DATA LINKING AND KNOWLEDGE DISCOVERY IN RDF DATA: METHODS AND SOME FEEDBACK FROM AGRONOMIC APPLICATIONS

FATIHA SAÏS

LAHDAK@LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

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

1ST DATAIA DAYS « LIFE SCIENCES & AI», DEC. 4TH 2019

LINKED OPEN DATA

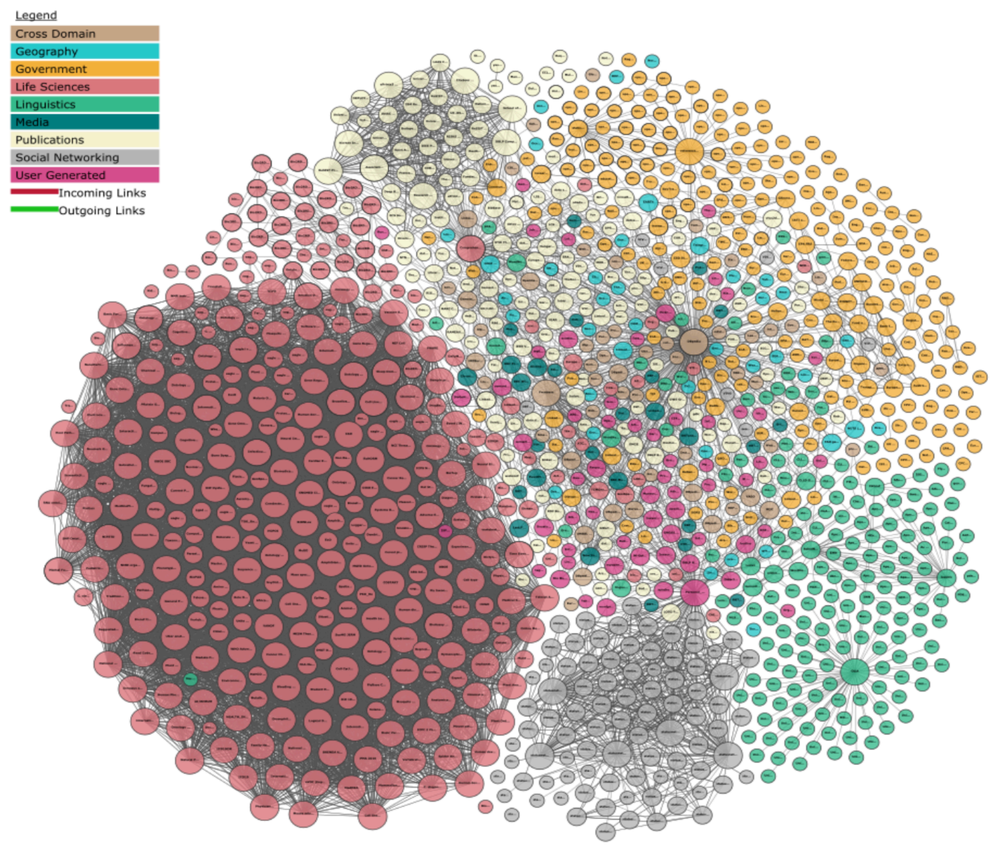
Linked Open Data (LOD)

FAIR Principles

Linked Data - Datasets under an open access

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

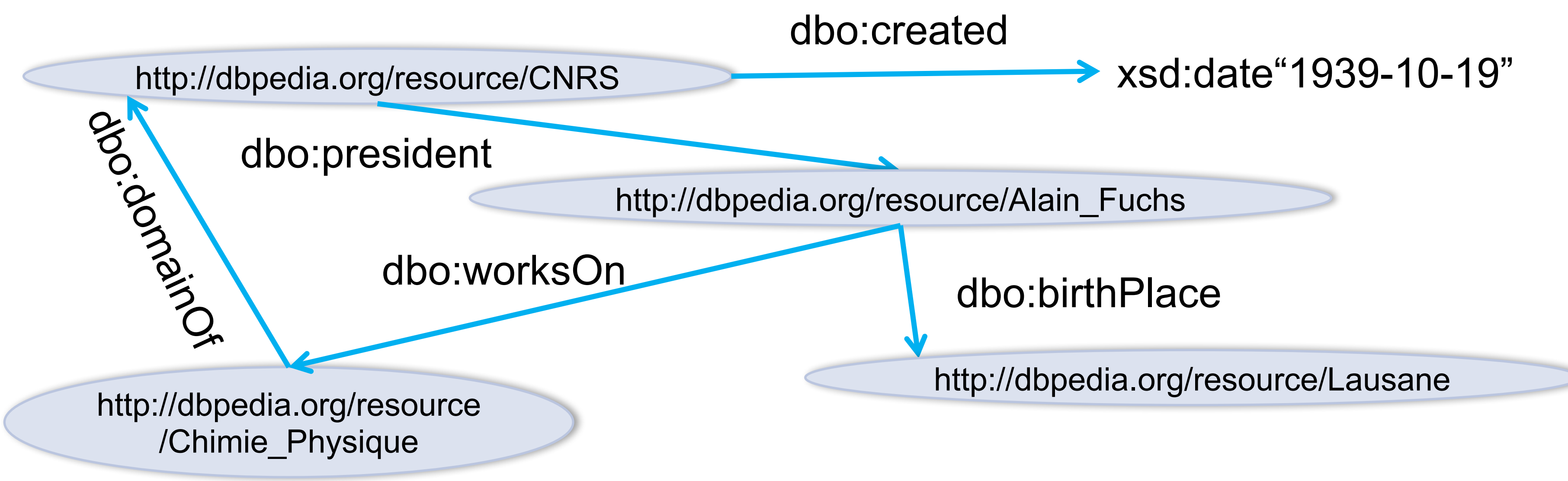
Gene Ontology: 807473 triples
Lipid Ontology: 15406 triples



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

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.

SEMANTIC WEB: ONTOLOGIES

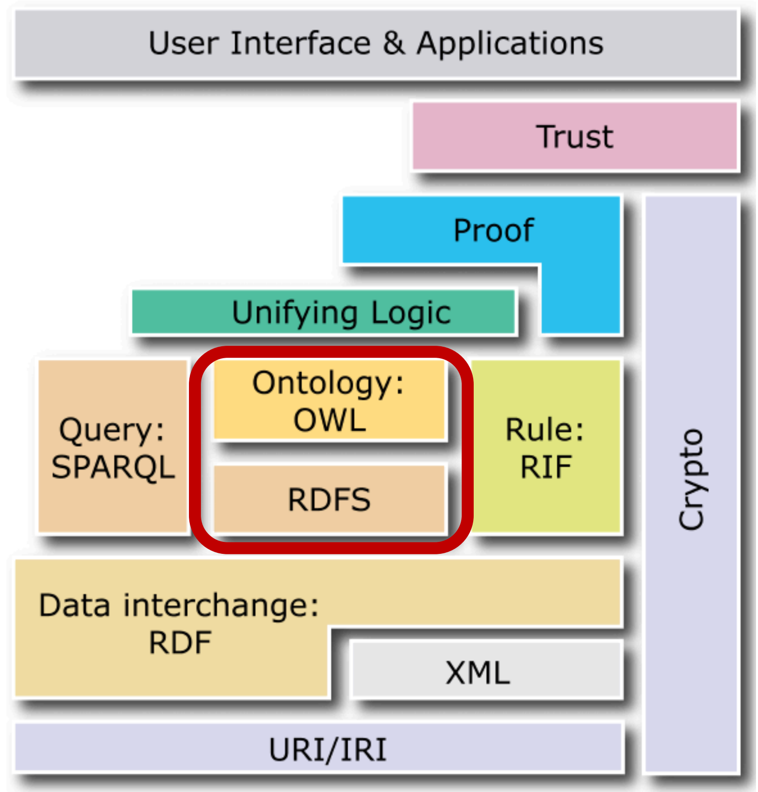
"An ontology is an **explicit**, **formal** specification of a shared conceptualization."
[Thomas R. Gruber, 1993]

