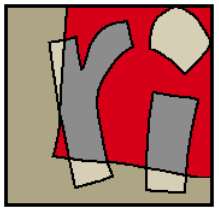
DATA LINKING AND KEY DISCOVERY IN KNOWLEDGE GRAPHS

FATIHA SAÏS

LRI, UNIVERSITÉ PARIS SUD, CNRS, UNIVERSITÉ PARIS SACLAY

MÉTASÉMINAIRE DU DÉPARTEMENT CEPIA@INRA



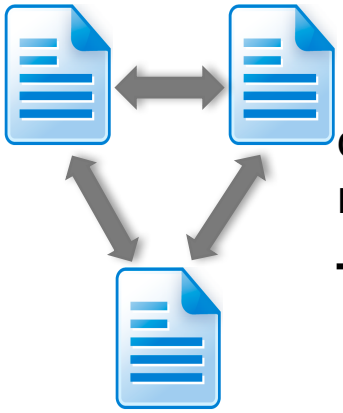
OUTLINE

- **Linked Open Data**
- **Knowledge Graphs**
- **Technical Part**
  1. Data Linking
  2. Key Discovery

- **Conclusion and Future Challenges**

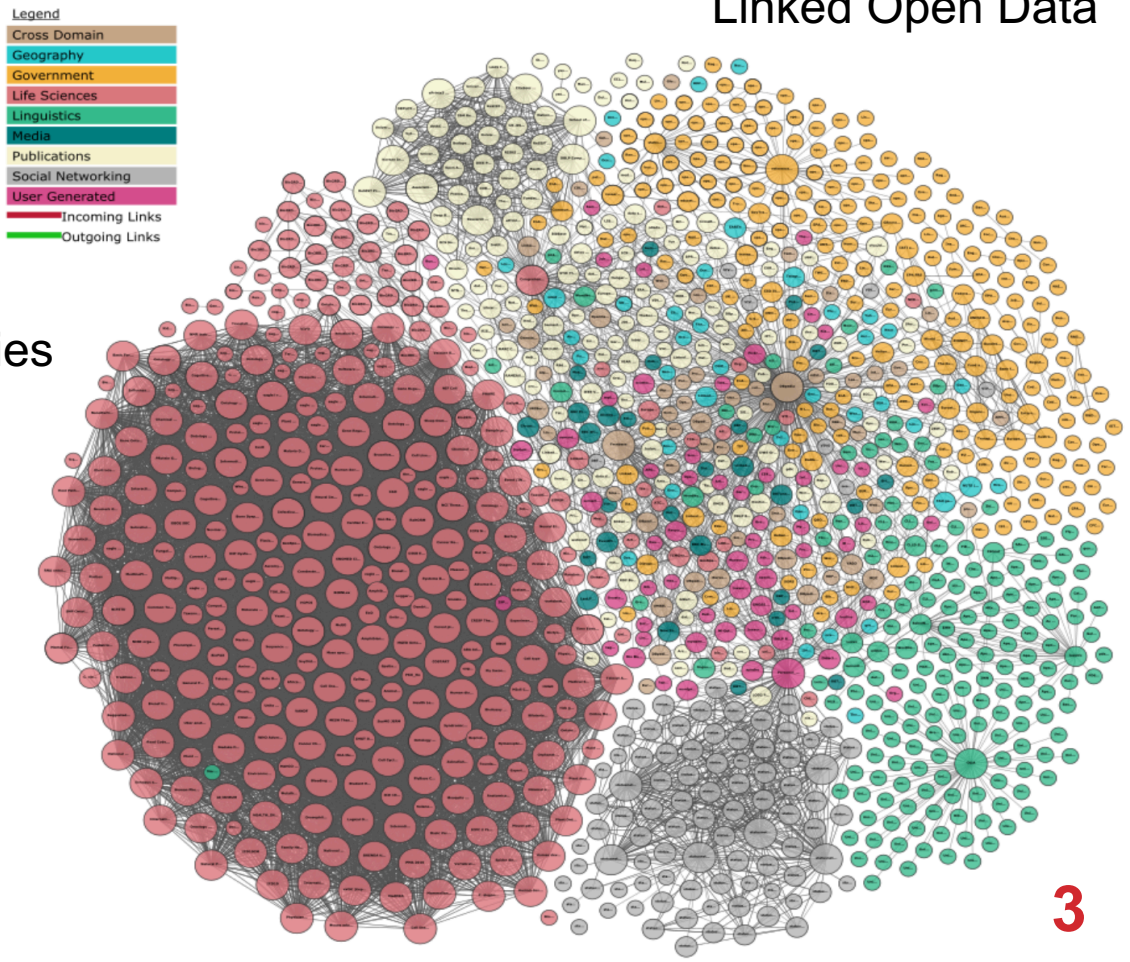
FROM THE WWW TO THE WEB OF DATA

- applying the principles of the WWW to data



data is links,
not only properties

Linked Open Data



LINKED DATA PRINCIPLES

① **Use HTTP URIs as identifiers for resources**

→ so people can look up the data

② **Provide data at the location of URIs**

→ to provide data for interested parties

③ **Include links to other resources**

→so people can discover more things

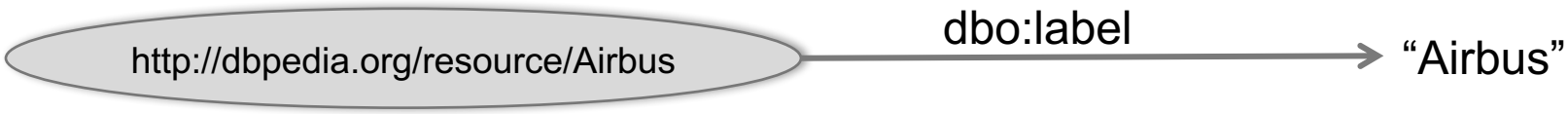
→bridging disciplines and domains

➔ Unlock the potential of island repositories

Tim Berners Lee, 2006

RDF – RESOURCE DESCRIPTION FRAMEWORK

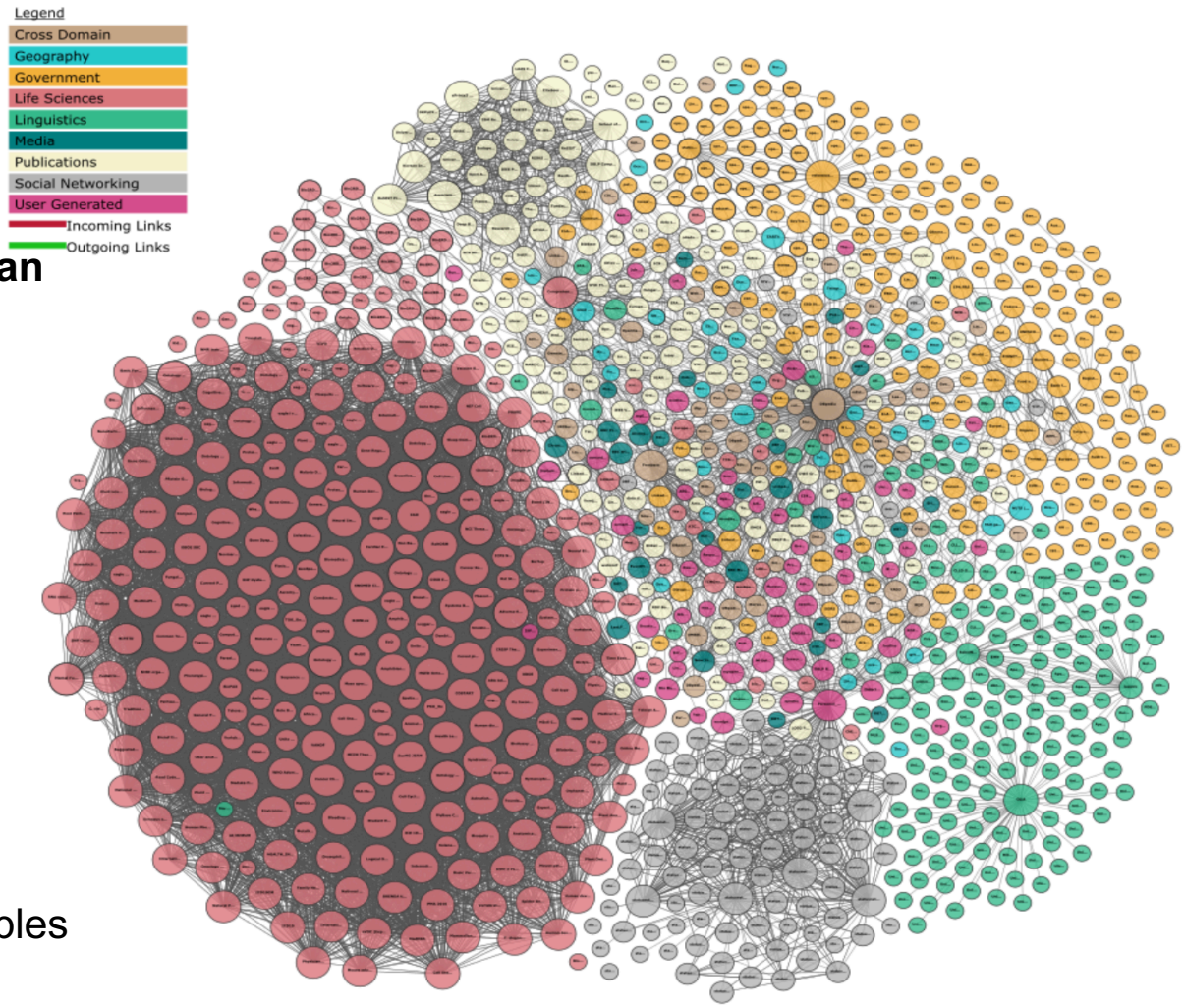
- **Statements of < subject predicate object >**



Subject Predicate Object

… is called a triple

LINKED OPEN DATA



Linked Data - Datasets under an open access

- **1,139 datasets**
- **over 100B triples**
- **about 500M links**
- **several domains**

Examples :

- DBPedia : **1.5 B** triples
- Gene Ontology : **807473** triples
- Lipid Otology : **15406** triples

"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

NEED OF KNOWLEDGE

THE ROLE OF KNOWLEDGE IN AI

[Artificial Intelligence 47 (1991)]

ON THE THRESHOLDS OF KNOWLEDGE

Douglas B. Lenat

MCC
3500 W. Balcones Center
Austin, TX 78759

Edward A. Feigenbaum

Computer Science Department
Stanford University
Stanford, CA 94305

Abstract

We articulate the three major fmdings of AI to date: (1) The Knowledge Principle: if a program is to perform a complex task well, it must know a great deal about the world in which it operates. (2) A plausible extension of that principle, called the Breadth Hypothesis: there are two additional abilities necessary for intelligent behavior in unexpected situations: falling back on increasingly general knowledge, and analogizing to specific but far-flung knowledge. (3) AI as Empirical Inquiry: we must test our ideas experimentally, on large problems. Each of these three hypotheses proposes a particular threshold to cross, which leads to a qualitative change in emergent intelligence. Together, they determine a direction for future AI research.

opponent is Castling.) Even in the case of having to search for a solution, the *method* to carry out the search may be

there is some minimum knowledge needed for one to even formulate it.

*The knowledge principle: "if a program is to perform a complex task well, **it must know a great deal about the world** in which it operates."*

ONTOLOGY, A DEFINITION

"An ontology is an **explicit**, **formal specification** of a **shared conceptualization.**"

[Thomas R. Gruber, 1993]

Conceptualization: abstract model of domain related expressions

Specification: domain related

Explicit: semantics of all expressions is clear

Formal: machine-readable

Shared: consensus (different people have different perceptions)

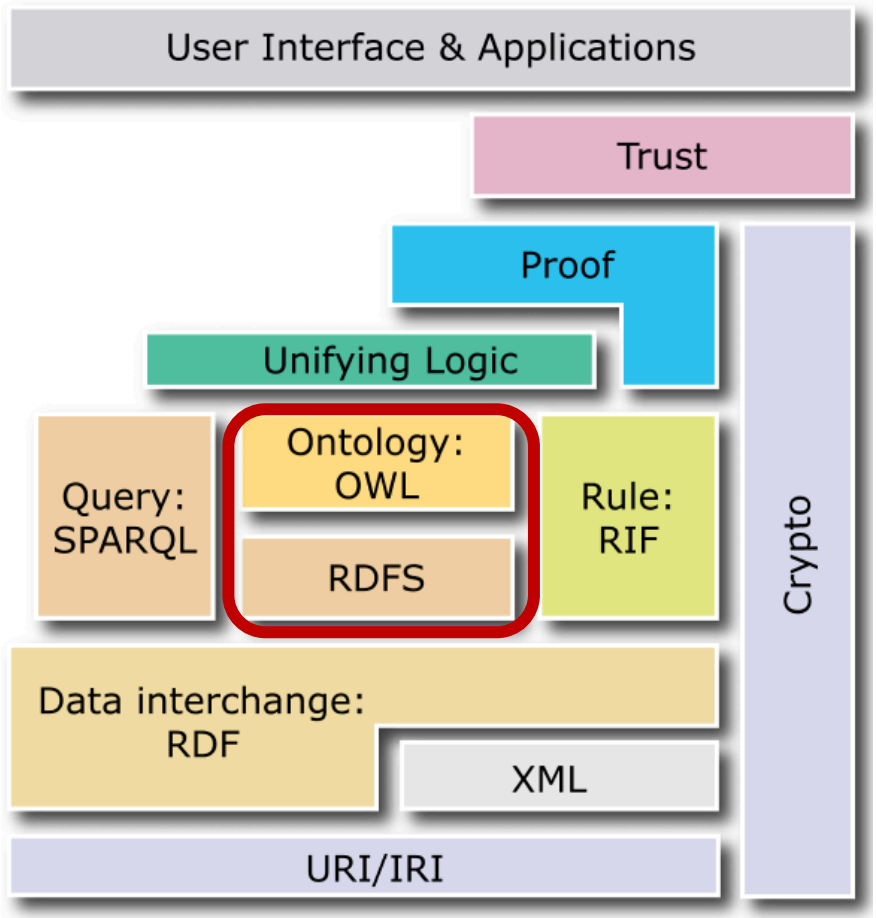
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema

- Lightweight ontologies

OWL – Web Ontology Language

- Expressive ontologies



Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png

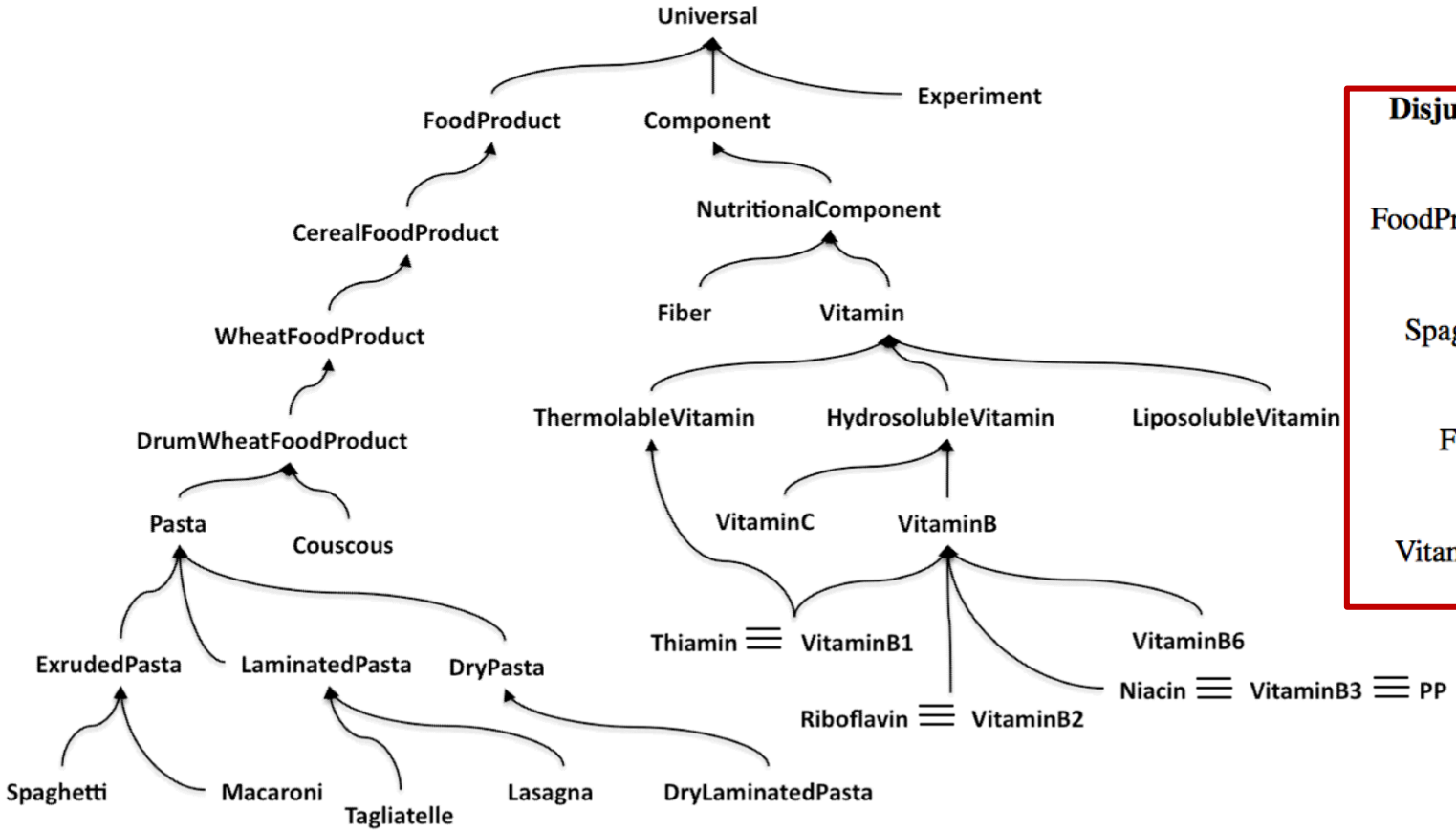
OWL ONTOLOGY

OWL – Web Ontology Language

- Represents rich and complex knowledge about things
- Based on First Order Logic (FOL)
- Can be used to verify the consistency of knowledge
- Can make implicit knowledge explicit

- **Classes:** concepts or collections of objects (individuals)
- **Properties:**
  - owl:DataTypeProperty (attribute)
  - owl:ObjectProperty (relation)
- **Hierarchy**
  - owl:subClassOf
  - owl:subPropertyOf
- **Individuals:** ground-level of the ontology (instances)

OWL ONTOLOGY



Disjunction Constraints

FoodProduct ⊥ Component

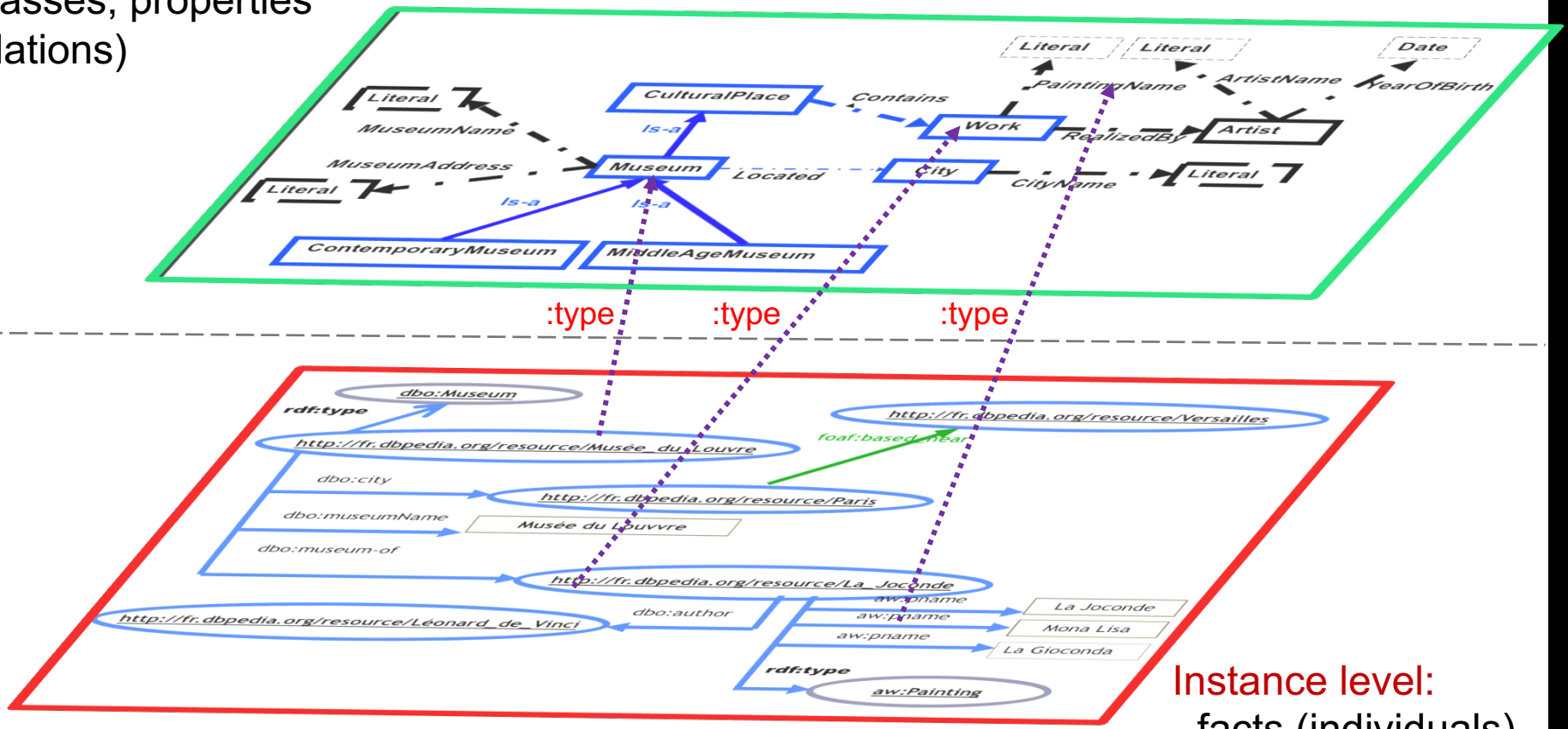
Spaghetti ⊥ Macaroni

Fiber ⊥ Vitamin

Vitamin C ⊥ Vitamin B

ONTOLOGY LEVELS

Conceptual level:
- classes, properties (relations)



Instance level:
- facts (individuals)

OWL ONTOLOGY - AXIOMS

- **Axioms:** knowledge definitions in the ontology that were **explicitly defined** and have **not been proven true**.
  - **Reasoning over an ontology**
    → Implicit knowledge can be made explicit by logical reasoning
- **Example:**

NutriComp12 is a **VitaminB1**	<NutriComp12 rdf:type VitaminB1> .
NutriComp13 is a **Fiber**	<NutriComp13 rdf:type Fiber> .
FoodProduct1 contains **NutriComp12**	<FoodProd ao:contains *NutriComp12*> .
FoodProduct1 contains **NutriComp13**	<FoodProd ao:contains *NutriComp13*> .

- **Infer that:**

➔ **NutriComp12** is a Vitamin because VitaminB1 < Vitamin

➔ **NutriComp12** is a Hydrosoluble Vitamin because Vitamin < HydrosolubleVitamin

➔ **NutriComp12** is a NurtitionalComponent
because HydrosolubleVitamin < NurtitionalComponent

➔ **NutriComp13** is-differentFrom **NutriComp12** because Vitamin != Fiber

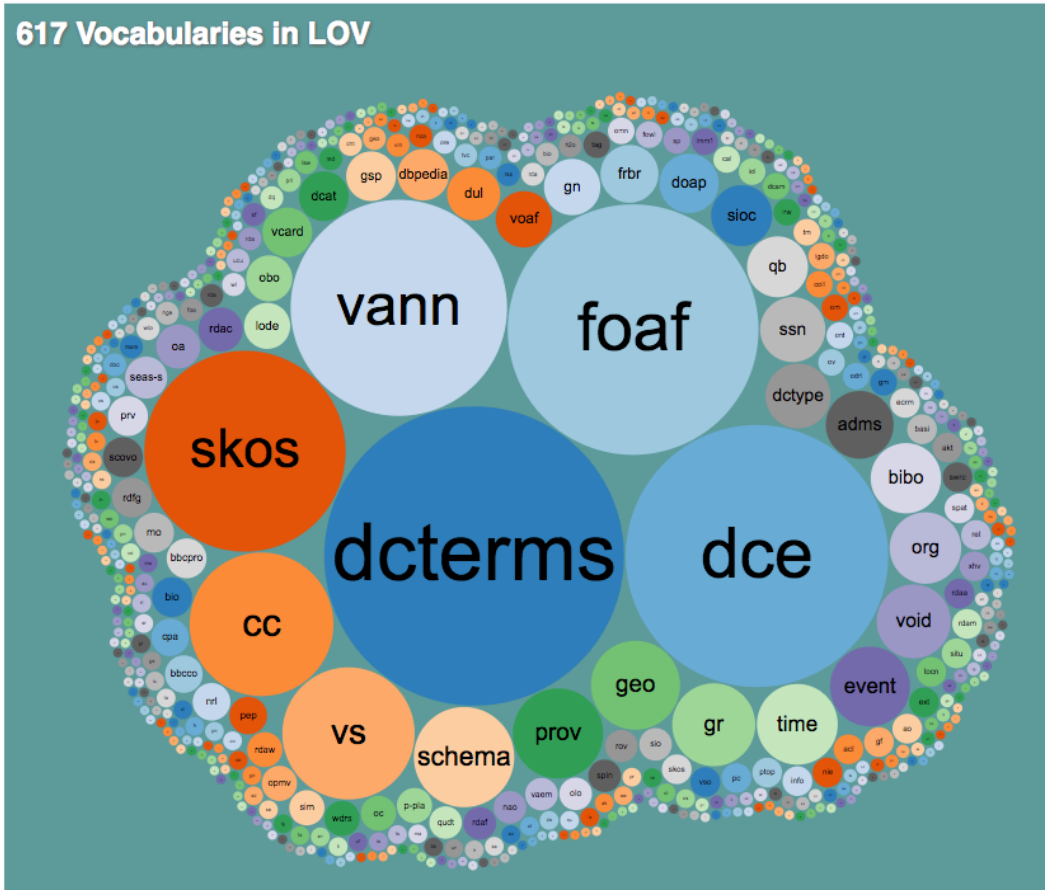
LINKED OPEN VOCABULARIES (LOV)

Linked Open Vocabularies

- Keeps track of available open ontologies and provides them as a graph
- Search for available ontologies, open for reuse

- Example:

http://lov.okfn.org/dataset/lov/vocabs/foaf

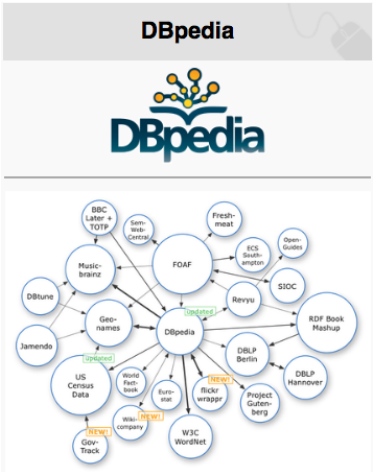


OUTLINE

- **Linked Open Data**
- **Knowledge Graphs**
- **Technical Part**
  - Data Linking
  - Key Discovery
- **Conclusion and Future Challenges**

