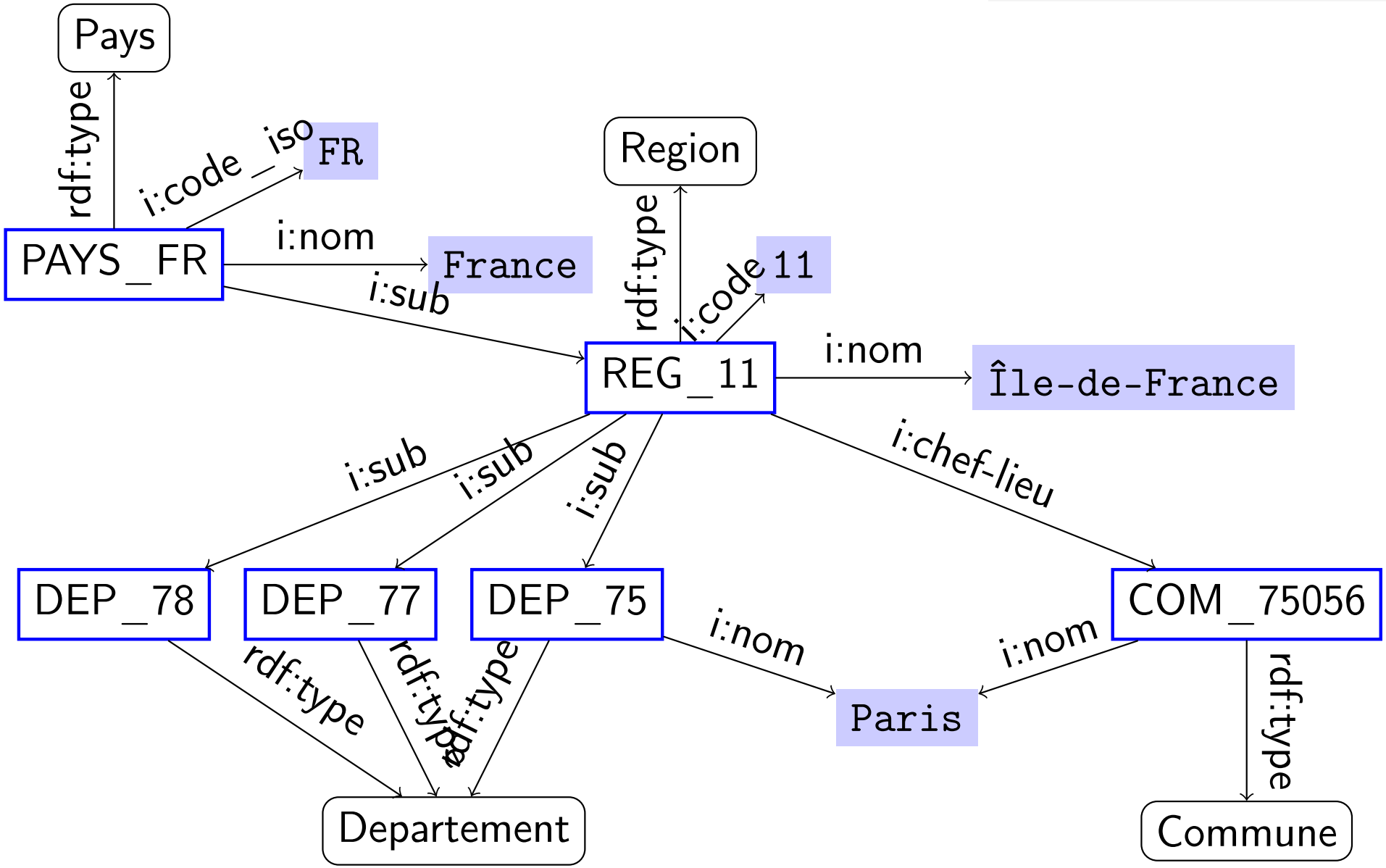
région	département
11	75
11	77
11	78
11	91
11	92
11	93



Example: Administrative ontology



Example: RDF


Pays
rdf:type
i:code_iso
FR
PAYS_FR
i:nom
France
i:sub
Region
rdf:type
i:code
11
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: 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	

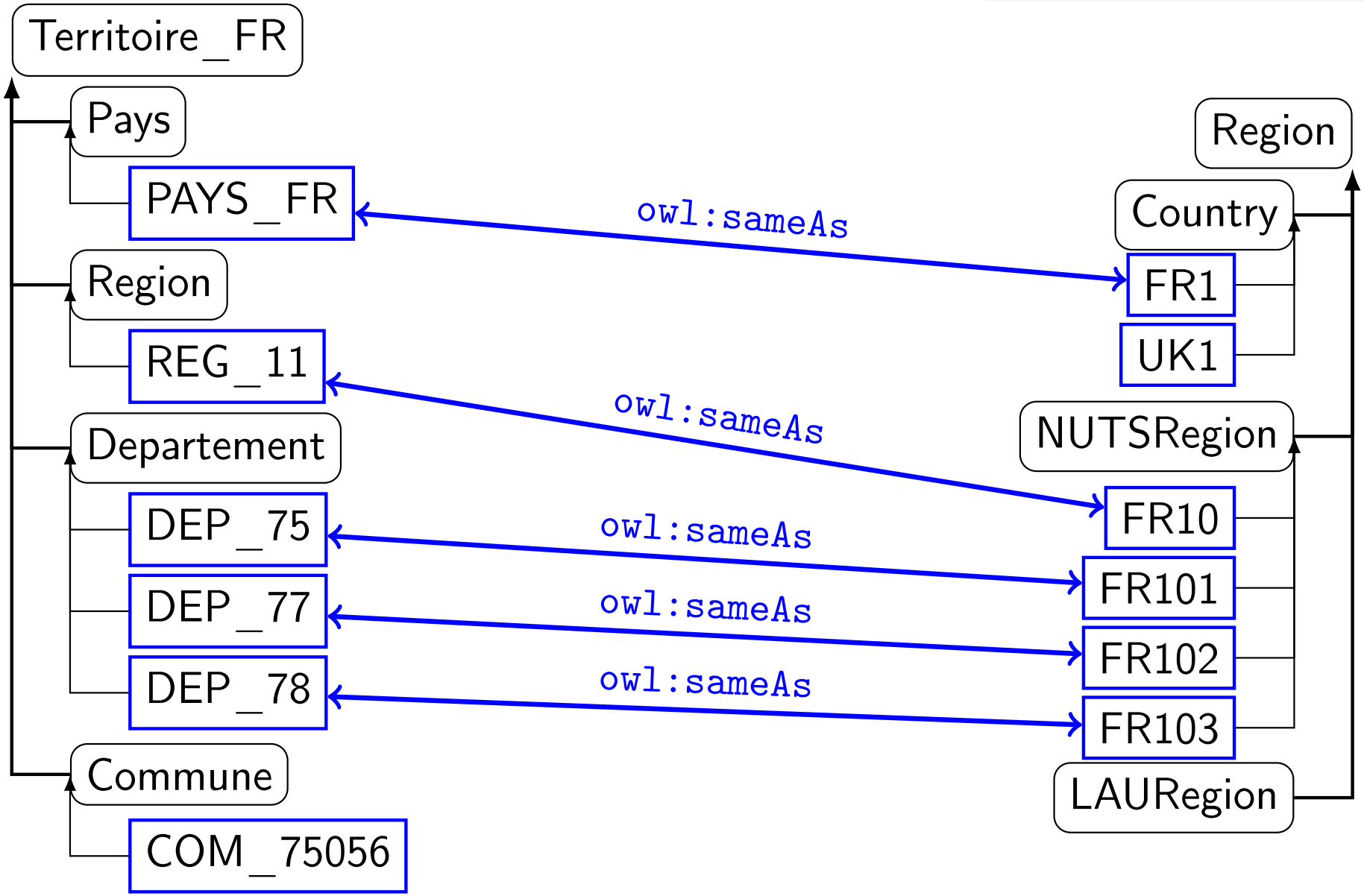
Example: Linking INSEE and NUTS

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	

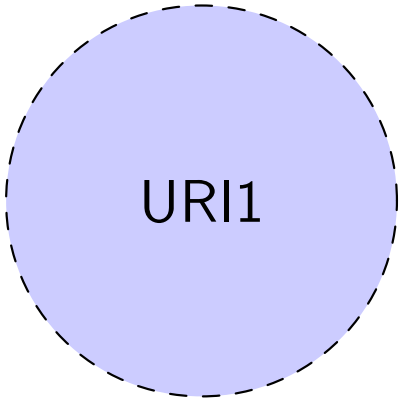
Å vs. Saint-Rémy-en-Bouzemont-Saint-Genest-et-Isson or Montbonnot Saint-Martin

