KNOWLEDGE GRAPH REFINEMENT
KEY DISCOVERY AND LINK INVALIDATION

FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

Joint work with: N. Pernelle, L. Papaleo, J. Raad and D. Symeonidou

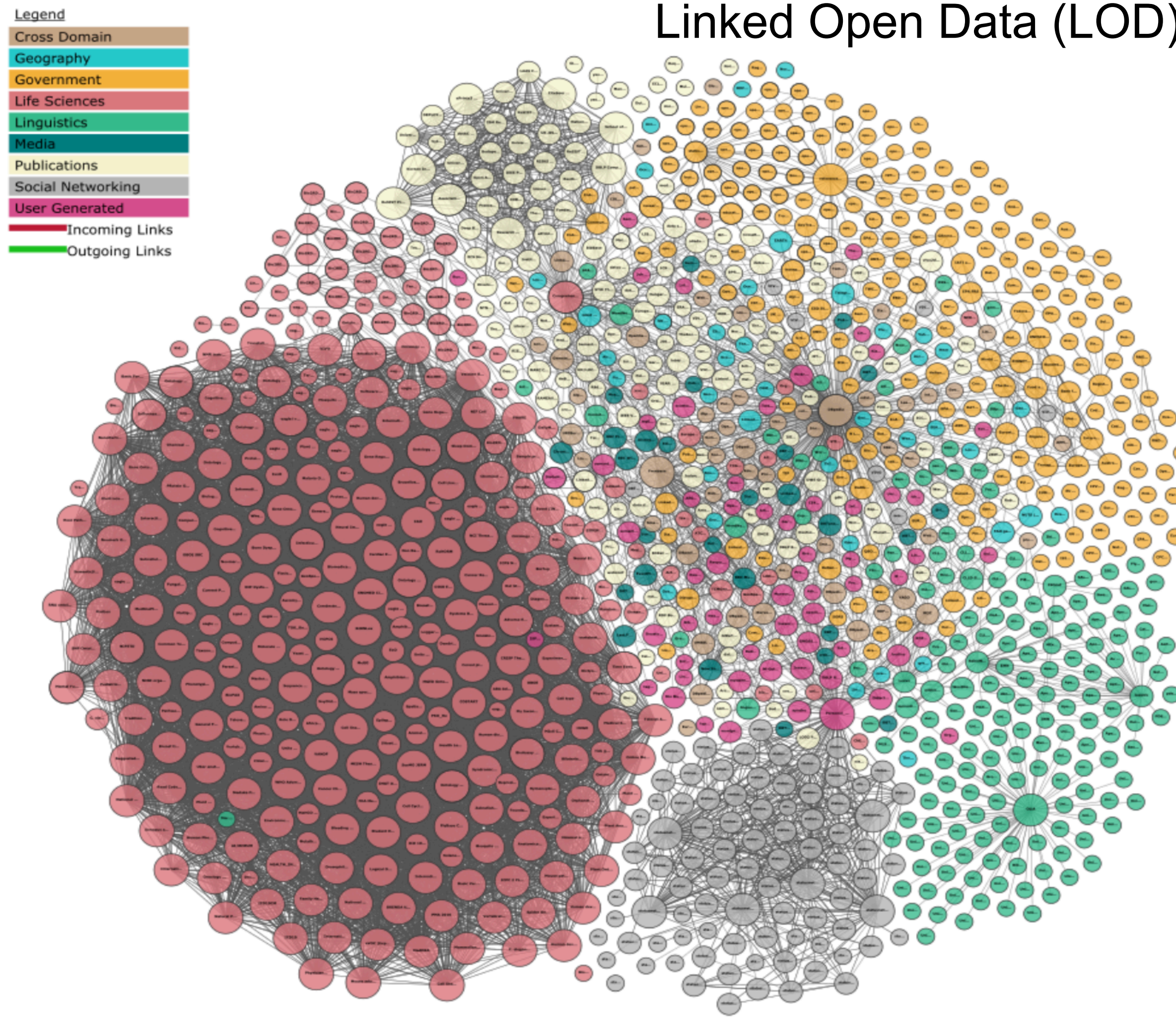
3ÈME JOURNÉE RI-IA SOUTENUE PAR L'AFIA ET ARIA, PARIS 2019

OUTLINE

- **Introduction**
  - Linked Data
  - Knowledge graphs
  - Knowledge graph refinement
- **Key discovery**
- **Link invalidation**
- **Conclusion**

LINKED OPEN DATA

Linked Open Data (LOD)



Linked Data - Datasets under an open access

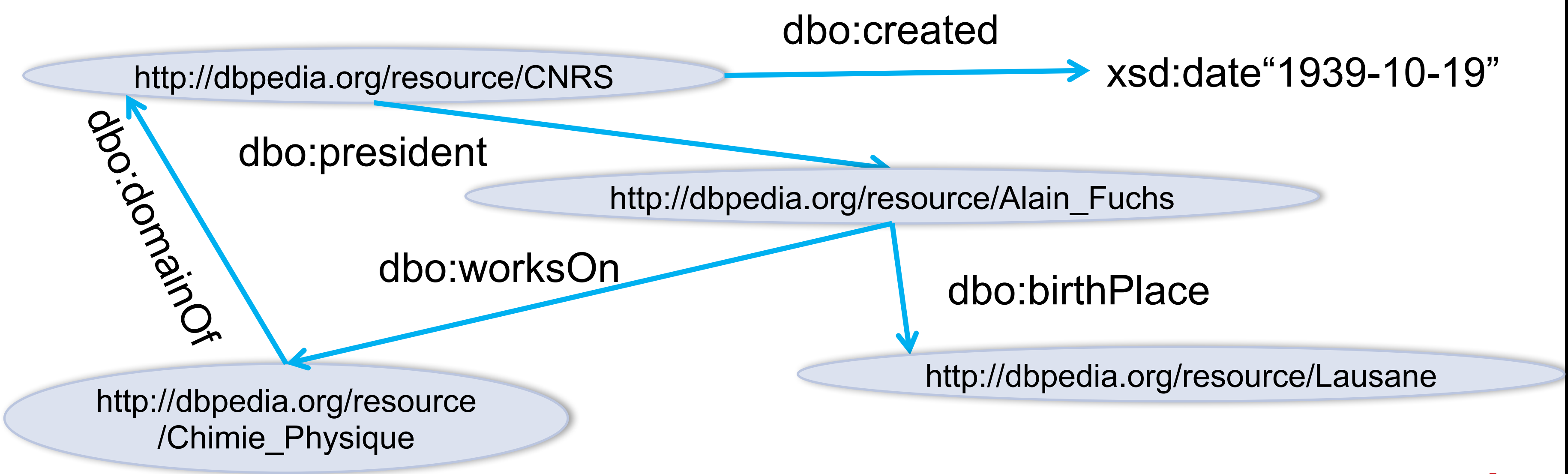
- **1,139 datasets**
- **over 100B triples**
- **about 500M links**
- **several domains**

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/"

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **An RDF Graph** is a set of triples.
  - Its **nodes** are (labelled by) the subjects and objects appearing in the triples.
  - Its **edges** are labelled by the properties



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

*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."*

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

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)

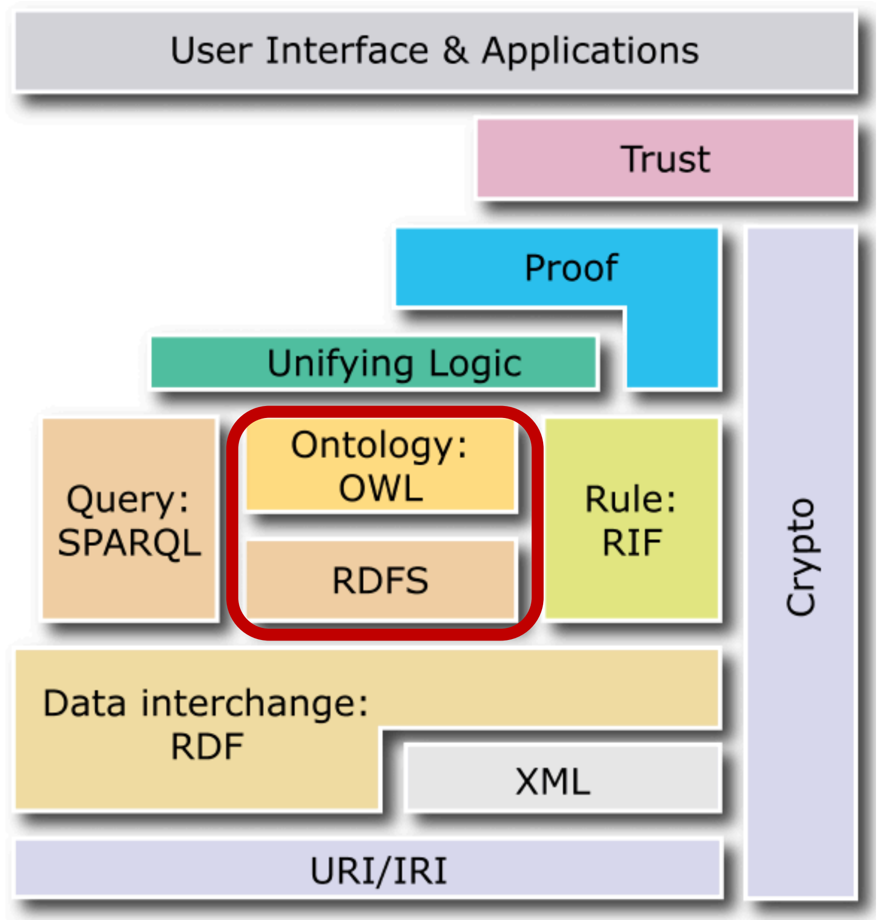
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema

- Lightweight ontologies

OWL – Web Ontology Language

- Expressive ontologies

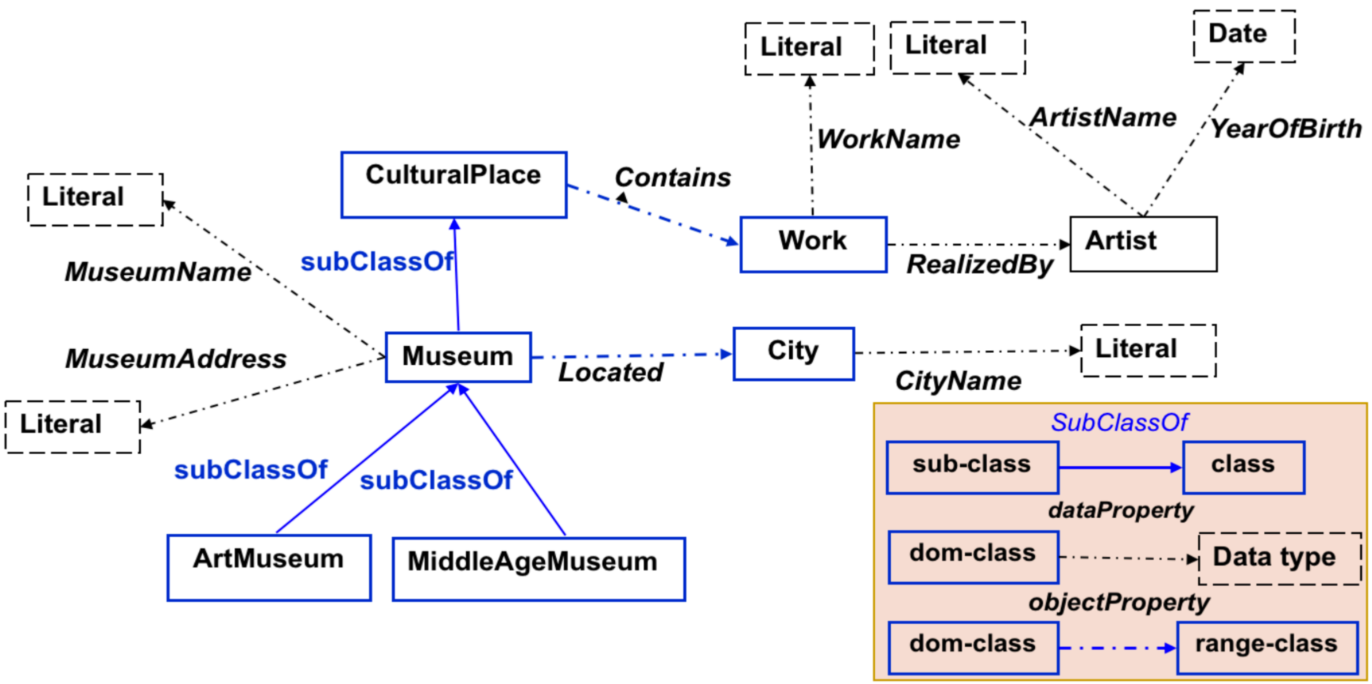


Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png

OWL – WEB ONTOLOGY LANGUAGE

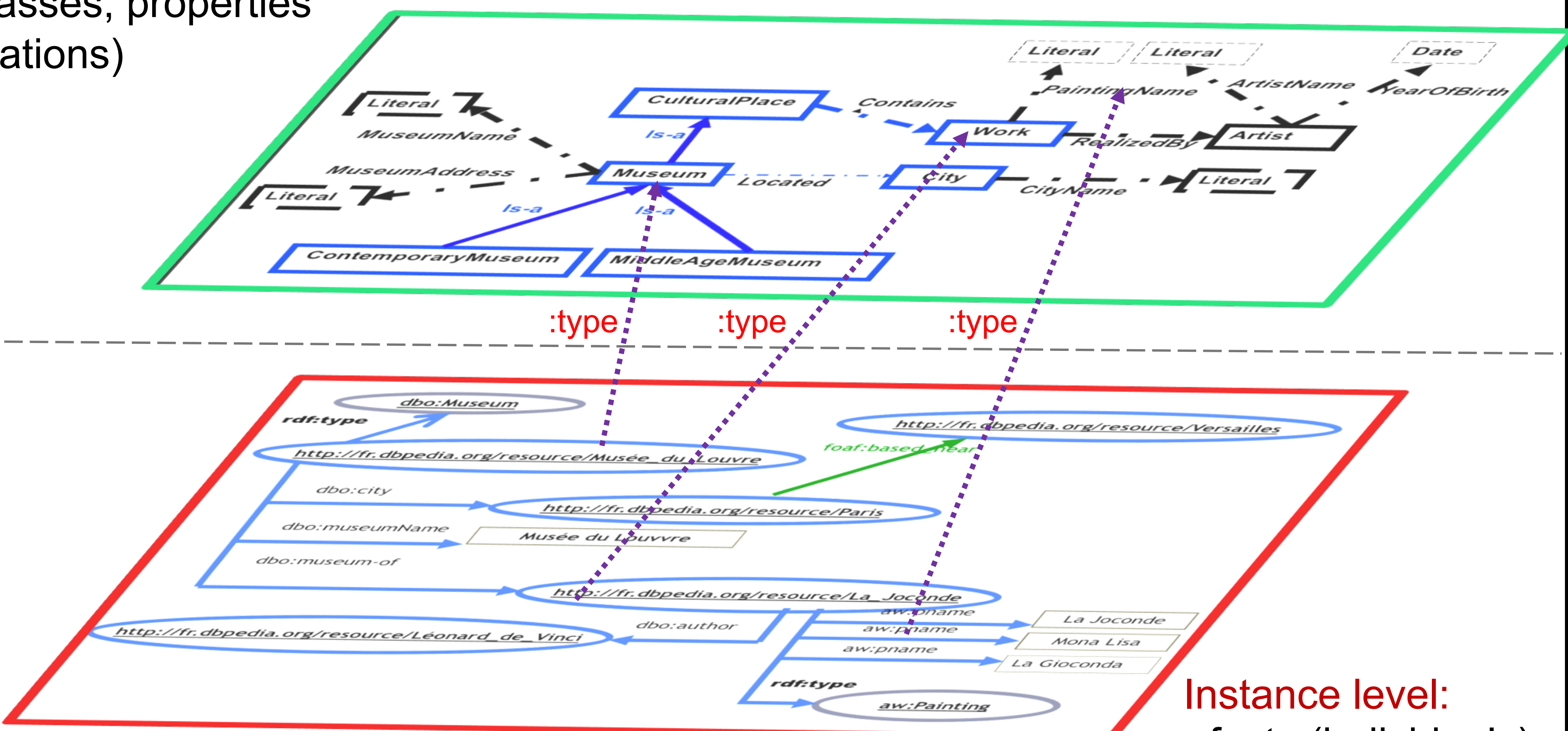
- **Classes:** concepts or collections of objects (individuals)
- **Properties**:
  - owl:DataTypeProperty (attribute)
  - owl:ObjectProperty (relation)
- **Individuals:** ground-level of the ontology (instances)