RDFS – Resource Description Framework Schema

- Lightweight ontologies

OWL – Web Ontology Language

- Expressive ontologies

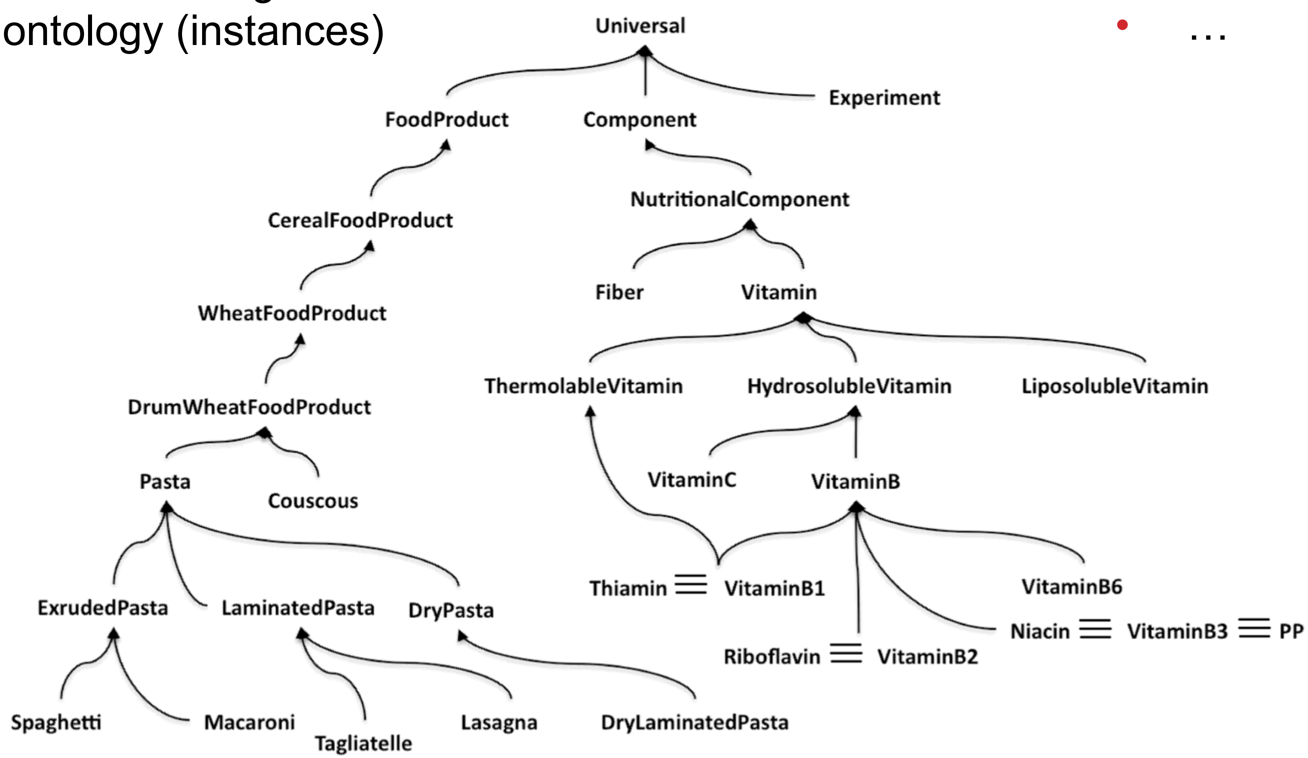


Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png

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
  - …



Disjunction Constraints

FoodProduct $\perp$ Component

Spaghetti $\perp$ Macaroni

Fiber $\perp$ Vitamin

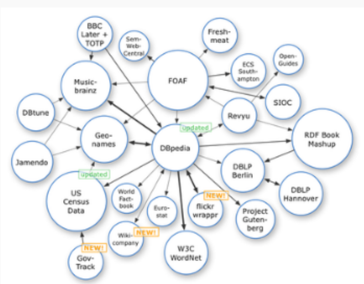
Vitamin C $\perp$ Vitamin B

KNOWLEDGE GRAPHS

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

DBpedia



2007

yago select knowledge

2012

WIKIDATA

2007

Freebase™

Academic side

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

2007

yago select knowledge

2012

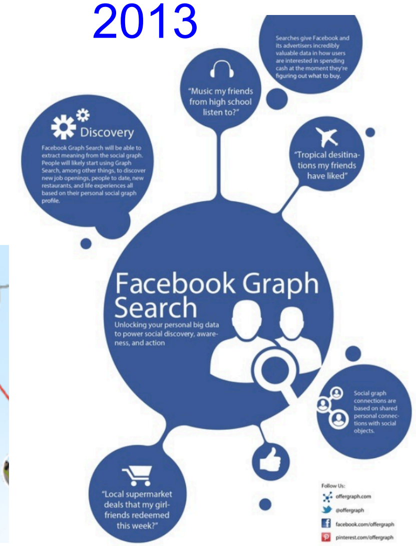
WIKIDATA

2007

Freebase™

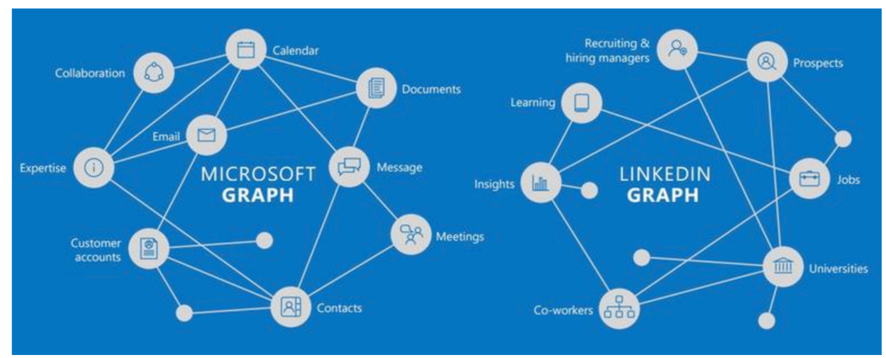
2012

2013



2015

2016



2013

Yahoo!™

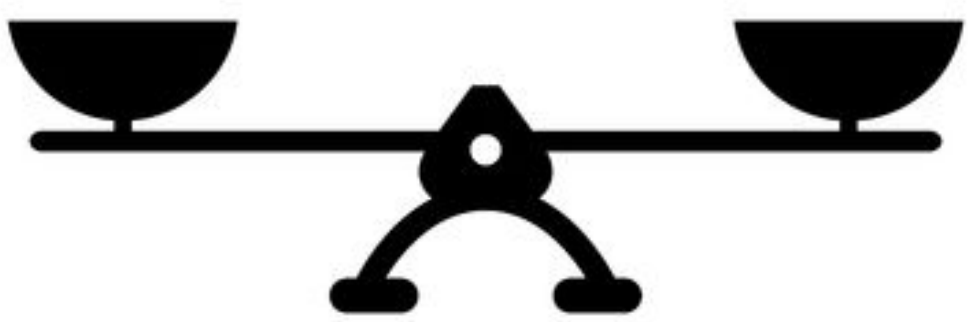
Yahoo's new SERP designs mobile and knowledge graph

