KNOWLEDGE GRAPH COMPLETION
PART 2: DATA LINKING

FATIHA SAÏS [1]

NATHALIE PERNELLE[1]

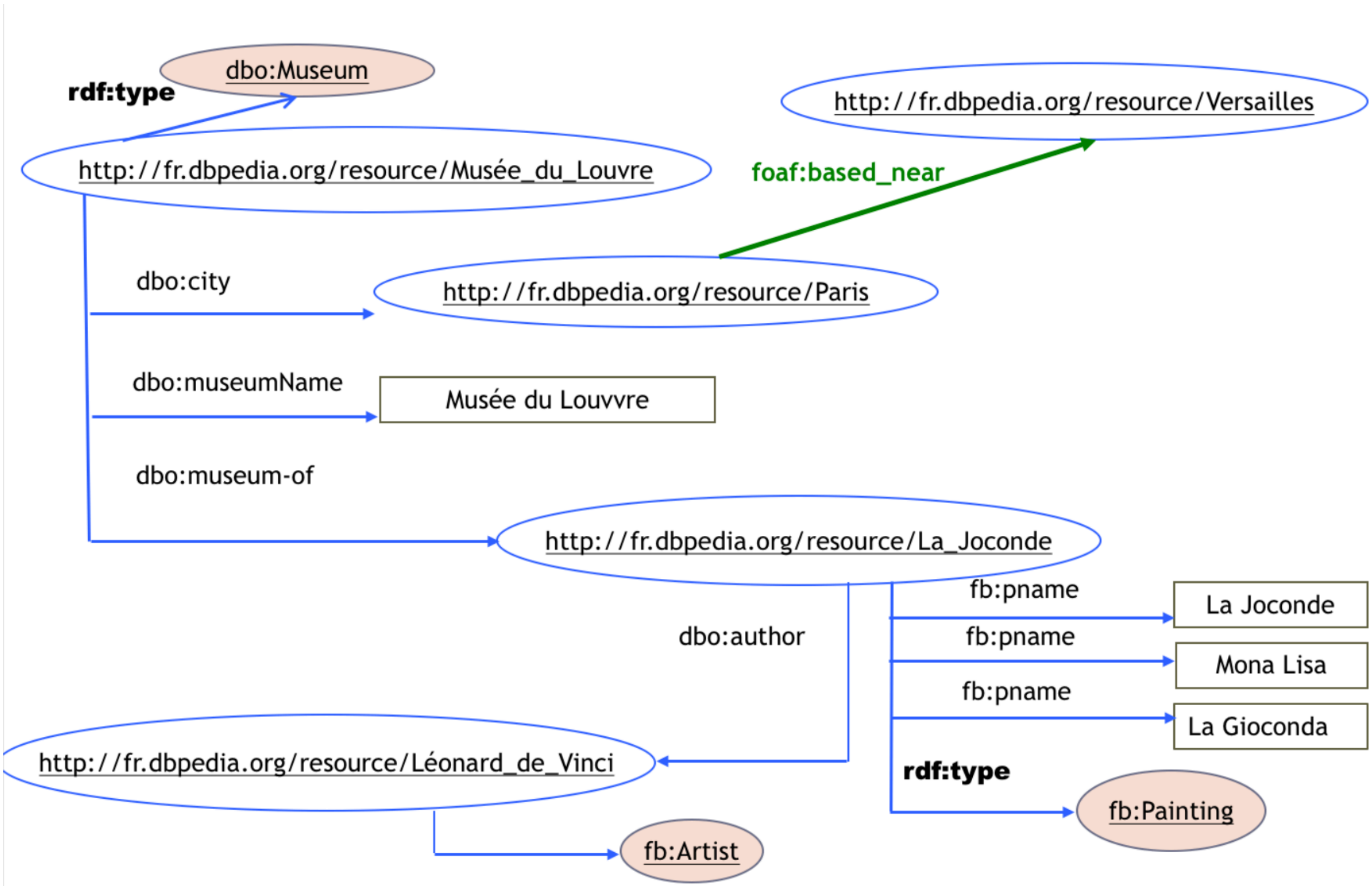
DANAI SYMEONIDOU[2]

[1] LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

[2] INRA, GAMMA TEAM

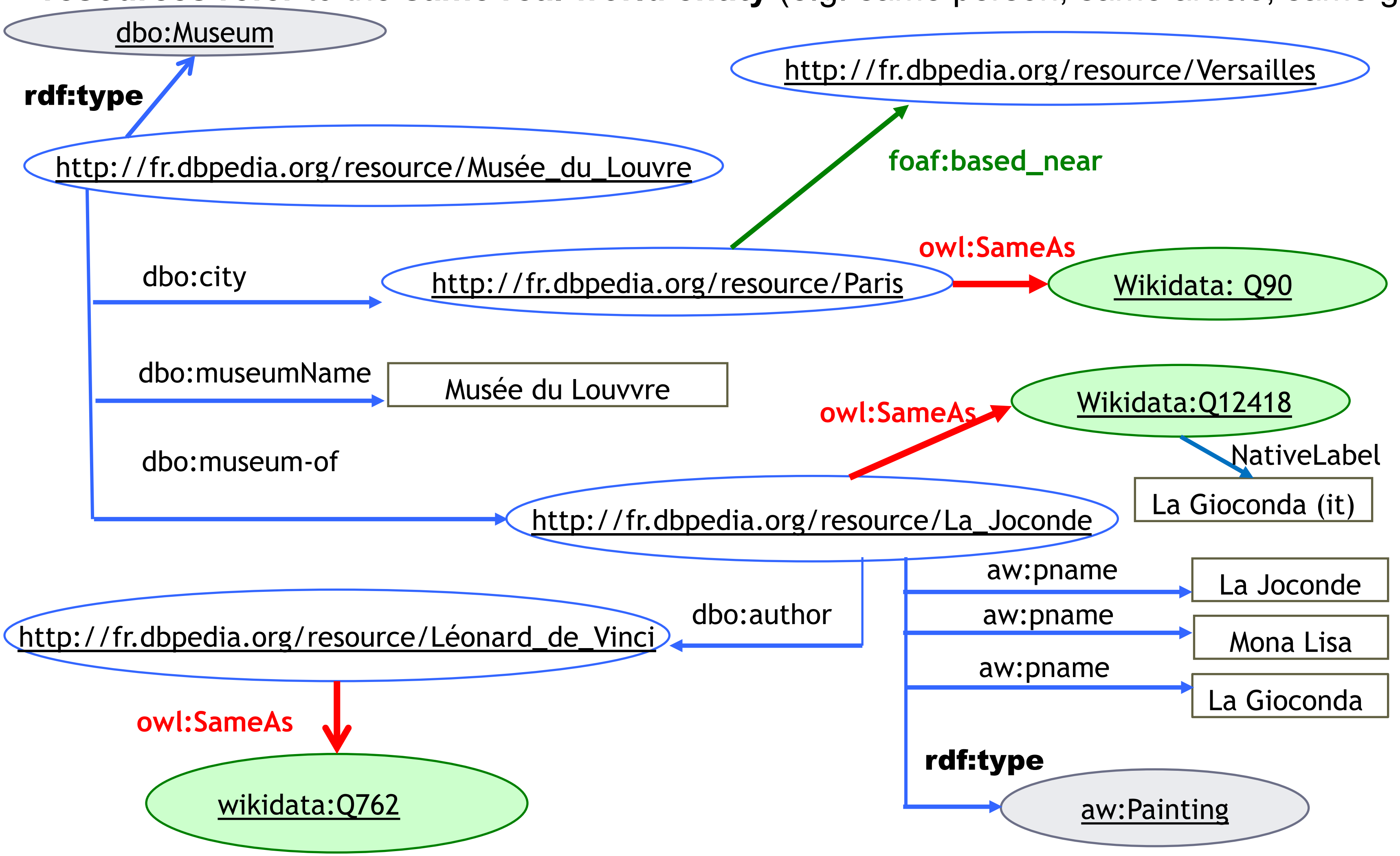
DATA LINKING

- **Data linking or Identity link detection** consists in detecting whether two descriptions of **resources refer** to the **same real world entity** (e.g. same person, same article, same gene).