- **Axioms**
  - owl:subClassOf
  - owl:subPropertyOf
  - owl:inverseProperty
  - owl:FunctionalProperty
  - owl:minCardinality
  - …



ONTOLOGY LEVELS: KNOWLEDGE ENGINEERING VIEW

Conceptual level:
- classes, properties (relations)



Instance level:
- facts (individuals)

KNOWLEDGE GRAPHS

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

2007

yago
select knowledge

2012

WIKIDATA

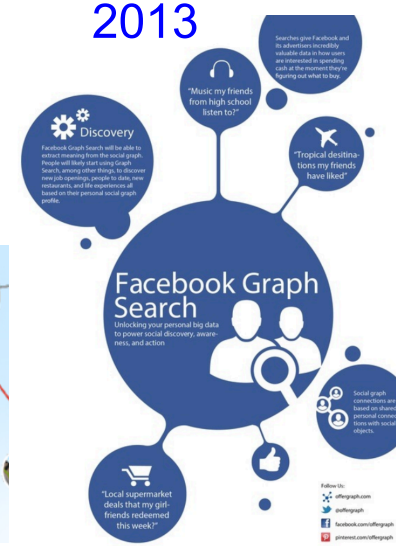
2007

Freebase™

Academic side

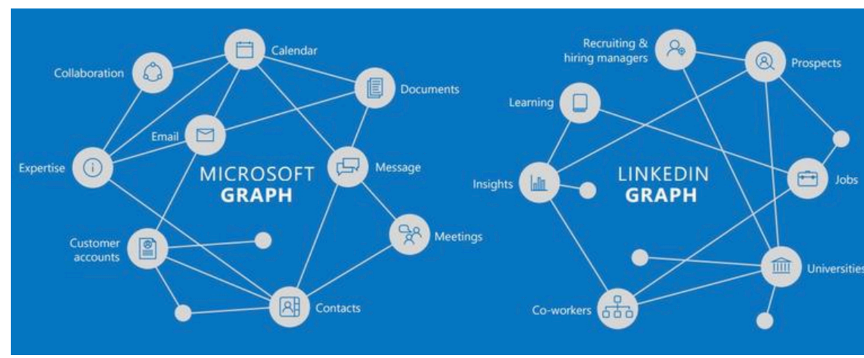
2012

2013



2015

2016

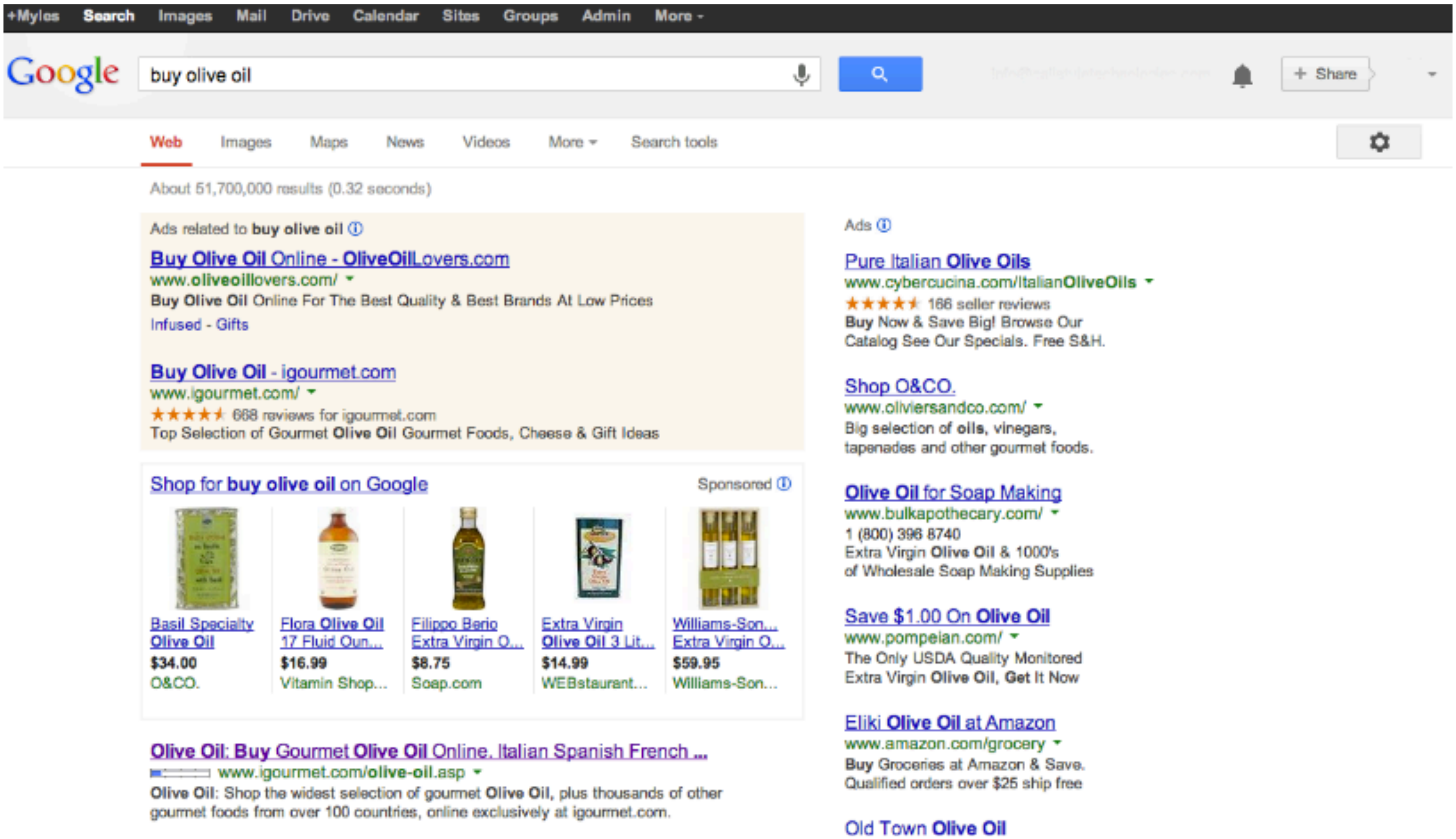


2013

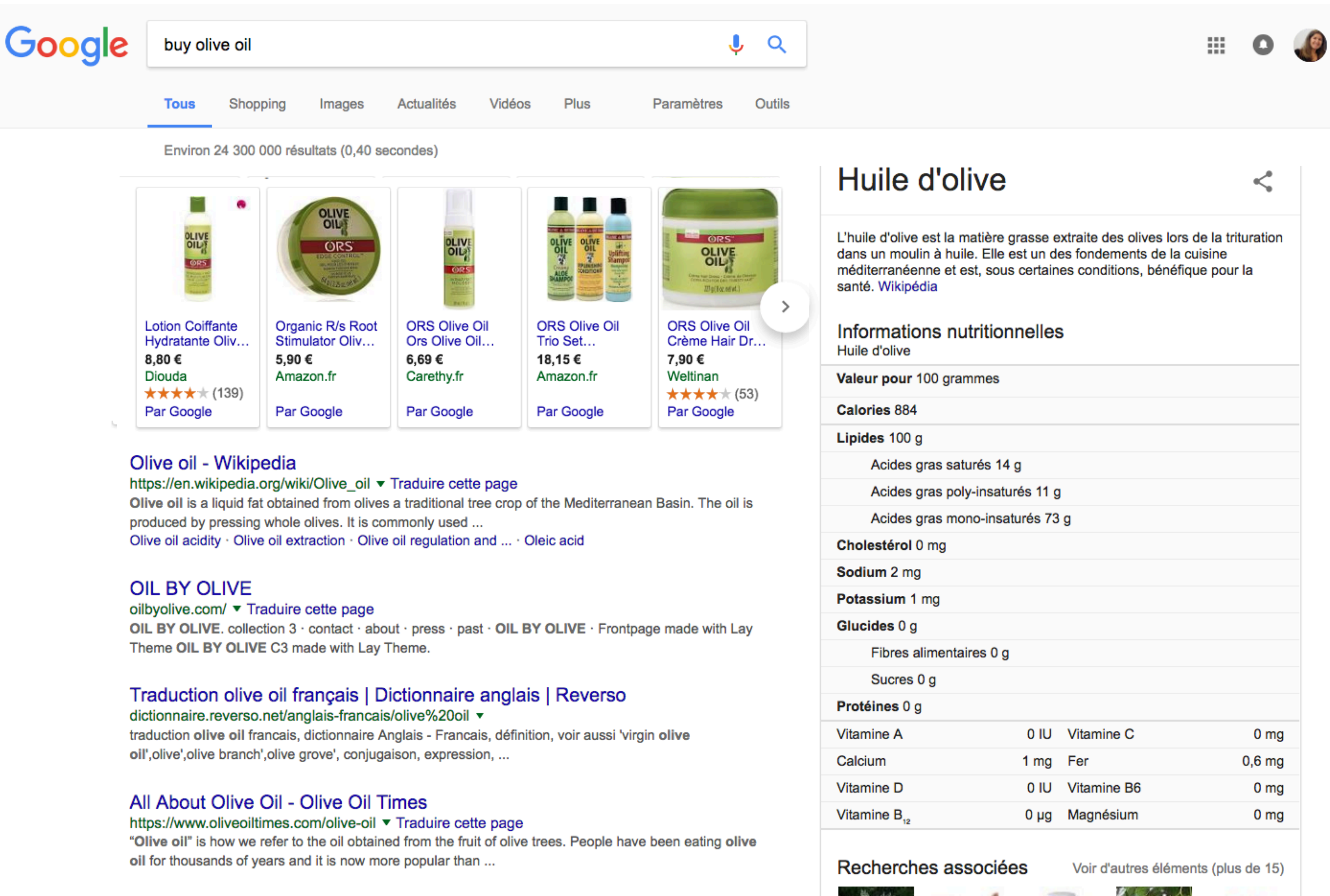
Yahoo's new SERP designs mobile and knowledge graph