Academic side

Commercial side

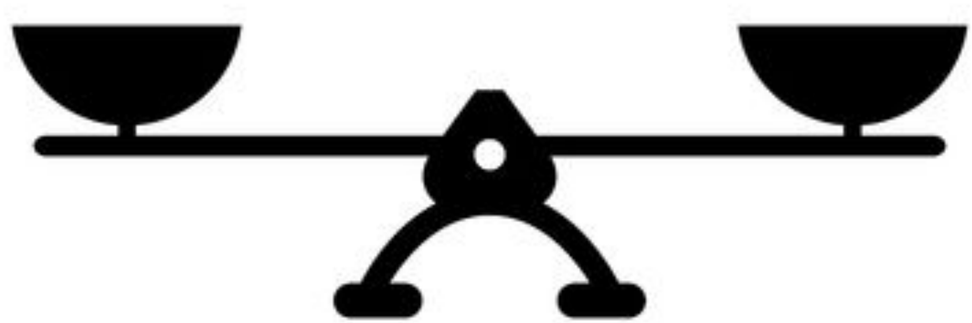
KNOWLEDGE GRAPH REFINEMENT

Completeness Correctness



KNOWLEDGE GRAPH REFINEMENT

Completeness Correctness



Data Linking
Key discovery
Data Fusion

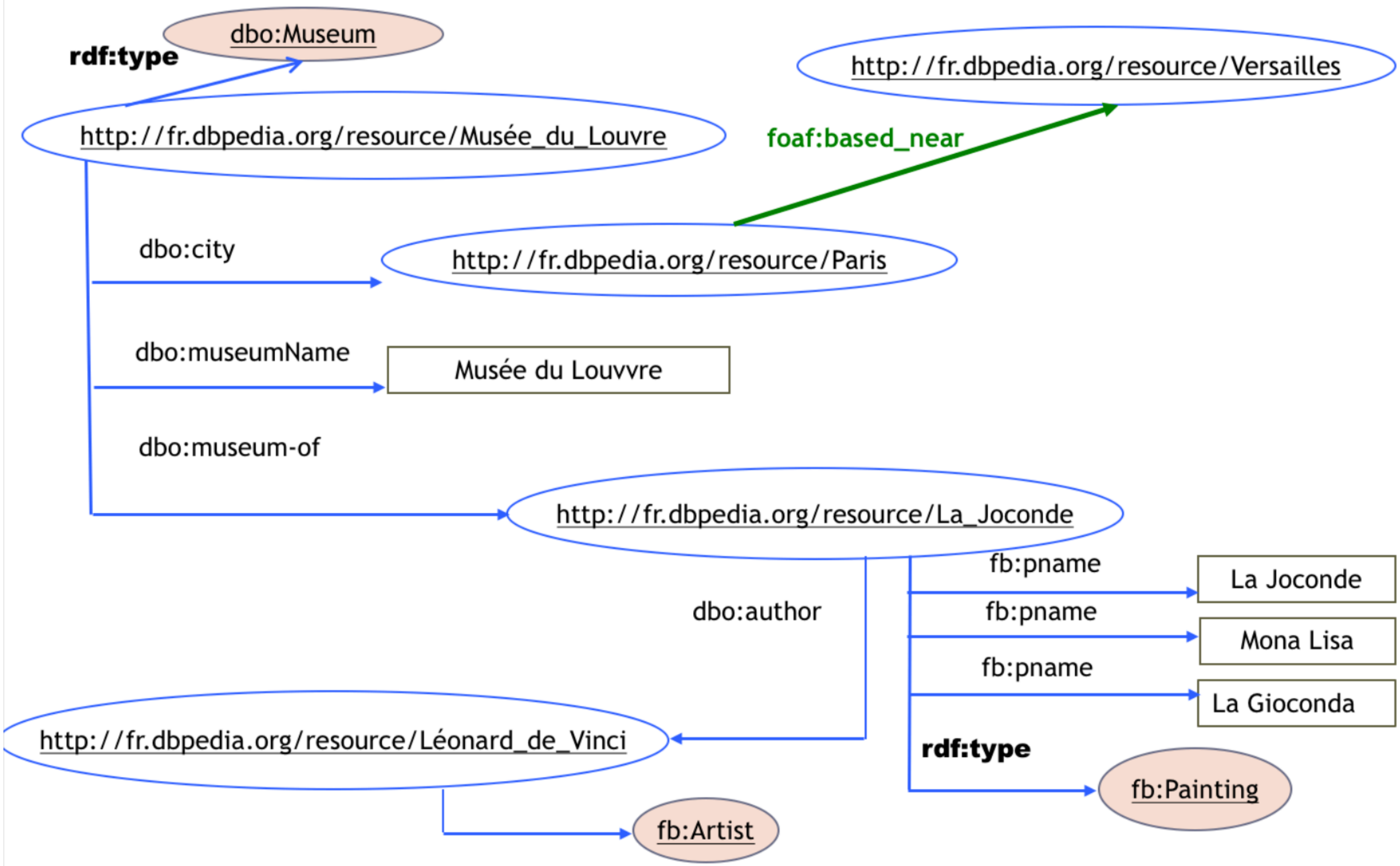
Link Invalidation
Contextual identity

OUTLINE

- Introduction
- Key discovery for data linking
- Link Invalidation
- Contextual identity
- Conclusion

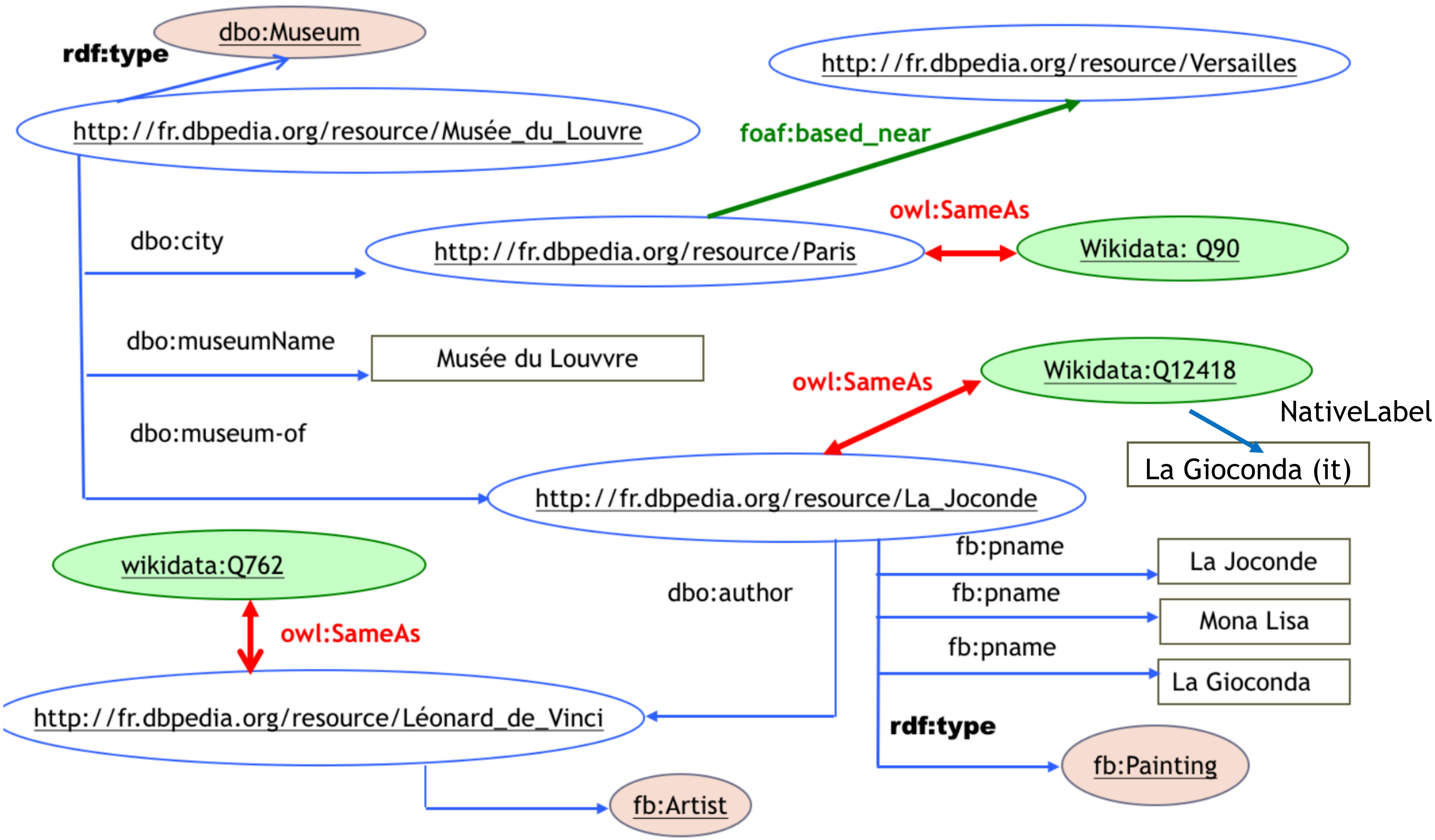
DATA LINKING

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, protein).