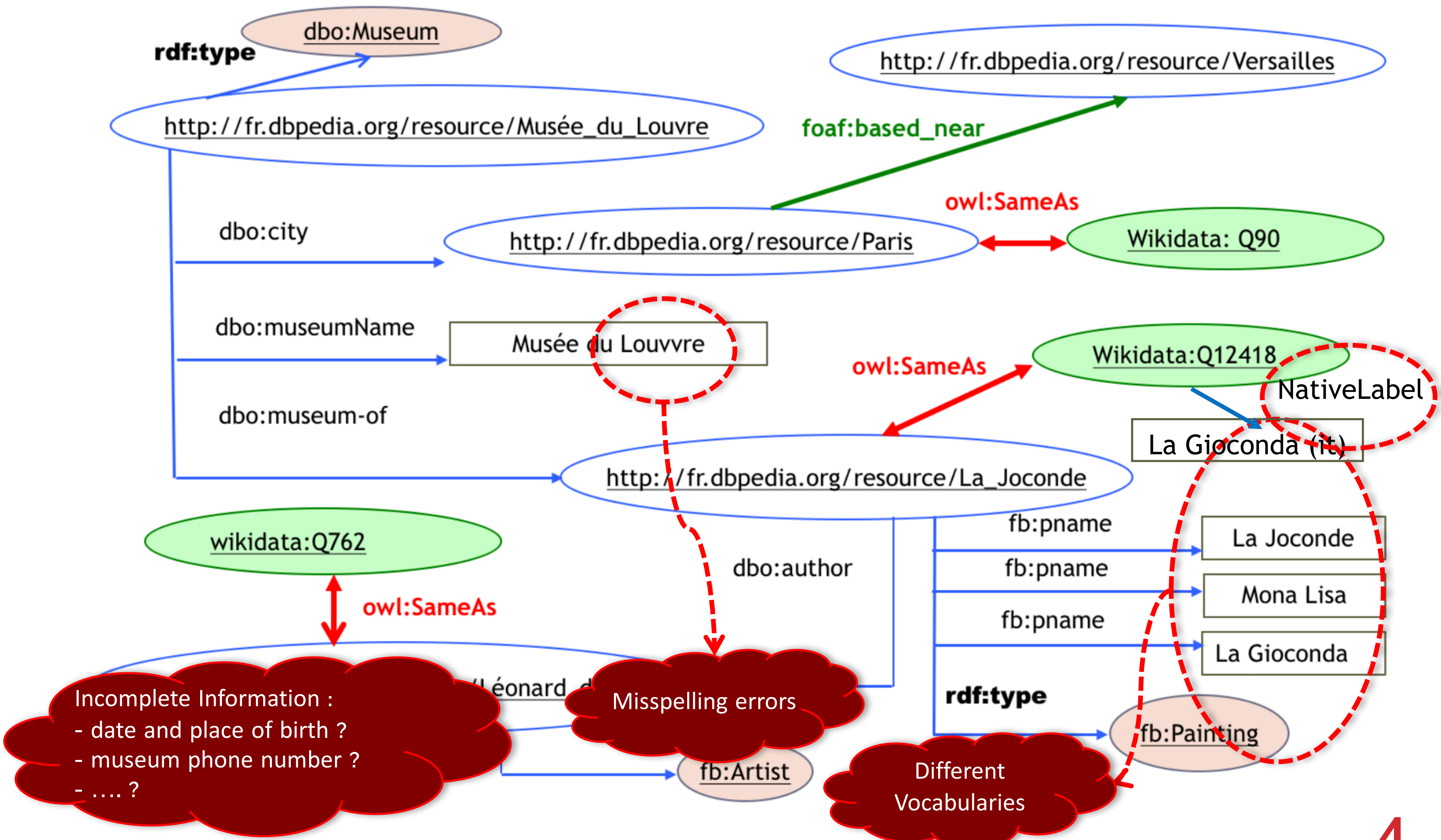
DATA LINKING

- **Data linking or Identity link detection** consists in detecting whether two descriptions of **resources refer** to the **same real world entity** (e.g. same person, same article, same gene)



DATA LINKING: DIFFICULTIES

- **Data linking or Identity link detection** consists in detecting whether two descriptions of **resources refer** to the **same real world entity** (e.g. same person, same article, same gene).



IDENTITY LINK DETECTION PROBLEM

- **Identity link detection:** detecting whether two descriptions of resources refer to the same real world entity (e.g. same person, same article, same gene).

- **Definition (Link Discovery)**
  - Given two sets U_1 and U_2 of resources
  - Find a partition of U_1 x U_2 such that :
    - S = {(s,t) $\in$ U1 × U2: owl:sameAs(s,t)} and
    - D = {(s,t) $\in$ U1 × U2: owl:differentFrom(s,t)}

- A method is **total** when $(S \cup D) = (U_1 \times U_2)$
- A method is **partial** when $(S \cup D) \subset (U_1 \times U_2)$
- **Naïve complexity** $\in O(U_1 \times U_2)$, i.e. $O(n^2)$

SOME OF HISTORY …

Problem which exists since the data exists … and under different terminologies: *record linkage, entity resolution, data cleaning, object coreference, duplicate detection, data linkage ….*

Automatic Linkage of Vital Records*

[NKAJ, Science 1959]

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

The term *record linkage* has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (*1*). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.

***Record linkage*: used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family.**

portance of repeated natural mutations on the one hand, and of fertility dif-

occurred with frequencies of about 10 percent of all record linkages involving live births and 25 percent of all link-

DATA LINKING IS MORE COMPLEX FOR GRAPHS THAN TABLES (WHY?)

	Databases	Semantic Web
Schema/Ontologies	**Same schema**	**Possibly different schema or ontologies**
Multiple types	**Single relation**	**Classes, hierarchically organized**
Open World Assumption	**NO**	**YES**
UNA-Unique Name Assumption	**Yes**	**May be no**
Data volume	**XX Thousands**	**XX Millions/Billions** **(e.g., DBpedia has 1.5 billion triples)**
Multiple values for a property	**NO**	**YES** P1 hasAuthor “Michel Chein” P1 hasAuthor “Marie-Christine Rousset”

- Can **propagate** similarity decisions ➔ more **expensive** but **better performance**
- Can be **generic** and use **domain knowledge**, e.g. ontology axioms

DATA LINKING APPROACHES: DIFFERENT CONTEXTS

- Datasets conforming to the same ontology
- Datasets conforming to different ontologies
- Datasets without ontologies

DATA LINKING APPROACHES

- **Local approaches**: consider properties to compare pairs of instances independently

versus

- **Global approaches**: consider data type properties as well as object properties to propagate similarity scores/linking decisions (collective data linking)

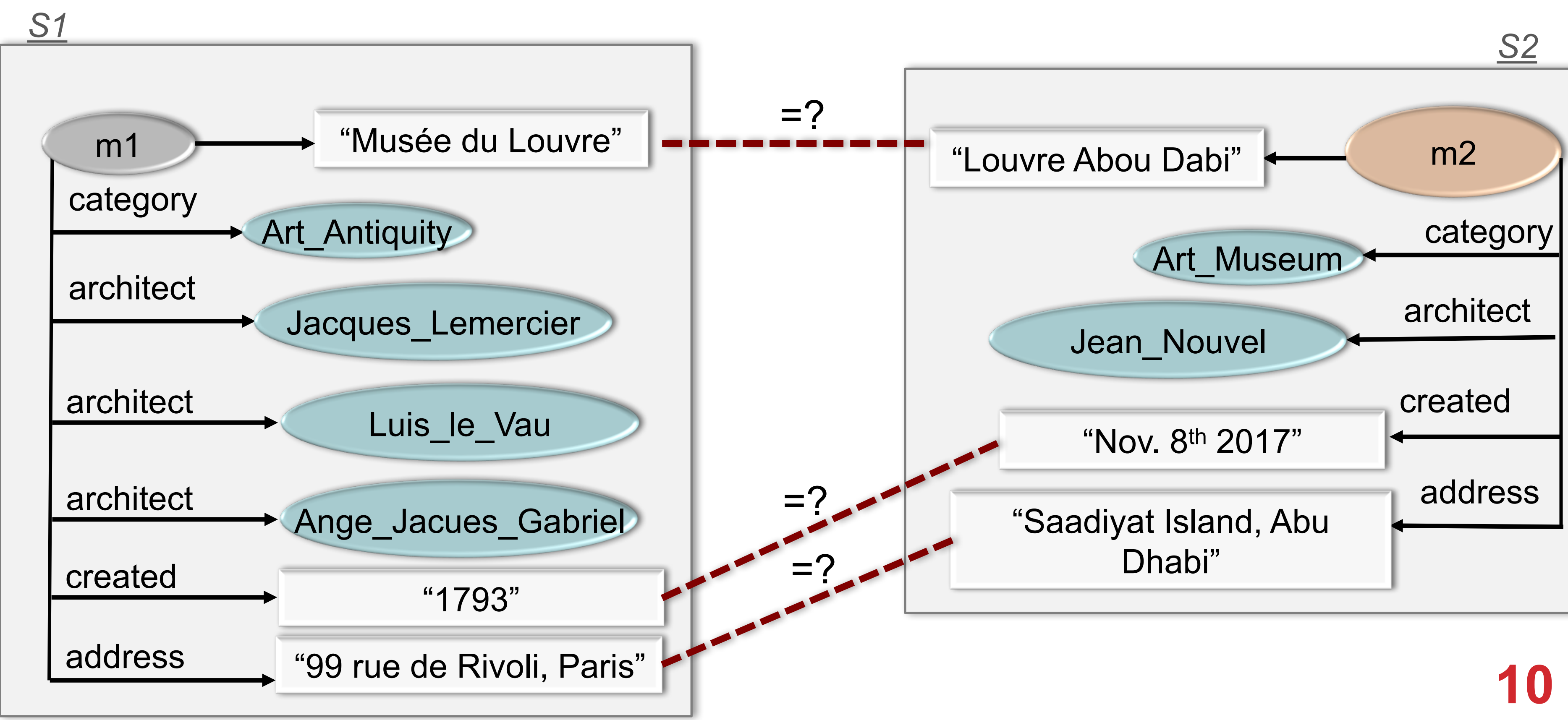
- **Supervised approaches**: need samples of linked data to learn models, or need interactions with expert

versus

- **Informed approaches**: need knowledge to be declared in the ontology or in other format

