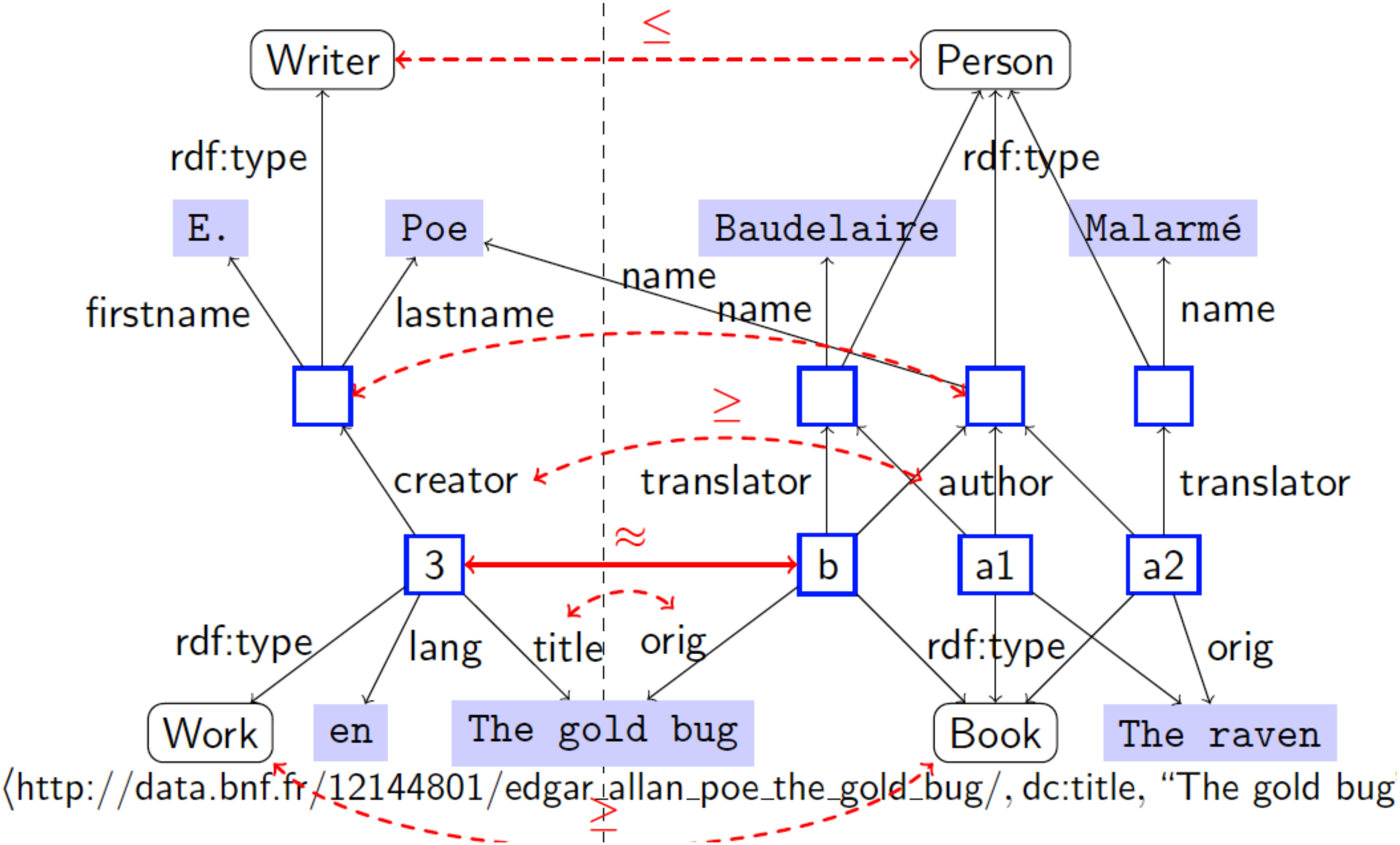
COURSE 3- DATA LINKING

FATIHA SAÏS

UNIVERSITÉ PARIS SACLAY
MASTER 2 OF COMPUTER SCIENCE – DATA SCIENCE

THE PROBLEM: RDF DATA LINKING

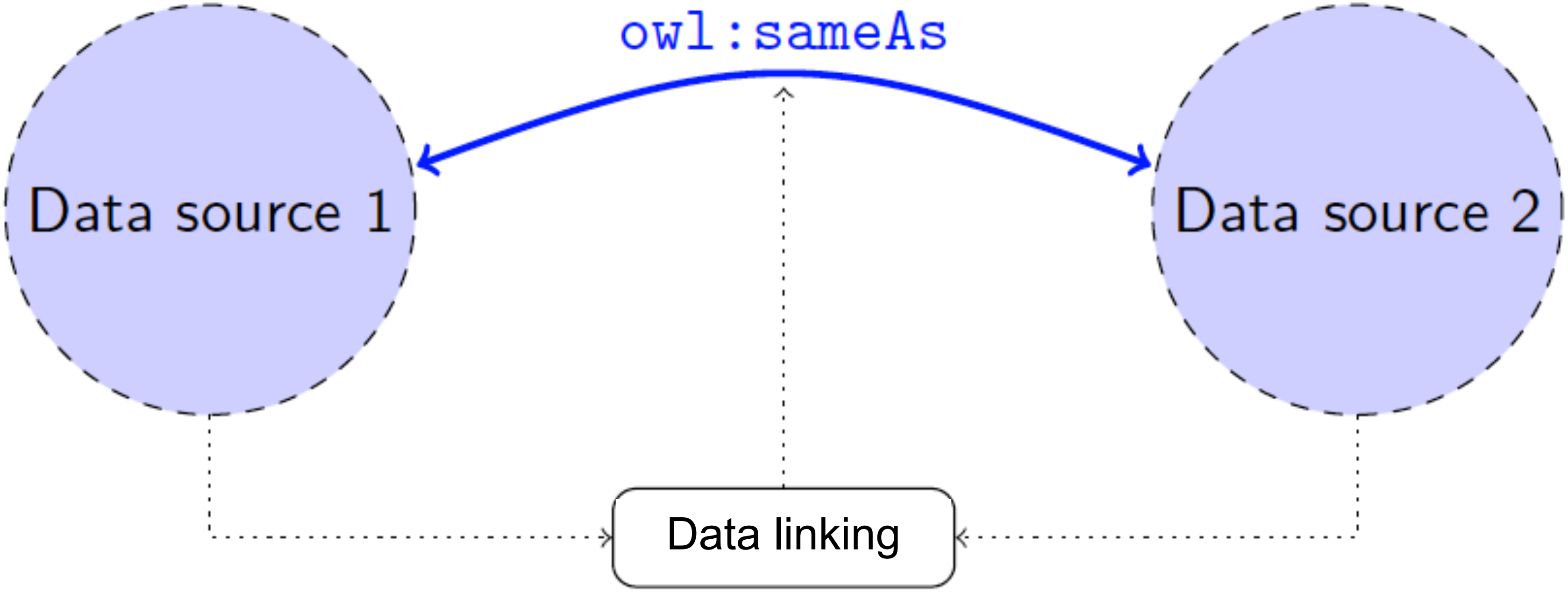


⟨http://data.bnf.fr/12144801/edgar_allan_poe_the_gold_bug/, dc:title, “The gold bug

RDF DATA LINKING PROBLEM

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

- usually in different datasets but can be in one dataset in case of redundant entities.



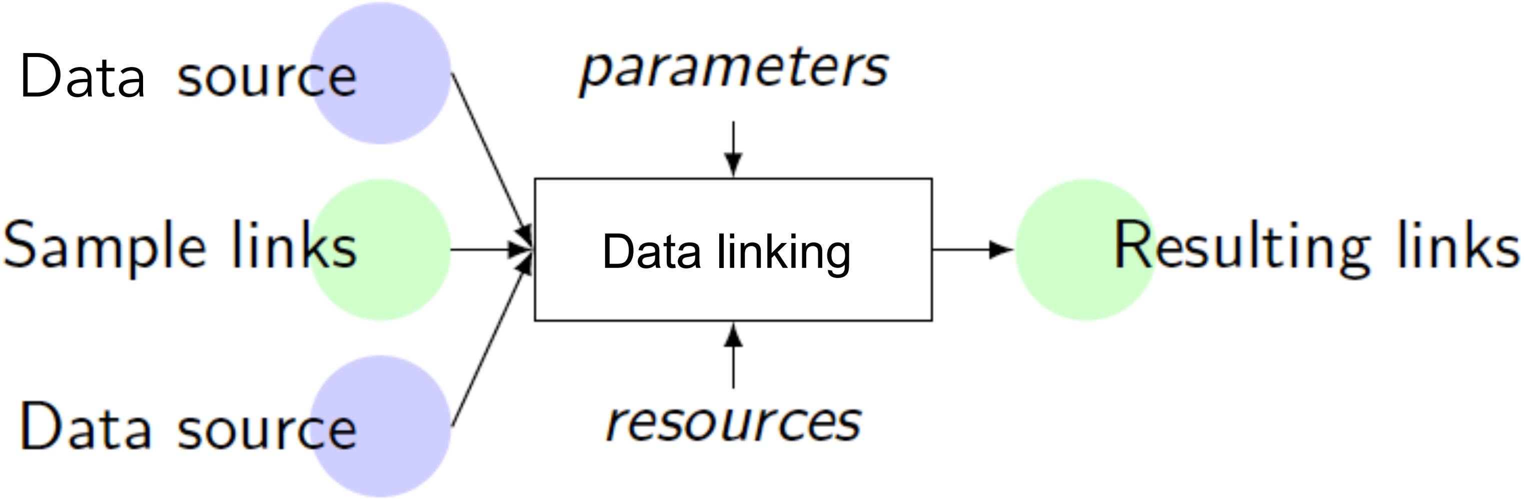
RDF DATA LINKING PROBLEM DEFINITION

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

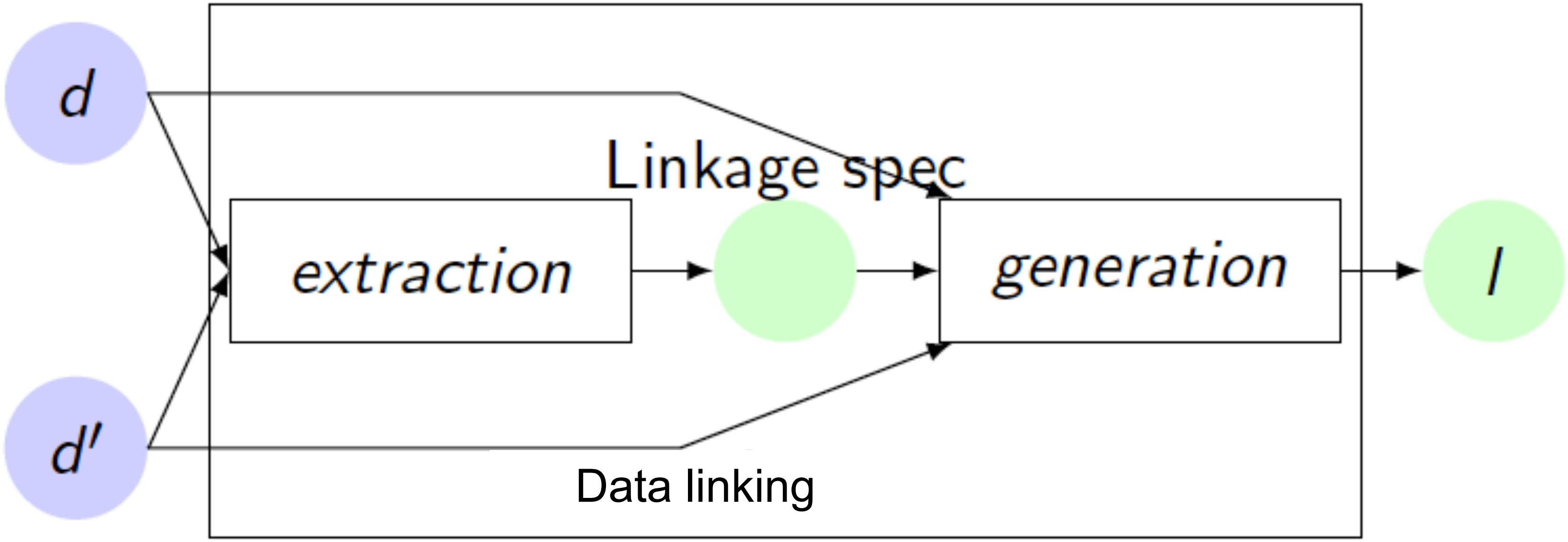
- **Definition (Link Discovery)**
  - Given two sets U_1 and U_2 of resources
  - Find a partition of U_1 x U_2 such that :
    - S = {(u1,u2) ∊ u1 × u2: owl:sameAs(s,t)} and
    - D = {(u1,u2) ∊ u1 × u2: owl:differentFrom(s,t)}

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

RDF DATA LINKING PROCESS

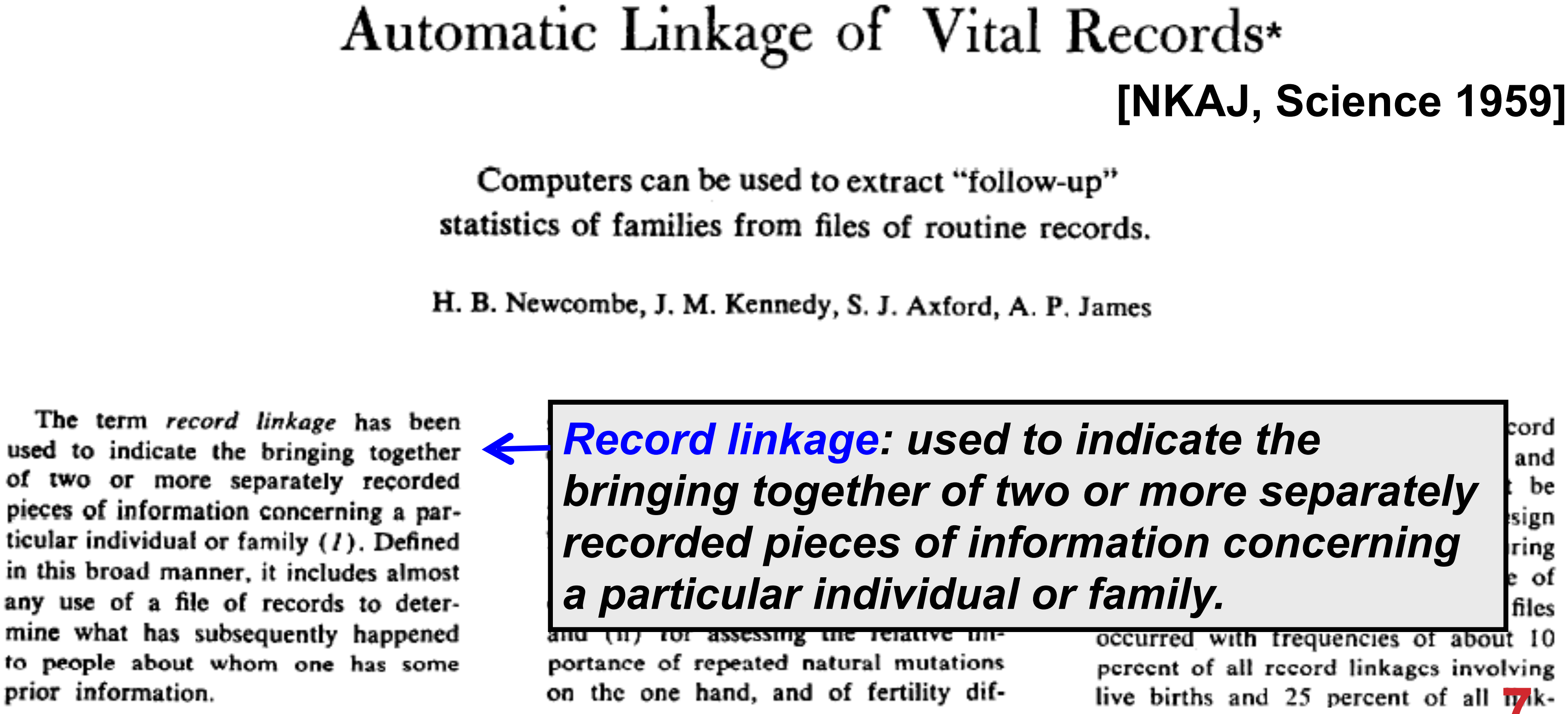


RDF DATA LINKING PROCESS



SOME OF HISTORY …

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



Automatic Linkage of Vital Records*

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.