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007



2008

yago
select knowledge

2012

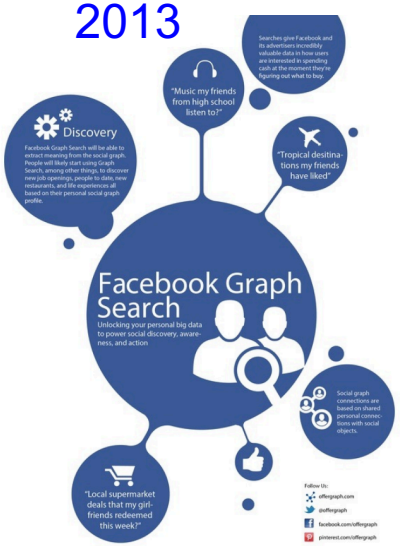
WIKIDATA

2007

Freebase™

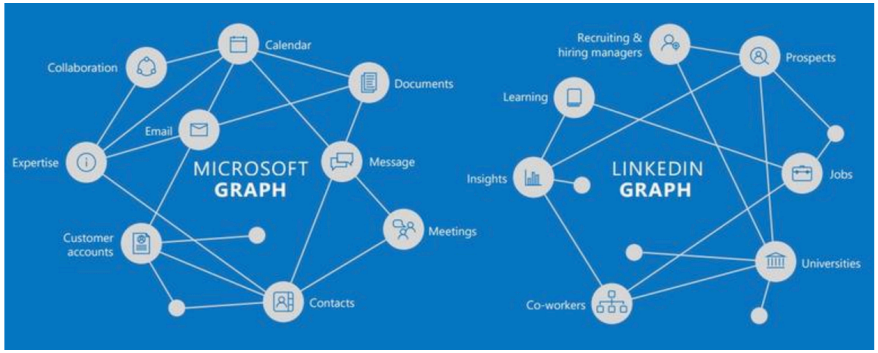
2012

2013



2015

2016



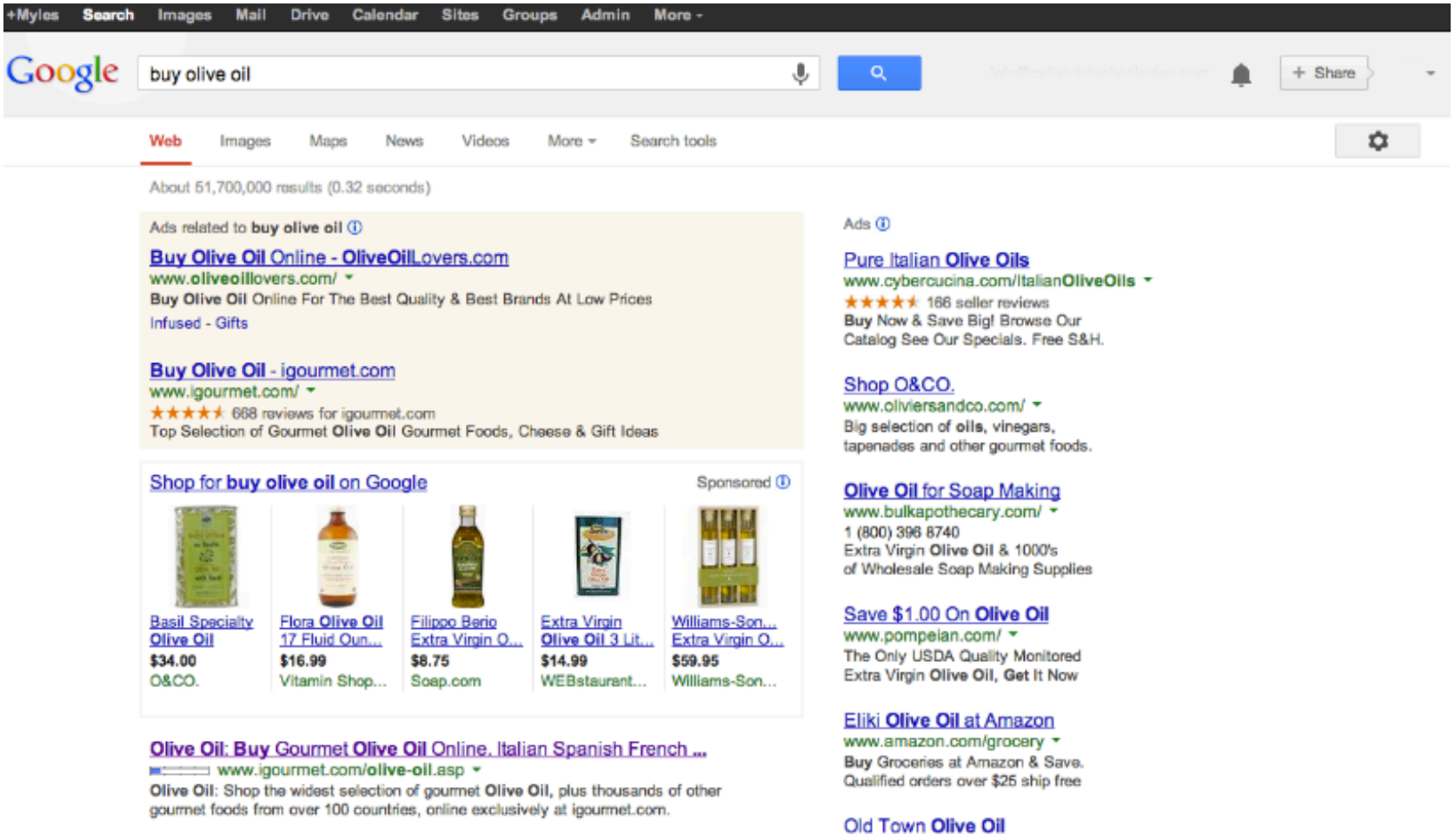
2013

Yahoo's new SERP designs mobile and knowledge graph

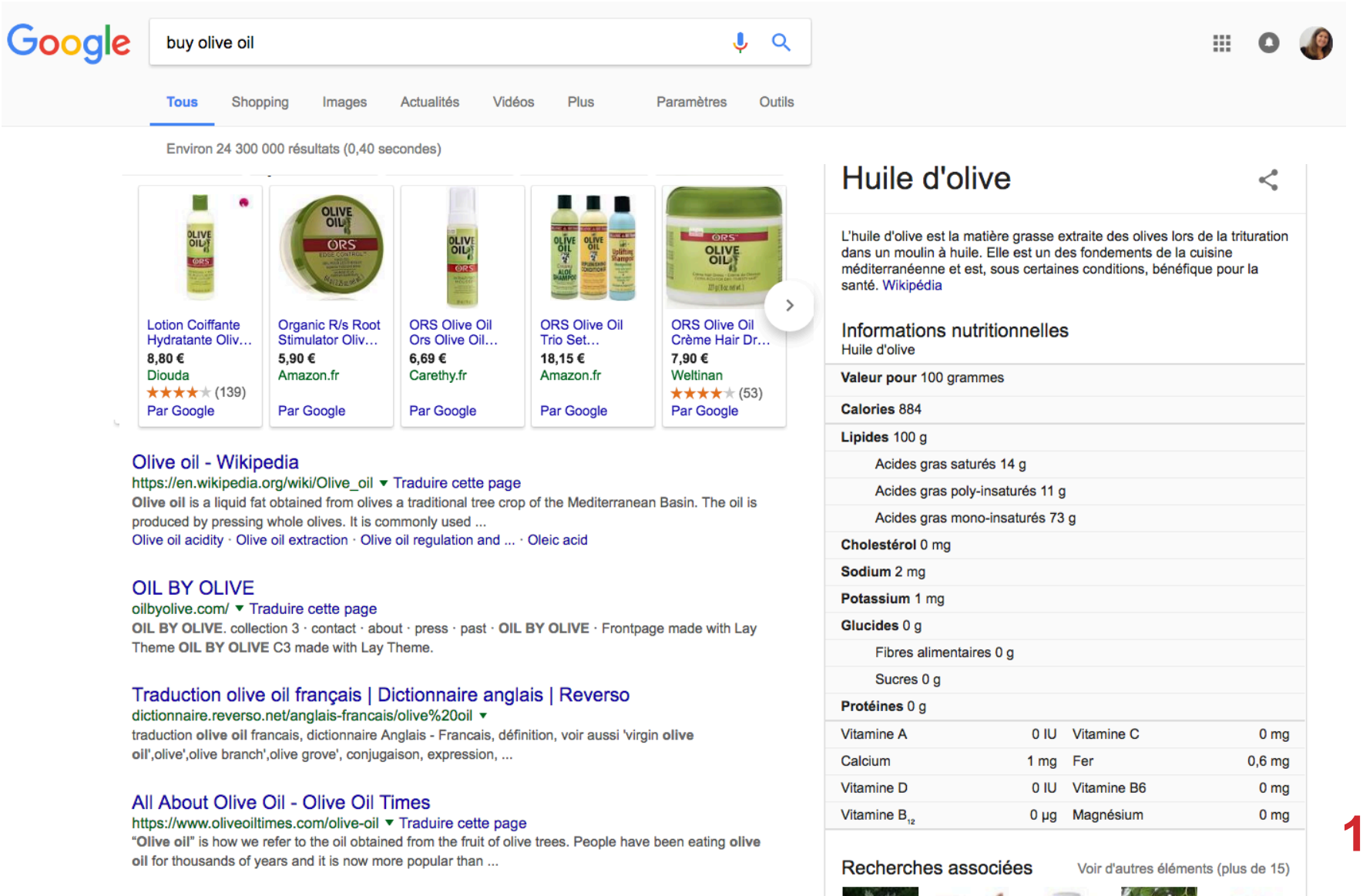
Academic side

Commercial side

WEB SEARCH WITHOUT KNOWLEDGE GRAPHS

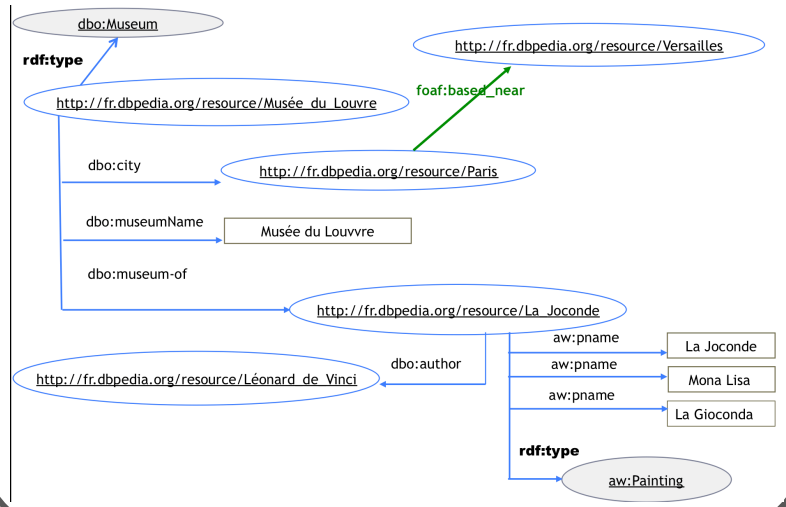


WEB SEARCH WITH KNOWLEDGE GRAPHS

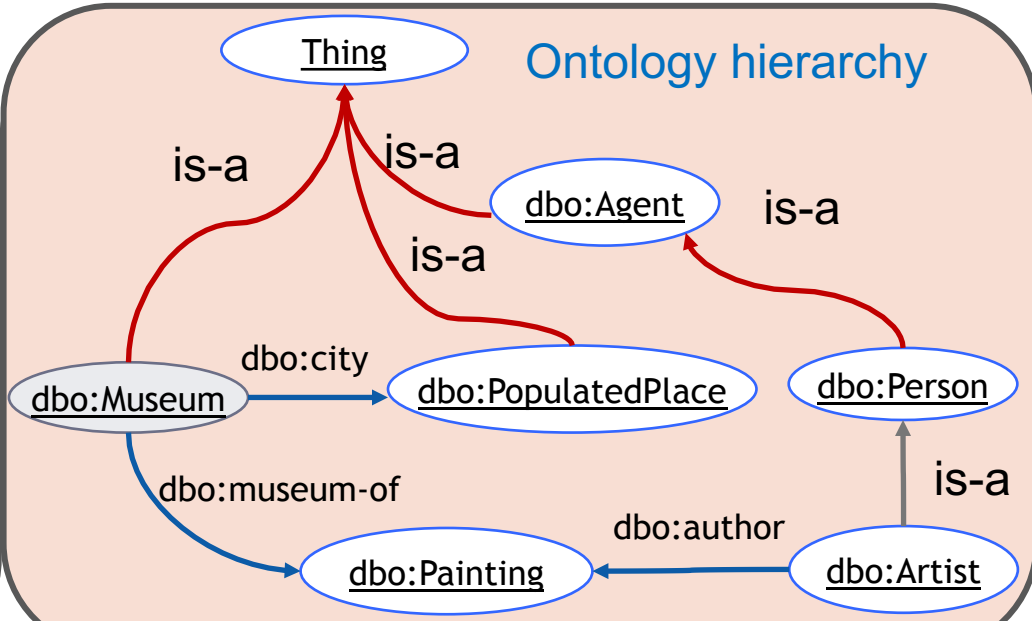


KNOWLEDGE GRAPH (KG)

RDF Graphs



Ontology hierarchy



Querying (SPARQL)

```
PREFIX dbo: <http://dbpedia.org/ontology#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?m ?p
WHERE { ?m rdf:type dbo:Museum . ?m dbo:musuem-of ?p .}
```

Reasoners: (Pellet, Fact++, Hermit, etc.)