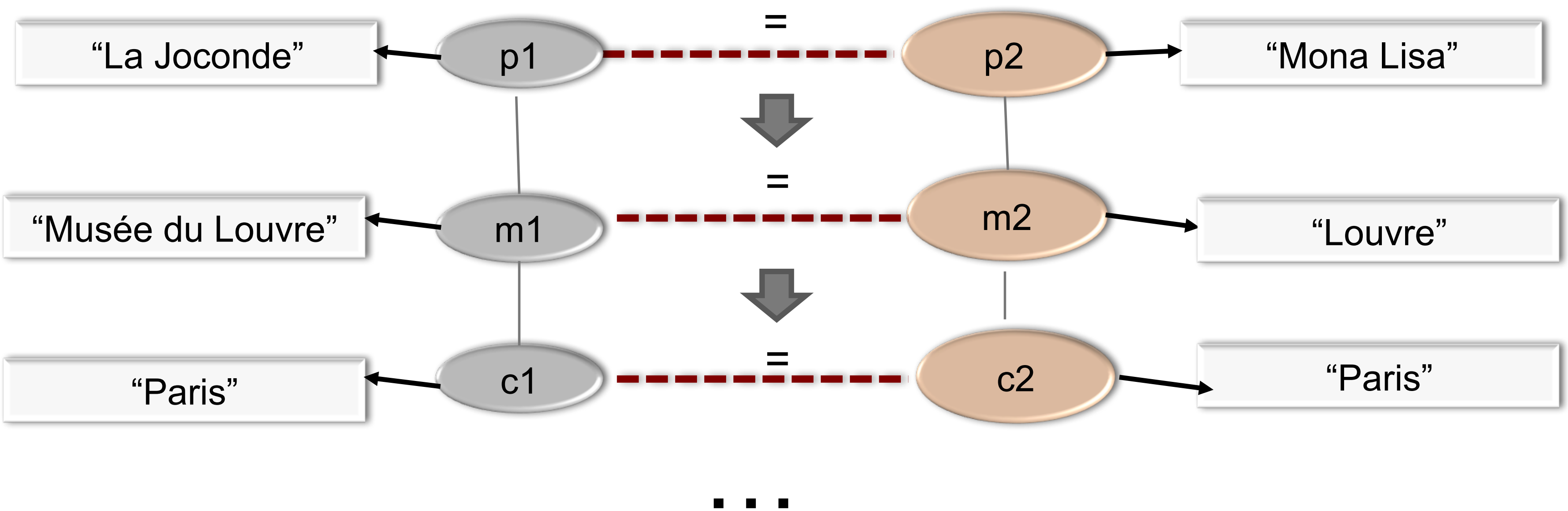
LOCAL APPROACHES

- Consider (path of) properties to compare pairs of instances independently



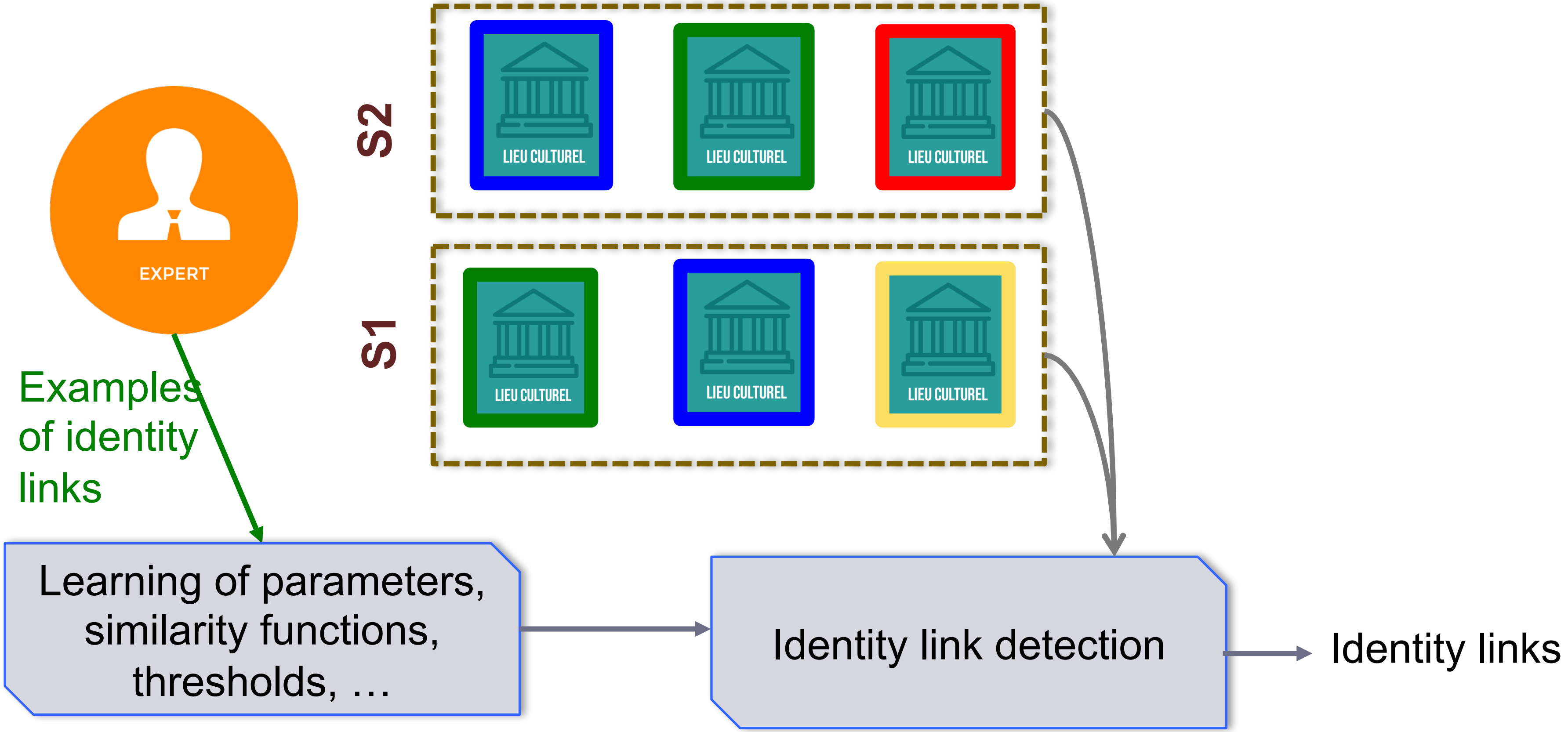
GLOBAL APPROACHES

- **Global approaches** (collective data linking): propagate similarity scores/linking decisions



SUPERVISED APPROACHES

- Need an expert to build samples of identity links to train models (or interactive approaches)



INFORMED APPROACHES

- **Informed approaches: need knowledge to be declared in the ontology or in other format**

If you know that an Home page is a key for the class Restaurant :

$$\text{homepage}(w1, y) \land \text{homepage}(w2, y) \rightarrow \text{sameAs}(w1, w2)$$

sameAs(Restaurant11, Restaurant21)
sameAs(Restaurant12, Restaurant22)
sameAs(Restaurant13, Restaurant23)

	…	homepage		homepage	…	
Restaurant11		www.kitchenbar.com	**SameAS**	www.kitchenbar.com		**Restaurant21**
Restaurant12		www.jardin.fr	**SameAS**	www.jardin.fr		**Restaurant22**
Restaurant13		www.gladys.fr	**SameAS**	www.gladys.fr		**Restaurant23**
Restaurant14		…		…		**Restaurant24**

KNOWLEDGE

Used to construct Logical Rules, numerical rules, complex similarity functions that infer sameAs, differentFrom or string equivalences

… or used to prune the search space (blocking).