Commercial side

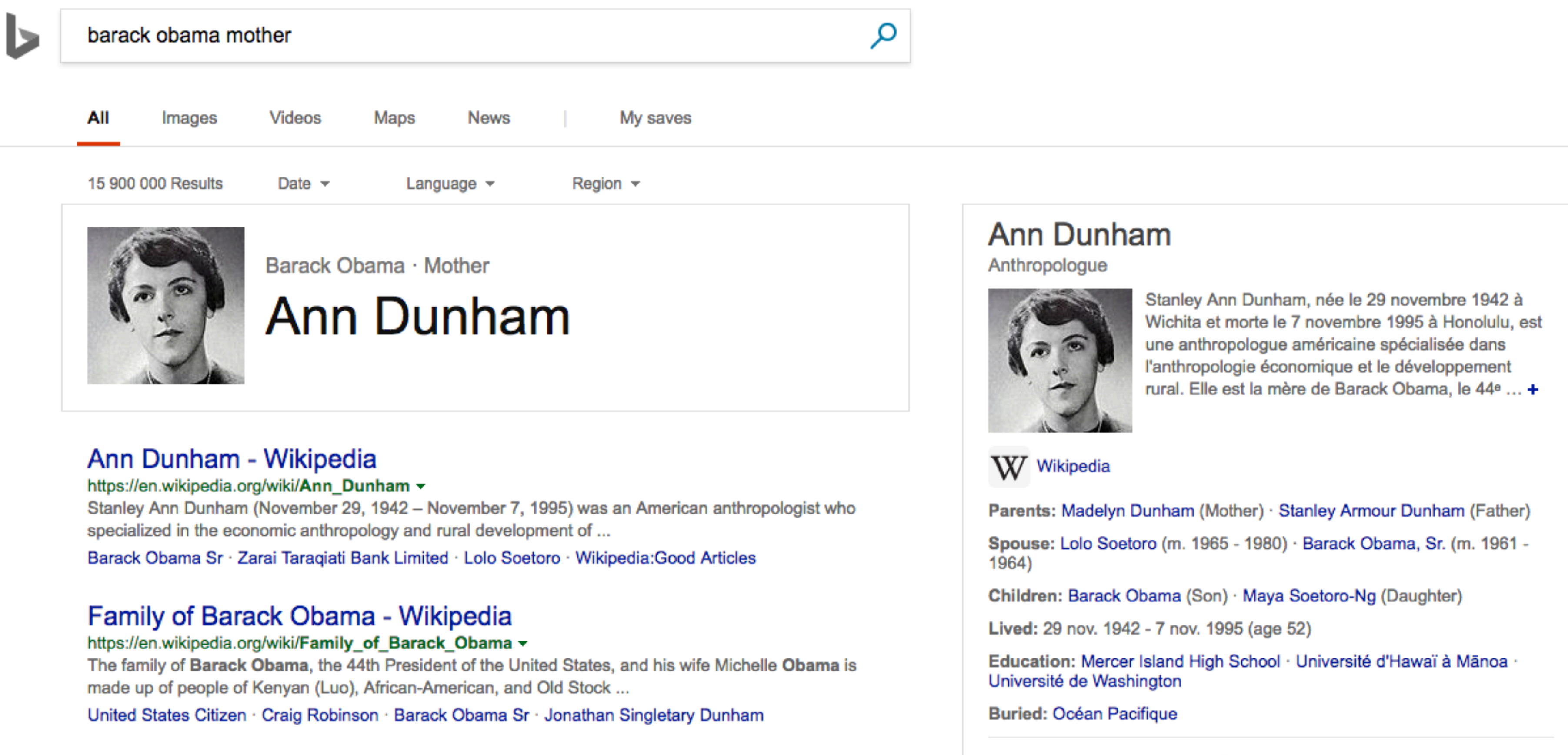
WEB SEARCH WITHOUT KNOWLEDGE GRAPHS



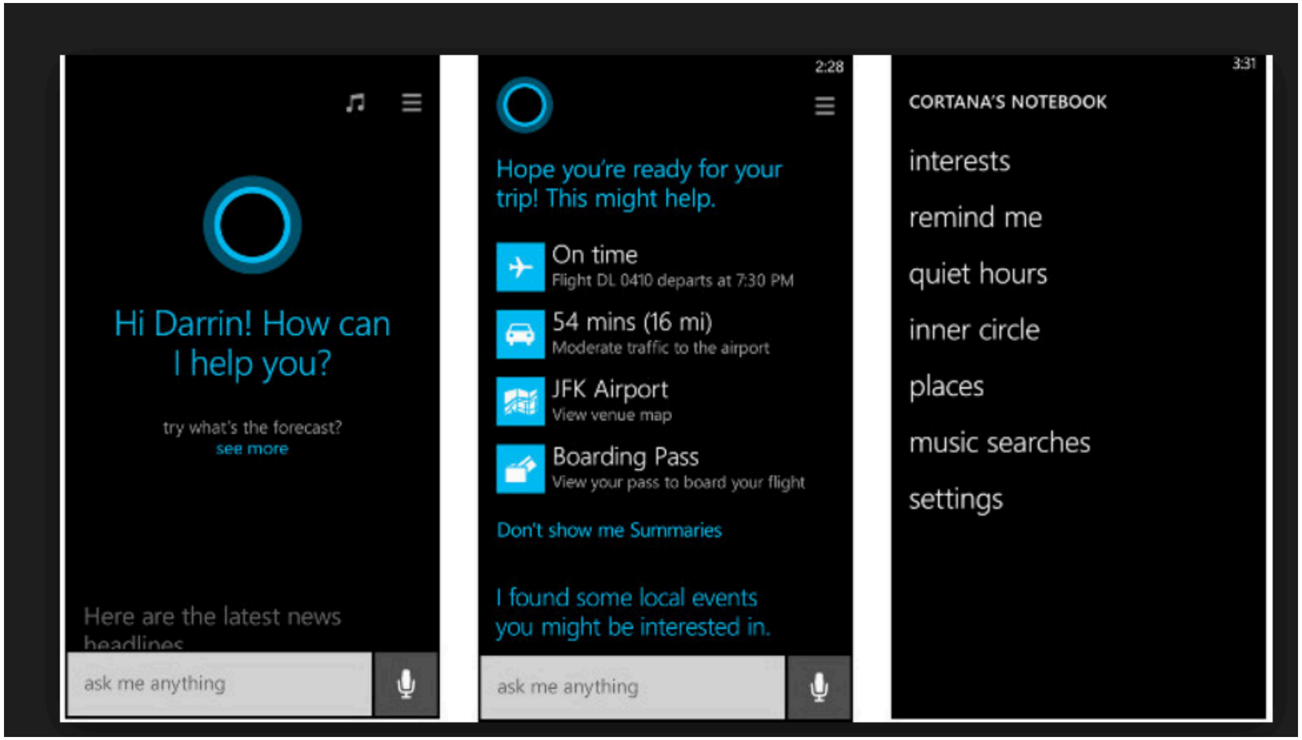
WEB SEARCH WITH KNOWLEDGE GRAPHS



QUESTION ANSWERING WITH KNOWLEDGE GRAPHS



TOWARDS A KNOWLEDGE-POWERED DIGITAL ASSISTANT



Cortana (Microsoft)

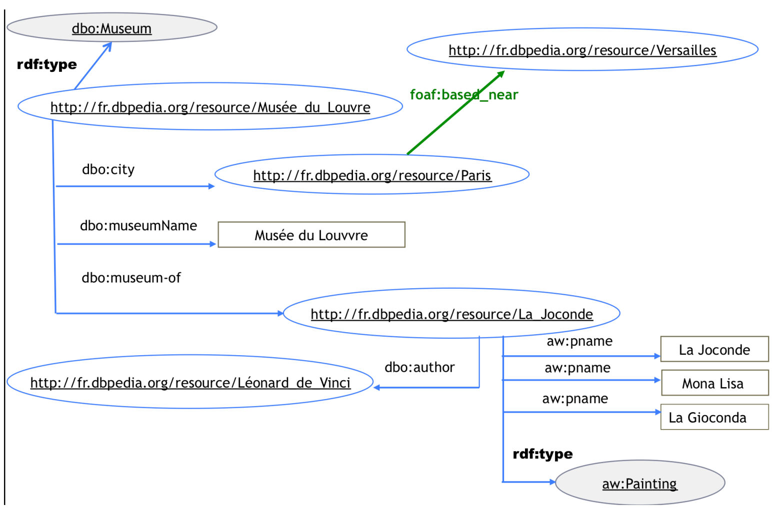
- Natural access and storage of knowledge
- Chat bots
- Personalization
- Emotion

KNOWLEDGE GRAPH ADOPTION [2019]

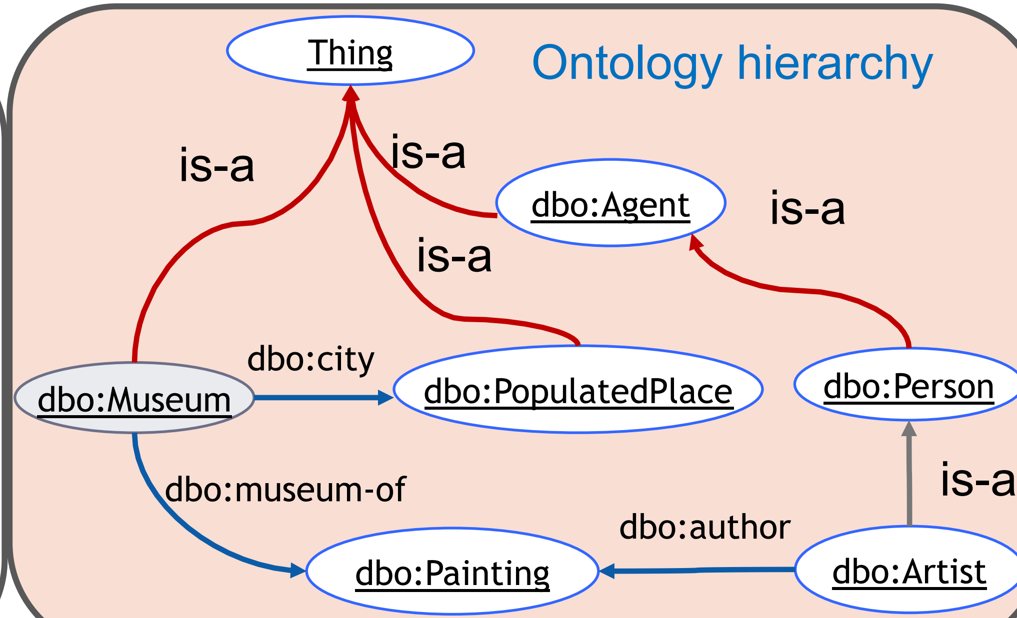
source: https://fr.slideshare.net/Frank.van.Harmelen/adoption-of-knowledge-graphs-mid-2019

