IDENTITY MANAGEMENT IN THE WEB OF DATA

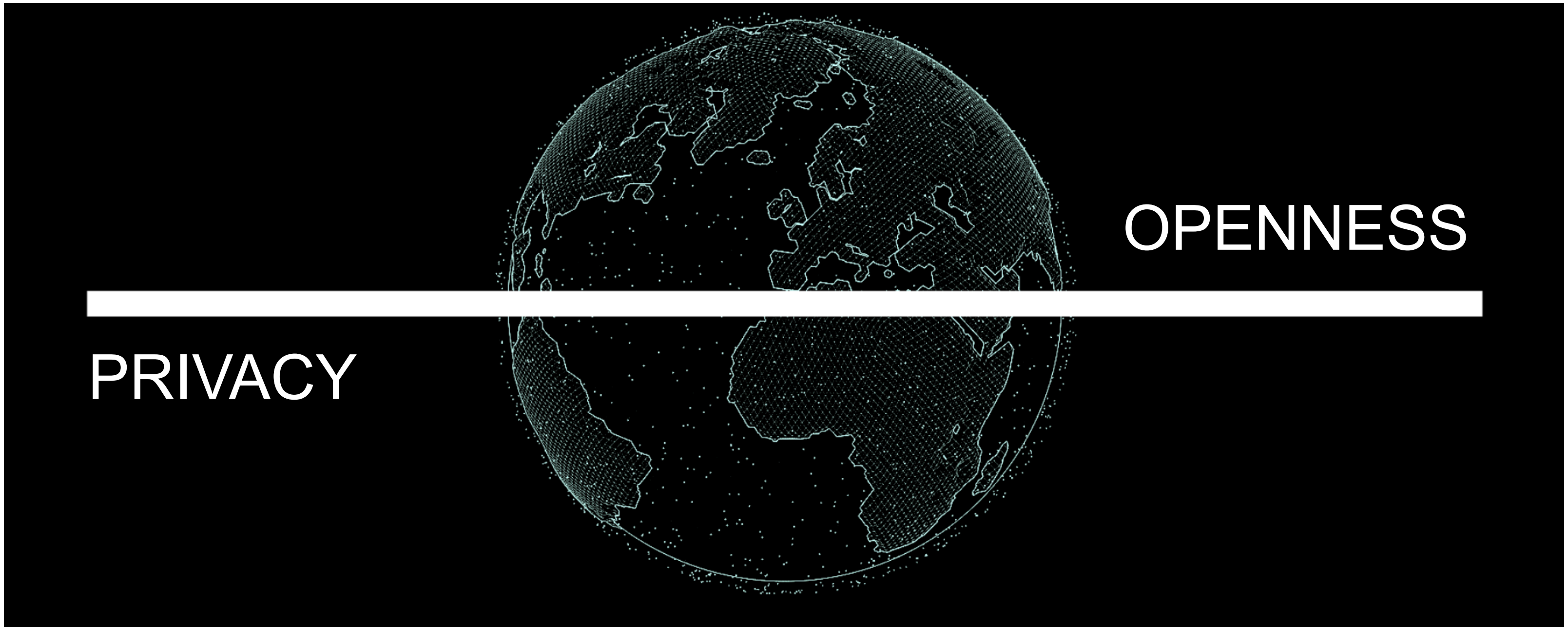
FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, ORSAY, FR

E-EGC2019

EGC2019 du 21 au 25 janvier 2019 à Metz

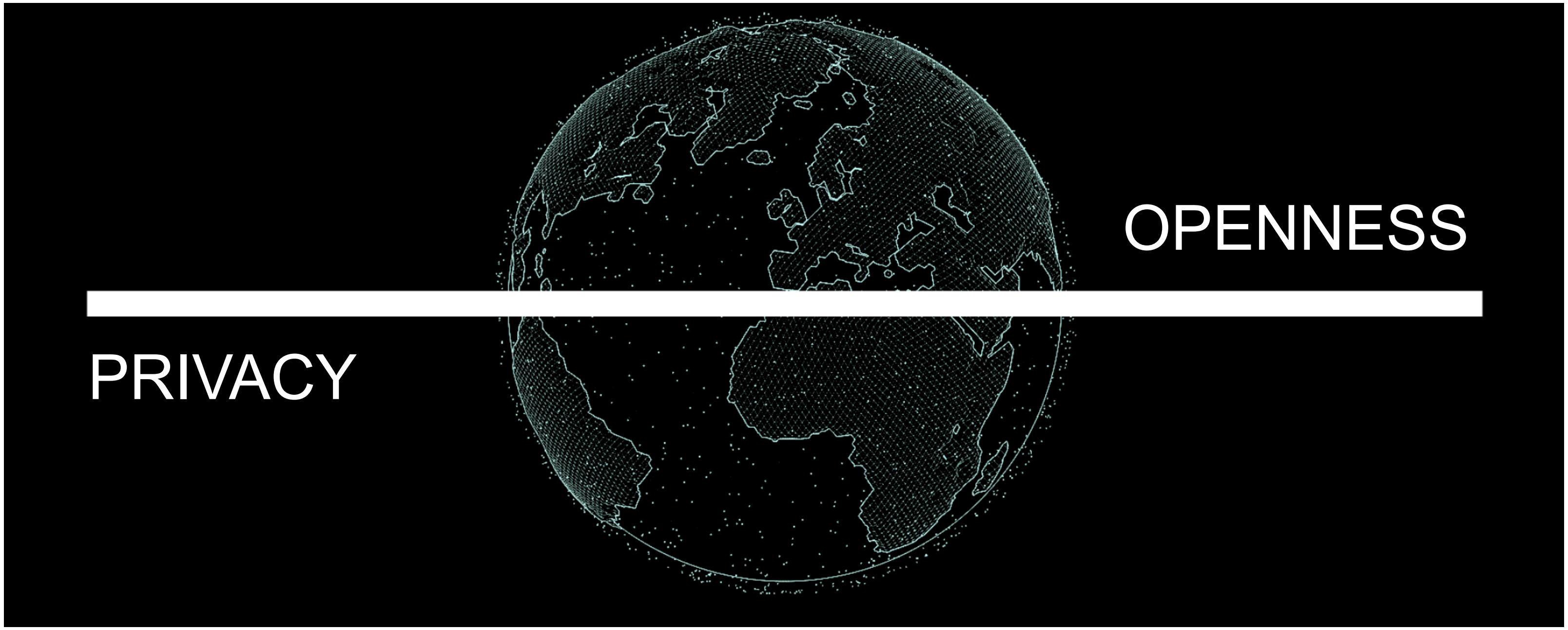
OPENNESS AND PRIVACY BALANCE



[source1]

[source1] https://cdn-images-1.medium.com/max/2600/1*xHzO_5cSSVetWnjpAbQABw.png

OPENNESS AND PRIVACY BALANCE



[source1]

Privacy

Open data contains the most detailed information, granular data often includes **personally sensitive** information.

Openness

Open data enables varied and detailed analyses, granular data is the most **interesting and useful** for businesses, policymakers, researchers, and the public.

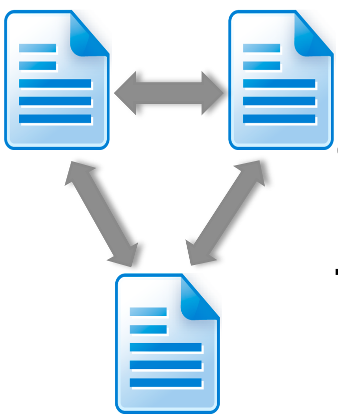
[source1] https://cdn-images-1.medium.com/max/2600/1*xHzO_5cSSVetWnjpAbQABw.png

OPEN DATA

LINKED OPEN DATA

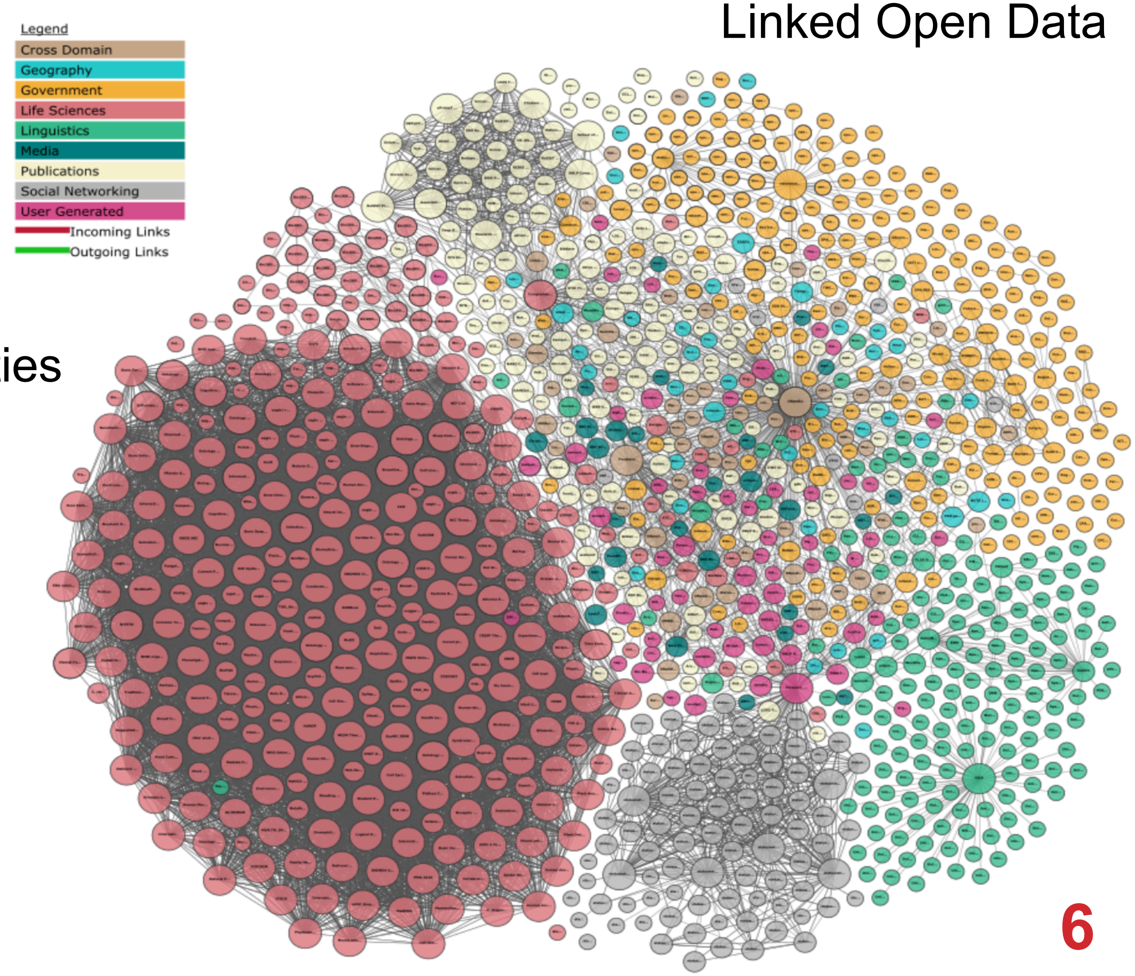
FROM THE WWW TO THE LINKED OPEN DATA

- applying the principles of the WWW to data



data is links,
not only properties

Linked Open Data



LINKED DATA PRINCIPLES

Tim Berners Lee, 2006

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**

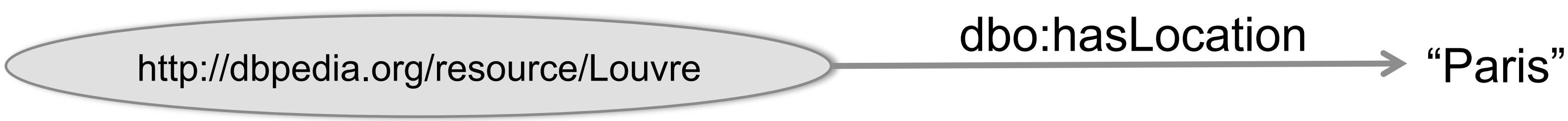
→so people can discover more information

→bridging disciplines and domains

➔ unlock the potential of isolated repositories (islands)

RDF – RESOURCE DESCRIPTION FRAMEWORK

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

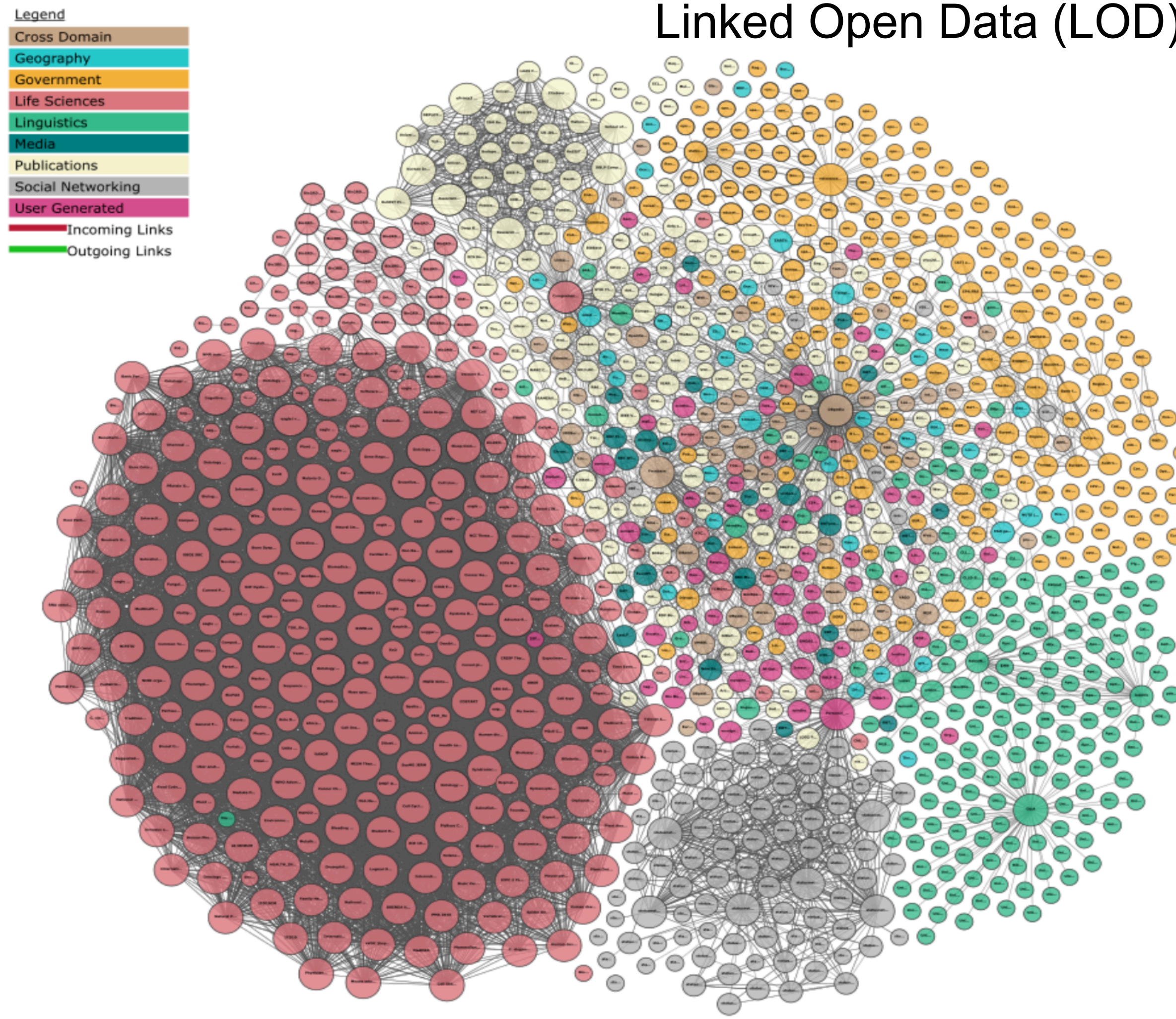


Subject Predicate Object

… is called a triple

LINKED OPEN DATA

Linked Open Data (LOD)



Linked Data - Datasets under an open access

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

Ex. DBPedia : 1.5 B 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

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)

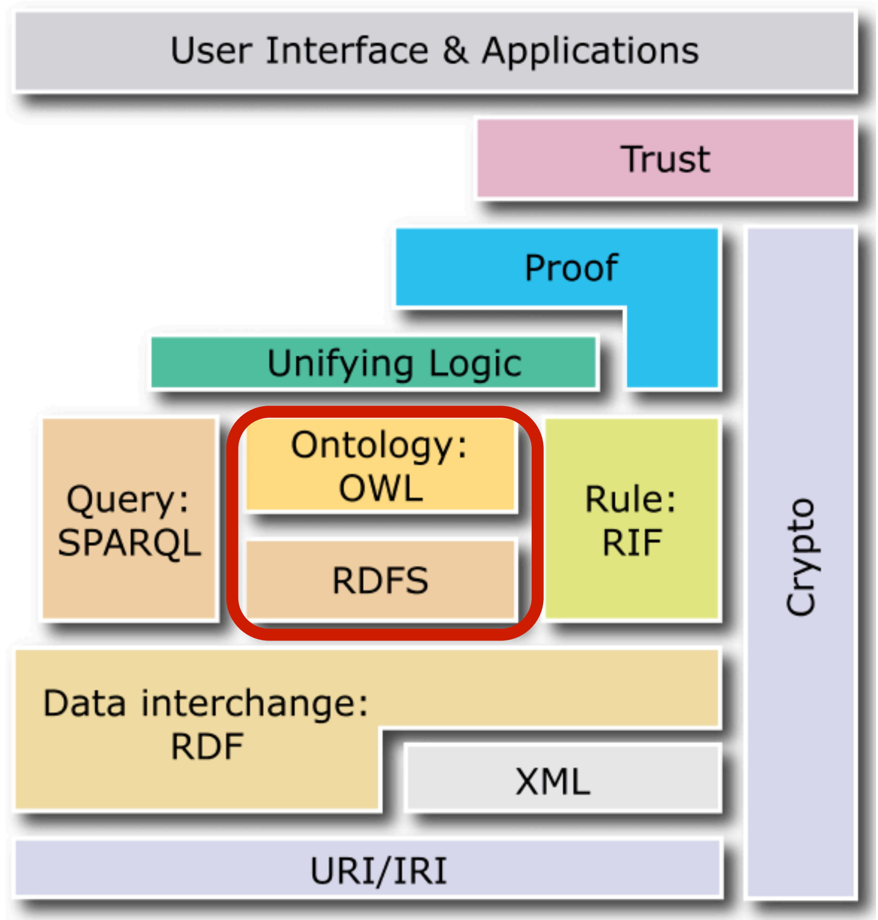
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema

- Lightweight ontologies

OWL – Web Ontology Language

- Expressive ontologies



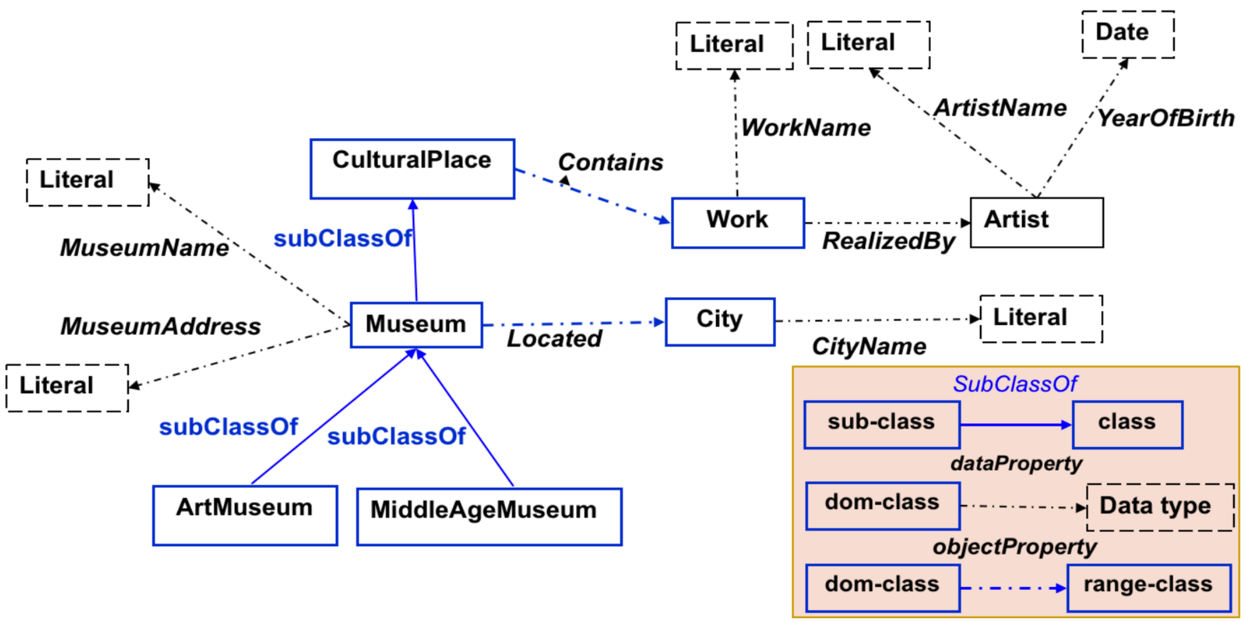
Source:
https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png