- **Semantics of owl:sameAs or owl:differentFrom** (transitivity …)
- **Ontology axioms/rules about classes or properties**

  Equivalent or disjoint classes, subsumption

  (Inverse)functional properties, composite keys, graph patterns

  Linkage rules with built-in predicates

- **Referring expressions** that identify one particular instance
- **Assumptions about the datasets**

  Unique Name Assumption (UNA) or Local-UNA for properties

FROM KNOWLEDGE TO LOGICAL RULES

Keys
Example: Address + city is a composite key for the class Restaurant
$Restaurant(r1) \wedge Restaurant(r2) \wedge address(r1, a) \wedge address(r2, a) \wedge city(r1,c) \wedge city(r2,c) \rightarrow sameAs(r1, r2)$

Disjoint classes $C1(x) \wedge C2(y) \rightarrow differentFrom(x,y)$

Functional DataType properties
$sameAs(r1,r2) \wedge city(r1,c1) \wedge city(r2,c2) \rightarrow equivalentString(c1,c2)$

Local-UNA
Example: For one publication, in one dataset, all the authors are distinct (the inverse may be untrue)
$authored(p, a1) \wedge authored(p, a2) \rightarrow differentFrom(a1,a2)$

Referring expression
Example: profession+name is not a key … but there is only one president named *Obama*
name(p1,'Obama') ∧ profession(p1,'president') → sameAs(p1, http://…81)

FROM KNOWLEDGE TO RULES (OR FUNCTIONS)

- **Complex Rules with built-in predicates**

Example: Address+city is a composite key

IF min(Jaccard(address(w1),address(w2)),jaro(city(w1),city(w2)) > 0.8 then sameAs(w1, w2)

Example:Two keys for a book : ISBN, title+year

Score(book1,book2) = Max(sim(isbn(book1), isbn(book2)), min(sim(title(book1),title(book2)), sim(year(book1),year(book2))

OWL2 KEY (S-KEY)

OWL2 Key for a class: a combination of property expressions that uniquely identify each instance of a class expression

hasKey(CE (OPE$_1$ ... OPE$_m$) (DPE$_1$ ... DPE$_n$))

$$\forall X, \forall Y, \forall Z_1, \ldots, Z_n, \forall T_1, \ldots, T_m \wedge ce(X) \wedge ce(Y) \bigwedge_{i=1}^{n} (ope_i(X, Z_i) \wedge ope_i(Y, Z_i))$$

$$\bigwedge_{i=1}^{m} (dpe_i(X, T_i) \wedge dpe_i(Y, T_i)) \Rightarrow X = Y$$

hasKey(Book(Author) (Title)) means:

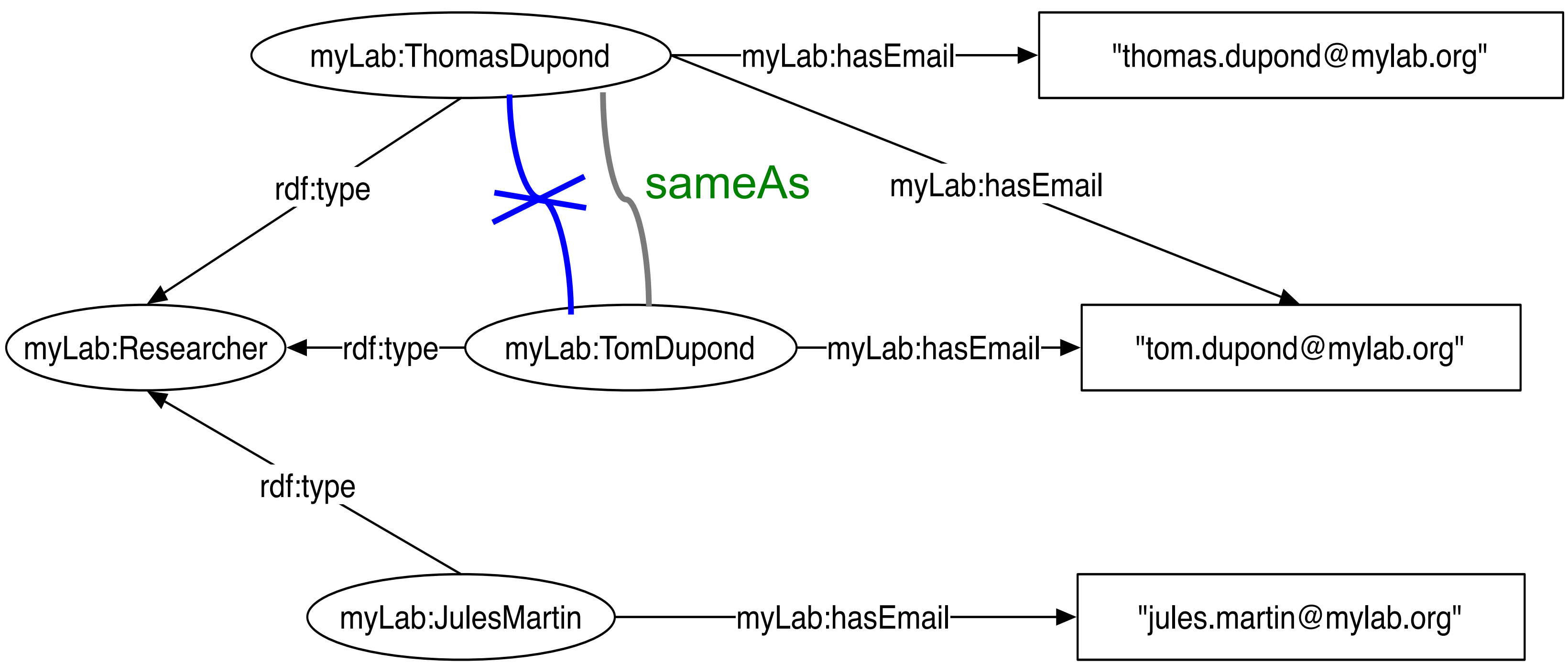
$Book(x_1) \wedge Book(x_2) \wedge Author(x_1, y) \wedge Author(x_2, y) \wedge Title(x_1, w) \wedge Title(x_2, w)$

➔ $sameAs(x_1, x_2)$

Inheritance : a key declared for persons is valid for researchers.

ALTERNATIVE KEY SEMANTICS: F-KEY, SF KEYS

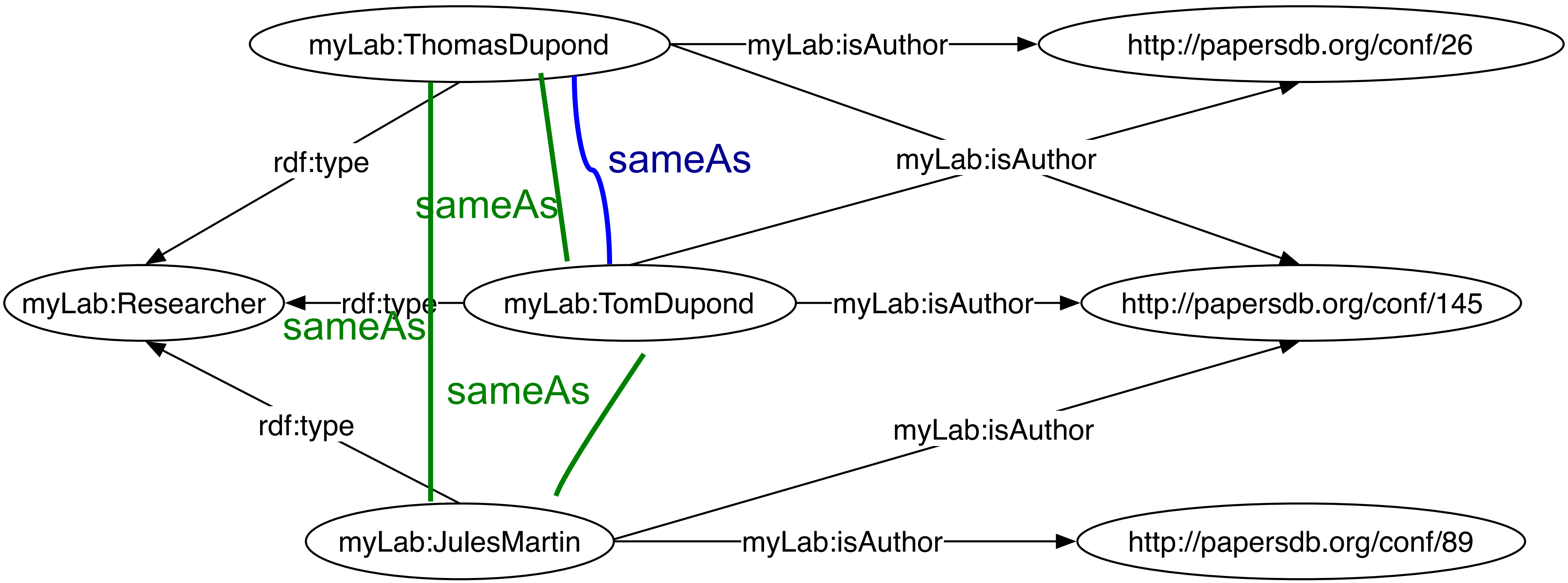
S-Key (Researcher, (e-mail)) ([pernelle12, Symeonidou14], Owl2 keys)
one shared e-mail is sufficient to decide
SF-Key (Researcher, (e-mail)) [Atencia12], or F-Key (Researcher, (e-mail)) [Soru15]
the sets of e-mail values must be identical



ALTERNATIVE KEY SEMANTICS: F-KEY, SF KEYS

SF-Key (Researcher, (isAuthor)), F-key(Researcher, (isAuthor))