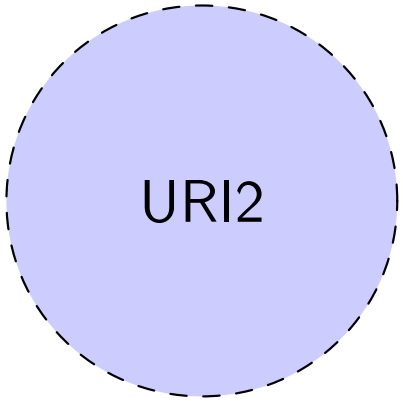
Example: Linking INSEE and NUTS



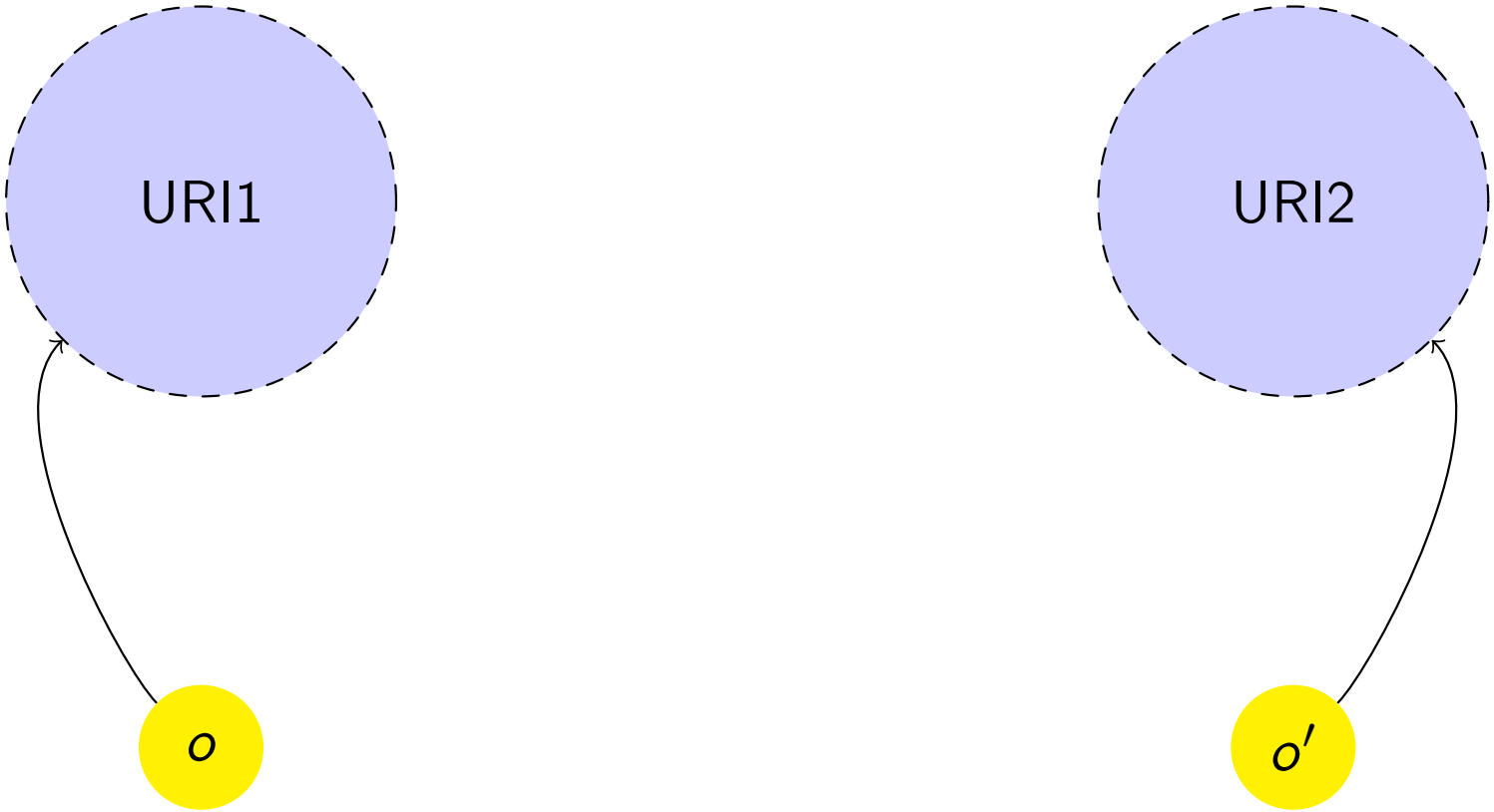
Outline

The data interlinking problem

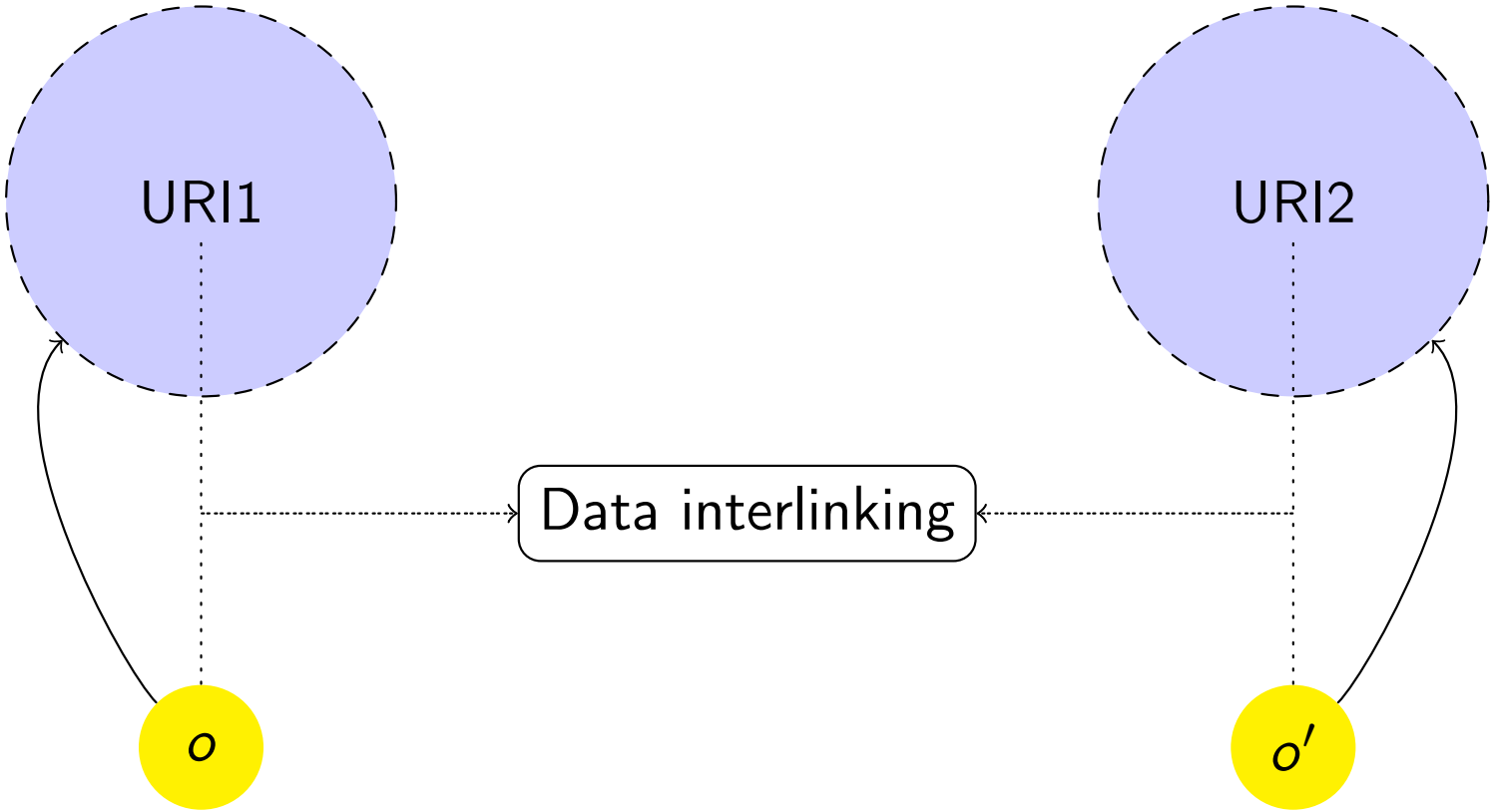




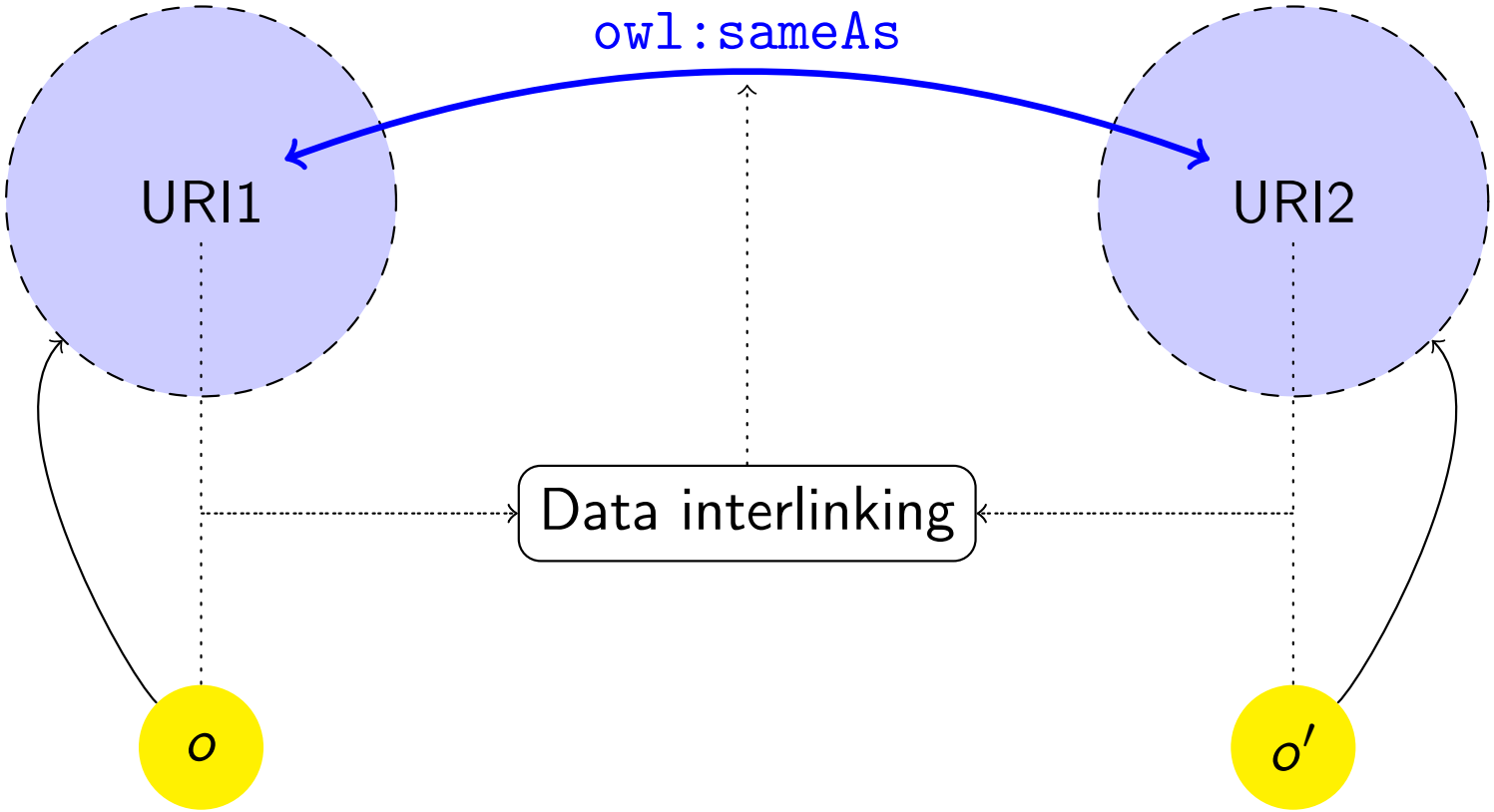
The data interlinking problem



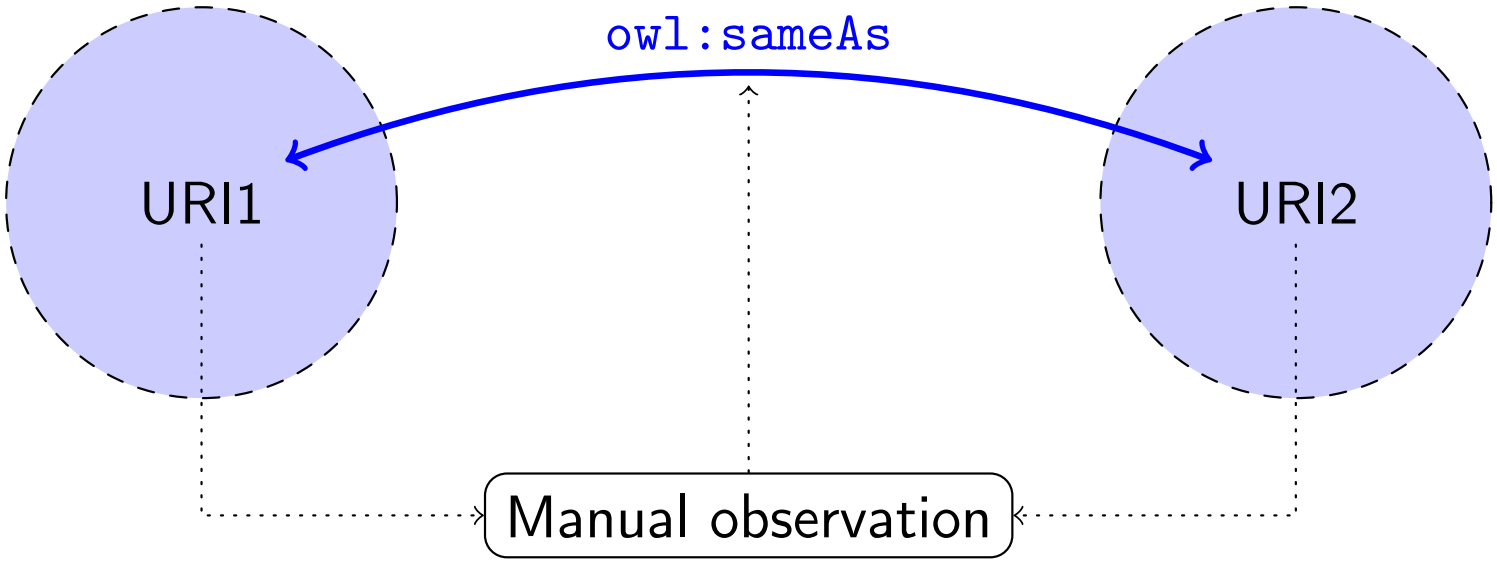
The data interlinking problem



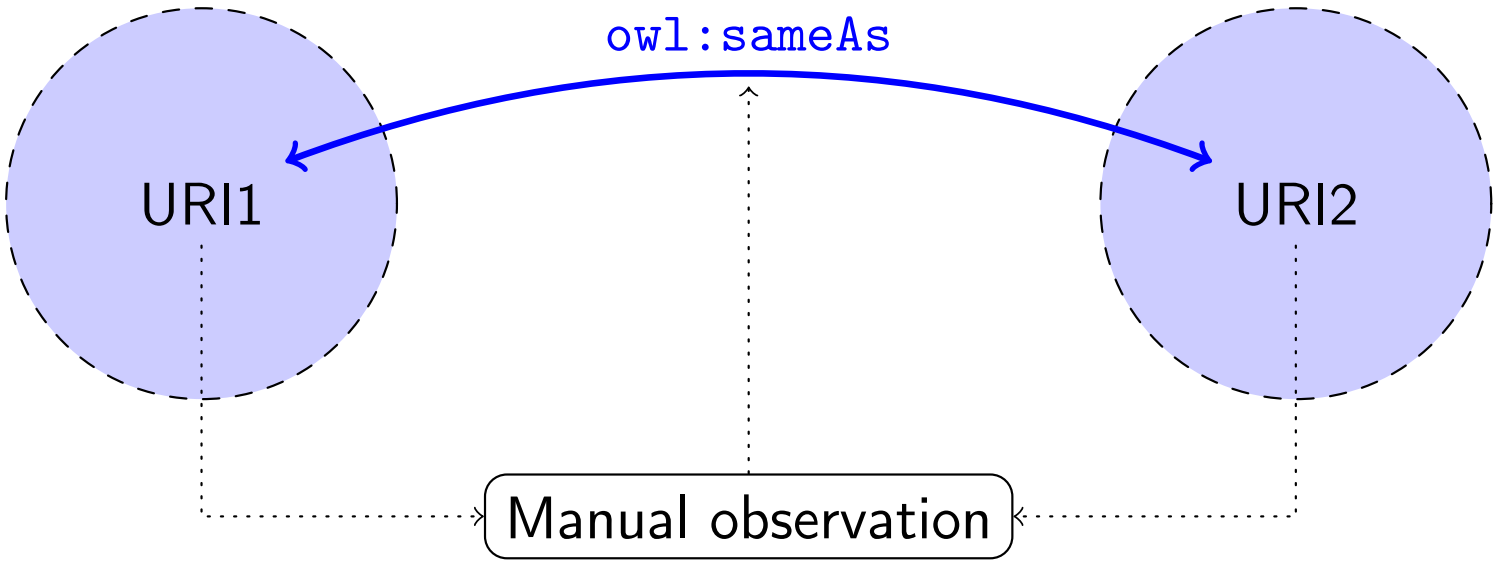
The data interlinking problem



Manual resource matching

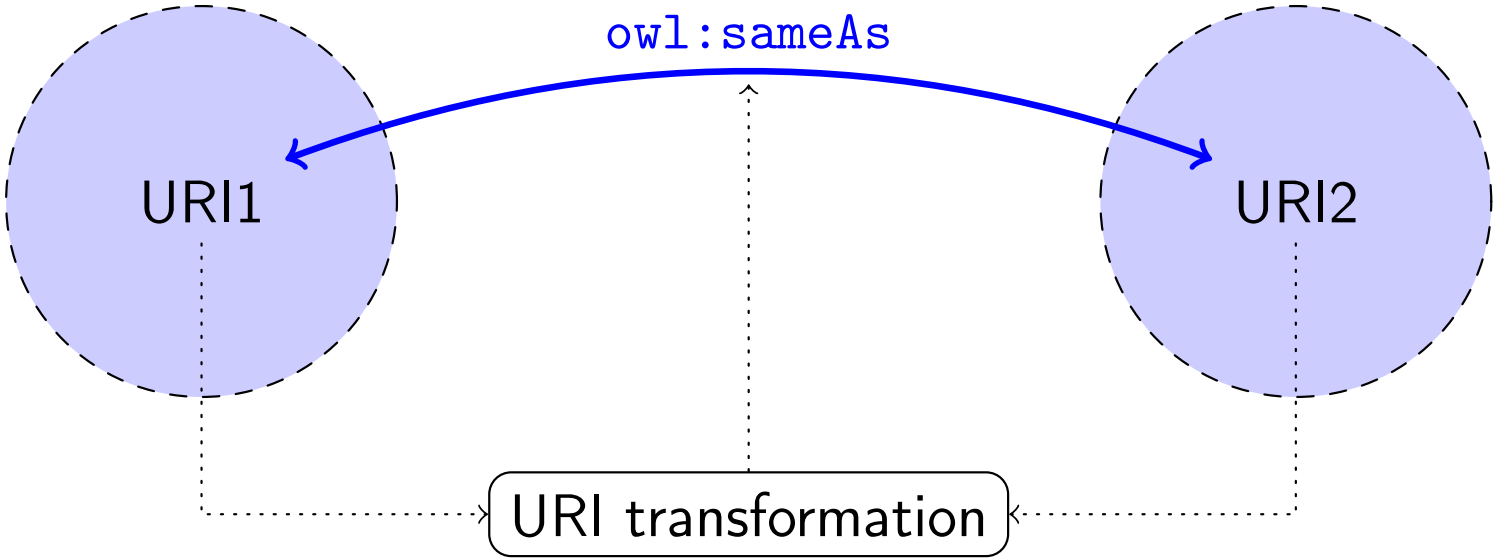


Manual resource matching

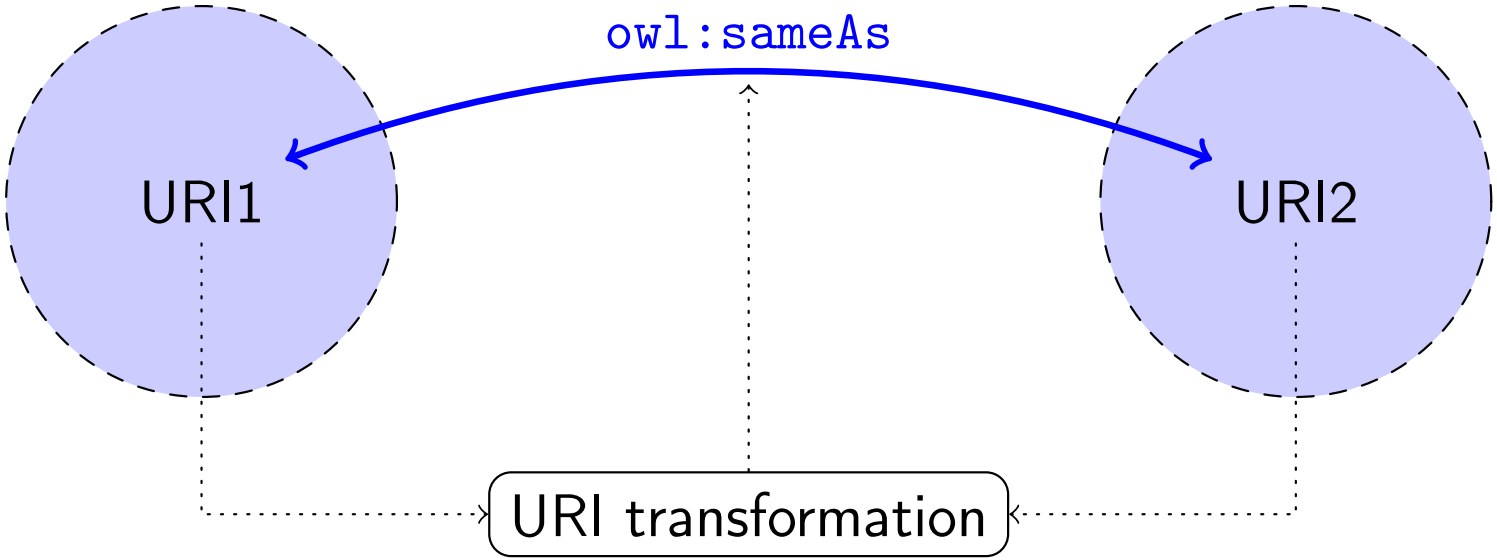


This does not scale.