OWL ONTOLOGY

OWL – Web Ontology Language

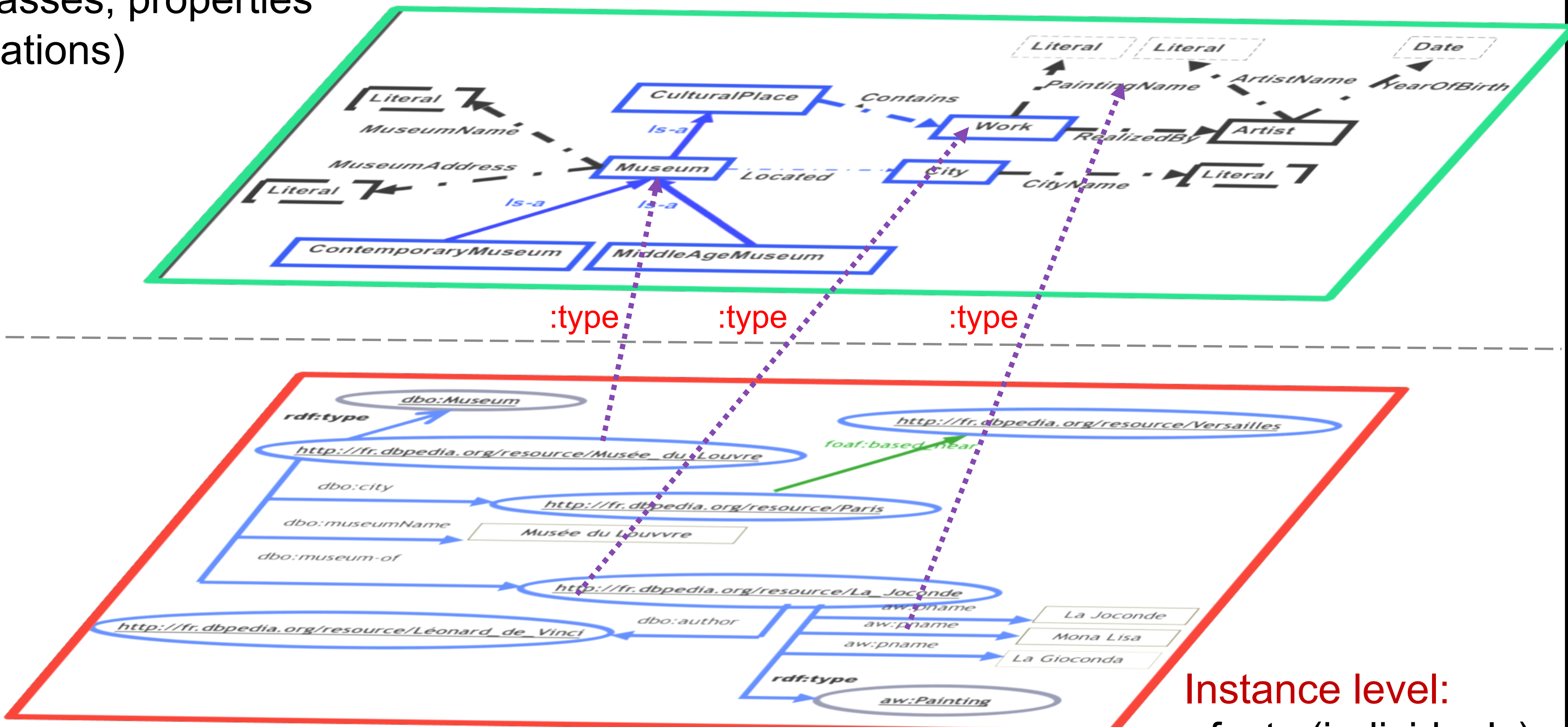
- Represents rich and complex knowledge about things
- Based on Description Logic
- Can be used to verify the consistency of knowledge
- Can make implicit knowledge explicit

- **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)



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:**

  Pompidou museum is an **Art Museum**

  < Pompidou_museum rdf:type ArtMuseum**> .**

  Pompidou museum contains ***Musicircus***

  < Pompidou_museum ao:contains ***Musicircus*****> .**

- **Infer that:**

➔ Pompidou museum is a CulturalPlace

< Pompidou_museum rdf:type CulturalPlace> .

Because: **Museum** subsumes **ArtMuseum** and **CulturalPlace** subsumes **Museum**

➔ ***Musicircus*** is a Work

<*Musicircus* rdf:type ao:Work> .

Because: the **range** of the object property **contains** is the class **Work**.

IDENTITY MANAGEMENT

- Detection of identity links between different descriptions of entities
- Discovery of identification rules, such as keys
- Detection of erroneous identity links and propose alternate links

OUTLINE

- Introduction

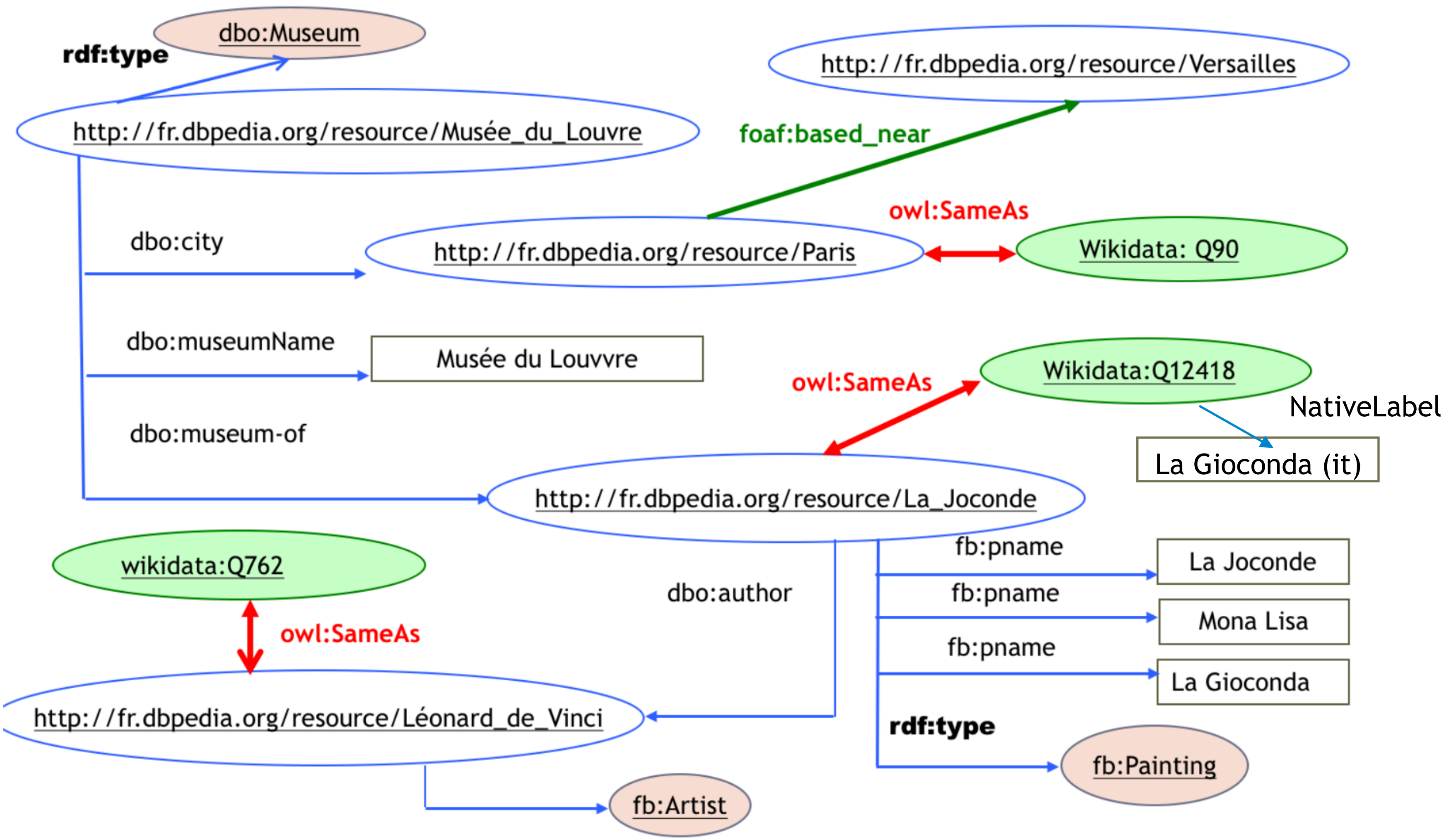
- Part 1: Data Linking
- Part 2: Key Discovery
- Part 3: Identity Link Invalidation

- Summary and Future Challenges

PART 1: DATA LINKING

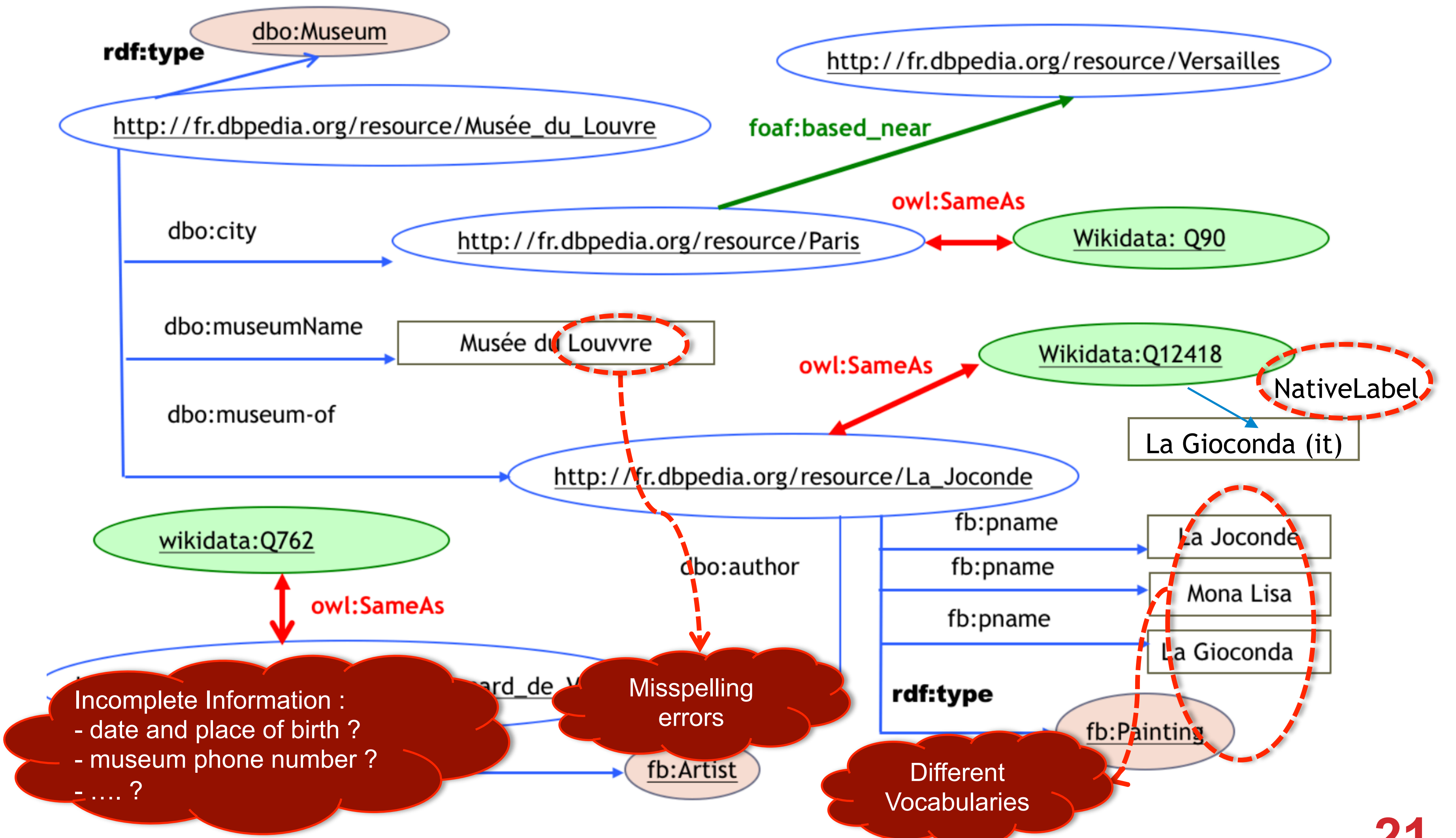
DATA LINKING

- **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: 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).



DATA LINKING 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) ∈ U1 × U2: owl:sameAs(s,t)} and
    - D = {(s,t) ∈ 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-

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 data type 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)

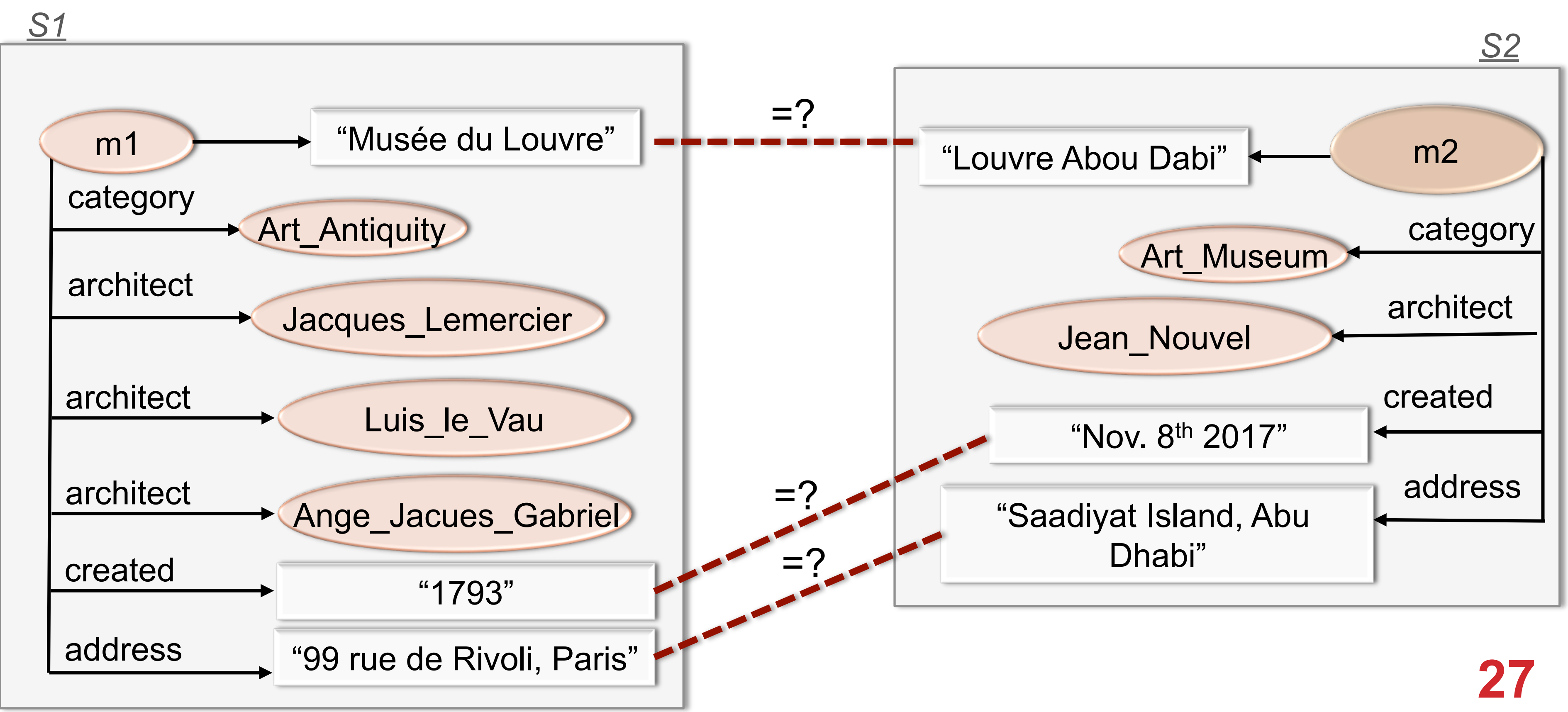
- **Supervised approaches**: need samples of linked data to learn models, or need interactions with expert

versus

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

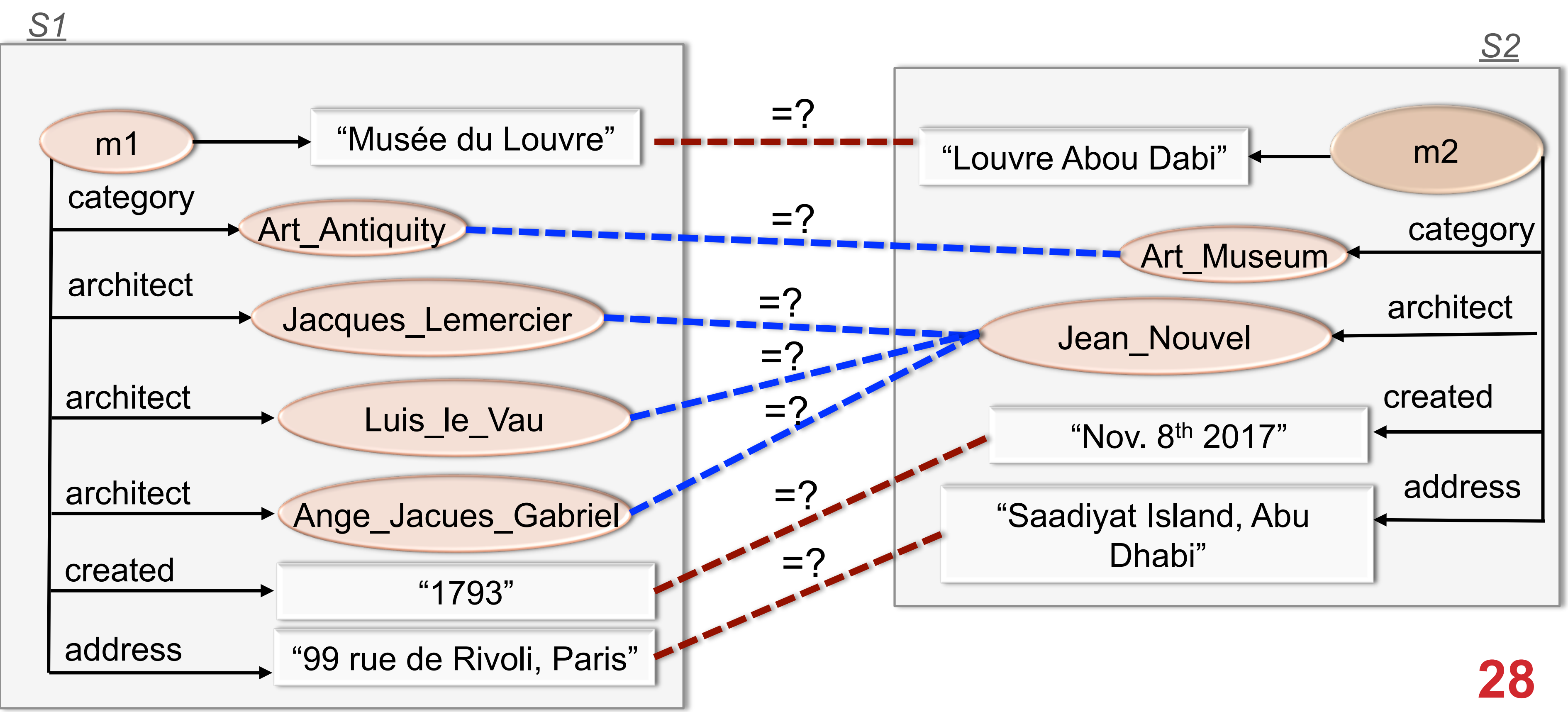
LOCAL APPROACHES

- Consider (path of) properties to compare pairs of instances <u>independently</u>