- KG saturation: infer whatever can be inferred from the KG.
- KG consistency checking: no contradictions
- KG repairing
- …

Ontology axioms and rules

owl:equivalentClass(dbo:Municipality, dbo:Place)
owl:equivalentClass(dbo:Place, dbo:Wikidata:Q532)
owl:equivalentClass(dbo:Village, dbo:PopulatedPlace)
owl:equivalentClass(dbo:PopulatedPlace,dbo:Municipality)
owl:disjointClass(dbo:PopulatedPlace, dbo:Artist)
owl:disjointClass(dbo:PopulatedPlace, dbo:Painting)
owl:FunctionalProperty(dbo:city)
owl:InverseFunctionalProperty(dbo:museum-of)

dbo:birthPlace(X, Y) => dbo:citizsenOf(X, Y)
dbo:parentOf(X, Y) => dbo:child(Y, X)

KNOWLEDGE GRAPH EXPANSION AND ENRICHMENT

- **Expansion: knowledge graphs are incomplete**
  - **Data linking (entity resolution, duplicate detection, reference reconciliation)**
  - Link prediction: add relations
  - Ontology matching: connect graphs
  - Missing values prediction/inference
- **Enrichment: can new knowledge emerge from knowledge graphs?**
  - **Knowledge discovery**
  - Automatic reasoning and planning

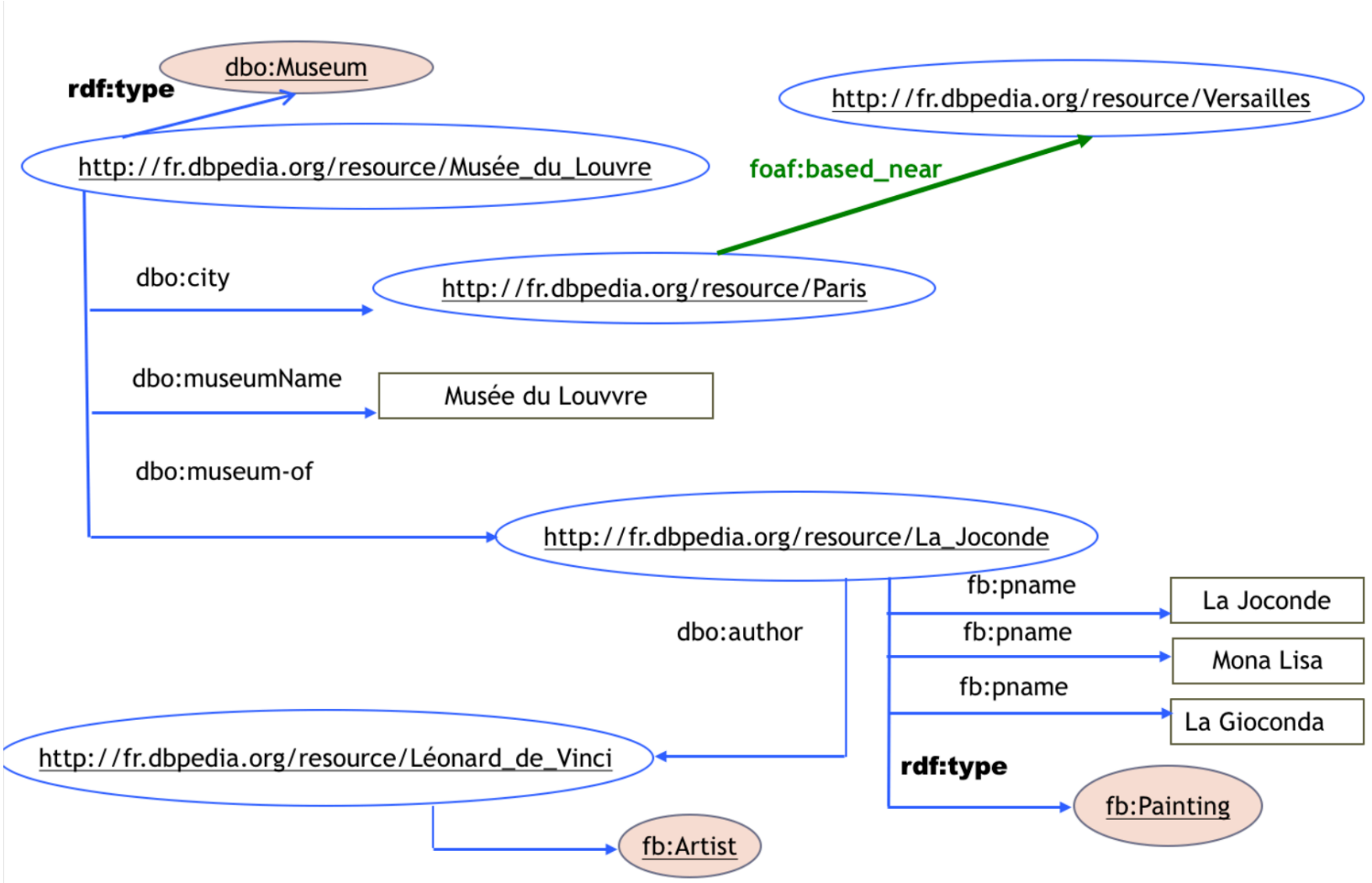
OUTLINE

- **Linked Open Data**
- **Knowledge Graphs**
- **Technical Part**
  1. Data Linking
  2. Key Discovery
- **Conclusion and Future Challenges**

1. DATA LINKING

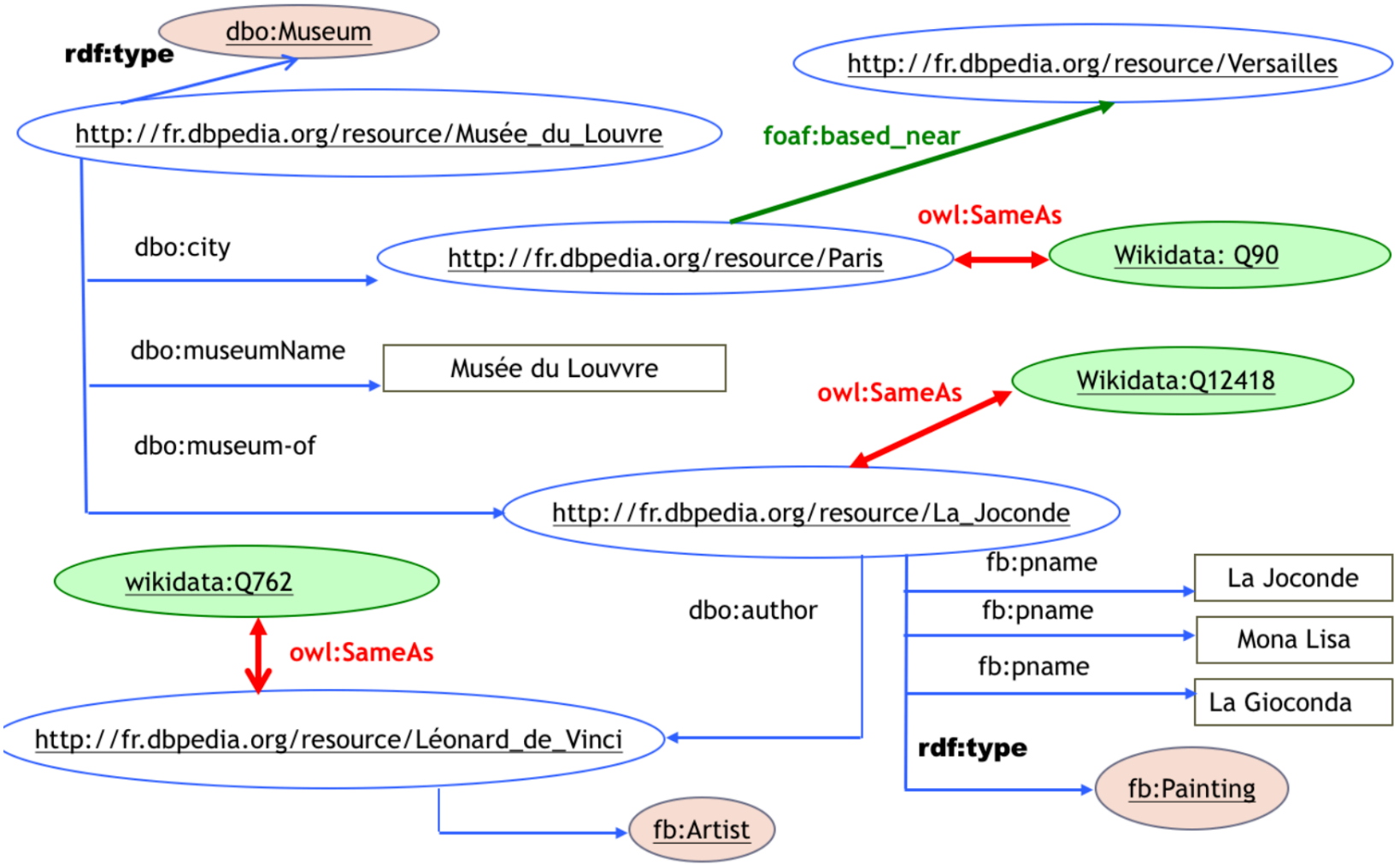
DATA LINKING

- **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).



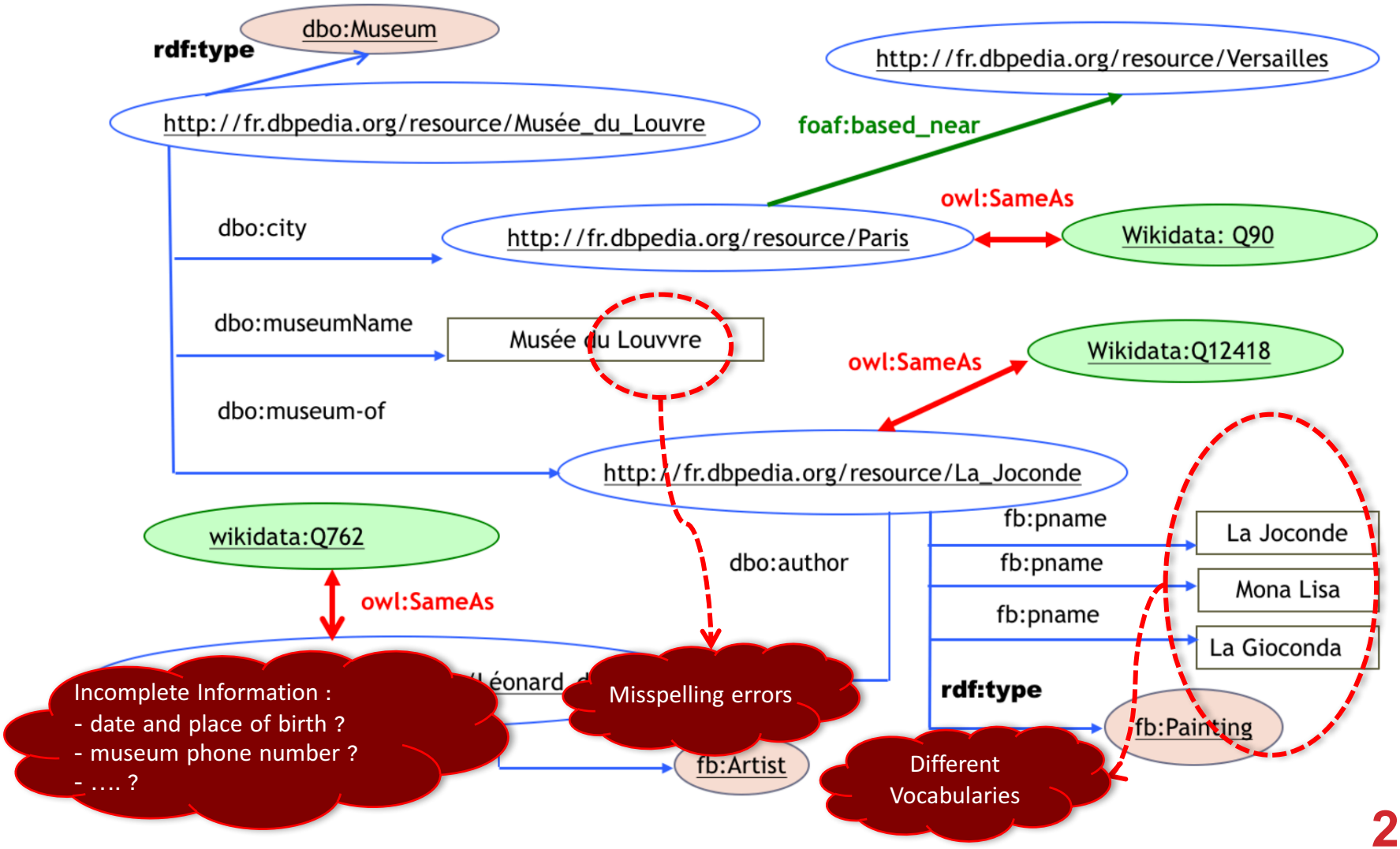
DATA LINKING

- **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** consists in 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 = {(u1,u2) ∈ u1 × u2: owl:sameAs(s,t)} and
    - D = {(u1,u2) ∈ u1 × u2: owl:differentFrom(s,t)}

- A method is said **total** when $(S \cup D) = (U_1 \times U_2)$
- A method is said **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, ...*

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.