KNOWLEDGE GRAPH - DEFINITION

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 .}
```

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)

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

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

KNOWLEDGE GRAPH COMPLETENESS?

	Name	Instances	Facts	Types	Relations
public	DBpedia (English)	4,806,150	176,043,129	735	2,813
	YAGO	$\boxed{4,595,906}$	25,946,870	$\boxed{488,469}$	77
	Freebase	49,947,845	3,041,722,635	26,507	37,781
	Wikidata	15,602,060	65,993,797	23,157	1,673
	NELL	2,006,896	432,845	285	425
	OpenCyc	118,499	2,413,894	45,153	18,526
private	Google's Knowledge Graph	$\boxed{570,000,000}$	18,000,000,000	$\boxed{1,500}$	35,000
	Google's Knowledge Vault	45,000,000	271,000,000	1,100	4,469
	Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

Heiko Paulheim. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web 8:3(2017), pp 489-508.

KNOWLEDGE GRAPH CORRECTNESS?

About: Donald Trump

An Entity of Type : person, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

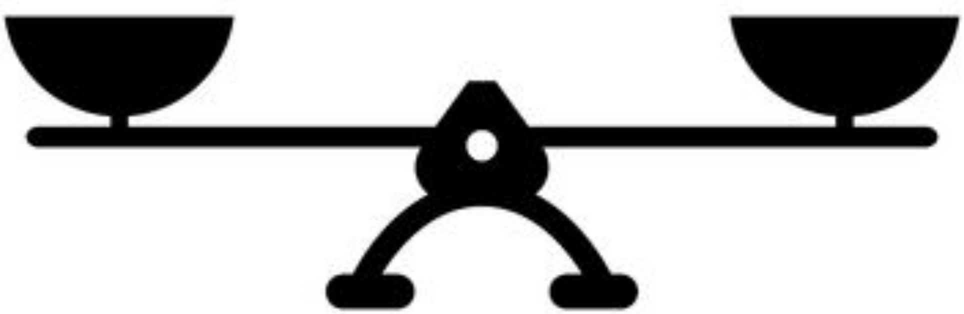
Donald John Trump (born June 14, 1946) is an American businessman, author, television producer, politician, and the Republican Party nominee for President of the United States in the 2016 election. He is the chairman and president of The Trump Organization, which is the principal holding company for his real estate ventures and other business interests. During his career, Trump has built office towers, hotels, casinos, golf courses, an urban development project in Manhattan, and other branded facilities worldwide.

dbo:birthName	▪ Donald John Trump (en)
dbo:birthPlace	▪ dbr:Queens ▪ dbr:New_York_City
dbo:birthYear	▪ 1946-01-01 (xsd:date)
dbo:child	▪ dbr:Donald_Trump_Jr. ▪ dbr:Tiffany_Trump ▪ dbr:Eric_Trump ▪ dbr:Ivanka_Trump ▪ dbr:Donald_Trump

Donald Trump is the child of himself!

KNOWLEDGE GRAPH REFINEMENT

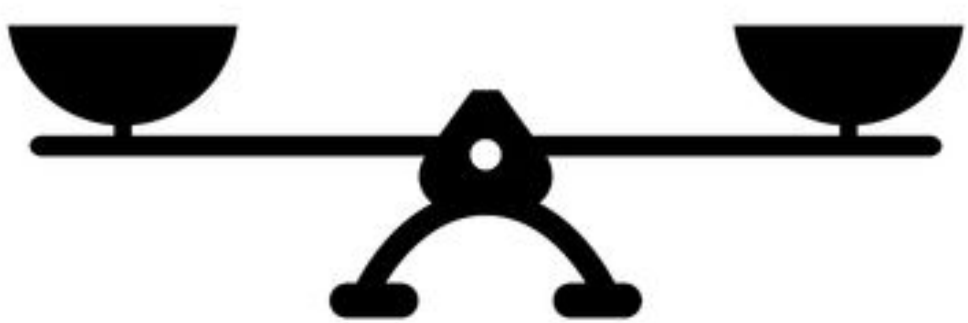
Completeness Correctness



KNOWLEDGE GRAPH REFINEMENT

Completeness

Correctness



Key discovery

Data Linking

Data Fusion

Link Invalidation

Contextual identity

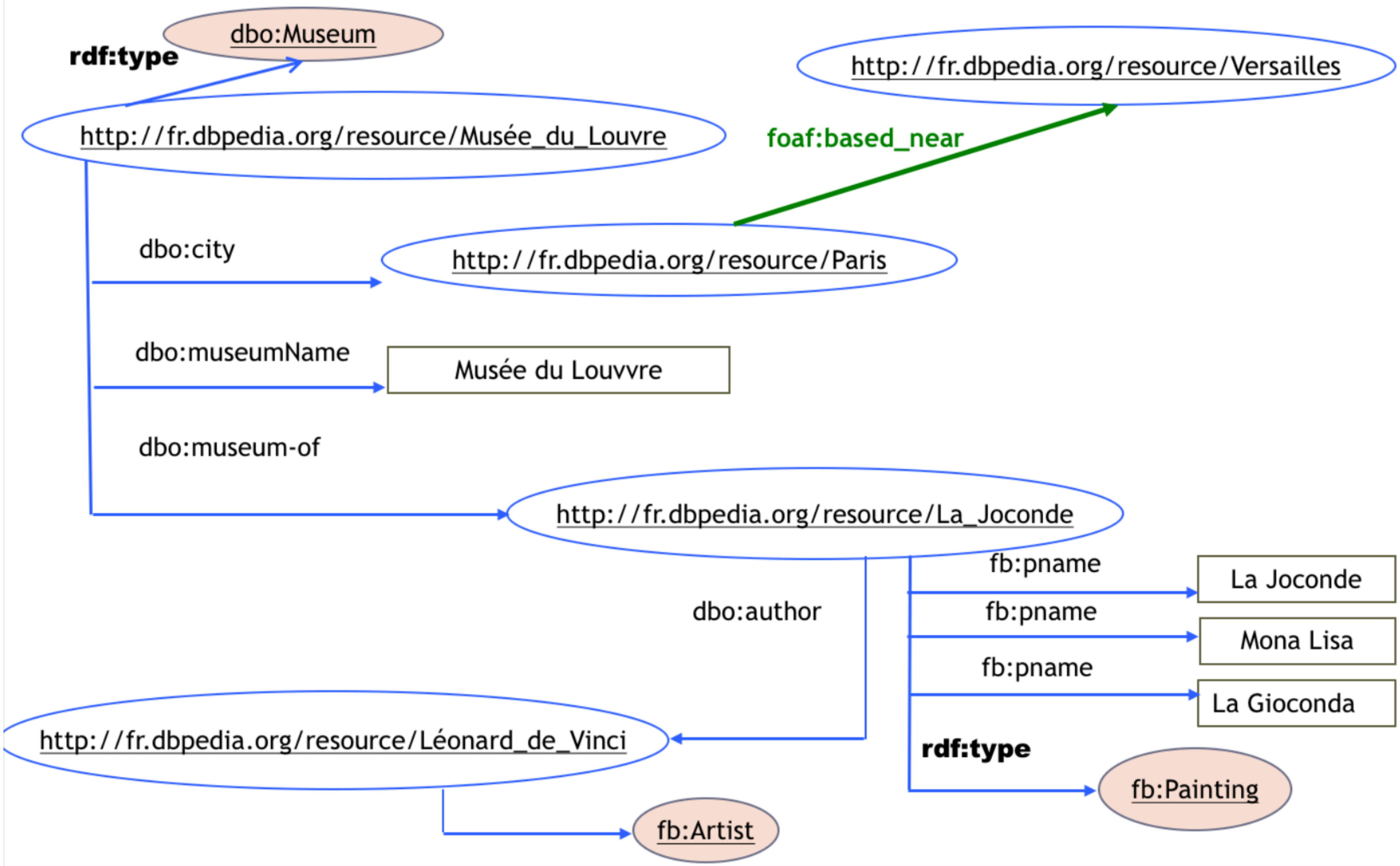
OUTLINE

- Introduction
- Key discovery
- Link invalidation
- Conclusion

KEY DISCOVERY FOR DATA LINKING

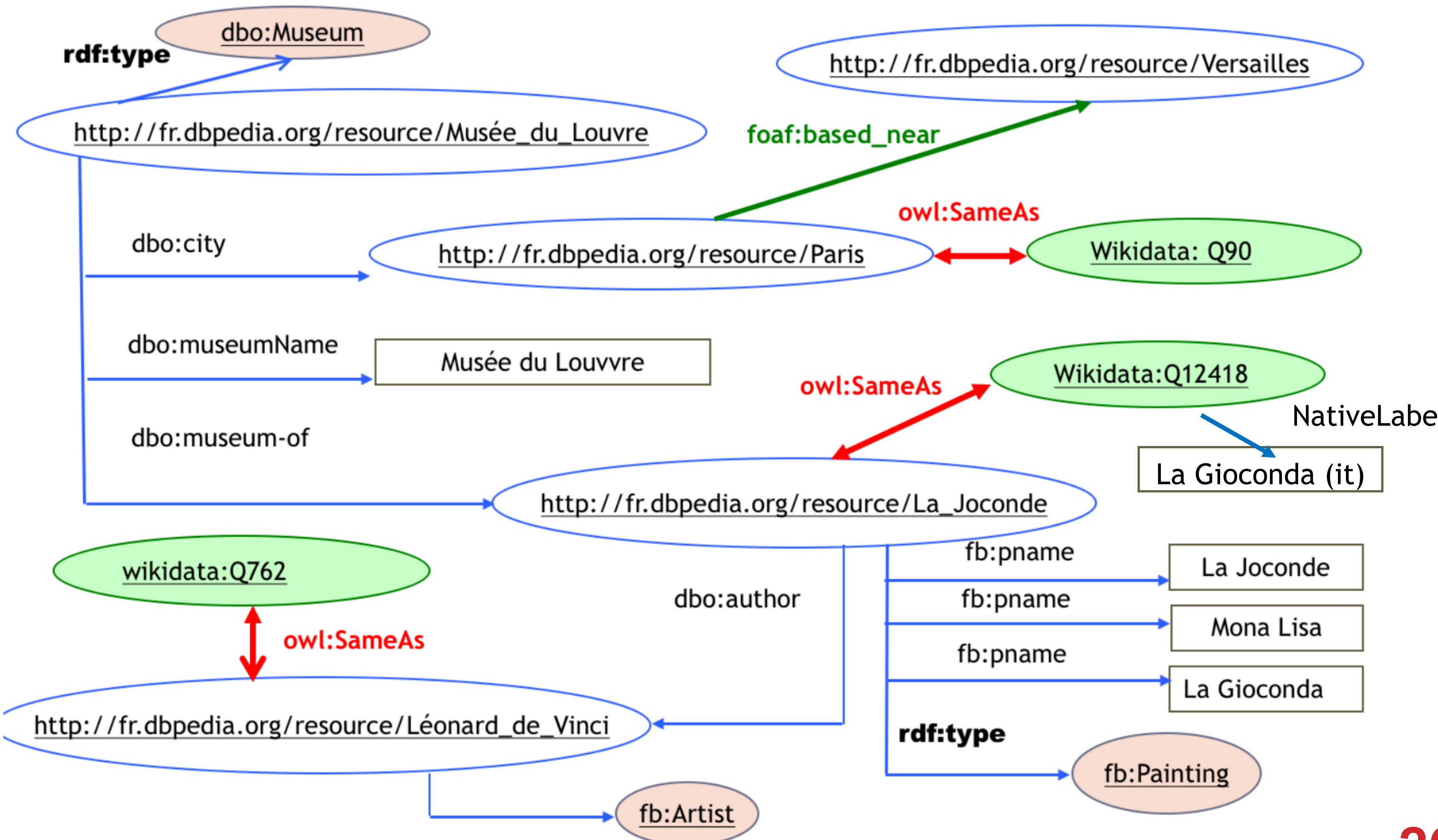
DATA LINKING

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



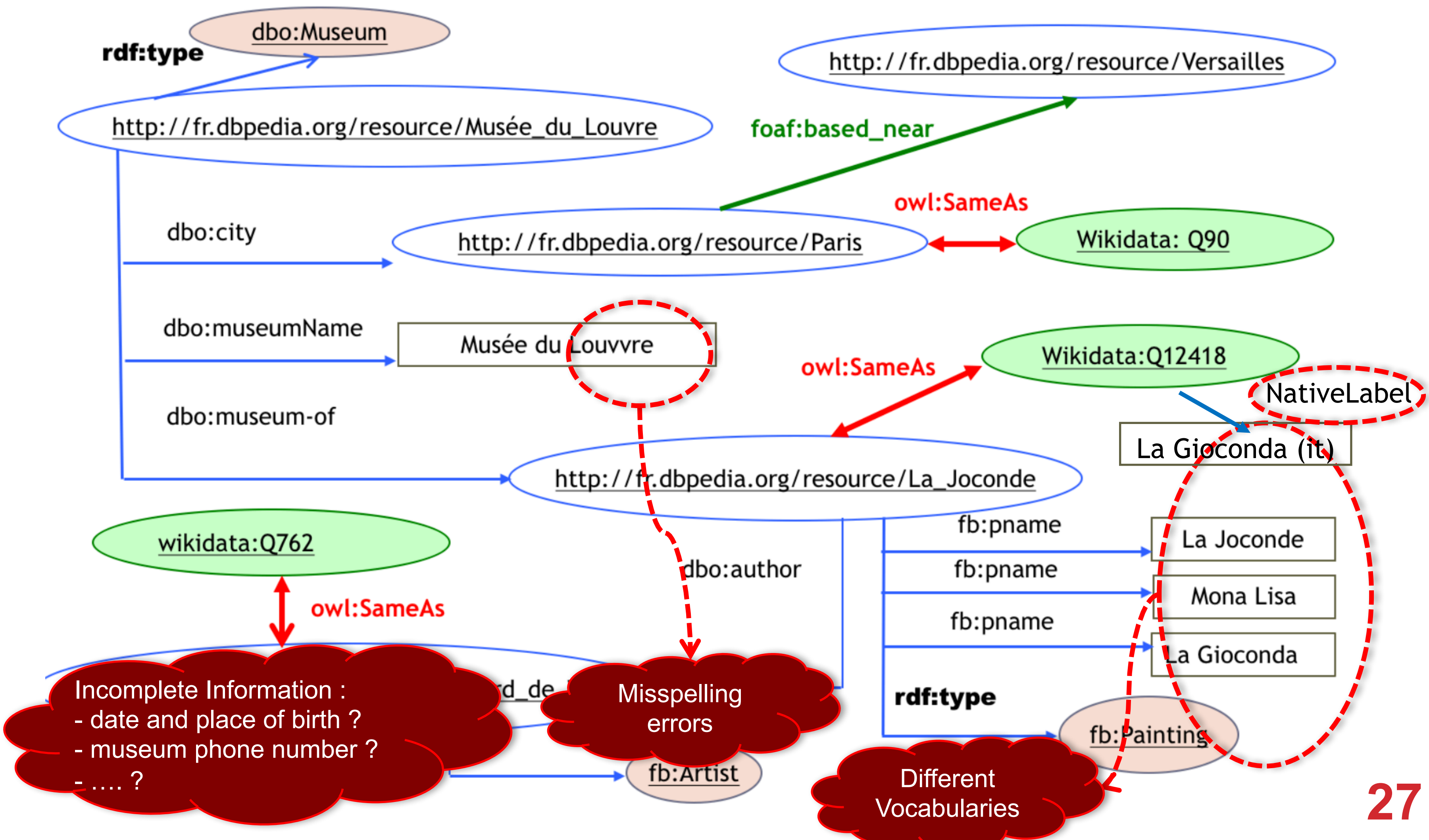
DATA LINKING

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



KEY DISCOVERY FOR DATA LINKING

Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for knowledge to be declared in an ontology language or other languages.

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

Then we may infer:

sameAs(museum11, museum21)
sameAs(museum12, museum22)
sameAs(museum13, museum23)

A key: is a set of properties that uniquely identifies every instance of a class

	…	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**

KEY DISCOVERY FOR DATA LINKING

Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for knowledge to be declared in an ontology language or other languages.

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

Then we may infer:

sameAs(museum11, museum21)
sameAs(museum12, museum22)
sameAs(museum13, museum23)

A key: is a set of properties that uniquely identifies every instance of a class

	…	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?

KEY VALIDITY: EXACT KEYS

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
Person4	Simon	Roche	06/12/90	Teacher
Person4	Jane	Ser	15/05/87	Teacher, Researcher
Person4	Sara	Khan	27/10/84	Teacher
Person4	Theo	Martin	06/12/90	Teacher, Researcher
Person4	Marc	Blanc	27/10/84	Teacher

Is [LastName] a key? ✖

Is [FirstName,LastName] a key? ✔

Exact keys

KEY VALIDITY: KEYS WITH EXCEPTIONS

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
Person4	Simon	Roche	06/12/90	Teacher
