KNOWLEDGE GRAPH REFINEMENT

FATIHA SAÏS

HDR-HABILITATION À DIRIGER DES RECHERCHES

JUNE 20th, 2019, ORSAY, FRANCE

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®
cnrs dépasser les frontières
université PARIS-SACLAY

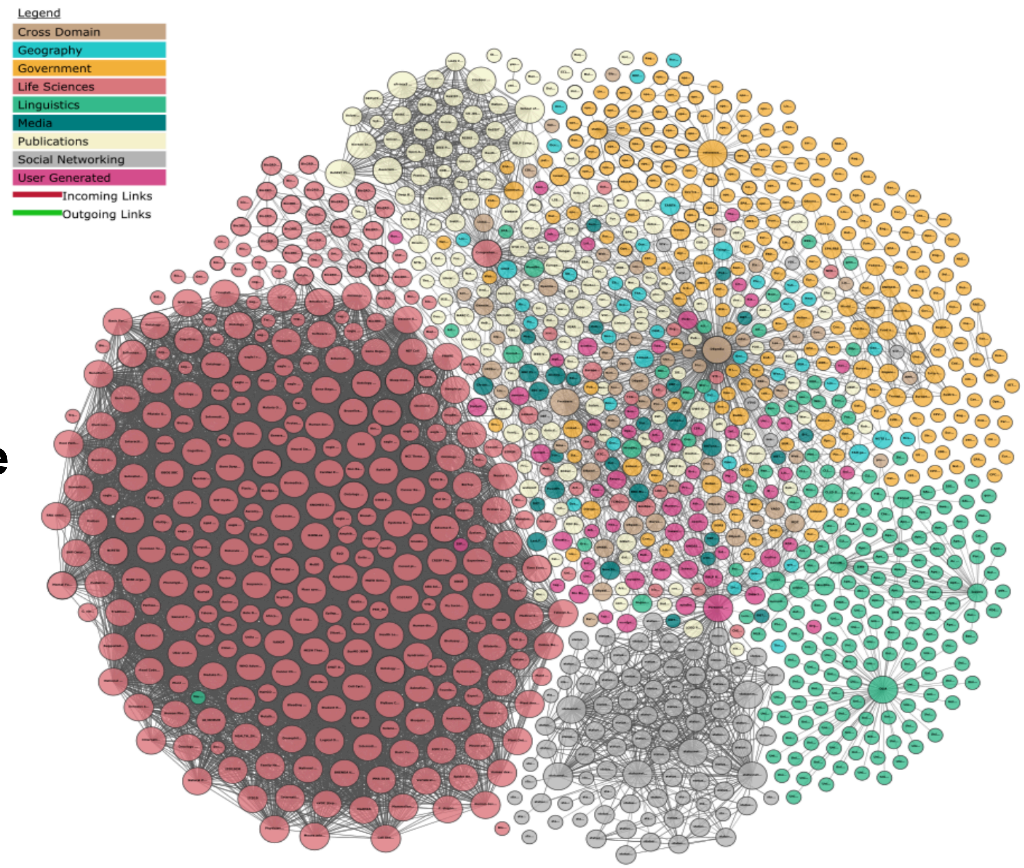
LINKED DATA

Tim Berners Lee, 2006

RDF Datasets publicly available

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

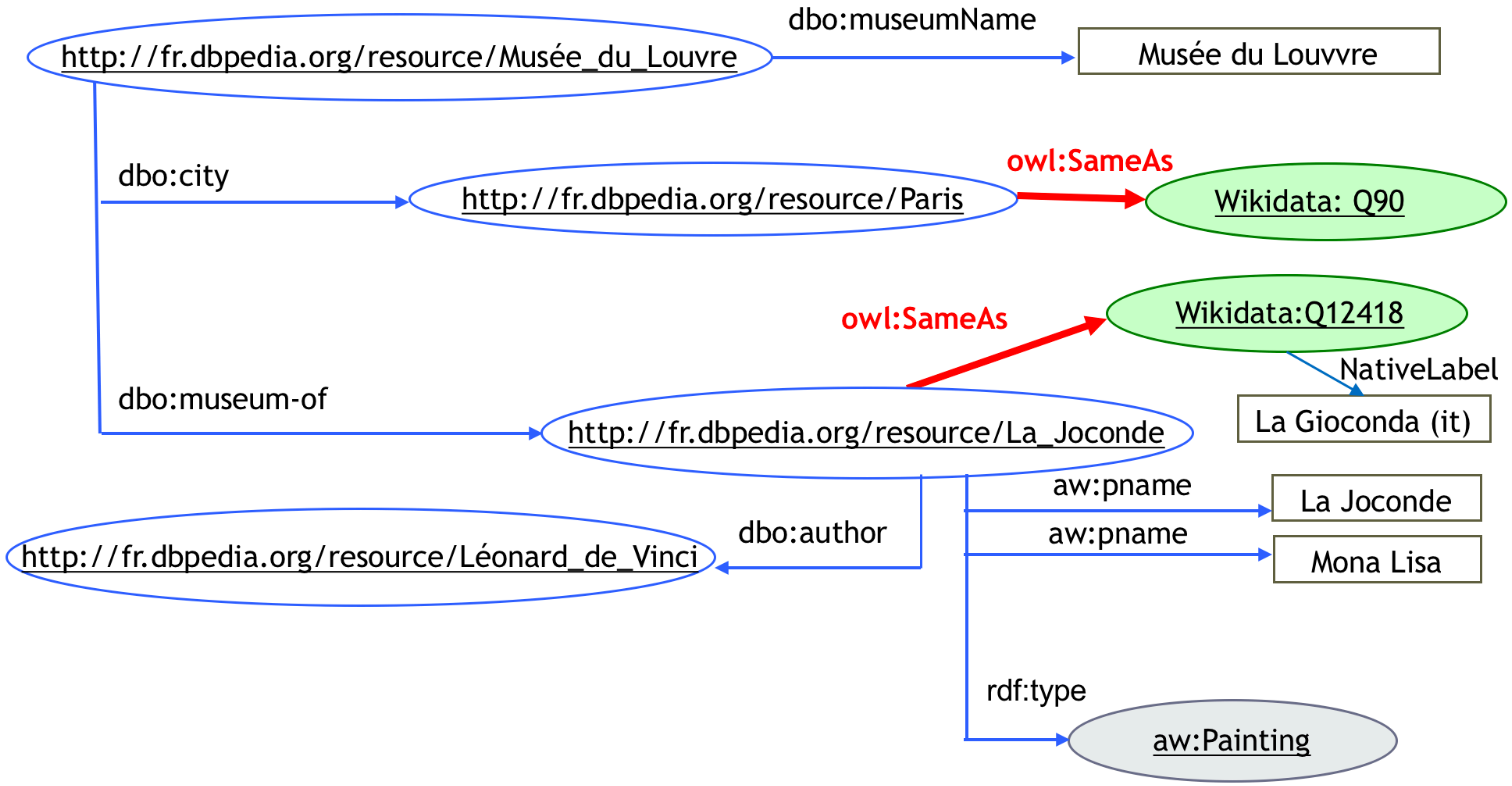
LOD – Linked Data Cloud



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

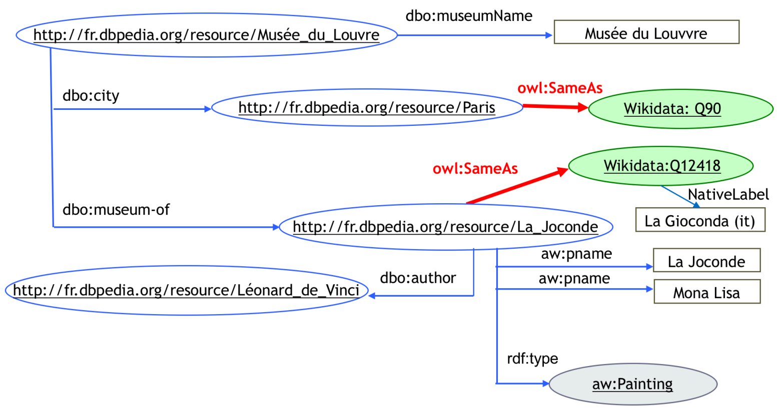
KNOWLEDGE GRAPHS (KG)

RDF Graphs



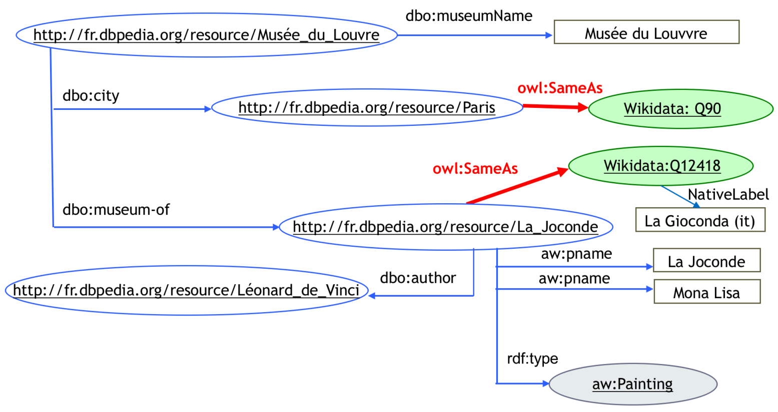
KNOWLEDGE GRAPHS (KG)

RDF Graphs

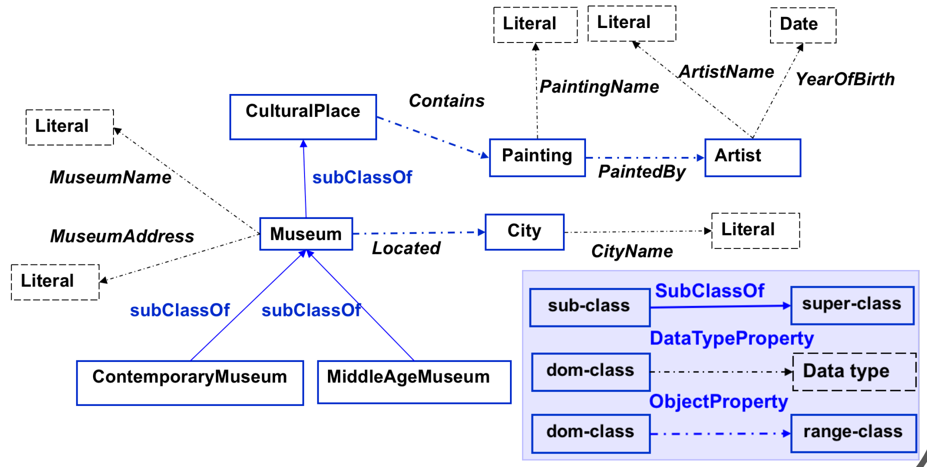


KNOWLEDGE GRAPHS (KG)

RDF Graphs



OWL Ontology



Ontology axioms and rules

- Disjunction between classes/properties
- (inverse) Functionality of properties
- Symmetry
- Keys
- Logical rules
- ...

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2012

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

2008

yago select knowledge

2012

WIKIDATA

2007

Freebase™

2012

Academic side

Commercial side

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

2008

yago select knowledge

2012

WIKIDATA

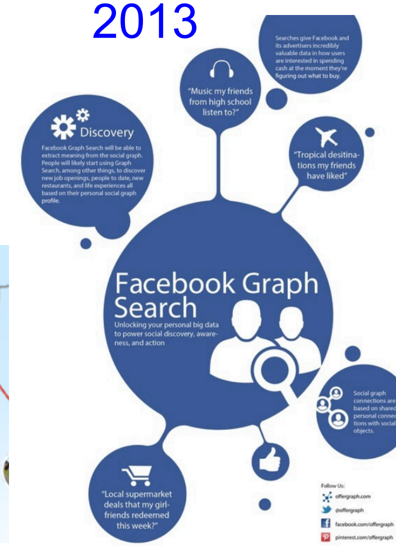
2007

Freebase™

Academic side

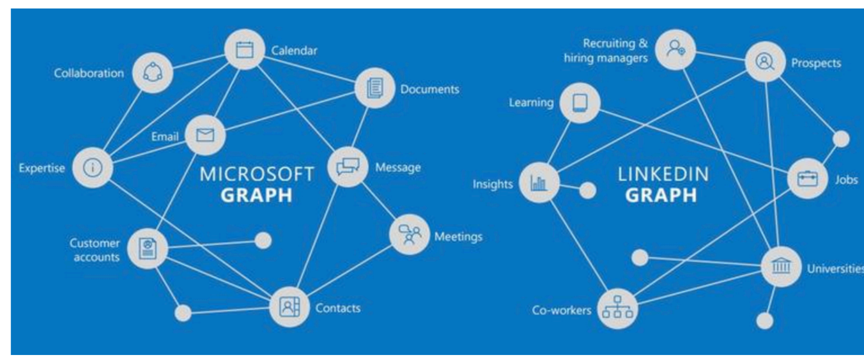
2012

2013



2015

2016



Yahoo!™

Yahoo's new SERP designs mobile and knowledge graph

2013

Commercial side

KNOWLEDGE GRAPH COMPLETENESS?