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

cord and 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 link

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

DATA LINKING APPROACHES: DIFFERENT CONTEXTS

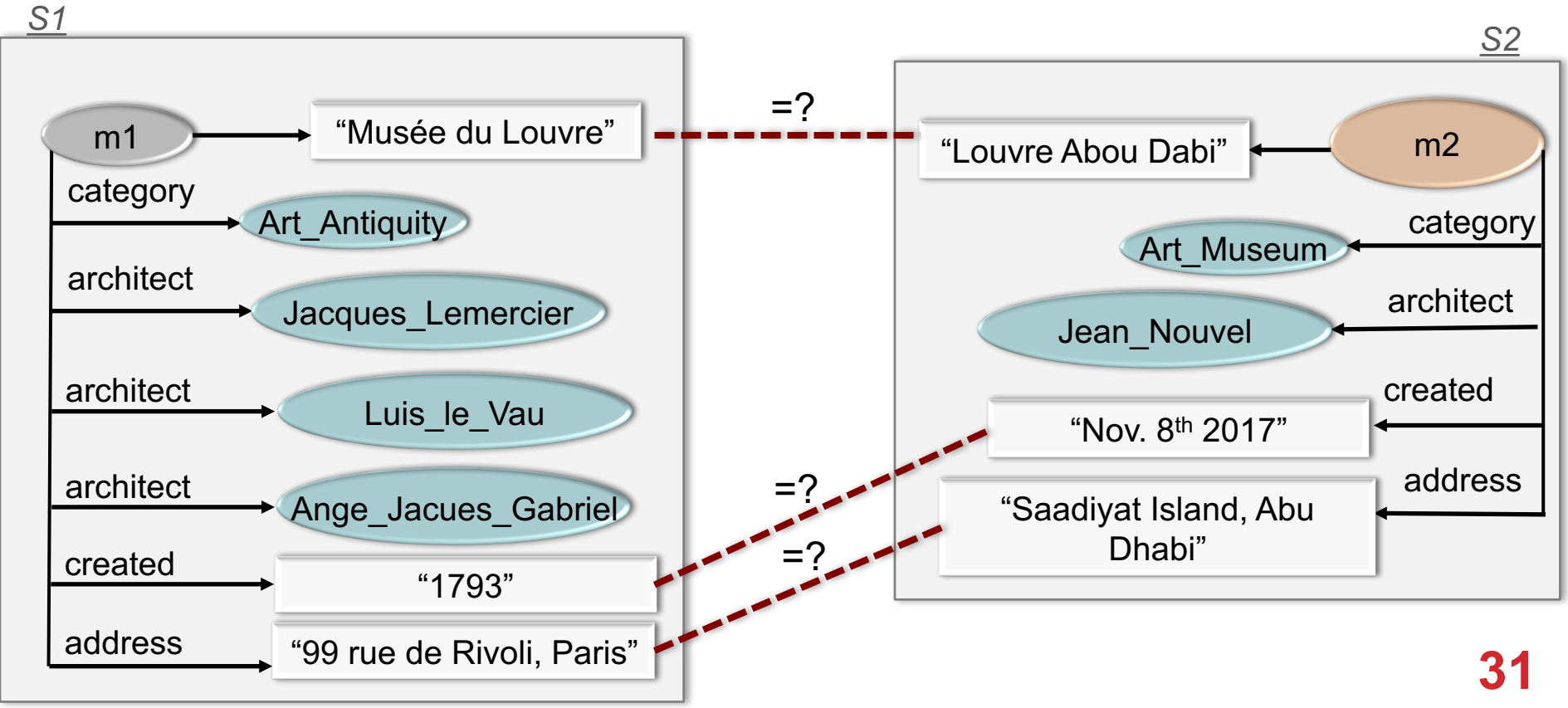
- Datasets conforming to the same ontology
- Datasets conforming to different ontologies
- Datasets without ontologies

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

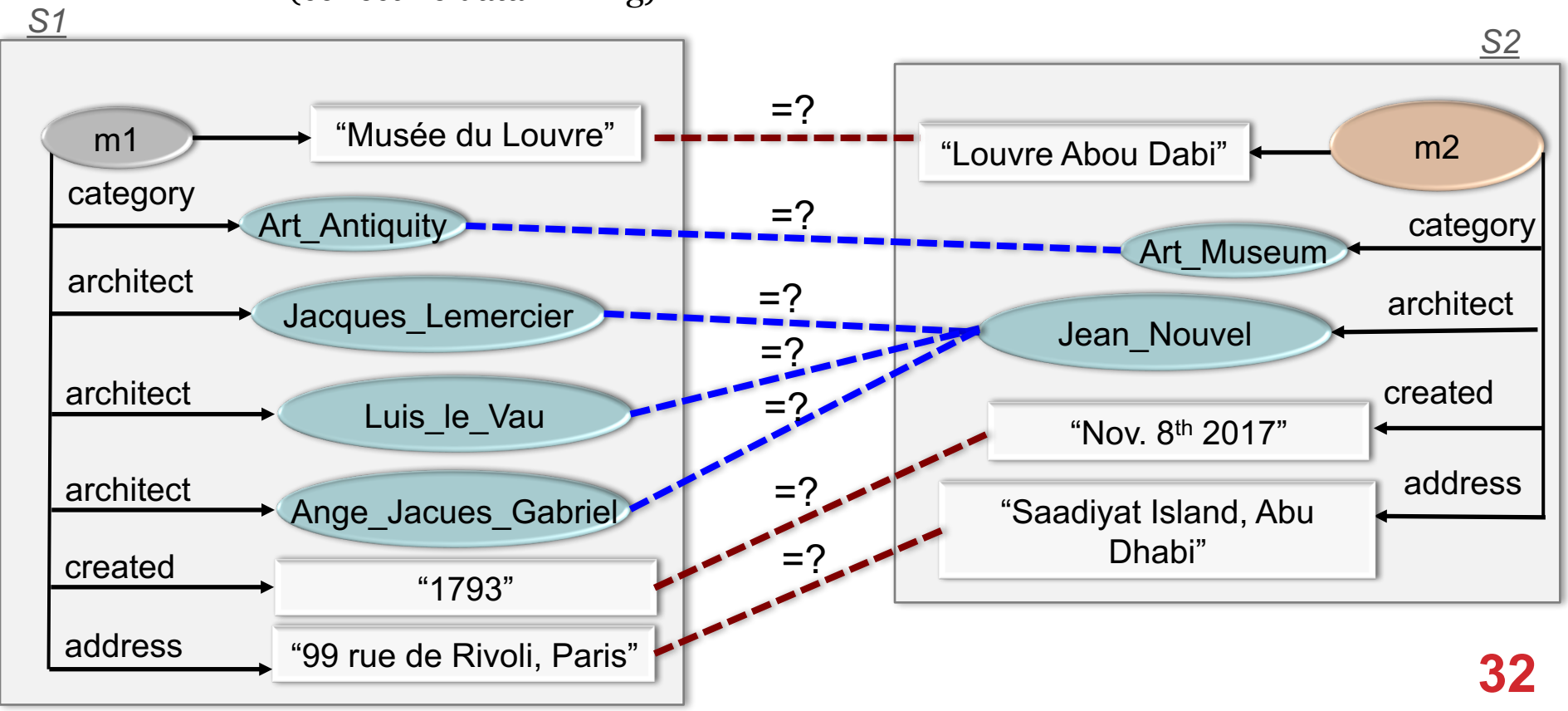
DATA LINKING APPROACHES

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



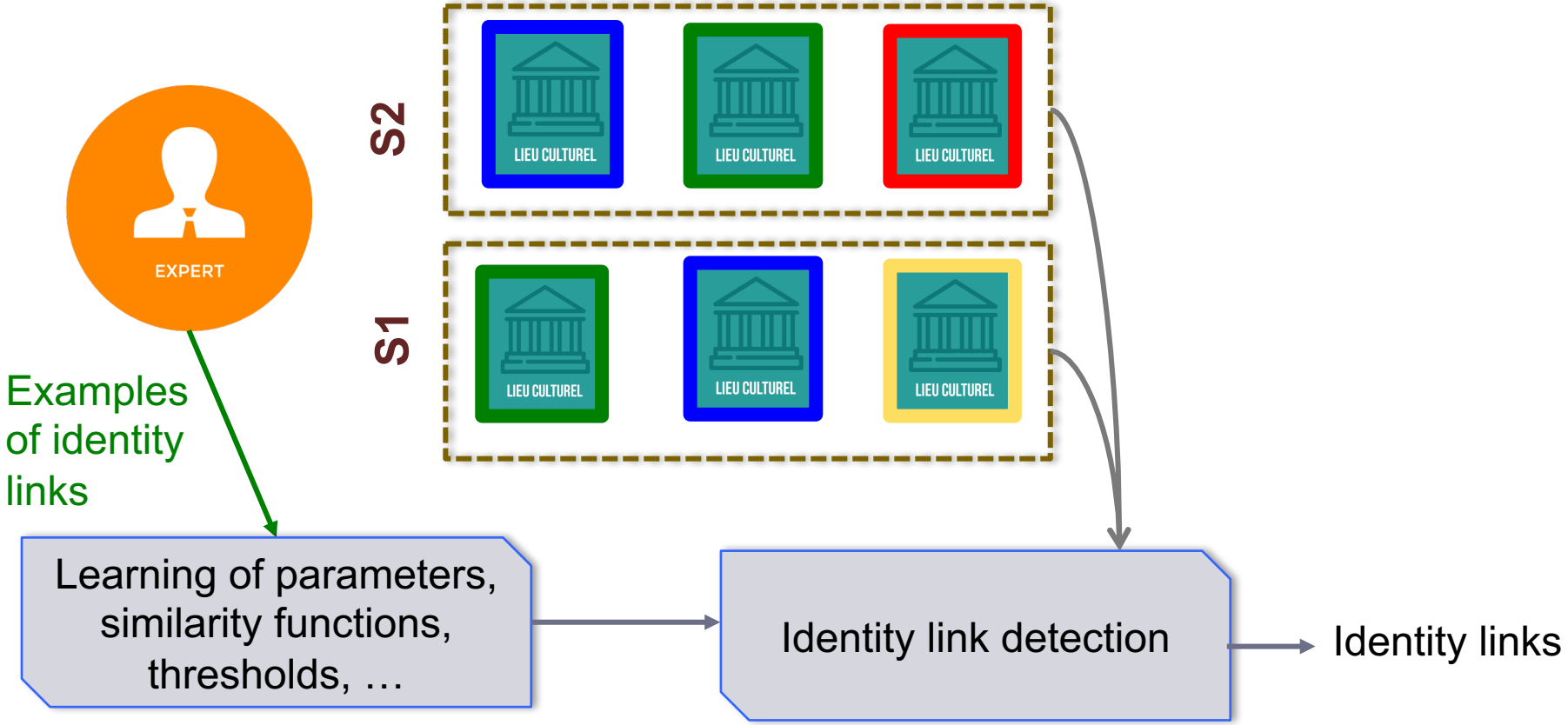
DATA LINKING APPROACHES

- **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		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**

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

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.

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)

[Saïs et al' 07, Saïs et al'09]

Ontology axioms

- Disjunction axioms between classes, DISJOINT(C, D)
- Functional properties axioms, PF(P)
- Inverse functional properties axioms, PFI(P)
- Key axioms, Key(P1, P2, P3, …)

Assumptions on the data

- Unique Name Assumption, UNA(src1)

N2R: A NUMERICAL METHOD FOR REFERENCE RECONCILIATION

[Saïs et al'09]

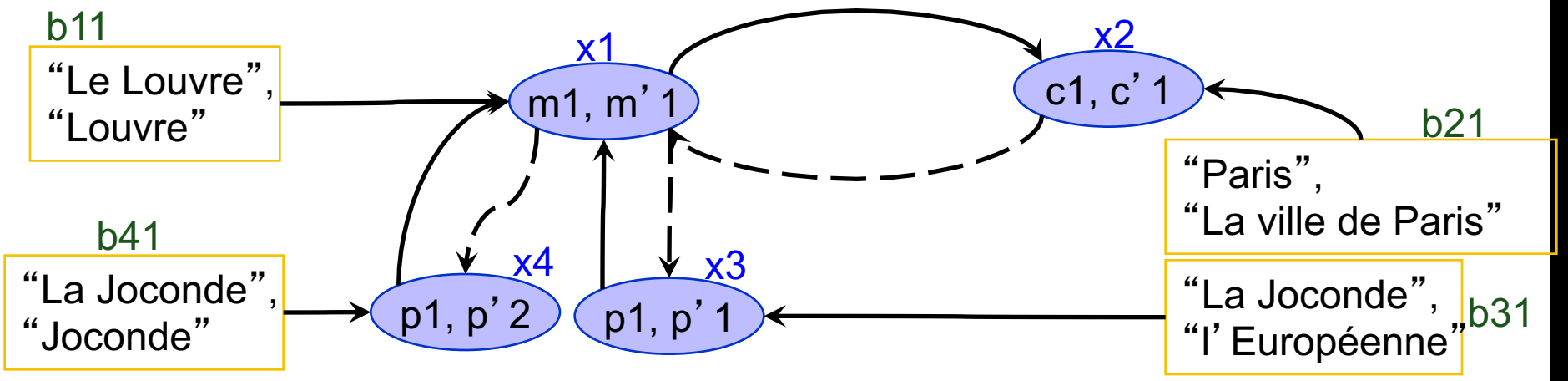
N2R: A NUMERICAL METHOD FOR REFERENCE RECONCILIATION

[Saïs et al'09]

- **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: ILLUSTRATION

[Saïs et al'09]



	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

x1 = **max(max(b11, x3), x4), λ * x2)**

x2 = **max(b21, x1)**

x3 = **max(b31, λ* x1)**

x4 = **max(b41 , λ * x1)**

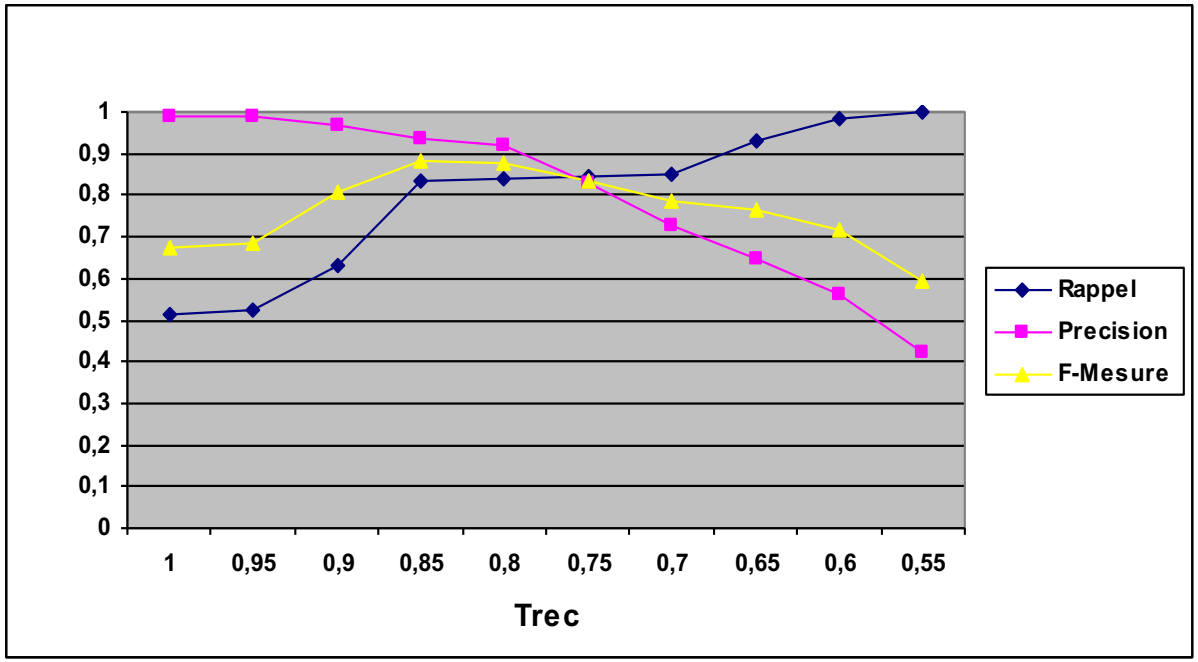
λ= 1/(| CAttr | + | CRel |) ε = 0.02

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

N2R EXPERIMENTS

N2R: RESULTS ON CORA [Saïs et al'09]



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 %.