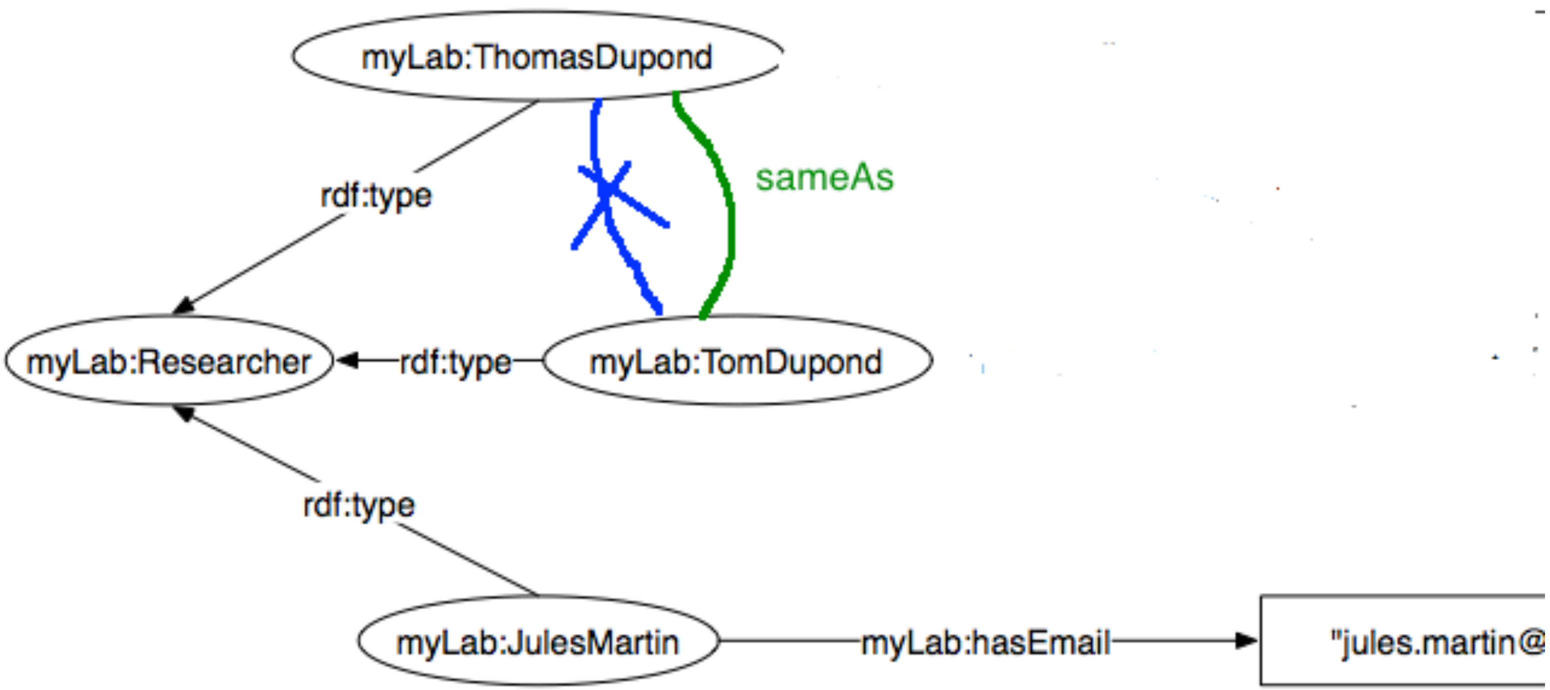
S-Key (Researcher, (isAuthor))



ALTERNATIVE KEY SEMANTICS : F-KEY, SF KEYS

SF-Key(Researcher, (e-mail)) or S-Key(Researcher, (e-mail))

F-key(Researcher, (e-mail)) (empty sets of values are considered)



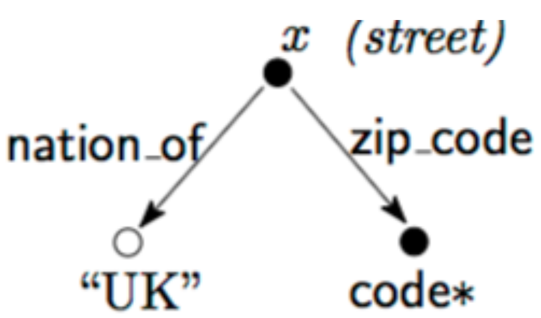
SF-Keys, F-keys are interesting when a local completeness is known.

GRAPH PATTERNS

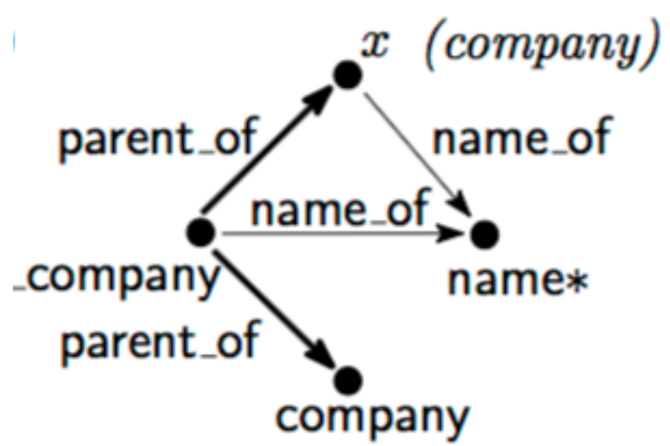
More generally, a key can be expressed as a graph pattern: topological constraints and value bindings that are needed for identifying entities [Fan et al 15]

A street in the UK can be identified by its zip code

(not in France or US)



A company split from a parent company of the same name is identified by the company name and another child company**.**



also called Conditionnal key [Symeonidou et al 17]

DATA LINKING APPROACHES: EVALUATION

- Effectiveness: evaluation of linking results in terms of recall and precision
  - Recall = (#correct-links-sys) /(#correct-links-groundtruth)
  - Precision = (#correct-links-sys) /(#links-sys)
  - F-measure (F1) = (2 x Recall x Precision) / (Recall +Precision)

- Efficiency: in terms of time and space (i.e. minimize the linking search space and the interaction actions with an expert/user).
- Robustness: override errors in the data
- Genericity: applicable to different datasets and different domains
- Use of benchmarks, like those of OAEI (Ontology Alignment Evaluation Initiative) or Lance

LOCAL DATA LINKING APPROACHES

FRAMEWORK SILK [Volz et al'09]

A Local, Informed, Unsupervised Rule-based approach

- Allows specifying linking conditions between two datasets (not limited to sameAs)
- Provides a Link Specification Language(LSL)
- The linking conditions may be expressed in terms of:

  Data transformation functions (e.g. removeBlanks)

  Elementary similarity measures (e.g. Jaro, maxSimilarityInSets, setSimilarity)

  Aggregation functions of the similarity scores (e.g. max, weighted average)

  Mappings between classes and properties

- Can be used for **S-keys** and some **SF-keys** (multivalued datatype properties)

SIMILARITY MEASURES IN SILK

EXTRACT

[Volz et al'09]

Metric	Description
jaroSimilarity	String similarity based on Jaro distance metric
jaroWinklerSimilarity	String similarity based on Jaro-Winkler metric
qGramSimilarity	String similarity based on q-grams
stringEquality	Returns 1 when strings are equal, 0 otherwise
numSimilarity	Percentual numeric similarity
dateSimilarity	Similarity between two date values
uriEquality	Returns 1 if two URIs are equal, 0 otherwise
taxonomicSimilarity	Metric based on the taxonomic distance of two concepts

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<Interlinks>
    <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>
```

Link types

Entities to be linked

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<LinkageRule>
        <Aggregate type="average">
          <Compare metric="levenshteinDistance" threshold="1">
            <Input path="?a/rdfs:label" />
            <Input path="?b/gn:name" />
          </Compare>
          <Compare metric="num" threshold="1000" >
            <Input path="?a/dbpedia:populationTotal" />
            <Input path="?b/gn:population" />
          </Compare>
        </Aggregate>
      </LinkageRule>

      <Filter limit="1" />
```

Aggregation function

Similarity measures

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<Outputs>
        <Output type="file" minConfidence="0.95">
          <Param name="file" value="accepted_links.nt" />
          <Param name="format" value="ntriples" />
        </Output>
        <Output type="file" maxConfidence="0.95">
          <Param name="file" value="verify_links.nt" />
          <Param name="format" value="alignment" />
        </Output>
      </Outputs>
    </Interlink>
  </Interlinks>