Person4	Jane	Ser	15/05/87	Teacher, Researcher
Person4	Sara	Khan	27/10/84	Teacher
Person4	Theo	Martin	06/12/90	Teacher, Researcher
Person4	Marc	Blanc	27/10/84	Teacher

Is [FirstName,LastName] a key? ✔

Is [Birthdate] a key? ✖

Is [Birthdate] a key with 2 exceptions? ✔

Exact keys

Almost keys

KEY VALIDITY: CONDITIONAL KEYS

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
Person4	Simon	Roche	06/12/90	Teacher
Person4	Jane	Ser	15/05/87	Teacher, Researcher
Person4	Sara	Khan	27/10/84	Teacher
Person4	Theo	Martin	06/12/90	Teacher, Researcher
Person4	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 FOR DATA LINKING: KEY SEMANTICS

OWL2 Semantics

- **A 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) $\wedge$ 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)

KEY DISCOVERY FOR DATA LINKING

Related Work

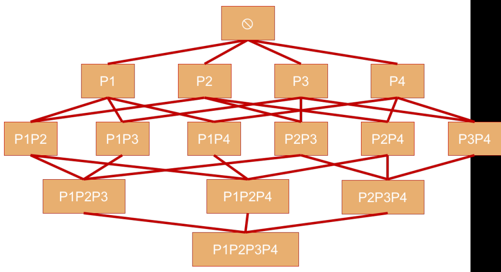
- **Approaches in relational databases are not applicable**
  - Closed world assumption
  - Do not consider multi-valued properties
  - No ontologies (semantics cannot be used)

Contributions

- **KD2R [ISSW 2011, JWS 2013]: exact key** discovery
  - Danai Symeonidou PhD, Qualinca ANR Project (2012-2016)
- **SAKey [ISWC 2014]: n-almost key** discovery
  - Danai Symeonidou PhD, Qualinca ANR Project (2012-2016)
- **VICKEY [ISWC 2017]: conditional key** discovery
  - Collaboration with INRA, Telecom ParisTech and Aalborg University (Danemark).

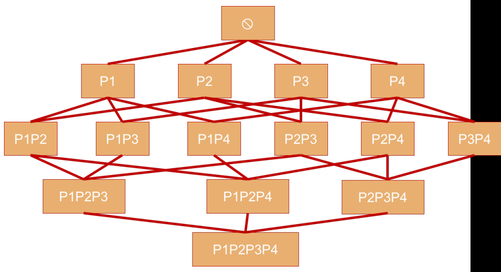
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least 2^n property combinations
  - ➢ need of efficient filtering and prunings



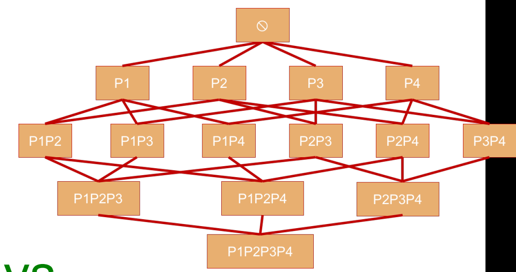
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least 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 at least 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	Phone	Profession
Person1	Anne	Tompson	0169154259	Actor, Director
Person2	Marie	Tompson	0169154226	Actor
Person3	Marie	David	0425154012	Actor
Person4	Vincent	Solgar	0425154009	Actor, Director
Person5	Simon	Roche	0321455823	Teacher
Person6	Jane	Ser	0425462914	Teacher, Researcher
Person7	Sara	Khan	0425462915	Teacher
Person8	Theo	Martin	0321455823	Teacher, Researcher
Person9	Marc	Blanc	0169154228	Teacher

KD2R

SAKEY

VICKEY

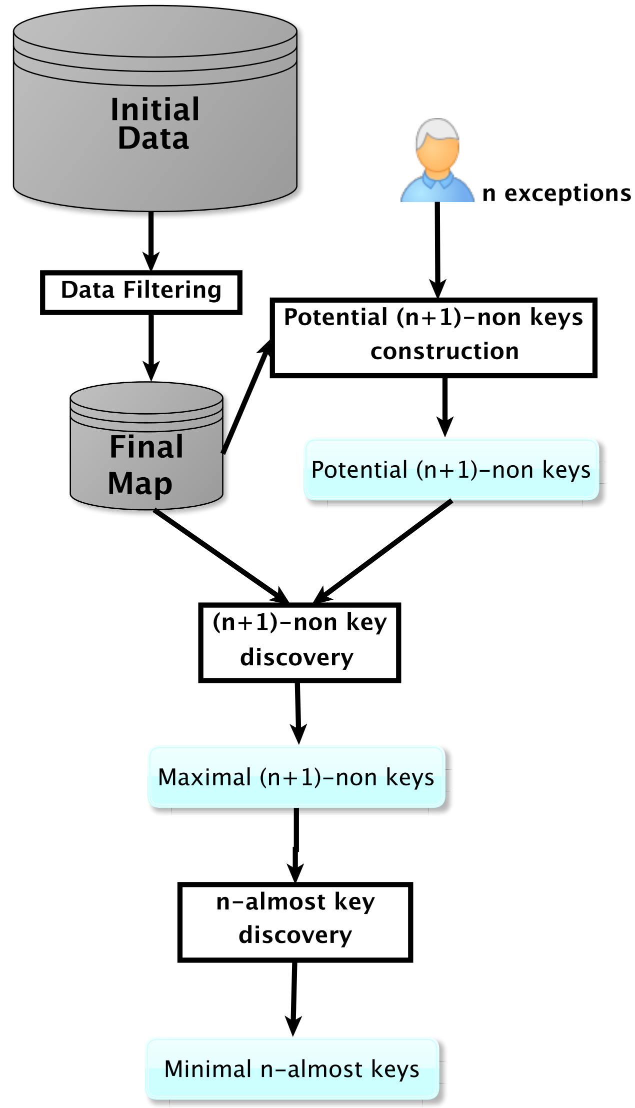
Is [LastName] a non-key? ➔ scan only a part of the data

SAKEY: N-ALMOST KEY DISCOVERY

- SAKey allows ***n*** **exceptions** in the data
- **Exception set E_P:** set of instances that share values for the set of properties P
- **n-almost key: a set of properties where $|E_P| \leq n$**
- **n-non key: a set of properties where $|E_P| \geq n+1$**

n=4

5-non keys	4-almost keys
All sets of properties that contain at least 5 exceptions	All sets of properties that contain at most 4 exceptions



SAKEY: EVALUATION

Evaluation on **13 different datasets** (OAEI, Qualinca project, Dbpedia, …)

Scalability

- Big classes (dbo:NaturalPlace more than 16 million triples and 243 properties): non-key discovery in 1min and key derivation 5min)

Quality

- **Data linking with SAKey keys**: obtains close or better results than expert keys
- **Exceptions**: important increase of recall and weak decrease of the precision.

# 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%

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

VICKEY: CONDITIONAL-KEY DISCOVERY

To discover even more keys in a dataset

VICKEY: CONDITIONAL-KEY DISCOVERY

To discover even more keys in a dataset

Conditional key: a key, valid for instances of a class satisfying a specific condition

Instances of the class Person

	FirstName	LastName	Gender	Lab	Nationality
instance1	Claude	Dupont	Female	Paris-Sud	France
instance2	Claude	Dupont	Male	Paris-Sud	Belgium
instance3	Juan	Rodríguez	Male	**INRA**	Spain, Italy
instance4	Juan	Salvez	Male	**INRA**	Spain
instance5	Anna	Georgiou	Female	**INRA**	Greece, France
instance6	Pavlos	Markou	Male	Paris-Sud	Greece