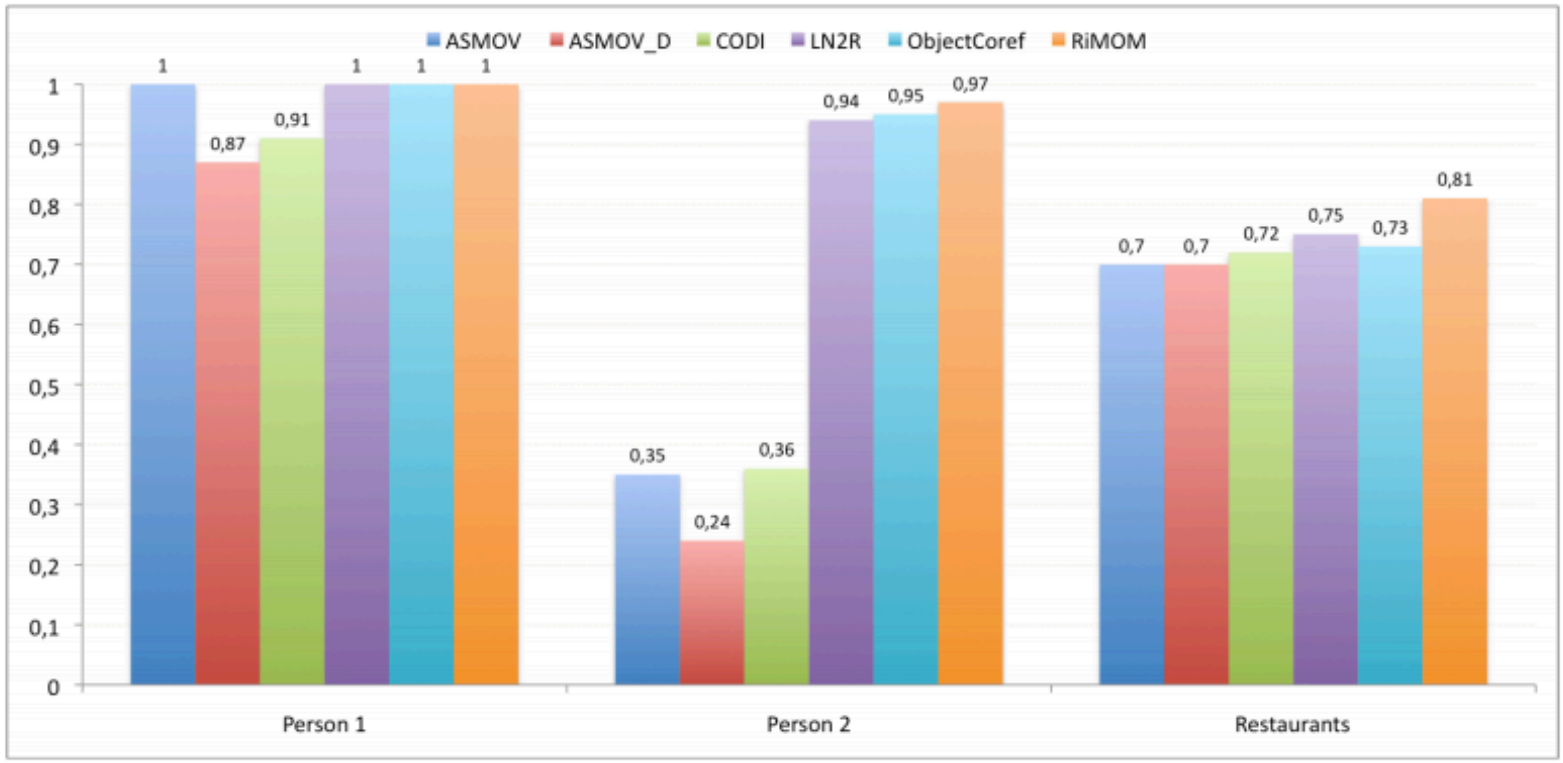
Trec = 0.85, the F-measure is of 88 %:

- Better than the results obtained by the supervised method of [Singla and Domingos'05]
- Worst than those (97 %) obtained by the supervised method of [Dong et al.'05]

N2R: RESULTS IN OAEI[2] 2010

[Saïs et al'09]

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



[2] Ontology Alignment Evaluation Initiative

OUTLINE

- **Linked Open Data**
- **Knowledge Graphs**
- **Technical Part**
  1. Data Linking
  2. Key Discovery
- **Conclusion and Future Challenges**

KEY DISCOVERY

[Qualinca ANR project (2012-2016)]

PHD OF DANAI SYMEONIDOU IN COLLABORATION WITH:

NATHALIE PERNELLE (LRI, UNIV. PARIS SUD),

DATA LINKING USING RULES

homepage(w1, y) ∧ homepage(w2, y) ➔ **sameAs(w1, w2)**

	…	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 USING RULES

homepage(w1, y) ∧ homepage(w2, y) ➔ 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**

DATA LINKING USING RULES

Rules

- **Logical Rules**
  - Ex. For instances of the class Restaurant

    homepage(w1, y) ∧ homepage(w2, y) ➔ sameAs(w1, w2)

    ↓

    {homepage}: discriminative property

- **Complex Rules**
  - Ex. For instances of the class Restaurant

    max(Jaccard(lat(w1, n);lat(w2, m));jaro(long(w1, x);long(w2, y)) > 0.8
    ➔sameAs(w1, w2)

    ↓

    {lat, long}: discriminative property set

Rules contain discriminative properties => keys

KEYS

Key: A set of properties that uniquely identifies every instance in the data

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor
Person2	Marie	Tompson	02/09/75	Researcher
Person3	Marie	David	15/02/85	Teacher
Person4	Vincent	Solgar	06/12/90	Teacher

Is [FirstName] a key? ✖

Is [LastName] a key? ✖

KEYS

Key: A set of properties that uniquely identifies every instance in the data

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor
Person2	Marie	Tompson	02/09/75	Researcher
Person3	Marie	David	15/02/85	Teacher
Person4	Vincent	Solgar	06/12/90	Teacher

Is [FirstName] a key? ✖

Is [LastName] a key? ✖

Is [FirstName,LastName] a key? ✔

KEYS - KEY MONOTONICITY

Key monotonicity: When a set of properties is a key, all its supersets are also keys

Minimal Key: A *key* that by removing one property stops being a key

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor
Person2	Marie	Tompson	02/09/75	Researcher
Person3	Marie	David	15/02/85	Teacher
Person4	Vincent	Solgar	06/12/90	Teacher

Is [FirstName,LastName, Birthday] a key? ✔

*Is [FirstName,LastName, Birthday] a **minimal** key?* ✖

***Minimal keys**: [[FirstName,LastName],[FirstName,Profession] [Birthdate], [LastName,Profession]]*

KEYS DECLARED BY EXPERTS FOR DATA LINKING

Not an easy task:

- Experts are not aware of all the keys

  Ex. {SSN}, {ISBN} easy to declare
  Ex. {Name, DateOfBirth, BornIn} **is it a key for the class Person?**

- Erroneous keys can be given by experts

- As many keys as possible
  - More keys => More linking rules

Goal: Discover keys automatically

OWL2 KEY, KEY IN THE OPEN WORLD

OWL2 Key for a class: a combination of properties that uniquely identify each instance of a class

- 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:

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

SAKEY

[Symeonidou et al. 14]

SAKey: Scalable Almost Key discovery approach for:

- Incomplete data (optimistic heuristic)
- Errors
- Duplicates
- Large datasets

Discovers *almost keys*

- Sets of properties that are not keys due to few exceptions

ALMOST KEY DISCOVERY STRATEGY

- **Naive automatic way to discover keys**
  - Examine all the possible combinations of properties
  - Scan all instances for each candidate key

    - **Example**: Class described by 15 properties ➔2^{15} = 32767 candidate keys

- **Discover keys efficiently by:**
  - Reducing the combinations
  - Partially scanning the data

KEY DISCOVERY

SAKey: Key discovery approach for very large RDF datasets that may contain erroneous data or duplicates

Producer is a key?

	Region	**Producer**	**Colour**
Wine1	Bordeaux	Dupont	White
Wine2	Bordeaux	Baudin	Rose
Wine3	Languedoc	Dupont	Red
Wine4	Languedoc	Faure	Red

KEY DISCOVERY

SAKey: Key discovery approach for very large RDF datasets that may contain erroneous data or duplicates

Region is a non key?

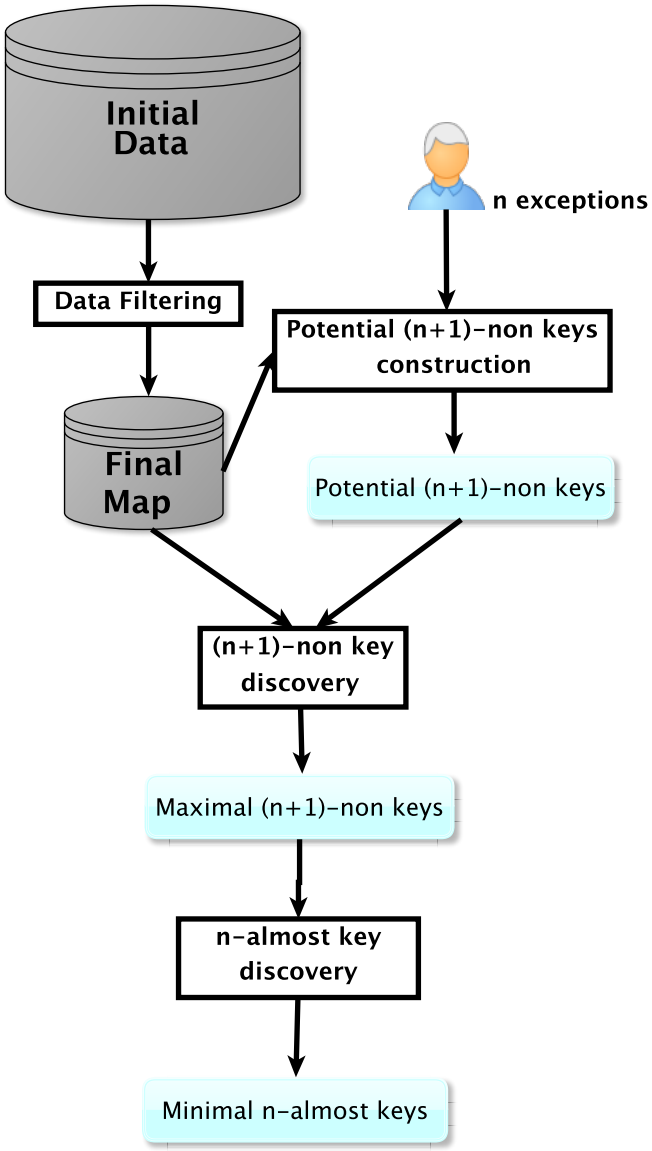
	Region	Producer	Colour
Wine1	Bordeaux	Dupont	White
Wine2	Bordeaux	Baudin	Rose
Wine3	Languedoc	Dupont	Red
Wine4	Languedoc	Faure	Red

- Interested only in maximal non keys:
- All the sets of properties that are not maximal non keys are keys
- **Example:** class described by the properties p1, p2, p3, p4

Maximal non key = {{p1, p2}} ➔ keys = {{p3}, {p4}}

SAKEY–GENERAL ARCHITECTURE

[Symeonidou et al. 14]



KEY DISCOVERY

[Symeonidou et al. 14]

SAKey: Key discovery approach for very large RDF datasets that may contain erroneous data or duplicates

- ***n*-almost key:** A set of properties for which at most *n* instances are identical values

	Region	**Producer**	**Colour**
Wine1	Bordeaux	Dupont	White
Wine2	Bordeaux	Baudin	Rose
Wine3	Languedoc	Dupont	Red
Wine4	Languedoc	Faure	Red

- Examples of keys
  {Region, Producer}:
  0-almost key

KEY DISCOVERY

[Symeonidou et al. 14]

SAKey: Key discovery approach for very large RDF datasets that may contain erroneous data or duplicates

- ***n*-almost key:** A set of properties for which at most *n* instances are identical values

	Region	Producer	Colour
Wine1	Bordeaux	**Dupont**	White
Wine2	Bordeaux	Baudin	Rose
Wine3	Languedoc	**Dupont**	Red
Wine4	Languedoc	Faure	Red

- Examples of keys

  {Region, Producer}:
  0-almost key

  {Producer}:
  2-almost key

Tool available in https://www.lri.fr/sakey

N-ALMOST KEYS

[Symeonidou et al. 14]

Exception of a key: an instance that shares values with another instance for a given set of properties *P*

Films	HasName	HasActor	HasDirector	ReleaseDate	HasWebsite	HasLanguage
f1	"Ocean's 11"	"B. Pitt" "J. Roberts"	"S. Soderbergh"	"3/4/01"	www.oceans11.com	---
f2	"Ocean's 12"	"B. Pitt" "G. Clooney" "J. Roberts"	"S. Soderbergh" "R. Howard"	"2/5/04"	www.oceans12.com	---
f3	"Ocean's 13"	"B. Pitt" "G. Clooney"	"S. Soderbergh" "R. Howard"	"30/6/07"	www.oceans13.com	---
f4	"The descendants"	"N. Krause" "G. Clooney"	"A. Payne"	"15/9/11"	www.descendants.com	"english"
f5	"Bourne Identity"	"D. Liman"	---	"12/6/12"	www.bourneIdentity.com	"english"
f6	"Ocean's 12"	---	"R. Howard"	"2/5/04"	---	---