	Name	Instances	Facts	Types	Relations
public	DBpedia (English)	4,806,150	176,043,129	735	2,813
	YAGO	4,595,906	25,946,870	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	570,000,000	18,000,000,000	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

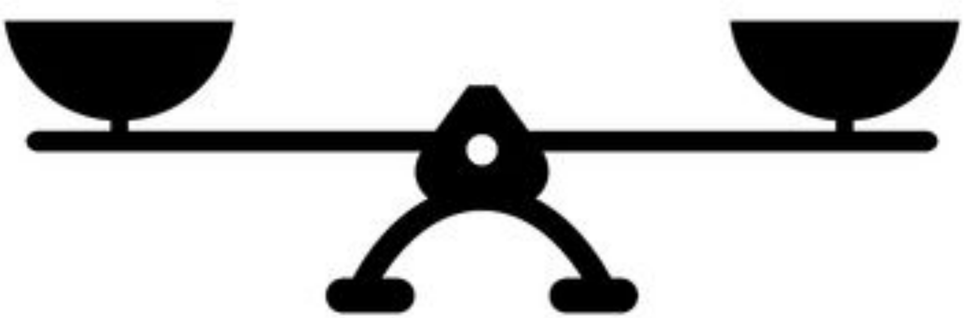
Donald John Trump (born June 14, 1946) is an American businessman, author, television producer, politician, and the Republican Party nominee for President of the United States in the 2016 election. He is the chairman and president of The Trump Organization, which is the principal holding company for 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

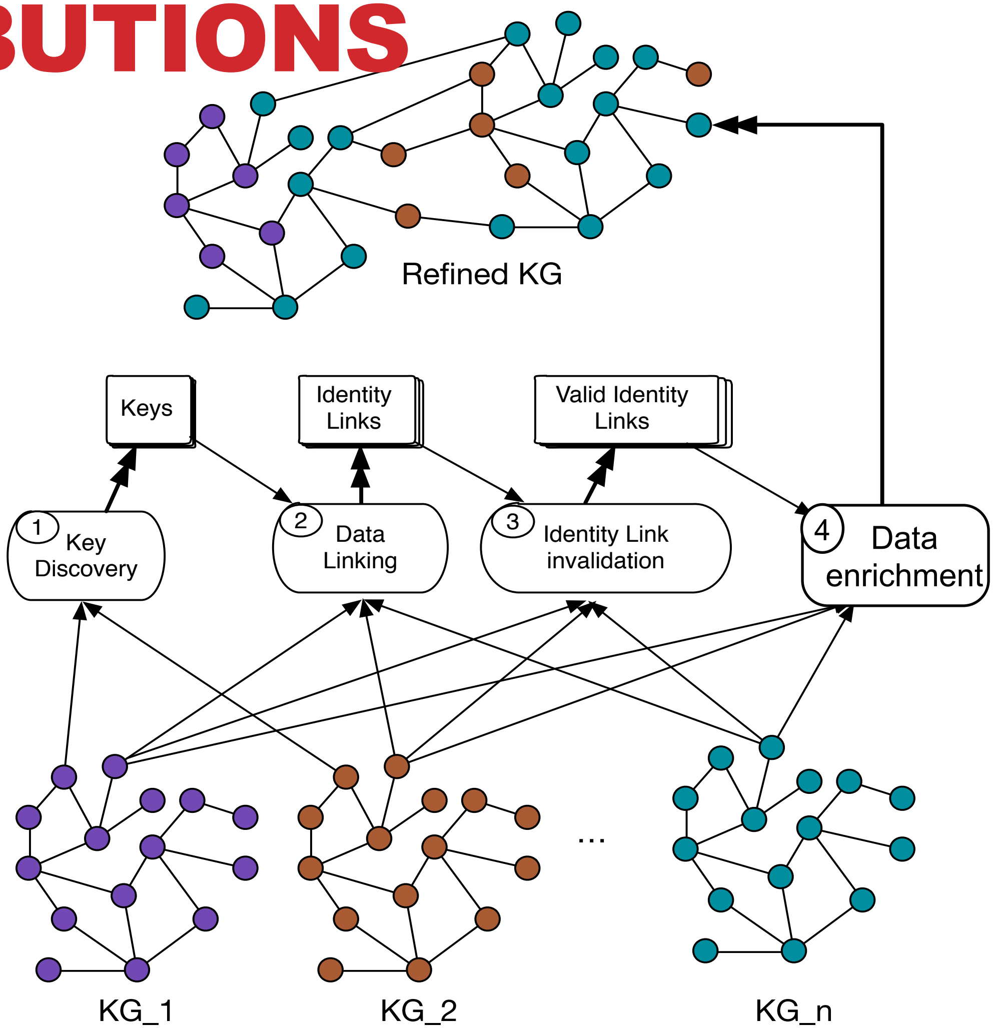
Donald Trump is the child of himself!

KNOWLEDGE GRAPH REFINEMENT

Completeness   Correctness



KNOWLEDGE GRAPH REFINEMENT: SOME CONTRIBUTIONS



Identity management

- **Data Linking**: contextual identity links detection (Completeness)
- **Identity Link Invalidation (Correctness)**

Key discovery

- **Key axiom enrichment**

Data Enrichment

- **Data Fusion:** Property value enrichment
- **Missing value prediction:** Property value enrichment

OUTLINE

- **Introduction**
- **Contributions**
  - Part 1: Identity Management
  - Part 2: Key Discovery
  - Part 3: Data Enrichment
- **Summary and Future Directions**

OUTLINE

- **Introduction**
- **Contributions**
  - **Part 1: Identity Management**
  - Part 2: Key Discovery
  - Part 3: Data Enrichment
- **Summary and Future Directions**

IDENTITY IN KNOWLEDGE GRAPHS

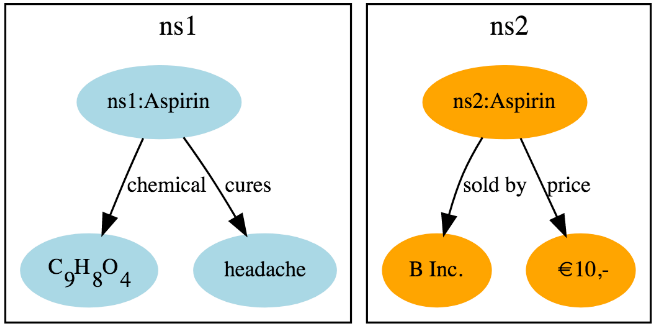
- Indicates that two different descriptions refer to the same entity

- owl:sameAs predicate: a standard for identity representation
- 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 MANAGEMENT IS COMPLEX …

- Data linking tools are rarely **100% precise**
- **Many erroneous owl:sameAs links:**
  - [Halpin et al., 2010] ~**21%** and [Hogan et al., 2012]~**2.8%**, manual evaluation of samples of owl:sameAs links from the Web
- **Identity is context-dependent:**



IDENTITY MANAGEMENT IS COMPLEX …

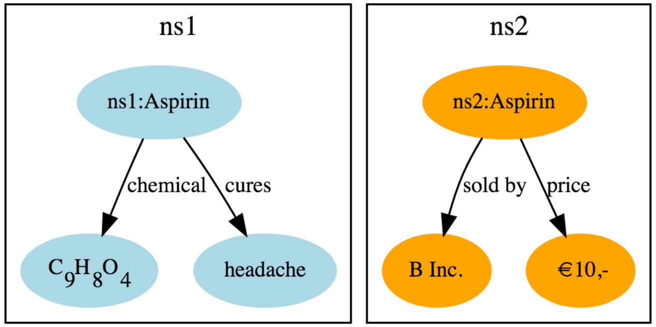
- Data linking tools are rarely **100% precise**
- **Many erroneous owl:sameAs links:**

➔ Link Invalidation problem

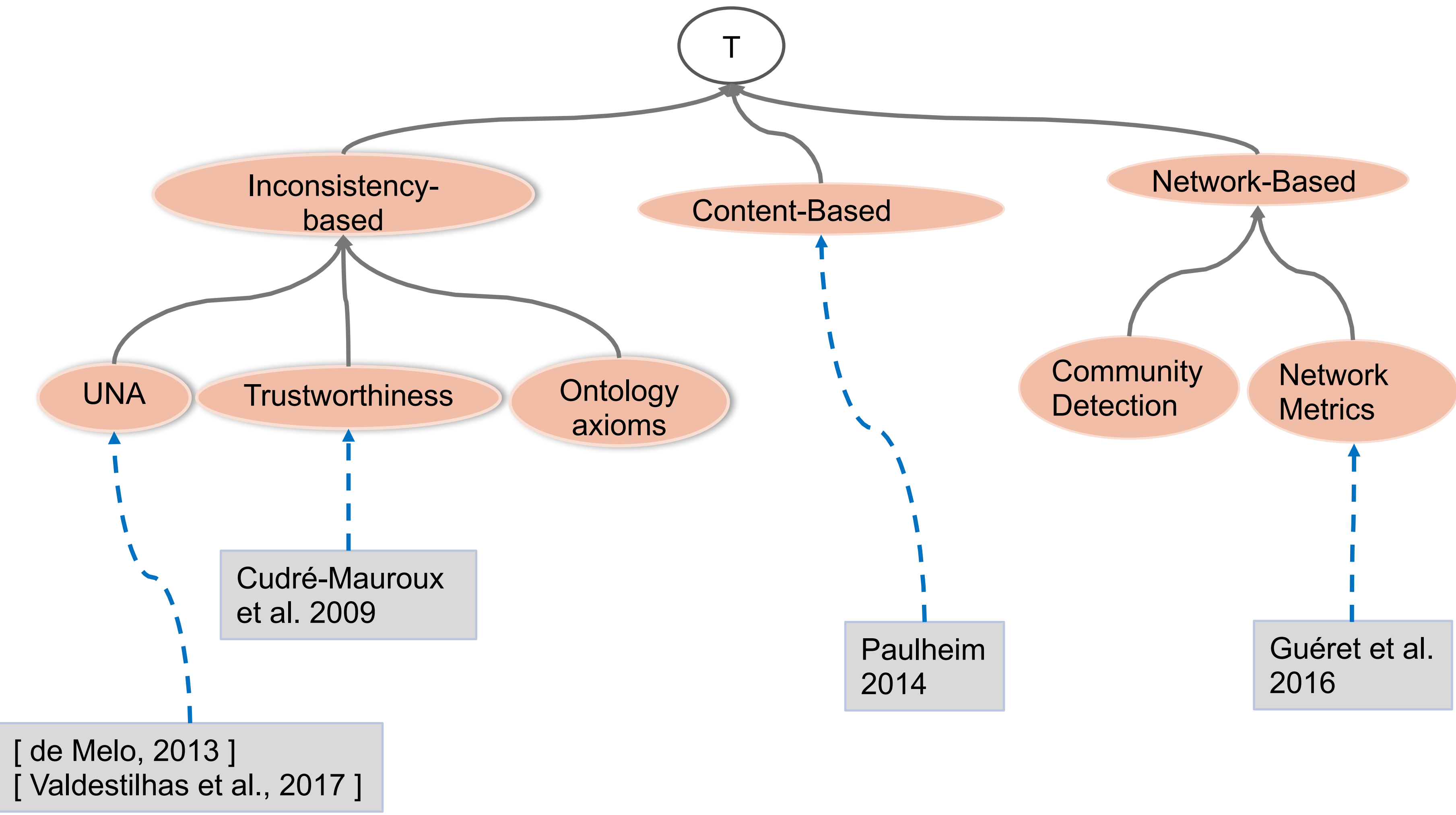
  - [Halpin et al., 2010] ~**21%** and [Hogan et al., 2012]~**2.8%,** manual evaluation of samples of owl:sameAs links from the Web