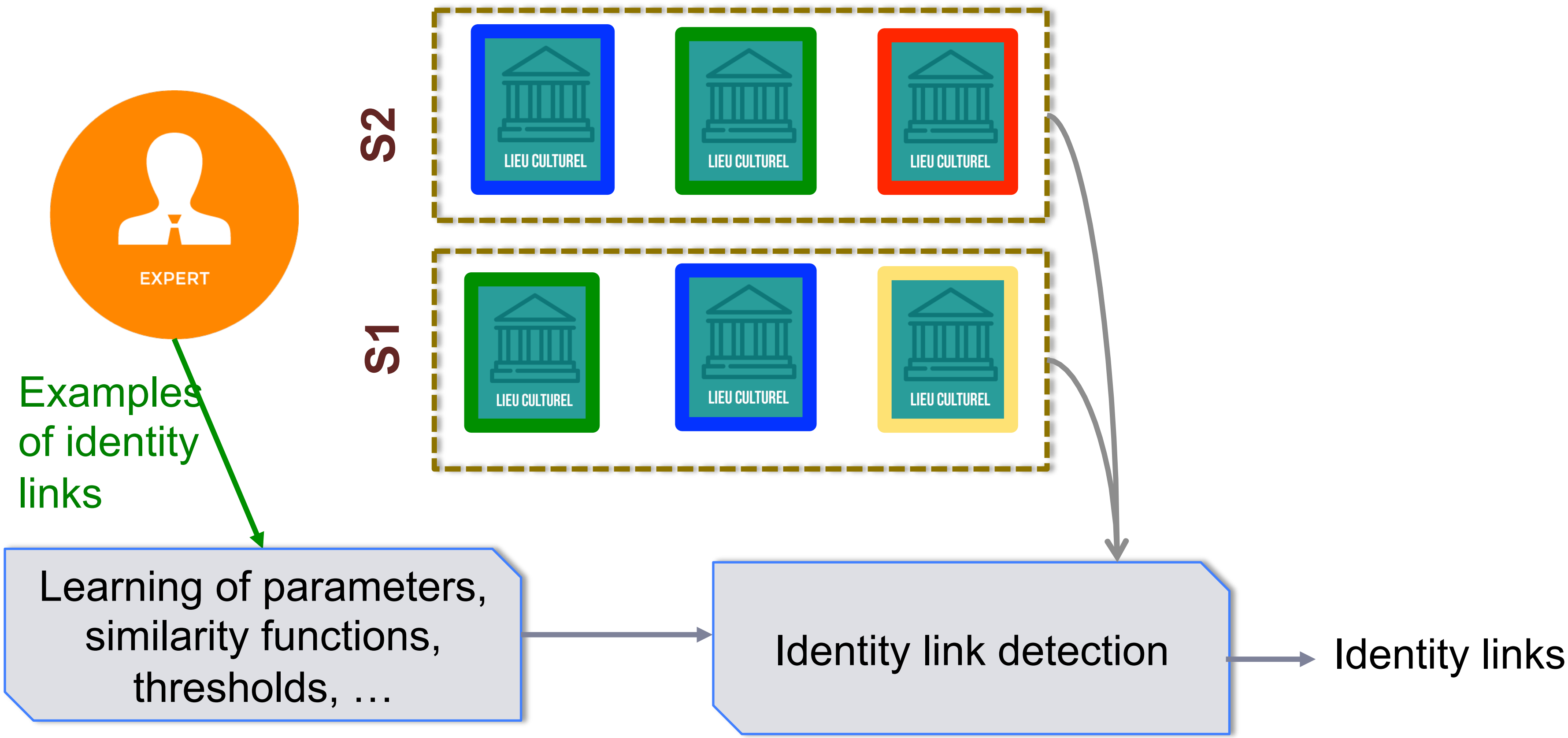
GLOBAL APPROACHES

- **Graph-based 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)



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)

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
- Generality: applicable to different datasets and different domains
- Use of benchmarks, like those of OAEI (Ontology Alignment Evaluation Initiative) or Lance

EXAMPLE: KNOFUSS (LOCAL, 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

See [Ferrara et al 2013] for a survey

EXAMPLE: KNOFUSS (LOCAL, UNSUPERVISED)

[Nikolov et al'12]

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

RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

homepage(X, Y) ∧ homepage(Z, Y) ➔ sameAs(X, Z)

	…	homepage
museum11		www.louvre.com
museum12		www.musee-orsay.fr
museum13		www.quai-branly.fr
museum14		…

homepage	…	
www.louvre.com		**museum21**
www.musee-orsay.fr		**museum22**
www.quai-branly.fr		**museum23**
…		**museum24**

RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

homepage(X, Y) ∧ homepage(Z, Y) ➔ sameAs(X, Z)

Then we may infer:

sameAs(museum11, museum11)
sameAs(museum12, museum22)
sameAs(museum13, museum23)

	…	homepage		homepage	…	
museum11		www.louvre.com	**SameAs**	www.louvre.com		**museum21**
museum12		www.musee-orsay.fr	**SameAs**	www.musee-orsay.fr		**museum22**
museum13		www.quai-branly.fr	**SameAs**	www.quai-branly.fr		**museum23**
museum14		…		…		**museum24**

RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

homepage(X, Y) ∧ homepage(Z, Y) ➔ sameAs(X, Z)

Then we may infer:

sameAs(museum11, museum11)
sameAs(museum12, museum22)
sameAs(museum13, museum23)

A key: is a set of properties that uniquely identifies every instance in the KG

	…	homepage
museum11		www.louvre.com
museum12		www.musee-orsay.fr
museum13		www.quai-branly.fr
museum14		…

SameAs
SameAs
SameAs

homepage	…	
www.louvre.com		**museum21**
www.musee-orsay.fr		**museum22**
www.quai-branly.fr		**museum23**
…		**museum24**

RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

$$\text{homepage}(X, Y) \land \text{homepage}(Z, Y) \rightarrow \text{sameAs}(X, Z)$$

Then we may infer:

sameAs(museum11, museum11)
sameAs(museum12, museum22)
sameAs(museum13, museum23)

A key: is a set of properties that uniquely identifies every instance in the KG

	…	homepage		homepage	…	
museum11		www.louvre.com	SameAs	www.louvre.com		**museum21**
museum12		www.musee-orsay.fr	SameAs	www.musee-orsay.fr		**museum22**
museum13		www.quai-branly.fr	SameAs	www.quai-branly.fr		**museum23**
museum14		…		…		**museum24**

How to automatically discover keys from KGs?

OUTLINE

- Introduction
- Part 1: Data Linking
- Part 2: Key Discovery
- Part 3: Identity Link Invalidation
- Summary and Future Challenges

PART 2:
KEY DISCOVERY

KEY SEMANTICS

- **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))

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

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

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

Book(x_1) ∧ Book(x_2) ∧ Author(x_1, y) ∧ Author (x_2, y) ∧ Title(x_1,w) ∧ Title(x_2, w)
➔ sameAs(x_1, x_2)

KEY VALIDITY

A key is a set of properties that uniquely identifies every instance in the data

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor, Director
Person2	Marie	Tompson	02/09/75	Actor
Person3	Marie	David	15/02/85	Actor
Person4	Vincent	Solgar	25/01/72	Actor, Director
Person5	Simon	Roche	06/12/90	Teacher
Person6	Jane	Ser	15/05/87	Teacher, Researcher
Person7	Sara	Khan	27/10/84	Teacher
Person8	Theo	Martin	06/12/90	Teacher, Researcher
Person9	Marc	Blanc	27/10/84	Teacher

Is [LastName] a key? ✖

Is [FirstName,LastName] a key? ✔

Exact keys

KEY VALIDITY

A key is a set of properties that uniquely identifies every instance in the data

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor, Director
Person2	Marie	Tompson	02/09/75	Actor
Person3	Marie	David	15/02/85	Actor
Person4	Vincent	Solgar	25/01/72	Actor, Director
Person5	Simon	Roche	06/12/90	Teacher
Person6	Jane	Ser	15/05/87	Teacher, Researcher
Person7	Sara	Khan	27/10/84	Teacher
Person8	Theo	Martin	06/12/90	Teacher, Researcher
Person9	Marc	Blanc	27/10/84	Teacher

Is [FirstName,LastName] a key? ✔

Exact keys

Is [Birthdate] a key with 2 exceptions? ✔

Almost keys

KEY VALIDITY

A key is a set of properties that uniquely identifies every instance in the data

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor, Director
Person2	Marie	Tompson	02/09/75	Actor
Person3	Marie	David	15/02/85	Actor
Person4	Vincent	Solgar	25/01/72	Actor, Director
Person5	Simon	Roche	06/12/90	Teacher
Person6	Jane	Ser	15/05/87	Teacher, Researcher
Person7	Sara	Khan	27/10/84	Teacher
Person8	Theo	Martin	06/12/90	Teacher, Researcher
Person9	Marc	Blanc	27/10/84	Teacher

Is [FirstName,LastName] a key? ✔

Exact keys

Is [Birthdate] a key with 2 exceptions? ✔

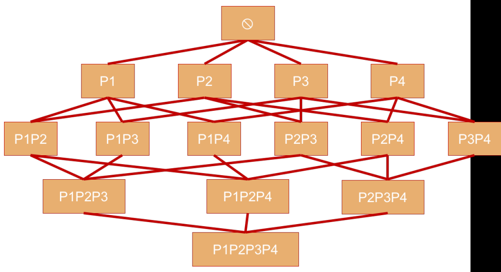
Almost keys

*Is [Birthdate and (**Profession =“Actor”**)] a key?* ✔

Conditional keys

KEY DISCOVERY: A COMPLEX PROBLEM

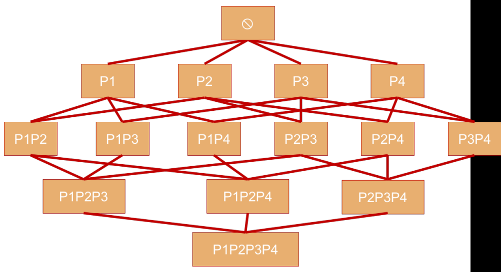
- Find all the minimal keys requires 2^n property combinations



- For each combination scan all the instances

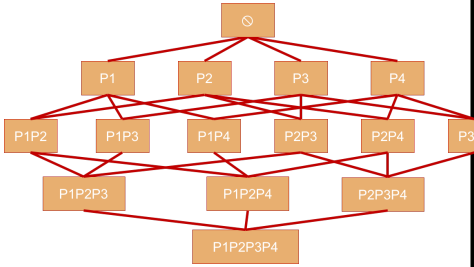
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires 2^n property combinations
  need of efficient filtering and prunings
- For each combination scan all the instances



KEY DISCOVERY: A COMPLEX PROBLEM

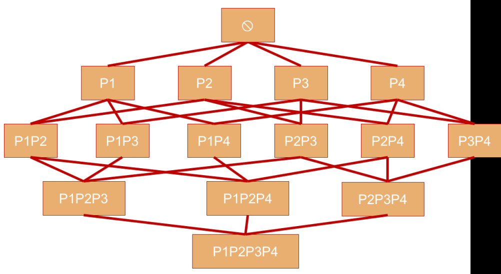
- Find all the minimal keys requires 2^n property combinations
  need of efficient filtering and prunings
- For each combination scan all the instances



maximal non-keys $\xrightarrow{\text{derive}}$ minimal keys

KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires 2^n property combinations
  need of efficient filtering and prunings
- For each combination scan all the instances



maximal non-keys $\xrightarrow{\text{derive}}$ minimal keys

	FirstName	LastName	Birthdate	Profession
Person1	Anne	Tompson	15/02/88	Actor
Person2	Marie	Tompson	02/09/75	Actor
Person3	Marie	David	15/02/85	Actor
Person4	Vincent	Solgar	25/01/72	Actor
Person4	Simon	Roche	06/12/90	Teacher
Person4	Jane	Ser	15/05/87	Teacher
Person4	Sara	Khan	27/10/84	Teacher
Person4	Theo	Martin	06/12/90	Teacher
Person4	Marc	Blanc	27/10/84	Teacher

Is [LastName] a non-key?

➔ scan only a part of the data

Is [FirstName,LastName] a key?

➔ scan all the data

KEY DISCOVERY IN KNOWLEDGE GRAPHS

- **Data characteristics:**
  - Conforms to an ontology
  - Multi-valued properties
  - Incomplete data
  - Errors
  - Big datasets
  - No exact key can be discovered

- **Assumptions:**
  - UNA-Unique Name assumption
  - OWA-Open World assumption

KEY DISCOVERY IN KNOWLEDGE GRAPHS: CONTRIBUTIONS

- **Data characteristics:**
  - Conforms to an ontology
  - Multi-valued properties
  - Incomplete data
  - Errors
  - Big datasets
  - No exact key can be discovered
- **Assumptions:**
  - UNA-Unique Name assumption
  - OWA-Open World assumption

- KD2R: Exact keys discovery [Pernelle et al. '13]
  - Multi-valued properties
  - Incomplete data

UNA

OWA

- SAKey: Almost-keys discovery [Symeonidou et al. 2014]
  - Few errors

UNA

OWA

- VICKey: Conditional keys discovery [Symeonidou et al. 2017]
  - No exact key is valid

UNA

OWA

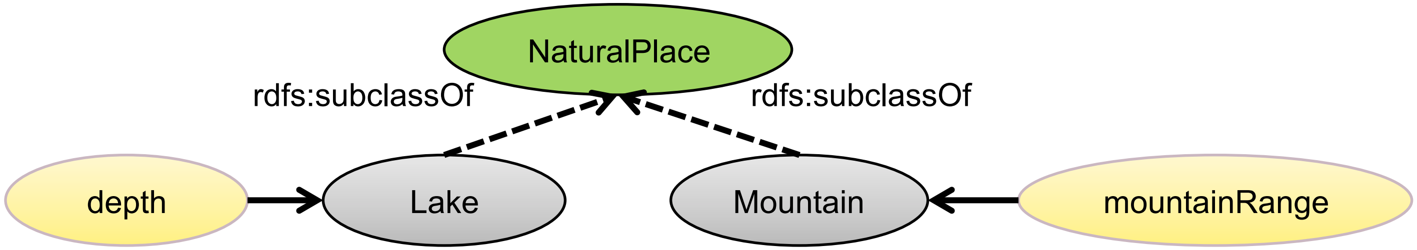
KEY DISCOVERY IN KNOWLEDGE GRAPHS: CONTRIBUTIONS

- KD2R: Exact keys discovery [Pernelle et al. 2013]
  - Derives minimal keys from maximal non-keys
  - Key inheritance pruning
  - Key monotonicity and non-key anti-monotonicity prunings
- SAKey: Almost-keys discovery [Symeonidou et al. 2014]
  - Derives minimal n-almost keys from maximal (n+1) non-keys
  - Key monotonicity and non-key anti-monotonicity prunings
  - Singleton pruning and single key pruning
  - Potential (n+1) non-key computation
  - Semantic dependencies pruning
- VICKey: Conditional keys discovery [Symeonidou et al. 2017]
  - Derives minimal conditional keys from maximal non-keys
  - Key monotonicity and non-key anti-monotonicity prunings
  - Key support and coverage

SAKEY: N-NON-KEY DISCOVERY (SAKEY)

- **(n+1) potential non-key construction:** filtering of combinations of properties not needed be explored

  - Incomplete data
  - Properties referring to different classes



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

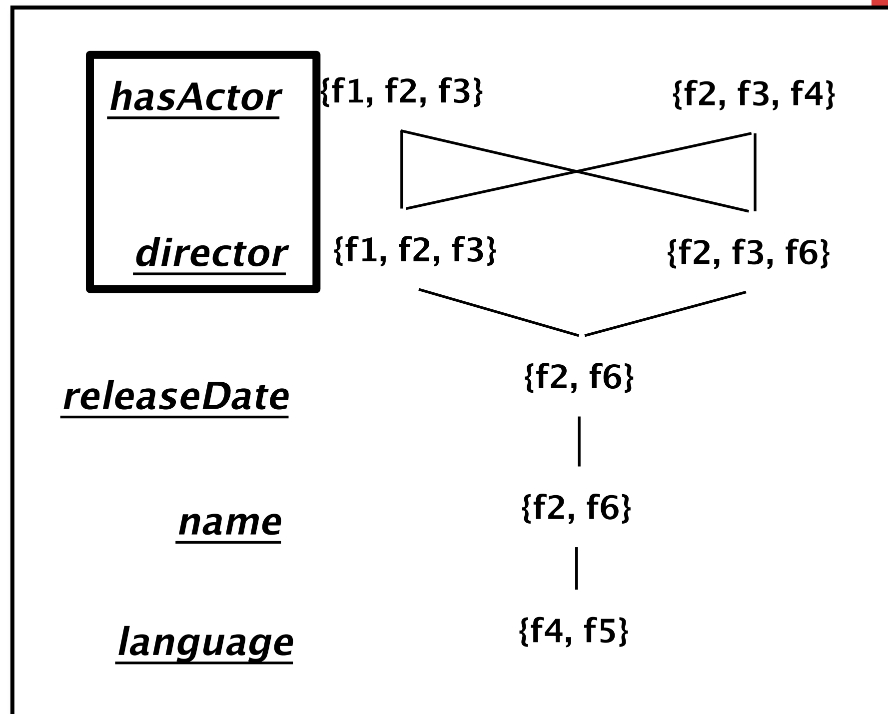
SAKEY: N-NON-KEY DISCOVERY (SAKEY)

(n+1)-non key discovery:

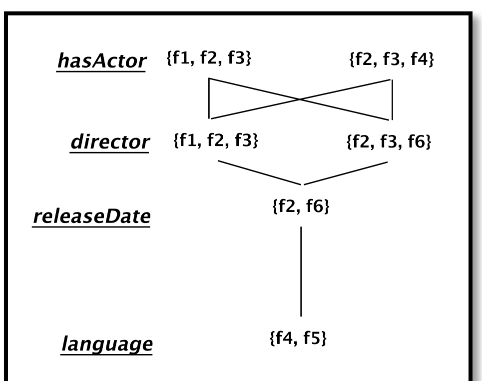
Intersections between sets of properties

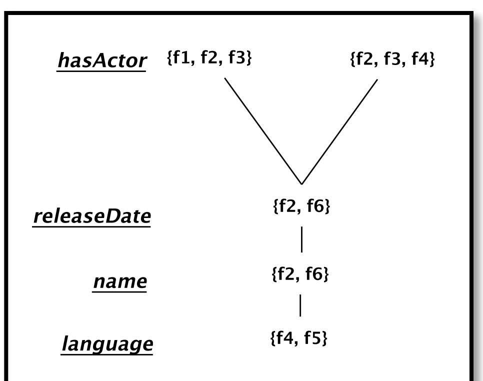
Final Map

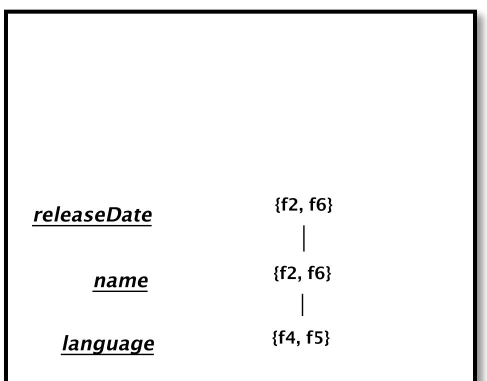
HasActor	{{f1, f2, f3}, {f2, f3, f4}}
HasDirector	{{f1,f2,f3}, {f2, f3, f6}}
ReleaseDate	{{f2, f6}}
HasName	{{f2, f6}}
HasLanguage	{{f4, f5}}



{hasActor, director} → 3-non key







. . .

KD2R, SAKEY AND VICKEY: EVALUATION

- **Scalability and runtime evaluation**
  - SAKey can handle size classes much larger than KD2R (DB: Natural Place more than 16 million triples and 243 properties (non-key discovery in 1min and key derivation 5min)
  - The use of prunings decreases the number of nodes to explore (e.g. a decrease of 50% for KD2R on DBpedia person) and the runtime (e.g. a decrease of 23% of runtime in SAKey)

KD2R, SAKEY AND VICKEY: EVALUATION

- **Scalability and runtime evaluation**
  - SAKey can handle size classes much larger than KD2R (DB: Natural Place more than 16 million triples and 243 properties (non-key discovery in 1min and key derivation 5min)
  - The use of prunings decreases the number of nodes to explore (e.g. a decrease of 50% for KD2R on DBpedia person) and the runtime (e.g. a decrease of 23% of runtime in SAKey)

- **Relevance of keys for the data linking (with equality of literals)**
  - When keys (KD2R on OAEI2010:Person) are used F-Measure increases from 0.24 to 0.76
  - When conditional keys (VICKEY on Dbpedia and Yago) are used F-Measure increases from 0.08 to 0.55
  - When 3-almost keys (SAKey on OAEI 2013:Film) are used the F-measure is of 0.81

KEY DISCOVERY IN KGS: CHALLENGES

- Choose the good **key semantics** using the data characteristics (e.g. completeness)
- Define **holistic approaches** to discover different kinds of dependency constraints (e.g., denial constraints, key graphs and referring expressions)
- Define **incremental approaches** taking into account data evolution

OUTLINE

- Introduction
- Part 1: Data Linking
- Part 2: Key Discovery
- Part 3: Identity Link Invalidation
- Summary and Future Challenges

PART 3: IDENTITY LINK INVALIDATION

IDENTITY IS COMPLEX …

"Lessons Learned: Managing Identity is Hard"

Jamie Taylor
in ISWC 2017

Google

Knowledge Graph

"Biggest Problem: Identity"

Alan Patterson
in ISWC 2018

ebay

Knowledge Graph

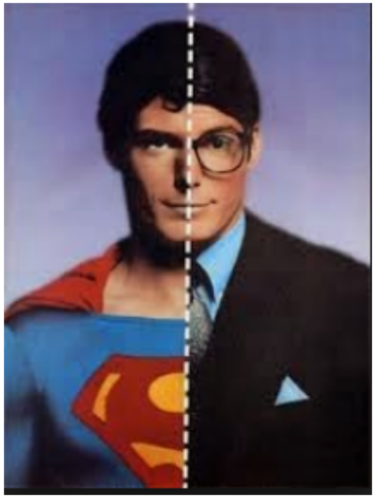
Source: Aaron Bradley
Twitter, October 26^{th}, 2018

IDENTITY IS COMPLEX ...