N-ALMOST KEYS

[Symeonidou et al. 14]

Exception of a key: an instance that shares values with another instance for a given set of properties *P*

- *f1, f2 and f3* are three exceptions for the property set {HasActor}

Films	HasName	HasActor	HasDirector	ReleaseDate	HasWebsite	HasLanguage
f1	"Ocean's 11"	**"B. Pitt"** "J. Roberts"	"S. Soderbergh"	"3/4/01"	www.oceans11.com	---
f2	"Ocean's 12"	**"B. Pitt"** "G. Clooney" "J. Roberts"	"S. Soderbergh" "R. Howard"	"2/5/04"	www.oceans12.com	---
f3	"Ocean's 13"	**"B. Pitt"** "G. Clooney"	"S. Soderbergh" "R. Howard"	"30/6/07"	www.oceans13.com	---
f4	"The descendants"	"N. Krause" "G. Clooney"	"A. Payne"	"15/9/11"	www.descendants.com	"english"
f5	"Bourne Identity"	"D. Liman"	---	"12/6/12"	www.bourneIdentity.com	"english"
f6	"Ocean's 12"	---	"R. Howard"	"2/5/04"	---	---

N-ALMOST KEYS

[Symeonidou et al. 14]

Exception of a key: an instance that shares values with another instance for a given set of properties *P*

- *f1, f2 and f3* are three exceptions for the property set {HasActor}

Exception Set E_P: set of exceptions for *P*

- E_P = {f1, f2, f3} ∪ {f2, f3, f4} = {f1, f2, f3, f4} for {HasActor}

Films	HasName	HasActor	HasDirector	ReleaseDate	HasWebsite	HasLanguage
f1	"Ocean's 11"	**"B. Pitt"** "J. Roberts"	"S. Soderbergh"	"3/4/01"	www.oceans11.com	---
f2	"Ocean's 12"	**"B. Pitt"** **"G. Clooney"** "J. Roberts"	"S. Soderbergh" "R. Howard"	"2/5/04"	www.oceans12.com	---
f3	"Ocean's 13"	**"B. Pitt"** **"G. Clooney"**	"S. Soderbergh" "R. Howard"	"30/6/07"	www.oceans13.com	---
f4	"The descendants"	"N. Krause" **"G. Clooney"**	"A. Payne"	"15/9/11"	www.descendants.com	"english"
f5	"Bourne Identity"	"D. Liman"	---	"12/6/12"	www.bourneIdentity.com	"english"
f6	"Ocean's 12"	---	"R. Howard"	"2/5/04"	---	---

N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

- {HasActor} is a 4-almost key

N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

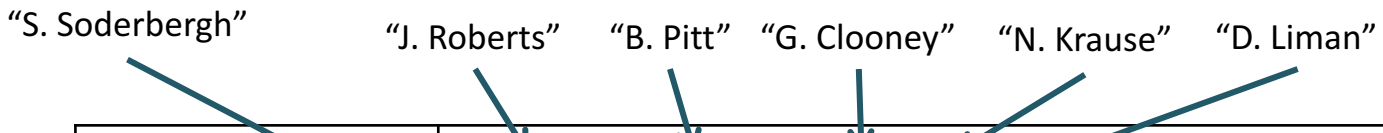
- {HasActor} is a 4-almost key

***n*-non key: a set of properties where $|E_P| \geq n$**

- Using all the maximal *n*-non keys we can derive all the minimal (*n*-1)-almost keys

N-NON KEY DISCOVERY: INITIAL MAP

[Symeonidou et al. 14]

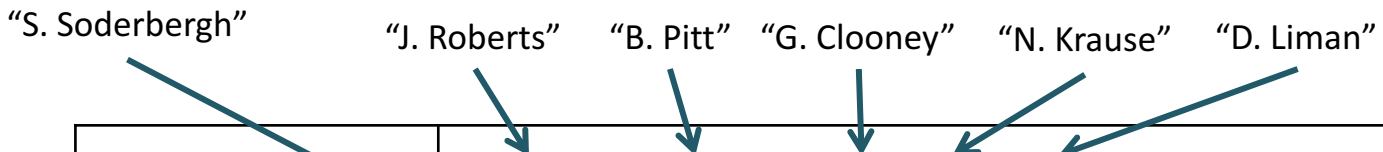


HasActor	{{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}, {f4}}
ReleaseDate	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasName	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasLanguage	{{f4, f5}}
HasWebsite	{{f1}, {f2}, {f3}, {f4}, {f5}, {f6}}

N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

Singleton filtering

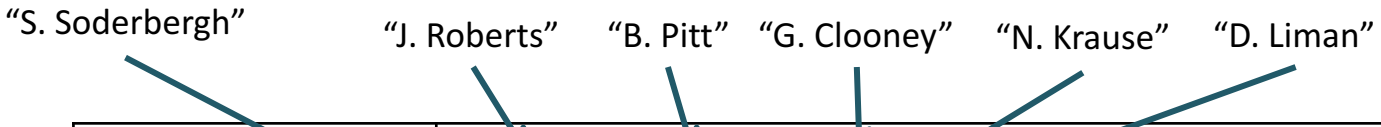


HasActor	{{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}, {f4}}
ReleaseDate	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasName	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasLanguage	{{f4, f5}}
HasWebsite	{{f1}, {f2}, {f3}, {f4}, {f5}, {f6}}

N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

Singleton filtering

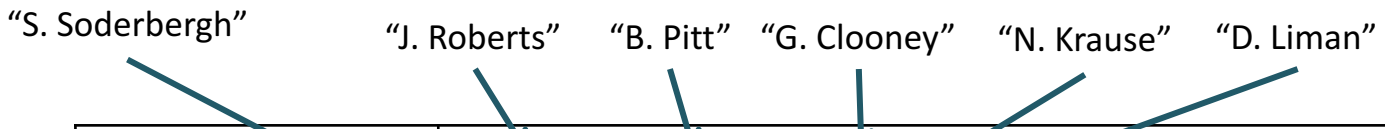


HasActor	{{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}, {f4}}
ReleaseDate	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasName	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasLanguage	{{f4, f5}}
HasWebsite	{{f1}, {f2}, {f3}, {f4}, {f5}, {f6}}

N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

Singleton filtering



HasActor	{{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}, {f4}}
ReleaseDate	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasName	{{f1}, {f2, f6}, {f3}, {f4}, {f5}}
HasLanguage	{{f4, f5}}
HasWebsite	{{f1}, {f2}, {f3}, {f4}, {f5}, {f6}}

Single key

N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

Singleton filtering

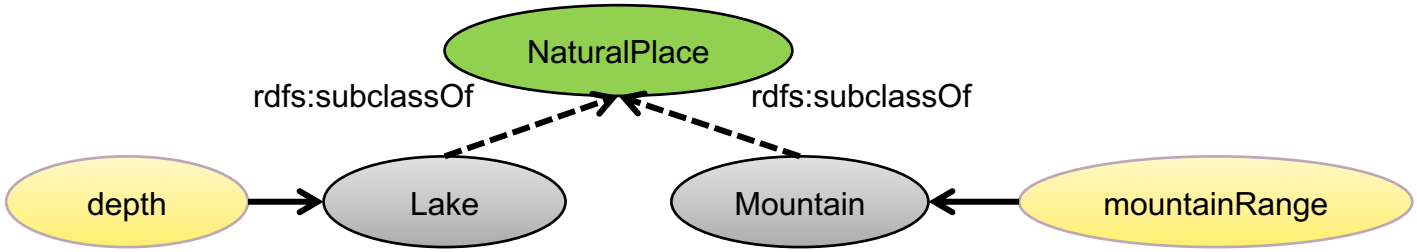
HasActor	{{f1, f2, f3}, {f2, f3, f4}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}}
ReleaseDate	{{f2, f6}}
HasName	{{f2, f6}}
HasLanguage	{{f4, f5}}

N-NON KEY DISCOVERY: POTENTIAL *N*-NON KEYS

[Symeonidou et al. 14]

Combinations of properties not needed to be explored

- Incomplete data
- Properties referring to different classes

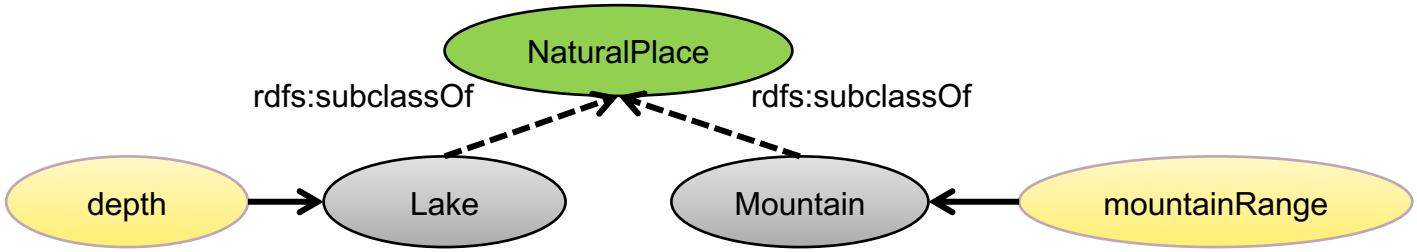


N-NON KEY DISCOVERY: POTENTIAL *N*-NON KEYS

[Symeonidou et al. 14]

Combinations of properties not needed to be explored

- Incomplete data
- Properties referring to different classes



- **Potential *n*-non keys**: Sets of properties that possibly refer to *n*-non keys

N-NON KEY DISCOVERY [Symeonidou et al. 14]

HasActor

- {f1, f2, f3} ∪ {f2, f3, f4} = {f1, f2, f3, f4} => 4-non key

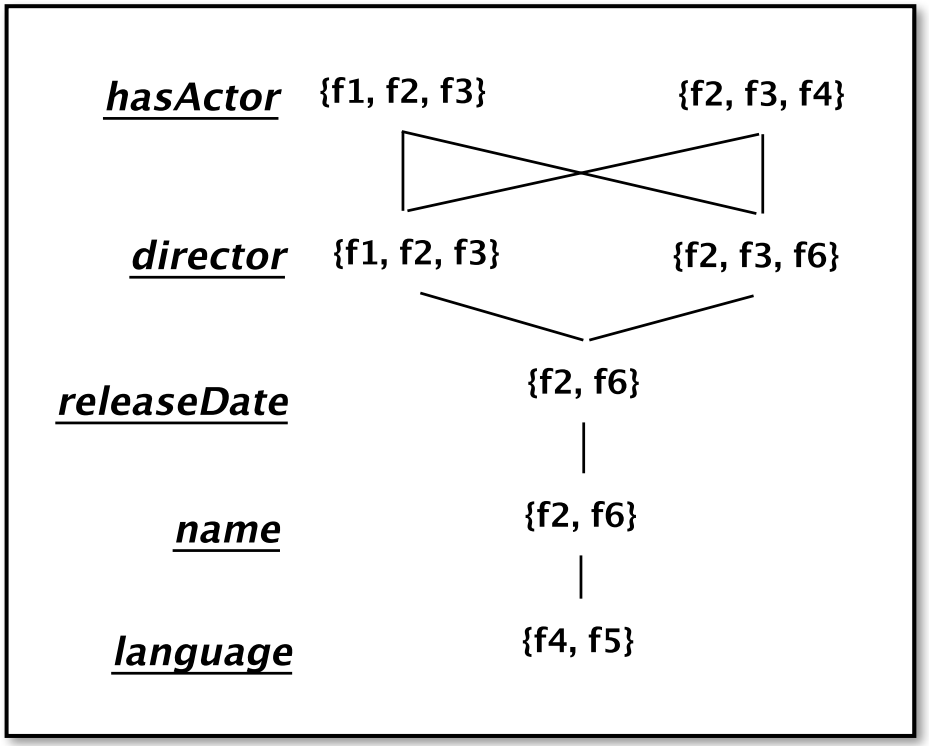
Composite *n*-non keys

- Intersections between sets of different properties

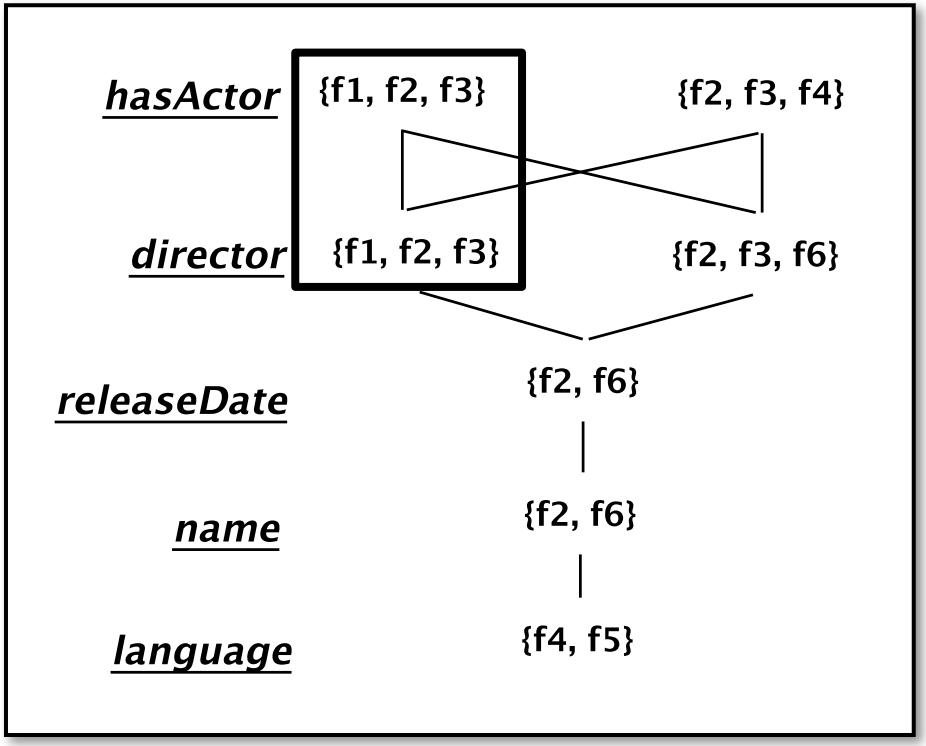
HasActor	{{f1, f2, f3}, {f2, f3, f4}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}}
ReleaseDate	{{f2, f6}}
HasName	{{f2, f6}}
HasLanguage	{{f4, f5}}