- **Identity is context-dependent:**

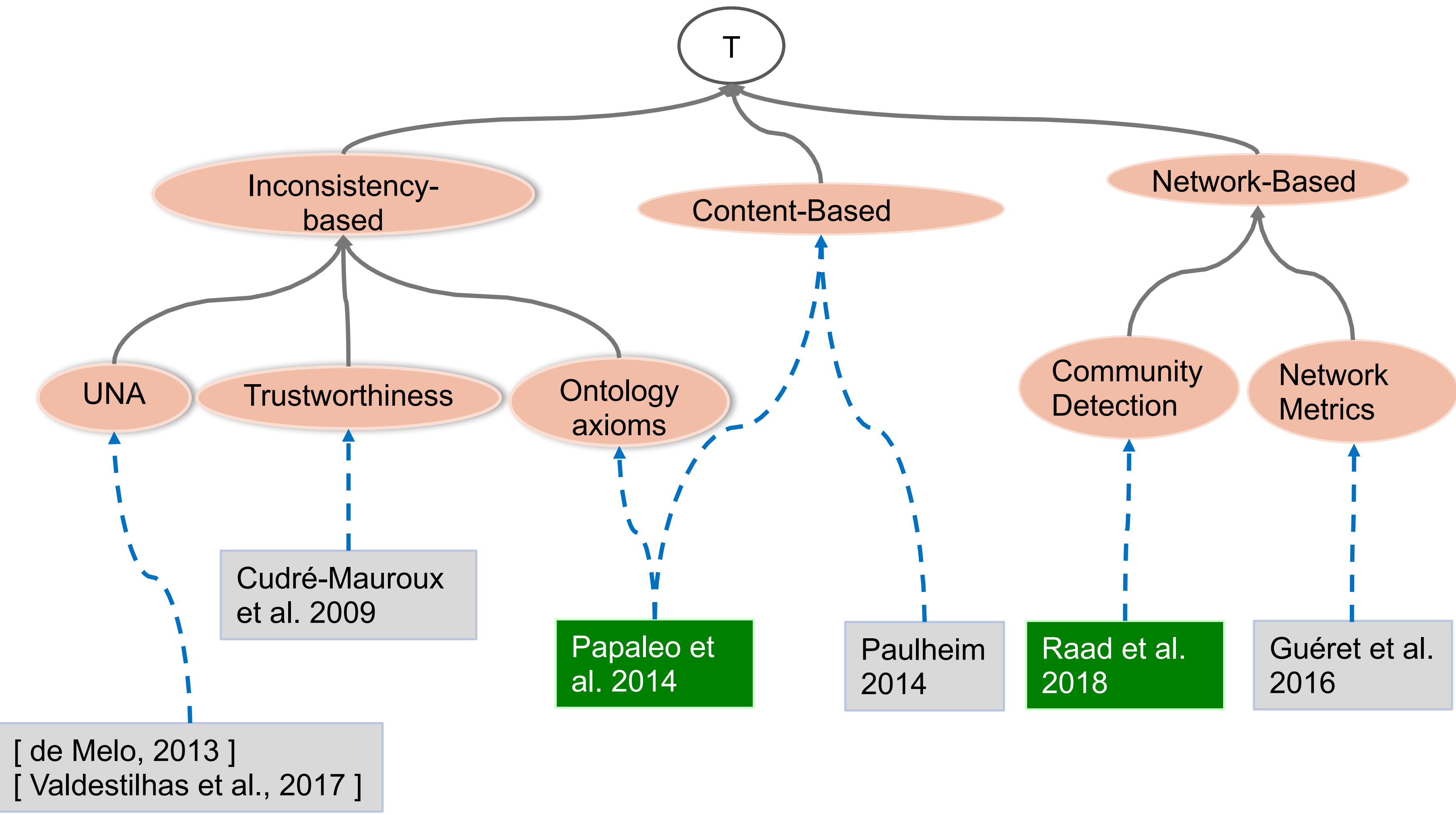
➔ Contextual identity



LINK INVALIDATION: RELATED WORK



LINK INVALIDATION: CONTRIBUTIONS



LINK INVALIDATION: CONTRIBUTIONS

- **Axiom-based for (small) datasets conforming to the same ontology**

L. Papaleo Post-Doc

QUALINCA ANR Project [2012-2016]

- **Network-based for big datasets without any assumption**

J. Raad PhD, co-supervised with N. Pernelle, J. Dibie and L. Ibanescu

Collaboration with VU Amsterdam (NL)

LIONES project from CDS Paris Saclay [2015-2018]

LINK INVALIDATION: AXIOM-BASED APPROACH

[Papaleo et al. 2014]

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

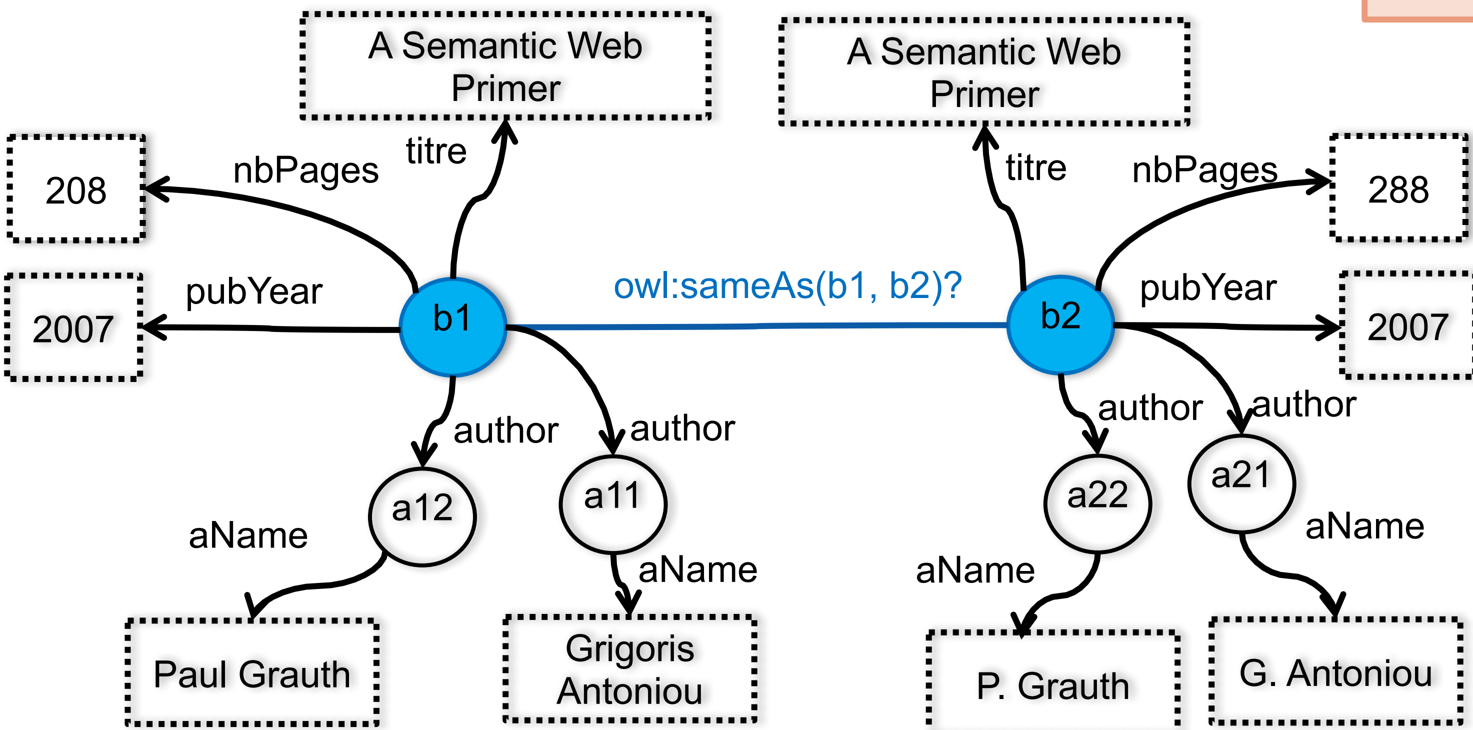
Axiom: nbPages is a Functional Property

LINK INVALIDATION: AXIOM-BASED APPROACH

[Papaleo et al. 2014]

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

Axiom: nbPages is a Functional Property

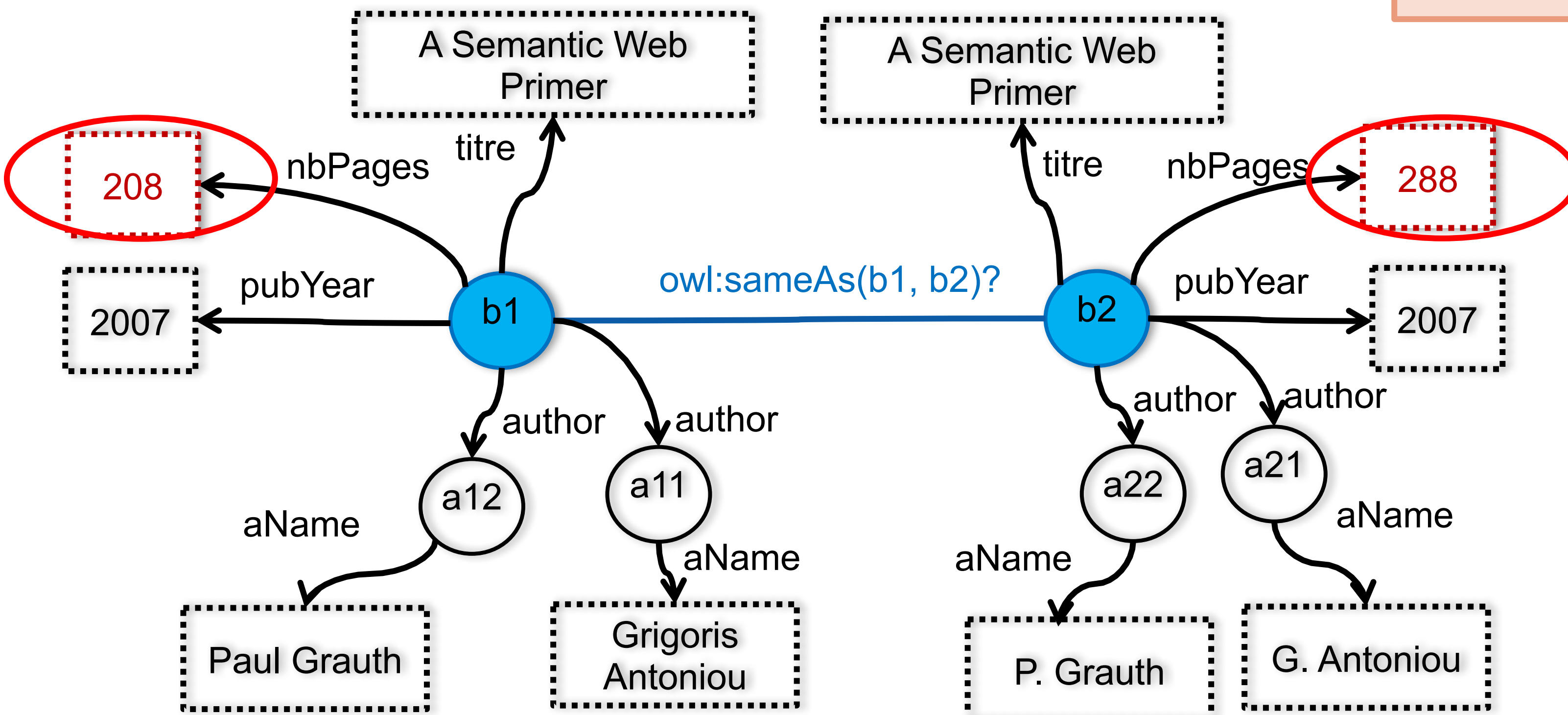


LINK INVALIDATION: AXIOM-BASED APPROACH

[Papaleo et al. 2014]

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

Axiom: nbPages is a Functional Property



LINK INVALIDATION: AXIOM-BASED APPROACH

[Papaleo et al. 2014]

Algorithm: builds a sub-graph «around» each one of the two resources involved in the owl:sameAs by exploiting ontology axioms

- Applies a logical reasoning based on Unit Resolution on:
  - Facts: set of RDF facts of the sub-graph and initial inequalities between literals
  - Rules: rules expressing the axiom semantics

$$- R_{1_{FDP}} : sameAs(x,y) \wedge p_i(x,w_1) \wedge p_i(y,w_2) \rightarrow synVals(w$$

sameAs(x,y) $\wedge$nbPages(x,w_1) $\wedge$ nbPages(y,w_2) → SynVals(w_1,w_2)

LINK INVALIDATION: AXIOM-BASED APPROACH EVALUATION

[Papaleo et al. 2014]

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools.