From a Philosophical Point of View [Beek, 2018]

① **Identity does not hold across modal contexts**

- *Allowing Lois Lane to believe that Superman saved her without requiring her to believe that Clark Kent saved her.*



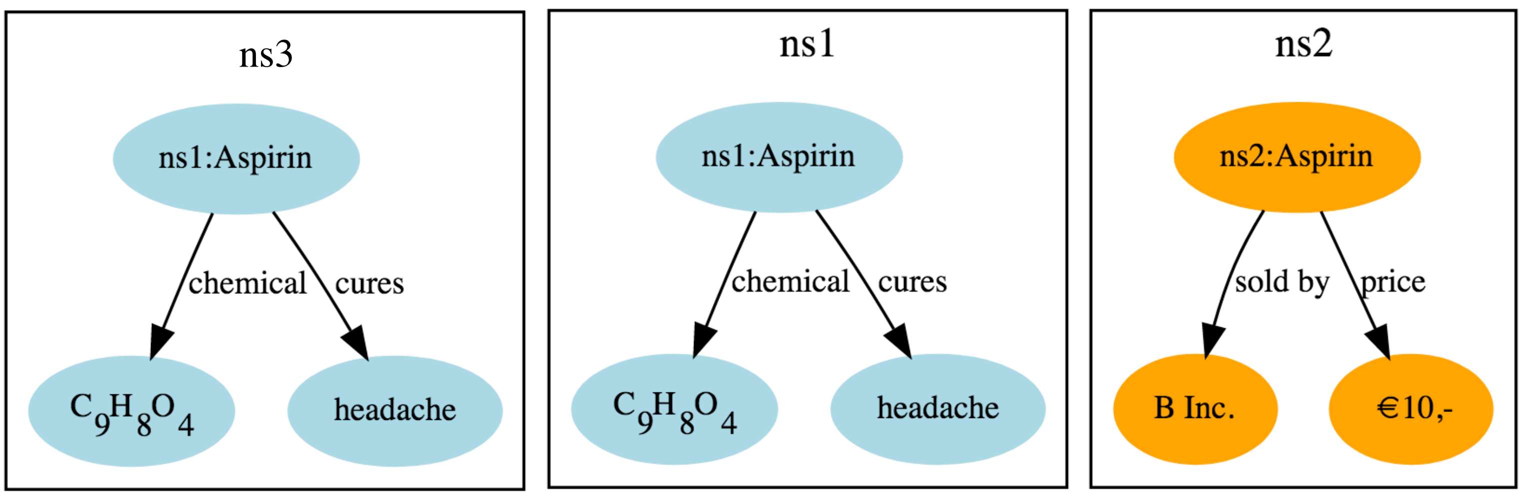
IDENTITY IS COMPLEX …

From a Philosophical Point of View [Beek, 2018]

① Identity does not hold across modal contexts

② **Identity is context-dependent [Geach, 1967]**

- *Allowing two medicines with the same chemical structure to be considered the same in a scientific context, but different in a commercial context (e.g., because they* are produced by different companies).



IDENTITY IS COMPLEX ...

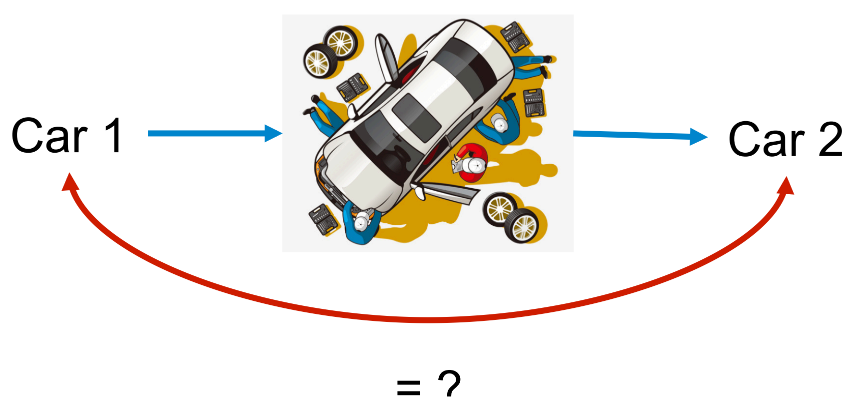
From a Philosophical Point of View [Beek, 2018]

① Identity does not hold across modal contexts

② Identity is context-dependent [Geach, 1967]

③ **Identity over time poses problems**

- since a car may be considered the same car, even though some (or even all) of its original components have been replaced by new ones.



IDENTITY IS COMPLEX ...

From an Operational Point of View

① **Unless two resources are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical**

- *Only **3.6K owl:differentFrom** triples compared to **558M owl:sameAs** (LOD-a-lot dataset, 2015 crawl of the LOD Cloud)*

IDENTITY IS COMPLEX …

From an Operational Point of View

① Unless two resources are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

② **Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource**

- *Barack Obama the person* ***vs*** *URL referring to his Web page (Problem of Sense and Reference [Halpin, 2010])*

IDENTITY IS COMPLEX …

From an Operational Point of View

① Unless two things are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

② Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource

③ **Modelers have different opinions about whether two objects are the same**

- *From a set of 250 owl:sameAs links, one Semantic Web expert judged that only **73** are correct identity links, whilst two other experts have judged **132** and **181** as true identity links, respectively [Halpin et al., 2010]*

IDENTITY IS COMPLEX …

From an Operational Point of View

① Unless two things are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

② Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource

③ Modelers have different opinions about whether two objects are the same

④ **Data linking approaches are rarely 100% precise**

- *Precision usually between 67% and 86% [OAEI 2017, OAEI 2018]*

IDENTITY IS COMPLEX …

From an Operational Point of View

① Unless two things are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

② Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource

③ Modelers have different opinions about whether two objects are the same

④ Data linkage approaches are rarely 100% precise

⑤ **Lack of alternative well-defined and standardized identity predicates**

- *rdfs:seeAlso, skos:exactMatch, etc. → Lack of formal semantics*

THE 'SAMEAS PROBLEM'

Web of Data contains a large* number of erroneous owl:sameAs

***~21%** **[Halpin et al., 2010]**	***~2.8%** **[Hogan et al., 2012]**	***~4%** **[Raad, 2018]**
Manual evaluation of 250 owl:sameAs from the Web	Manual evaluation of 1K identical pairs from the Web	Manual evaluation of 300 owl:sameAs from the LOD Cloud + error degree distribution of 558M owl:sameAs

THE ‘SAMEAS PROBLEM’

The largest identity set contains 177,794 terms that 'should' refer to the same real world entity

However:

http://dbpedia.org/resource/Albert_Einstein
http://dbpedia.org/resource/Basketball
http://dbpedia.org/resource/Coca-Cola
http://dbpedia.org/resource/Deauville
http://dbpedia.org/resource/Italy
http://dbpedia.org/resource/Lists_of_christian_religions
...

Full list at: https://sameas.cc/term?id=4073

HOW TO LIMIT THIS 'SAMEAS PROBLEM'?

- **Detect erroneous identity links / Validate correct ones**
  - Inconsistency-based Approaches
  - Content-based Approaches
  - Network-based Approaches

- **Propose alternative semantics for identity**
  - Weak-Identity and Similarity predicates
  - Contextual Identity

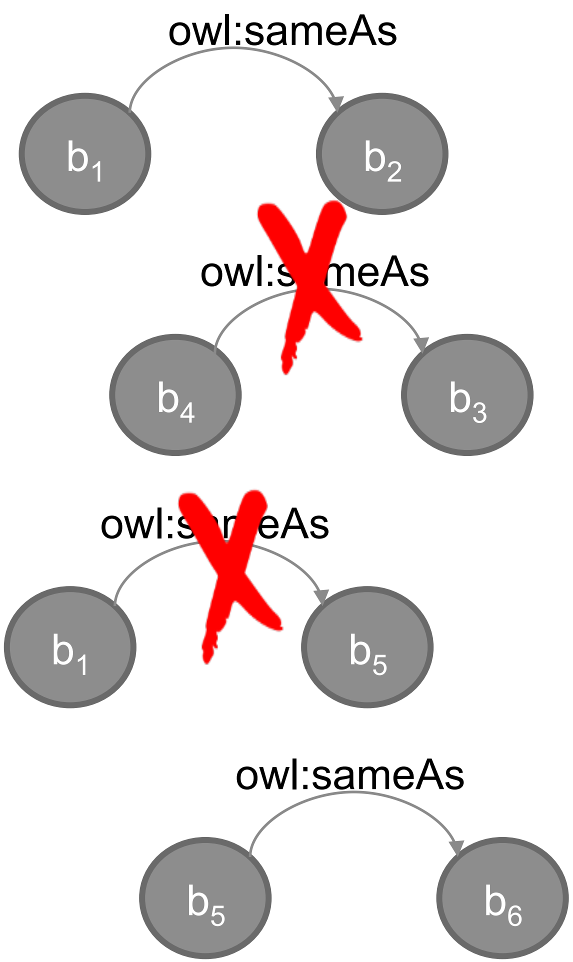
OWL:SAMEAS PREDICATE

- owl:sameAs, indicates that two different descriptions refer to the same entity
- a strict semantics,
  1) Reflexive,
  2) Symmetric,
  3) Transitive and
  4) Fulfils property sharing:

$$\forall X \forall Y \ \text{owl:sameAs}(X, Y) \wedge p(X, Z) \Rightarrow p(Y, Z)$$

IDENTITY LINK INVALIDATION

- **Link invalidation** consists in determining whether an identity link is **erroneous**
- **Different kinds of information can be used:**
  - Resource descriptions
  - Consistency constraints
  - Source trustworthiness
  - Identity network metrics



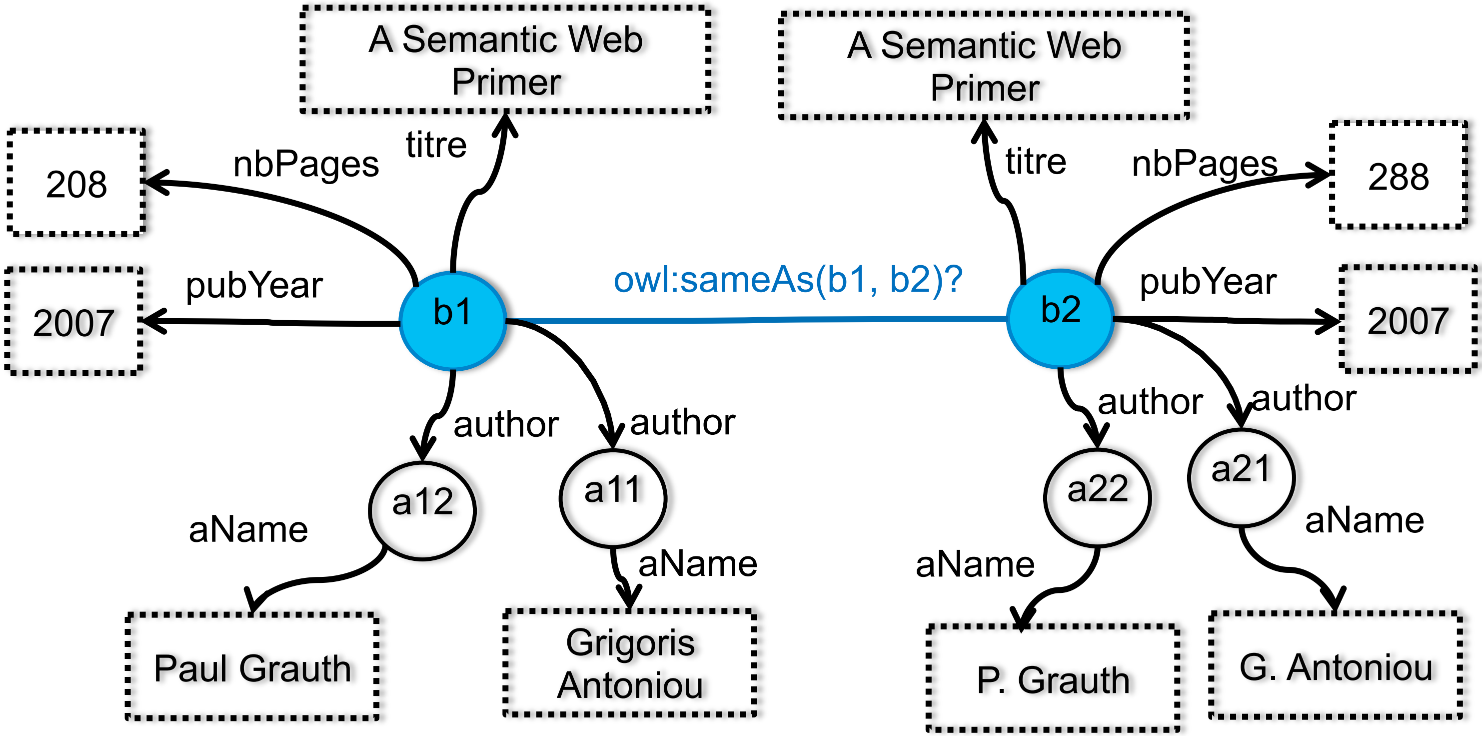
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

1. DETECTION OF ERRONEOUS IDENTITY LINKS

Content

Which kind of information to use for detecting erroneous Identity links?

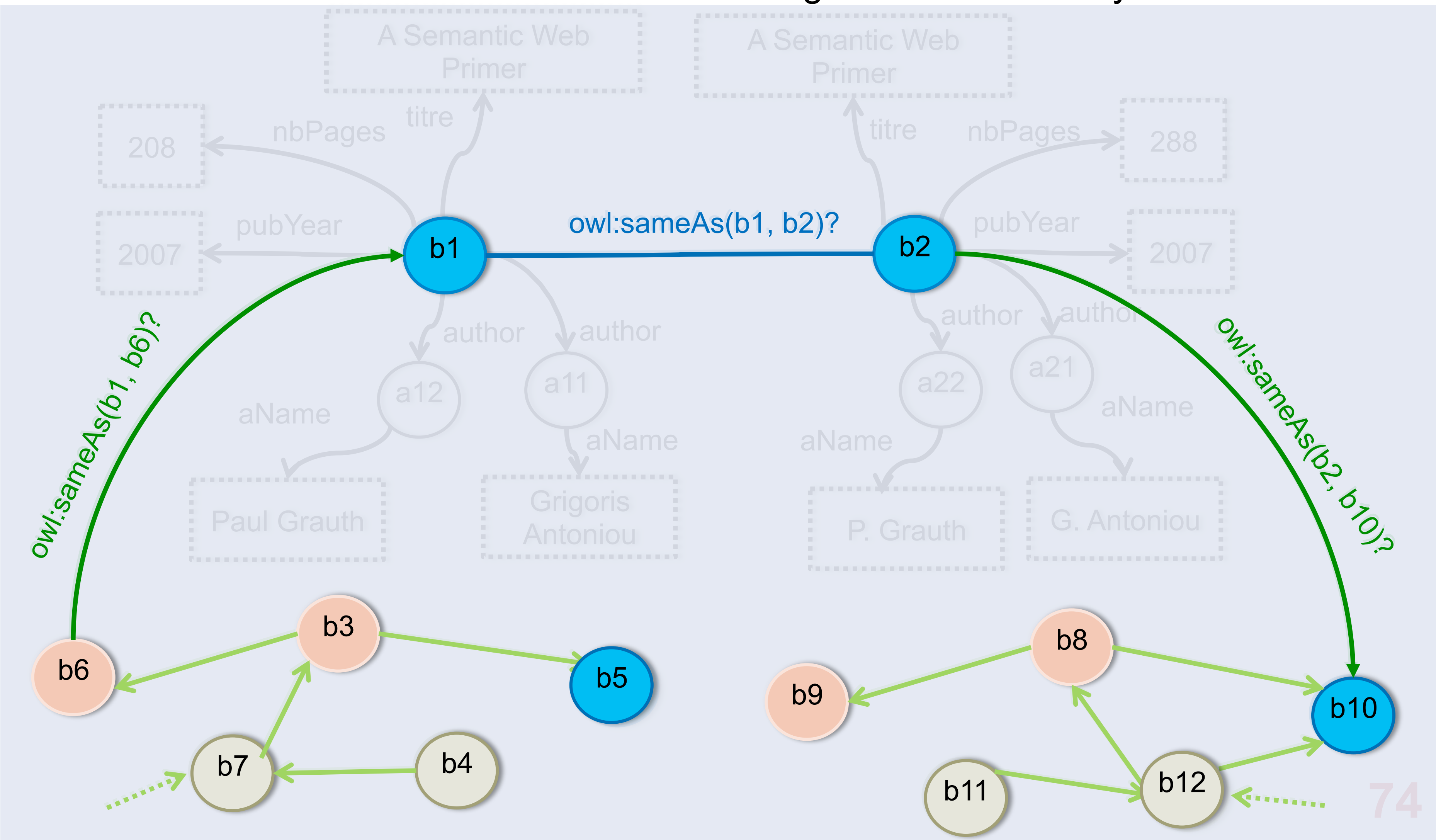


1. DETECTION OF ERRONEOUS IDENTITY LINKS

Content

Identity Network

Which kind of information to use for detecting erroneous Identity links?



1. DETECTION OF ERRONEOUS IDENTITY LINKS

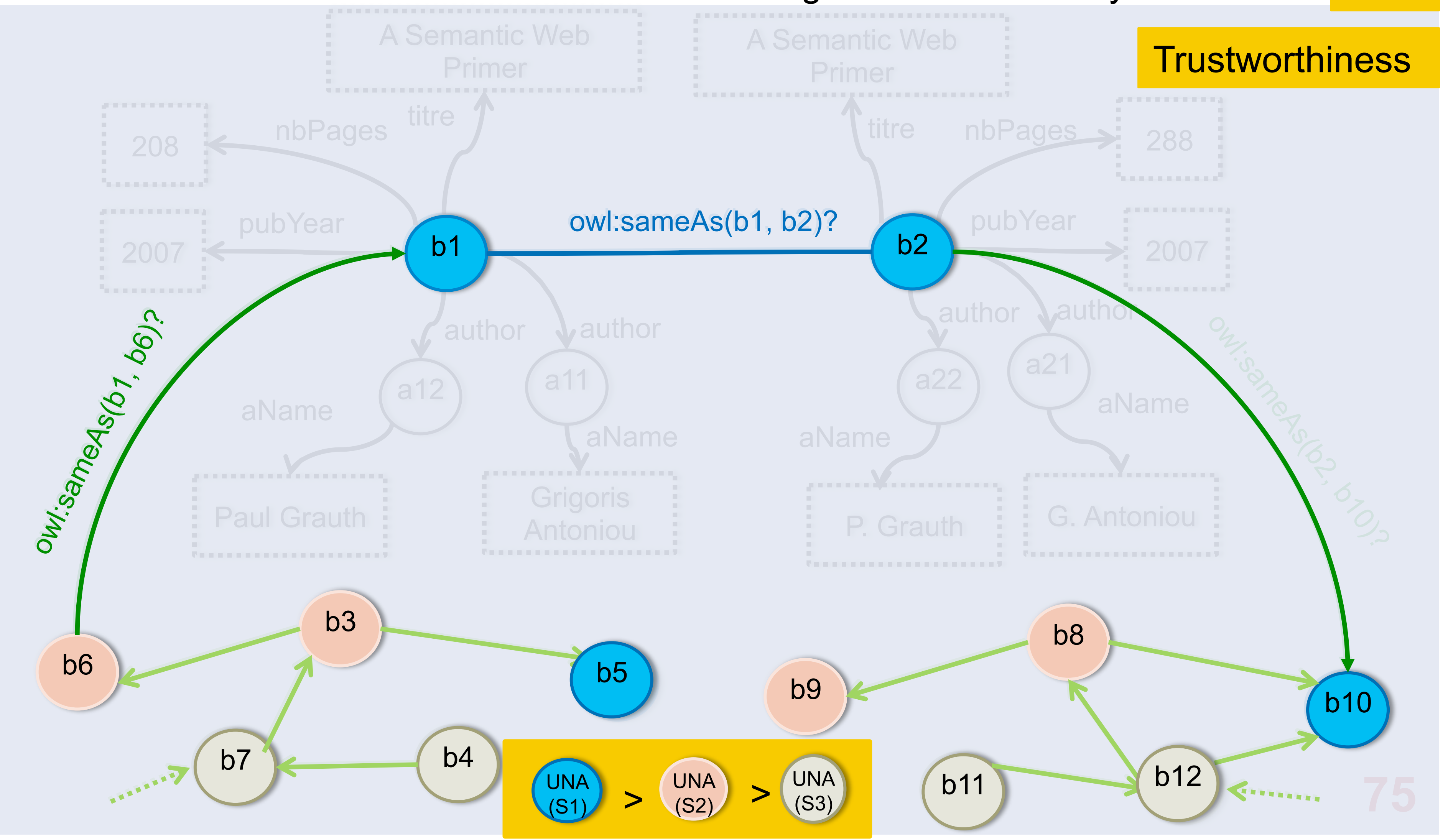
Content

Identity Network

UNA

Trustworthiness

Which kind of information to use for detecting erroneous Identity links?