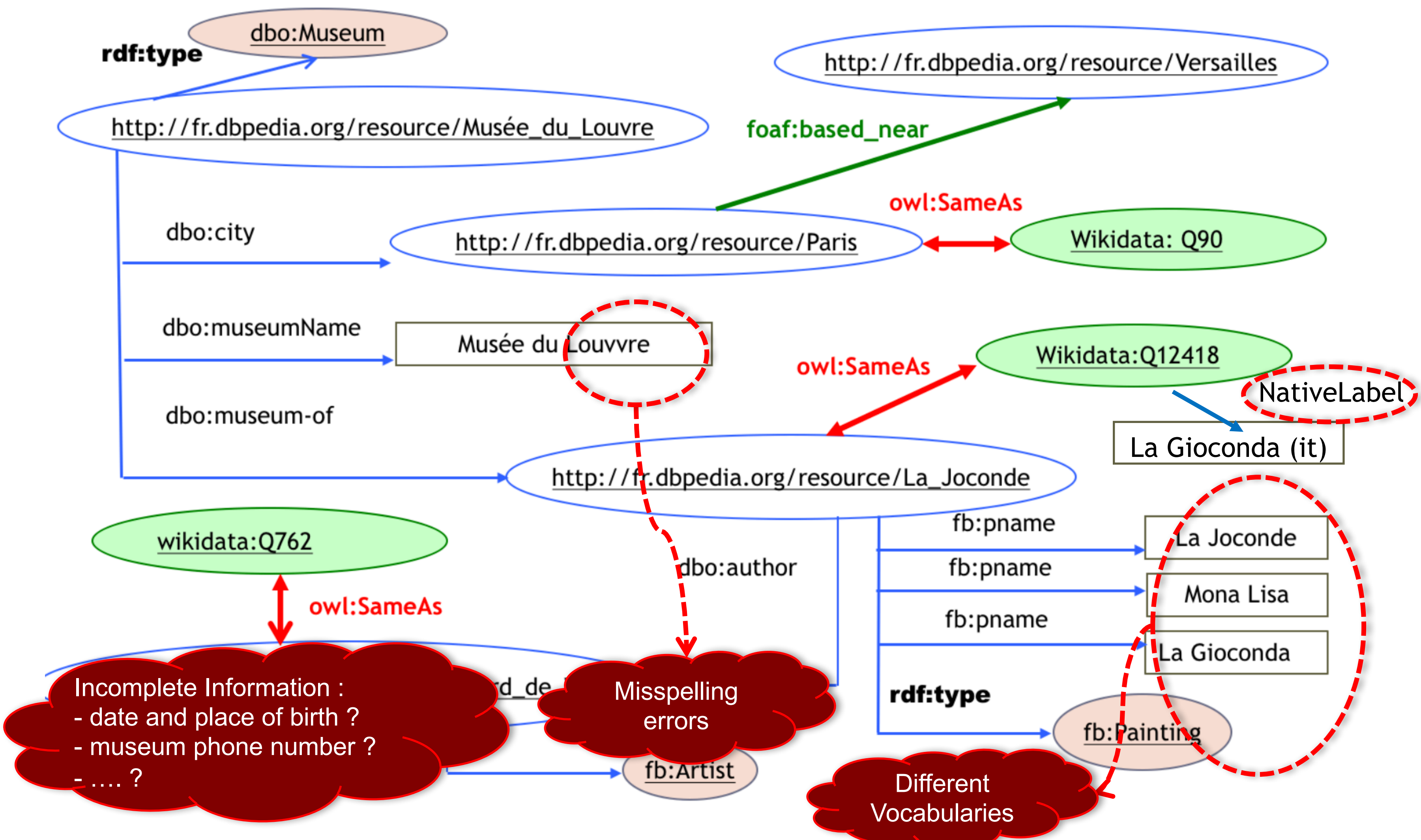
DATA LINKING

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, protein).



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, protein).



DATA LINKING APPROACHES

- **Local approaches**: consider properties to compare pairs of instances independently

versus

- **Global approaches**: consider data type properties (attributes) as well as object properties (relations) 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

- **Some surveys:**

1. Alfio Ferrara, Andriy Nikolov, François Scharffe:Data Linking. J. Web Semant. 23: 1 (2013)
2. Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, Erhard Rahm:A survey of current Link Discovery frameworks. Semantic Web 8(3): 419-436 (2017)

KNOWLEDGE-BASED 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.

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

Then we may infer:

sameAs(museum11, museum21)

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**

KNOWLEDGE-BASED 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**

KNOWLEDGE-BASED 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: KEYS WITH EXCEPTIONS

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

	FirstName	LastName	City	Profession
Person1	Anne	Tompson	Paris	Actor, Director
Person2	Marie	Tompson	Berlin	Actor
Person3	Marie	David	Toulouse	Actor
Person4	Vincent	Solgar	Rome	Actor, Director
Person4	Simon	Roche	Montpellier	Teacher
Person4	Jane	Ser	Paris	Teacher, Researcher
Person4	Sara	Khan	London	Teacher
Person4	Theo	Martin	Lyon	Teacher, Researcher
Person4	Marc	Blanc	Nantes	Teacher

Is [FirstName,LastName] a key? ✔

Is [City] 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	City	Profession
Person1	Anne	Tompson	Paris	Actor, Director
Person2	Marie	Tompson	Berlin	Actor
Person3	Marie	David	Toulouse	Actor
Person4	Vincent	Solgar	Rome	Actor, Director
Person4	Simon	Roche	Montpellier	Teacher
Person4	Jane	Ser	Paris	Teacher, Researcher
Person4	Sara	Khan	London	Teacher
Person4	Theo	Martin	Lyon	Teacher, Researcher
Person4	Marc	Blanc	Nantes	Teacher

Is [FirstName,LastName] a key? ✔

Is [City] a key? ✖

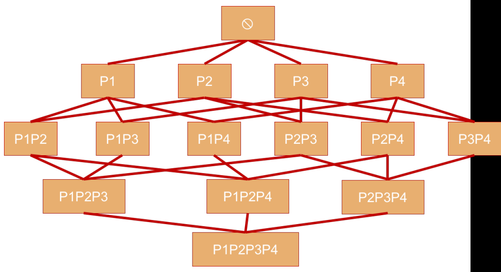
Is [City] a key with 2 exceptions? ✔

Exact keys

Almost keys

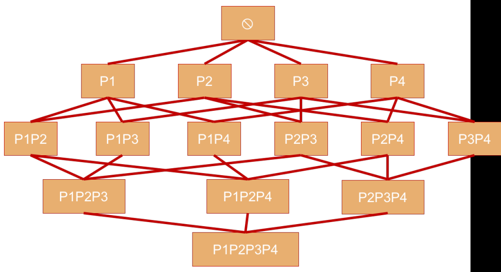
KEY DISCOVERY: A COMPLEX PROBLEM

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



KEY DISCOVERY: A COMPLEX PROBLEM

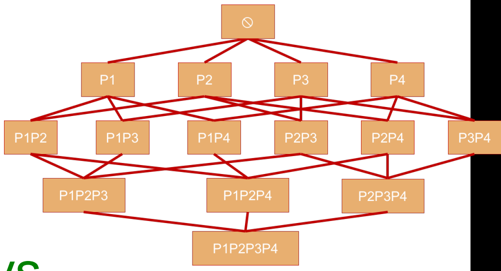
- Find all the minimal keys requires at least 2^n property combinations
  - ➢ need of efficient filtering and prunings



- 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	City	Profession
Person1	Anne	Tompson	Paris	Actor, Director
Person2	Marie	Tompson	Berlin	Actor
Person3	Marie	David	Toulouse	Actor
Person4	Vincent	Solgar	Rome	Actor, Director
Person4	Simon	Roche	Montpellier	Teacher
Person4	Jane	Ser	Paris	Teacher, Researcher
Person4	Sara	Khan	London	Teacher
Person4	Theo	Martin	Lyon	Teacher, Researcher
Person4	Marc	Blanc	Nantes	Teacher

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

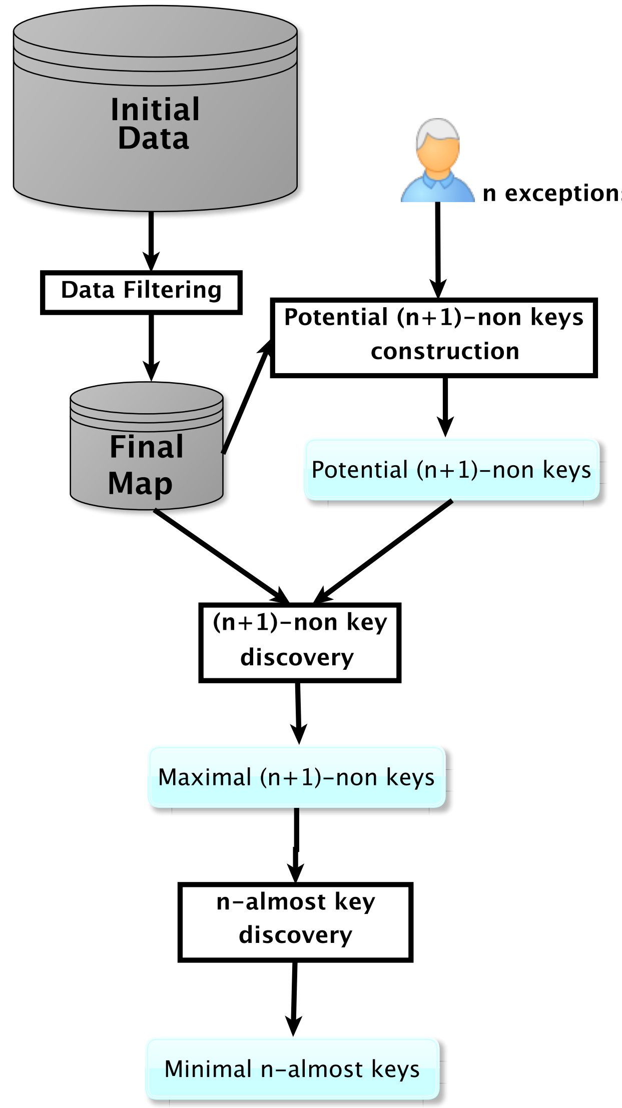
SAKEY: N-ALMOST KEY DISCOVERY

- SAKey allows ***n*** **exceptions** in the data

- **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



APPLICATION TO SCIENTIFIC DATA

- **Many scientific numerical data**
  - Sensor data
  - Experimental data..

- Difficult to interpret numerical data
  - Different levels of precision
  - Different measure units…

- Better understand the numerical data

Danai Symeonidou, Isabelle Sanchez, Madalina Croitoru, Pascal Neveu, Nathalie Pernelle, Fatiha Saïs, Aurelie Roland-Vialaret, Patrice Buche, Aunur-Rofiq Muljarto, Remi Schneider:Key Discovery for Numerical Data: Application to Oenological Practices. ICCS 2016: 222-236

APPLICATION TO SCIENTIFIC DATA

Discover keys in numerical data

- **Keys:** combinations of properties that discriminate a resource

Evaluate their quality

- Experimental numerical data in 3 wine flavour datasets (2011-2014)

How do we discriminate the wines??