IM: Invalidation method, LM: Linking method

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%

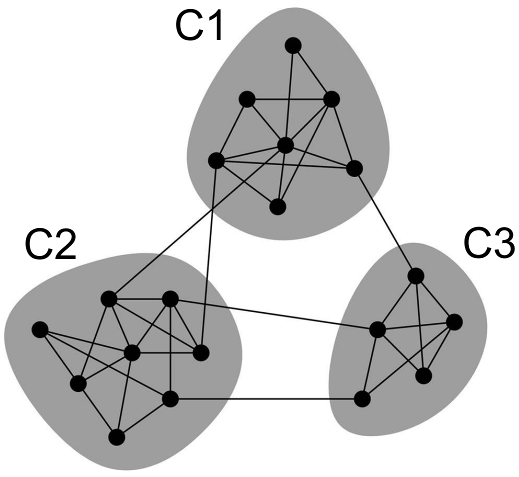
Precision Improvement up to 25%

- **Limitations**
  - Not scalable (evaluation on some thousands of instances)
  - Strong assumptions: same ontology and axioms available

LINK INVALIDATION: NETWORK-BASED APPROACH

[Raad et al. 2018]

Algorithm: uses the density of the **community structure** of the **identity graph** to assign each link an **error degree**.



Identity graph

LINK INVALIDATION: NETWORK-BASED APPROACH

[Raad et al. 2018]

Algorithm: uses the density of the **community structure** of the **identity graph** to assign each link an **error degree**.

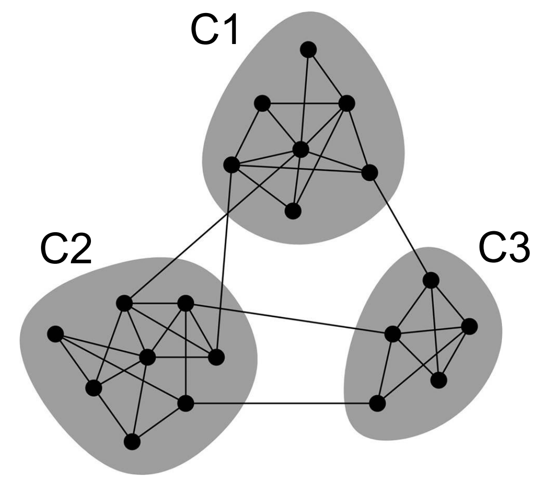
Intra-community link

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

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

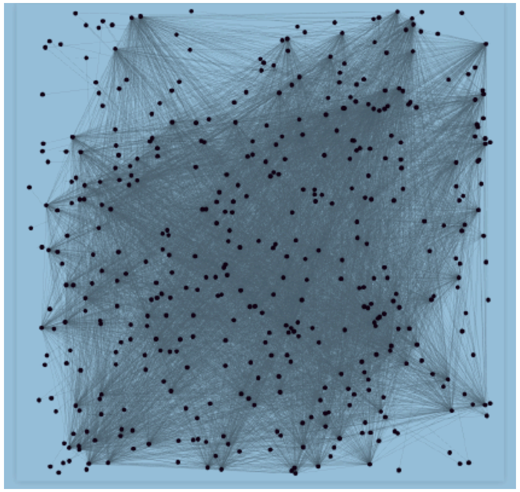


Identity graph

LINK INVALIDATION: NETWORK-BASED APPROACH EVALUATION [Raad et al. 2018]

Experimentation - Dataset

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of **28 Billion** triples from a LOD 2015 crawl
- Identity graph of **558.9 Million** owl:sameAs links (179M nodes)
- Partitioned into **48.9 Million** non singleton **equality sets**



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

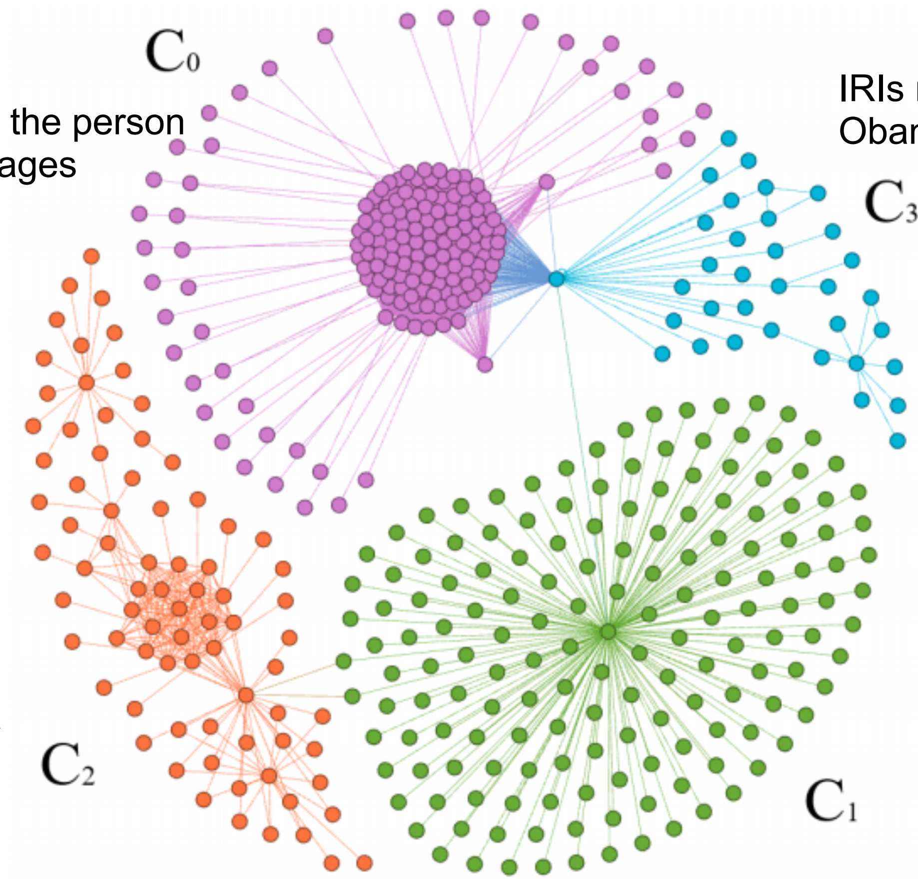
LINK INVALIDATION: NETWORK-BASED APPROACH EVALUATION

[Raad et al. 2018]

The community structure of the *Barack Obama's* Equality Set

DBpedia IRIs referring to the person Obama in different languages

IRIs referring to the person Obama, his senator career



IRIs referring to the Obama administration, government

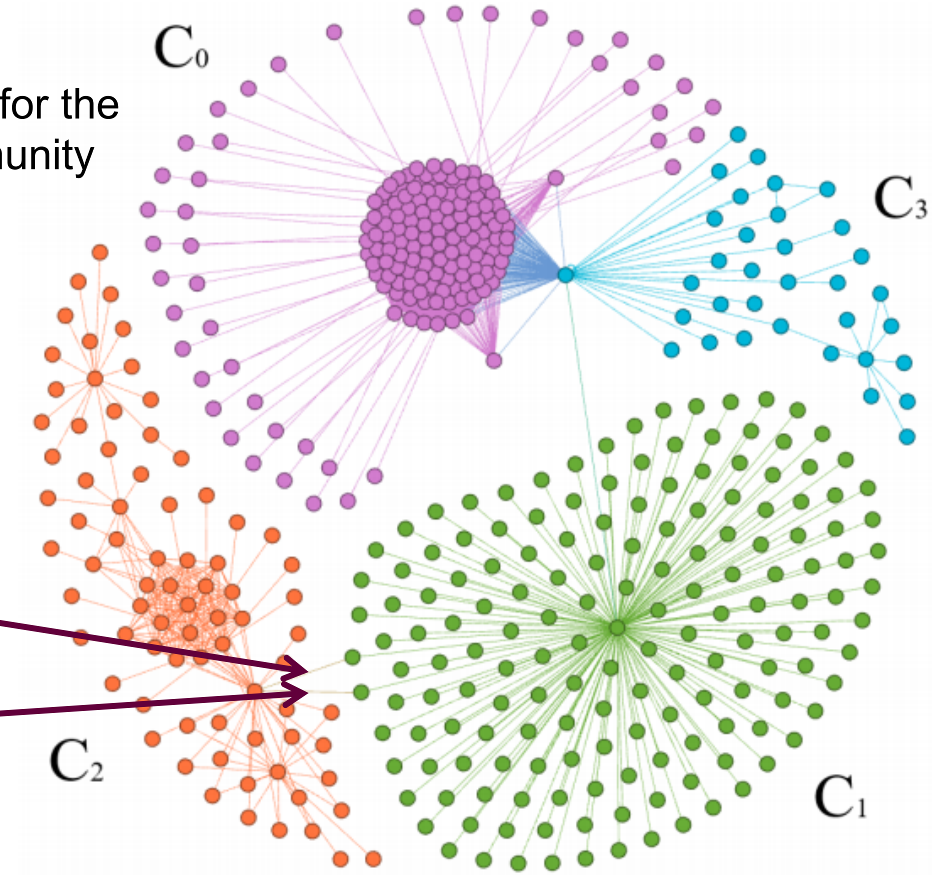
IRIs referring to the person Obama in different functions

LINK INVALIDATION: NETWORK-BASED APPROACH EVALUATION

[Raad et al. 2018]

Error degrees

Low error degrees for the links of this community



err(e)= 1
For these 2 links

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 %**

CONTEXTUAL IDENTITY

- Need to distinguish **weak identity** from **genuine identity**

CONTEXTUAL IDENTITY: RELATED WORK

- Need to distinguish **weak identity** from **genuine identity**
- **Existing alternate links**
  - Similarity ontology (SO) [Halpin et al., 2010]:13 different predicates including 8 new ones
  - UMBEL[1] vocabulary introduces umbel:isLike "*to assert a link between similar individuals who may be believed to be identical*"
  - x No formal semantics
  - x No algorithm proposed for their discovery
- **Weaker** kinds of **identity** expressed as a subset of properties [Beek et al. 2016]

[1] http://umbel.org

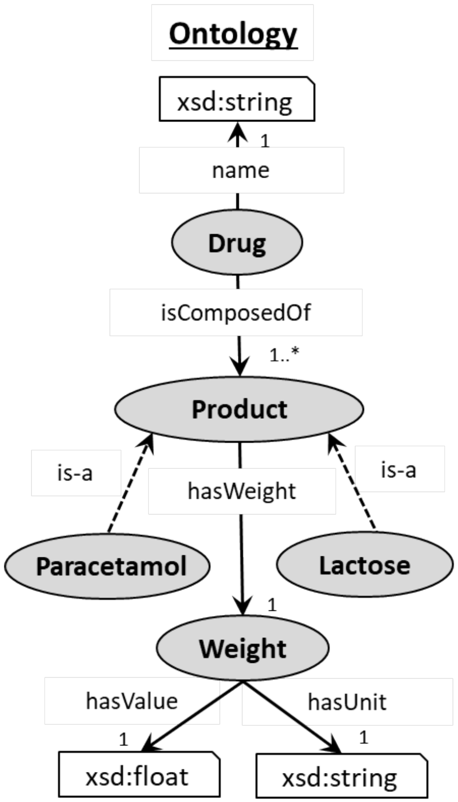
CONTEXTUAL IDENTITY: CONTRIBUTIONS

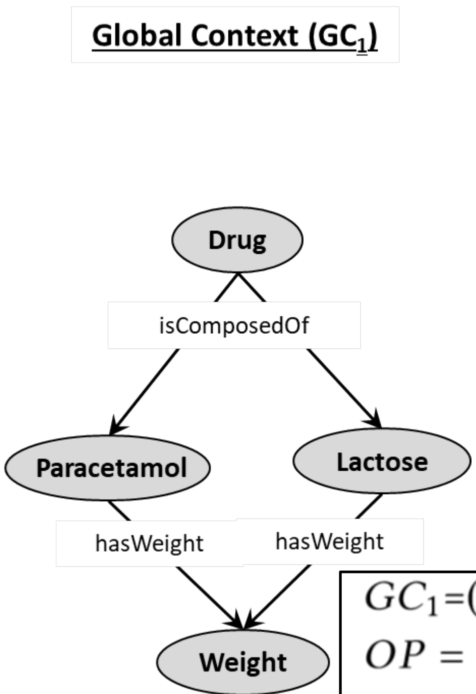
[Raad et al. 2017]