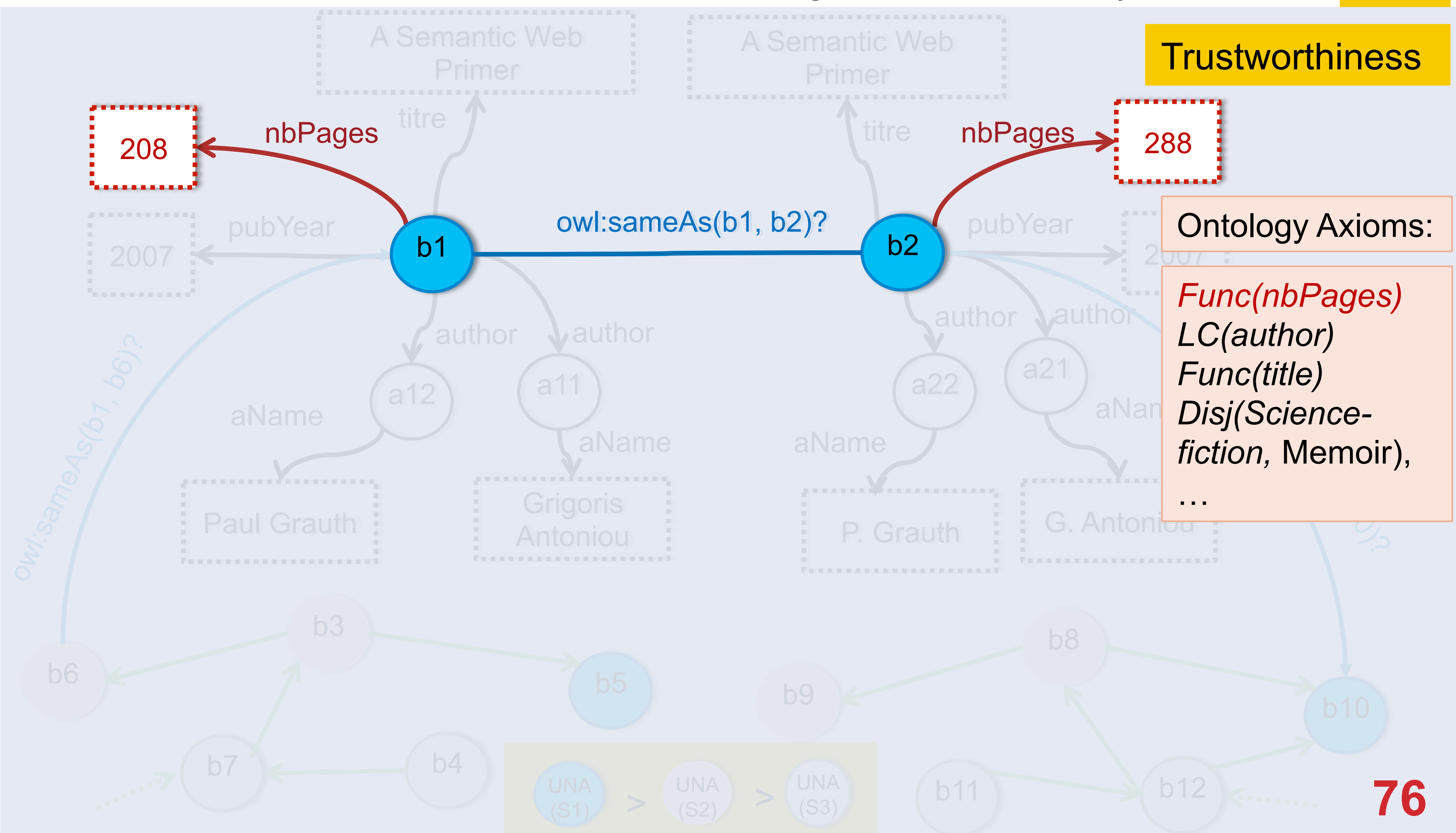
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Content

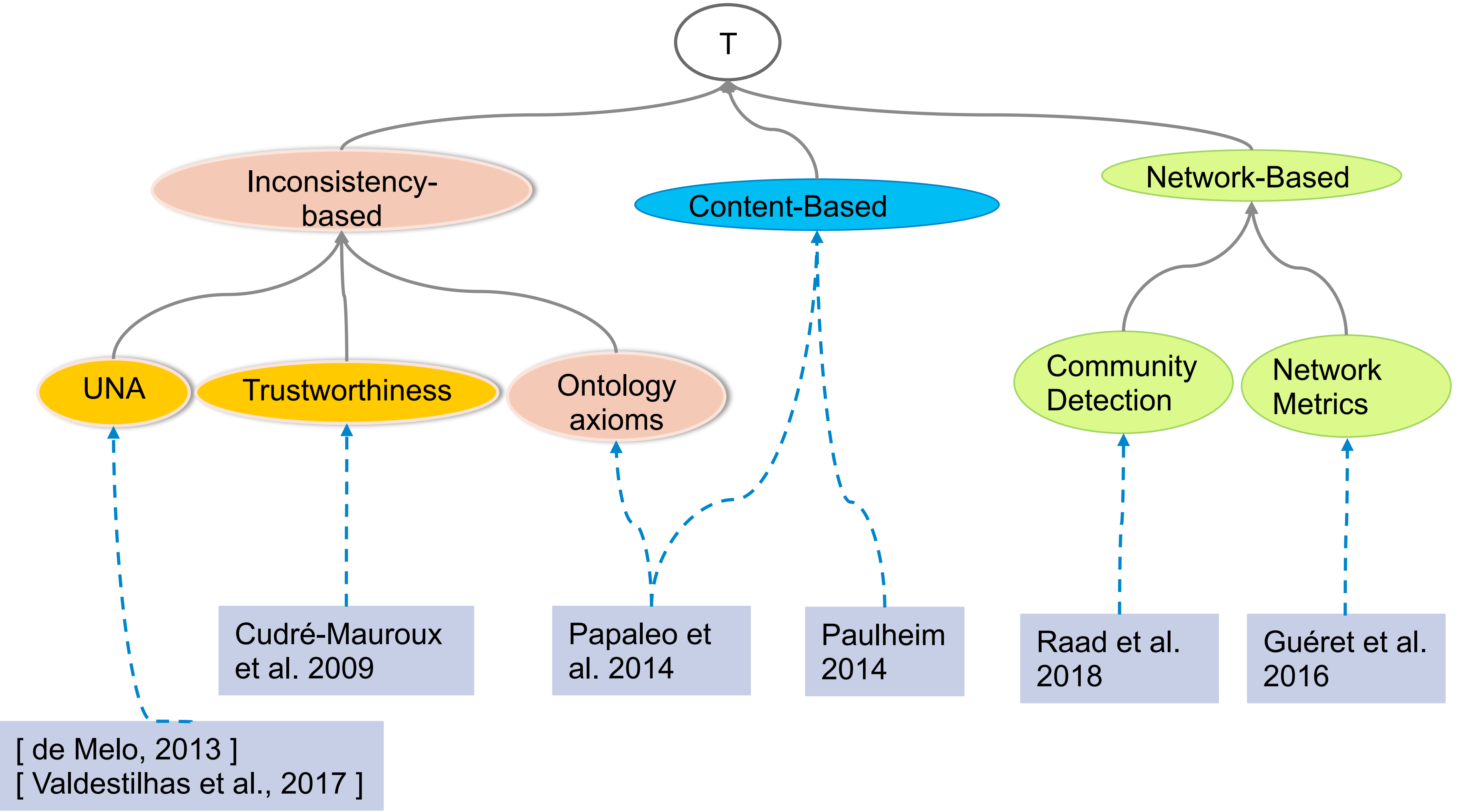
Identity Network

UNA

Which kind of information to use for detecting erroneous Identity links?

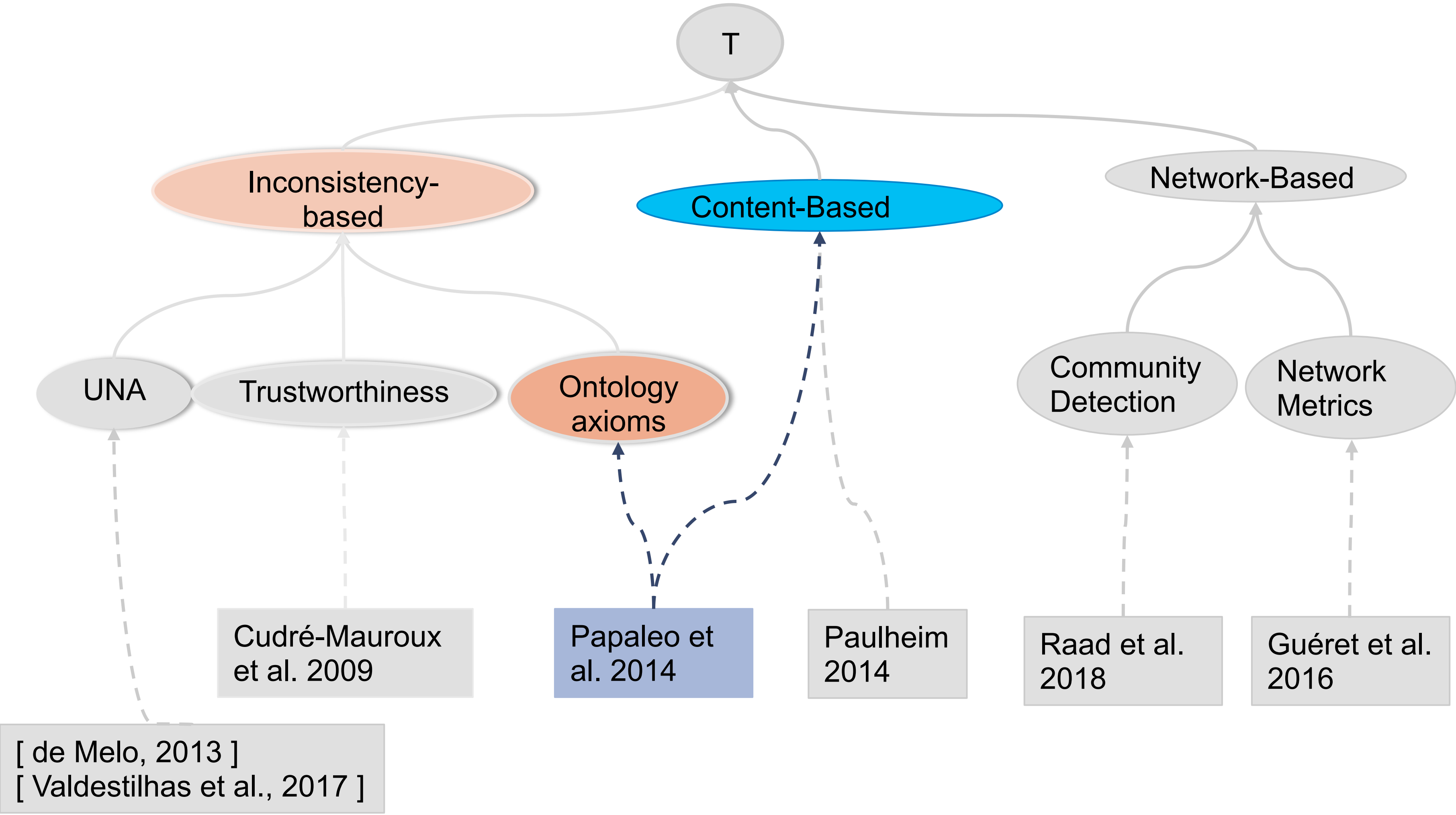


1. DETECTION OF ERRONEOUS IDENTITY LINKS



INCONSISTENCY-BASED

1. DETECTION OF ERRONEOUS IDENTITY LINKS

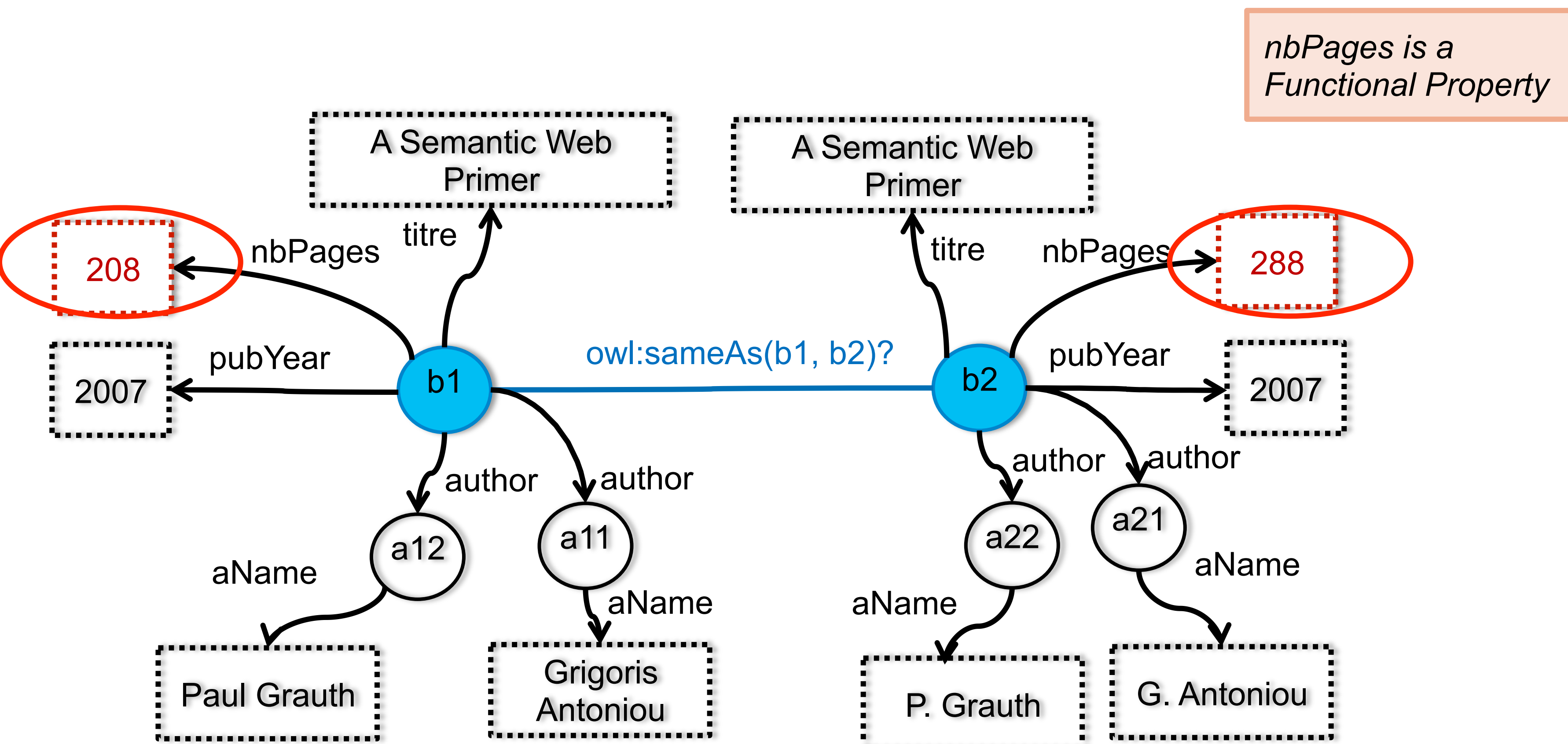


INCONSISTENCY-BASED AND CONTENT-BASED

[Papaleo *et al.*, 2014]
[Hogan et al. 2012]

ONTOLOGY AXIOM VIOLATION

Principle: use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies and possible errors in the linked resources.



ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]

- A logical **ontology-based method** to detect invalid sameAs statements
- Builds a contextual graph «around» each one of the two resources involved in the sameAs by exploiting ontology axioms on:
  - functionality and inverse functionality of properties and
  - local completeness of some properties, e.g., the author list of a book.
- Exploit the descriptions provided in these contextual graphs to eventually detect inconsistencies or high dissimilarities.

ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]

Apply Unit Resolution on {F ∪ R}.
[*F set of facts, R set of rules*]

R the set of rules

(inverse) functional properties

- $R1_{FDP}: sameAs(x,y) \wedge p_i(x,w_1) \wedge p_i(y,w_2) \rightarrow synVals(w_1,w_2)$
- $R2_{FOP}: sameAs(x,y) \wedge p_j(x,w_1) \wedge p_j(y,w_2) \rightarrow sameAs(w_1,w_2)$
- $R3_{IFP}: sameAs(x,y) \wedge p_k(w_1,x) \wedge p_k(w_2,y) \rightarrow sameAs(w_1,w_2)$

sameAs(x,y) ∧ nbPages(x,w_1) ∧ nbPages(y,w_2) → SynVals(w_1,w_2)

local complete properties

- $R4_{LC}: sameAs(x,y) \wedge p(x,w_1) \rightarrow p(y,w_1)$

sameAs(x,y) ∧ hasAuthor(x,w_1) → hasAuthor(y,w_1)

ONTOLOGY AXIOM VIOLATION

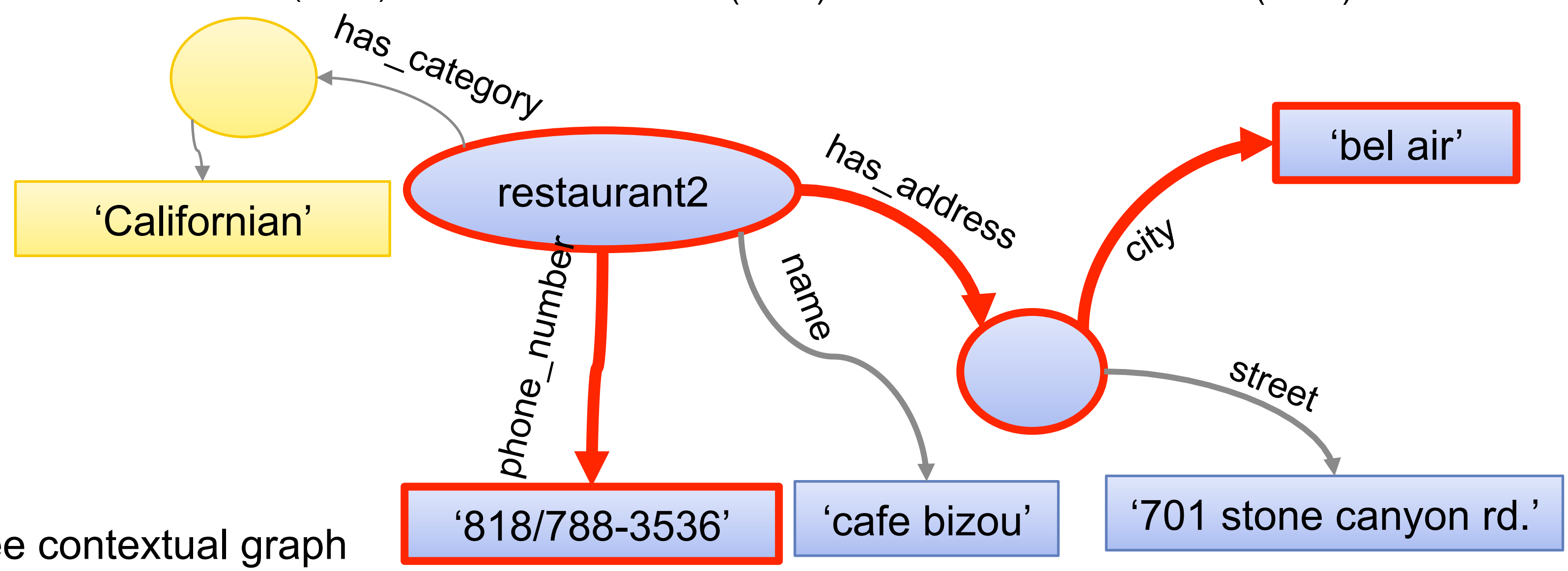
[Papaleo et al. 2014]

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].