PROBLEM STATEMENT

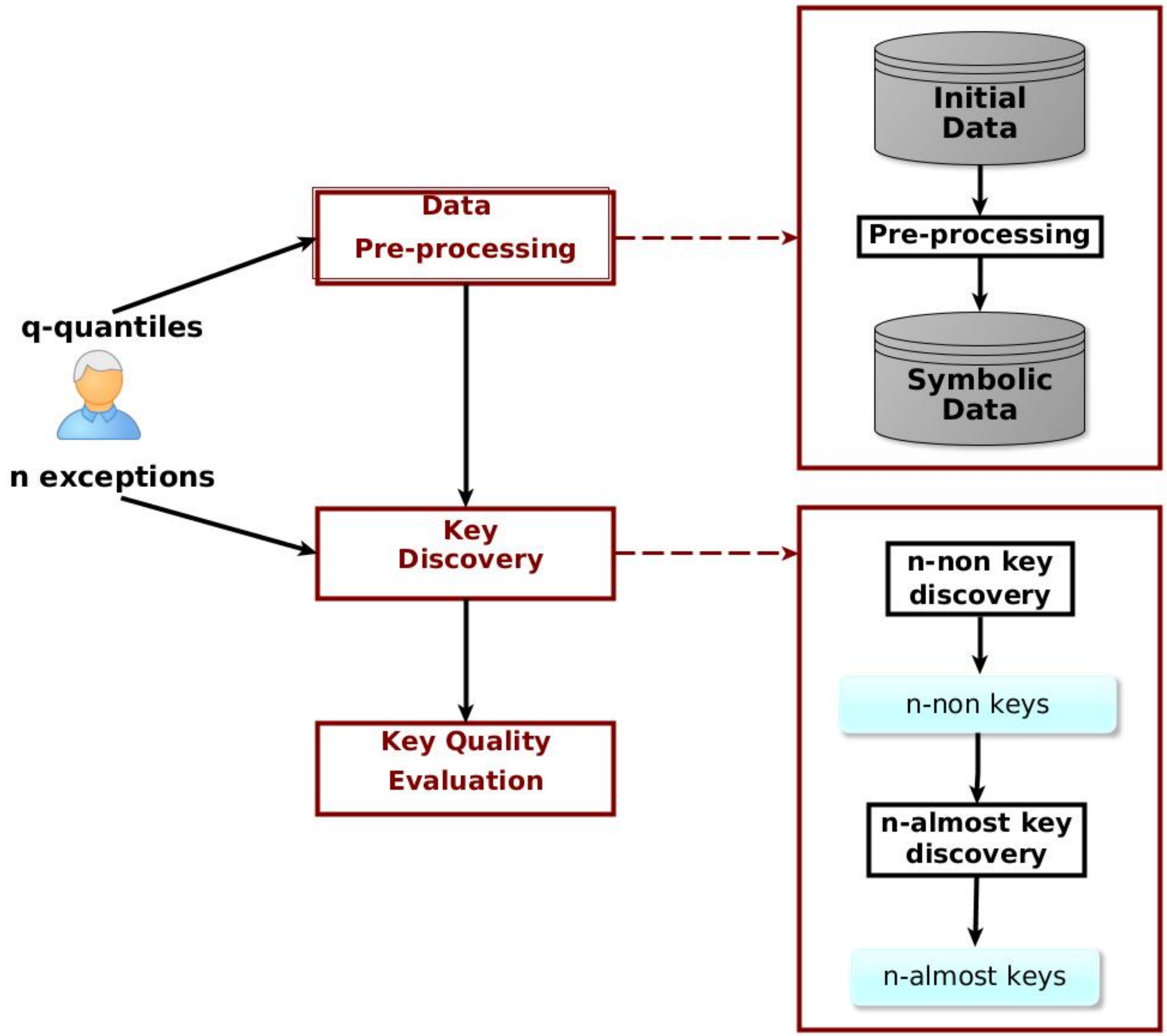
Key discovery approaches consider all the values as symbolic

- Ex. PH(Wine1, **3.455**),PH(Wine2, **3.457**)

$\neq$

Key discovery in raw numerical data: Many not-significant keys can be found

PROPOSED METHOD STEPS



DATA PRE-PROCESSING

Objective: Interpret numerical data in a symbolic way

Solution: Use *quantiles* to group data values

- **Quantiles:** Cut points dividing a set of observations into equal-sized groups
    - Many quantiles ➔ Discovery of false keys
    - Few quantiles ➔ Lose of true keys

Initial Data

	PH
Wine1	3.15
Wine2	3.22
Wine3	3.23
Wine4	3.24
Wine5	3.56
Wine6	3.68

Quantiles

	PH
Wine1	3.15
Wine2	3.22
Wine3	3.23
Wine4	3.24
Wine5	3.56
Wine6	3.68

Final Data

	PH
Wine1	1
Wine2	1
Wine3	2
Wine4	2
Wine5	3
Wine6	3

KEY QUALITY MEASURES

1) **Key support**
   - **Intuition:** The higher is the support the more sure is a key

2) **Key exceptions**
   - **Intuition:** A 0-almost key is considered more reliable than a 10-almost key

3) **Key size**
   - **Intuition:** Keys composed of few properties are preferred (easier to interpret)

4) **Property correlation:** The dependence of properties co-appearing in a key
   - **Intuition:** The less correlated are the properties participating in a key, the more informative the key is

EXPERIMENTAL DATA

Chemical families	Concentration levels in wine	Analyzed molecules	Analysis methodology
Thiols	ppt	3MH = 3 mercaptohexanol 3MHA = 3 mercaptohexylacetate	LC-MS/MS
Esters	ppm	2PHEN= 2-phenylethanol AH= hexyl acetate AI= isoamyl acetate ABPE= phenethyl acetate DE= ethyl decanoate HE= ethyl hexanoate OE= ethyl octanoate BE= ethyl butyrate 2HPE= Ethyl lactate 3HBE= Ethyl 3-hydroxybutyrate 2MBE= Ethyl 2-methylbutyrate 2MPE= Ethyl isobutyrate 2HICE= Ethyl leucate	GC-MS/MS
C13-noriprenoïds	Ppb	BDAM= beta-damascenone BION= beta-ionone	GC-MS/MS
PDMS	Ppb	Dimethylsulfide potential= S-methylmethione + others compounds	GC-MS/MS
GSH	ppm	Glutathione	LC-MS/MS

EXAMPLES OF KEYS

	Year	Quantiles	Support	Probability	Size
[3MHA, BDAM, GSH, 2MPE, 3MH]	2014	5	73%	26%	5
[3MHA, GSH, OE, 2HICE, 3MH]	2014	5	73%	26%	5
[3MHA, AI, PDMS, 2MPE, 3MH]	2014	5	100%	26%	5
[3MHA, BE, PDMS, 2MPE, 3MH]	2014	5	100%	26%	5
[BDAM, OE, PDMS, 3MH]	2012	10	100%	17%	4
[GSH, OE, PDMS, 2PHEN]	2012	10	100%	17%	4
[AI, BDAM, 2HICE, 3MH]	2012	10	100%	17%	4
[3MHA, BDAM, GSH, 2MPE]	2013	5	63%	94%	4
[3MHA, BDAM, GSH]	2013	12	63%	64%	3
[AH, BDAM, GSH]	2013	12	63%	64%	3
[BE, 2HICE, 3MH]	2013	12	100%	64%	3
[BDAM, GSH, 3MH]	2013	12	63%	63%	3
[GSH, PDMS, 3HBE]	2013	10	63%	83%	3
[BDAM, GSH, 3MH]	2013	10	63%	83%	3
[GSH, PDMS, 3HBE]	2014	10	73%	63%	3
[PDMS, 3HBE, 3MH]	2014	12	100%	44%	3
[3MHA, GSH, PDMS]	2014	12	73%	44%	3
[BE, GSH, 3MH]	2014	12	73%	44%	3

VALIDATED KEYS

- **18 out of 104 keys (18%) were validated by the expert**
- **Support from 63% to 100%**
  - Keys with low support can be as well significant

- **Evaluated keys contain from 3 to 5 properties**
  - Expert chose keys with big size (on contrary to the initial intuition)

- **Example: Key {AI, BDAM, 2HICE, 3MH}**
    - Correlations from 0.05 to 0.42
  - Properties not highly correlated ➔ interesting keys

- **First step for predicting wine taste and wine component concentration**

OUTLINE

- Introduction
- Key discovery for data linking
- Link Invalidation
- Contextual identity
- Conclusion

OAEI*: RECENT RESULTS

- **Data Linking results for OAEI 2018 - SPIMBENCH Track**

SPIMBENCH Sandbox

	Precision	Recall	F-measure	Time in ms
AML	0.8348	0.8963	0.8645	6220
Lily	0.8494	1.0	0.9185	1960
LogMap	0.9382	0.7625	0.8413	5887