N-NON KEY DISCOVERY

[Symeonidou et al. 14]



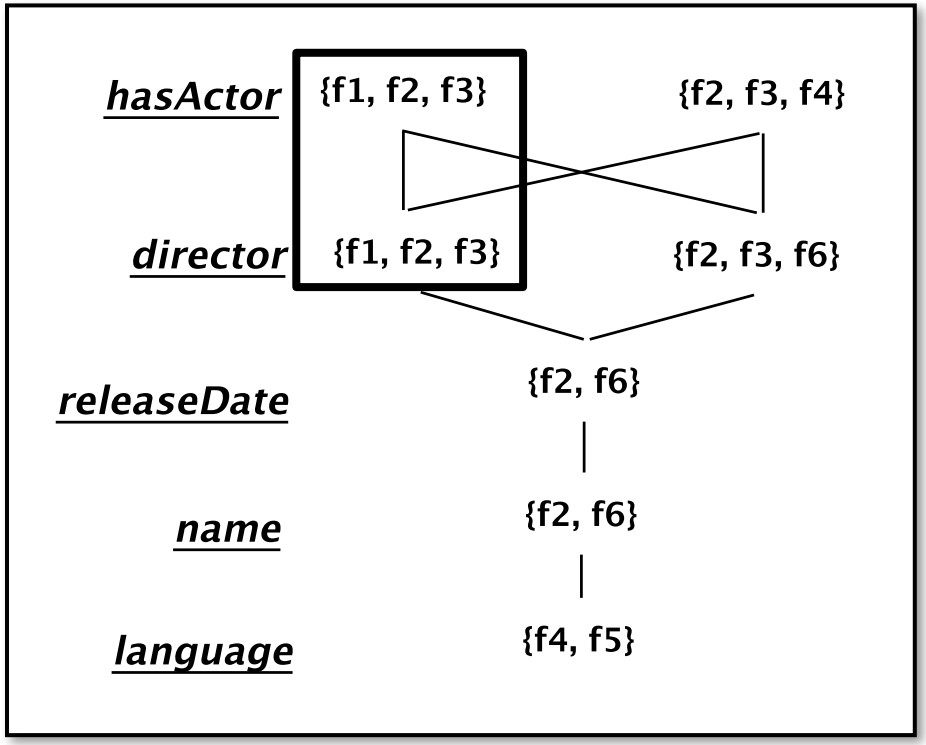
N-NON KEY DISCOVERY [Symeonidou et al. 14]



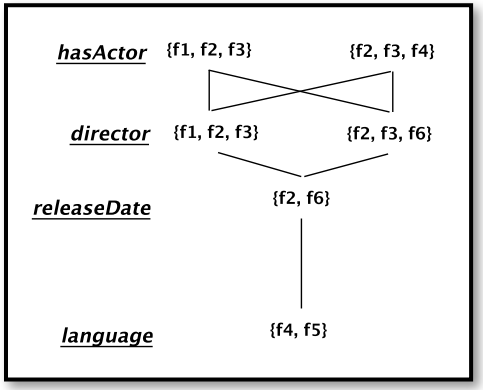
{hasActor, director} ➔ 3-non key

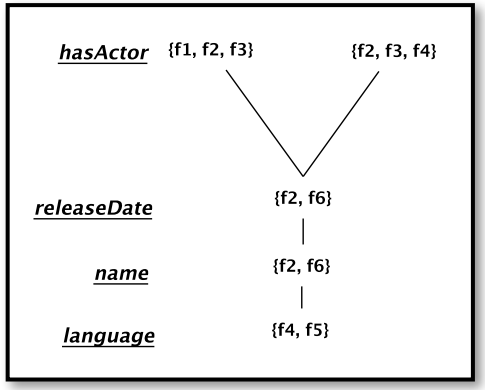
N-NON KEY DISCOVERY

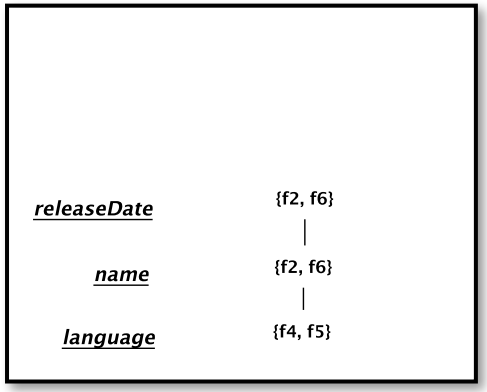
[Symeonidou et al. 14]



{hasActor, director} ➔ 3-non key







...

KEY DERIVATION

[Symeonidou et al. 14]

Interested only in maximal non keys

- All the sets of properties that are not maximal non keys are keys
  - **Example:** class described by the properties p1, p2, p3, p4

Maximal non key = [[p1, p2]] ➔ keys = [[p3], [p4]]

- Key derivation process :
  - Compute the complement set of each non key
  - Compute the Cartesian product of all the complement sets
  - Keep the minimal sets ➔ all the minimal keys.

DATA LINKING USING ALMOST KEYS

[Symeonidou et al. 14]

Goal: Compare linking results using almost keys with different *n*

Evaluation of linking using

- Recall
- Precision
- F-Measure

Datasets

- OAEI 2010
- OAEI 2013

Conclusion

- Linking results using *n*-almost keys are the better than using keys

EXAMPLE: DATA LINKING USING ALMOST KEYS

[Symeonidou et al. 14]

OAEI 2013 - Person

- BirthName, BirthDate, award, comment, label, BirthPlace, almaMater, doctoralAdvisor

	Almost keys	Recall	Precision	F-Measure
0-almost key	{BirthDate, award}	9.3%	100%	17%
2-almost key	{BirthDate}	32.5%	98.6%	49%

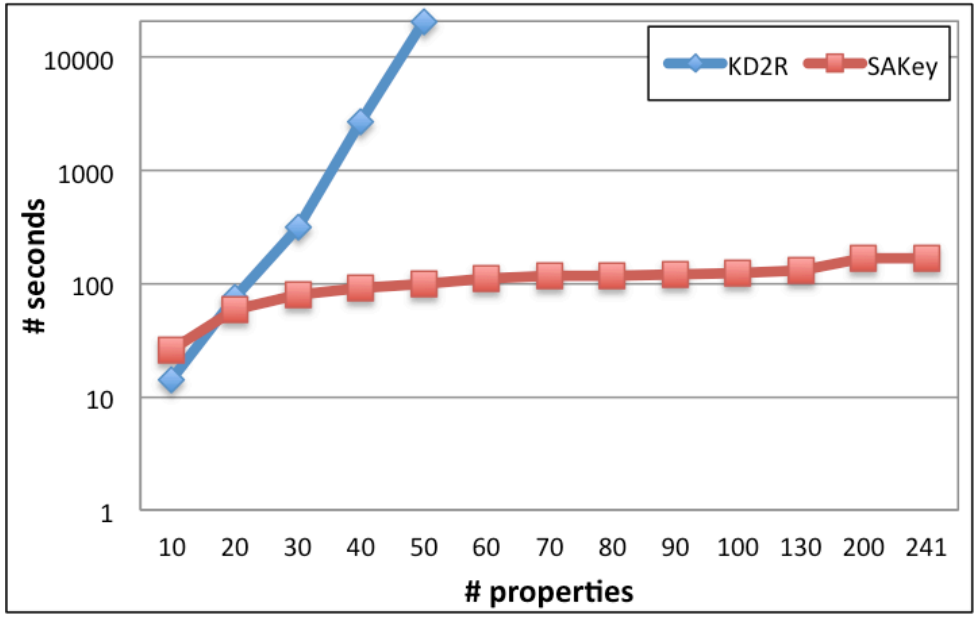
# exceptions	Recall	Precision	F-measure
0, 1	25.6%	100%	41%
2, 3	47.6%	98.1%	64.2%
4, 5	47.9%	96.3%	63.9%
6, ..., 16	48.1%	96.3%	64.1%
17	49.3%	82.8%	61.8%

KD2R VS. SAKEY - NON KEY DISCOVERY

[Symeonidou et al. 14]

Class	# triples	# Instances	#Properties	KD2R Runtime	SAKey Runtime (*n*=0)
DB:Website	8506	2870	66	13min	1s
YA:Building	114783	54384	17	26s	9s
DB:BodyOfWater	1068428	34000	200	outOfMem.	37s
DB:NaturalPlace	1604348	49913	243	outOfMem.	1min10s

Dbpedia class= DB:NaturalPlace

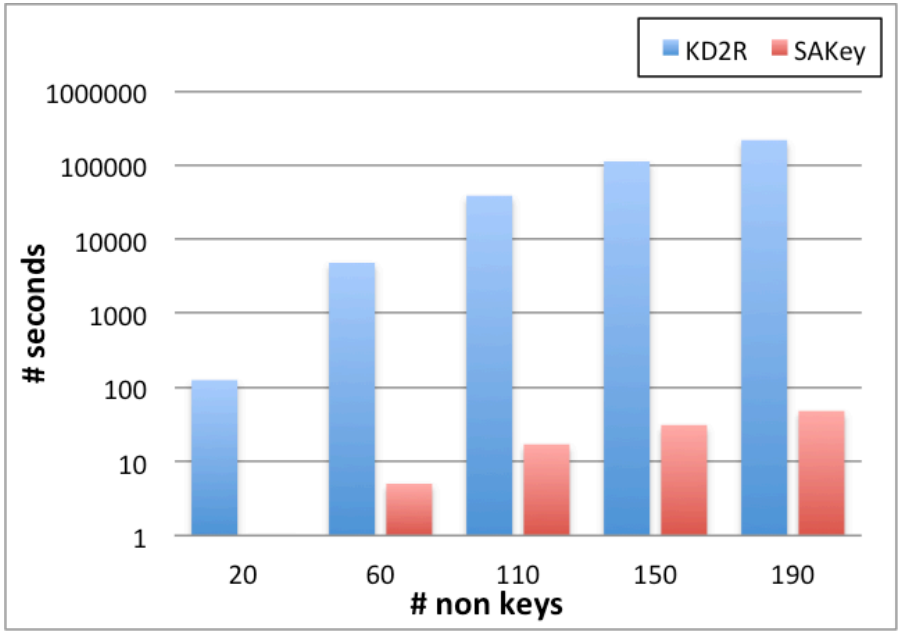


KD2R VS. SAKEY - KEY DERIVATION

[Symeonidou et al. 14]

Class	# non keys	# keys	KD2R	SAKey (*n*=0)
DB:Lake	50	480	1min10s	1s
DB:Mountain	49	821	8min	1s
DB:BodyOfWater	220	3846	> 1 day	66s
DB:NaturalPlace	302	7011	> 2 days	5min

Dbpedia class= DB:BodyOfWater



SOME FUTURE CHALLENGES

- **Data evolution**
  - ontology axiom evolution
  - link evolution
  - temporal data linking
- **Data veracity ➔ what is the truth?**
- **Knowledge discovery in complex data**
  - causality rules
- **Explanation problem**
  - provenance of the data and the data/knowledge processes

THANKS!

REFERENCES (1)

[Atencia et al. 2014] Data interlinking through robust Linkkey extraction.

Atencia, Manuel, Jérôme David, and Jérôme Euzenat. ECAI, 2014.

[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.

[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

[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. Graph2008

REFERENCES (2)

[Lenat and Feigenbaum 1991] On the threshold of knowledge.

Douglas B. Lenat and Edward A. Feigenbaum

In Artificial Intelligence 47 (1991)

[Nikolov et al'12] *Unsupervised Learning of Link Discovery Configuration*

Andriy Nikolov, Mathieu d'Aquin, Enrico Motta. In ESWC 2012.

[Pernelle et al.'13] An Automatic Key Discovery Approach for Data Linking.

Nathalie Pernelle, Fatiha Saïs. and Danai Symeounidou.

In Journal of Web Semantics

[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.

[Shvaiko,Euzenat13] Ontology Matching: State of the Art and Future Challenges,

Pavel Shvaiko, Jérôme Euzenat. In TKDE 2013

REFERENCES (3)

[Suchanek11] PARIS: Probabilistic Alignment of Relations, Instances, and Schema

Fabian Suchanek, Serge Abiteboul, Pierre Senellart. In VLDB 2011.

[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.

[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.

[Papageorgiou et al. 2017*] Approche numérique pour l'invalidation de liens d'identité (owl:SameAs)*. Dimitrios Christaras Papageorgiou, Nathalie Pernelle and Fatiha Saïs. In IC 2017.

REFERENCES (4)

[Papaleo et al. 2014] Logical Detection of Invalid SameAs Statements in RDF Data,

Laura Papaleo, Nathalie Pernelle, Fatiha Saïs and Cyril Dumont. In EKAW 2014

[Volz et al'09] Silk – A Link Discovery Framework for the Web of Data.

Julius Volz, Christian Bizer et al. In WWW 2009.

[*Zheng et al. 2013*] Results for OAEI 2013

Qian Zheng, Chao Shao, Juanzi Li, Zhichun Wang and Linmei Hu. OM@ISWC 2010

INSTANCE-BASED DATA LINKING APPROACHES

FRAMEWORK SILK [Volz et al'09]

- 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

[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

KNOFUSS (INSTANCE-BASED, UNSUPERVISED) [Nikolov et al'12]

- 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]

SUMMARY

Data linking: numerous and different approaches …

- Supervised approaches: needs samples of linked data

  → It can be avoided by using assumptions like (UNA)

- Graph-based approaches: decision propagation (good recall but highly time consuming)

- Logical approaches: good precision but partial

  ➔ Few approaches generate differentFrom(i1,i2) or use dissimilarity evidence

- Informed approaches: need knowledge to be declared in the ontology (generality) and/or ad-hoc knowledge given by an expert (a selection of properties, similarity functions)

  →This kind of knowledge are not always available but can be learnt/discovered from the data (e.g., key/rule discovery approaches) [Symeonidou et al. 14, Symeonidou et al. 17, Galarraga et al. 13]