URI matching

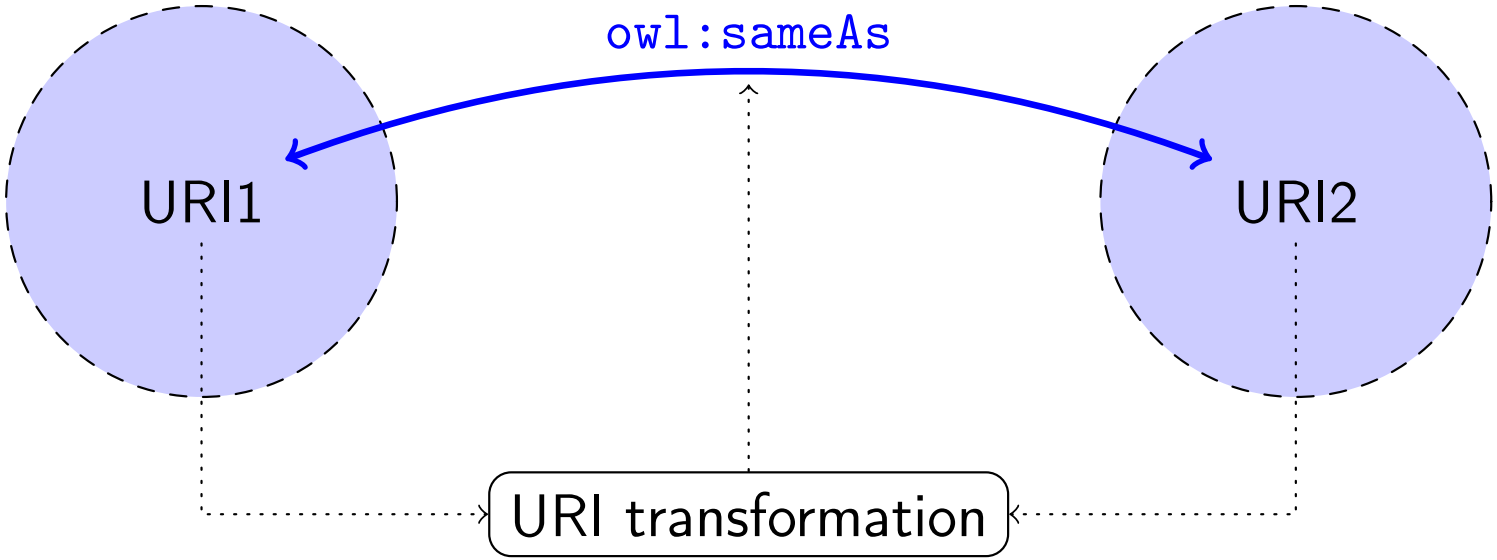


URI matching




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

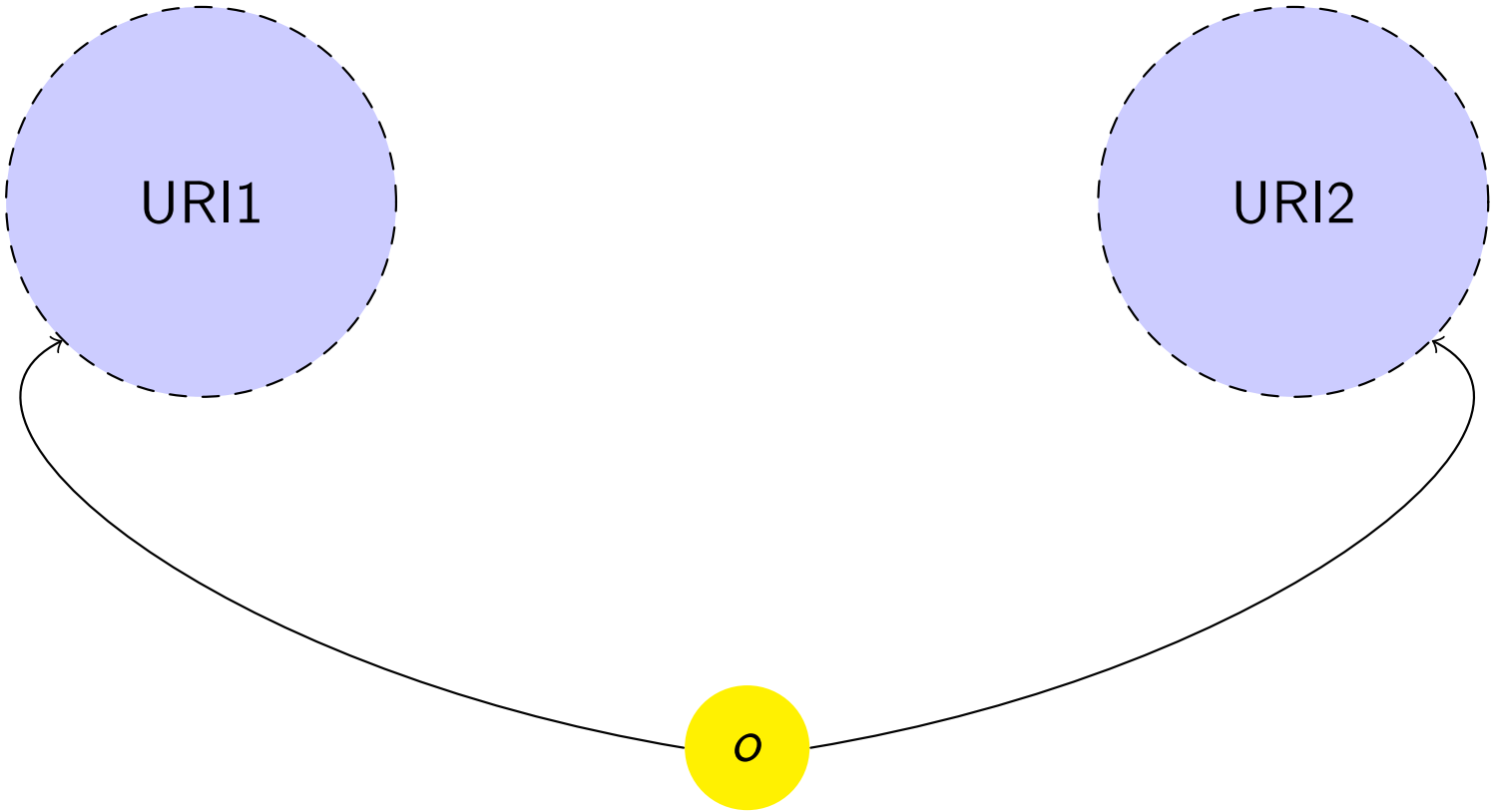
URI matching



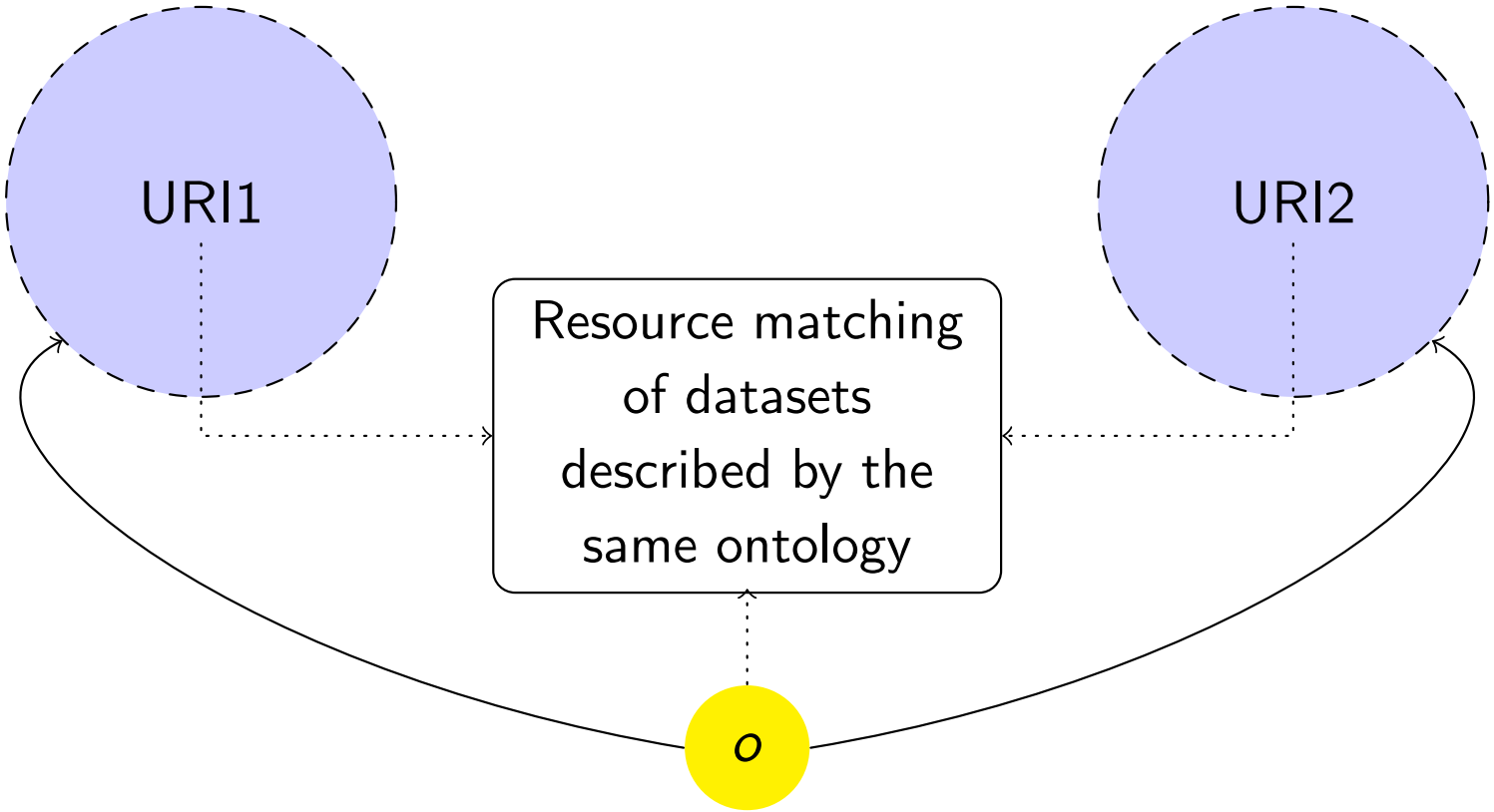

```
http://dbpedia.org/resource/Johann_Sebastian_Bach owl:sameAs
    http://www.lastfm.fr/music/Johann+Sebastian+Bach

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

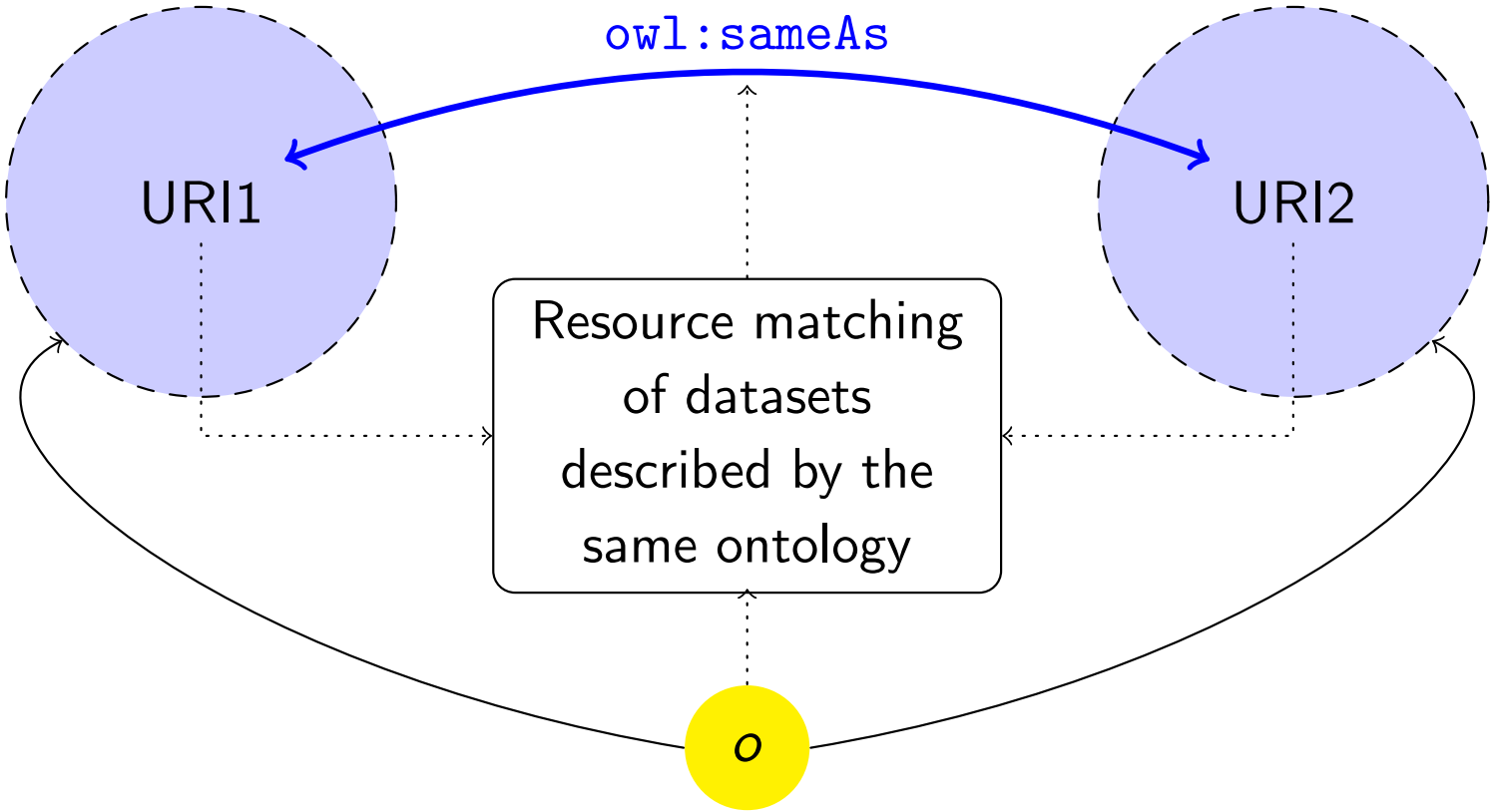
Data matching through a common ontology



Data matching through a common ontology



Data matching through a common ontology



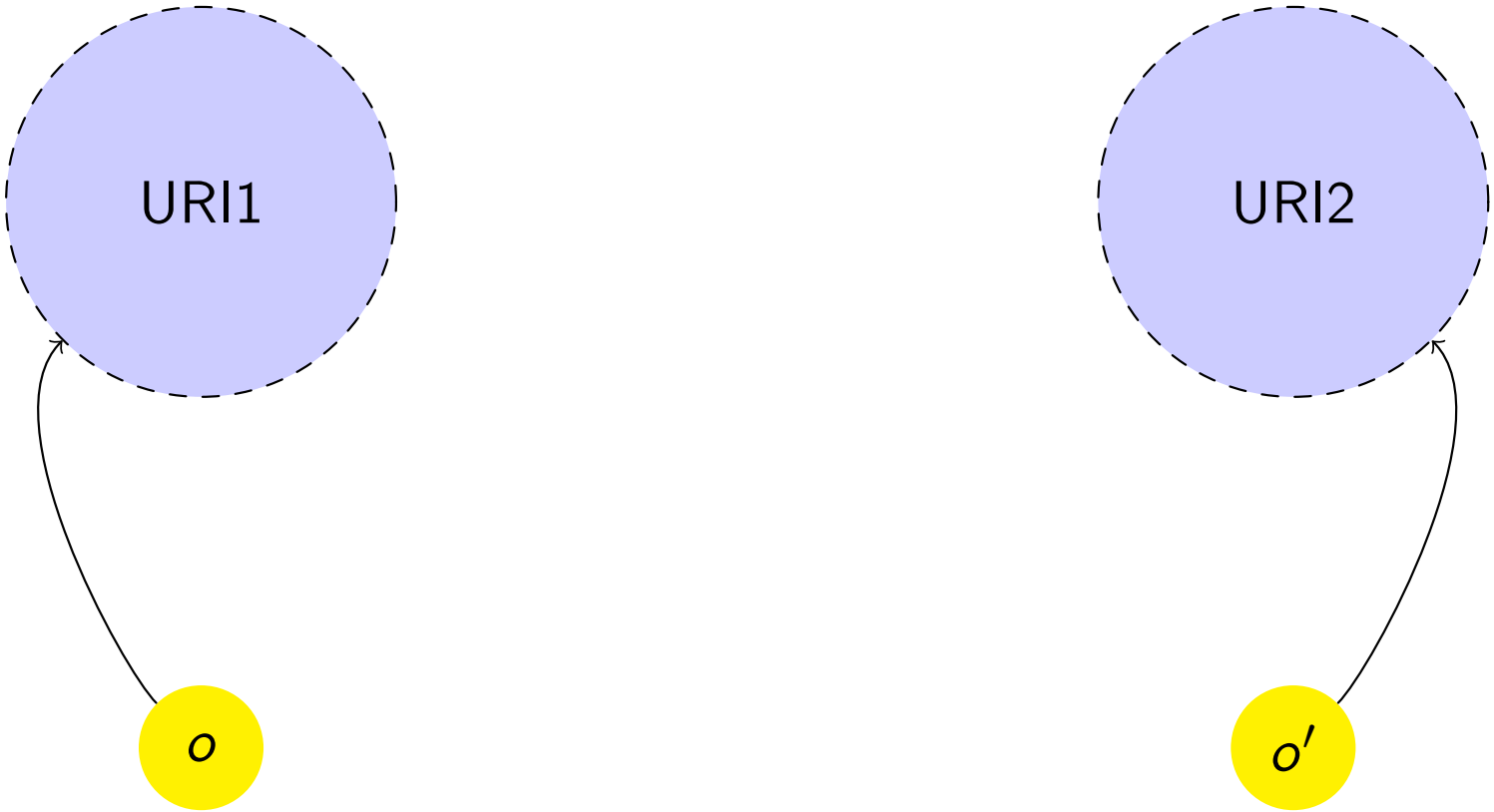
Matching with a common ontology

+ 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.

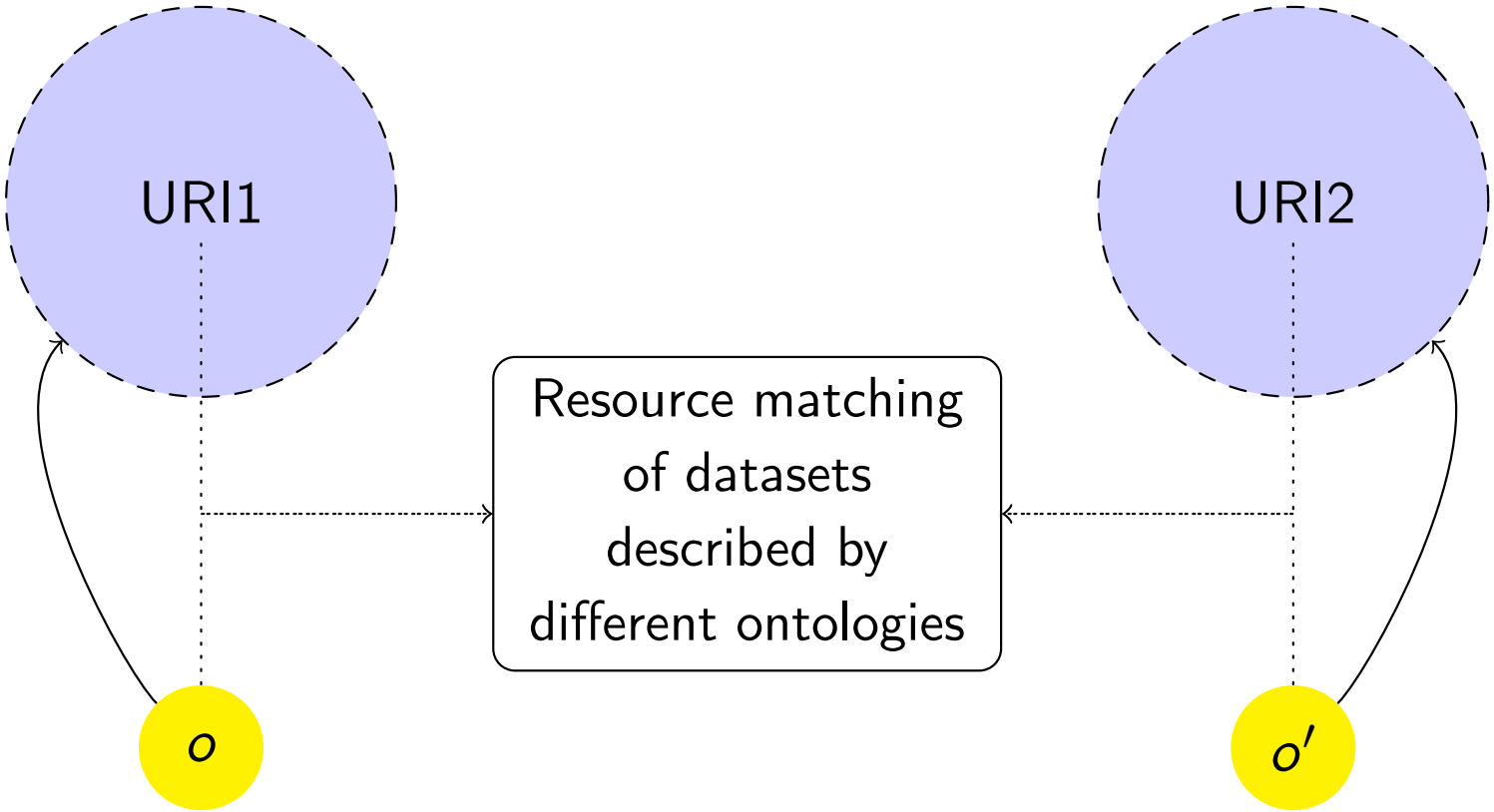
Record linkage

Having a common ontology does not solve the problem.

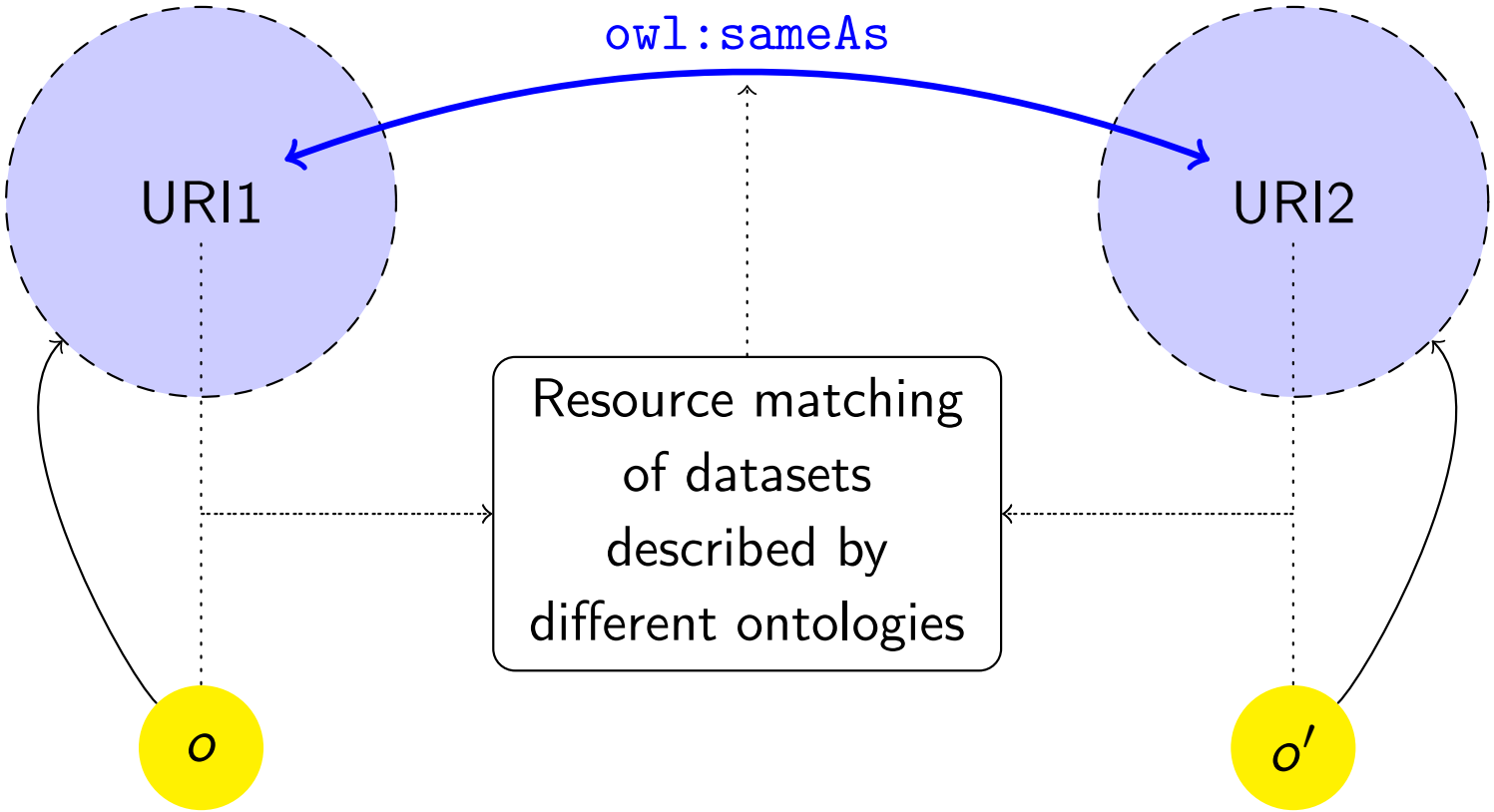