and (ii) for assessing the relative importance of repeated natural mutations on the one hand, and of fertility dif-

cord and t be sign ring e of files occurred with frequencies of about 10 percent of all record linkages involving live births and 25 percent of all mk-

[NKAJ, Science 1959]

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

ASIDE: DETECTING IDENTITY LINKS

Ironically “Identity link detection” has many duplicates

Record linkage

Duplicate detection

Coreference resolution

Reference reconciliation

Entity resolution

Fuzzy match

Object consolidation

Object identification

Deduplication

Data linking

Entity clustering

Approximate match

Data interlinking

Identity uncertainty

Merge/purge

Household matching

Identity link detection

Hardening soft databases

Householding

Reference matching

Instance matching

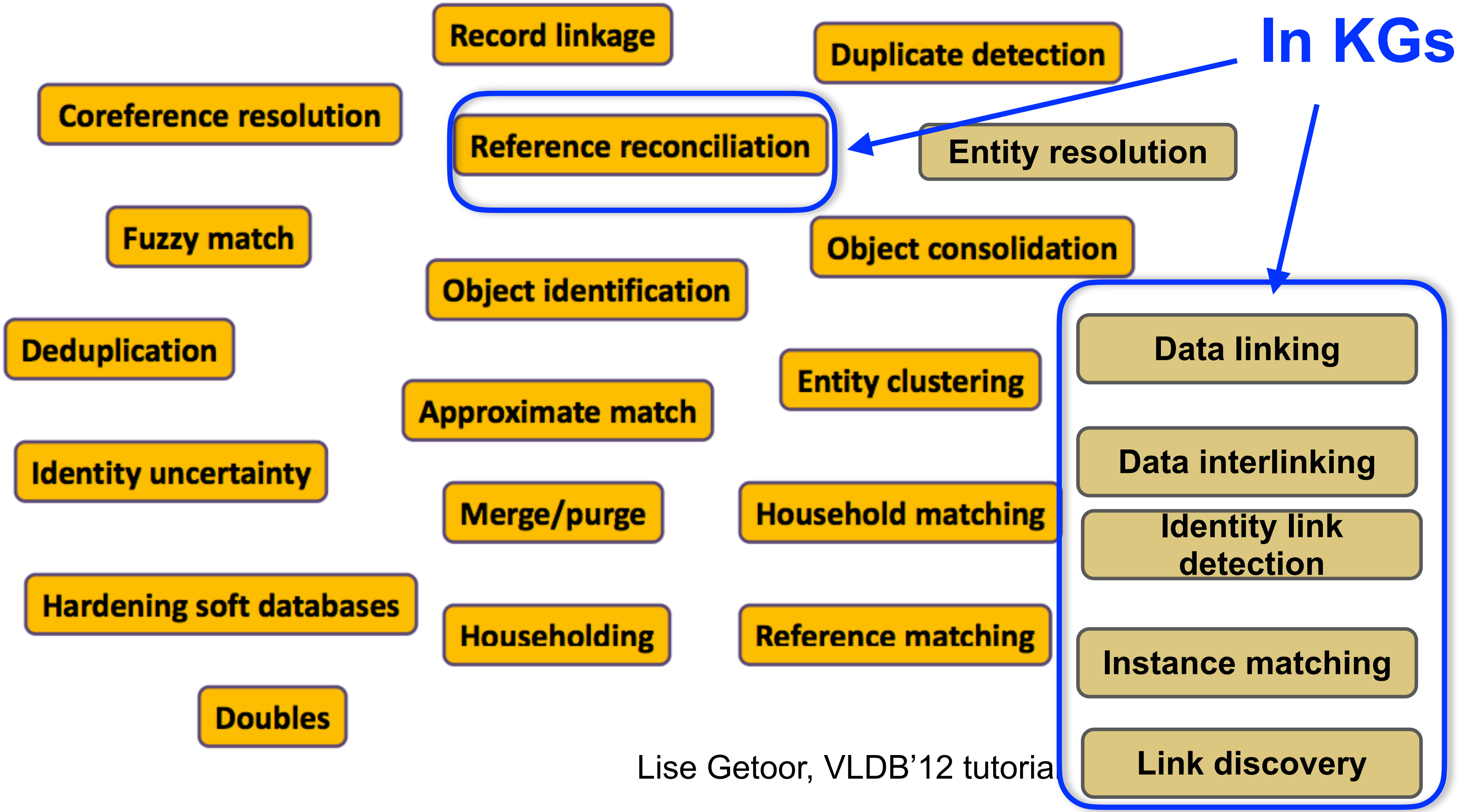
Doubles

Lise Getoor, VLDB’12 tutorial

Link discovery

ASIDE: DETECTING IDENTITY LINKS

Ironically "Identity link detection" has many duplicates



Record linkage

Duplicate detection

In KGs

Coreference resolution

Reference reconciliation

Entity resolution

Fuzzy match

Object consolidation

Object identification

Deduplication

Entity clustering

Approximate match

Identity uncertainty

Merge/purge

Household matching

Hardening soft databases

Householding

Reference matching

Doubles

Data linking

Data interlinking

Identity link detection

Instance matching

Link discovery

Lise Getoor, VLDB'12 tutorial

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

	Databases	Semantic Web
Schema/Ontologies	**Same schema**	**Possibly different ontologies in the same dataset**
Multiple types	**Single relation**	**Several classes**
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

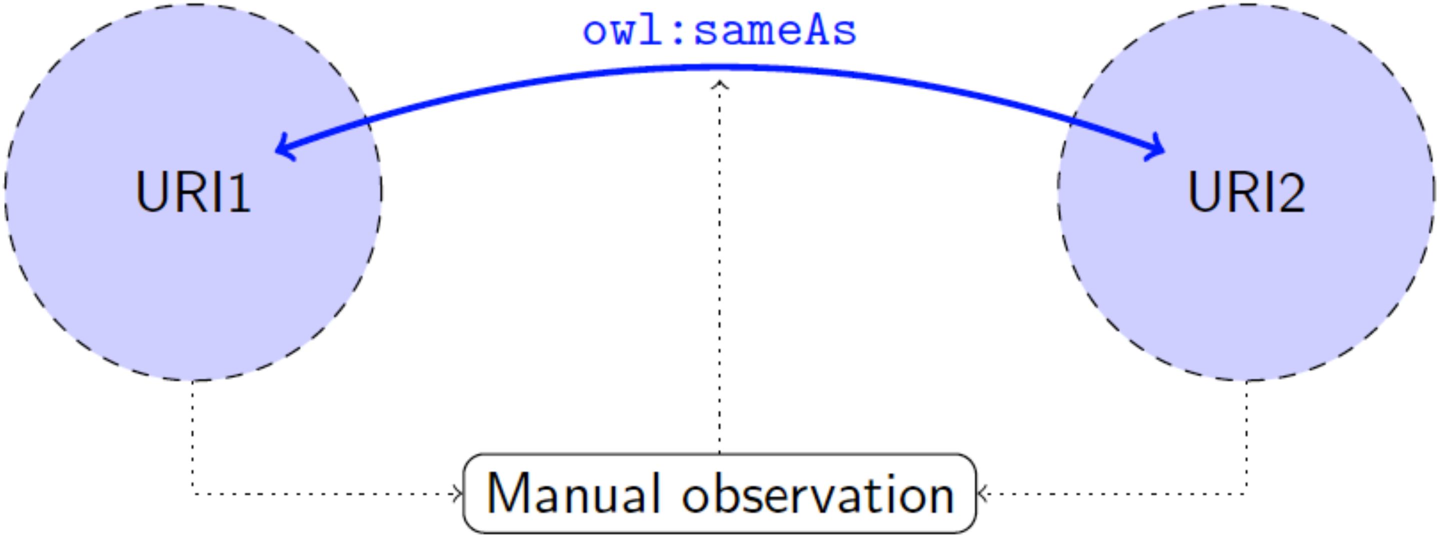
- Datasets conforming to the same ontology
- Datasets conforming to different ontologies
- Datasets without ontologies

DATA LINKING: WHAT INFORMATION TO USE?

Data linking techniques may be based on:

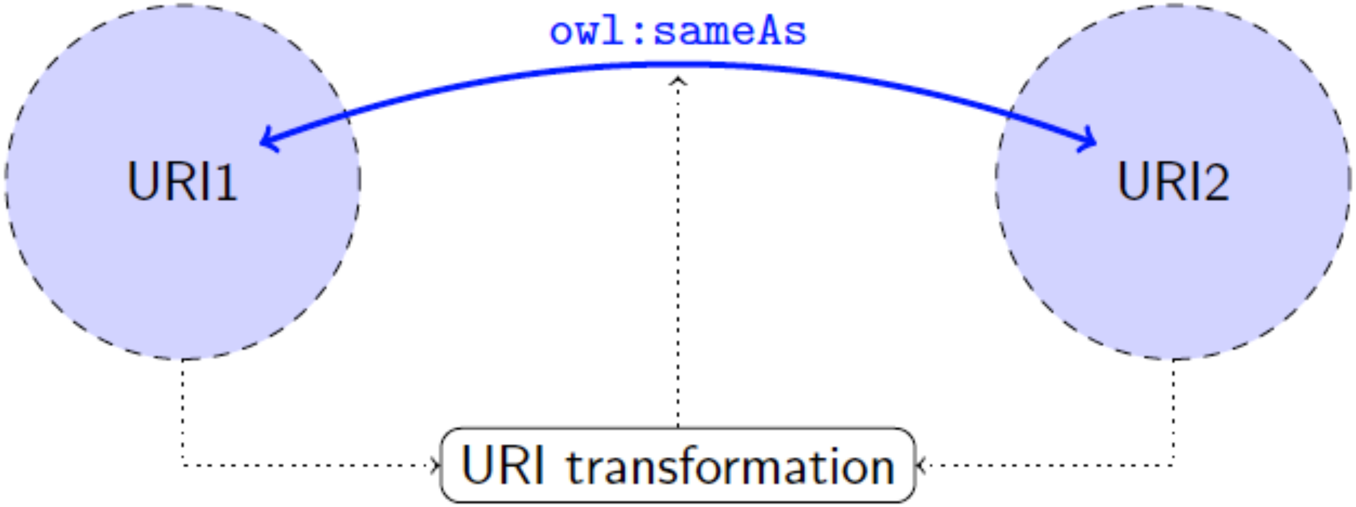
- Data ID (URIs)
- Linking rules and Keys
- External relations: (explicit or implicit) links to other resources
- Data description (content)

MANUAL DESCRIPTION MATCHING



- This does not scale.
- But may be good for a first sample or reference.
- Crowdsourcing?

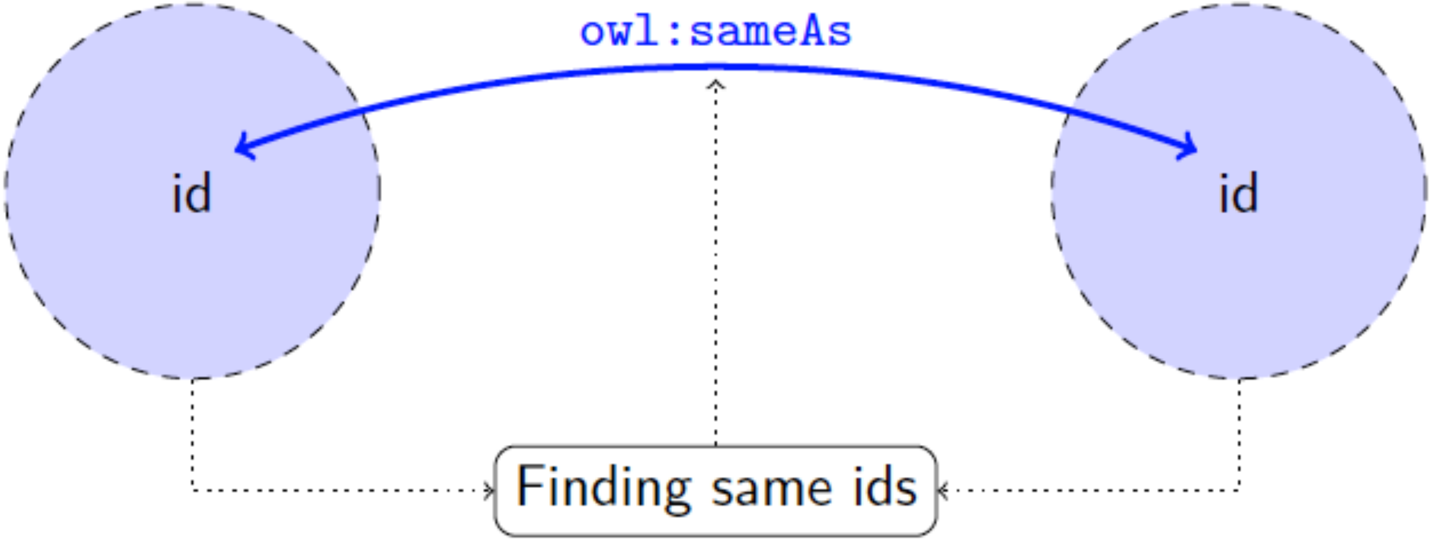
URI MATCHING



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

ID MATCHING

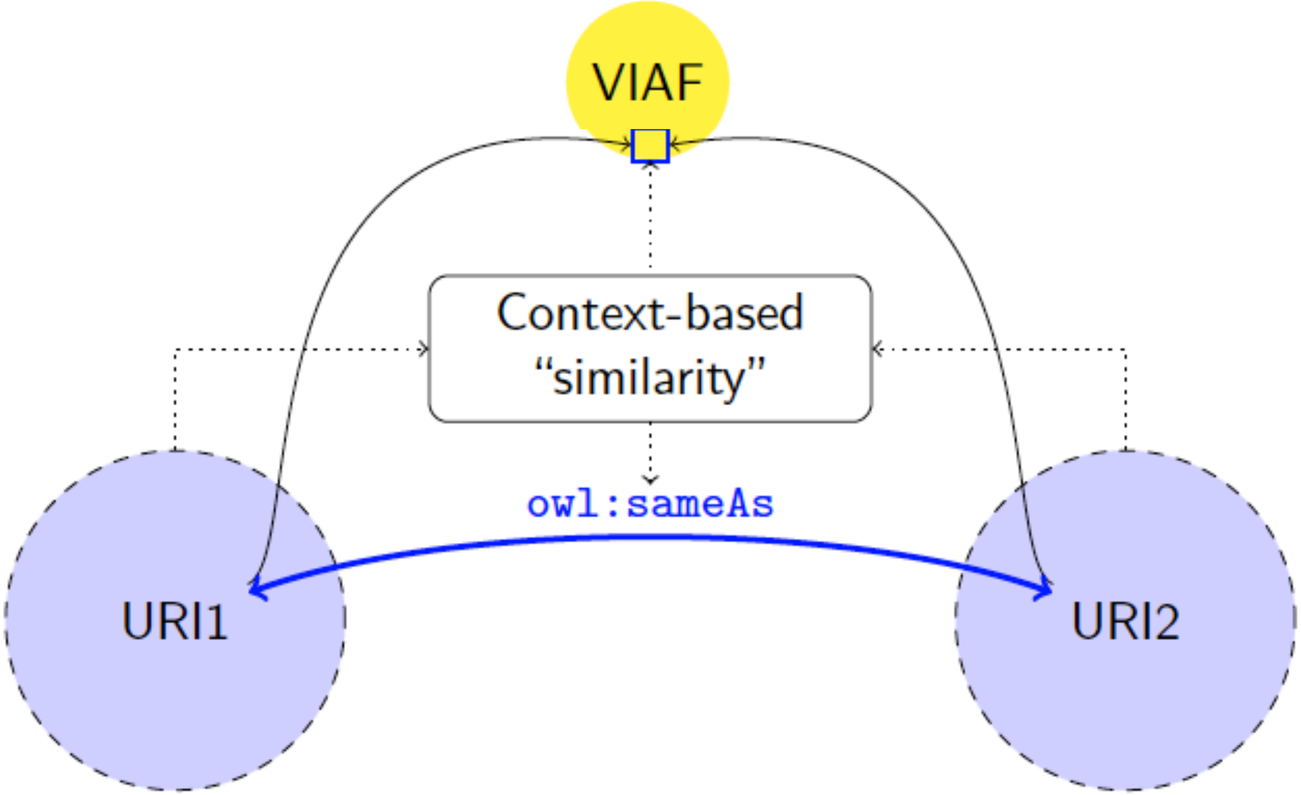


You can find such types of ids:

- Social security numbers
- ISBN, SSN, DOI, MAC addresses, etc.
- authorities: ISO (countries, languages), IATA (airports)

Most databases are built on such **identifiers**. . . but they are often **local** to the database.

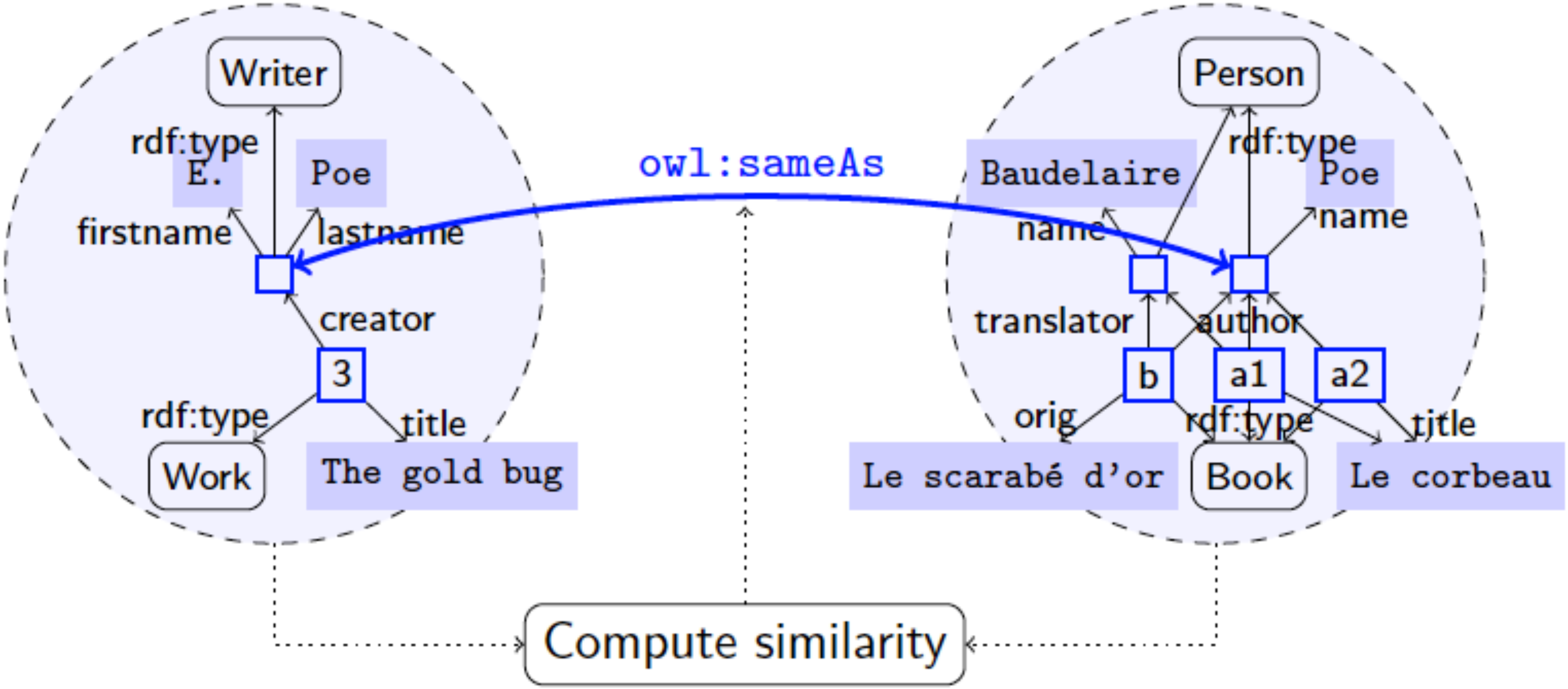
CONTEXT-BASED MATCHING



Process:

- Project your data into another resource (DBPedia, geonames, viaf, etc.)
- Assess relations between considered terms
- Import the relation in the dataset

CONTENT-BASED MATCHING

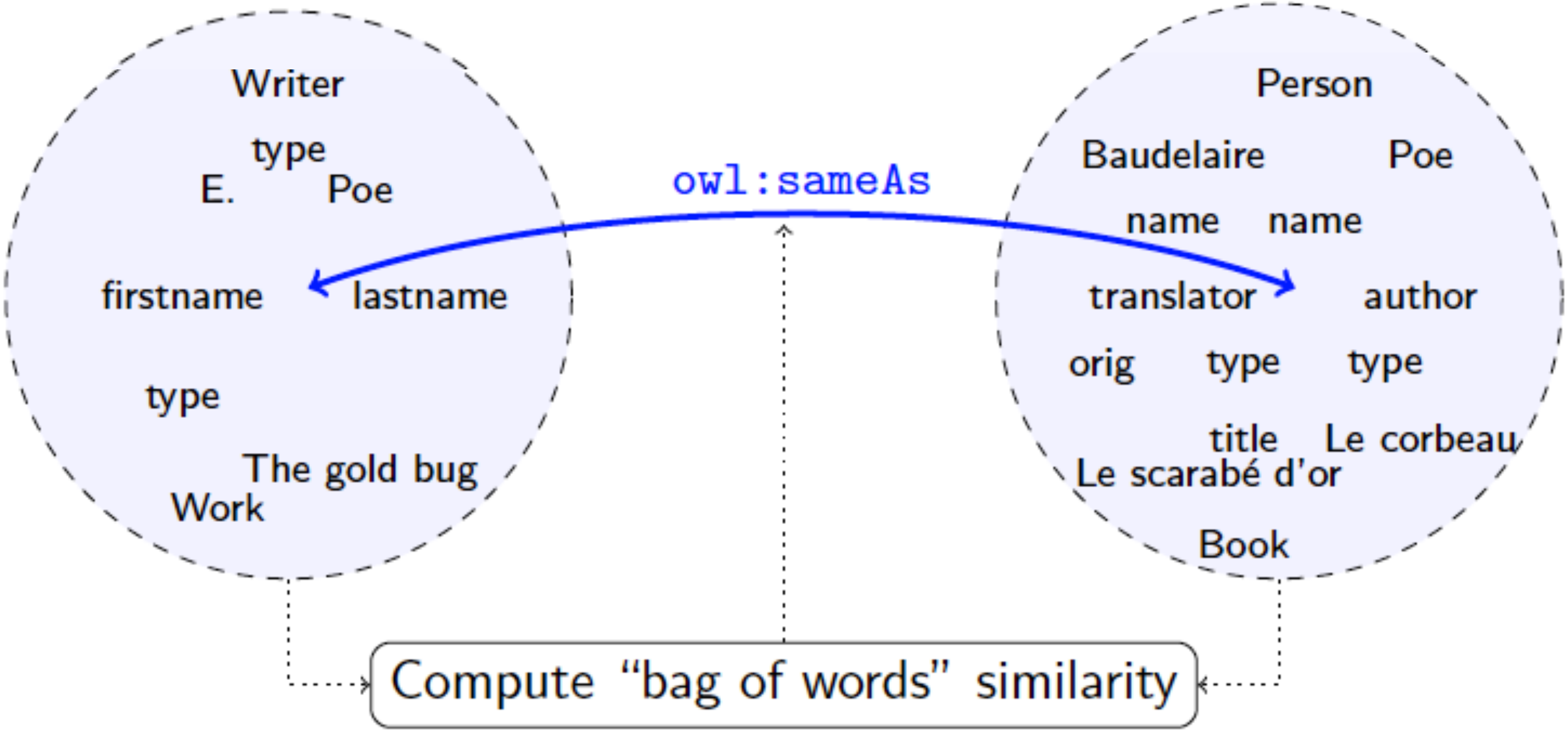


Two main approaches:

- bag of text
- structured similarity

Hypothesis: ↑ similarity ↑ probability that it is the same entity

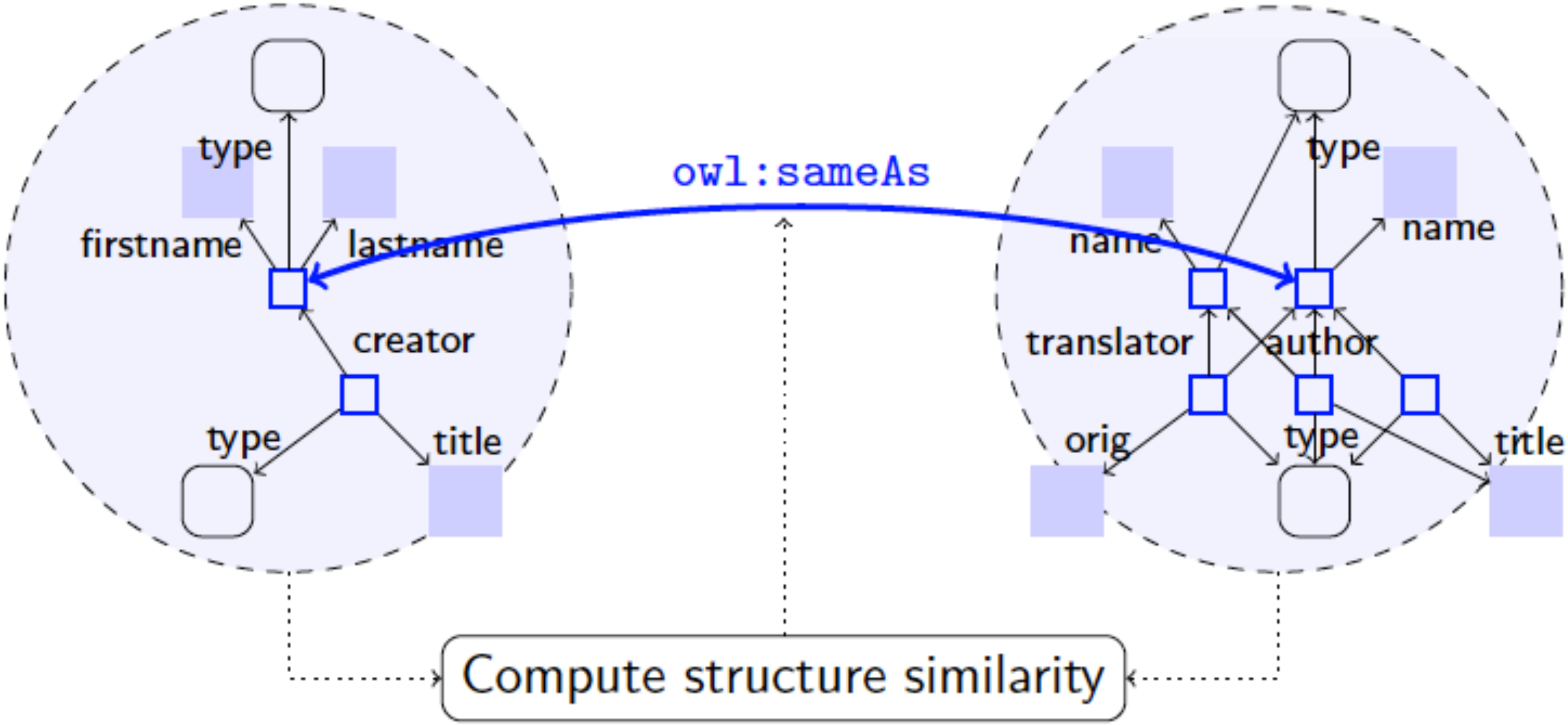
TERM-BASED MATCHING



Various tools:

- Normalisation (Stemmer, Tokenizers)
- Use of linguistic resources (Wordnet)
- Translation
- Many similarity measures, especially from IR (e.g., cosinus, n-grams)