instance7	Marie	Legendre	Female	**INRA**	France

VICKEY: CONDITIONAL-KEY DISCOVERY

To discover even more keys in a dataset

Conditional key: a key, valid for instances of a class satisfying a specific condition

Instances of the class Person

	FirstName	LastName	Gender	Lab	Nationality
instance1	Claude	Dupont	Female	Paris-Sud	France
instance2	Claude	Dupont	Male	Paris-Sud	Belgium
instance3	Juan	Rodríguez	Male	**INRA**	Spain, Italy
instance4	Juan	Salvez	Male	**INRA**	Spain
instance5	Anna	Georgiou	Female	**INRA**	Greece, France
instance6	Pavlos	Markou	Male	Paris-Sud	Greece
instance7	Marie	Legendre	Female	**INRA**	France

{LastName} is a *key* under the *condition* **{Lab=INRA}**

Conditional keys

Algorithm: discovers minimal conditional keys from maximal non-keys (SAKey)

VICKEY: EVALUATION

Goal: evaluate the quality of data linking using:

- Classical keys discovered by SAKey
- Conditional keys discovered by VICKEY
- Both classical keys and conditional keys

Use of **Yago** and **Dbpedia** datasets (**9 classes**) **:** Actor, Album, Book, Film, Mountain, Museum, Organization, Scientist, University

VICKEY: EVALUATION

Goal: evaluate the quality of data linking using:

- Classical keys discovered by SAKey
- Conditional keys discovered by VICKEY
- Both classical keys and conditional keys

Use of **Yago** and **Dbpedia** datasets (**9 classes**) **:** Actor, Album, Book, Film, Mountain, Museum, Organization, Scientist, University

Class		Recall	Precision	F-Measure
Actor	SAKey Keys	0.27	0.99	**0.43**
	Conditional keys	0.57	0.99	0.73
	SAKey Keys + Conditional keys	0.6	0.99	**0.75**
Album	SAKey Keys	0	1	**0.00**
	Conditional keys	0.15	0.99	0.26
	SAKey Keys + Conditional keys	0.15	0.99	**0.26**
Film	SAKey Keys	0.04	0.99	**0.08**
	Conditional keys	0.38	0.96	0.54
	SAKey Keys + Conditional keys	0.39	0.98	**0.55**

x 1.75

x 869

x 7.1

KEY DISCOVERY: SUMMARY

- **Different methods** (KD2R, SAKey, VICKEY, Linkkey [Atencia et al. 2014], Rocker [Soru et al. 2015]) that discover three different kinds of keys
- **Relevance** of exact-keys, n-almost and conditional keys for **data linking**
- Relying on the strategy of **non-key search first** prevents the use of **well-known quality metrics** to prune the search space (e.g., support)

Possible improvements

- **More expressive keys** such as key graphs or referring expressions may be discovered
- **Different key semantics** can co-exist: how to choose the good key semantics using the data characteristics (e.g. completeness)

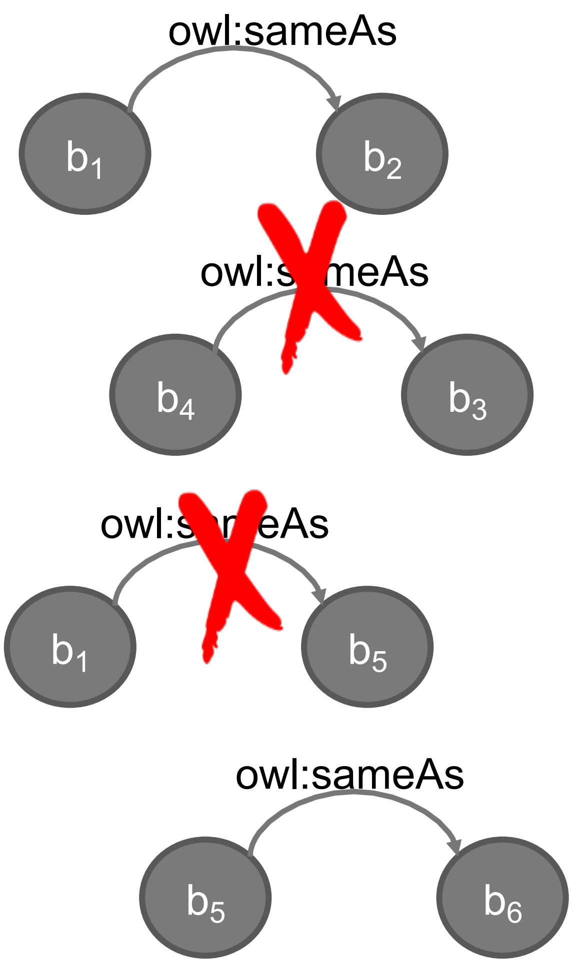
OUTLINE

- Introduction
- Key discovery
- Link invalidation
- Conclusion

IDENTITY PROBLEM

- [Halpin et al. 2010] showed that 37% of owl:sameAs links randomly selected among 250 identity links between books were incorrect.
- In [Jaffri et al., 2008], the authors discuss how erroneous use of owl:sameAs in the interlinking of the DBpedia and DBLP datasets has resulted in publications becoming incorrectly assigned to different authors.

- Automatic data linking tools do not guarantee 100% precision, because of:
  - Errors, missing information, data freshness, etc.



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:

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

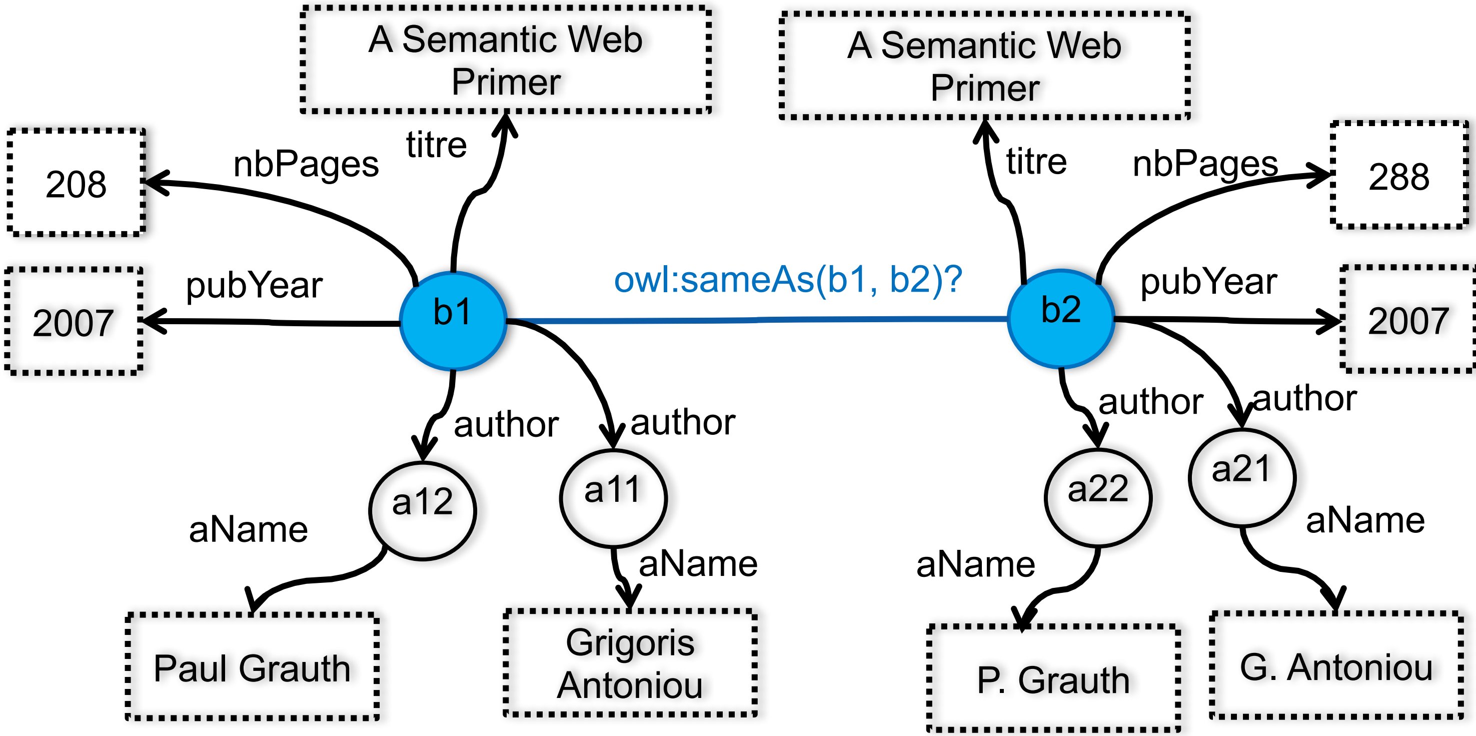
DETECTION OF ERRONEOUS IDENTITY LINKS

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

DETECTION OF ERRONEOUS IDENTITY LINKS

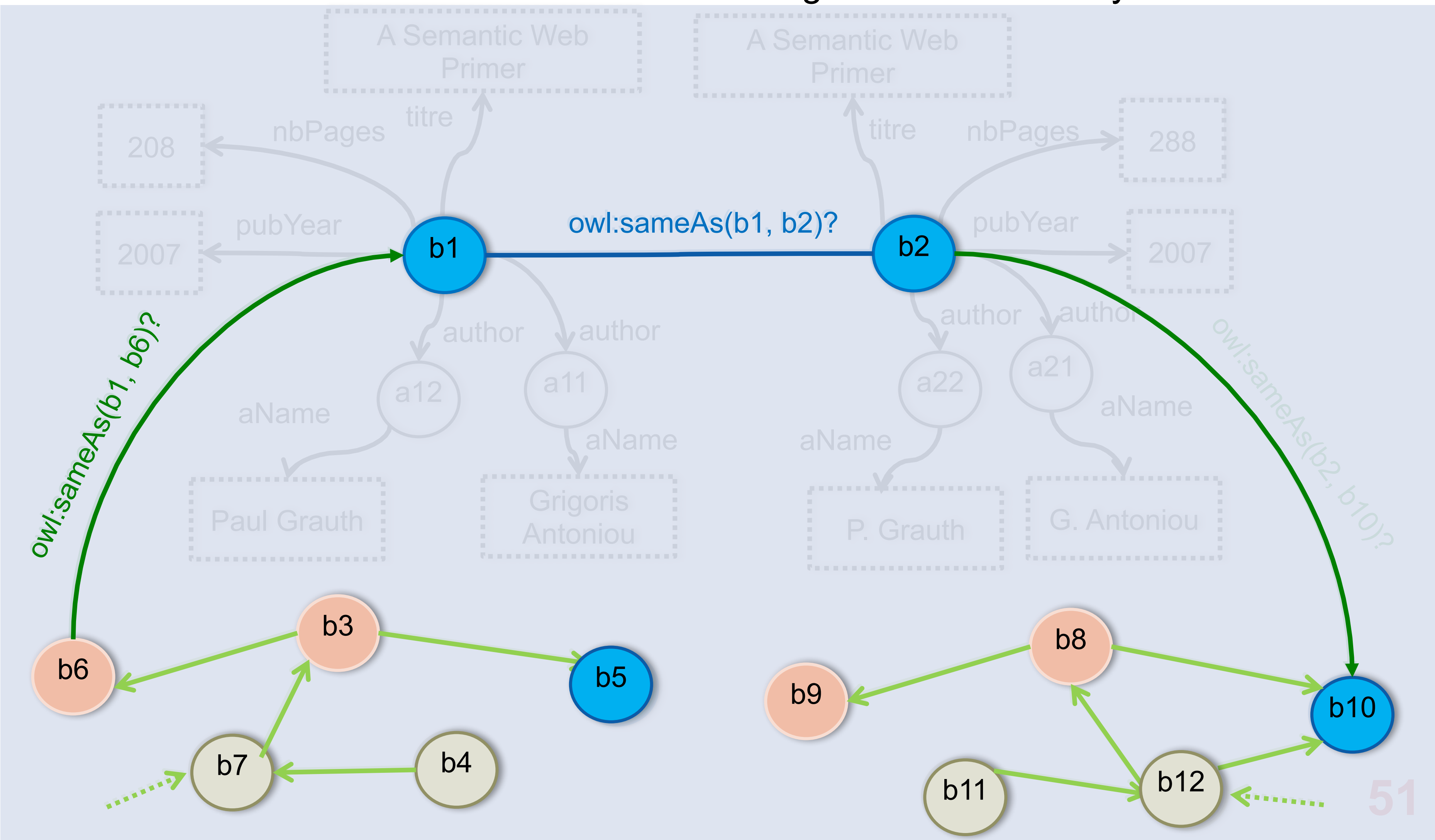
Content

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



DETECTION OF ERRONEOUS IDENTITY LINKS

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



DETECTION OF ERRONEOUS IDENTITY LINKS

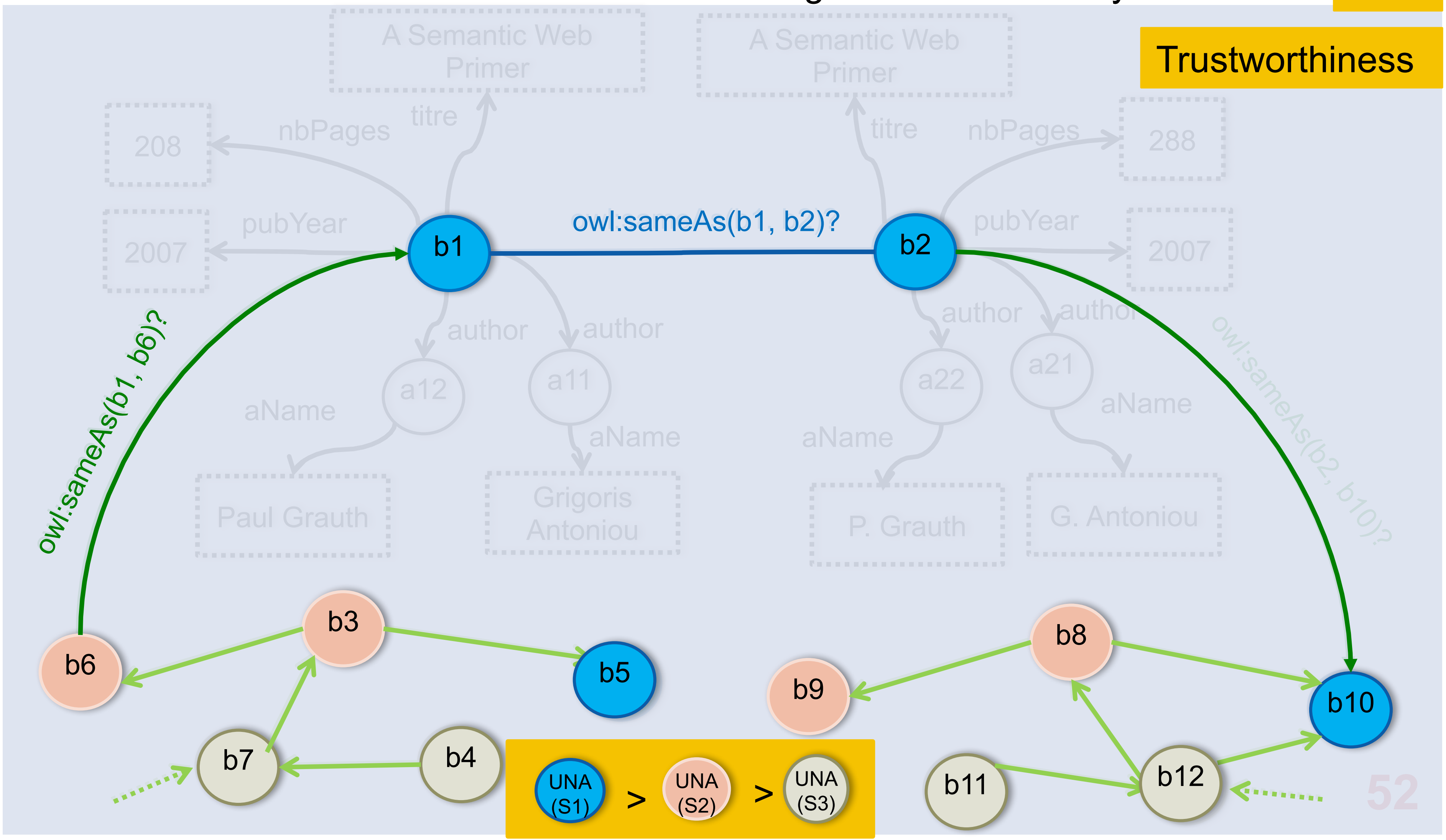
Content

Identity Network

UNA

Trustworthiness

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



DETECTION OF ERRONEOUS IDENTITY LINKS

Content

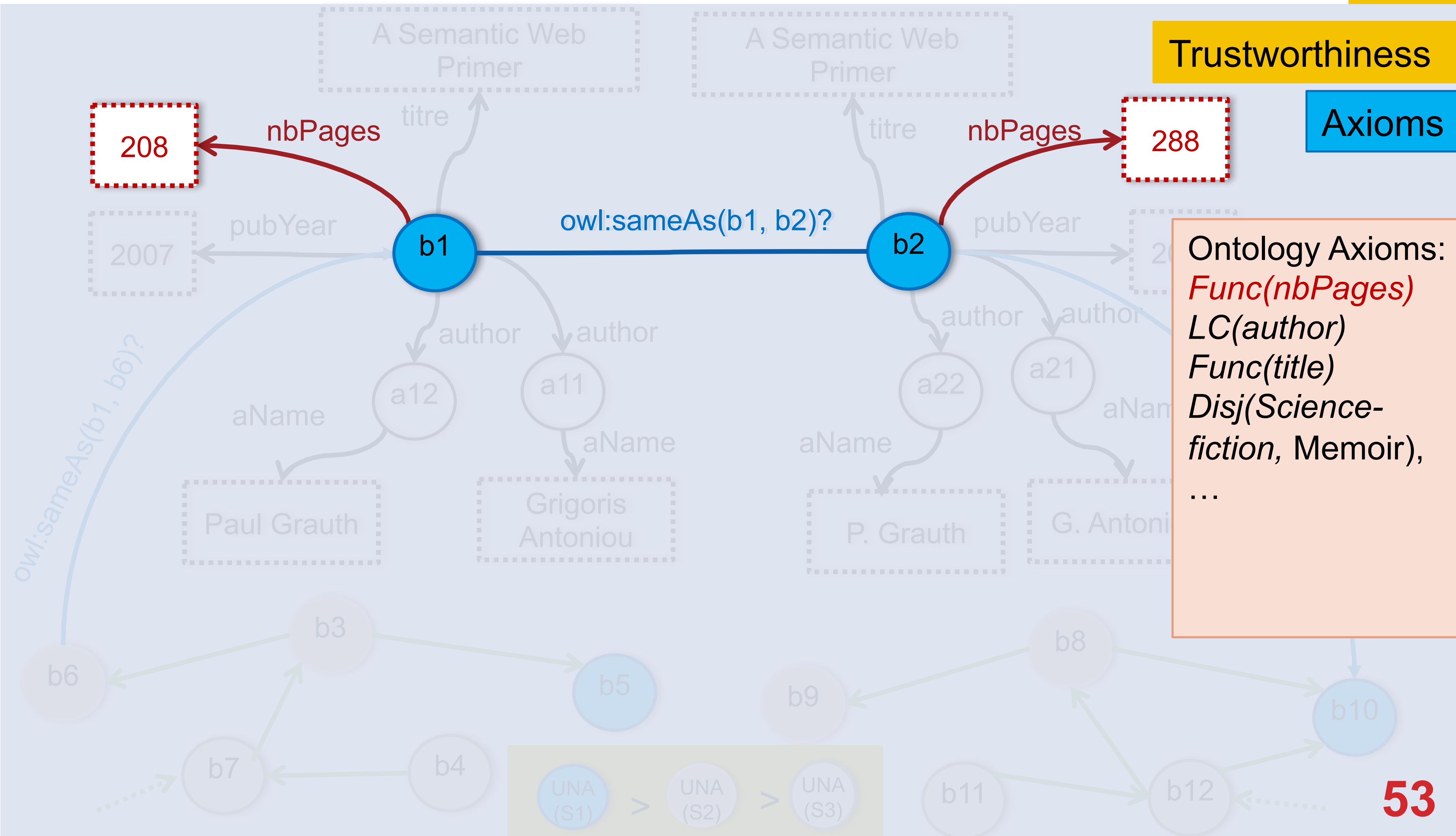
Identity Network

UNA