Data matching with different ontologies (implicit alignment)



Data matching with different ontologies (implicit alignment)



Data matching with different ontologies (implicit alignment)



Independent key computation

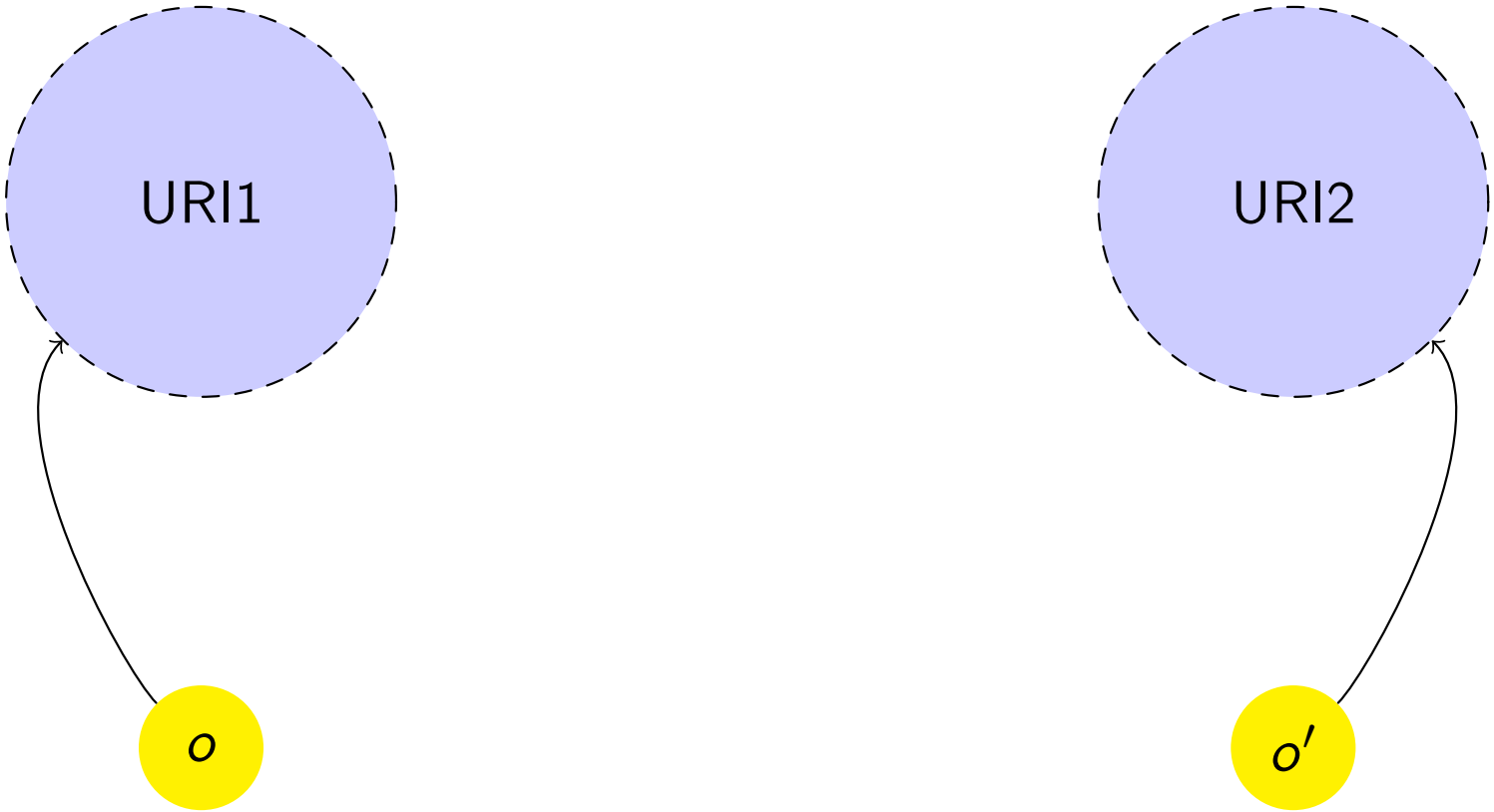
- Different span requires different key (France is not a key for INSEE);
- Differences in schema and depths makes difference in what is a key ("Paris" is both a department name (DEP_75) and a municipality name (COM_75056) for INSEE while the region name may be a key for NUTS)
- Keys are often meaningless: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

Independent key computation

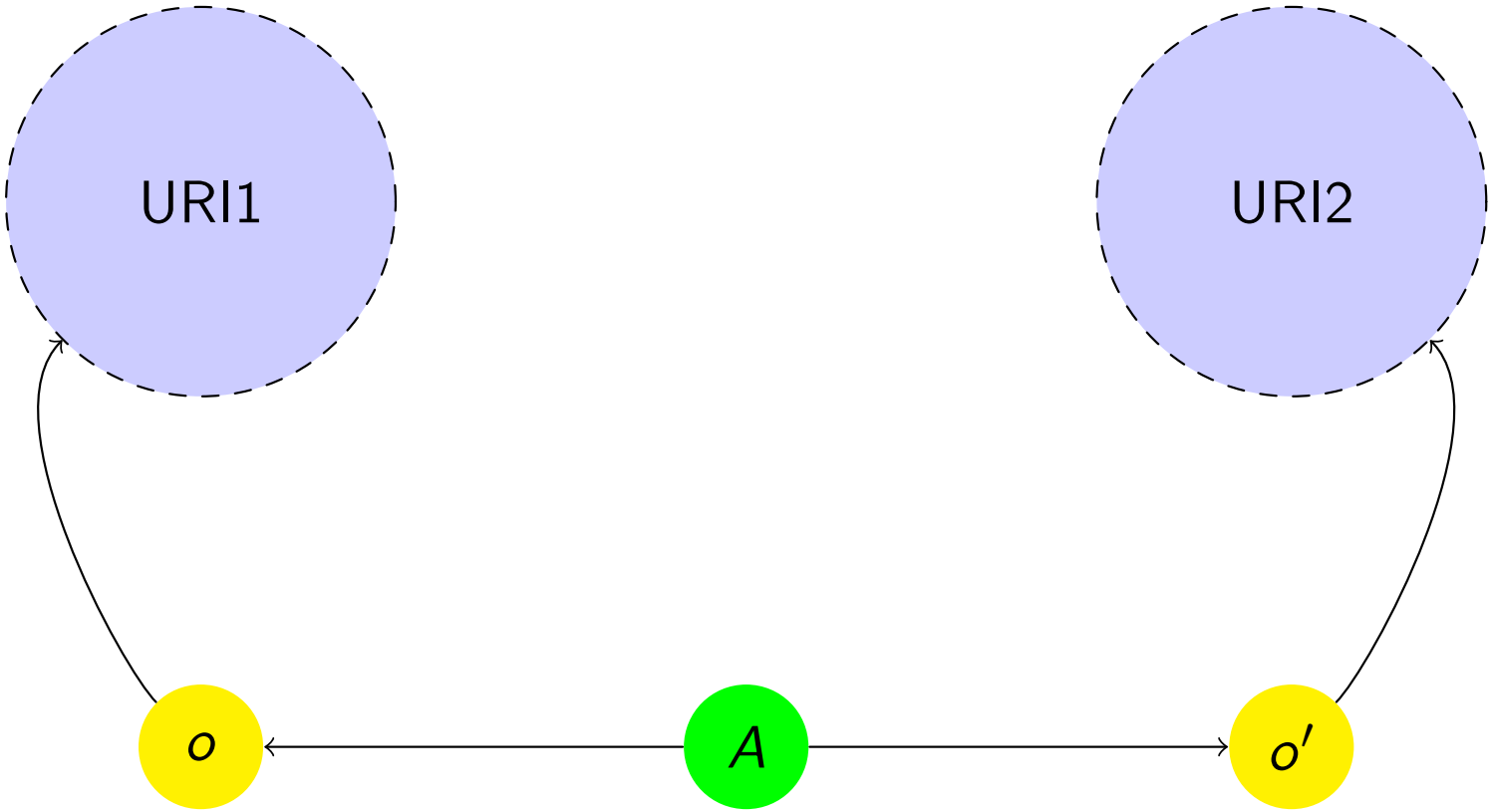
- Different span requires different key (France is not a key for INSEE);
- Differences in schema and depths makes difference in what is a key ("Paris" is both a department name (DEP_75) and a municipality name (COM_75056) for INSEE while the region name may be a key for NUTS)