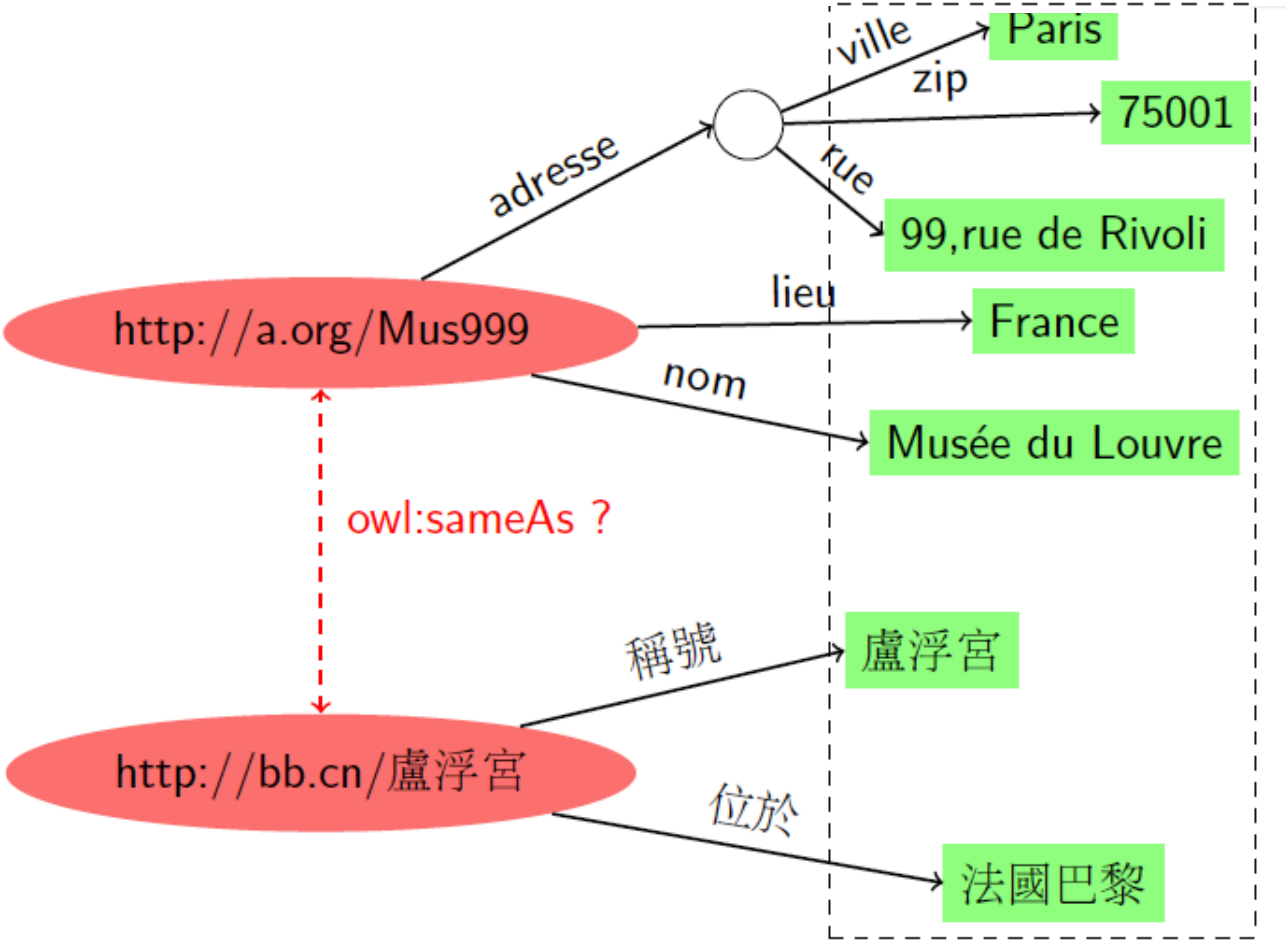
STRUCTURE-BASED MATCHING



Techniques:

- Based on graph matching techniques
- Can be used to learn weights on properties (but need matching)
- Problem: scalability

CROSS-LINGUAL RDF DATA LINKING

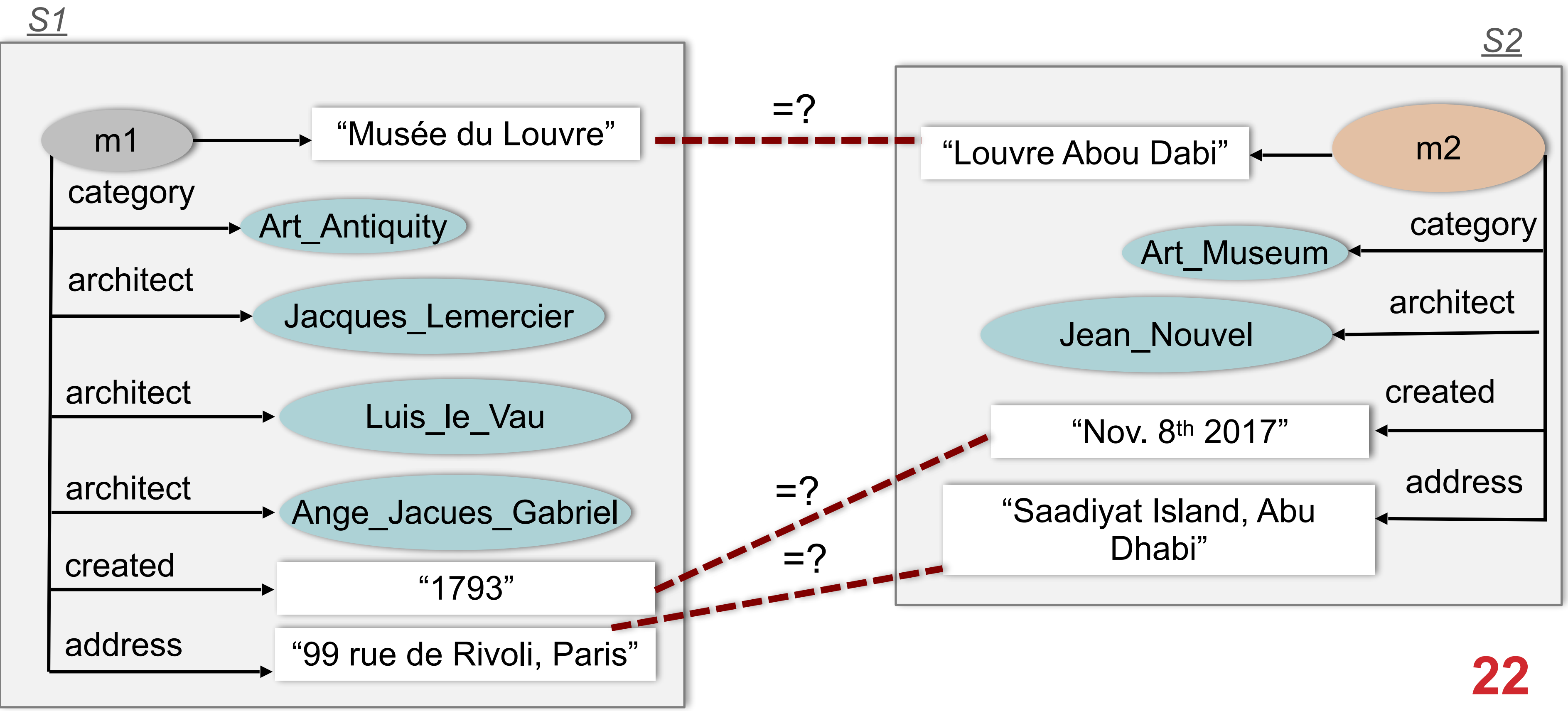


DATA LINKING APPROACHES

- **Instance-based approaches**: consider only data type properties (attributes)
- **Graph-based approaches**: consider data type properties (attributes) as well as object properties (relations) to propagate similarity scores/linking decisions (collective data linking)
- **Supervised approaches**: need an expert to build samples of linked data to train models (manual and interactive approaches)
- **Rule-based approaches**: need knowledge to be declared in the ontology or in other format given by an expert

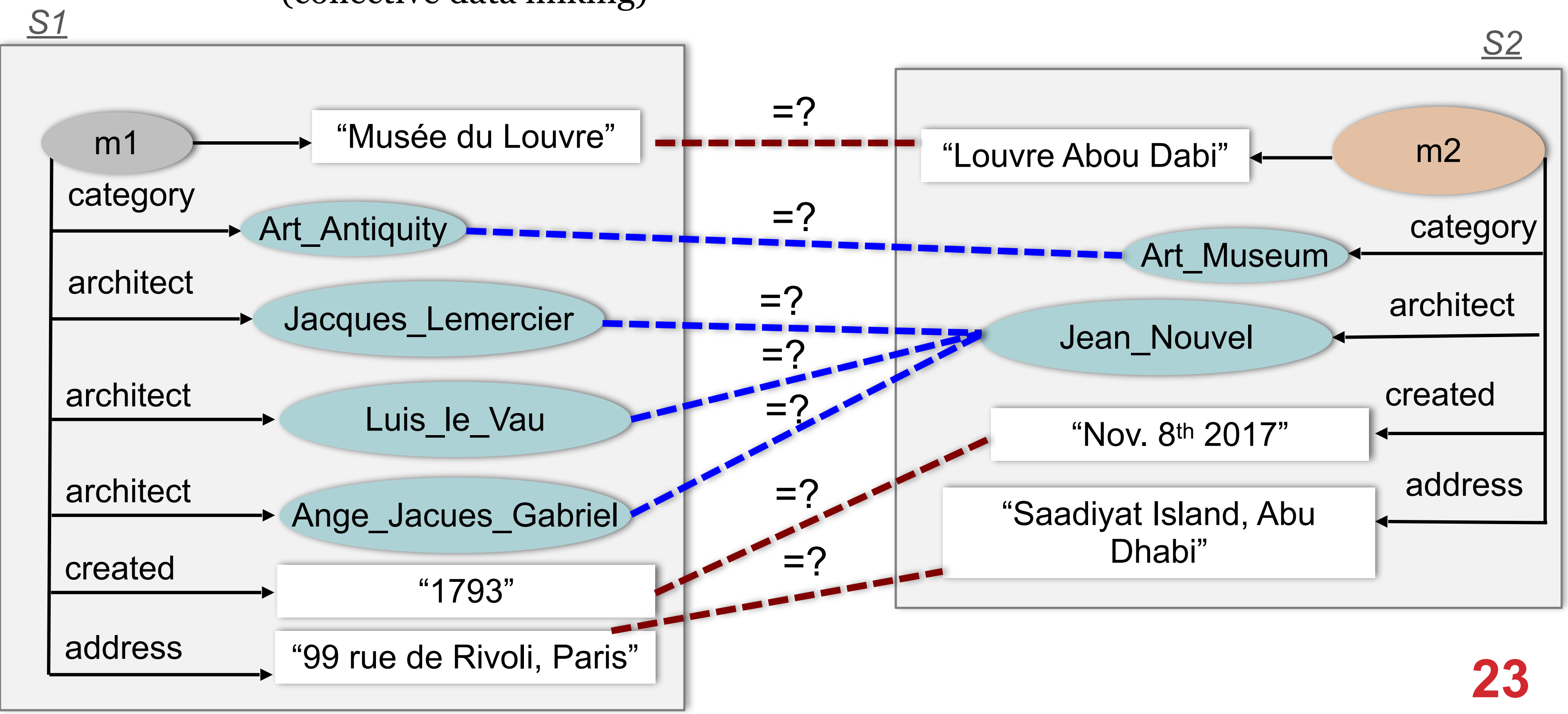
DATA LINKING APPROACHES

- **Instance-based approaches**: consider only data type properties (attributes)
    - String comparison



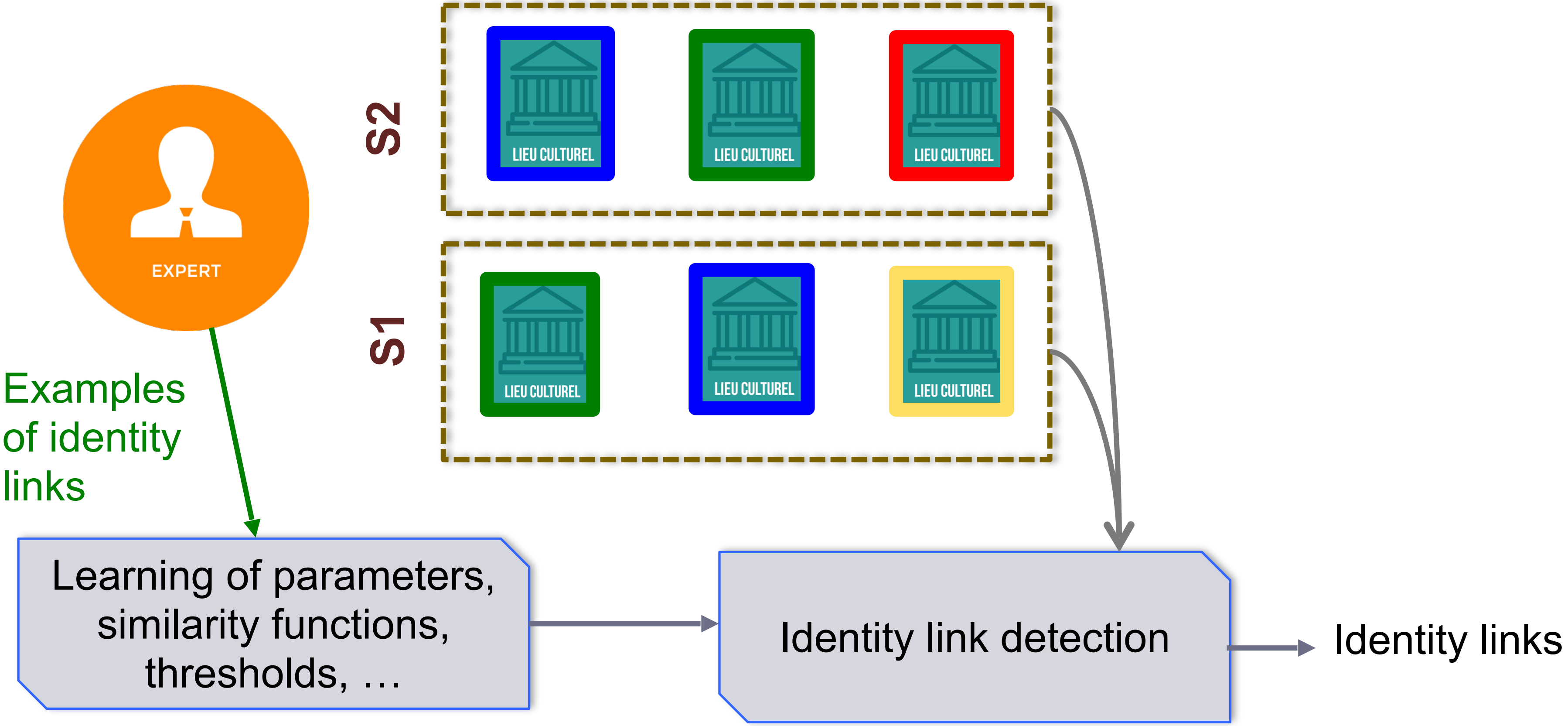
DATA LINKING APPROACHES

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



DATA LINKING APPROACHES

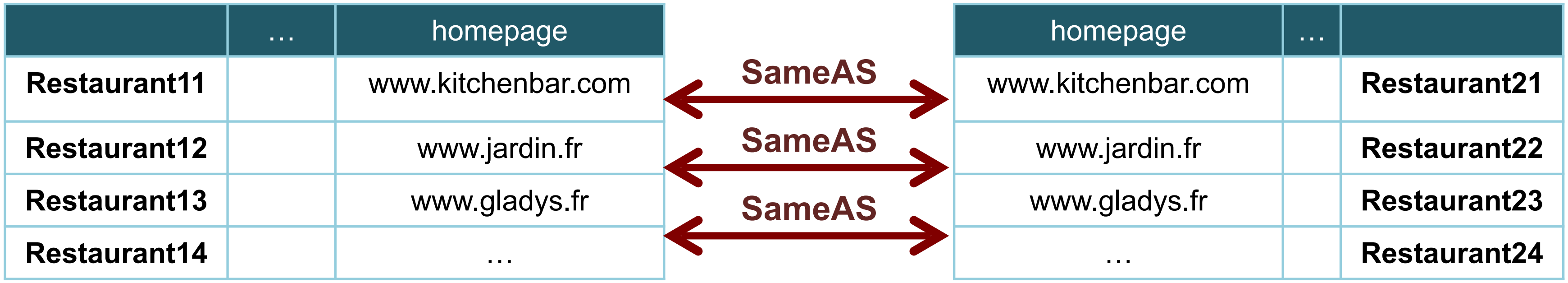
- **Supervised approaches**: need an expert to build samples of identity links to train models (manual and interactive approaches)



DATA LINKING APPROACHES

- **Rule-based approaches: need knowledge to be declared in the ontology or in other format given by an expert**

- **homepage(w1, y) ∧ homepage(w2, y) ➔ sameAs(w1, w2)**
  - *sameAs(Restaurant11, Restaurant21)*
  - *sameAs(Restaurant12, Restaurant22)*
  - *sameAs(Restaurant13, Restaurant23)*