SPIMBENCH Mainbox

	Precision	Recall	F-measure	Time in ms
AML	0.8385	0.8835	0.8604	37190
Lily	0.8546	1.0	0.9216	3103
LogMap	0.8925	0.7094	0.7905	23494

*** OAEI: Ontology Alignment Evaluation Initiative**

OAEI*: RECENT RESULTS

- **Data Linking results for OAEI 2018 - SPIMBENCH Track**

SPIMBENCH Sandbox

	Precision	Recall	F-measure	Time in ms
AML	0.8348	0.8963	0.8645	6220
Lily	0.8494	1.0	0.9185	1960
LogMap	0.9382	0.7625	0.8413	5887

SPIMBENCH Mainbox

	Precision	Recall	F-measure	Time in ms
AML	0.8385	0.8835	0.8604	37190
Lily	0.8546	1.0	0.9216	3103
LogMap	0.8925	0.7094	0.7905	23494

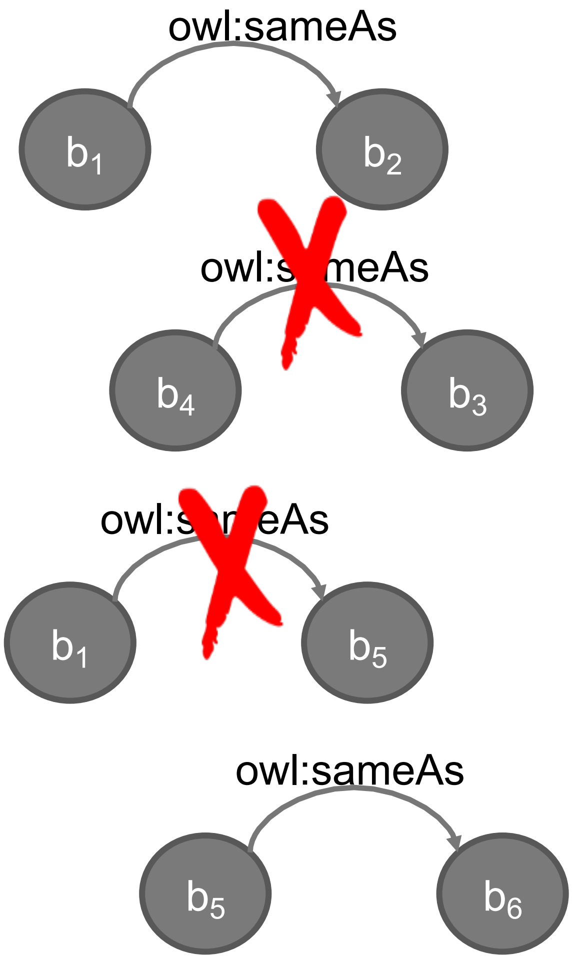
*** OAEI: Ontology Alignment Evaluation Initiative**

IDENTITY PROBLEM

- 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.

- [Halpin et al. 2010] showed that 37% of owl:sameAs links randomly selected among 250 identity links between books were incorrect.

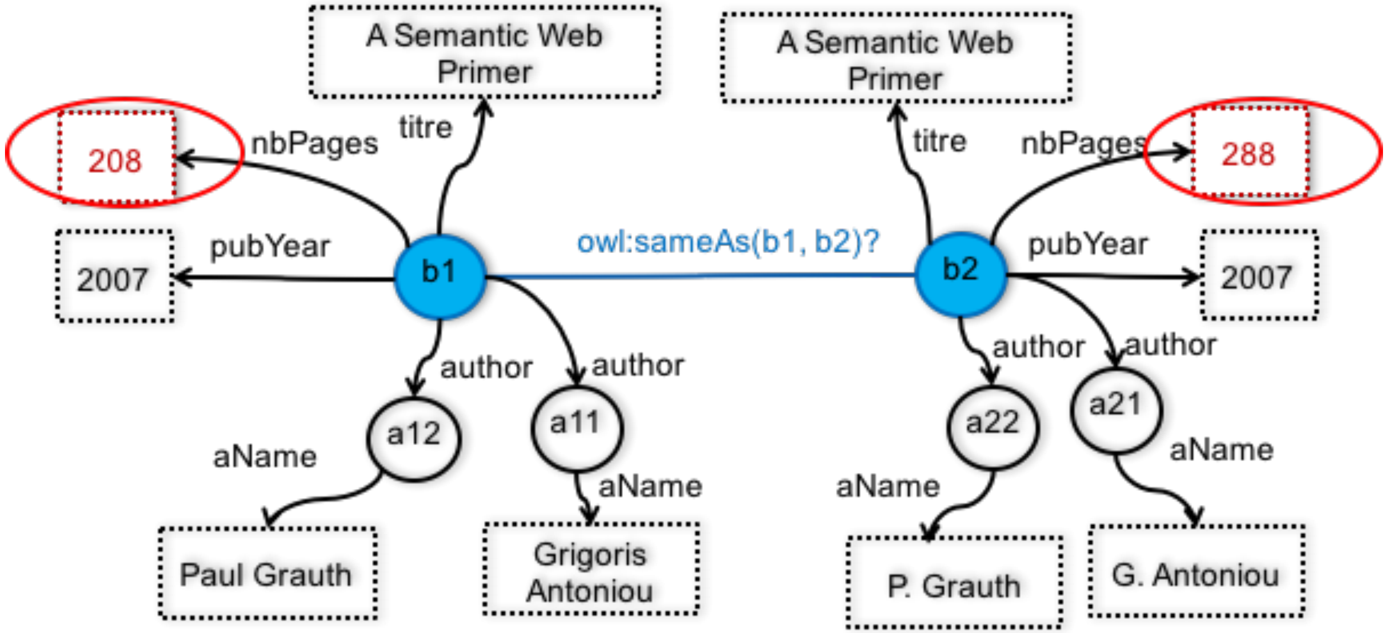
- Automatic data linking tools do not guarantee 100% precision, because of:
  - Errors, missing information, data freshness, etc.



IDENTITY LINK INVALIDATION

[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.



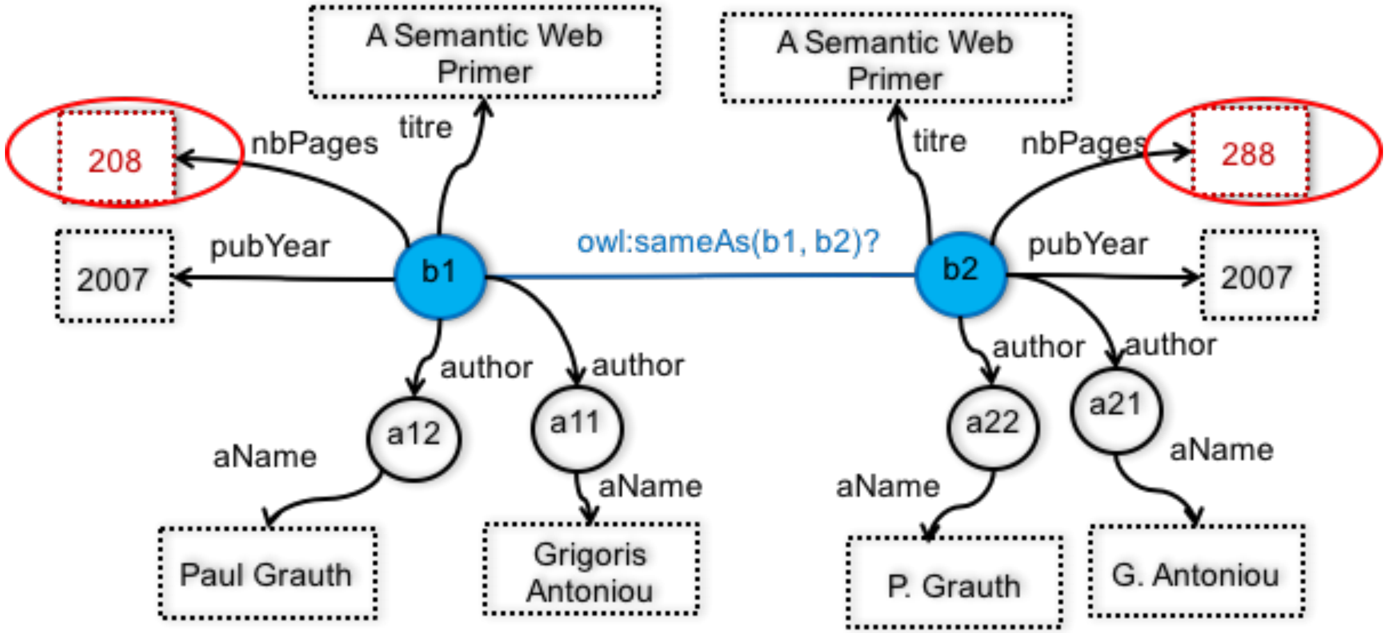
nbPages is a Functional Property

Laura Papaleo, Nathalie Pernelle, Fatiha Saïs, Cyril Dumont:Logical Detection of Invalid SameAs Statements in RDF Data. EKAW 2014: 373-384