</Silk>
```

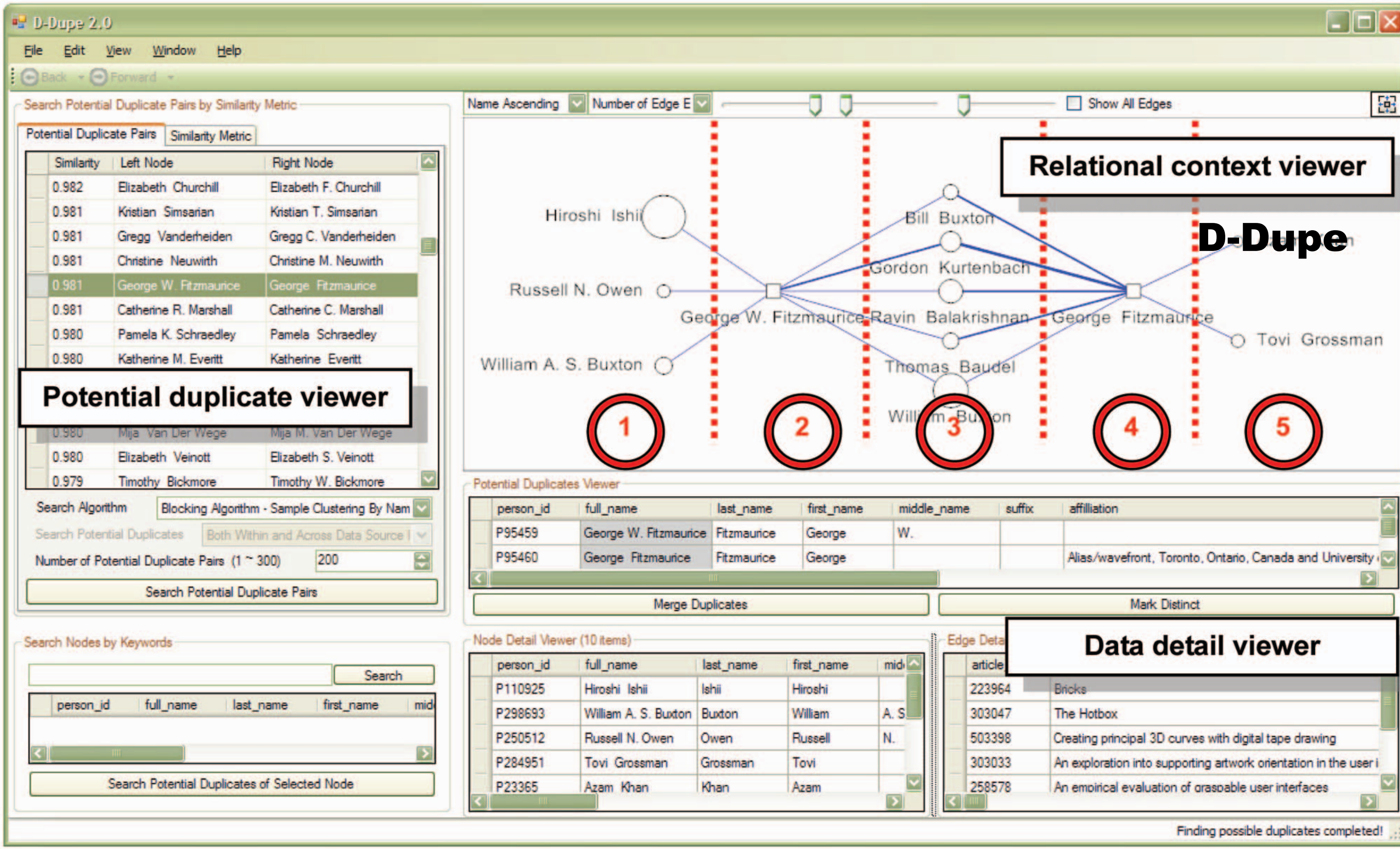
Linking threshold

Possible links

GLOBAL DATA LINKING APPROACHES

GLOBAL AND INTERACTIVE APPROACH

[Kang et al' 08]



OBJECTCOREF [HU ET AL. 2011]

- **A Global, then Local, (informed), semi-supervised approach**
- Learn to detect new links from a set of existing links or links inferred thanks to ontology axioms (semi-supervised)

- D : a RDF graph that represents a set of equivalent instances

  H : a RDF graph that represents new instances

Iterate (1), (2) et (3)

(1) Exploits D to learn property mappings (similarities of values):

geoalternateName / rdfs:label

(2) D and H are used to learn a discriminative (property,value) pair for the instance (e.g. rdfs:label, '*Beijing*' is discriminative for the city of beijing)

(3) Exploits the discriminative (property,value) pair to discover links with new instances and add them to D.

Considered entity		
Dbpedia:Beijing	rdfs:label	'Beijing'
	Owl:sameAs	geo:1816670

geo: 1816670	wgs84-pos:long	'116'
	wgs84-pos:lat	'40'
	geo:alternateName	'Beijing'
	geo:alternateName	'Pékin'

D

semweb:Beijing	rdfs:label	'Beijing »
	wgs84-pos:long	'116'
	wgs84-pos:lat	'40'

H

First discriminative (property,value) pair = referring expression:

(rdfs:label mapped to geo:alternateName, 'Beijing')

Discriminative:

(#instances with this pair in D) / (#instances with this pair in H) > given threshold.

→ New instance discovered in H : *semweb:Beijing … next property = latitude*

OBJECTCOREF - EXPERIMENTS

- Restaurants/Persons (benchmark OAEI'2010)

  D: 20 links of the goldstandard

Approche	F-Mesure
ObjectCoref	0.95
LN2R	0.95

- Discriminative properties for persons: SSN, phoneNumber then age

  Discriminative properties for restaurants: phoneNumber
- Results can be incorrect when there are two many iterations.

  Frequent pairs of properties can improve the precision

  (e.g. more complex referring expressions s.t *latitude* +*longitude*).

LN2R

[Saïs et al'09]

A global, unsupervised, informed approach that combines two methods:

- **L2R, a Logical method: applies rules to infer *sure* owl:sameAs, owl:differentFrom and equivalences or differences of literals.**

  Rules are automatically generated from the ontology axioms, and from the declared assumptions on the dataset.

  Forward chaining (unit resolution).

- **N2R, a Numerical method: computes similarity scores for each pair of instances**

  An equation system models dependencies between similarity scores. Automatically constructed from the dataset, the ontology axioms and the assumptions on the dataset. Iteratively solved (non linear, fix point, convergence). Results of L2R can be considered.

- **Assumptions**
  - The datasets are conforming to the same ontology
  - The ontology contains axioms

LN2R

[Saïs et al'09]

Considered Knowledge

- **Ontology axioms**

  Disjunction between classes, (L2R)
  (Inverse)Functional properties, (L2R, N2R)
  Composite keys, (L2R, N2R)

- **Expert knowledge**

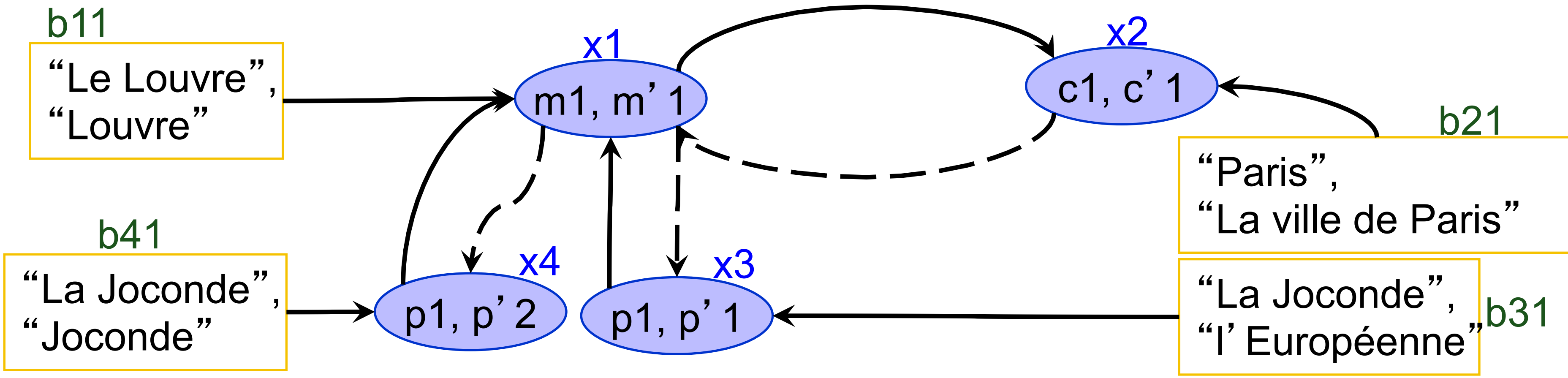
  Similarity functions declared for each property, (N2R)

- **Assumptions on the data**

  Unique Name Assumption (UNA) (L2R)
  Local-UNA (L2R)

N2R: ILLUSTRATION

[Saïs et al'09]



x1 = max(max(b11, x3), x4), λ * x2)

x2 = max(b21, x1)

x3 = max(b31, λ* x1)

x4 = max(b41 , λ * x1)

	x1	x2	x3	x4
Initialization	0.0	0.0	0.0	0.0
Iteration 1	0.8	0.3	0.1	0.7
Iteration 2	0.8	0.8	0.4	0.7
Iteration 3	0.8	0.8	0.4	0.7

λ= 1/(| CAttr | + | CRel |) ε = 0.02

b11 = 0.8, b21 = 0.3, b31 = 0.1, b41 = 0.7

Solution: x1 = 0.8
x2 = 0.8
x3 = 0.4
x4 = 0.7

LN2R - EXPERIMENTS

- **L2R**

Precision of *100% (by construction).*

A recall that varies depending on the heterogeneity of the vocabulary (e.g. 52 % for CORA dataset, 54% for Orange hotel descriptions)

Many differentFrom can be generated thanks to UNA, local-UNA, and non equivalent literals involved in functional properties (recall >90% on Cora).

Sensible to errors.

- ***N2R***

95% of F-mesure in OAEI restaurant/person benchmark

Not efficient.

IMPORT BY QUERY

[Al Bakri et al 15]

A **global**, **informed**, **rule-based** approach based on a backward-chaining algorithm that combines **:**

- Local reasoning (forward reasoning)
- External querying to bypass local data incompleteness (backward chaining)

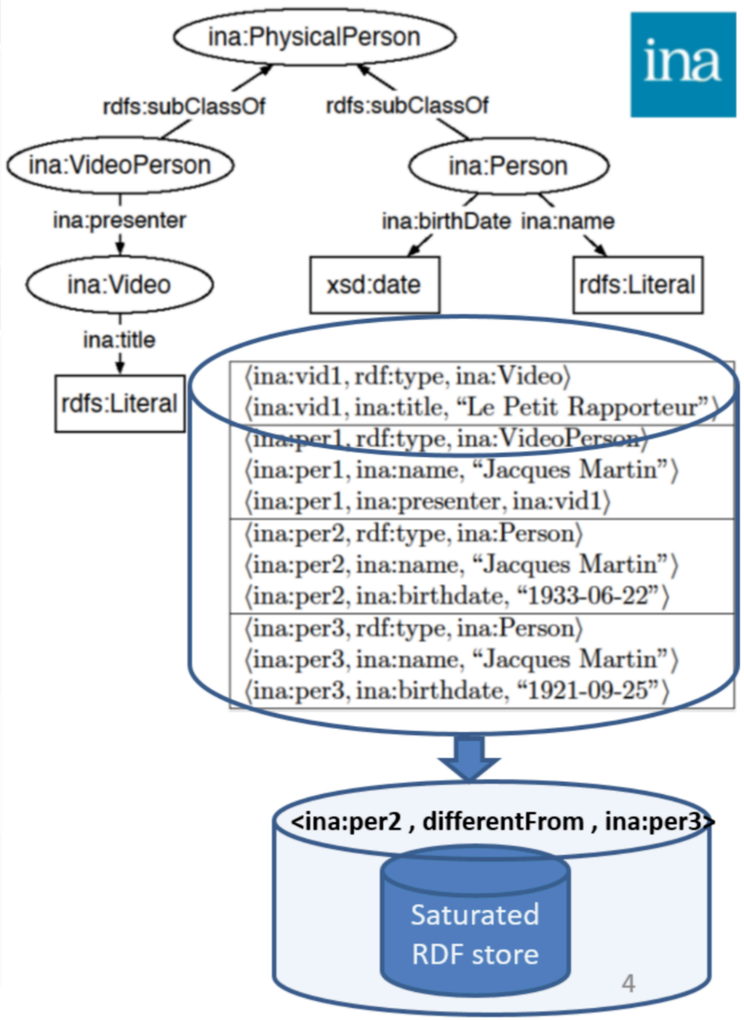
To infer a target owl:sameAs or contradict it.

Knowledge : (inverse) functional properties, composite keys, semantics of owl:sameAs (transitivity) and owl:differentFrom.