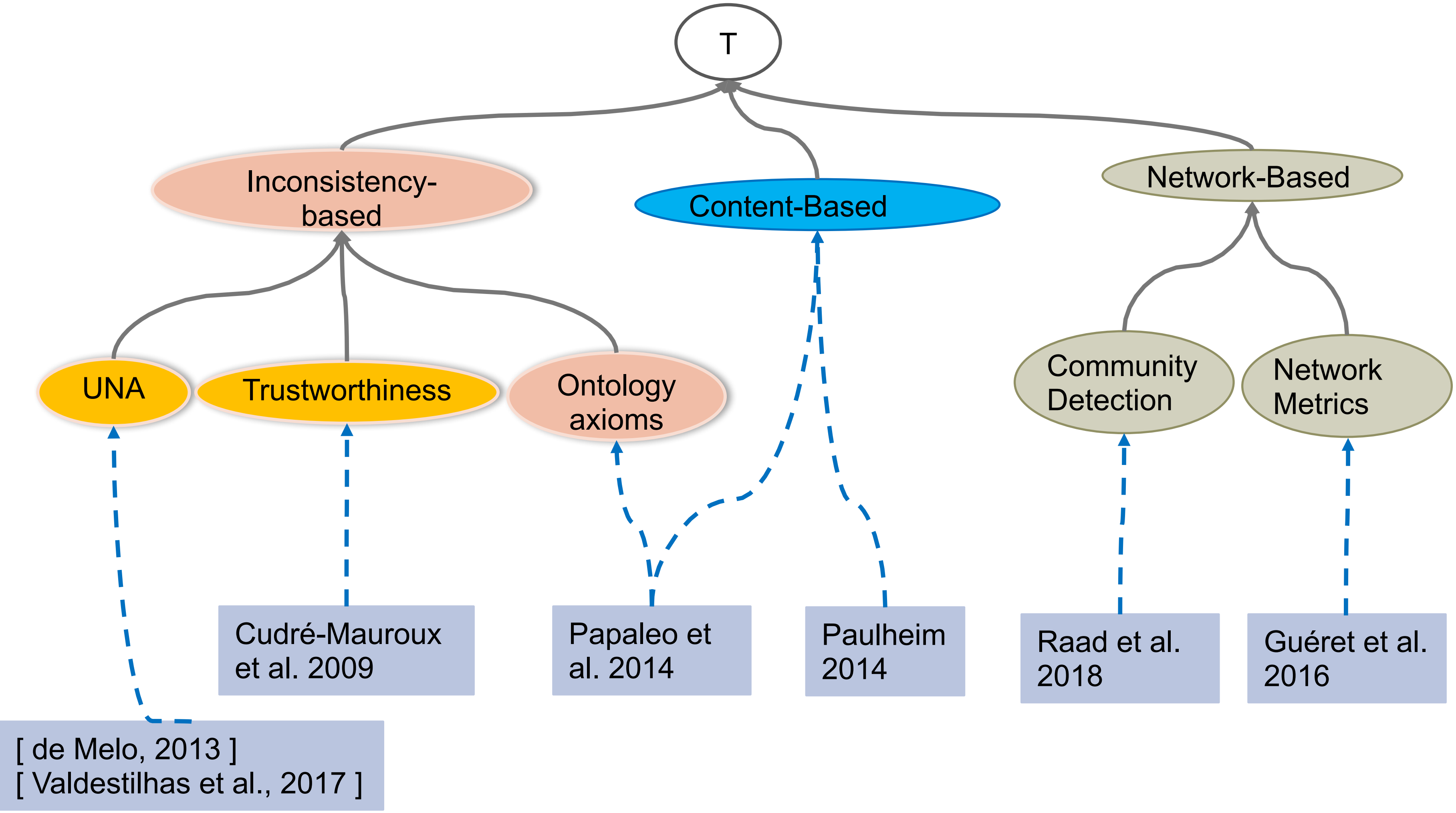
Trustworthiness

Axioms

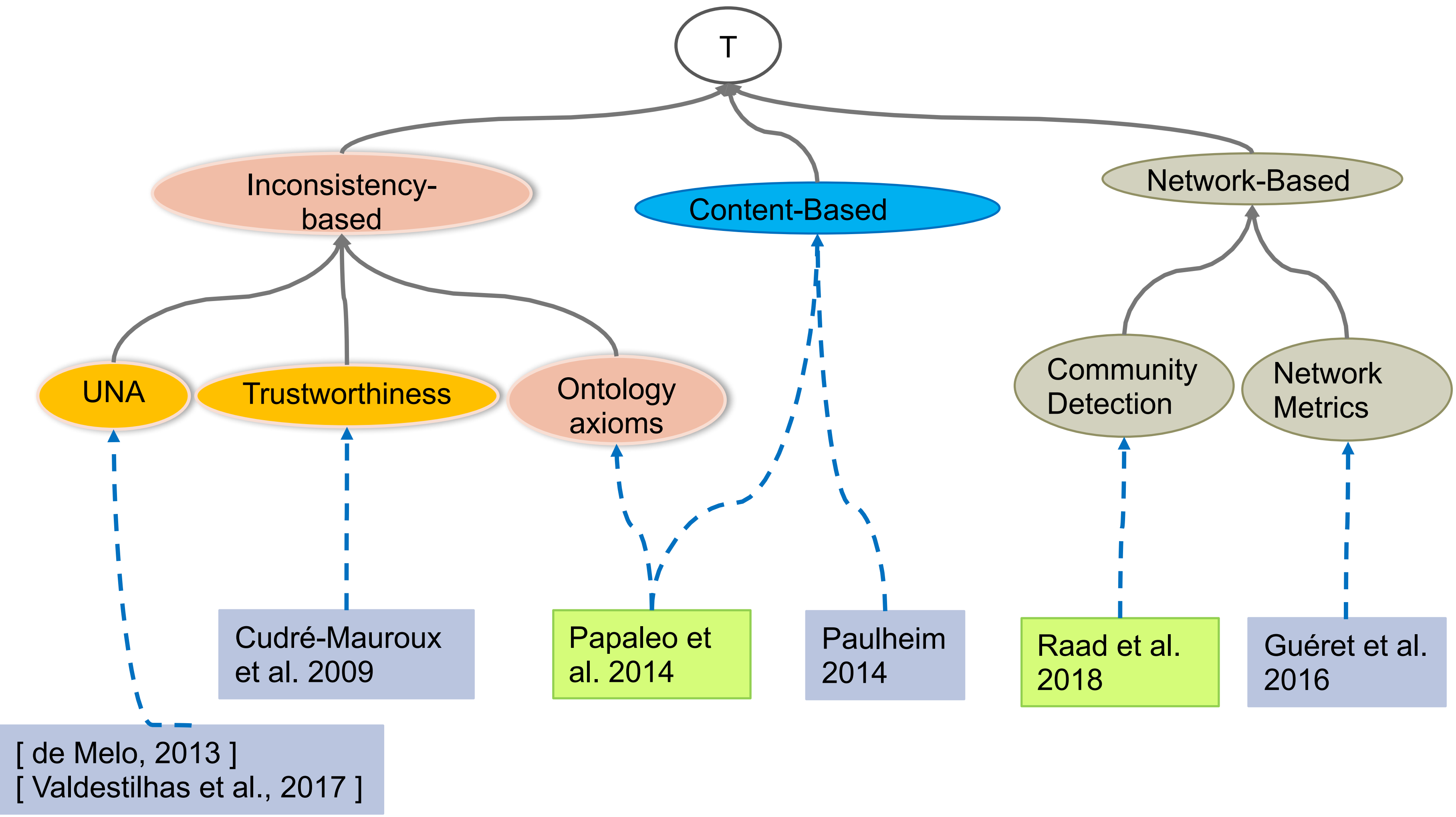
Which kind of information to use for detecting erroneous Identity links?



DETECTION OF ERRONEOUS IDENTITY LINKS



DETECTION OF ERRONEOUS IDENTITY LINKS



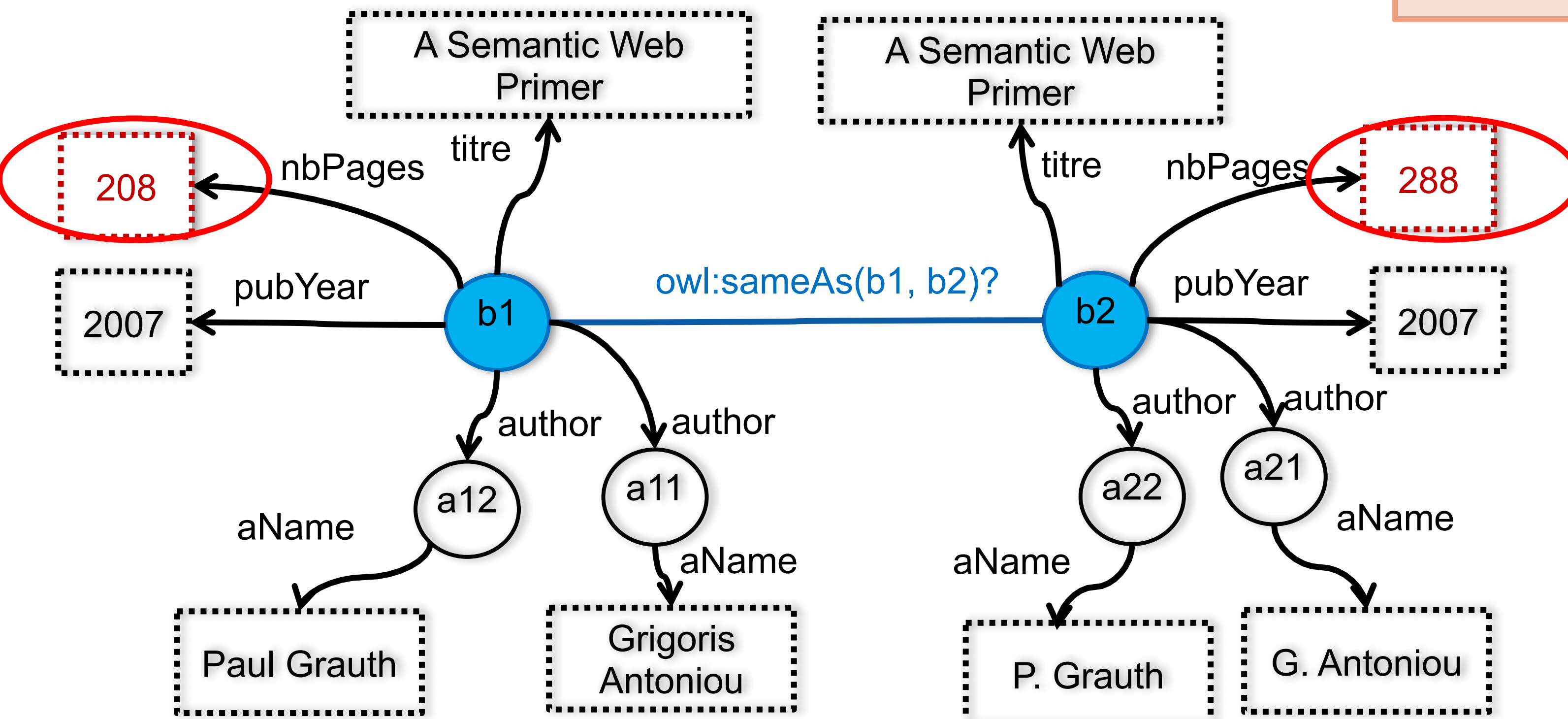
AXIOM-BASED APPROACHES

ONTOLOGY AXIOM VIOLATION

[Papaleo *et al.*, 2014]

Principle: use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies or error candidates in the linked resources descriptions.

nbPages is a Functional Property



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]

F is the set of RDF facts enriched by a set of ¬synVals facts in the form

¬synVals(w_1, w_2)

w_1 and w_2, being literals and different.

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

EXAMPLES:
- notSynVals('231','100')
for a functional property *nbPages*

-notSynVals('New York', 'Paris')
for a functional property *cityName*

… knowledge from expert or extracted.

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

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

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

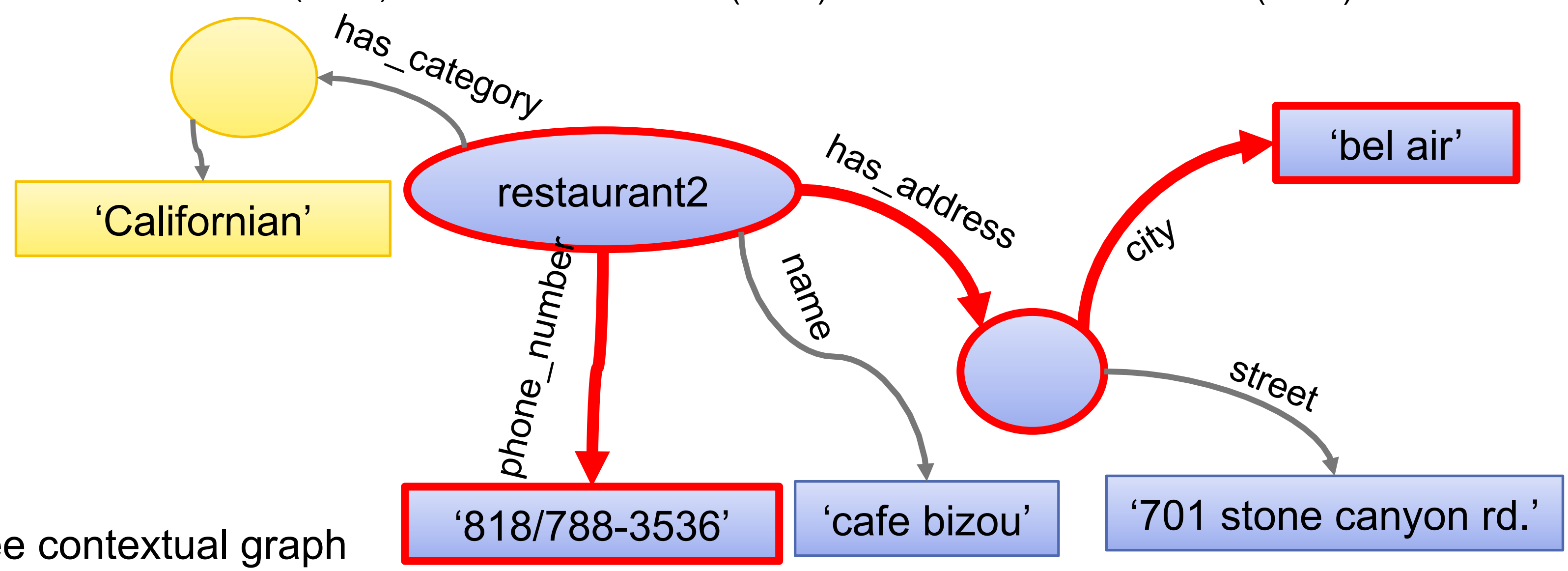
[Papaleo et al. 2014]

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

[1] Sais 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].

LM	LM Precision	linkInv precision	LM+linkInv precision
2	95.55%	37%	**98.85%**
1	69.71%	88.4%	**95.19%**
3	90.17%	42.30%	**100%**

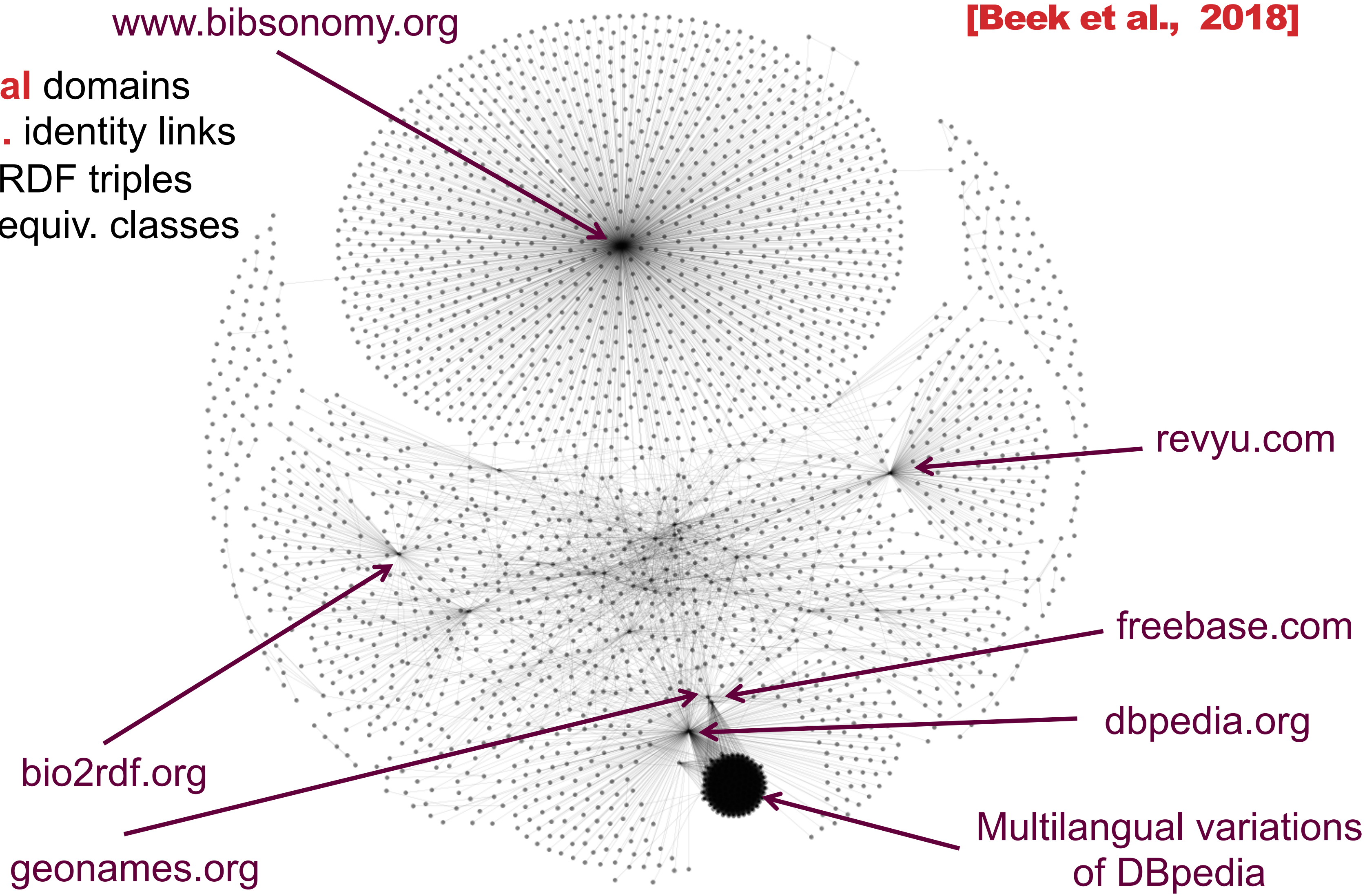
Improvement in precision

Limits:
- Scalability problems
- Need of uniform vocabulary in datasets

IDENTITY PROBLEM AT LOD SCALE

[Beek et al., 2018]

> **Several** domains
> **558 M.** identity links
> **28 B.** RDF triples
> **48 K.** equiv. classes



http://sameas.cc/explicit/img

IDENTITY PROBLEM AT LOD SCALE

[Beek et al., 2018]

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

SameAs.cc Documentation Identity sets Terms Triples

Terms for identity set 4073

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

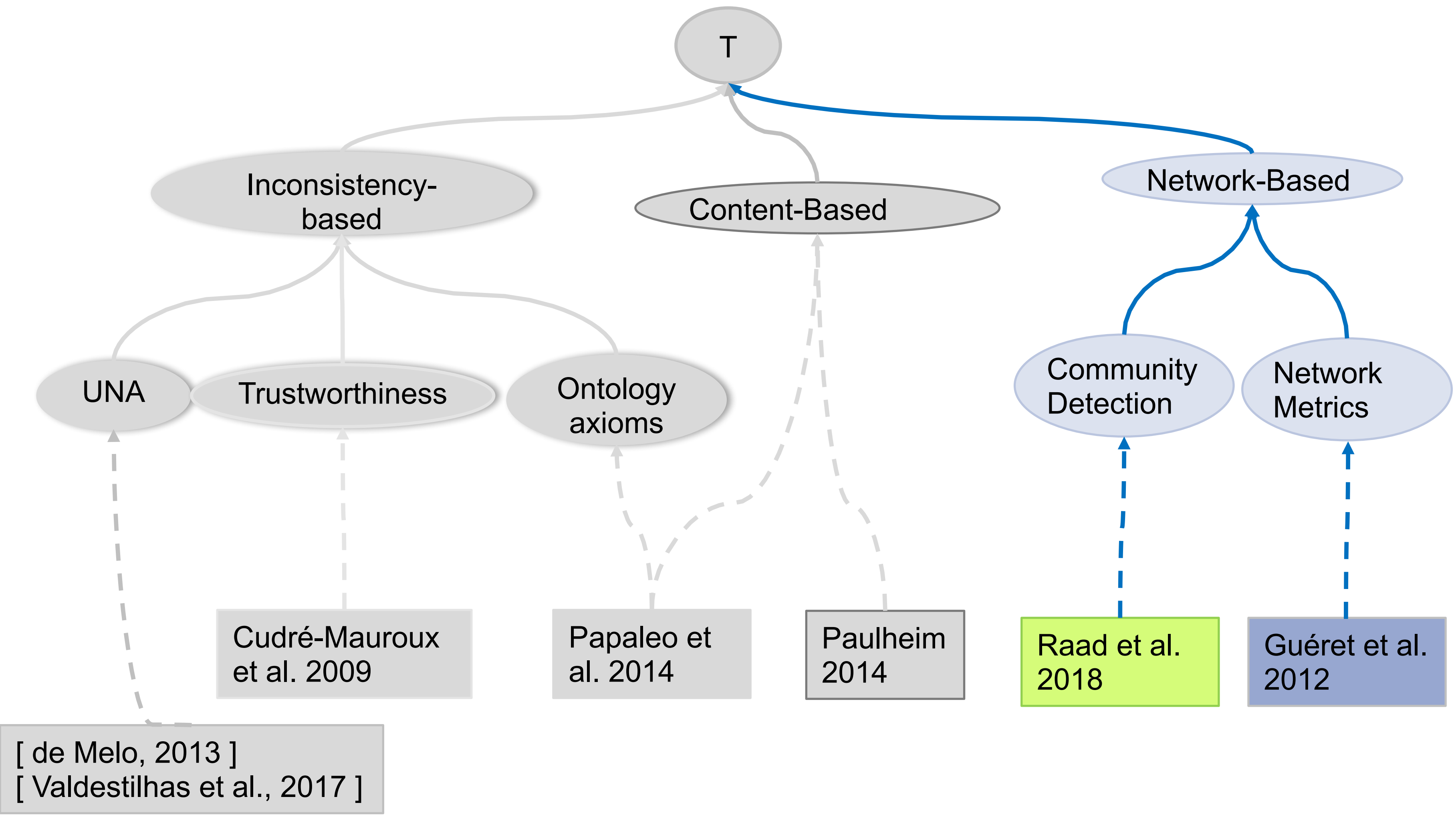
Previous results 1 to 20 (of 177,794) Next

The largest identity set contains 177 794 terms:

Different countries
Different cities
Albert Enstein

→ quality problems

IDENTITY LINK INVALDIATION



NETWORK BASED LINK INVALIDATION

[Guéret *et al.*, 2012]

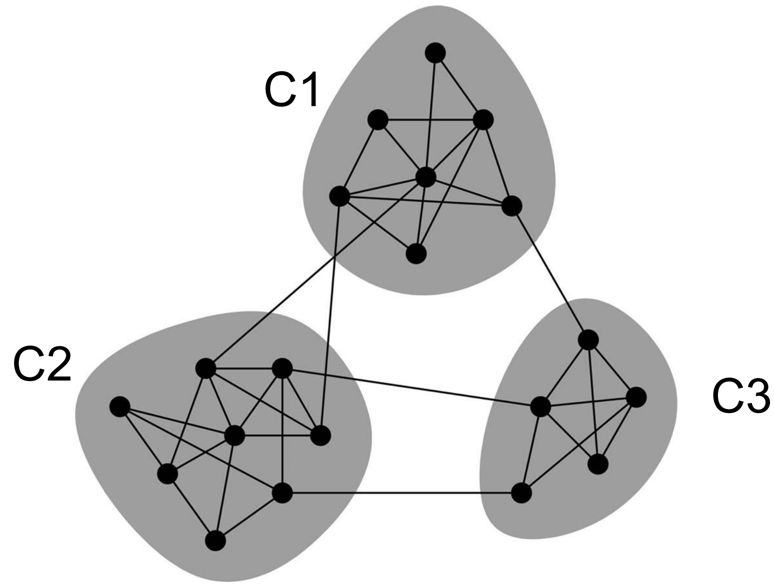
[Raad *et al.*, ISWC 2018]

Principle

- The quality of a link can be determined based on **how connected a node** is within the **network** in which it appears.
- Use of **network metrics and structures** can help to detect erroneous links?

NETWORK BASED

[Raad *et al.*, ISWC 2018]



- Considers the **identity network** build from the **explicit identity network** of sameAs links: removing of symmetric and reflexive links.
- Uses of Louvain **community detection** algorithm to detect subgraphs in the **identity network** that are highly connected.
- Defines a **ranking score** for each (intra-community and inter-community) identity link based on the **density of the community**.

NETWORK BASED

[Raad *et al.*, ISWC 2018]

Ranking of identity links

intra-community erroneousness degree

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

inter-community erroneousness degree

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

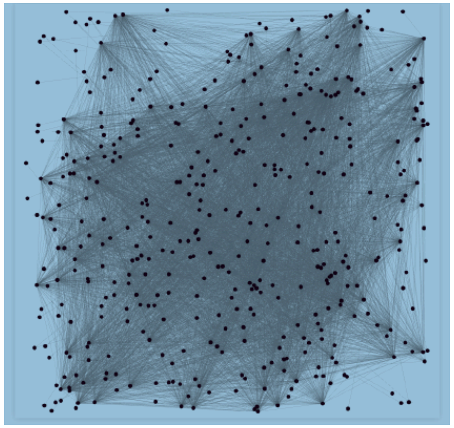
Correct link 0 — Error degree — 1 Erroneous link

NETWORK BASED

[Raad *et al.*, ISWC 2018]

Dataset

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of 28B triples from LOD 2015 crawl
- An **explicit identity network** of 558.9M edges (links) and 179M nodes (resources)

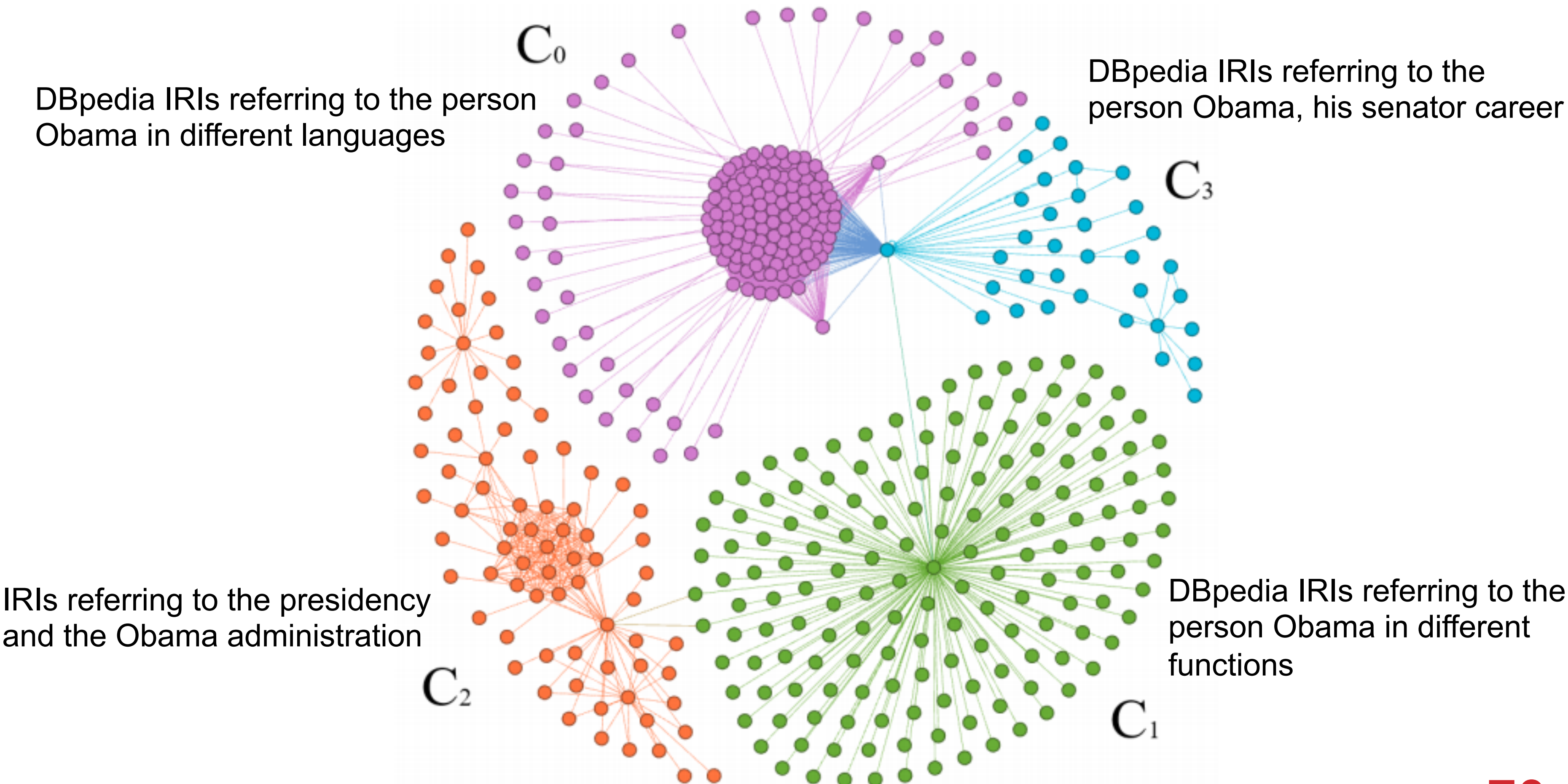


Example: The *B. Obama* equality set that contain 440 nodes

NETWORK BASED

[Raad *et al.*, ISWC 2018]

Barack Obama's Equality Set

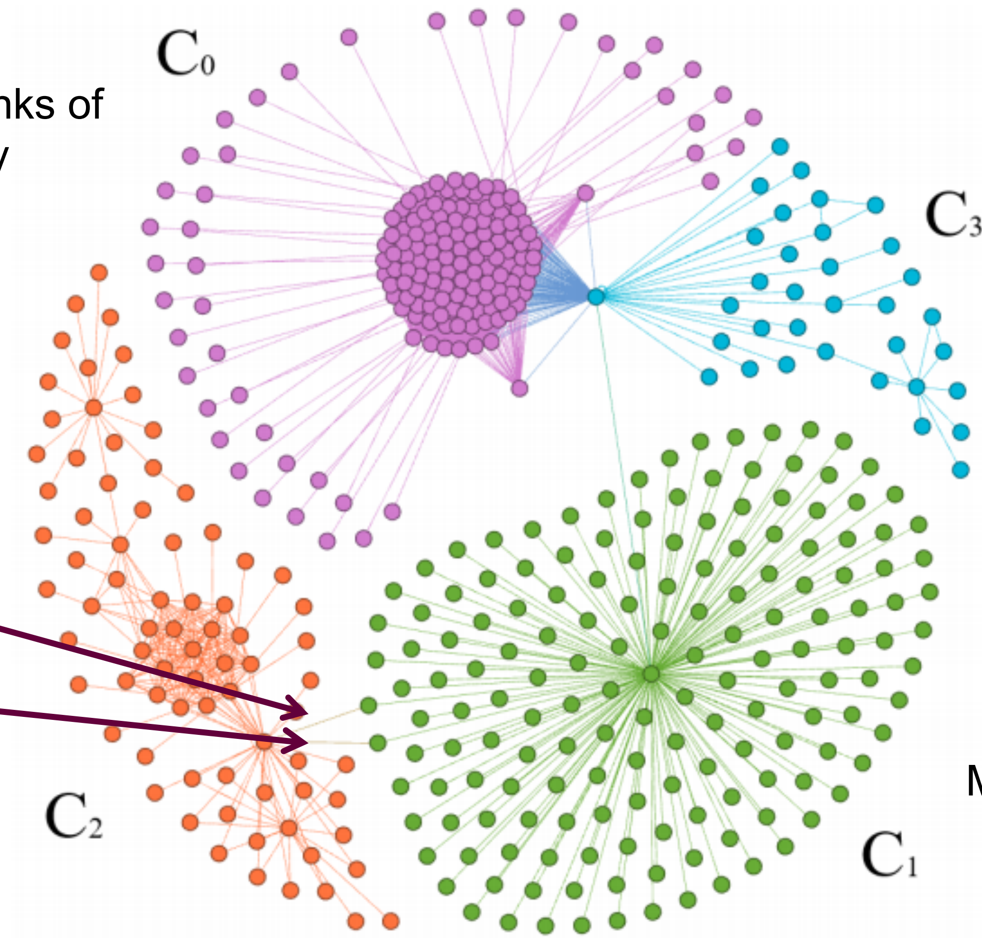


NETWORK BASED

[Raad *et al.*, ISWC 2018]