- Keys are often meaningless: 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.

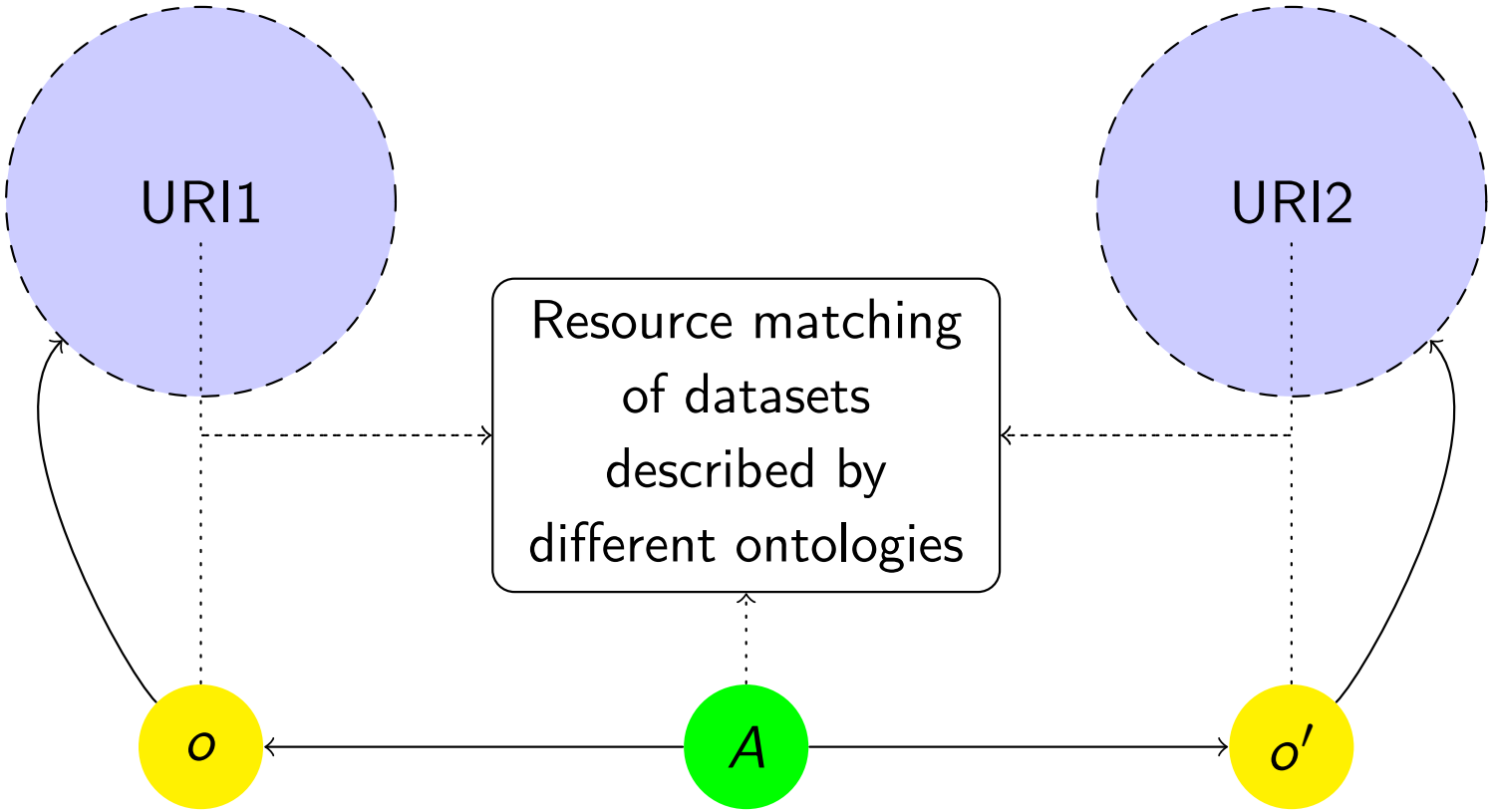
Data matching with different ontologies (explicit alignment)



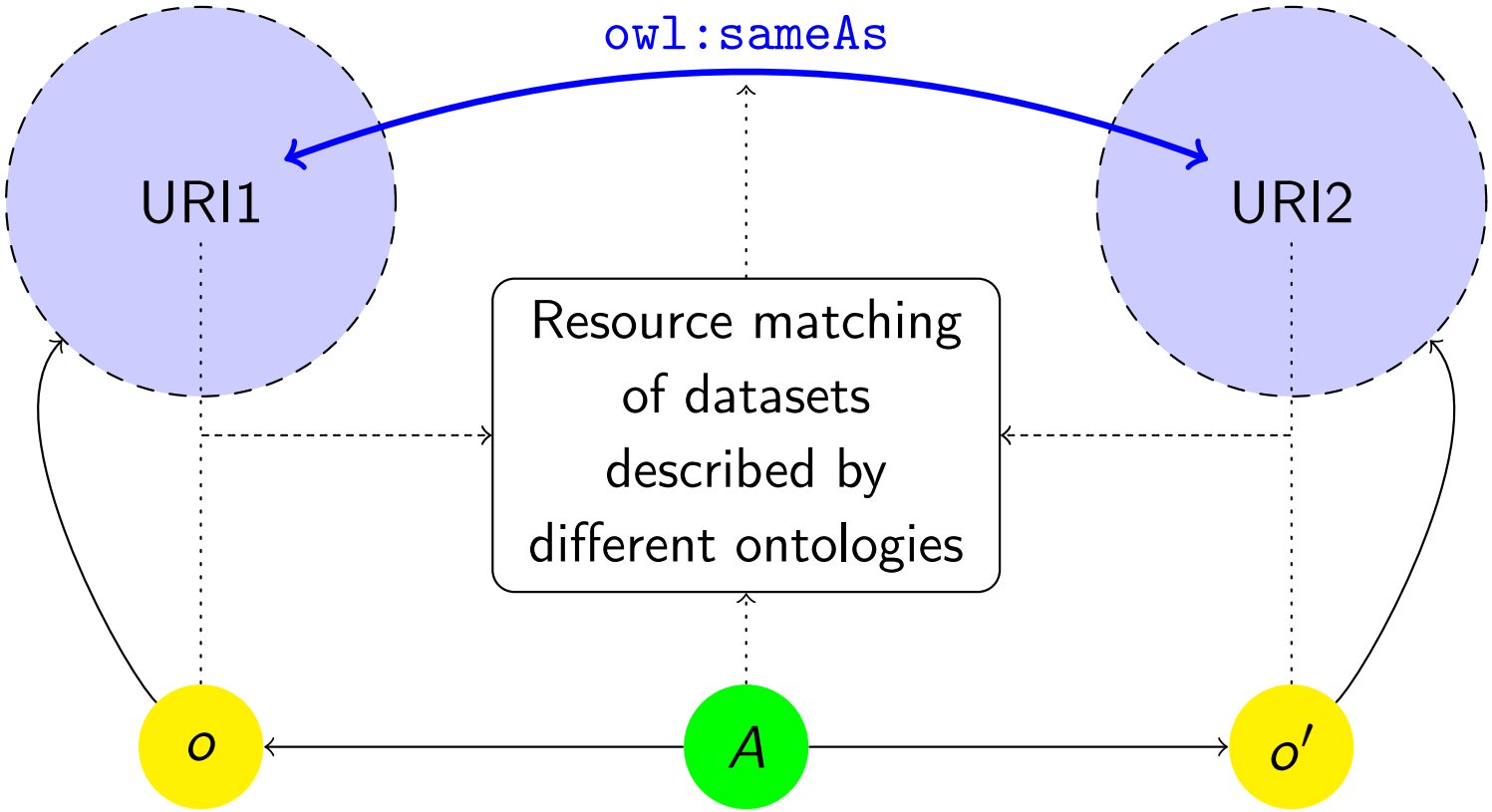
Data matching with different ontologies (explicit alignment)



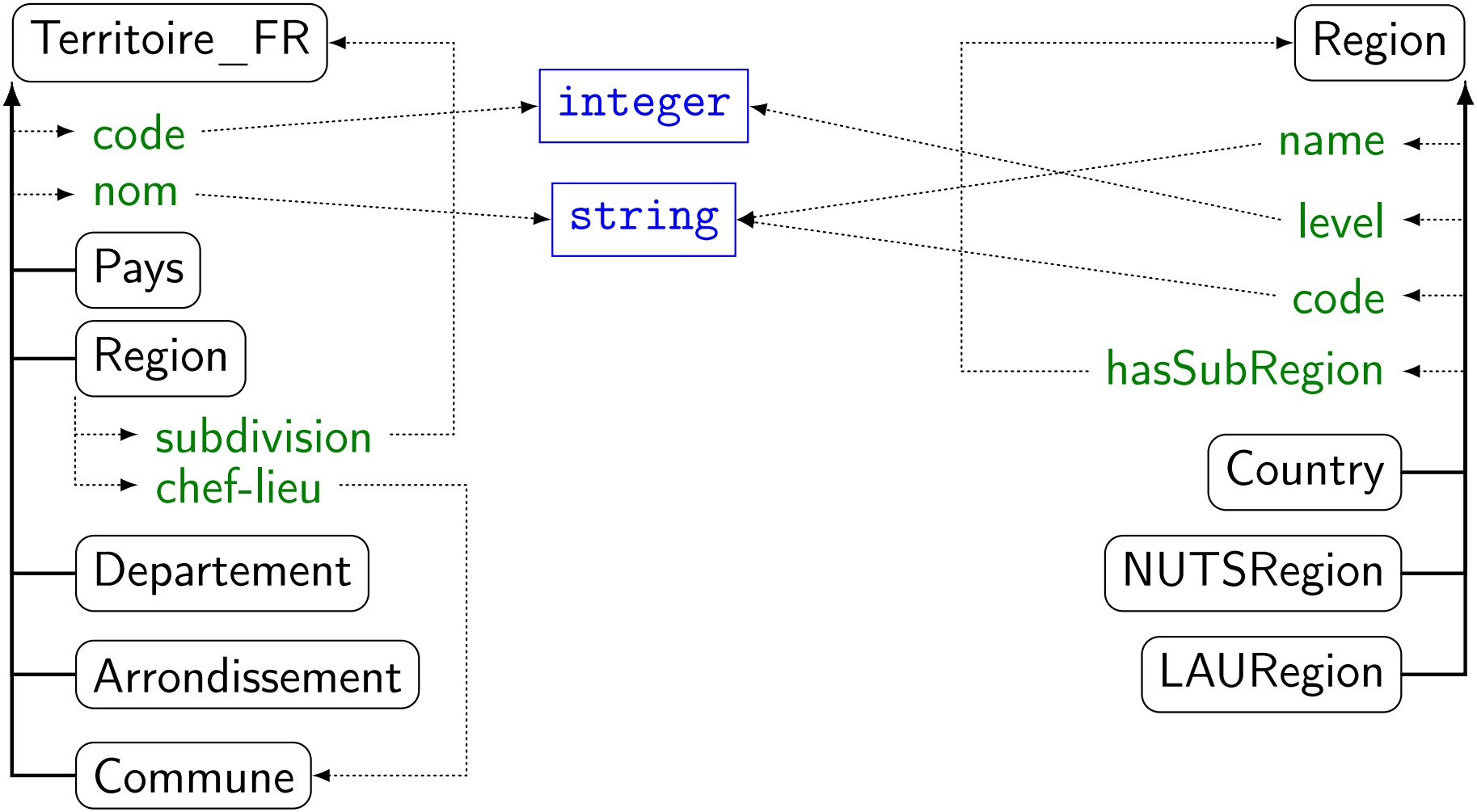
Data matching with different ontologies (explicit alignment)



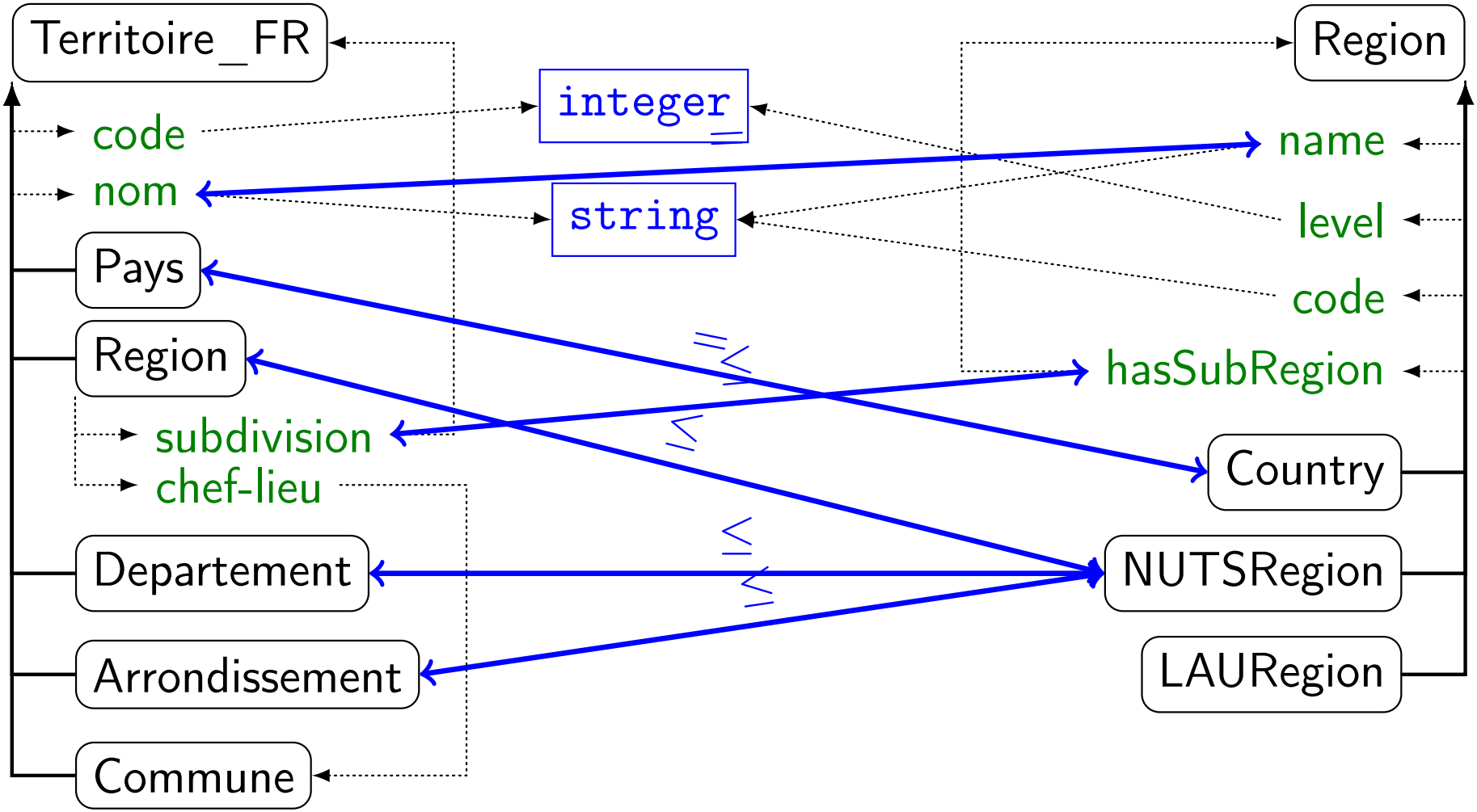
Data matching with different ontologies (explicit alignment)



INSEE and NUTS: ontology alignment



INSEE and NUTS: ontology alignment



Ontology alignment

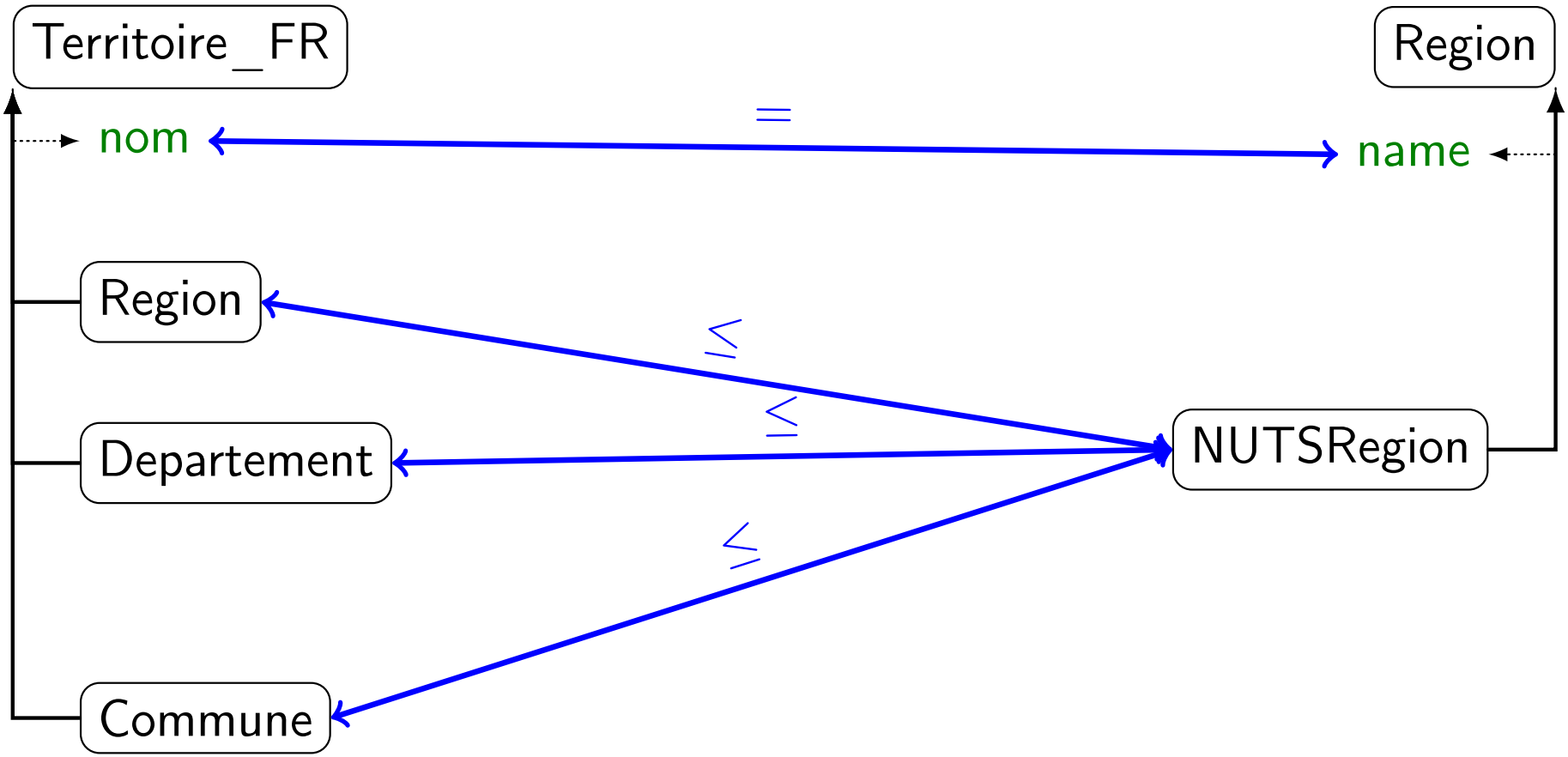
+ Ontology alignment restore the benefits of common ontology: it helps focussing the search;
– It is not exact science! (but alignments may be available);
+ Ontology alignment and data linking reinforce each others.