IMPORT BY QUERY

[Al Bakri et al 15]

	IF	THEN
R1	?p1 name ?name ?p1 birthdate ?d ?p2 name ?name ?p2 birthdate ?d	?p1 same_as ?p2
	IF	**THEN**
R2	?p1 name ?name ?p1 ina:presenter ?v1, ?v1 title ?t ?p2 name ?name ?p2 db:presenter ?t	?p1 same_as ?p2
R3	?p1 birthdate ?d1 ?p2 birthdate ?d2 ?d1 <> ?d2	?p1 differentFrom?p2
R4	?x1 same_as ?x2 ?x2 same_as ?x3	?x1 same_as ?x3
R5	?x1 same_as ?x2 ?x2 differentFrom ?x3	?x1 differentFrom ?x3



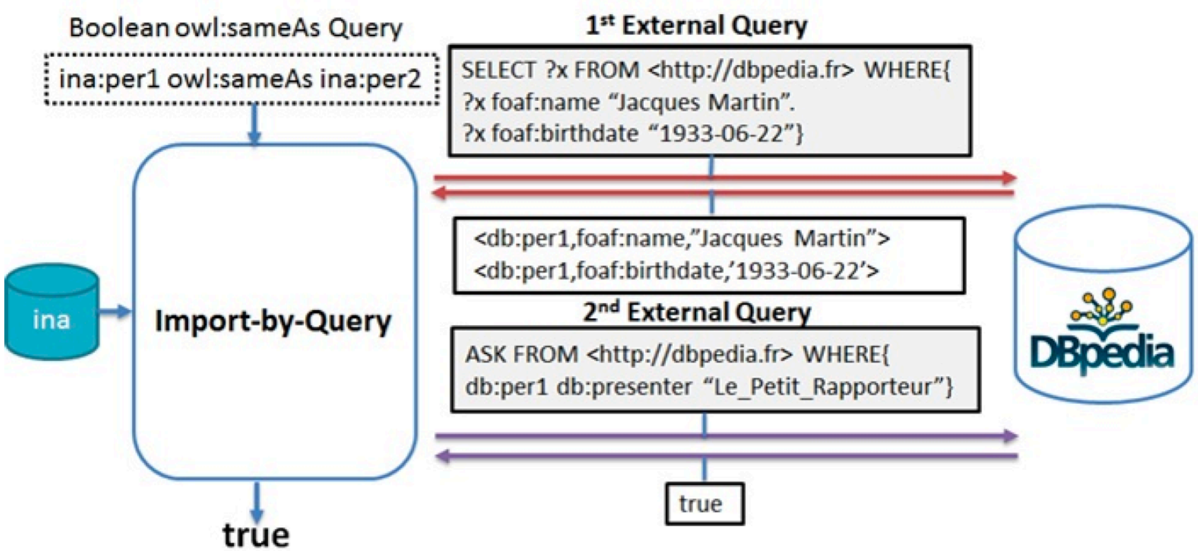
BUT <ina:per1, same_as, ina:per2> ? STILL UNKNOWN

M.C. Rousset, ICFCA'17

IMPORT BY QUERY

[Al Bakri et al 15]

Build on demand queries to some entry points of Linked Data
Alternates subquery rewriting steps based on backward chaining and external query evaluation (adaptation of Query-Subquery algorithm).



IMPORT BY QUERY - EXPERIMENTS

[Al Bakri et al 15]

1.5 million RDF facts, provided by a french national audiovisual institute (INA)
35 rules (built with the help of INA experts), 0.5 million external facts (DBPedia).

	IF	THEN
r7	$\langle ?x1$,`foaf:name`,$?name1\rangle$, $\langle ?x2$, `skos:altLabel`, $?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$
r8	$\langle ?x1$,`foaf:name`,$?name1\rangle$, $\langle ?x2$, `skos:prefLabel`, $?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$
r9	$\langle ?x1$,`rdfs:label`,$?name1\rangle$, $\langle ?x2$, `skos:prefLabel`, $?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$
r10	$\langle ?x1$,`rdfs:label`,$?name1\rangle$, $\langle ?x2$, `skos:altLabel`, $?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$
r11	$\langle ?x1$,`prop-fr:nom`,$?name1\rangle$, $\langle ?x2$, `skos:prefLabel`, $?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$
r12	$\langle ?x1$,`prop-fr:nom`,$?name1\rangle$, $\langle ?x2$, `skos:altLabel`, $?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$

	IF	THEN
r13	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$, $\langle ?x1$, `dbpedia:birthYear`, $?Y1\rangle$, $\langle ?x2$, `ina:birthYear`, $?Y1\rangle$ $\langle ?x1$, `dbpedia:deathYear`, $?Y2\rangle$, $\langle ?x2$, `ina:deathYear`, $?Y2\rangle$	$\langle ?x1$, `ina:sameAs`, $?x2\rangle$
r14	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$, $\langle ?x1$, `dbpedia:birthYear`, $?Y1\rangle$, $\langle ?x2$, `ina:birthYear`, $?Y2\rangle$ notEqual($Y1, Y2$)	$\langle ?x1$, `ina:differentFrom`, $?x2\rangle$
r15	$\langle ?x1$, `ina:sameNameDBp`, $?x2\rangle$, $\langle ?x1$, `dbpedia:deathYear`, $?Y1\rangle$, $\langle ?x2$, `ina:deathYear`, $?Y2\rangle$ notEqual($Y1, Y2$)	$\langle ?x1$, `ina:differentFrom`, $?x2\rangle$

IMPORT BY QUERY - EXPERIMENTS

[Al Bakri et al 15]

- External information can be useful to link Data

  2 links (108 differentFrom) with INA

  versus 4,884 links (resp.9,700) with DBPEDIA
- 100 % precision if the facts and rules are correct

  500 have been manually checked
- Reasoning allows to discover more links

  Silk only discovered 2% of the sameAs links discovered by the forward reasoner.
- Low number of imported facts

  Only 6,000 facts are needed (among 500,000 facts of the DBPedia extract)
- Efficient : 191s forward chaining, 7s per query (in average)

PROBFR

[Al Bakri et al 15]

- **A global, informed approach that model uncertainty as probabilities**

    Uncertain rules, Uncertain facts, Uncertain mappings

- Based on Probabilistic Datalog

    Facts and rules are associated with a symbolic event e

    An event expression is computed for each inferred fact during the saturation process (provenance)

    ex. $Prov_{R,F}((i1 \text{ sameAs } i2)) = (e(r1) \wedge e(f1)) \vee (e(r2) \wedge e(f3))$

    where f_i is a fact, r_i is a rule.

Probabilities are then computed thanks to the event expressions (and can be reevaluated easily, if some probabilities are updated).

PROBFR - EXPERIMENTS

[

[Al Bakri et al 15]

- MusicBrainz (122 million triples), DBpedia (73 million triples)

  20 certain rules, 36 uncertain rules (probabilities from 0.3 to 0.9)

- Runtime: < 2 hours

- When uncertain information is used, the recall increases very significantly (checked on samples)

Certain rules

	Precision	Recall
Person	100%	8%
Band	100%	12%

All the rules (probability > 0.9)

	Precision	Recall
$\text{Band}_{\geqslant 0.9}$	100%	80%
$\text{Song}_{\geqslant 0.9}$	100%	44%

WHAT IF THE ONLY APPLIED RULE IS TRANSITIVITY OF SAMEAS ?

[Beek et al.18]

LOD-a-lot dataset

(lod-a-lot.lod.labs.vu.nl)

[Fernández et al., 2017]

~**4.5 hours**