IDENTITY LINK INVALIDATION

[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

- Improvements in data linking **precision** up to **25%**

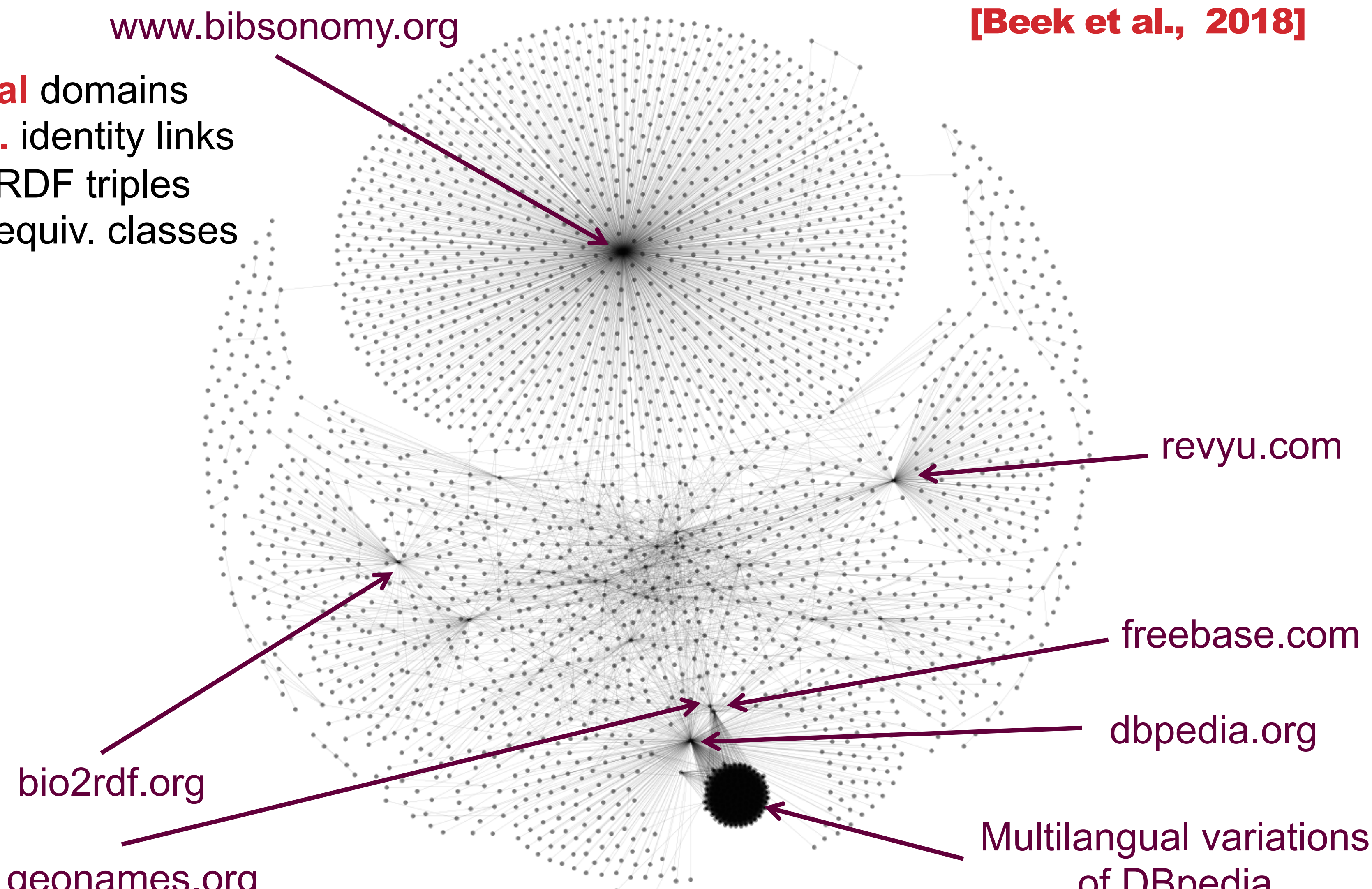
Limits:

- Scalability problems and 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

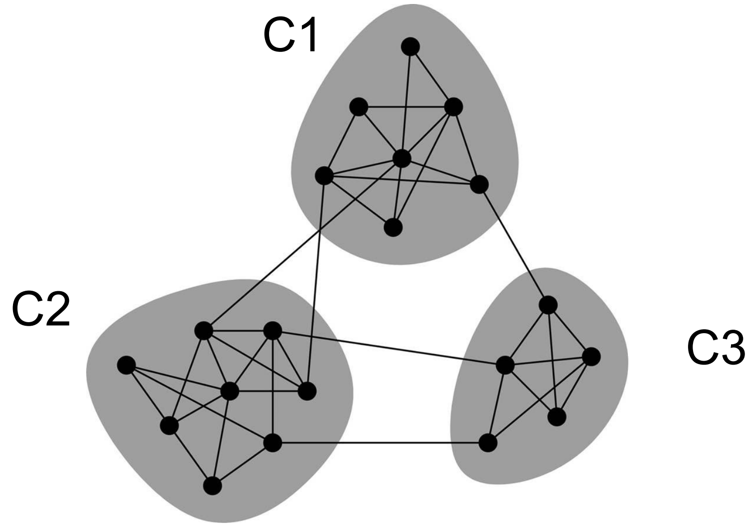
The largest identity set contains 177 794 terms:

Different countries
Different cities
Albert Enstein

→ quality problems

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**.

Joe Raad, Wouter Beek, Frank van Harmelen, Nathalie Pernelle, Fatiha Saïs:Detecting Erroneous Identity Links on the Web Using Network Metrics. International Semantic Web Conference (1) 2018: 391-407

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

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

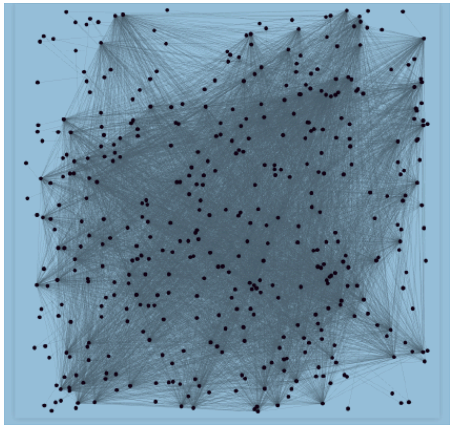
Correct link 0 — Error degree — 1 Erroneous link

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

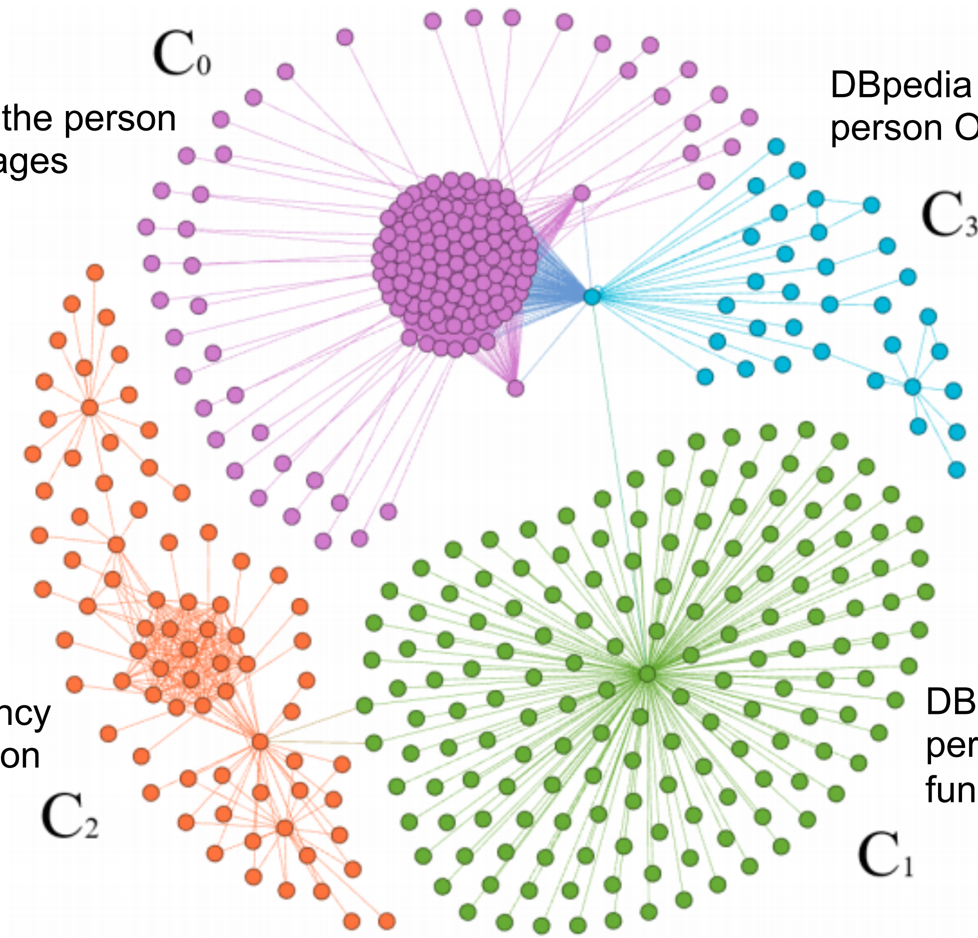
NETWORK BASED

[Raad *et al.*, ISWC 2018]