A simple algorithm

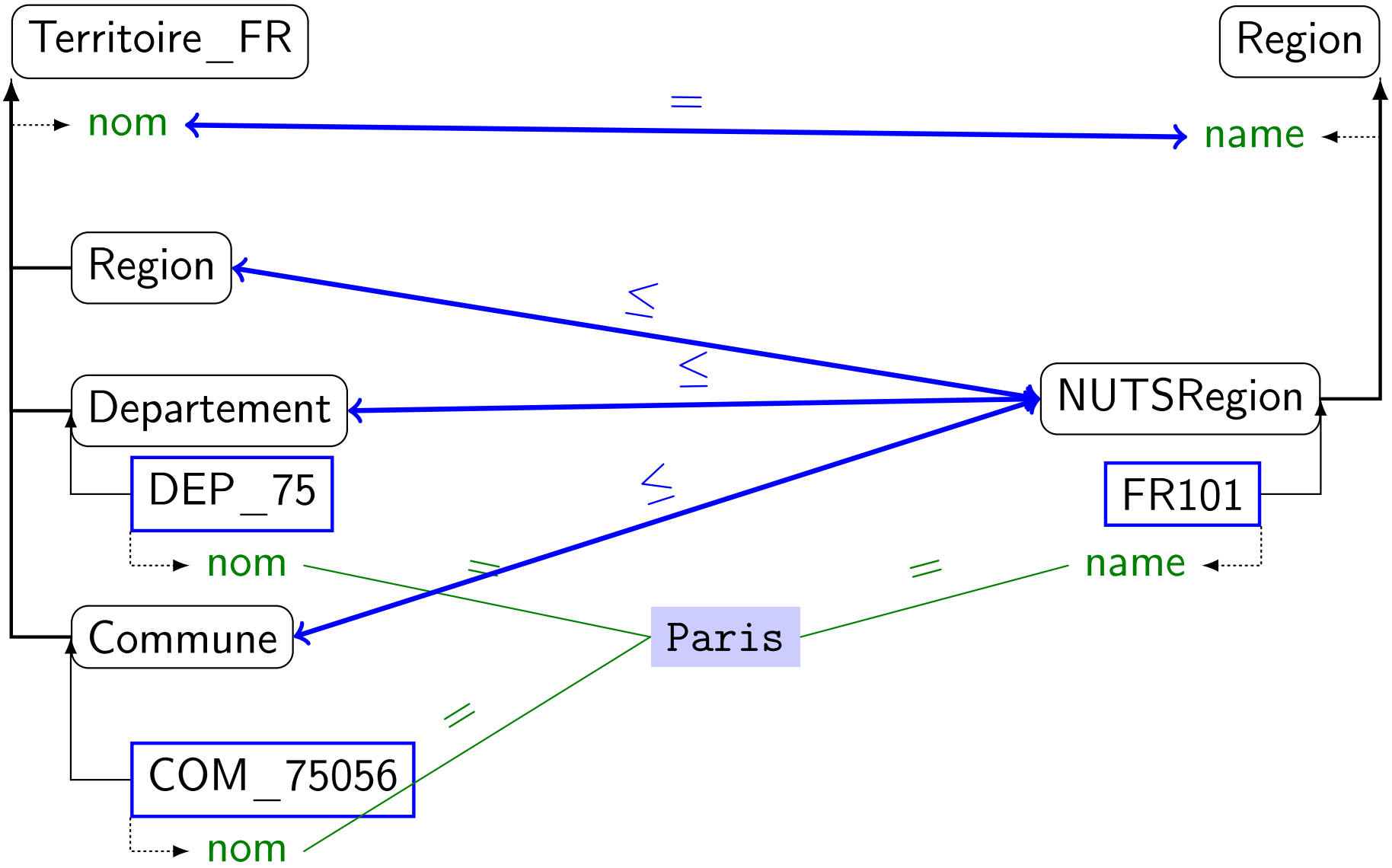
- Find matching concepts [concept matching];
- For each of them, determine matching properties based on the similarity between their values in both datasets [property matching];
- From them find property combinations identifying corresponding entities [key extraction];
- Link corresponding entities [link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

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

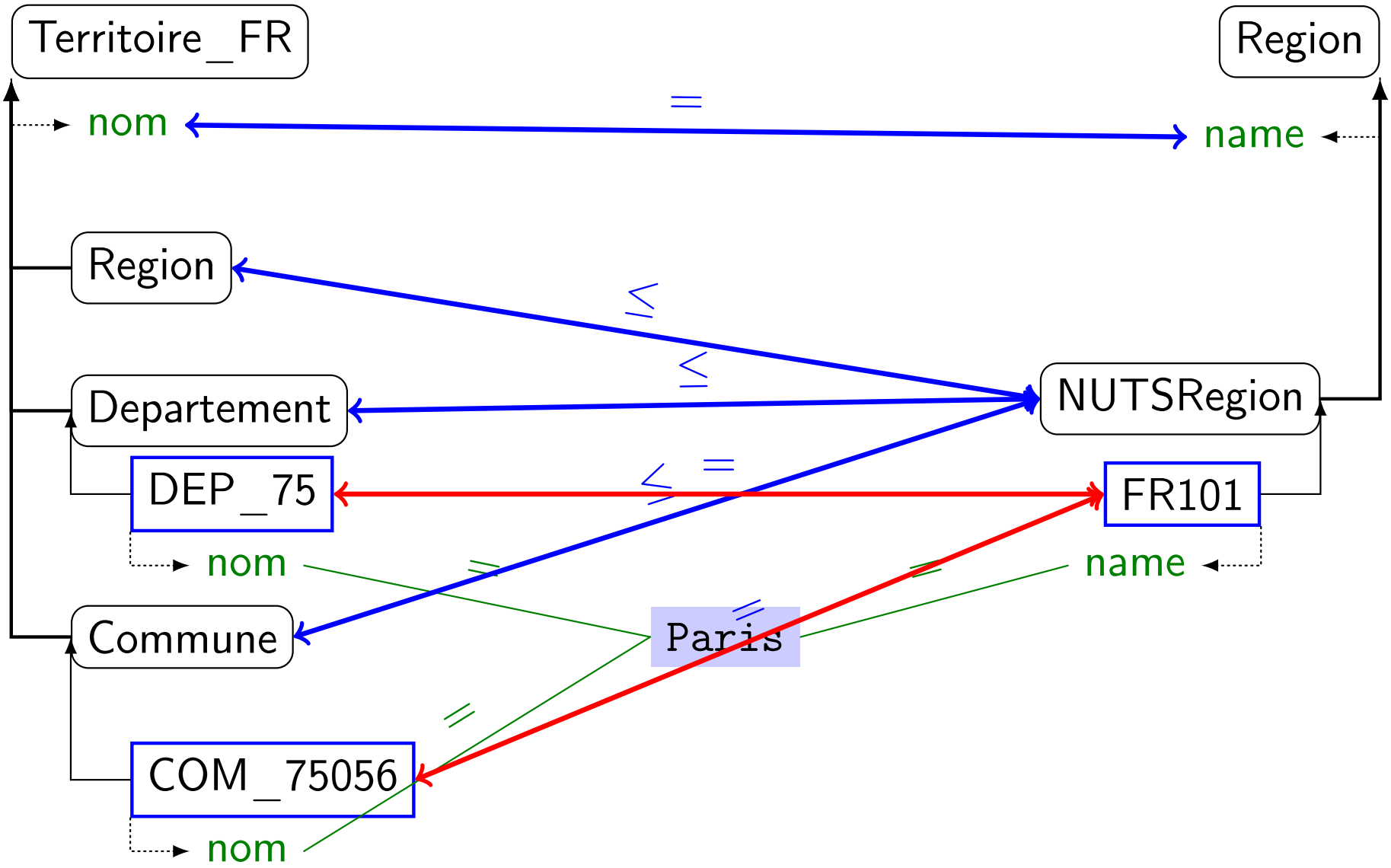
Simple alignments are not sufficient



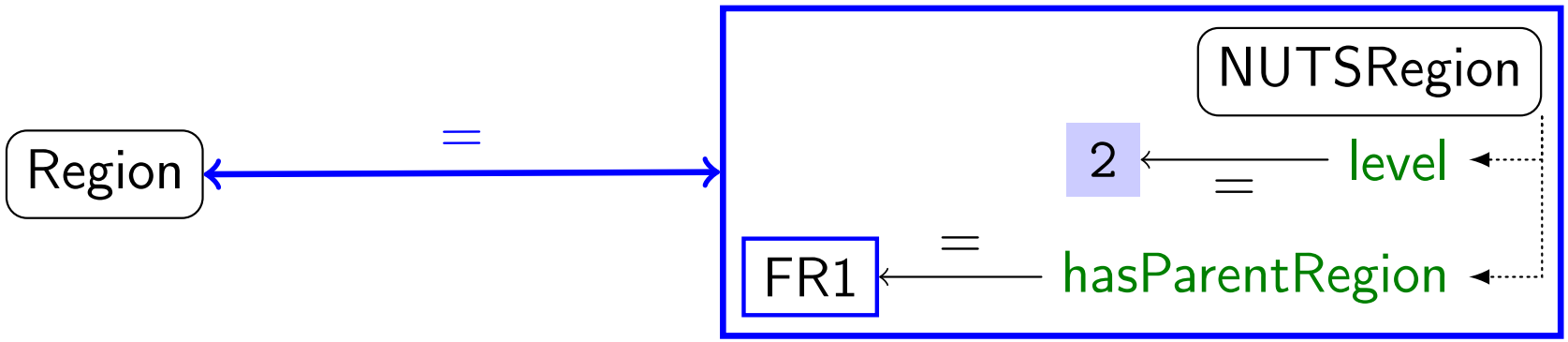
Simple alignments are not sufficient



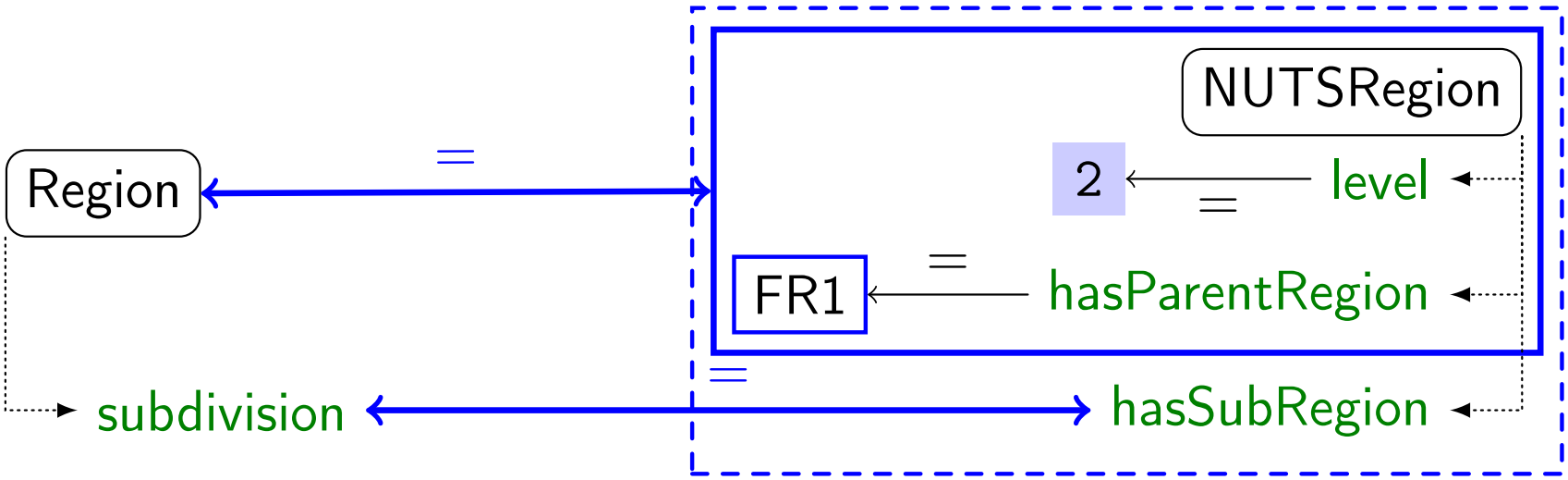
Simple alignments are not sufficient



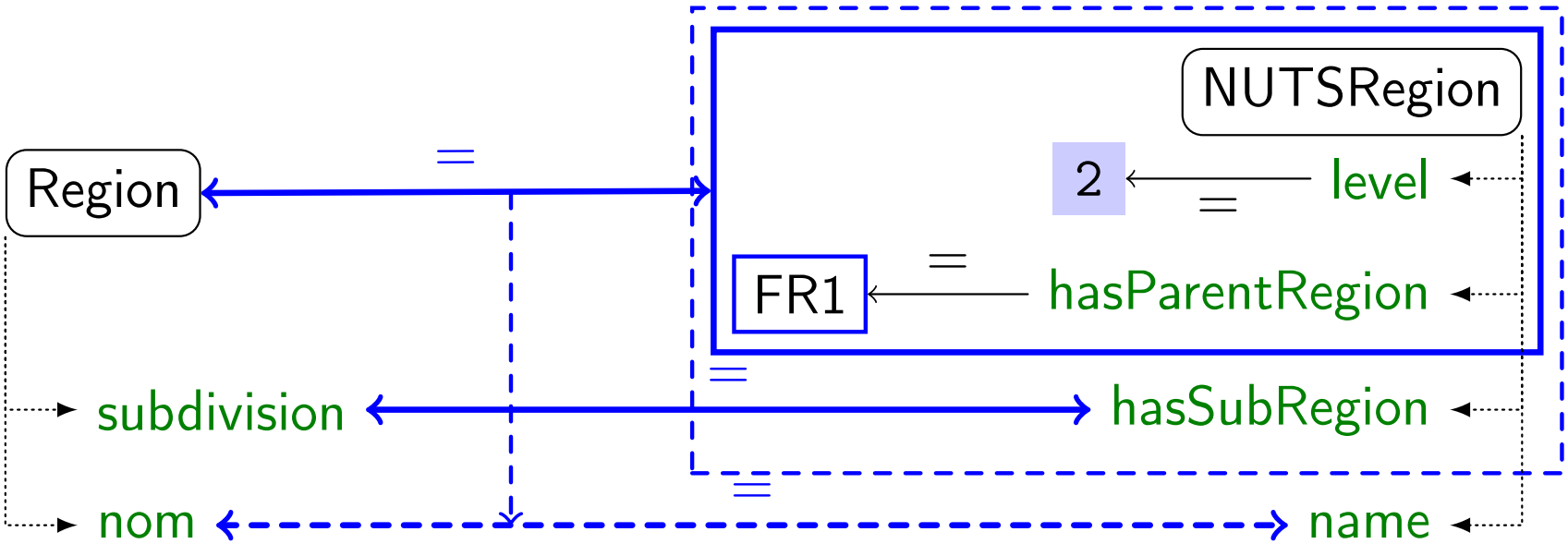
Expressive alignments are necessary



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 .
}
```