	…	homepage
Restaurant11		www.kitchenbar.com
Restaurant12		www.jardin.fr
Restaurant13		www.gladys.fr
Restaurant14		…

homepage	…	
www.kitchenbar.com		**Restaurant21**
www.jardin.fr		**Restaurant22**
www.gladys.fr		**Restaurant23**
…		**Restaurant24**

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/mistakes in the data
- Use of benchmarks, like those of OAEI (Ontology Alignment Evaluation Initiative) or **Lance**

ONTOLOGY ALIGNMENT EVALUATION INITIATIVE - OAEI

« *OAEI is a coordinated **international initiative** to forge the a consensus for the **evaluation of the numerous methods** available for ontology and instance matching* »

Since 2004, it proposes every year a set of datasets:

- **Synthetic** datasets
- **Real** datasets (with some adaptations – injecting heterogeneity)

For both **Ontology alignment** and **Instance matching (data linking)**

The tools are run and compared on common platforms like **Hobbit*** and **Seals**#

* https://hobbit-project.github.io/index.html

# http://oaei.ontologymatching.org/2020/seals/index.html#tutorial

OUTLINE

- **Introduction to Data linking**
- **Overview of the well-know approaches**
  - Instance-based Data linking approaches
  - Graph-based Data linking approaches
  - Combined instance and ontology matching
- **Summary**

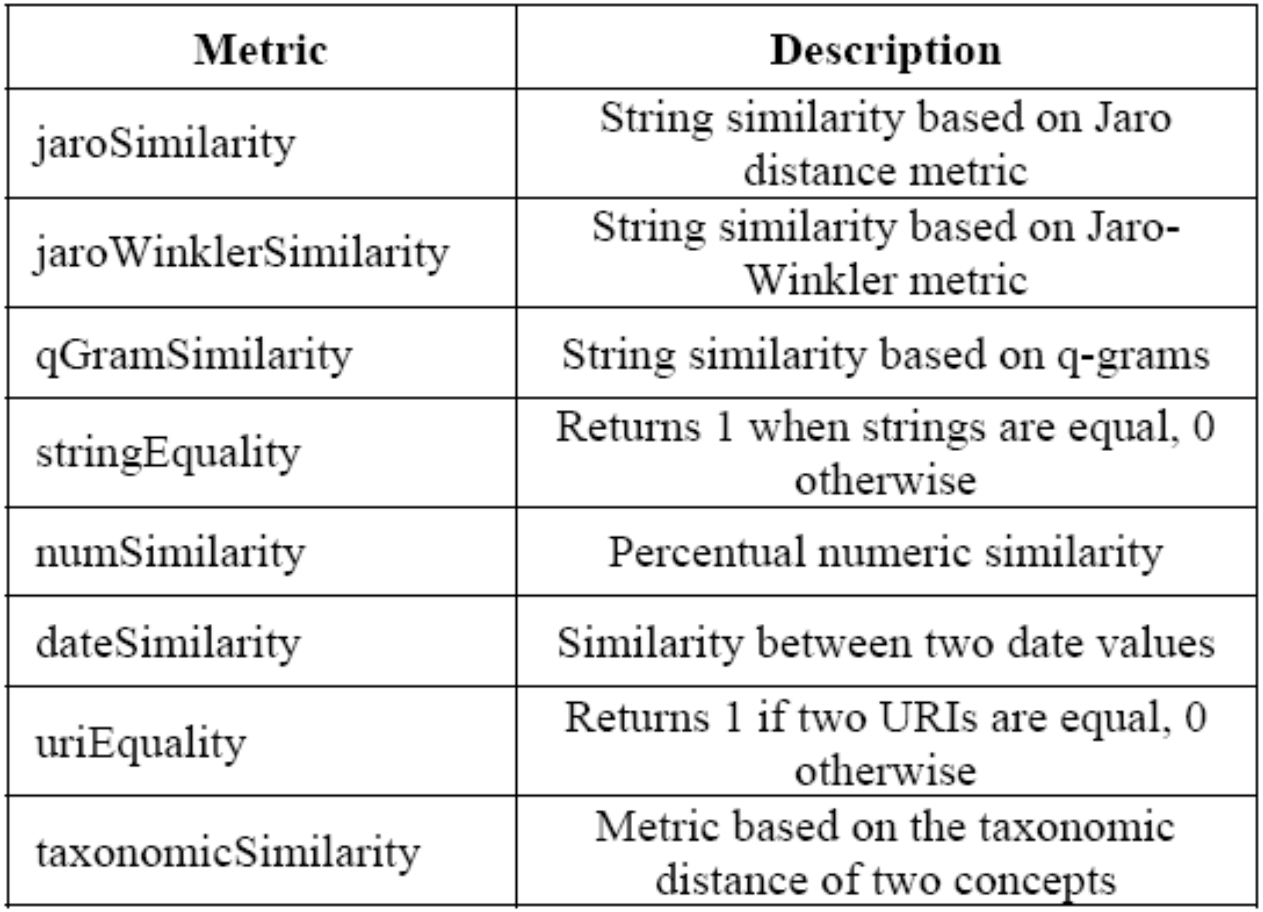
INSTANCE-BASED DATA LINKING APPROACHES

FRAMEWORK SILK [1]

- Provides a Link Specification Language(LSL)
- Allows specifying linking conditions between two datasets

- The linking conditions may be expressed in terms of:
  - Elementary similarity measures (e.g., Jaccard, Jaro) and
  - Aggregation functions (e.g. max, average) of the similarity scores

SIMILARITY MEASURES IN SILK



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

SILK: EXAMPLE OF LSL SPECIFICATION

```
<Silk>
  <Prefixes>
    <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#" />
  </Prefixes>

  <DataSources>
    <DataSource id="dbpedia">
      <Param name="endpointURI" value="http://demo_sparql_server1/sparql" />
      <Param name="graph" value="http://dbpedia.org" />
    </DataSource>

    <DataSource id="geonames">
      <Param name="endpointURI" value="http://demo_sparql_server2/sparql" />
      <Param name="graph" value="http://sws.geonames.org/" />
    </DataSource>
  </DataSources>
```

Prefixes

SPARQL endpoints

EXAMPLE OF LSL SPECIFICATION

```
<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

```
<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

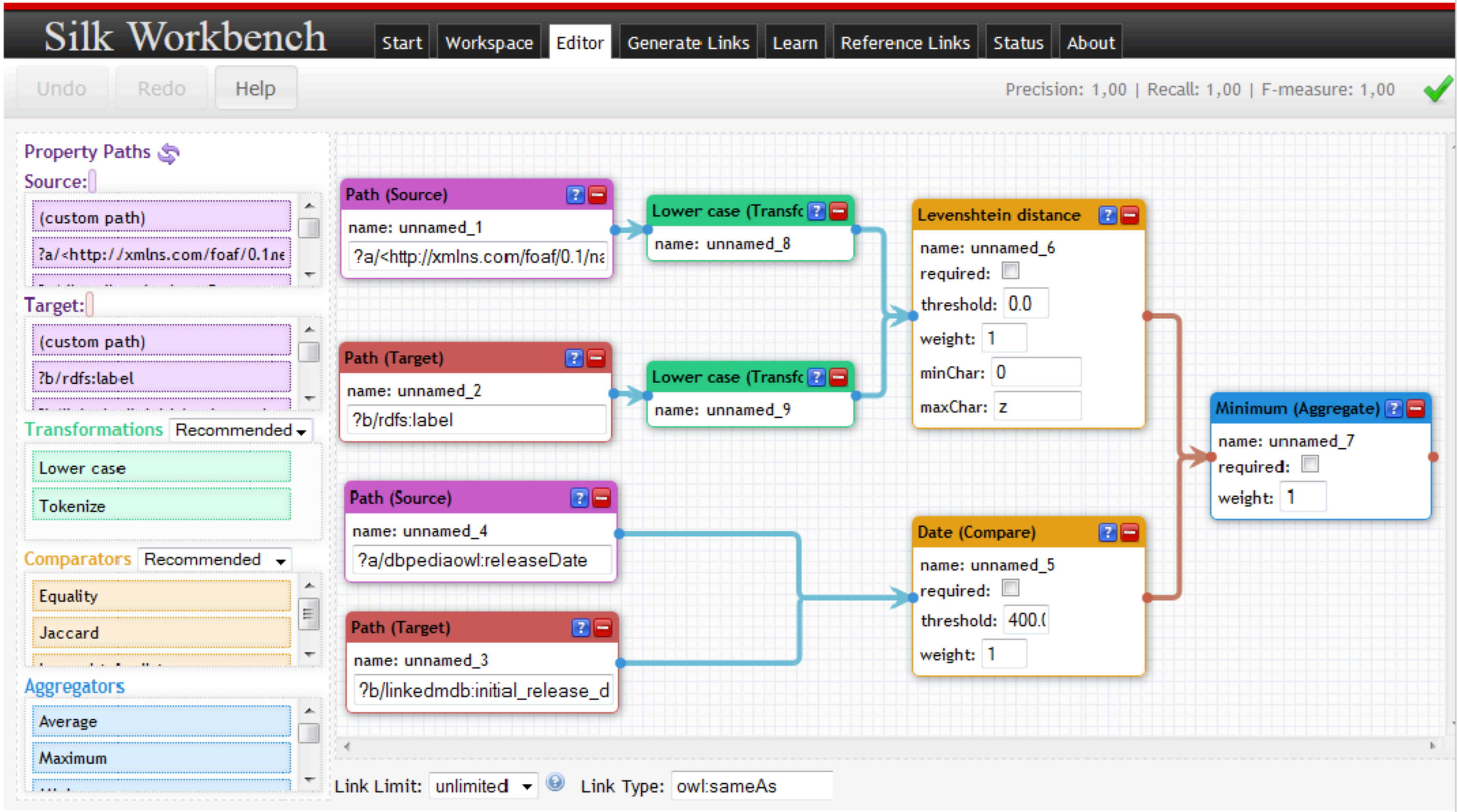
```
<Outputs>
        <Output type="file" minConfidence="0.95">
          <Param name="file" value="accepted_links.nt" />
          <Param name="format" value="ntriples" />
        </Output>
        <Output type="file" maxConfidence="0.95">
          <Param name="file" value="verify_links.nt" />
   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

</Silk>
```

Linking threshold

Possible links

SILK WORKBENCH



Precision: 1,00 | Recall: 1,00 | F-measure: 1,00 ✓

Source:
- (custom path)
- ?a/<http://xmlns.com/foaf/0.1/na

Target:
- (custom path)
- ?b/rdfs:label

Transformations: Recommended
- Lower case
- Tokenize

Comparators: Recommended
- Equality
- Jaccard

Aggregators
- Average
- Maximum

Path (Source)
name: unnamed_1
?a/<http://xmlns.com/foaf/0.1/na

Path (Target)
name: unnamed_2
?b/rdfs:label

Lower case (Transfc
name: unnamed_8

Lower case (Transfc
name: unnamed_9

Levenshtein distance
name: unnamed_6
required: □
threshold: 0.0
weight: 1
minChar: 0
maxChar: z

Minimum (Aggregate)
name: unnamed_7
required: □
weight: 1

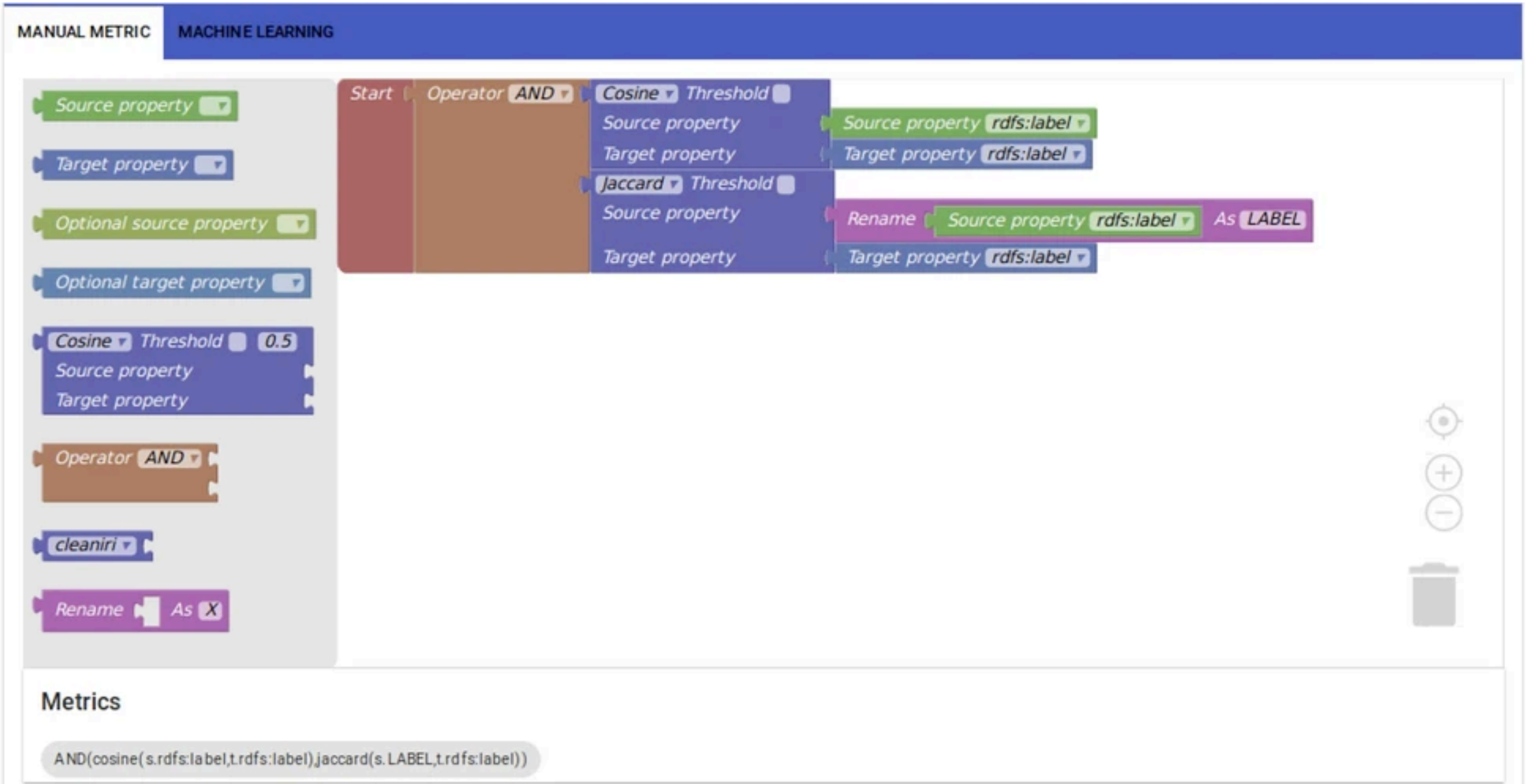
Path (Source)
name: unnamed_4
?a/dbpediaowl:releaseDate