[1] Saïs et al.: *LN2R a knowledge based reference reconciliation system: OAEI2010 results*. (2010)

[2] Symeonidou et al.: *SAKey: Scalable Almost Key Discovery in RDF Data*. (2014)

[3] Yves et al.: *Ontology matching with semantic verification*. (2009)



2-degree contextual graph
phone_number, hasAddress & city
(possible synvals computation)

ONTOLOGY AXIOM VIOLATION

[Papaleo et al. 2014]

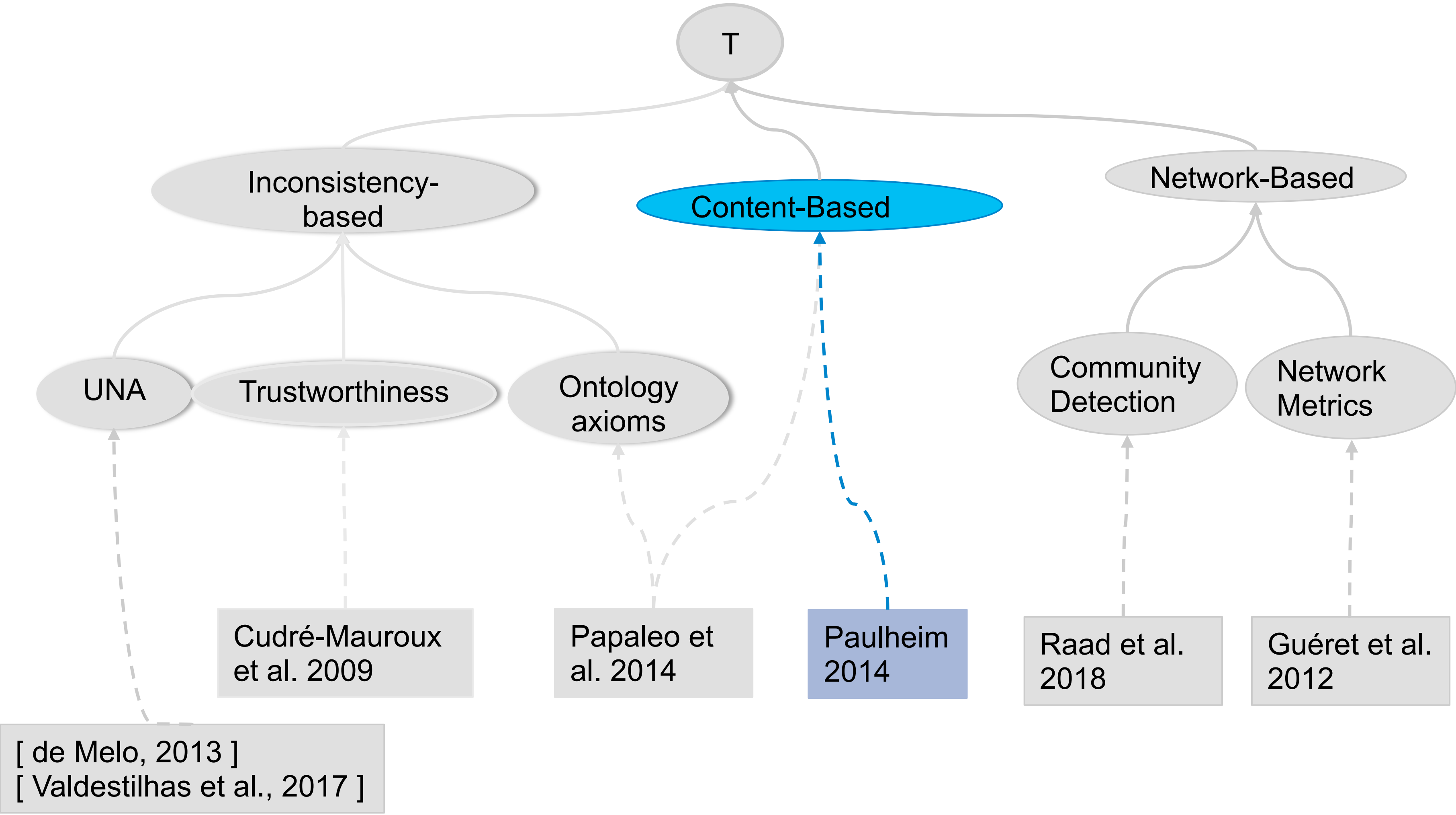
- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].

Linking Method	LM Precision	IM Recall	IM Precision	IM Accuracy	LM+IM precision
[120]	95.55%	75%	37%	93.34%	98.85%
[110]	69.71%	88.4%	88.4%	92.9%	95.19%
[138]	90.17%	100%	42.30%	86.60%	100%

IM: Invalidation method
LM: Linking method

Improvement in precision

1. DETECTION OF ERRONEOUS IDENTITY LINKS

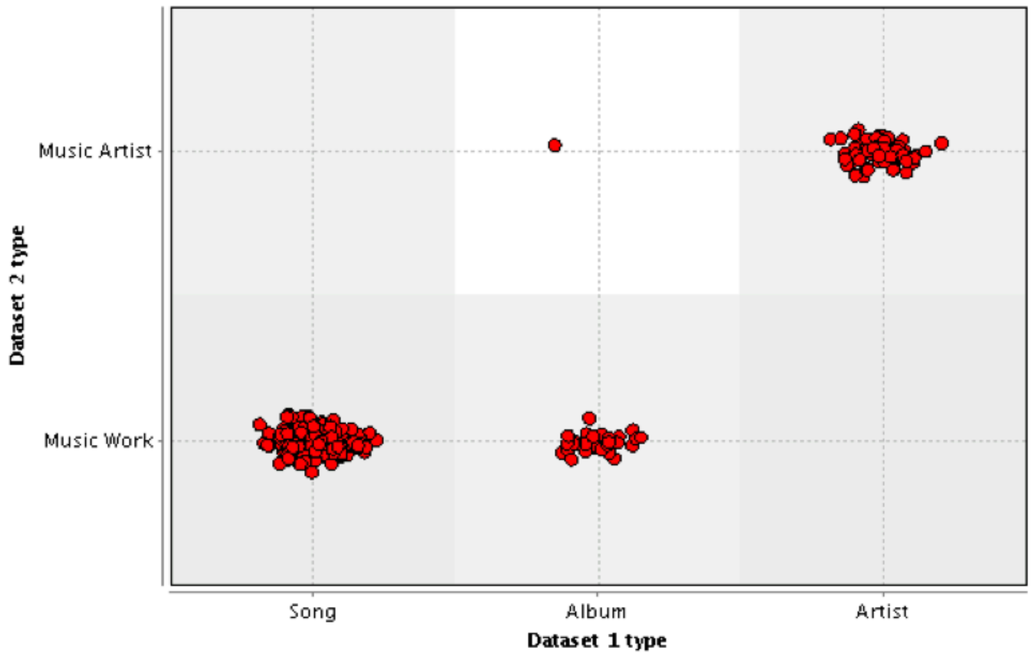


CONTENT BASED

[Paulheim, 2014]

Principle: links follow some patterns, links that violate those patterns are erroneous.

- A multi-dimensional and scalable **outlier detection** approach for finding **erroneous identity links**.
- Projection of links into **Vector Space**: each link is a point in an n-dimensional vector space

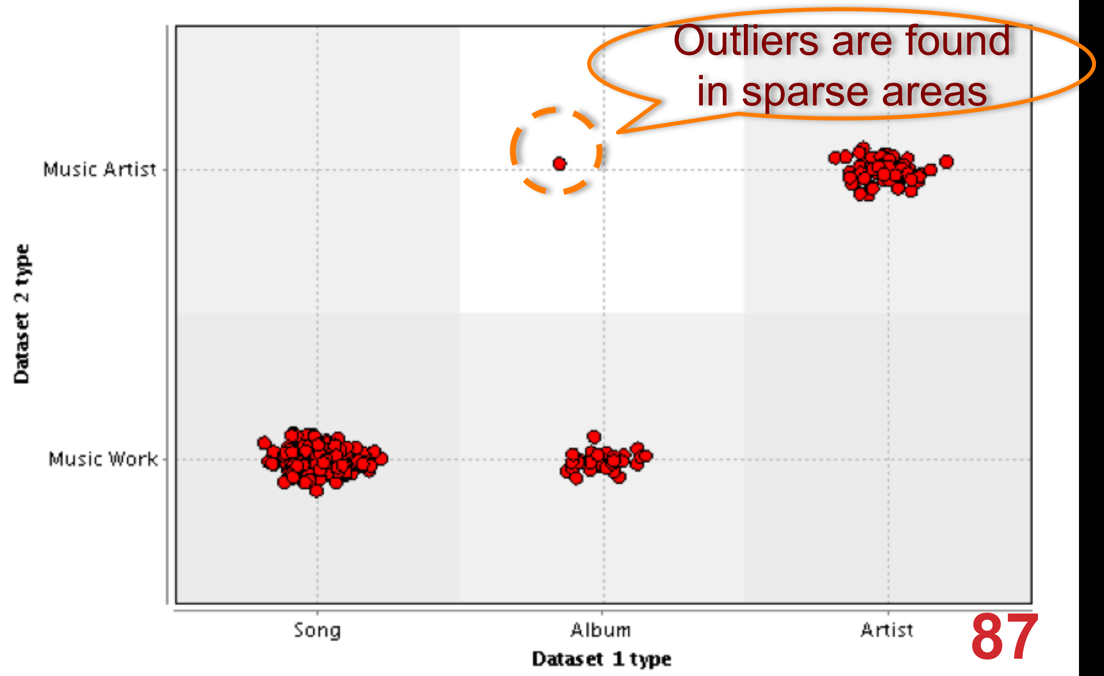


CONTENT BASED

[Paulheim, 2014]

Principle: links follow some patterns, links that violate those patterns are erroneous.

- A multi-dimensional and scalable **outlier detection** approach for finding **erroneous identity links**.
- Projection of links into **Vector Space**: each link is a point in an n-dimensional vector space



CONTENT BASED

[Paulheim, 2014]

- **Feature Vector**: resource types and ingoing/outgoing properties
  - e.g. LHS_foaf:based_near and RHS_foaf:based_near are distinct features.

- **Different strategies of creating vectors**: direct types only, all ingoing and outgoing properties, or a combination
- Several outlier detection methods were tested: LOF, CBLOF, LOP, 1-class SVM etc.
- Each method assign a score to each data point indicating the likeliness of being an outlier ➔ **incorrect link**.

CONTENT BASED

[Paulheim, 2014]

- **Dataset**

	D1		D2	
Dataset	Peel Session	DBpedia	DBTropes	DBpedia
# Links	2,087		4,229	
# Types	3	31	2	79
# Properties	4	56	18	124

- **Gold Standard**: 100 randomly sampled links from D1 and D2
- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)

CONTENT BASED

[Paulheim, 2014]

- **Dataset**

	D1		D2	
Dataset	Peel Session	DBpedia	DBTropes	DBpedia
# Links	2,087		4,229	
# Types	3	31	2	79
# Properties	4	56	18	124

- **Gold Standard**: 100 randomly sampled links from D1 and D2
- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)
- **Best performance on D1:**
  - 1-class SVM (**AUC = 0.857, F1= 0.471**)
- **Best performance on D2:**
  - LOF (**AUC = 0.619, F1= 0.5**)

CONTENT BASED

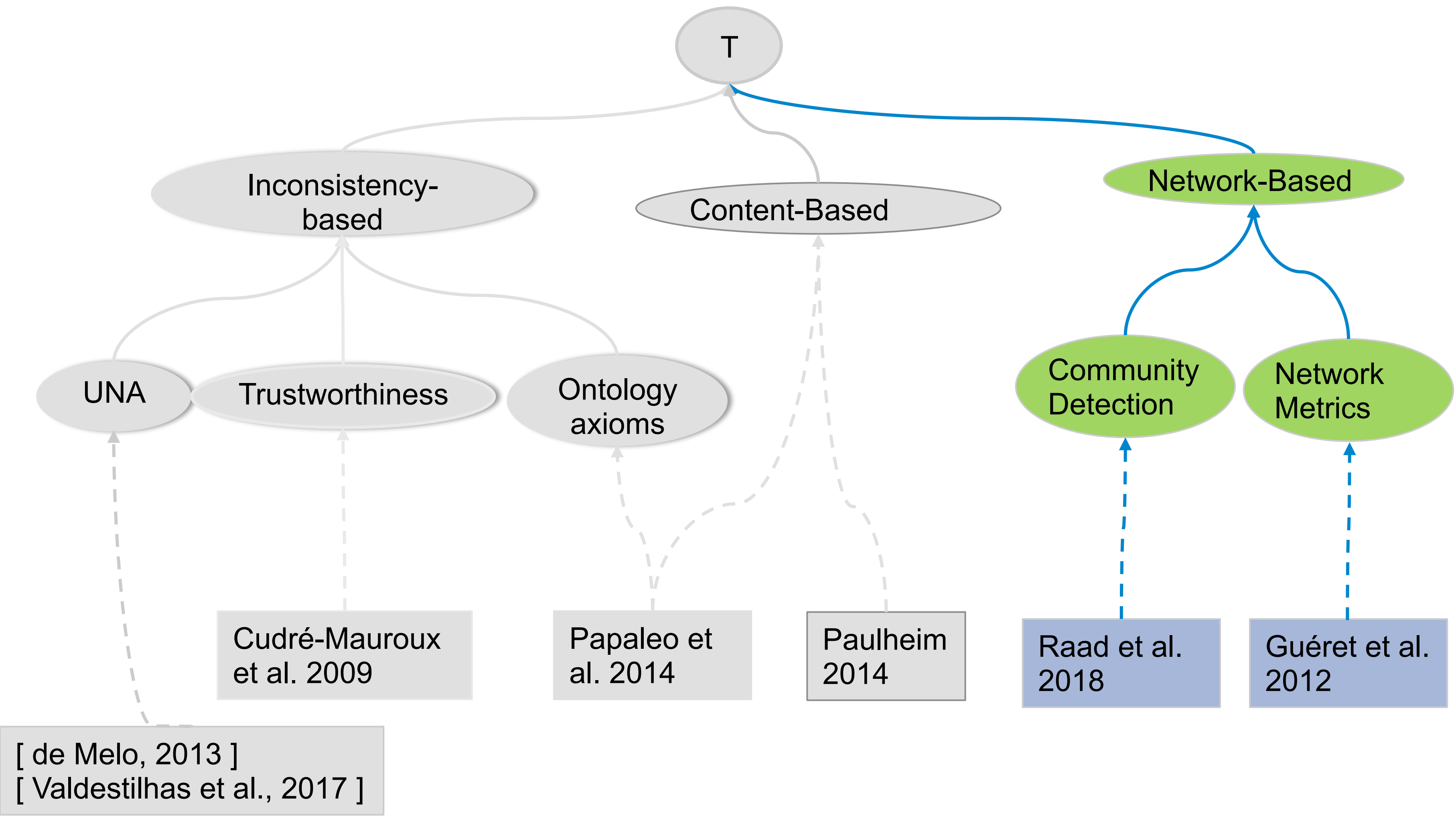
[Paulheim, 2014]

- **Dataset**

	D1		D2	
Dataset	Peel Session	DBpedia	DBTropes	DBpedia
# Links	2,087		4,229	
# Types	3	31	2	79
# Properties	4	56	18	124

- **Gold Standard**: 100 randomly sampled links from D1 and D2
- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)
- **Best performance on D1:**
  - 1-class SVM (**AUC = 0.857, F1= 0.471**)
- **Best performance on D2:**
  - LOF (**AUC = 0.619, F1= 0.5**)
- Examples of **typical source of errors** for D1:
  - Linking of songs to albums with the same name.
  - Linking of different persons of the same name.
    e.g., a blues musician named Jimmy Carter to the U.S. president.

1. DETECTION OF ERRONEOUS IDENTITY LINKS



NETWORK BASED

[Guéret *et al.*, 2012]

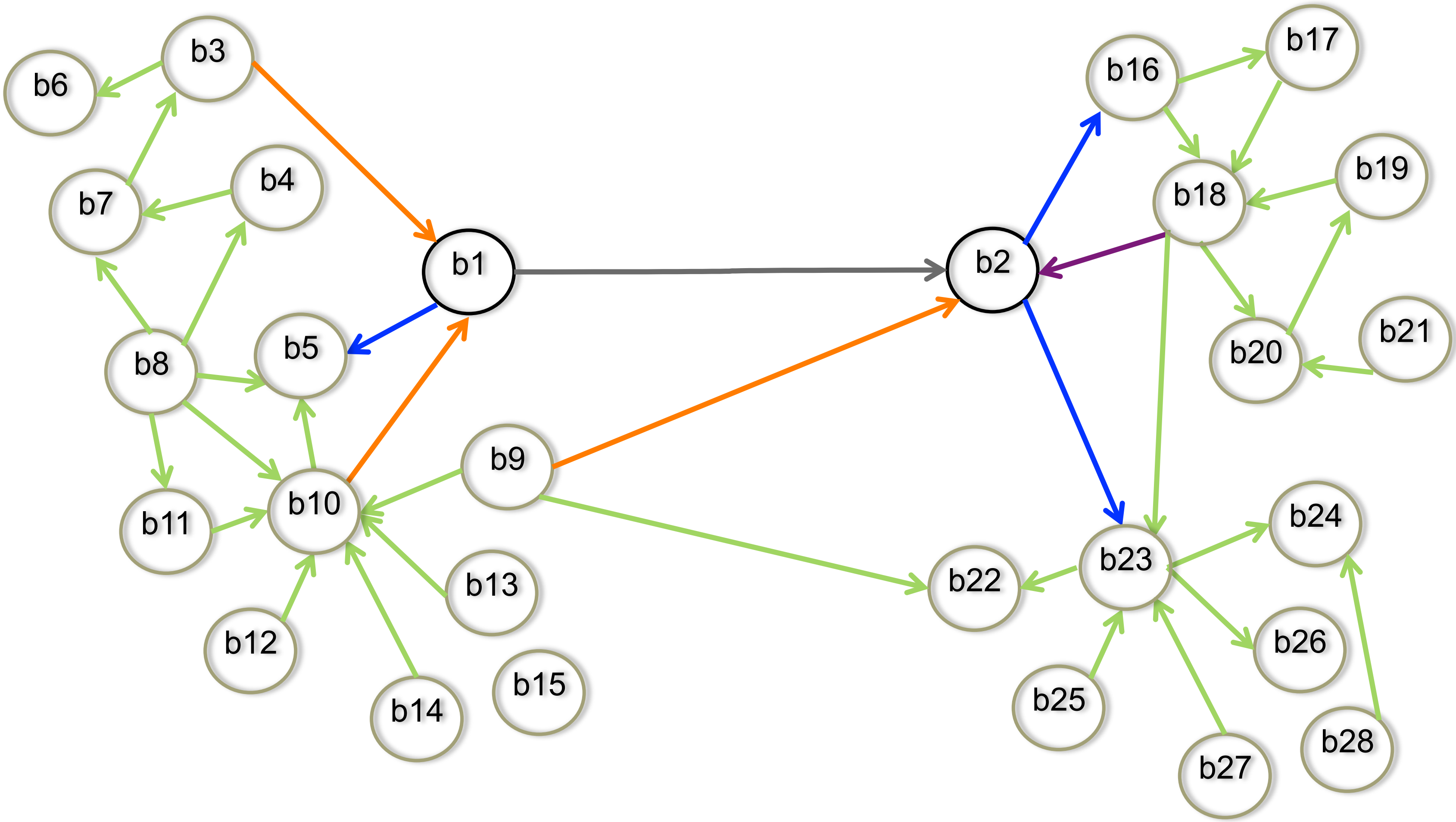
[Raad *et al.*, 2018]

Principle

- The quality of a link can be evaluated based on **how connected a node** is within the **network** (data graph, sameAs network) in which it appears.
- How **network metrics** can help to detect erroneous links?

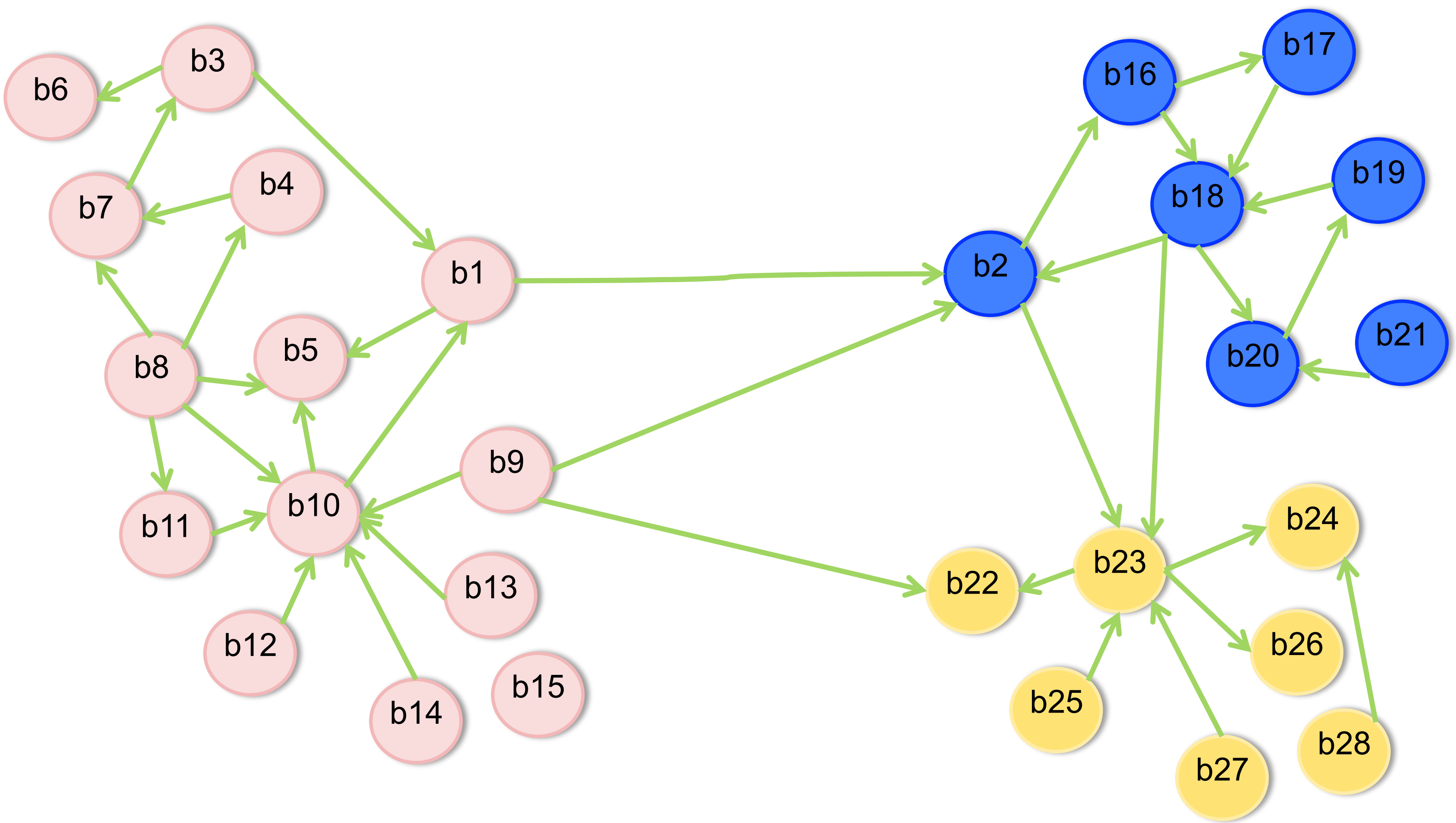
NETWORK BASED

Node in-degree and out-degree, Centrality, Clustering coefficient ….