Outline

What does this mean?

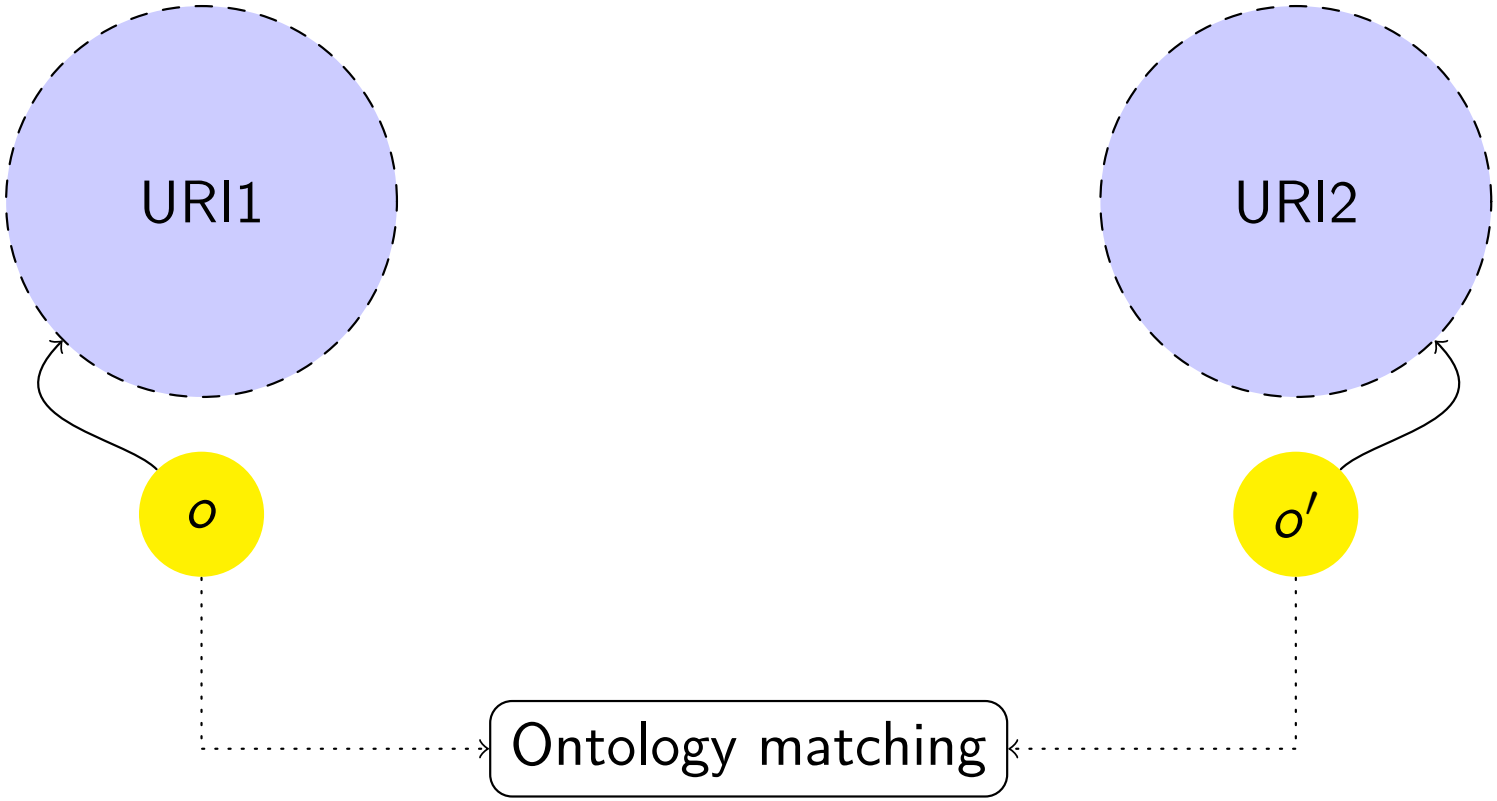
- Ontology alignments are schema-level expression of correspondences;
- They are useful for focussing the search;
- Expressive alignments are necessary;
- They can be turned into SPARQL-based link generators.

but it is also necessary to express instance level constraints:

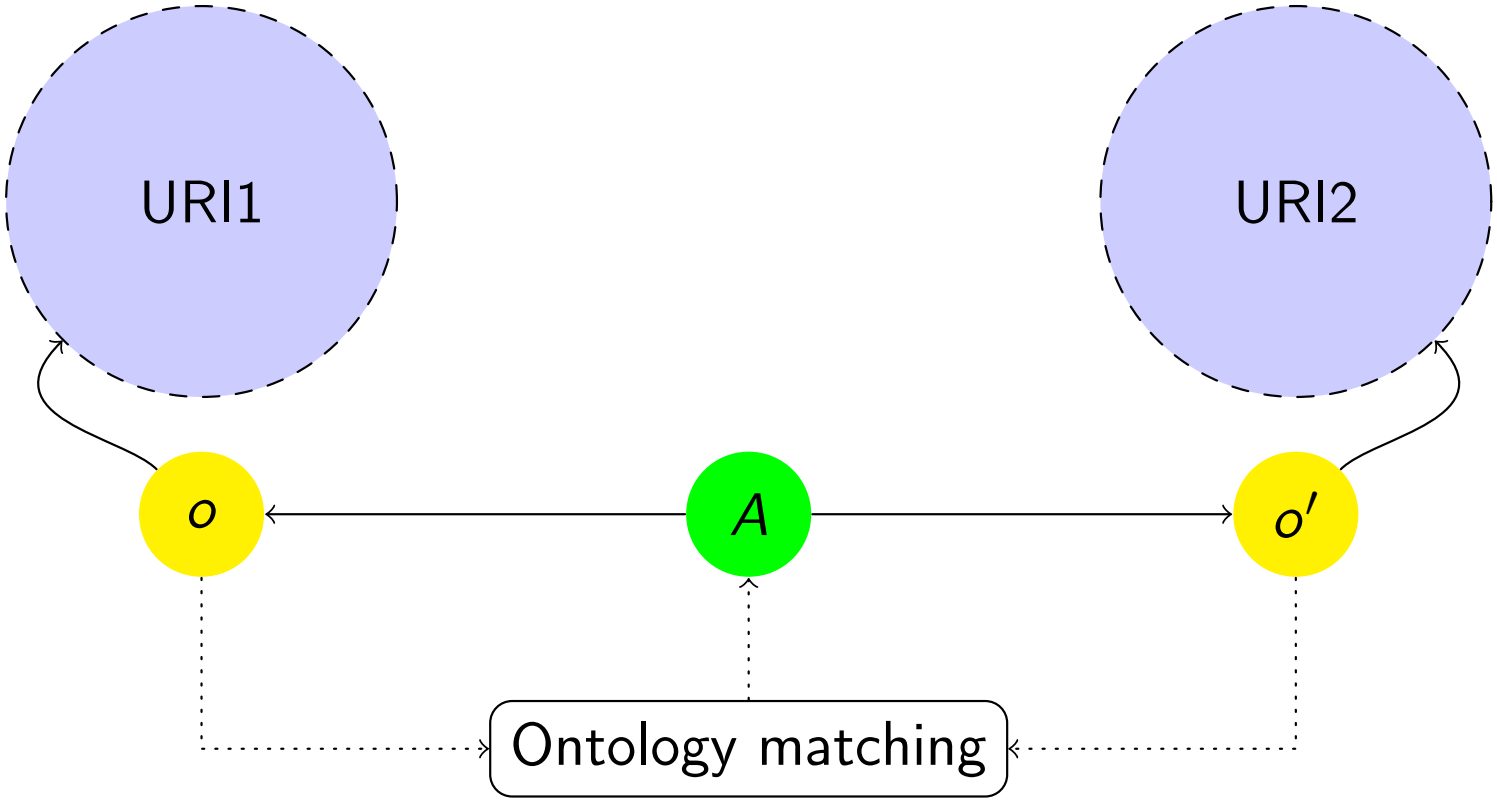
- for converting data (e.g., mph vs. m/s);
- for expressing matching constraint on data (e.g., similarity).