- A **context** defined as a **sub-ontology**
- New **contextual identity** predicate
- New **algorithm** for detecting the **most specific contexts** in which two instances (resources) are **identical**
  - Use of **semantic constraints** from domain experts
- All the possible contexts are organized in a **lattice** using an **order relation**

CONTEXTUAL IDENTITY: CONTRIBUTIONS

[Raad et al. 2017]

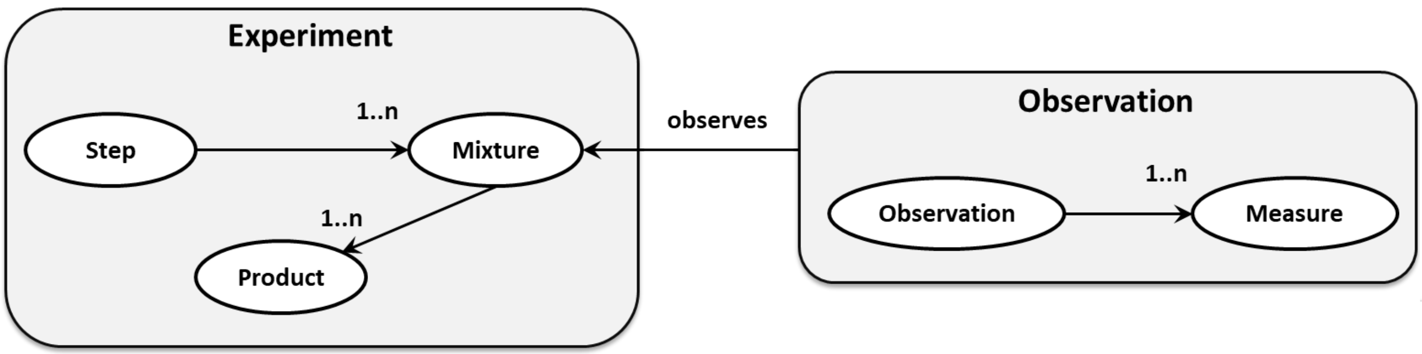




$$
\begin{aligned}
&GC_1 = (C = \{Drug, Paracetamol, Lactose, Weight\},\\
&OP = \{isComposedOf, hasWeight\}, DP = \emptyset,\\
&A = \{domain(isComposedOf) = Drug,\\
&range(isComposedOf) = Lactose \sqcup Paracetamol,\\
&domain(hasWeight) = Lactose \sqcup Paracetamol,\\
&range(hasWeight) = Weight\})
\end{aligned}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

CONTEXTUAL IDENTITY DETECTION APPROACH EVALUATION

[Raad et al. 2017]

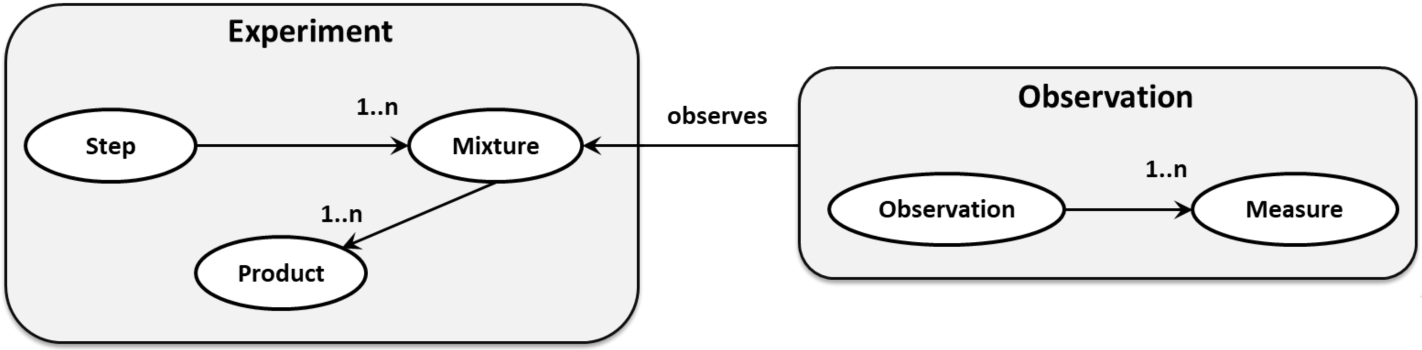


PO²
A Process and Observation Ontology
Carredas

- **Classes:** ≈ 5 000
- **Instances:** ≈ 42 000
- **Statements:** ≈ 237 000

- **Prediction rules**: generated for each context $\mathbf{C_i}$, and each observation result $\mathbf{m_i}$:
  $\boldsymbol{identiConTo_{<Ci>}(x, y) \wedge observes(x, m_1) \rightarrow observes(y, m_2)}$, with $m_1 \simeq m_2$

CONTEXTUAL IDENTITY DETECTION APPROACH EVALUATION [Raad et al. 2017]



A Process and Observation Ontology
Carredas

- Classes: ≈ 5 000
- Instances: ≈ 42 000
- Statements: ≈ 237 000

- **Prediction rules**: generated for each context $\mathbf{C_i}$, and each observation result $\mathbf{m_i}$:
  $\boldsymbol{identiConTo_{<Ci>}(x, y) \wedge observes(x, m_1) \rightarrow observes(y, m_2)}$, with $m_1 \simeq m_2$

Rule	*Error Rate*	*Support*
$identiConTo_{<GC_1>}(x, y) \rightarrow same(pH)$	6.19 %	57
$identiConTo_{<GC_3>}(x, y) \rightarrow same(Hardness)$	1.86 %	66
$identiConTo_{<GC_2>}(x, y) \rightarrow same(Friability)$	4.52 %	647

38 844 rules on Carredas dataset

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

IDENTITY MANAGEMENT: LESSONS LEARNED

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

Contextual identity

- An approach that detects contextual identity links in RDF KG while considering semantic constraints from domain experts
- Contexts used for value prediction in scientific KGs

IDENTITY MANAGEMENT: LESSONS LEARNED

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

Contextual identity

- An approach that detects contextual identity links in RDF KG while considering semantic constraints from domain experts
- Contexts used for value prediction in scientific KGs

Possible improvements

  - Need for hybrid approaches for link invalidation
  - Need for approaches for **difference links** detection: useful for inconsistency checking

OUTLINE

- **Introduction**
- **Contributions**
  - Part 1: Identity Management
  - **Part 2: Key Discovery**
  - Part 3: Data Enrichment
- **Summary and Future Directions**

PART 2: KEY DISCOVERY

PhD of Danai Symeonidou (2011-2014)

Co-supervised with N. Pernelle

Qualinca ANR Project (2012-2016)

Collaborations with: **LIG, LIRMM, Telecom ParisTech, INRA** and **Aalborg University (Danemark).**

KEY DISCOVERY FOR KNOWLEDGE GRAPH REFINEMENT

Rule-based data linking approaches [Saïs et al 2007, 2009]: need for knowledge to be declared in an ontology language or other languages.

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

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

	…	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**

KEY DISCOVERY FOR KNOWLEDGE GRAPH REFINEMENT

Rule-based data linking approaches [Saïs et al 2007, 2009]: 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
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**

KEY DISCOVERY FOR KNOWLEDGE GRAPH REFINEMENT

Rule-based data linking approaches [Saïs et al 2007, 2009]: 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 DISCOVERY FOR KNOWLEDGE GRAPH REFINEMENT: 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 KNOWLEDGE GRAPH REFINEMENT

Related Work

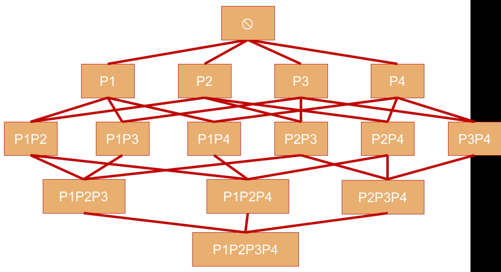
- In **2011**, **no key discovery** approach for **RDF** data
- **Approaches in relational databases are not applicable**
  - Closed world assumption
  - Do not consider multi-valued properties
  - No ontologies (semantics cannot be used)

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).

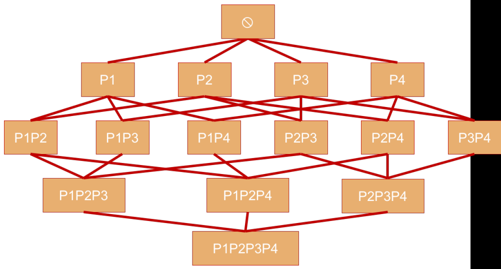
KEY DISCOVERY: A COMPLEX PROBLEM

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



KEY DISCOVERY: A COMPLEX PROBLEM

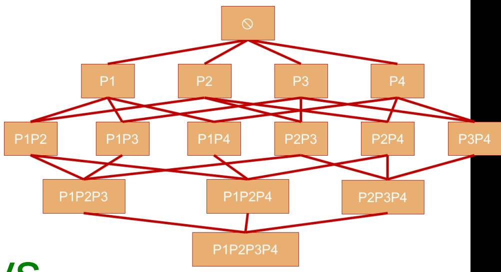
- Find all the minimal keys requires at least 2^n property combinations
  - need of efficient filtering and prunings