Path (Target)
name: unnamed_3
?b/linkedmdb:initial_release_d

Date (Compare)
name: unnamed_5
required: □
threshold: 400.(
weight: 1

Link Limit: unlimited
Link Type: owl:sameAs

LIMES: A FRAMEWORK FOR LINK DISCOVERY ON THE SEMANTIC WEB

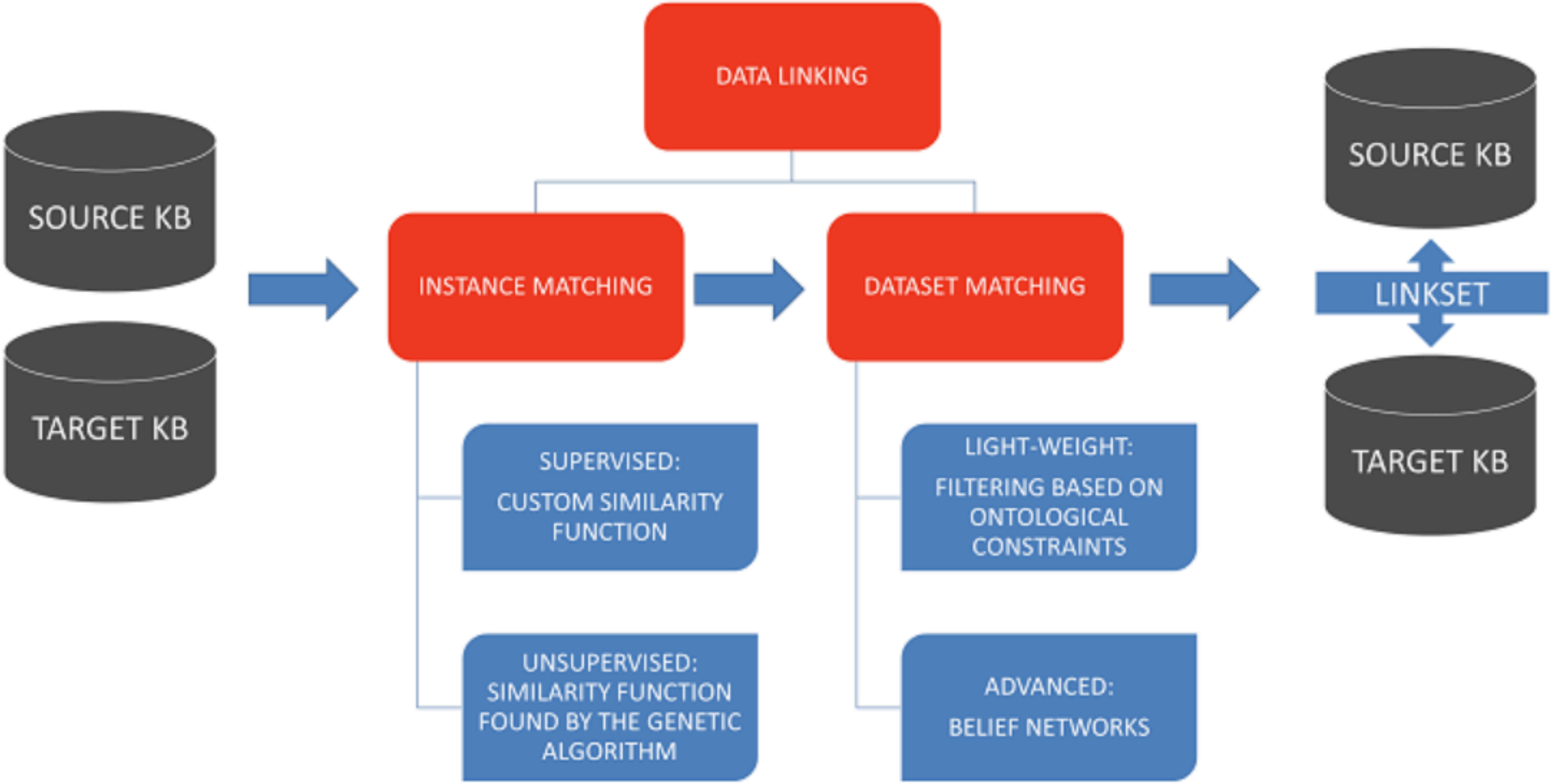


Metrics

AND(cosine(s.rdfs:label,t.rdfs:label),jaccard(s.LABEL,t.rdfs:label))

Ngonga Ngomo, AC., Sherif, M.A., Georgala, K. *et al.* LIMES: A Framework for Link Discovery on the Semantic Web. *Künstl Intell* **35,** 413–423 (2011). https://doi.org/10.1007/s13218-021-00713-x

KNOFUSS (INSTANCE-BASED, UNSUPERVISED) [2]



KNOFUSS (INSTANCE-BASED, UNSUPERVISED)

Instance matching

Aggregated attribute-based similarity

This method uses the classical approach to individual matching where:

- the **similarity between individuals** is calculated as an **aggregation** of similarities between their relevant properties
- the user can select the **properties** to **be compared, similarity functions, weights,** and the **cut-off threshold**

KNOFUSS (INSTANCE-BASED, UNSUPERVISED)

Instance matching

Unsupervised attribute-based similarity

This method also implements the aggregated attribute-based similarity

- instead of relying on the user to choose the parameters of the combined similarity function, it tries to **pick them automatically using a genetic algorithm.**
- in the absence of reliable training data, it **uses the desired distribution of resulting links to evaluate the fitness of candidate solutions**: e.g., the expected number of mappings

KNOFUSS (INSTANCE-BASED, UNSUPERVISED)

- Learns linking rules using genetic algorithms:

$$Sim(i1, i2) = f_{ag}(w_{11}sim_{11}(V11,V21), \ldots w_{mn}sim_{mn}(V1m,V2n))$$

  - F_{ag} : aggregation function for the similarity scores
  - sim_{ij}: similarity measure between values V1i and V2j
  - w_{ij}: weights in [0..1]

- **Assumptions:**
  - Unique name assumption (UNA), i.e., two different URIs refer to two different entities.
  - Good coverage rate between the two datasets
  - Normalized similarity scores in [0..1]

KNOFUSS (INSTANCE-BASED, UNSUPERVISED)

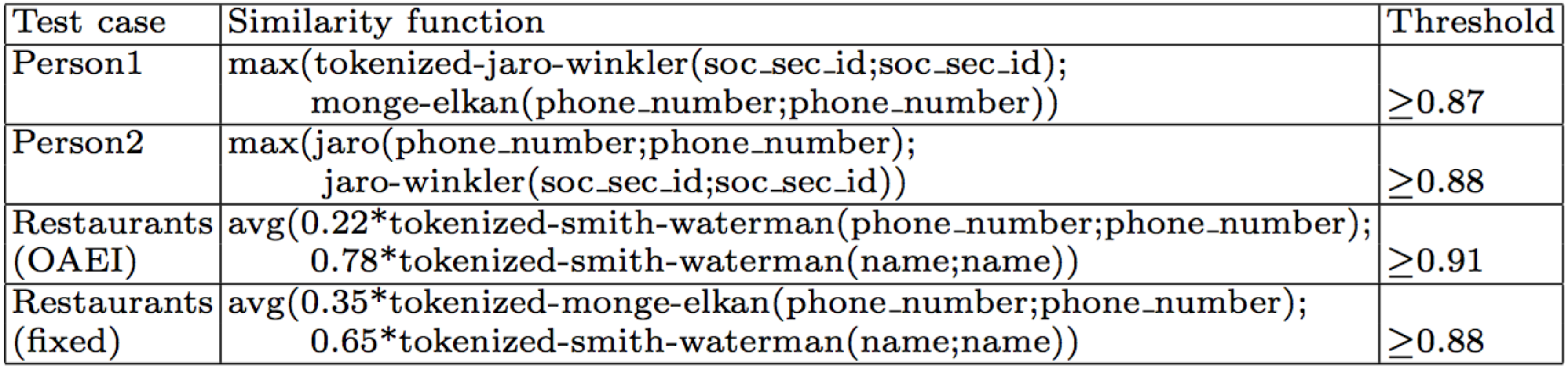
Dataset matching

Filtering based on ontological constraints:

uses explicitly defined ontological constraints (class disjointness, functionality and cardinality restrictions) to:

- **update** the original **set of mappings** provided by individual matching and
- **filter out** those which **violate** these constraints.

KNOFUSS (INSTANCE-BASED, UNSUPERVISED)



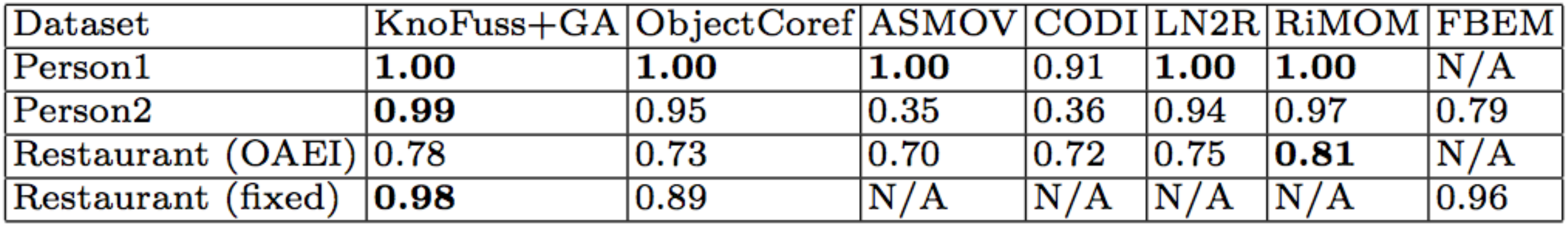
Test case	Similarity function	Threshold

Person1	max(tokenized-jaro-winkler(soc_sec_id;soc_sec_id);
	monge-elkan(phone_number;phone_number))
		≥0.87

Person2	max(jaro(phone_number;phone_number);
	jaro-winkler(soc_sec_id;soc_sec_id))
		≥0.88

Restaurants (OAEI)	avg(0.22*tokenized-smith-waterman(phone_number;phone_number); 0.78*tokenized-smith-waterman(name;name))	≥0.91
Restaurants (fixed)	avg(0.35*tokenized-monge-elkan(phone_number;phone_number); 0.65*tokenized-smith-waterman(name;name))	≥0.88

Examples of linking rules learned on the OAEI’10 benchmark



Dataset	KnoFuss+GA	ObjectCoref	ASMOV	CODI	LN2R	RiMOM	FBEM
Person1	**1.00**	**1.00**	**1.00**	0.91	**1.00**	**1.00**	N/A
Person2	**0.99**	0.95	0.35	0.36	0.94	0.97	0.79
Restaurant (OAEI)	0.78	0.73	0.70	0.72	0.75	**0.81**	N/A
Restaurant (fixed)	**0.98**	0.89	N/A	N/A	N/A	N/A	0.96

Results in term of F-Measure on OAEI’10

OUTLINE

- **Introduction to Data linking**

- **Overview of the well-know approaches**
  - Instance-based Data linking approaches
  - Graph-based Data linking approaches
  - Combined instance and ontology matching

- **Summary**

LN2R: A LOGICAL AND NUMERICAL METHOD FOR REFERENCE RECONCILIATON

[Saïs et al'07, Saïs et al'09]

LN2R

(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al'07, Saïs et al'09]

- A combination of two methods:
  - L2R, a Logical method for reference reconciliation: applies logical rules to infer sure owl:sameAs and owl:differentFrom links
  - N2R, a Numerical method for reference reconciliation: computes similarity scores for each pair of references
- **Assumptions**
  - The datasets are conforming to the same ontology
  - The ontology contains axioms

LN2R

(GRAPH BASED, UNSUPERVISED AND INFORMED)

Ontology axioms

- Disjunction axioms between classes, DISJOINT(C, D)
- Functional properties axioms, PF(P)
- Inverse functional properties axioms, PFI(P)
- A set of properties that is functional or inverse functional axioms

Assumptions on the data

- Unique Name Assumption, UNA(src1)
- Local Unique Name Assumption, LUNA(R)

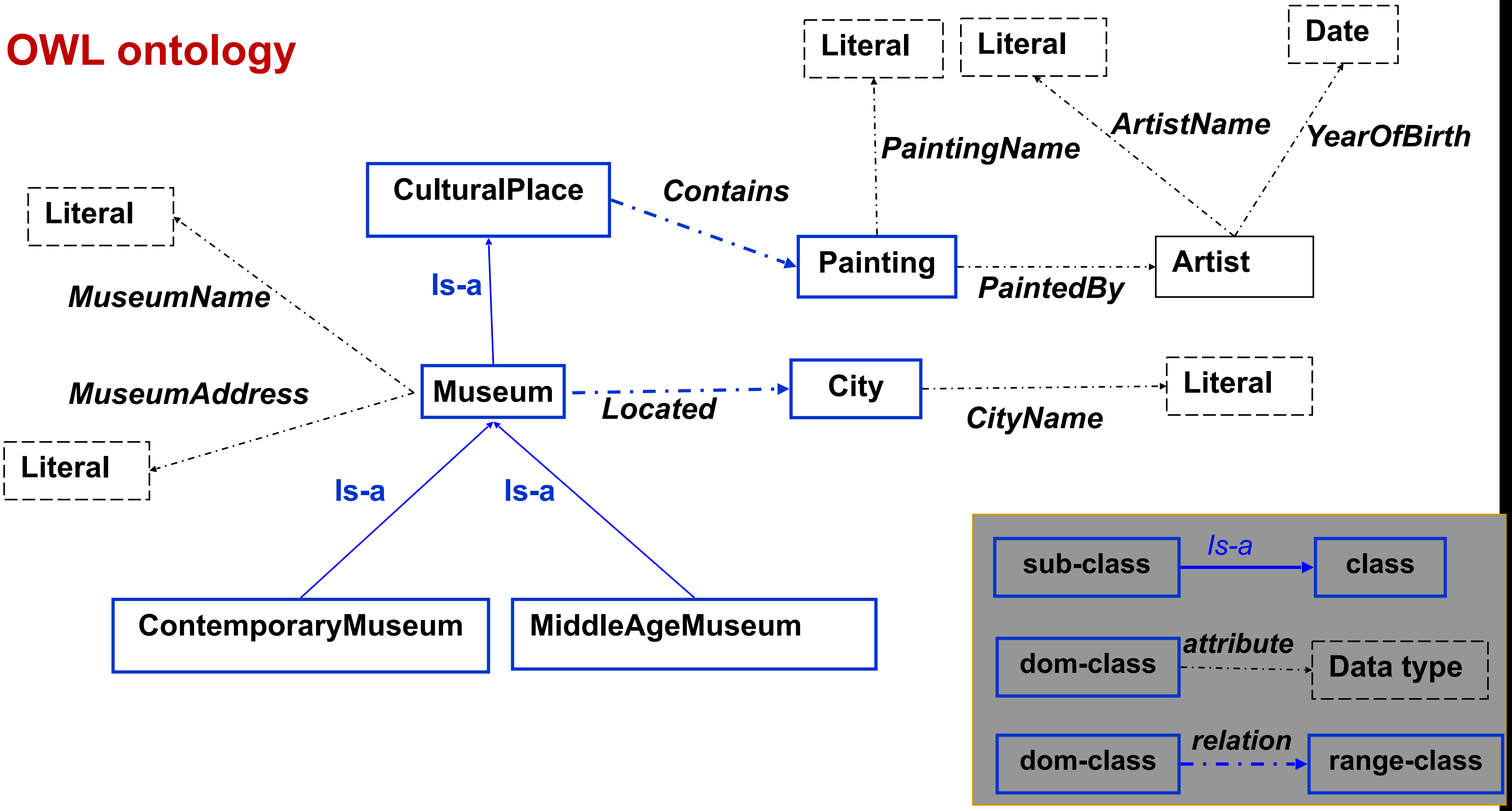
Example:

Authored(p, a1), Authored(p, a2), Authored(p, a3) …., Authored(p, an)
➔ (a1 ≠a2), (a1 ≠ a3), (a2 ≠ a3) , …

LN2R

(GRAPH BASED, UNSUPERVISED AND INFORMED)

OWL ontology

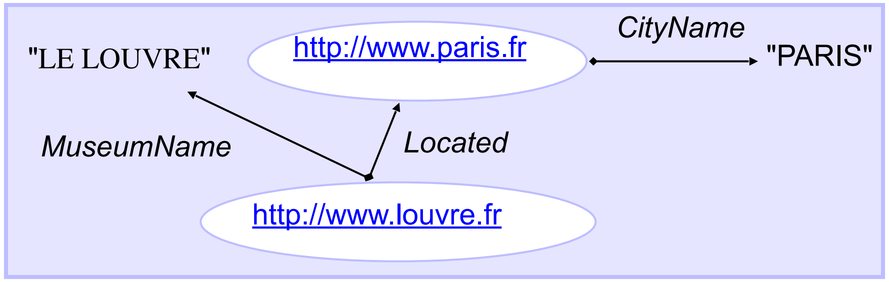


LN2R

(GRAPH BASED, UNSUPERVISED AND INFORMED)

RDF datasets

- **RDF Graphs:**



- **RDF Facts:**

Desc(http://www.louvre.fr)= {

Museum(http://www.louvre.fr),
Located(http://www.louvre.fr,http://www.paris.fr),
MuseumName(http://www.louvre.fr,"LE LOUVRE")}

Desc(http://www.paris.fr)= {

Located(http://www.louvre.fr,http://www.paris.fr),
CityName(http://www.paris.fr,"PARIS")}

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

Ontology axioms:

- Disjunction axioms between classes, DISJOINT(C, D)
- Functional properties axioms, PF(P)
- Inverse functional properties axioms, PFI(P)
- A set of properties that is functional or inverse functional axioms

Assumptions on the data

- Unique Name Assumption, UNA(src1)
- Local Unique Name Assumption, LUNA(R)

Example:

Authored(p, a1), Authored(p, a2), Authored(p, a3) …., Authored(p, an)
➔ (a1 ≠a2), (a1 ≠ a3), (a2 ≠ a3) , …

LN2R

(GRAPH BASED, UNSUPERVISED AND INFORMED)

- Disjunction axioms between classes DISJOINT(C, D), its logical semantics:

$$\forall X \quad C(X) \Rightarrow \neg D(X)$$

- Functional properties axioms, PF(P), its logical semantics:

$$\forall X, Y, Z \quad P(X,Y) \wedge P(X, Z) \Rightarrow Y=Z$$

- Inverse functional properties axioms, its logical semantics:

$$\forall X, Y, Z \quad P(Y,X) \wedge P(Z, X) \Rightarrow Y=Z$$

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

SWRL rules are used to generalize:

- Functionality axioms to a set of properties (relations and attributes) *{P1,…, Pn}*, PF(*P1,…, Pn)*, *its logical semantics:*

$$\forall X1,\ldots, Xn, Y, Z \bigwedge_{\forall i \in [1..n]} (Pi(Xi, Y) \wedge Pi(Xi, Z) \Rightarrow Y=Z$$

- Inverse functionality axioms to a set of properties (relations and attributes)*{P1,…, Pn}*, PF(*P1,…, Pn)*, *its logical semantics:*

$$\forall X1,\ldots, Xn, Y, Z \bigwedge_{\forall i \in [1..n]} (Pi(Y,Xi) \wedge Pi(Z, Xi) \Rightarrow Y=Z$$

L2R: A LOGICAL METHOD FOR REFERENCE RECONCILIATION

L2R: AUTOMATIC GENERATION OF INFERENCE RULES

Translation of UNA(src1)

$\mathrm{R1{:}src1}(X) \wedge \mathrm{src1}(Y) \wedge (X \neq Y) \Rightarrow \neg \mathrm{Reconcile}(X,Y)$; …

Translation of LUNA(R)

$\mathrm{R11}(R) : R(Z, X) \wedge R(Z, Y) \wedge (X \neq Y) \Rightarrow \neg \mathrm{Reconcile}(X,Y)$; …

Translation of DISJOINT(C, D):

$\mathrm{R5}(C, D) : C(X) \wedge D(Y) \Rightarrow \neg \mathrm{Reconcile}\ (X, Y)$

Translation of PF(R):

$\mathrm{R6.1}(R): \mathrm{Reconcile}(X, Y) \wedge R(X, Z) \wedge R(Y, W) \Rightarrow \mathrm{Reconcile}\ (Z, W)$

$\mathrm{R6.1}(\mathrm{Located}): \mathrm{Reconcile}(X, Y) \wedge \mathrm{Located}\ (X, Z) \wedge \mathrm{Located}\ (Y, W) \Rightarrow \mathrm{Reconcile}\ (Z, W)$

Translation of PF(A):

$\mathrm{R6.2}(A): \mathrm{Reconcile}(X, Y) \wedge A(X, Z) \wedge A(Y, W) \Rightarrow \mathrm{SynVals}(Z, W)$

$\mathrm{R6.2}(\mathrm{MuseumName}): \mathrm{Reconcile}(X,Y) \wedge \mathrm{MuseumName}\ (X, Z) \wedge \mathrm{MuseumName}\ (Y,W)$
$\Rightarrow \mathrm{SynVals}(Z, W)$

L2R: INFERENCE ALGORITHM

- Apply until saturation the resolution principle [Robinson'65], by following the unit strategy

  Resolution rule : $\dfrac{C_1 : (L_1), C_2 : (L_2 \vee C)}{C_{1,2} : (C_{\sigma})}$ Avec $L_{1\sigma} = \neg L_{2\sigma}$

- R ∪ F: Horn clauses without functions, where :

  - R: rules in the form of horn clauses
  - F: unit clauses fully instantiated,
    - Reference descriptions: RDF facts (class-facts, relation-facts and attribute-facts).
    - Facts that express the reference origin: src1(i) and src2(j)
    - Facts that express the synonymy and not synonymy between values: SynVals(v1, v2) or ¬ SynVals(v1, v2)

- Computation of the set SatUnit(R ∪ F)

L2R: ALGORITHM PROPERTIES

- Termination of the algorithm: guaranteed thanks to the absence of function symbols in the knowledge base
- Completeness: for the deduction of all the unit clauses fully instantiated, *Reconcile* and *SynVals*.

Theorem : *Let R be a set of un Horn clauses without functions. Let F be a set of unit clauses fully instantiated. If $R \cup F$ is satisfiable, then:*

$$\forall\, p(\mathbf{a}), \quad (R \cup F \models p(\mathbf{a})) \Rightarrow (p(\mathbf{a}) \in SatUnit(R \cup F))$$

With $p(\mathbf{a})$, a unit clause fully instantiated and $SatUnit(R \cup F)$ is the set of inferred clauses by applying the unit resolution until saturation on $R \cup F$.

L2R: EXAMPLE OF AXIOMS

Disjunction : {DISJOINT(MiddleAgeMuseum,ContemporaryMuseum), DISJOINT(Painting, Artist), DISJOINT(CulturalPlace, City), DISJOINT(CulturalPlace,Painting)}.

Functional properties: {PF(Located), PF(PaintedBy), PF(ArtistName), PF(YearOfBirth), PF(PaintingName), PF(CityName), PF(MuseumName), PF(MuseumAddress)}.

Inverse functional properties:

{PFI(PaintingName, PaintedBy), PFI(Contains), PFI(ArtistName), PFI(MuseumName), PFI(MuseumAddress), PFI(CityName)}.

L2R: EXAMPLE OF DATASETS

S1

CulturalPlace(S1_m1); Museum(S1_m2);
MiddeleAgeMuseum(S1_m3), Painting(S1_p1);
Painting(S1_p2); Painting(S1_p3) Artist(S1_a1);
Artist(S1_a2); City(S1_c1);
MuseumName(S1_m1,"musee du LOUVRE");
Contains(S1_m1,S1_p1);
MuseumName(S1_m2,"musee des arts premiers");

MuseumAddress(S1_m2, "quai branly");
Located(S1_m2,S1_c1); CityName(S1_c1,"Paris");
PaintingName(S1_p1, "La Joconde");
PaintedBy(S1_p1,S1_a1);

ArtistName(S1_a1, "Leonard De Vinci");

PaintingName(S1_p2,"La Cene");

PaintedBy(S1_p2, S1_a1);

S2

Museum(S2_m1); Museum(S2_m2);
Painting(S2_p1); ContemporaryMuseum(S2_m4)
Painting(S2_p2);Painting(S2_p3); Artist(S2_a1);
City(S2_c1); MuseumName(S2_m1,"Le LOUVRE");
Located(S2_m1,S2_c1); Contains(S2_m1,S2_p2);
Contains(S2_m1, S2_p1);
MuseumName(S2_m2,"Musée du quai Branly");
MuseumAddress(S2_m2, "37 quai branly, portail
Debilly"); Contains(S2_m1,S2_p3);
Located(S2_m2,S2_c1);
CityName(S2_c1, "Ville de paris");
PaintingName(S2_p2, "Vierge aux rochers");
PaintedBy(S2_p2,S2_a1);
ArtistName(S2_a1,"De Vinci");
PaintingName(S2_p3, "Sainte Anne, la vierge et
l'enfant jesus"); PaintingName(S2_p1, "la Joconde");

The UNA is stated in the two sources S1 and S2.

L2R: RUNNING EXAMPLE DE

Instantiated rules

R1, R2

R5(CulturalPlace, Painting)
R5(Artist, Painting)
R5(MiddleAgeMuseum, ContemporaryMuseum)
…

Fact set

scr1(S1_m2), scr1(S1_p1), scr1(S1_p2), scr2(S2_m1),
scr2(S2_p1), scr2(S2_p2),
CulturalPlace(S1_m1), Painting(S2_p1)
Artist(S1_a1), Painting(S2_p2)
MiddeleAgeMuseum(S1_m3),ContemporaryMuseum(S2_m4)
…

REC

NREC

¬Reconcile(S1_m1,S1_m2), ¬Reconcile(S1_p1,S1_p2),
¬Reconcile(S2_m1,S2_p1), ¬Reconcile(S2_p1, S2_p2)
¬Reconcile(S1_m1, S2_p1),
¬Reconcile(S1_a1, S2_p1)
¬Reconcile(S1_m3, S2_m4)

SynVals(“La Joconde”,” la joconde”)

L2R: RUNNING EXAMPLE DE

Instantiated rules

…
R7.2 (PaintingName)

Fact set

…
PaintingName(S1_p1,”La joconde”),
PaintingName(S2_p1,″ La Joconde″)

REC

Reconcile(S2_p1, S1_p1)

NREC

¬Reconcile(S1_m1,S1_m2), ¬Reconcile(S1_p1,S1_p2),
¬Reconcile(S2_m1,S2_p1), ¬Reconcile(S2_p1, S2_p2)
¬Reconcile(S1_m1, S2_p1),
¬Reconcile(S1_a1, S2_p1)
¬Reconcile(S1_m3, S2_m4)

SynVals(“La Joconde″,″ la joconde″)

L2R: RUNNING EXAMPLE DE

Instantiated rules

…
R7.1(Contains)
R4. “UNA”
R6.2(MuseumName)
R6.1(Located),
R6.2(CityName)

Fact set

…
Contains(S1_m1, S1_p1), Contains(S2_m1, S2_p1)
src1(S1 m1), src2 (S2 m1), scr2 (S2 m2),
MuseumName(S1 m1, ‘musee du LOUV RE”)
MuseumName(S2 m1, “LE LOUV RE”)
Located(S1 m1, S1 c1), Located(S2 m1, S2 c1)

REC

Reconcile(S2_p1, S1_p1)
Reconcile(S1_m1, S2_m1)
Reconcile(S1_c1, S2_c1)

SynVals(“La Joconde”,” la joconde”)
SynVals(“musee du LOUVRE”, “LE LOUVRE”)
SynVals(”ville de Paris”,”Paris”)

NREC

¬Reconcile(S1_m1,S1_m2), ¬Reconcile(S1_p1,S1_p2),
¬Reconcile(S2_m1,S2_p1), ¬Reconcile(S2_p1, S2_p2)
¬Reconcile(S1_m1, S2_p1),
¬Reconcile(S1_a1, S2_p1)
¬Reconcile(S1_m3, S2_m4)
¬Reconcile(S2_m2, S1_m1)

N2R: A Numerical Method for Reference Reconciliation

[4]

N2R: A NUMERICAL METHOD FOR REFERENCE RECONCILIATION

- **N2R computes a similarity score for pair of references obtained from their common description.**
    - Uses known similarity measures, e.g. Jaccard, Jaro-Winkler.
    - Exploits ontology knowledge in a way to be coherent with L2R.
    - May consider the results of L2R: *Reconcile(i, i')*, ¬*Reconcile(i, i')* , *SynVals(v, v')* and ¬*SynVals(v, v')*.

N2R: COMMON DESCRIPTION

- Common attributes for a reference pair (i, i’):

$$\mathrm{CAttr}(i, i') = \{ a \mid \exists\, v, v' \in \mathrm{Val}, \text{ st. } [a(i, v) \in \mathrm{Desc}(i) \text{ and } a(i', v') \in \mathrm{Desc}(i')]\}$$

- Common relations for a reference pair (i, i’):

$$\mathrm{CRel}(i, i') = \{ r \mid \exists\, j, j' \in I, \text{ st. } [r(i, j) \in \mathrm{Desc}(i) \text{ and } r(i', j') \in \mathrm{Desc}(i')] \text{ or } [r(j, i) \in \mathrm{Desc}(i) \text{ and } r(j', i') \in \mathrm{Desc}(i')] \}$$

- Set of values associated to a reference i:

$$a{+}(i) = \{ v \mid \forall\, v, \text{ st. } a(i,v) \in \mathrm{Desc}(i)\}$$

- Set of references associated to a reference i:

$$r{+}\,(i) = \{ j \mid \forall\, j,\ r(i, j) \in \mathrm{Desc}(i)\}$$

- Set of references to which a reference i is associated to a reference:
- $r{-}\,(i) = \{ j \mid \forall\, j,\ r(j, i) \in \mathrm{Desc}(i)\}$

SIMILARITY DEPENDENCY MODELLING

RDF facts in source S1:

Located(m1, c1), MuseumName(m1, “le Louvre”)

Contains(m1, p1), CityName(c1, “Paris”)

PaintingName(p1, “la Joconde”)

RDF facts in source S2 :

Located(m’1, c’1), MuseumName(m’1, “Louvre”)