```
prefix owl: <http://www.w3.org/2002/07/owl#>
select distinct ?x ?p ?y {                          (distinct)
  bind (owl:sameAs ?identity)                       (triples)
  ?x ?identity ?y
}
```

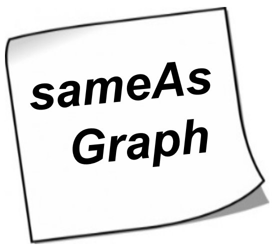
sameAs Graph

558.9 million triples
179.7 million entities

WHAT IF THE ONLY APPLIED RULE IS TRANSITIVITY OF SAMEAS ?

[Beek et al.18]

After transitive closure …



558.9 million triples

~5 hours

Identity Set 1

Identity Set 2

. . .

Identity Set 48 999 148

35.2 billion triples to materialize all the implicit owl:sameAs

49 million identity sets (non-singleton)

BUT …

https://sameas.cc/term?page=1&page_size=20&id=4073

SameAs.cc Documentation Identity sets Terms Triples

Terms for identity set 4073

- <http://af.dbpedia.org/resource/%D0%A7> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/%D1%A4> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/7> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Aandelebeurs> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Afghanistan> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Afrika> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Albanees> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Albani%C3%AB> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Albanië> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Albany,_New_York> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Albert_Einstein> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Algeri%C3%AB> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Algerië> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Amerikaans-Samoa> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Amerikaanse_Maagde-eilande> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Amerikas> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Andorra> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Andorra_la_Vella> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Angola> (↦ id) ⟨s, owl:sameAs, o⟩
- <http://af.dbpedia.org/resource/Anguilla_(eiland)> (↦ id) ⟨s, owl:sameAs, o⟩

Previous results 1 to 20 (of 177,794) Next

The largest identity set contains 177 794 terms:

Different countries
Different cities
Albert Enstein

→ quality problems

SUMMARY

Informed approaches can take into account many kinds of knowledge:

ontology axioms, expert knowledge, assumption on datasets, referring expressions …

Such approaches can easily be extended by new rules.

+ Local approaches: pairs compared independently are efficient, but do not allow to propagate decisions (recall can be lower).

+ Global approaches: decision can be propagated logically or numerically.

+ Logical approaches infer *sure* identity links, can be used to infer differentFrom.

+ Can deal with large datasets:

> forward chaining can be parallelized [Hogan et al. 12],
>
> backward chaining can be used efficiently (minimization of the number of imported facts from external sources).

SUMMARY

- Logical approaches are partial: they cannot decide for all pairs.

- Strong assumptions: data are clean, rules are certain (but even transitivity can lead to many wrong decisions !)

+ In **global and numerical approaches**, similarity scores can be propagated (equation system, probabilistic datalog).

+ Uncertainty can be modelled (similarity of literals, rules with exceptions, uncertain facts).

+- Similarity scores can be assigned to more instance pairs, but the decision is not guaranteed.

- The obtained scores are not so significant, thresholds are difficult to fix.

+ Probabilistic approaches can capture the provenance of an assigned score.

+- Linkage rules are not always available but can be discovered from the data (e.g., key discovery approaches)

REFERENCES (1)

[Al Bakri et al. 16] Uncertainty-Sensitive Reasoning for Inferring sameAs Facts in Linked Data. Mustafa Al-Bakri, Manuel Atencia, Jérôme David, Steffen Lalande, Marie-Christine Rousset, In ECAI 2016

[Al Bakri et al. 15] Inferring Same-As Facts from Linked Data: An Iterative Import-by-Query Approach. Mustafa Al-Bakri, Manuel Atencia, Steffen Lalande, Marie-Christine Rousset:. In AAAI 2015.

[Atencia et al.'12] Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking. Manuel Atencia, Jérôme David, François Scharffe. In EKAW 2012

[Cohen et al. 2003] A comparison of string distance metrics for name-matching tasks. William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. In IIWEB@AAAI 2003.

[Fan et al 15] Keys for Graphs Wenfei Fan, Zhe Fan, Chao Tian, Xin Luna Dong. In PVLDB 2015.

[Ferrara13] Evaluation of instance matching tools: The experience of OAEI. Alfio Ferrara, Andriy Nikolov, Jan Noessner, François Scharffe. OM@ISWC 2013

[Hu et al. 2011] A Self-Training Approach for Resolving Object Coreference on the Semantic Web. Wei Hu, Jianfeng Chen, Yuzhong Qu. In WWW 2011

REFERENCES (2)

[Kang et al. 2008] Interactive Entity Resolution in Relational Data: A Visual Analytic Tool and Its Evaluation. Kang, Getoor, Shneiderman, Bilgic, Licamele, In IEEE Trans. Vis. Comput. Graph 2008.

[Pernelle et al.'13] An Automatic Key Discovery Approach for Data Linking. Nathalie Pernelle, Fatiha Saïs. and Danai Symeounidou. In Journal of Web Semantics 2013.

[Saïs et al.07] L2R: a Logical method for Reference Reconciliation. Fatiha Saïs, Nathalie Pernelle and Marie-Christine Rousset. In AAAI 2007.

[Saïs et al.09] Combining a Logical and a Numerical Method for Data Reconciliation. Fatiha Saïs., Nathalie Pernelle and Marie-Christine Rousset. In Journal of Data Semantics 2009.

[Soru et al. 2015] ROCKER: a refinement operator for key discovery. Soru, Tommaso, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. In WWW, 2015.

[Symeonidou et al. 2014] SAKey: Scalable almost key discovery in RDF data. Symeonidou, Danai, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. In ISWC 2014.

REFERENCES (3)

[Symeonidou et al. 2017] VICKEY: Mining Conditional Keys on RDF datasets .
Danai Symeonidou, Luis Galarraga, Nathalie Pernelle, Fatiha Saïs and Fabian Suchanek. In ISWC 2017.

[Volz et al'09] Silk – A Link Discovery Framework for the Web of Data.
Julius Volz, Christian Bizer et al. In WWW 2009.

[Beek, et al. 2018] The Closure of 500M owl:sameAs Statements', sameAs.cc',
J. Raad, J. Wielemaker & F. van Harmelen. In ESWC 2018 (to appear)

SIMILARITY MEASURES

- **Token based (e.g. Jaccard, TF/IDF cosinus) :**

  The similarity depends on the set of tokens that appear in both S and T.

  ➔ Efficient, but sensitive to spelling errors

- **Edit based (e.g. Levenstein, Jaro, Jaro-Winkler) :**

  The similarity depends on the smallest sequence of edit operations which transform S into T.

  ➔ Less efficient, may deal with spelling errors, but sensitive to word order

- **Hybrids (e.g. N-Grams, Jaro-Winkler/TF-IDF, Soundex)**

For more details: William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. **A comparison of string distance metrics for name-matching tasks**. In *Proceedings of the 2003 International Conference on Information Integration on the Web* (IIWEB'03), Subbarao Kambhampati and Craig A. Knoblock (Eds.). AAAI Press 73-78.