- 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

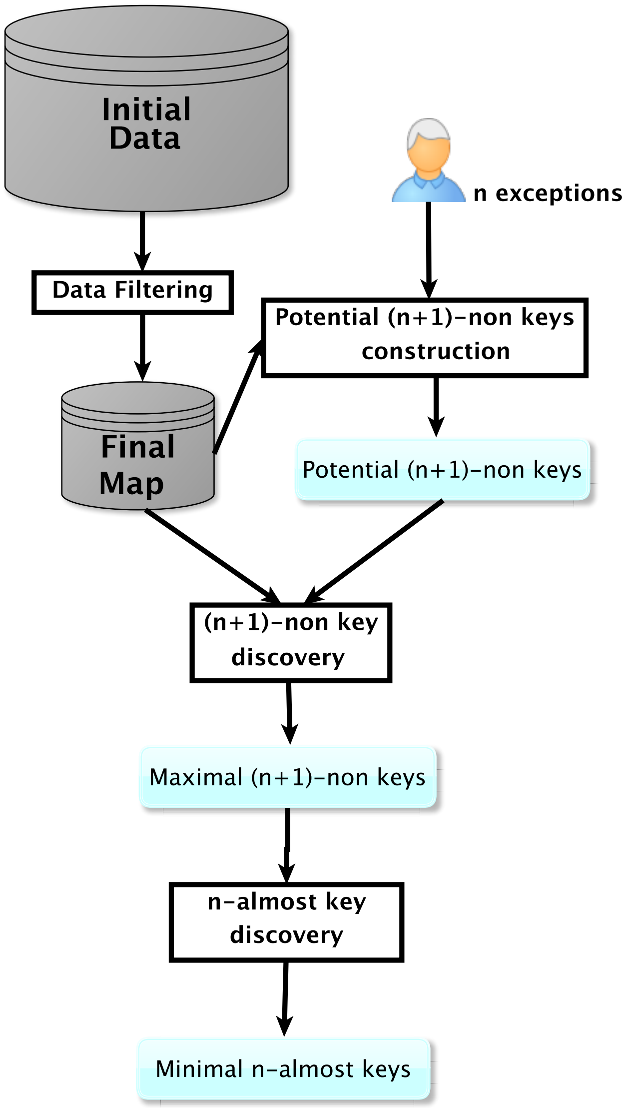
SAKEY

VICKEY

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

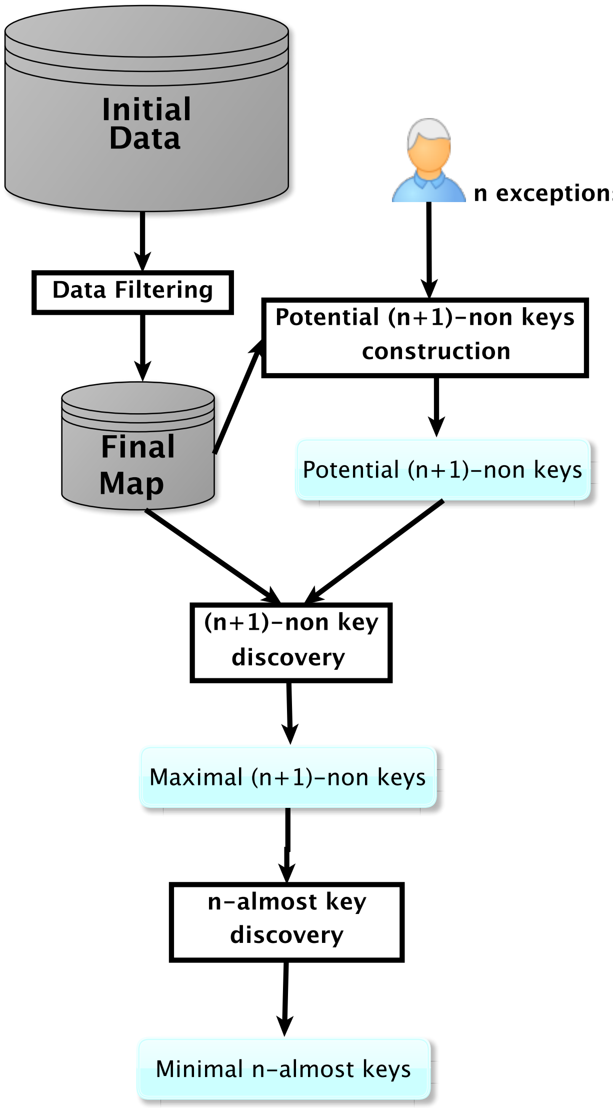
SAKEY: N-ALMOST KEY DISCOVERY

- **SAKey allows *n* exceptions in the data**
- **Exception set E_P:** set of instances that share values for the set of properties P



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 $|E_P| \geq n+1$**

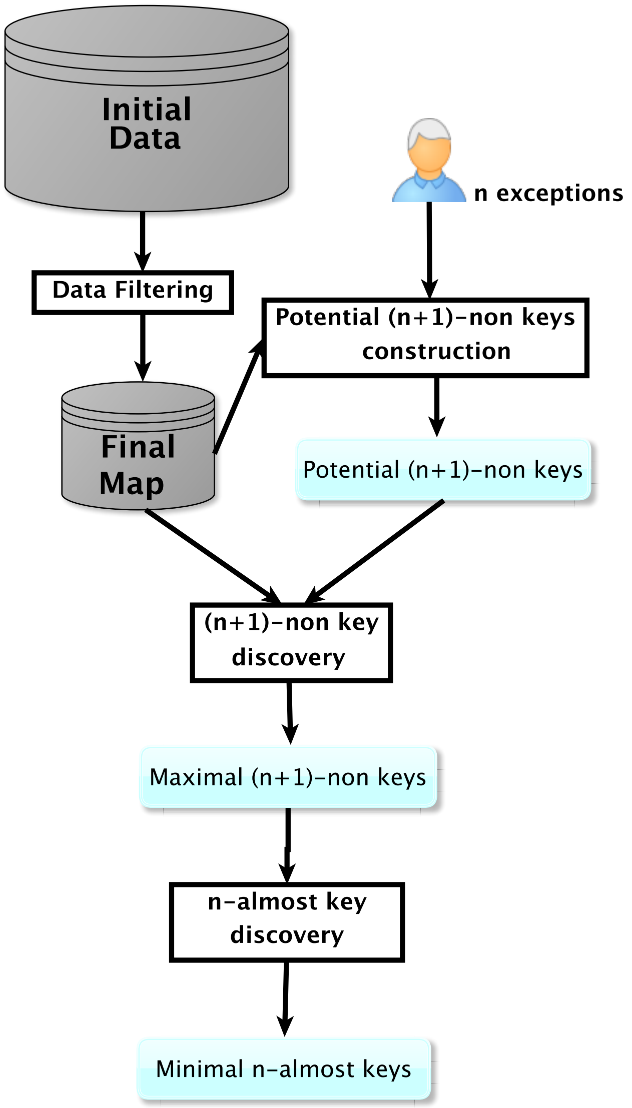


SAKEY: N-ALMOST KEY DISCOVERY

- 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



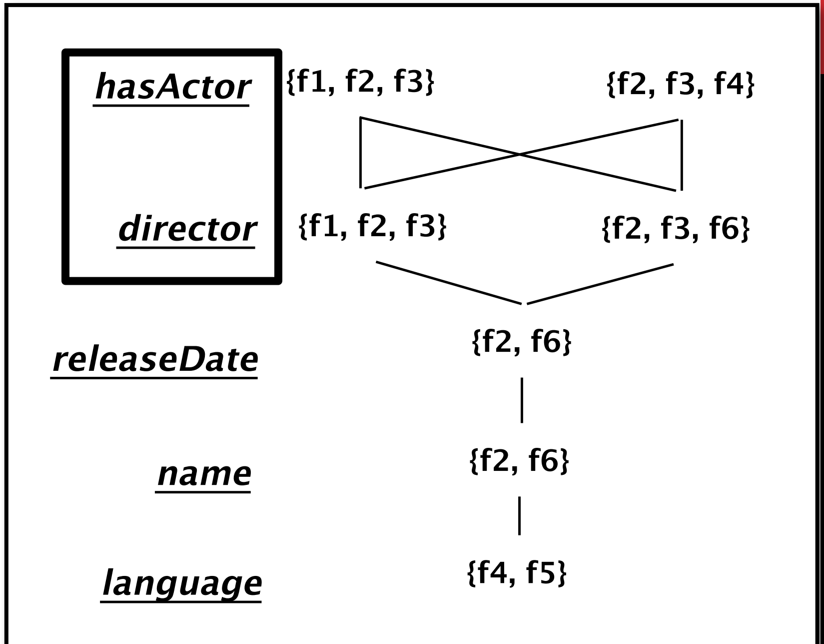
SAKEY: N-ALMOST KEY DISCOVERY

(n+1)- maximal non-key discovery:

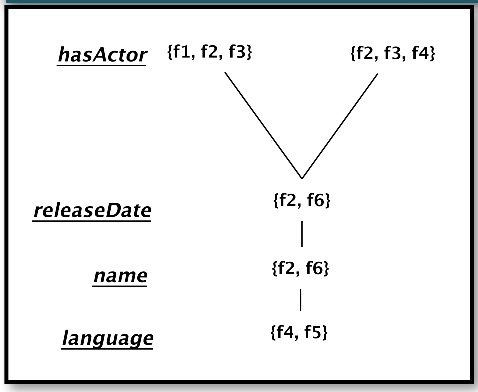
Intersections between sets of properties

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



…

Different prunings and filtering

Efficient n-almost key derivation

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	x 1.75
	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	x 869
	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	x 7.1
	SAKey Keys + Conditional keys	0.39	0.98	**0.55**	

KEY DISCOVERY: LESSONS LEARNED

- **Three different methods** (KD2R, SAKey, VICKEY) 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)

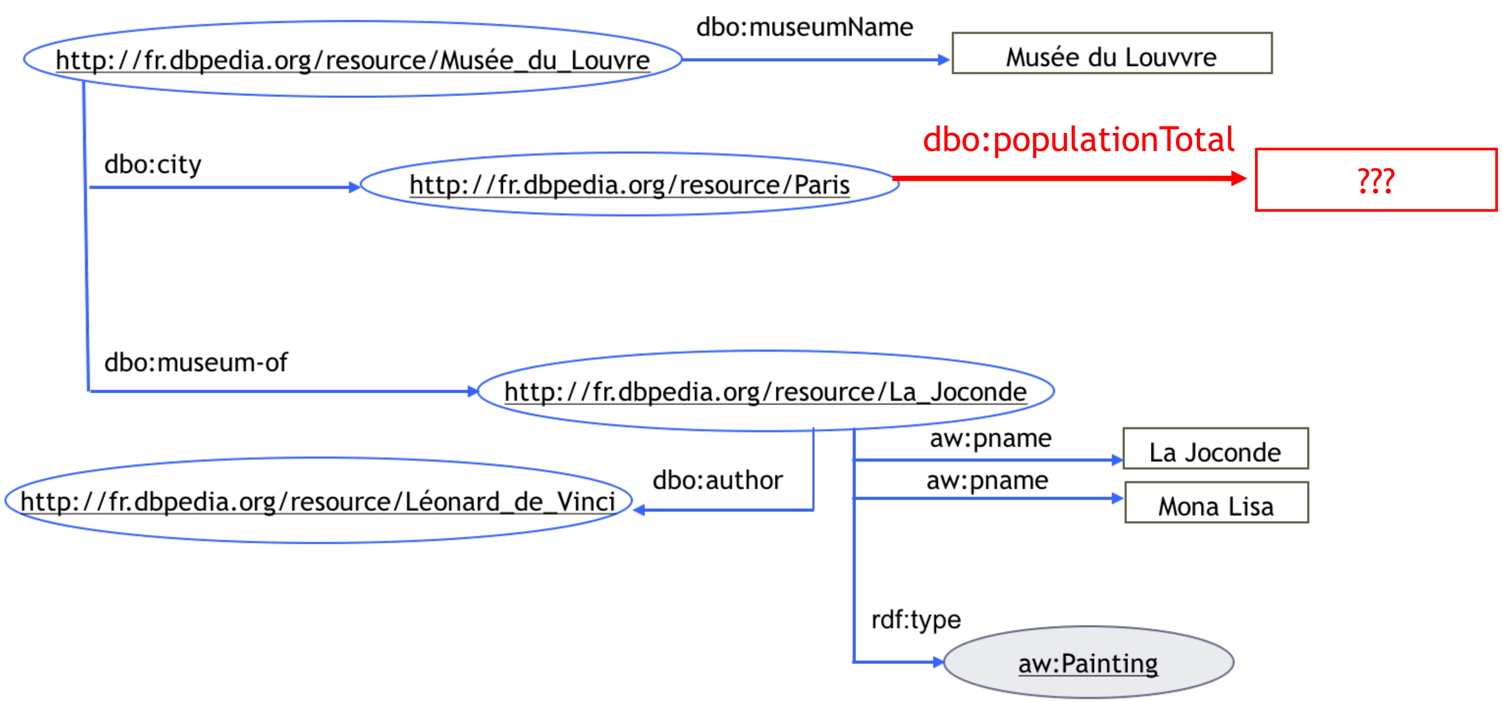
Possible improvements

- **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**
- **Contributions**
  - Part 1: Identity Management
  - Part 2: Key Discovery
  - **Part 3: Data Enrichment**
- **Conclusion and Future Directions**