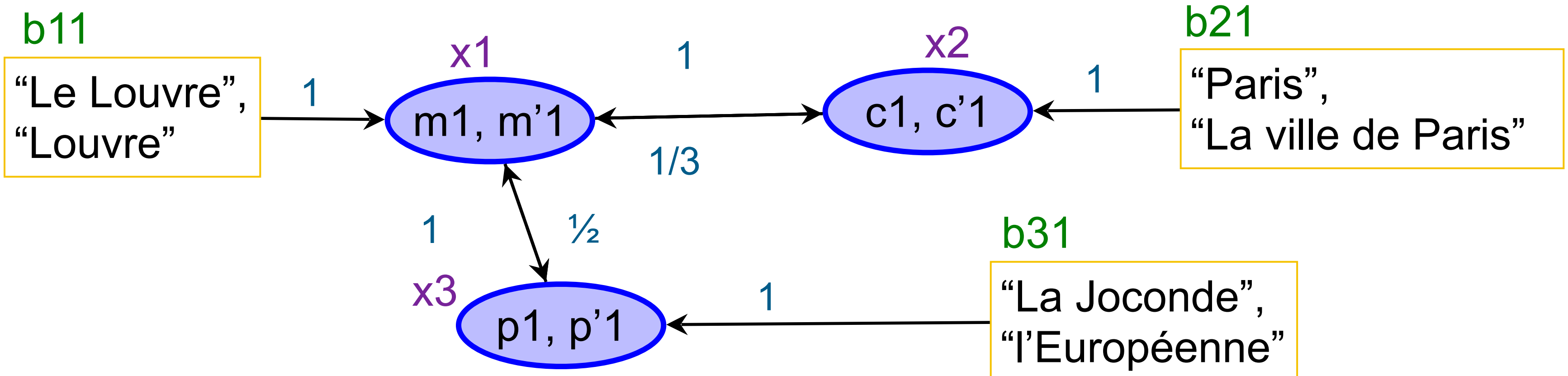
Contains(m’1, p’1), CityName(c’1, “la Ville de Paris”)

PaintingName(p’1, “l’Europèenne”)

CAttr(m1, m’1) = {MuseumName} ,
CAttr(c1, c’1)= {CityName},CAttr(p1,p’1)={PaintingName}
CRel(m1, m’1)= {Located, Contains}
CRel(c1, c’1) = {Located }, CRel(p1,p’1) = {Contains}

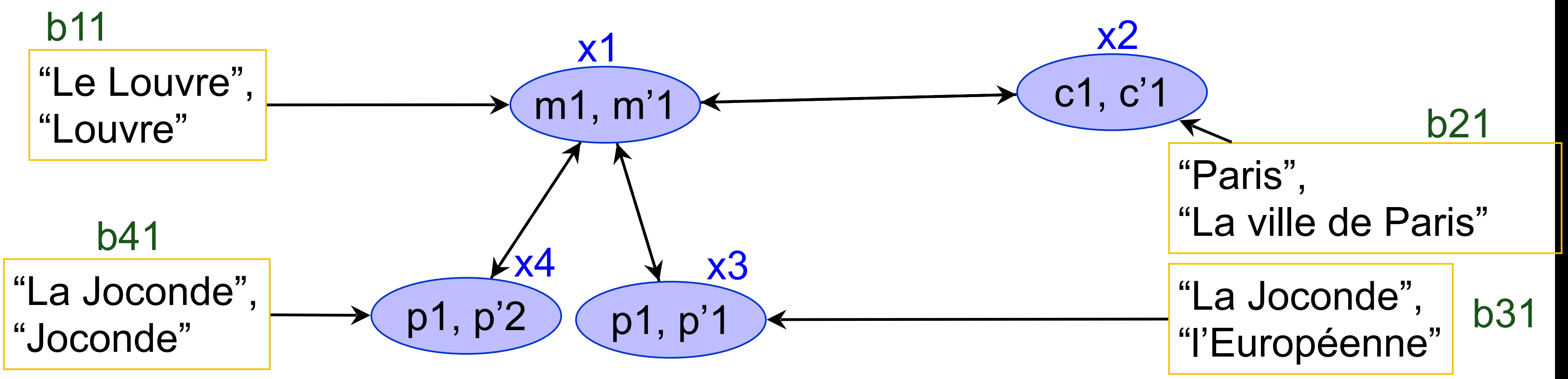
MuseumName+(m1) = {”Le Louvre”},
MuseumName+(m’1) = {”Louvre”},
Located+(m1) = {c1}, Located+(m’1) = {c’1},
Located-(c1) = {m1} , Located-(c’1) = {m’1}, ….

(c1, c’1) is functionally dependent on (m1, m’1)



➔ Equation system

N2R: ILLUSTRATION



b11

"Le Louvre", "Louvre"

x1

m1, m'1

x2

c1, c'1

b21

"Paris", "La ville de Paris"

b41

"La Joconde", "Joconde"

x4

p1, p'2

x3

p1, p'1

"La Joconde", "l'Européenne"

b31

$\mathbf{x1} = \max(\max(b11, x3), x4), \lambda * x2)$

$\mathbf{x2} = \max(b21, x1)$

$\mathbf{x3} = \max(b31, \lambda * x1)$

$\mathbf{x4} = \max(b41 , \lambda * x1)$

	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

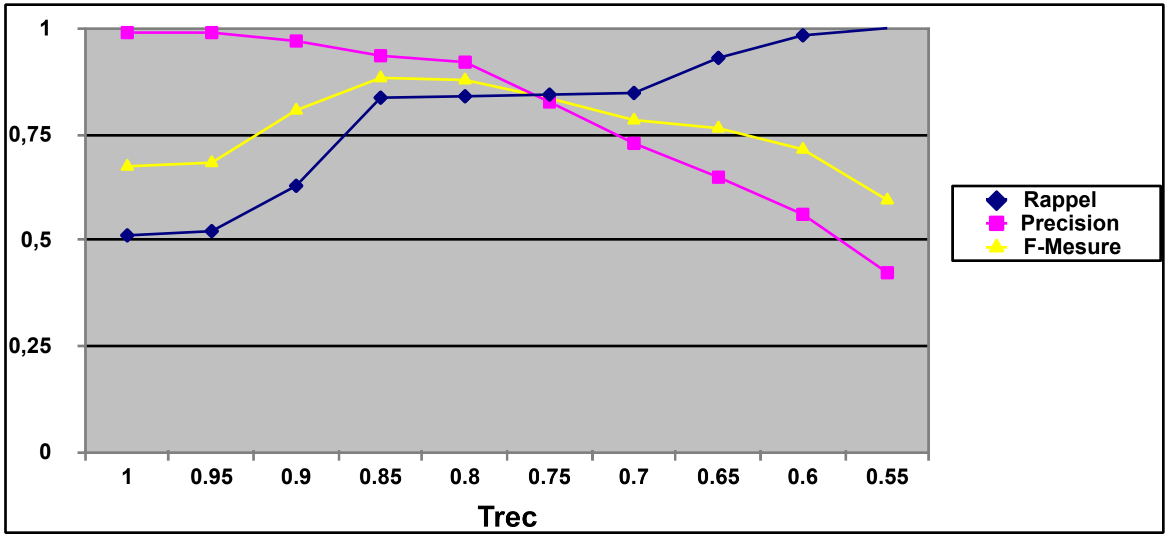
$\lambda = 1/(| CAttr | + | CRel |)$ $\varepsilon = 0.02$

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

N2R EXPERIMENTS

N2R: RESULTS ON CORA



Trec=1, all the reconciliations obtained by L2R are also obtained by N2R.

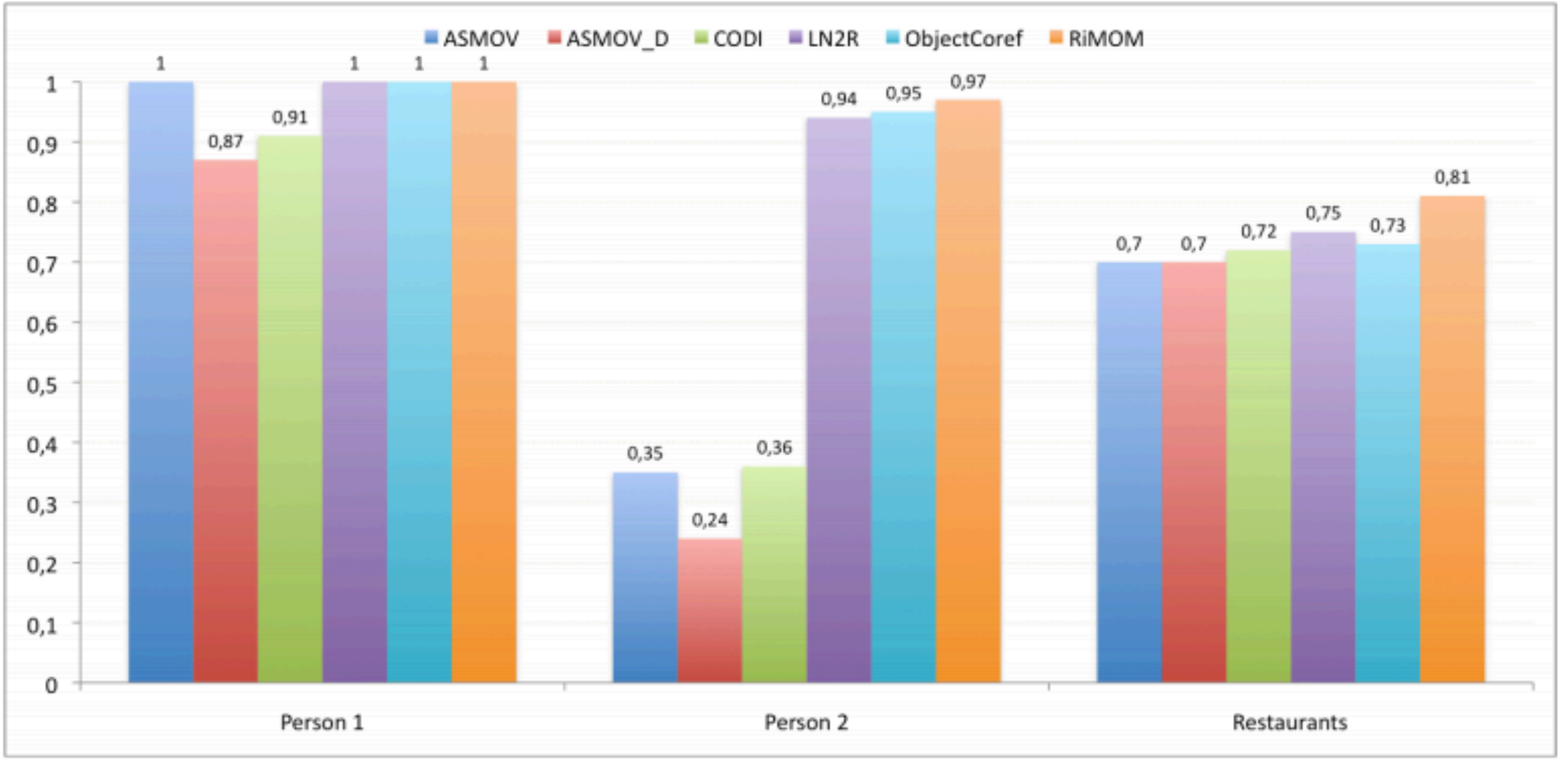
Trec=1 to Trec=0.85, the recall increases of 33 % while the precision decreases only of 6 %.

Trec = 0.85, the F-measure is of 88 %:

- Better than the results obtained by the supervised methods.

N2R: RESULTS IN OAEI[2] 2010

OAEI 2010 – Instance Matching track (PR), 2nd



[2] Ontology Alignment Evaluation Initiative

COMBINED INSTANCE AND ONTOLOGY MATCHING

ONTOLOGY MATCHING

- **Ontology alignment** [Shvaiko,Euzenat13]: computes a set A of mappings between elements (classes, properties) of two ontologies O1 and O2:

  f(O1,O2)=A

- The relations that are used to express a mapping can be: **owl:equivalentClass, owl:equivalentProperty, rdfs:subClassOf, skos:closeMatch, skos:broader**, etc.

- Example: A={(

  owl:equivalentClass(http://dbpedia.org/ontology/City,
  http://schema.org/City, *0.8*)}

KINDS OF INFORMATION

- **Terminology**: lexical information describing the ontology elements (i.e. labels, comments, …)
  - *Example: Way* vs *Underground way*

- **Structure**: hierarchy of classes and properties (relations/ attributes)
  - *Example: the sub-classes of Way are very similar to the sub-classes of Path*

- **Extension**: the existence of **common instances**!!

PARIS [5]

- Objective: instance-based ontology alignment and data linking (graph-based, unsupervised and probabilistic)
- Inputs: two populated RDFS ontologies with UNA (two different URI refer to two different entities)
- Principle:
  - Compute the similarities between literal values (“12 cm”=“12”)
  - Iterate (1) and (2) until a fix-point :
    - ① Compute the probability that two instances are linked

$$P(i_1 = i_2)$$

    - ① Compute the probabilities of subClassOf/subPropertyOf

$$P(C_i \subseteq C_j), P(P_i \subseteq P_j)$$

PARIS

- Property functionality degree (computed from data)
  - *The more a property is functional the more the probability of* X=Y will be.

- **Local functionality**: $\mathrm{Fun}(p,x) = 1 / \#y:p(x, y)$
- **Global functionality**: $\mathrm{Fun}(p) = (\#x : \exists y:p(x,y)) / (\#x,y : p(x,y))$

- **Example**:

  city(m1,Londres), city(m1,Orsay), city(m2,Tokyo)

  Fun(city,m1)= ½ Fun(city,m2)=1

  Fun(city)=2/3

➔ The same is done for **inverse functionality** (denoted fun^{-1})

PARIS

Link probability computation

- **Positive evidence (P1)**: if there exists a property that is highly inverse functional which has range values that are equal with a high probability

$$P_1(x=x') = 1 - \prod_{\substack{r(x,y)\\ r(x',y')}} (1 - Fun^{-1}(p).P(y=y'))$$

isbn(x,isbn1), isbn(x’,isbn2), P(isbn1=isbn2) = 1, fun^{-1}(isbn)=1 …

P1(x=x’) = 1 - ((1 - (1.1)) . …) = 1-(0. …) = 1

- **Negative evidence (P2)**: if there exists a property that is highly functional which has range values for the probability to be equal is very low.

- **Combination** : P(x=x’) = P1(x=x’).P2(x=x’)

PARIS

- The probabilities of the existence of a subsumption mapping between properties and between classes are also computed
- It is based on the proportion of common instances comparing to the number of instances of the general class

$$\boxed{\begin{array}{c} P(C \subseteq C') = \#(C \cap C') / \# C \\ P(p \subseteq p') = \#(p \cap p') / \# p \end{array}}$$

- To compute these probabilities, the probabilities of the existence of a sameAs link between instances are exploited.

PARIS - EXPERIMENTS

Ontology	#Instances	#Classes	#Relations
Yago	2 795 289	292 206	67
Dbpedia	2 365 777	318	1 109

Linking or mapping if the probability >0.4

Instances			Classes		Relations	
Précision	Rappel	F-Mesure	Yago ⊆ DBp Précision	DBp⊆ Yago Précision	Yago ⊆ DBp Précision	DBp⊆ Yago Précision
90%	73%	81%	-	-	100%	92%
90%	73%	81%	94%	84%	100%	92%

Instances: DBPedia and Yago uses the URIs of Wikipedia (recall and precision possible)
Classes/properties: sampling + expert
5h00 to compute the linking probabilities for instances in one iteration (2h for the classes and 20 minutes for the properties)

OAEI NOV. 2020: RESULTS

SPIMBENCH: contains

- Two ontologies TBox1 and TBox2
- Two conrresponding sets of instances

Task: Determine when two instances refer to the same **Creative Work**

MAINBOX (~1800 instances, ~50000 triples)					
System	**LogMap**	**Agrrement Maker**	**Lily**	**FTRLIM**	**REMiner**
F-measure	0,785	0,8604	0,995	0,921	**0,997**
Precision	0,880	0,838	0,990	0,855	0,998
Recall	0,709	0,883	1	0,998	0,996
TimePerformance	26782	38772	3899	2247	33966

DATA LINKING: SUMMARY

- **Knowledge-based 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.

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

- **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) [Al Bakri et al. 15].

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)