Barack Obama's Equality Set

Low err(e) for the links of this community



These two links have err(e)= 1

Most of the links have err (e) = 0.9

LINK INVALIDATION: NETWORK-BASED APPROACH EVALUATION

[Raad et al. 2018]

- **Scales** to a graph of **28 billion** triples: **11 hours for the 4 steps**

No **benchmark** for qualitative evaluation

Precision: manual evaluation of 200 links

- The higher the error degree is the most likely the link will be 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% (1.26M** owl:sameAs) have an **error degree** in [0.99, 1]

Recall: 780 incorrect links between **40 distinct** resources have been introduced in the explicit identity graph. **Recall** = **93 %**

IDENTITY MANAGEMENT: SUMMARY

Identity invalidation

- **Different kinds of information can be used for link invalidation:** axioms, resource descriptions and graph topology
- **The efficiency** of the proposed approaches depends on **the characteristics** of the knowledge graphs: volume, heterogeneity, ontology

IDENTITY MANAGEMENT: SUMMARY

Identity invalidation

- **Different kinds of information can be used for link invalidation:** axioms, resource descriptions and graph topology
- **The efficiency** of the proposed approaches depends on **the characteristics** of the knowledge graphs: volume, heterogeneity, ontology

Possible improvements

- Need for hybrid approaches for link invalidation
- Need for well-formalized **weak-identity**: contextual identity, similarity, …
- Need for approaches for **difference links** detection: useful for inconsistency checking

CONCLUSION

- **Semantic Web standards, data and many applications are there**
- **Promising applications are emerging for which reasoning on data is central:**
  - Web search, recommendation systems, chat-bots, …

- **Many challenges remain to handle at large scale the incomplete, uncretain and evolving knowledge graphs**
  - Combining numerical and symplic AI is challenging but worthwhile to investigate more deeply.

KNOWLEDGE GRAPH REFINEMENT

KEY DISCOVERY AND LINK INVALIDATION

FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

Joint work with: N. Pernelle, L. Papaleo, J. Raad and D. Symeonidou

3ÈME JOURNÉE RI-IA SOUTENUE PAR L'AFIA ET ARIA, PARIS 2019

REFERENCES (1)

[Atencia et al. 2014] Manuel Atencia, Jérôme David, Jérôme Euzenat:

Data interlinking through robust linkkey extraction. ECAI 2014: 15-20

[Al-Balkri et al. 2015] Mustafa Al-Bakri, Manuel Atencia, Steffen Lalande, Marie-Christine Rousset:

Inferring Same-As Facts from Linked Data: An Iterative Import-by-Query Approach. AAAI 2015: 9-15

[Al-Bakri et al 2016] Mustafa Al-Bakri, Manuel Atencia, Jérôme David, Steffen Lalande, Marie-Christine Rousset: Uncertainty-Sensitive Reasoning for Inferring sameAs Facts in Linked Data. ECAI 2016: 698-706

[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.

REFERENCES (2)

[Guéret et al. 2012] C. Guéret, P. Groth, C. Stadler, and J. Lehmann.

Assessing linked data mappings using network measures. In ESWC 2012

[Halpin et al., 2010] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson.

When owl:sameAs isn't the same: An analysis of identity in Linked Data. In ISWC 2010.

[Hogan et al., 2012] A. Hogan, A. Zimmermann, J. Umbrich, A. Polleres, and S. Decker.

Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. In JWS 2012.

[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.

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

REFERENCES (3)

[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

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

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

[Soru et al. 2015] Tommaso Soru, Edgard Marx, Axel-Cyrille Ngonga Ngomo:

ROCKER: A Refinement Operator for Key Discovery. WWW 2015: 1025-1033

[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.

[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.