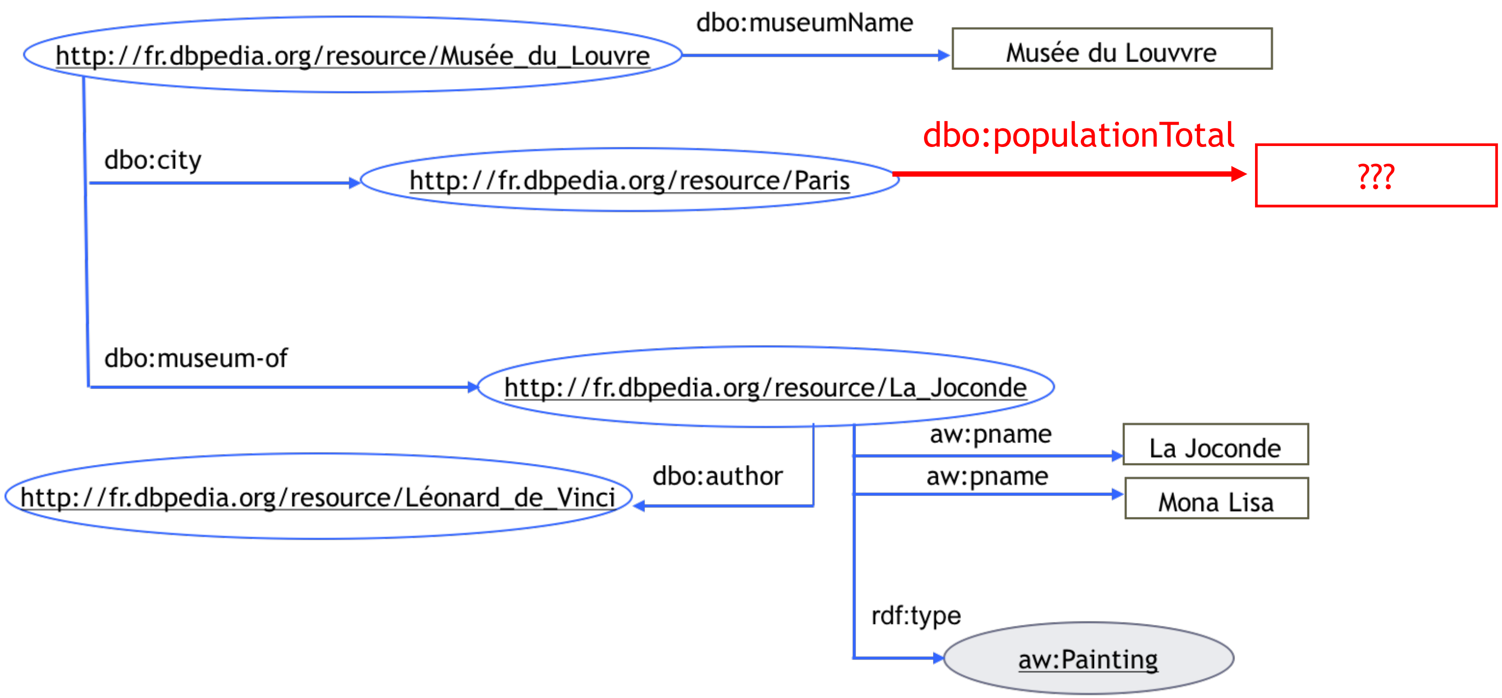
DATA ENRICHMENT



Contributions [Collaboration with R. Thomopoulos and S. Destercke]

- **Fusion of different RDF data sources** [ODBASE'08, LFA'09, ODBASE'10, EGC'15]
- **Prediction of missing values** [KBS'14, Chapter in Nova Science'15]

DATA ENRICHMENT



Contributions [Collaboration with R. Thomopoulos and S. Destercke]

- **Fusion of different RDF data sources** [ODBASE'08, LFA'09, ODBASE'10, EGC'15]
- **Prediction of missing values** [KBS'14, Chapter in Nova Science'15]

DATA ENRICHMENT: DATA FUSION

- **Merge information from entities linked by *identity links* *to* obtain a single homogenized representation**
- **Why fusion?**
  - Improve knowledge graphs completeness
  - Group together best quality information

DATA ENRICHMENT: DATA FUSION

- **Merge information from entities linked by *identity links* *to* obtain a single homogenized representation**
- **Why fusion?**
  - Improve knowledge graphs completeness
  - Group together best quality information

	title	nbPages	auteur	datePub	keywords
b1	A semantic Web Primer	238	G. Antoniou	01/05/2008	Semantic Web Artificial Intelligence
b2	A semantic Web Primer, second edition …	0	Grigoris Antoniou	January, 1st 2008	Semantic Web AI Knowledge Representation

sameAs

DATA ENRICHMENT: DATA FUSION

- **Merge information from entities linked by *identity links* *to* obtain a single homogenized representation**
- **Why fusion?**
  - Improve knowledge graphs completeness
  - Group together best quality information

	title	nbPages	auteur	datePub	keywords
b1	A semantic Web Primer	238	G. Antoniou	01/05/2008	Semantic Web Artificial Intelligence
sameAs	**conflit**	**conflit**	**conflit**	**conflit**	**conflit**
b2	A semantic Web Primer, second edition …	0	Grigoris Antoniou	January, 1st 2008	Semantic Web AI Knowledge Representation

DATA ENRICHMENT: DATA FUSION

- **Merge information from entities linked by *identity links to* obtain a single homogenized representation**
- **Why fusion?**
  - Improve knowledge graphs completeness
  - Group together best quality information

	title	nbPages	auteur	datePub	keywords
b1	A semantic Web Primer	238	G. Antoniou	01/05/2008	Semantic Web Artificial Intelligence
b2	A semantic Web Primer, second edition …	0	Grigoris Antoniou	January, 1st 2008	Semantic Web AI Knowledge Representation

sameAs

conflit conflit conflit conflit conflit

Data freshness and errors

Different vocabularies

DATA ENRICHMENT: DATA FUSION RELATED WORK

In **2008**, there was **no approach** that deals with **RDF data fusion**

In **relational databases** [survey in Bleiholder & Naumann, 2008]

Data quality independent strategies

- Keep the most frequent value (democratic vote)
- Aggregation functions: average, max, min, concatenation, intervals

Data quality driven strategies

- Keep the value getting the best confidence value (or / threshold)
- Trust a reliable source
- Apply a vote weighted by the source reliability

Not applicable to RDF data: OWA, multi-valued properties and no ontologies

DATA ENRICHMENT: DATA FUSION

[Saïs et al. 2008, 2010, 2015]

Multi-criteria and conservative data fusion approach

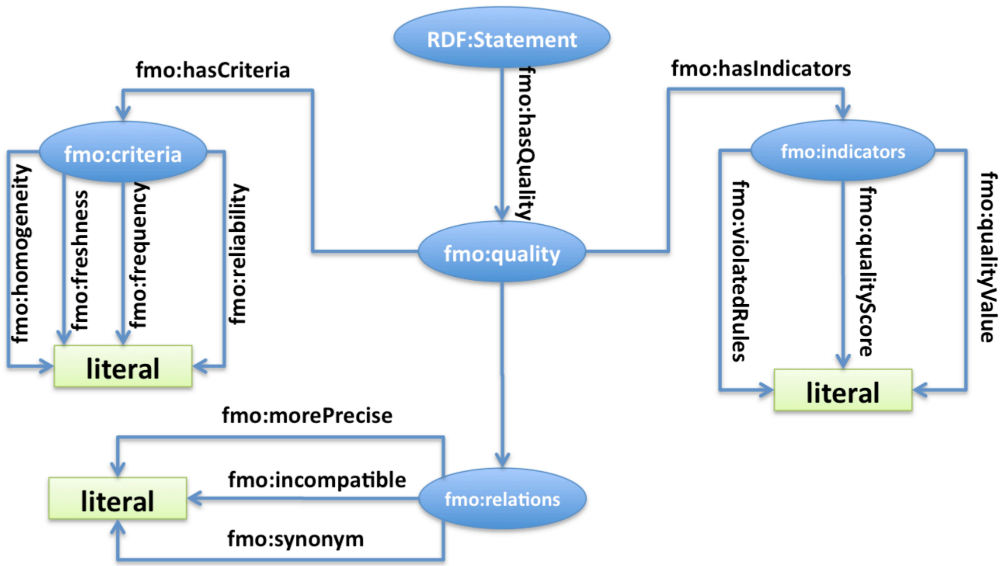
- **Detects implausible values** using expert constraints (age >0)
- **Computes quality score for plausible values**: frequency, homogeneity, source freshness and reliability

- **Discovers semantic relations (can affect the quality score) :**
  - More Precise: (Paris, France)
  - Synonyms: AI, Artificial Intelligence
  - Incompatible: R: reviewingDate < publicationDate

DATA ENRICHMENT: DATA FUSION

[Saïs et al. 2008, 2010, 2015]

- **Data Fusion Metadata Ontology**




```
Book-F1 dfa:name v1
v1 rdf:type Value
q1 rdf:type Quality
c1 rdf:type Criteria
c2 rdf:type Criteria
…

v1 dfa:hasValue 'Grigoris  Antoniou''
v1 dfa:isImplausible false
v1 fmo:hasQuality q1
q2 fmo:hasCriteria c1
c1 fmo:homogeneity 0.6
c2 fmo:freshness 0.99
…
```

Explanation of data fusion decisions [Saïs et al. 2018]

Explanations

DATA ENRICHMENT: DATA FUSION LESSONS LEARNED

Multi-criteria data fusion approach

- Uses ontology semantics for confidence degree and provenance
- Keeps all the values ranked to allow flexible querying (top-k)
- Uncertainty modelling: fuzzy sets and possibility theory

Possible improvements

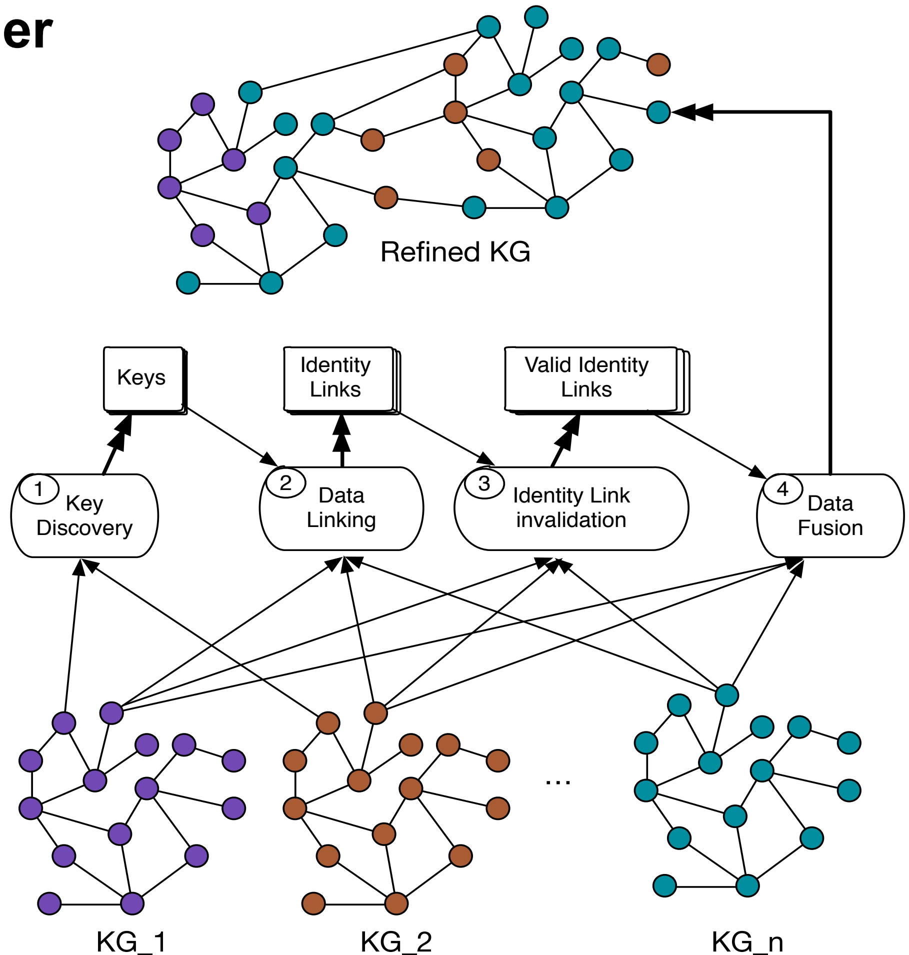
- **Object properties:** sameAs links, differentFrom links, information gain.
- **Multi-valued properties**: use information completeness
- **Evaluation** on big datasets: use of crowd-sourcing
- **Explanation** models for result interpretation and human validation

OUTLINE

- **Introduction**
- **Contributions**
  - Part 1: Identity Management
  - Part 2: Key Discovery
  - Part 3: Data Enrichment
- **Conclusion and Future Directions**