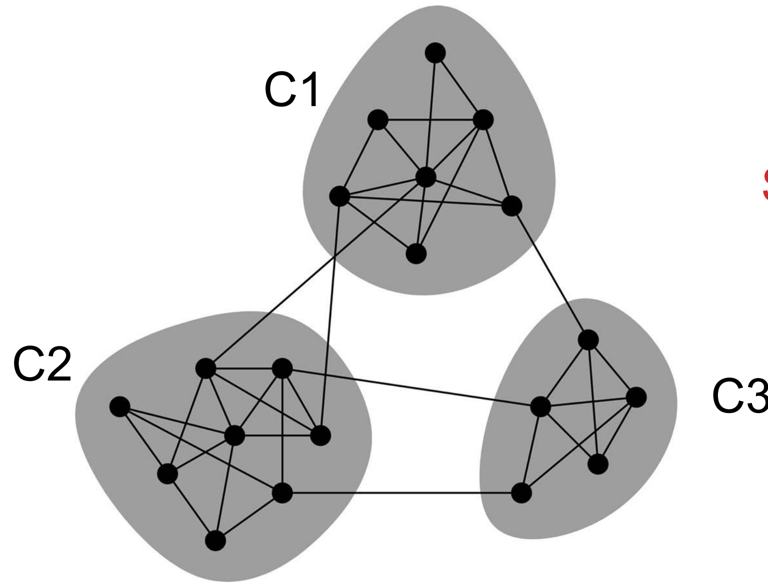
NETWORK BASED

Density of detected community structures



NETWORK BASED

[Raad *et al.*, 2018]



SameAs network

Overall idea

Use the community structure of the network containing solely owl:sameAs links to assign error degree for each link.

NETWORK BASED

[Raad *et al.*, 2018]

4 main steps

Step 1: Extraction of the explicit identity statements

Step 2: Partition into equality sets (theoritically…the same entity)

Step 3: Detection of the community structure of each equality set using the Louvain algorithm [Blondel et al. 2008]

Step 4: Assignation of an error degree to each sameAs

NETWORK BASED

[Raad *et al.*, 2018]

Error degree based on the weight (w=2 when the sameAs is symmetric) and on the density of the involved community(ies)

intra-community link

$$a)\ err(e_C) = \frac{1}{w(e_C)} \times \left(1 - \frac{W_C}{|C| \times (|C| - 1)}\right)$$

inter-community link

$$b)\ err(e_{C_{ij}}) = \frac{1}{w(e_{C_{ij}})} \times \left(1 - \frac{W_{C_{ij}}}{2 \times |C_i| \times |C_j|}\right)$$

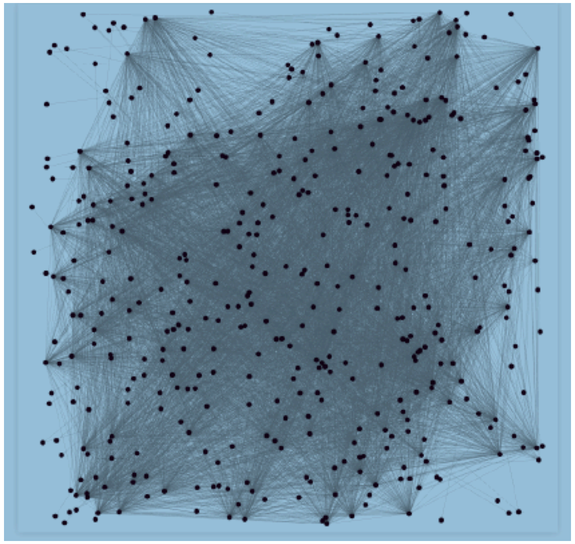
Correct link 0 Error degree 1 Erroneous link

NETWORK BASED

[Raad *et al.*, 2018]

Experimentation - Dataset

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of **28B** triples from LOD 2015 crawl
- Step 1: extraction of an **explicit identity network** of 558.9M sameAs links (179M nodes)
- Step 2: Partitionned in 48.9M of non singleton **equality sets**

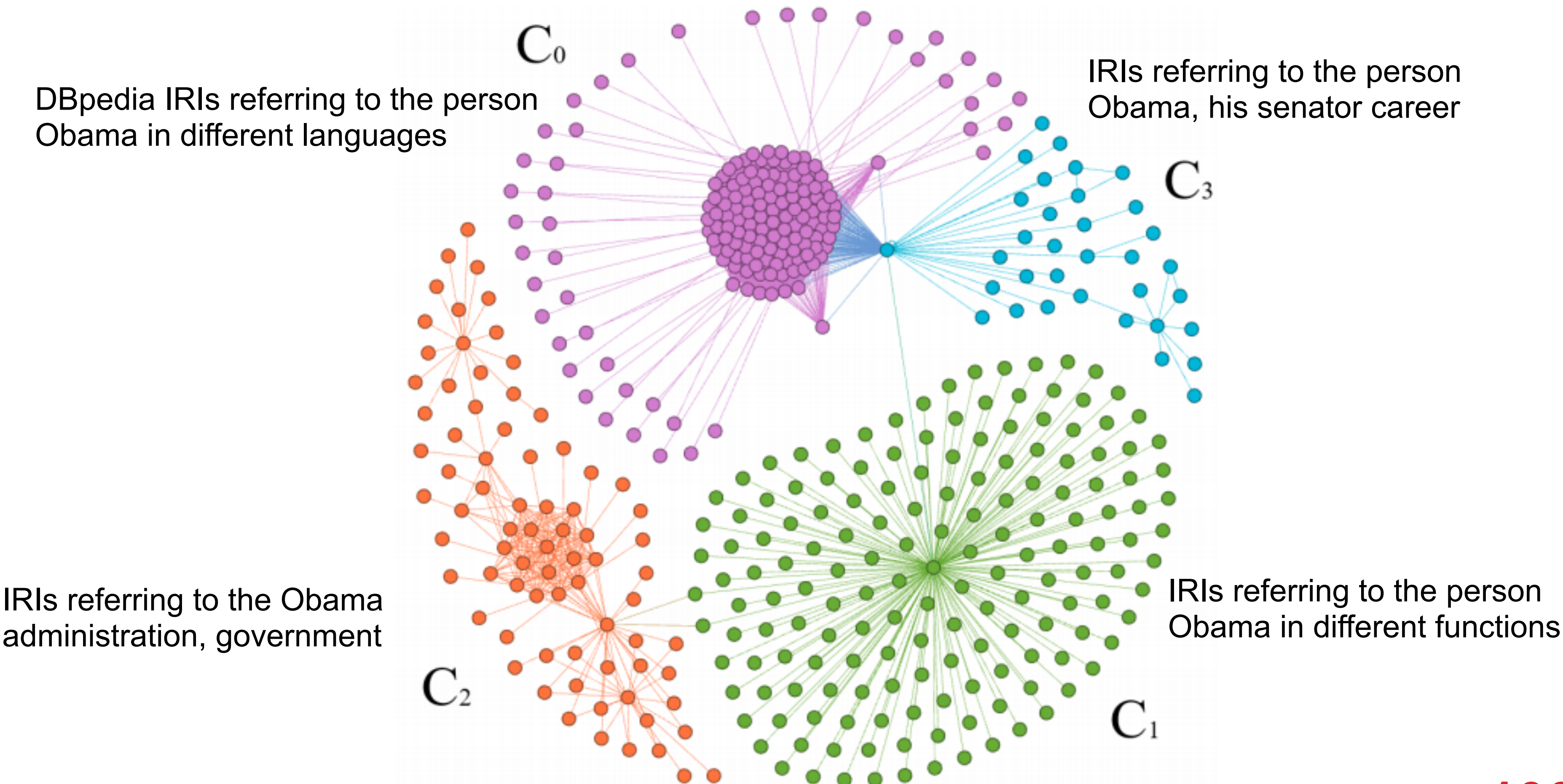


Example: The *B. Obama* equality set which contains 440 nodes

NETWORK BASED

[Raad *et al.*, 2018]

Step 3: Detection of the community structure for each equality set
Example: the community structure of the *Barack Obama's* Equality Set



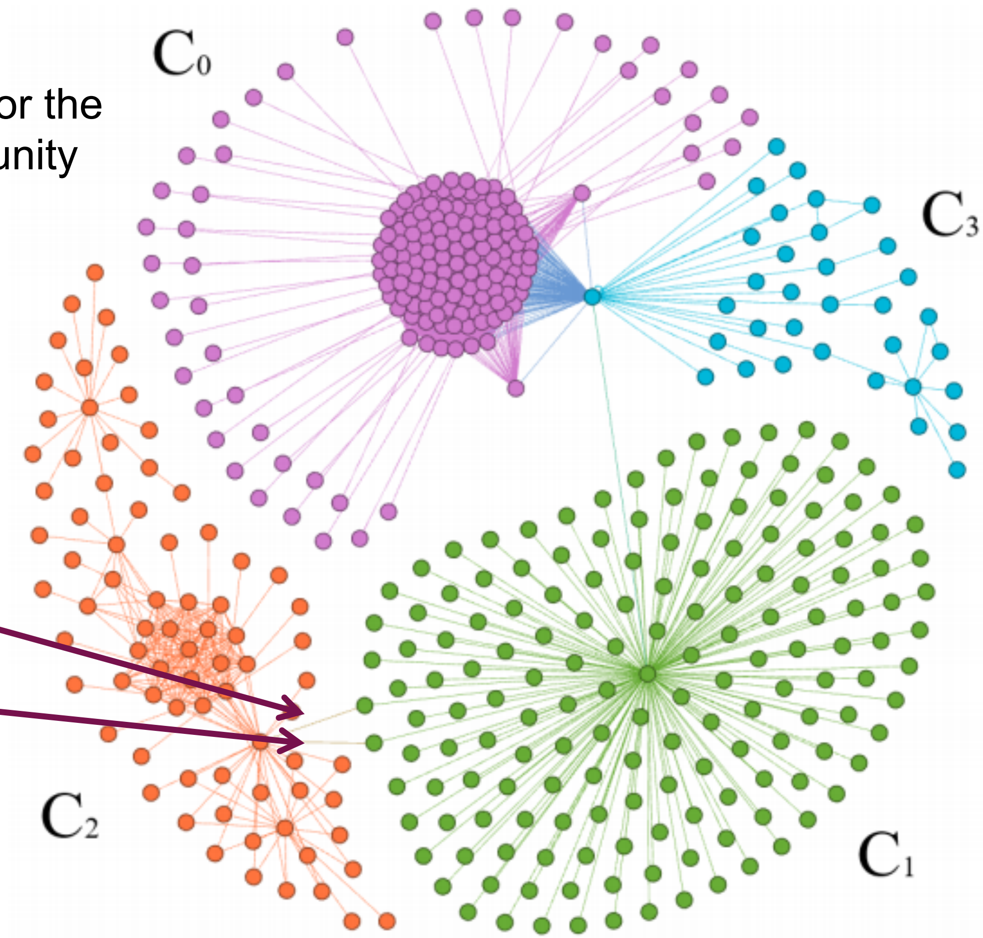
NETWORK BASED

[Raad *et al.*, 2018]

Step 4: Computation of the error degrees

Low error degrees for the links of this community

err(e)= 1
For these 2 links



NETWORK BASED

[Raad *et al.*, 2018,]

- **Scales up** to a graph of **28 billion** triples: **11 hours for the 4 steps**
- **Finding the threshold: manual evaluation of 200 randomly chosen links**

	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	total
same	35(100%)	22(100%)	18(85.7%)	7(77.7%)	15(68.1%)	**97**(88.9%)
related	0	0	2	2	2	**6**
unrelated	0	0	1	0	5	**6**
related + unrelated	0(0%)	0(0%)	3(14.2%)	2(22.2%)	7(31.8%)	**12**(11%)
can't tell	5	18	19	31	18	**91**
Total	**40**	**40**	**40**	**40**	**40**	**200**

The higher an error degree is the most likely a link is erroneous

- 100% of owl:sameAs with an **error degree <0.4** are correct

- Can theoretically **invalidate a large set of owl:sameAs links** on the LOD:
  - **1.26M** owl:sameAs have an **error degree** in [0.99, 1]

ERRONEOUS LINK DETECTION: SUMMARY

- **Different approaches**: consistency-based, content-based or network-based relaying on different kinds of information (UNA, axioms, mappings, textual values/types/properties or network metrics)
- Some approaches are global (collective), some are instance-based (pairs of resources are considered independently).

ERRONEOUS LINK DETECTION: SUMMARY

- **Different approaches**: consistency-based, content-based or network-based relaying on different kinds of information (UNA, axioms, mappings, textual values/types/properties or network metrics)
- Some approaches are global (collective), some are instance-based (pairs of resources are considered independently).

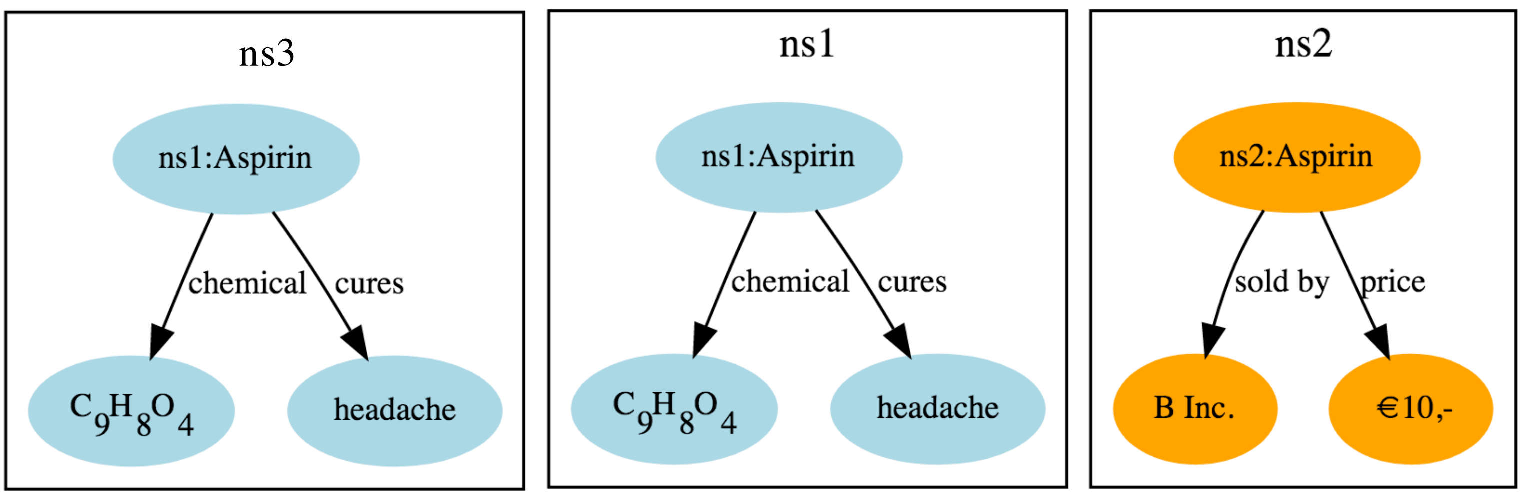
Limitations

- **Evaluation** are often conducted on few links, on specific datasets
- Some **assumptions** cannot be made on the LOD:
  - UNA is not always fulfilled
  - Ontology and Ontology axioms are not always available
  - Differences are rarely available: useful for inconsistency checking
  - Network-based approaches do not need such assumptions + scalable but they cannot decide for small equality sets, and higher precision is needed.
- Need of alternate links

CONTEXTUAL IDENTITY LINKS

CONTEXTUAL IDENTITY LINKS

- **Weaker** kinds of **identity** can be expressed by considering a **subset of properties** with respect to which two resources can be considered to be the same.
- Identity is **context-dependent** [Geach, 1967]
  - *allowing two medicines to be considered the same in terms of their chemical substance, but different in terms of their price (e.g., because they* are produced by different companies).



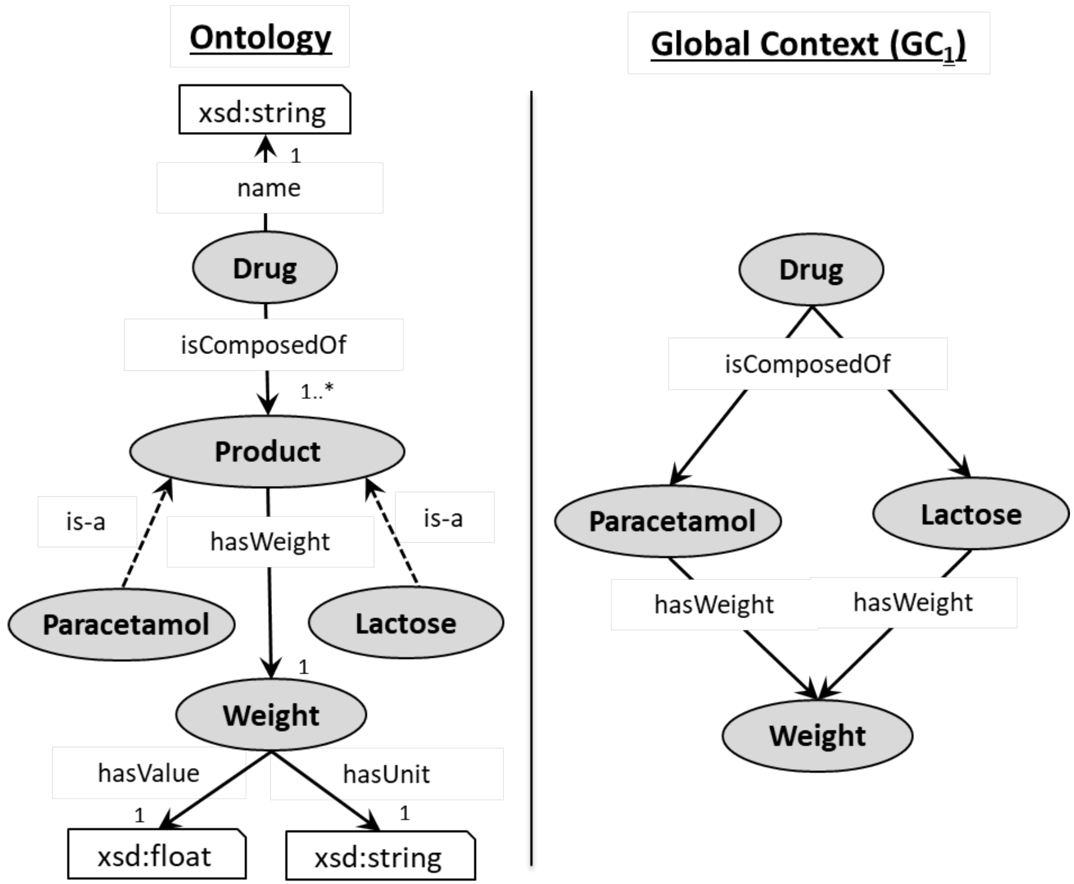
CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

- New **contextual identity** relation
- An **algorithm** for automatic detection of the **most specific contexts** in which two instances (resources) are identical
  - the detection process can further be guided by a set of **semantic constraints** that are provided by domain experts.

- Contexts are defined as a sub-ontology of the domain ontology
- All the possible contexts are organized in a lattice using an order relation.