General framework

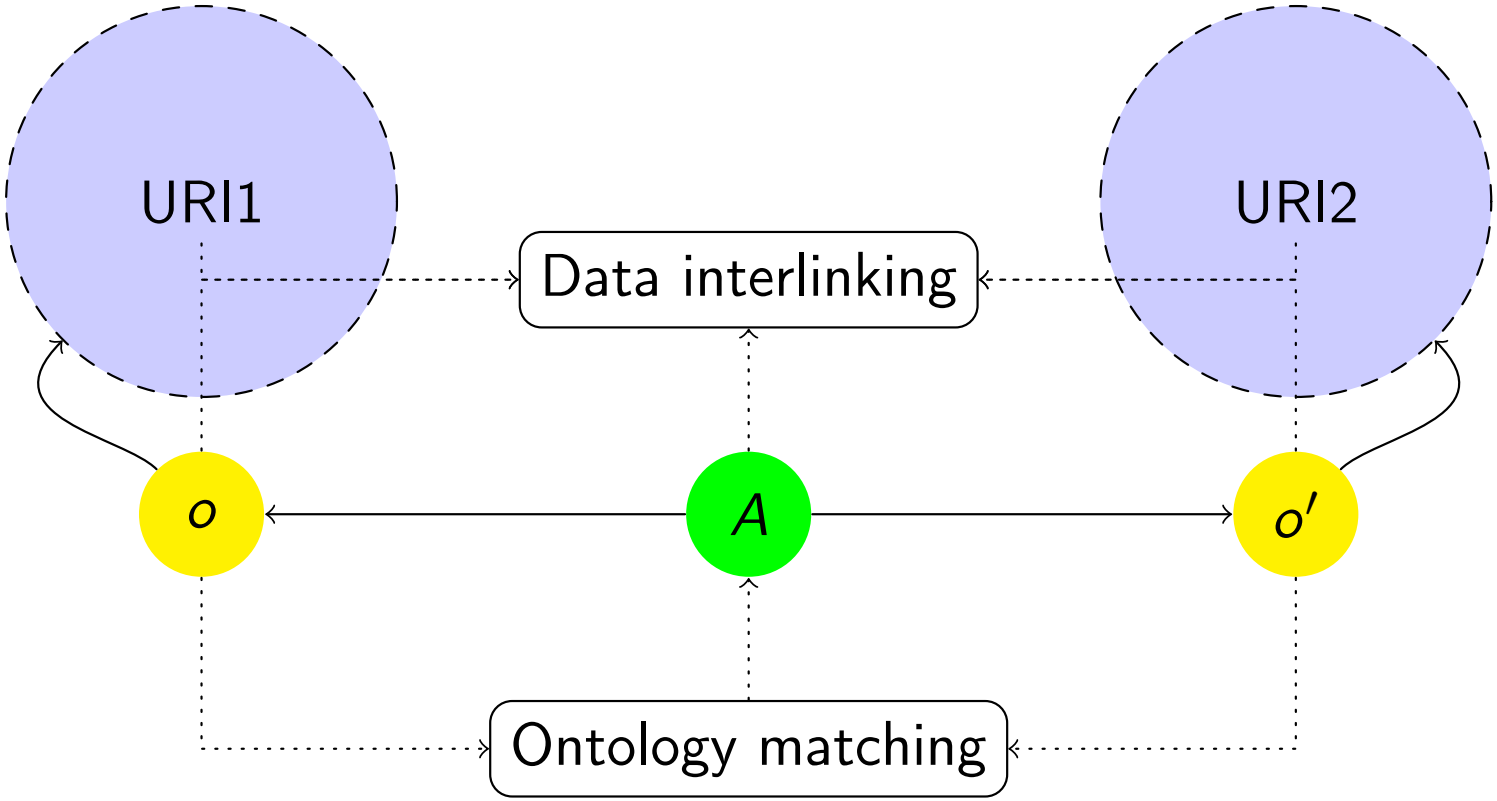
General framework



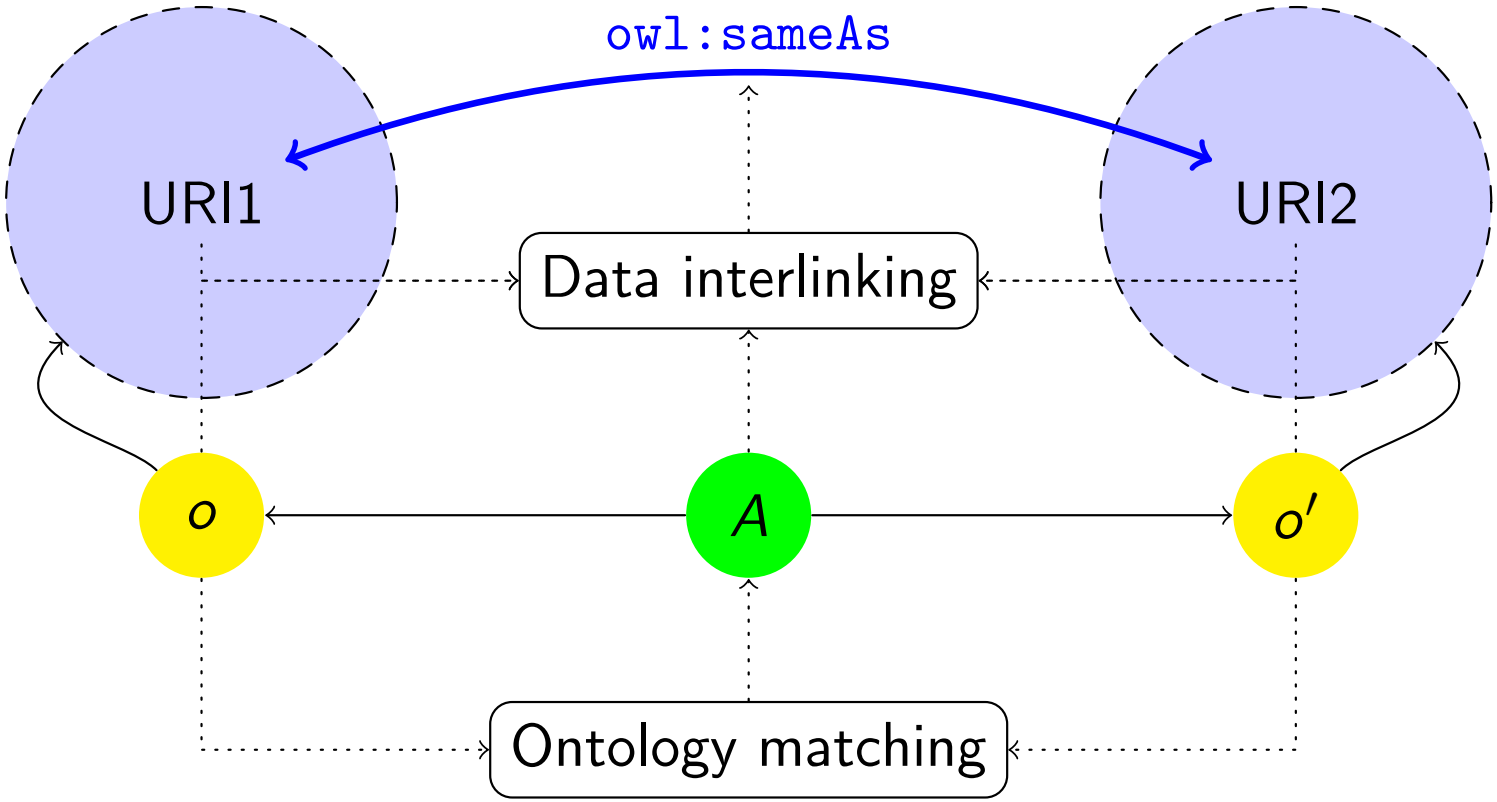
General framework



General framework



General framework



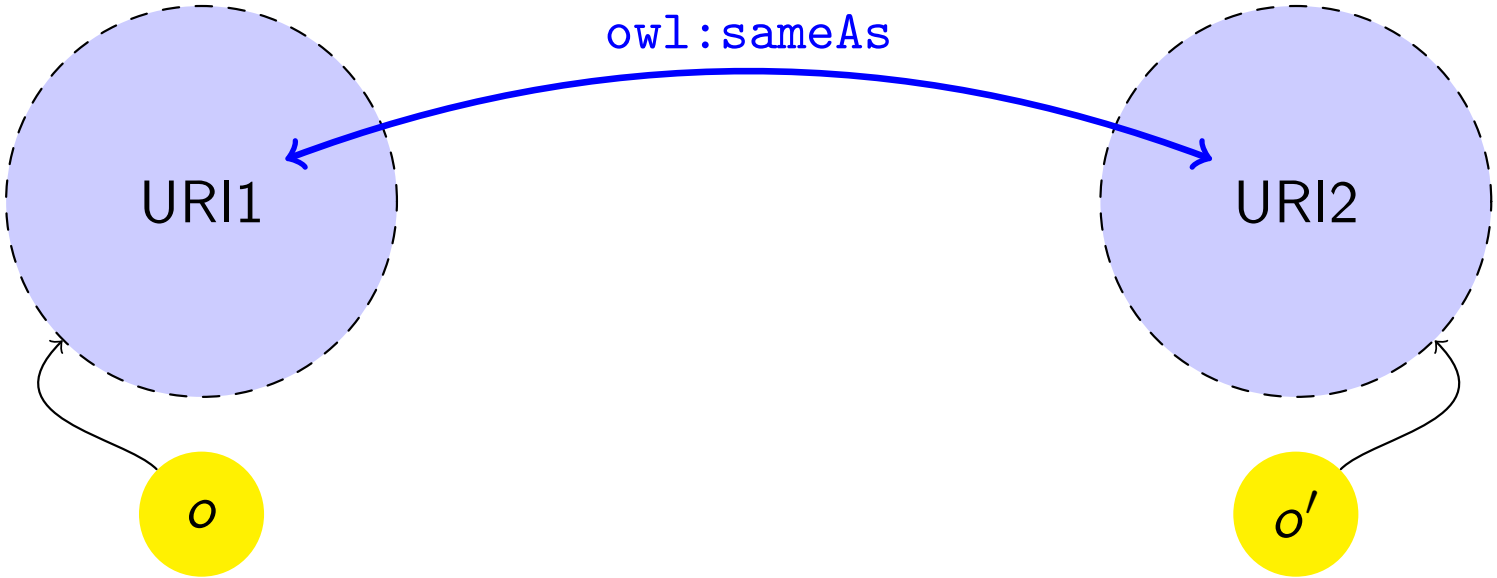
Six interlinking tools

RKB-CRS	Co-reference resolution for the RKB knowledge base.
LD-mapper	Linking tool for the Music Ontology.
ODD Linker	SQL-based linker.
RDF-AI	Dataset linker and merger.
Silk	Linking script engine based on explicit linking description.
KnoFuss	Alignment based linker.

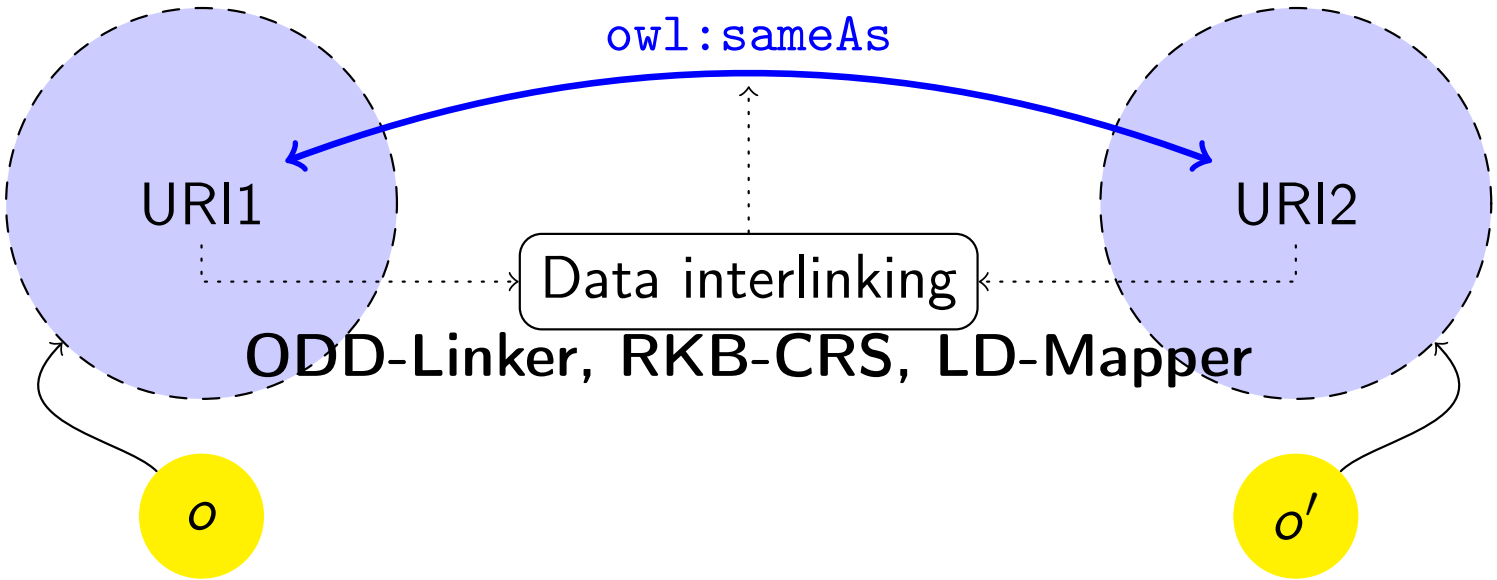
and others: ObjectCoRef, LN2R, LIMES...

`http://melinda.inrialpes.fr`

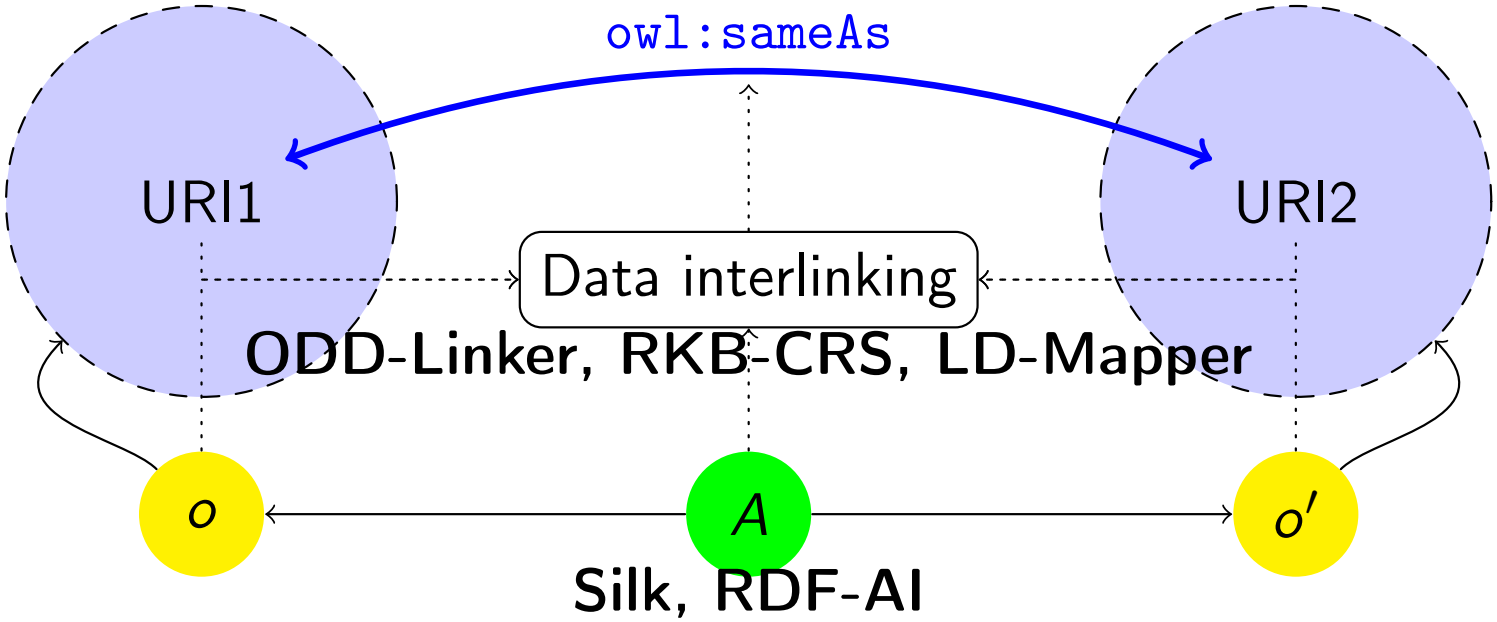
Six interlinking tools



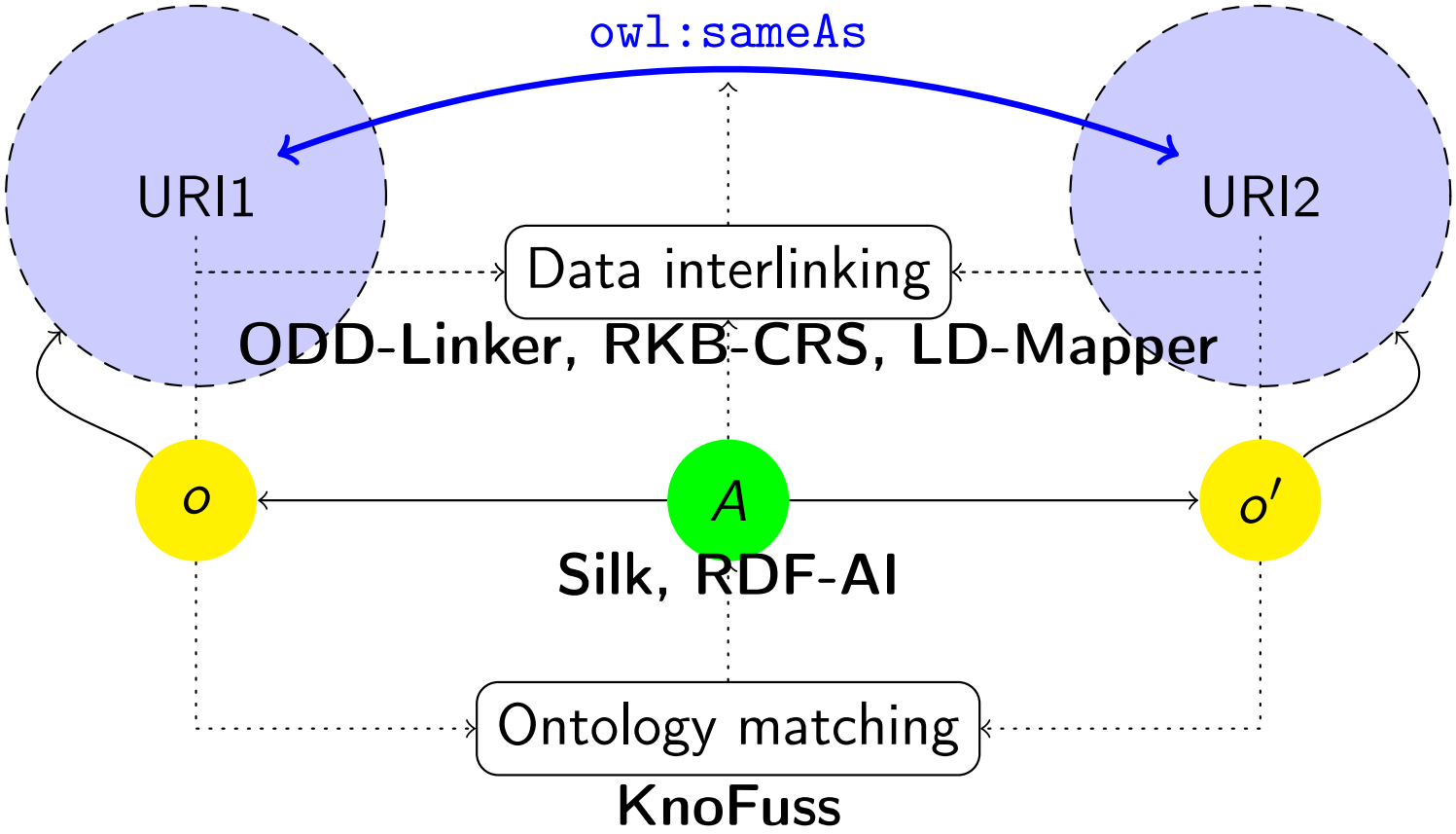
Six interlinking tools



Six interlinking tools



Six interlinking tools



Directions

- Combination between ontology matching and data interlinking;
- Explicit use of alignments or interlinking scripts;
- Data-driven algorithms (database key computation).
- Domain-specific techniques (geographic data).

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.

Questions?

http://exmo.inrialpes.fr

Jerome . Euzenat @ inria . fr