CONCLUSION

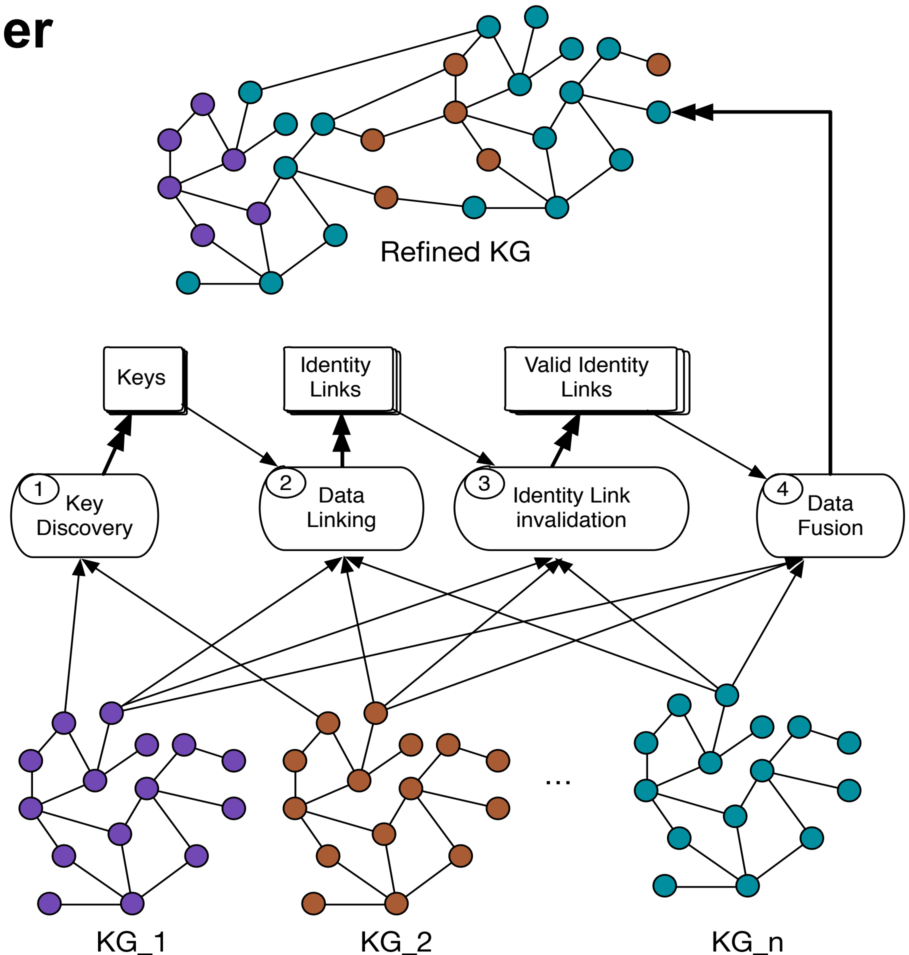
- **Knowledge graph refinement under**
  - ☑ Open World Assumption
  - ☑ Imperfect KGs
  - ☑ Complex KGs
  - ☑ Massive KGs



CONCLUSION

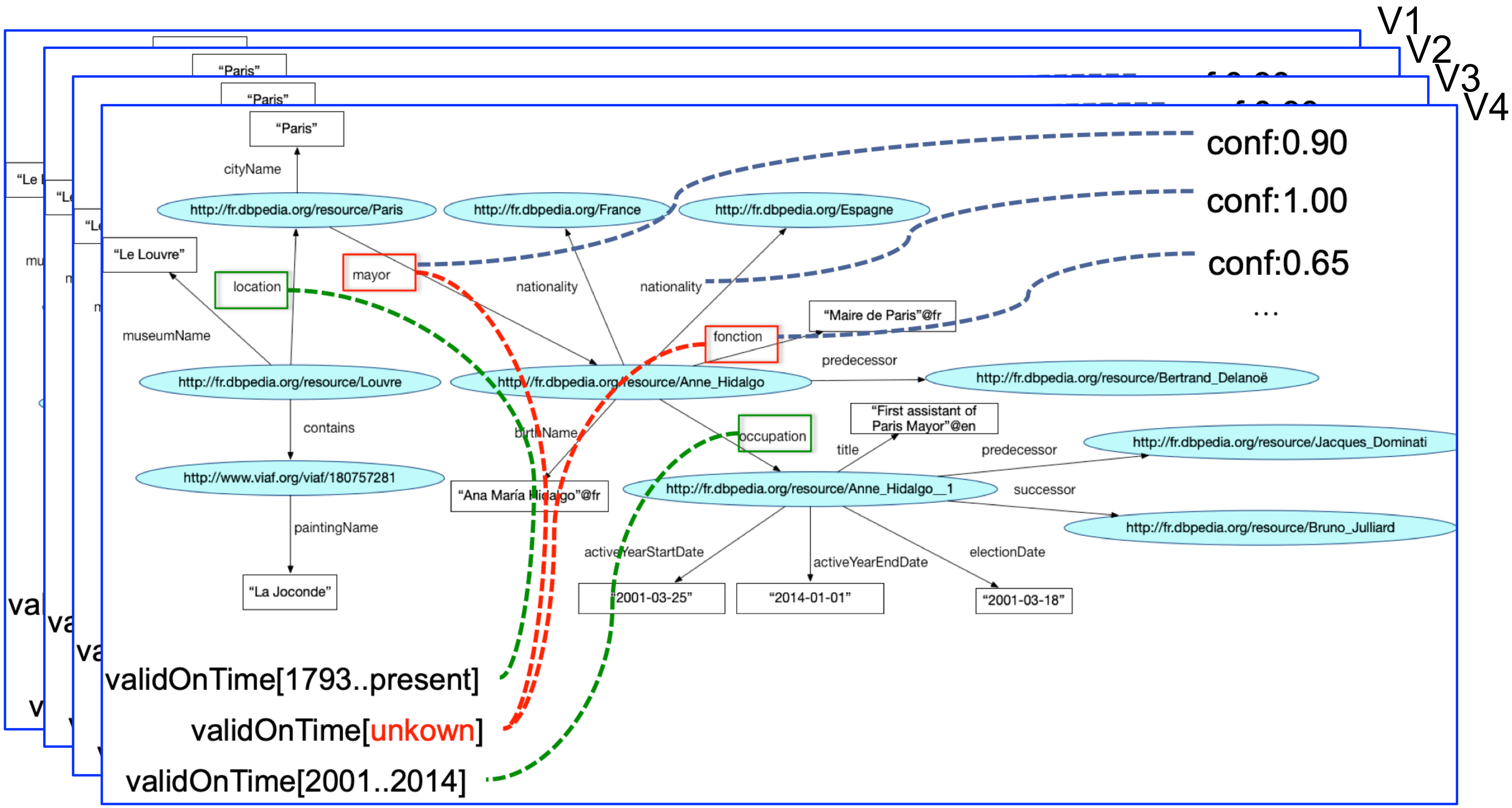
- **Knowledge graph refinement under**
  - ☑ Open World Assumption
  - ☑ Imperfect KGs
  - ☑ Complex KGs
  - ☑ Massive KGs

- **Efforts needed for**
  - ☐ Evolution
  - ☐ Uncertainty
  - ☐ Temporality



FUTURE DIRECTIONS

TEMPORAL, UNCERTAIN AND EVOLVING KG REFINEMENT: FUTURE DIRECTIONS



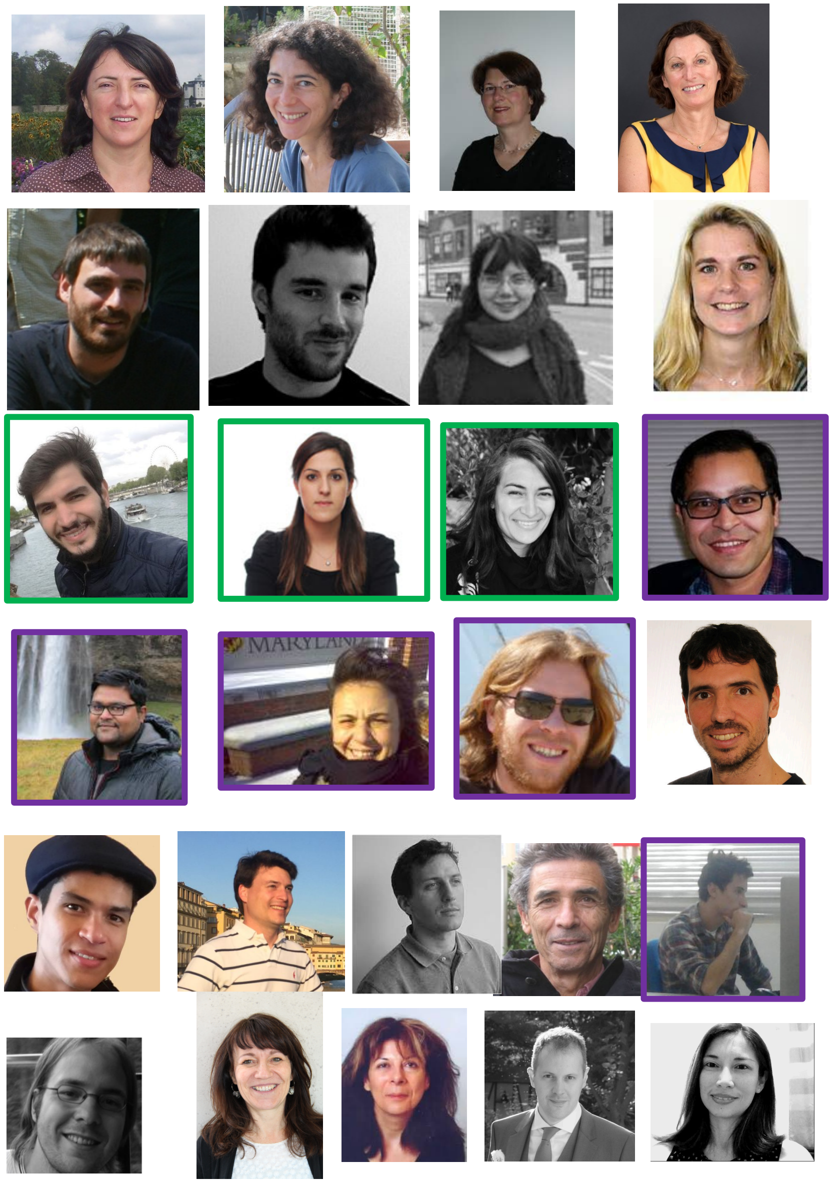
TEMPORAL, UNCERTAIN AND EVOLVING KG REFINEMENT: FUTURE DIRECTIONS

- **Data Evolution (Collaboration with N. Pernelle, C. Pruski (LIST))**
  - Semantic representation of changes on entities, topology, frequency of changes, the longevity of values, …
  - Incremental KG refinement: identity management, data fusion and knowledge discovery

- **Time (Collaboration G. Quercini)**
  - KG enrichment with temporal meta-facts [Malaverri post-doc 2019]
  - Time-aware veracity assessment [Sergey Konovalov Internship]
  - Temporal data linking

- **Knowledge Discovery (more expressive rules)**
  - Discovery of **referring expressions** [A. Khajeh Nassiri Internship]
  - Discovery of **causality rules** in scientific KGs [A. Filali Rotbi Internship]
    - PhD funding WarmRules project (2019-2021) from DATAIA
    - Combination of symbolic and statistical approaches

KNOWLEDGE GRAPH REFINEMENT



Identity Management [ANR Qualinca, LIONES]

- **Data Linking**: contextual identity link detection
- **Identity Link Invalidation**

  J. Raad PhD (N. Pernelle, J. Dibie, L. Ibanescu)

  L. Papaleo post-doc

  Publications: EKAW'15, K-Cap'17, ISWC'18, …

Key Discovery [ANR Qualinca]

- **Key axiom enrichment**

  D. Symeonidou PhD (N. Pernelle)

  Collaboration with: LIRMM, LIG, Telecom ParisTech

  Publications: SWW'11, JWS'13, ISWC'14, ICCS'14, ISWC'17, …

Data Enrichment [ANR Qualinca]

- **Data Fusion:** Property value enrichment
- **Missing value prediction:** Property value enrichment

  Collaboration with R. Thomopoulos, S. Destercke

  Publications: ODBASE'08, LFA'10, ODBASE'10, KBS'14, Nova Science Chapt.'15, WETICE'18, …