CONTEXTUAL IDENTITY LINKS



$$\begin{aligned}
&GC_1 = (C = \{Drug, Paracetamol, Lactose, Weight\},\\
&OP = \{isComposedOf, hasWeight\}, DP = \emptyset,\\
&A = \{domain(isComposedOf) = Drug,\\
&range(isComposedOf) = Lactose \sqcup Paracetamol,\\
&domain(hasWeight) = Lactose \sqcup Paracetamol,\\
&range(hasWeight) = Weight\})
\end{aligned}$$

CONTEXTUAL IDENTITY LINKS

$$GC_1 = (C = \{Drug, Paracetamol, Lactose, Weight\},$$
$$OP = \{isComposedOf, hasWeight\}, DP = \emptyset,$$
$$A = \{domain(isComposedOf) = Drug,$$
$$range(isComposedOf) = Lactose \sqcup Paracetamol,$$
$$domain(hasWeight) = Lactose \sqcup Paracetamol,$$
$$range(hasWeight) = Weight\})$$

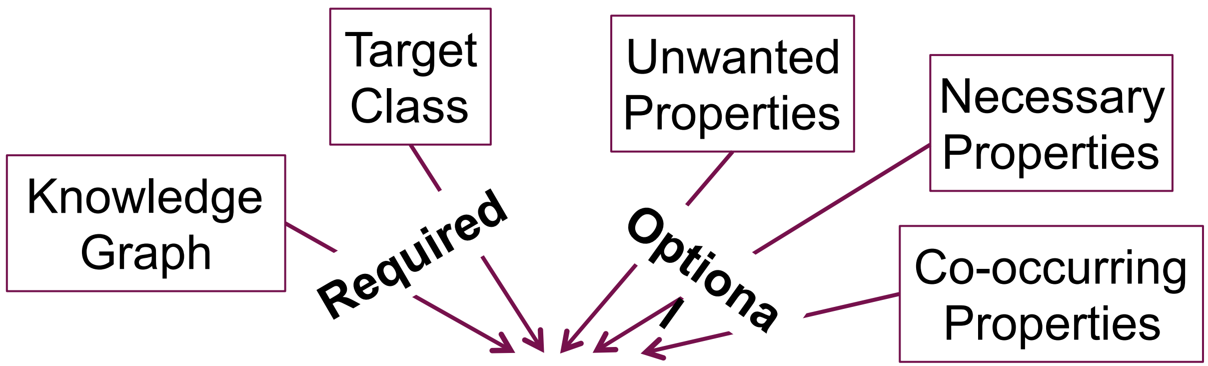
$$\leq$$

(more specific than)

$$GC_2 = (C = \{Drug, Lactose\},$$
$$OP = \{isComposedOf, DP = \{\emptyset\},$$
$$A = \{domain(isComposedOf) = Drug,$$
$$range(isComposedOf) = Lactose\})$$

CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]



DECIDE

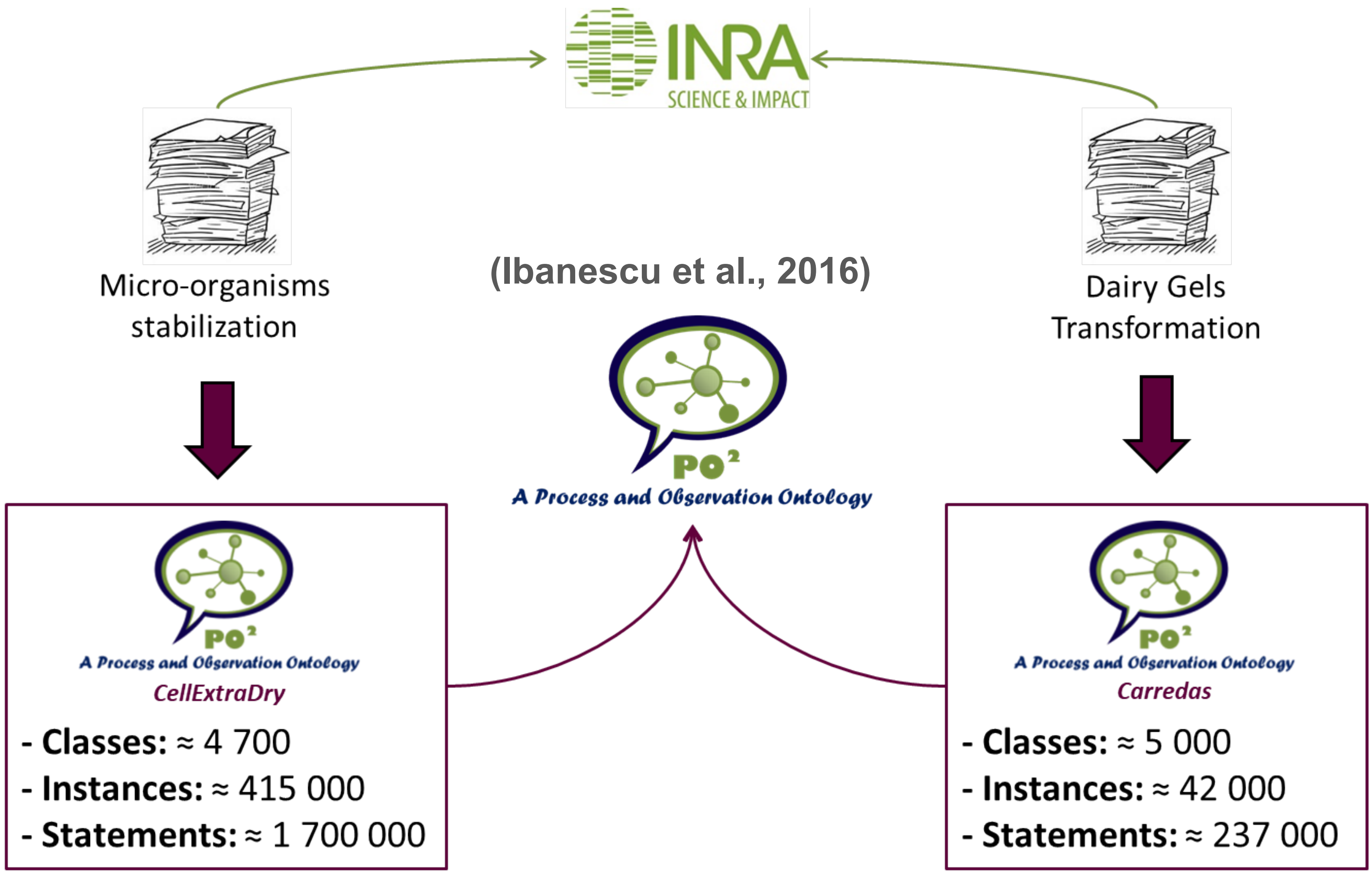
DEtection of **C**ontextual **IDE**ntity

For each pair of individuals (i_1, i_2) of the target class

set of the most specific global contexts in which (i_1, i_2) are identical

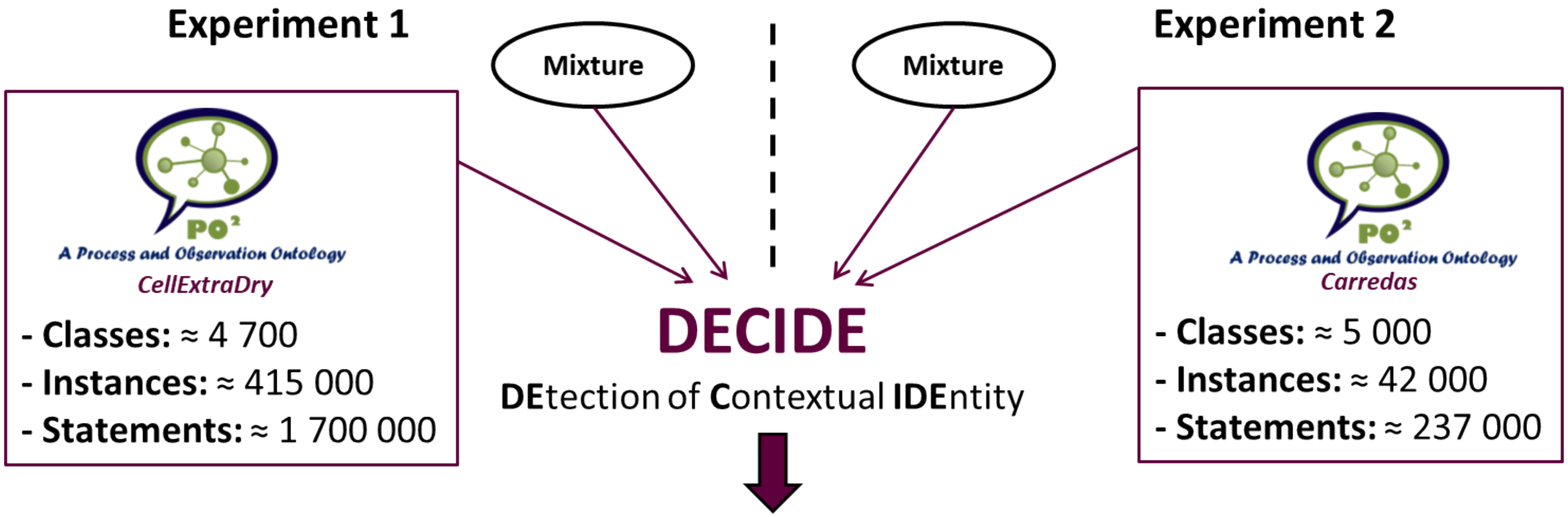
CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]



CONTEXTUAL IDENTITY LINKS

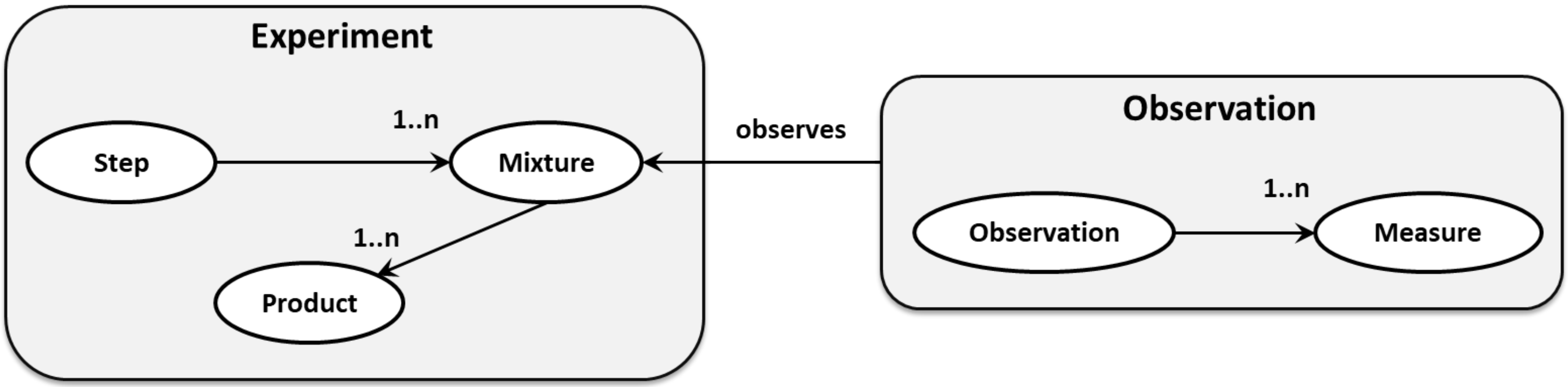
[Raad et al., 2017]



	CellExtraDry	Carredas
# Instances (type:Mixture)	210	619
# Possible Pairs	21 945	191 271
# Dependant Classes (Total Classes)	191 (208)	488 (555)
# Graph Nodes per pair	11	7
# Identity Links	33 092	239 410
# Identity Links per pair	1.41	1.25
# Different Global Contexts	28	233
Execution Time (approx. minutes)	2	26

CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]



Detect for each context $\mathbf{GC_i}$, the measures $\mathbf{m_i}$ where

$$\textit{identiConTo}_{<GCi>}(i_1, i_2) \cap \textit{observes}(i_1, m_1) \rightarrow \textit{observes}(i_2, m_2)$$

with $m_1 \simeq m_2$

$$\textit{identiConTo}_{<GCi>}(i1, i2) \rightarrow \text{same}(m_i)$$

CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Detection of 38 844 rules

Règle	*Taux d'erreur*	*Support*
$identiConTo_{<GC_1>}(x, y) \rightarrow$ same(pH)	6.19 %	57
$identiConTo_{<GC_3>}(x, y) \rightarrow$ same(Dureté)	1.86 %	66
$identiConTo_{<GC_2>}(x, y) \rightarrow$ same(Friabilité)	4.52 %	647

The domain experts have evaluated the plausibility of the best **20 rules** (in terms of error rate and support)

CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Detection of 38 844 rules

Règle	*Taux d'erreur*	*Support*
$identiConTo_{<GC_1>}(x, y) \rightarrow \text{same(pH)}$	6.19 %	57
$identiConTo_{<GC_3>}(x, y) \rightarrow \text{same(Dureté)}$	1.86 %	66
$identiConTo_{<GC_2>}(x, y) \rightarrow \text{same(Friabilité)}$	4.52 %	647

The domain experts have evaluated the plausibility of the best **20 rules** (in terms of error rate and support)

Impossible	Not very probable	Can't tell	Why not	Plausible
3	**5**	**4**	**5**	**3**

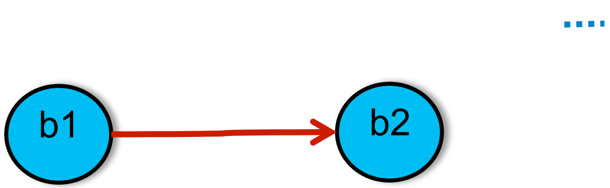
plausibility

The error rate decreases of 12% when a global context is replaced by a more specific global context

CONCLUSION AND FUTURE CHALLENGES

IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship

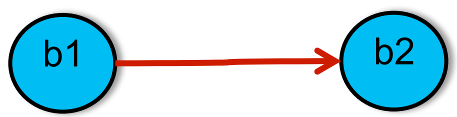


IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship

owl:sameAs — Genuine identity

lvont:somewhatSameAs — Near/weak identity

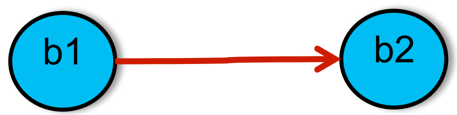


IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship

owl:sameAs → Genuine identity

lvont:somewhatSameAs → Near/weak identity



:sameBook

:sameExpression → Subjective identity

:sameArtWork

:dentiConTo$_{<\Pi a>}$ → Contextual identity

IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship
- Need of hybrid methods

Network Topology

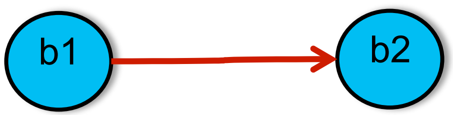
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

:dentiConTo$_{<\Pi a>}$

IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship
- Need of hybrid methods
- Link quality assessment is not a matter of one unique dimension

Network Topology

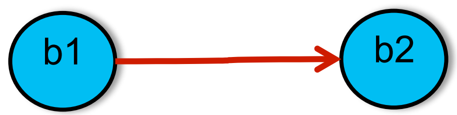
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

:dentiConTo$_{<\Pi a>}$

Link Validity: Inconsistent equivalent classes, Invalid links, Contextual links

Link Properties: Transitivity, symmetry, …

Link added-value: Information gain, reachability, …

Link meta-data: availability, evolution

IDENTITY MANAGEMENT: SUMMARY AND CHAL

What is about the **distinctness** relation?

- Different kinds of identity relationship
- Need of hybrid methods
- Link quality assessment is not a matter of one unique dimension

Network Topology

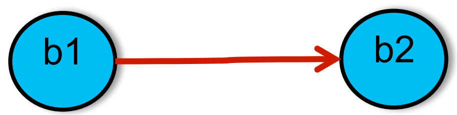
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

:dentiConTo$_{<\Pi a>}$

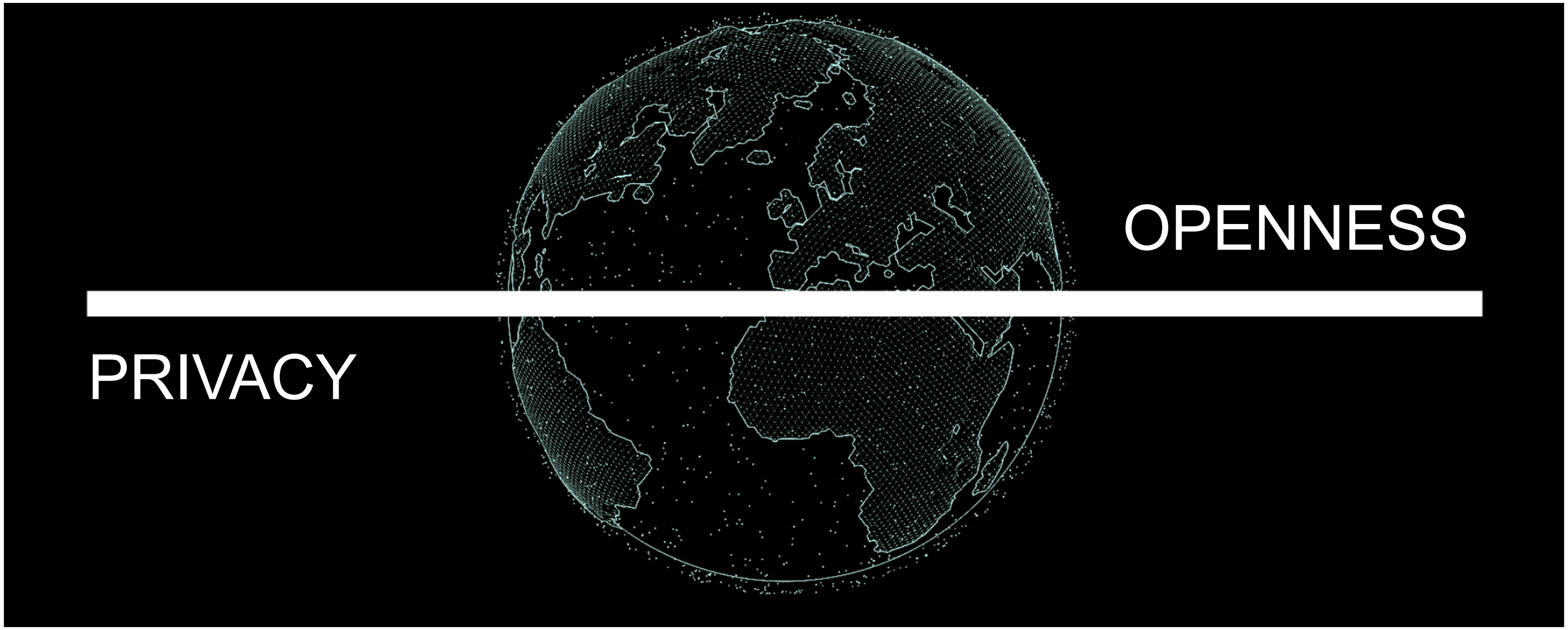
Link Validity: Inconsistent equivalent classes, Invalid links, Contextual links

Link Properties: Transitivity, symmetry, …

Link added-value: Information gain, reachability, …

Link meta-data: availability, evolution

OPENNESS AND PRIVACY BALANCE



[source1]

Mastering **open data** and **entity identification** technologies will lead to master **data access control** and and **data de-identification**.

[source1] https://cdn-images-1.medium.com/max/2600/1*xHzO_5cSSVetWnjpAbQABw.png

IDENTITY MANAGEMENT IN THE WEB OF DATA

FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, ORSAY, FR

SAIS@LRI.FR

E-EGC2019

EGC2019 du 21 au 25 janvier 2019 à Metz

REFERENCES (1)

[Beek et al., 2016] A contextualised semantics for owl: sameas.

W. Beek, S. Schlobach, and F. van Harmelen. In ESWC 2016

[CudreMauroux et al., 2009] idmesh: graph-based disambiguation of linked data.

P. CudreMauroux, P. Haghani, M. Jost, K. Aberer, and H. De Meer. In WWW 2009.

[de Melo, 2013] Not quite the same: Identity constraints for the web of linked data.

G. de Melo. In AAAI 2013.

[Geach, 1967] Identity. P. Geach. Review of Metaphysics, 21:3–12, 1967.

[Guéret et al. 2012] Assessing linked data mappings using network measures.

C. Guéret, P. Groth, C. Stadler, and J. Lehmann. In ESWC 2012

[Halpin et al., 2010] When owl:sameAs isn't the same: An analysis of identity in Linked Data.

H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. In ISWC 2010.

[Hogan et al., 2012] Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora.

A. Hogan, A. Zimmermann, J. Umbrich, A. Polleres, and S. Decker. In JWS 2012.

REFERENCES (2)

[Jaffri et al., 2008] URI disambiguation in the context of linked data.

A. Jaffri, H. Glaser, and I. Millard. In LDOW@WWW 2008.

[Paulheim, 2014] Identifying wrong links between datasets by multi-dimensional outlier detection.

H. Paulheim. In WoDOOM 2014.

[Papaleo et al., 2014] Logical detection of invalid sameas statements in rdf data.

L. Papaleo, N. Pernelle, F. Saïs, and C. Dumont. In EKAW 2014.

[Raad et al., 2017] Detection of contextual identity links in a knowledge base.

J. Raad, N. Pernelle, and F. Saïs. In K-CAP 2017.

[Raad et al., 2018] Detecting Erroneous Identity Links on the Web using Network Metrics. J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs. ISWC 2018

[Valdestilhas et al., 2017] Cedal: time-efficient detection of erroneous links in large-scale link repositories. A. Valdestilhas, T. Soru, and A.-C. N. Ngomo. In WI 2017.

REFERENCES (3)

[Pernelle et al. 2013] An Automatic Key Discovery Approach for Data Linking.

Nathalie Pernelle, Fatiha Saïs. and Danai Symeounidou. In Journal of Web Semantics, 2013

[Symeonidou et al. 2014] SAKey: Scalable almost key discovery in RDF data.

Danai Symeonidou, 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.

[Ferrara et al. 2013] Alfio Ferrara, Andriy Nikolov, François Scharffe: Data Linking. J. Web Sem. 23: 1 (2013)