Barack Obama's Equality Set

DBpedia IRIs referring to the person Obama in different languages

DBpedia IRIs referring to the person Obama, his senator career



IRIs referring to the presidency and the Obama administration

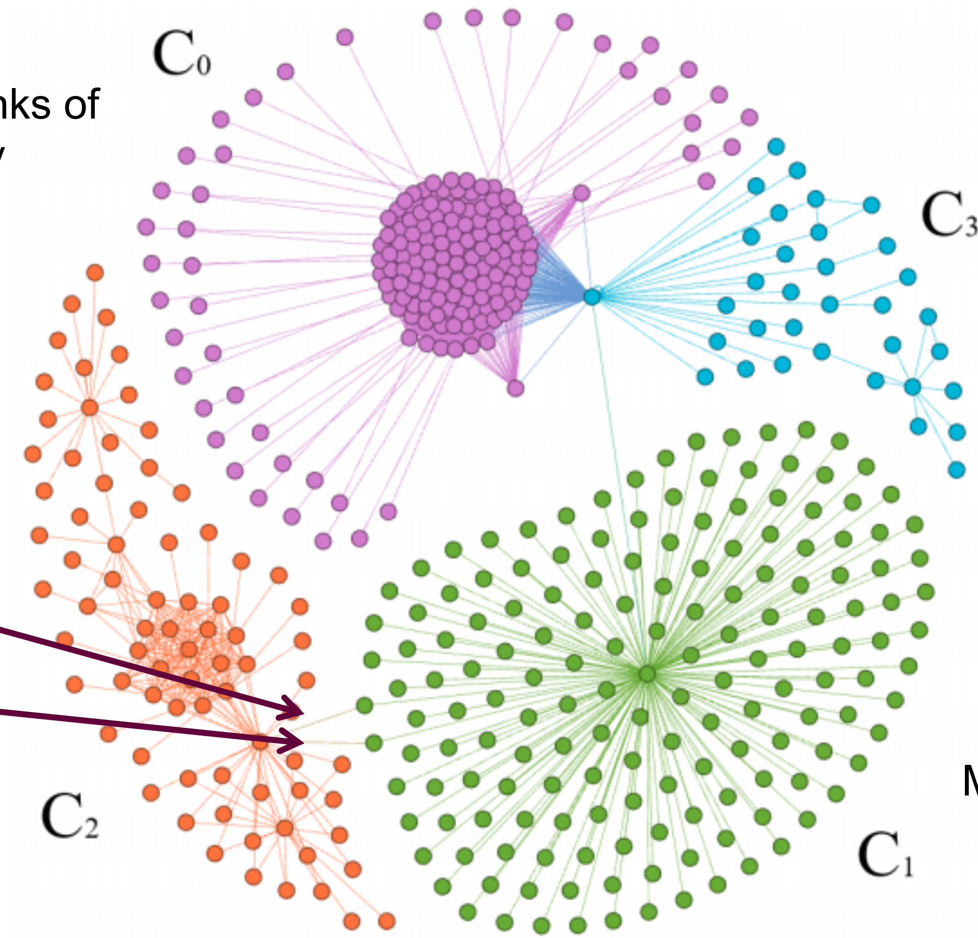
DBpedia IRIs referring to the person Obama in different functions

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

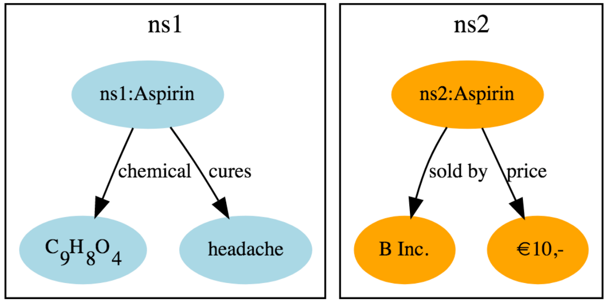
Precision**: manual evaluation of 200 links**

- 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: **1.26M** 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 %**

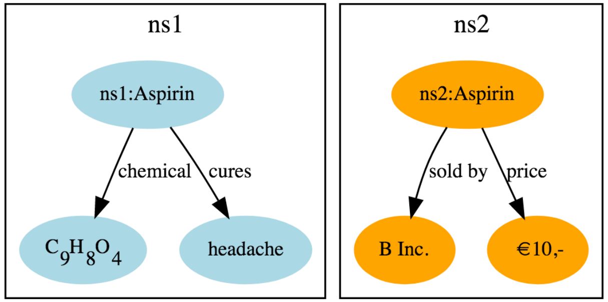
SOMETIMES, WE NEED WEAKER IDENTITY ...

- 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*



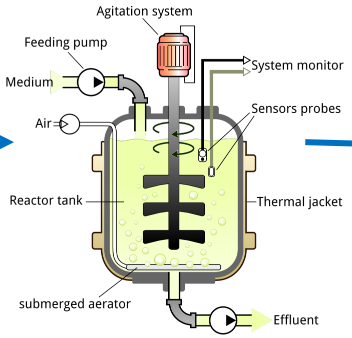
SOMETIMES, WE NEED WEAKER IDENTITY …

- 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*



- **Identity over time** poses problems
  - a material could it be considered the same, even though some (or even all) of its original components have been replaced by new ones.

Bioreactor 1 →



→ Bioreactor 2

Same as Bioreactor 1 ?

OWL:SAMEAS PREDICATE IS TOO STRICT

- 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:

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

OWL:SAMEAS PREDICATE IS TOO STRICT

- 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 **weak-identity links** ➔ **Contextual Identity**

CONTEXTUAL IDENTITY LINKS

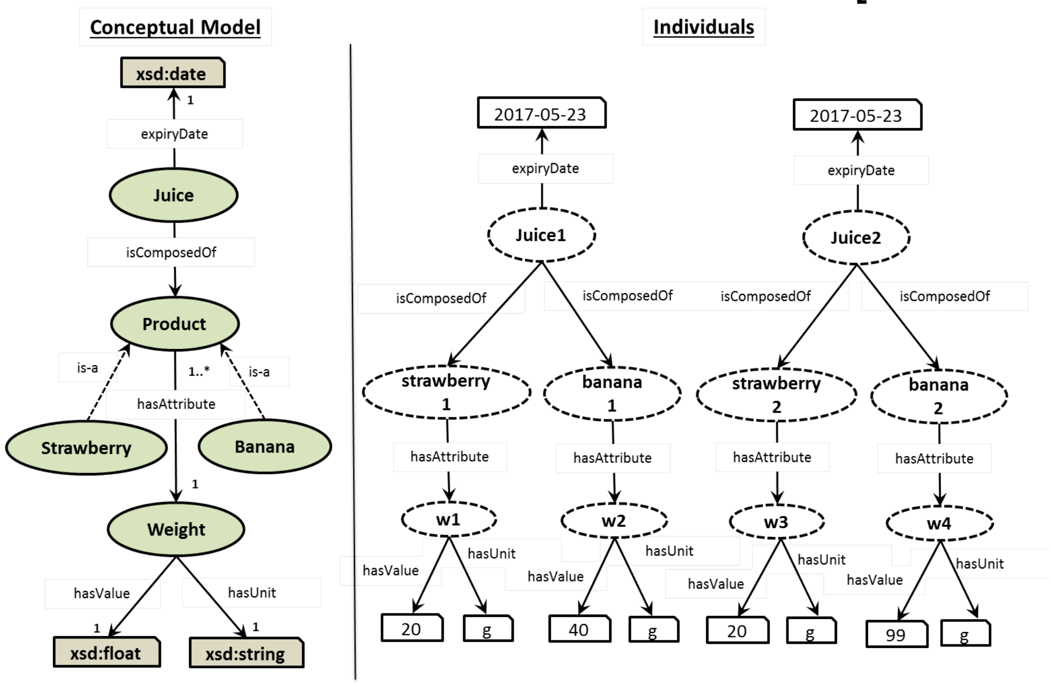
[Raad et al., 2017]

- New predicate *:<u>identiConTo</u>* for expressing **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

Joe Raad, Nathalie Pernelle, Fatiha Saïs: Detection of Contextual Identity Links in a Knowledge Base. K-CAP 2017: 8:1-8:8

CONTEXTUAL IDENTITY LINKS

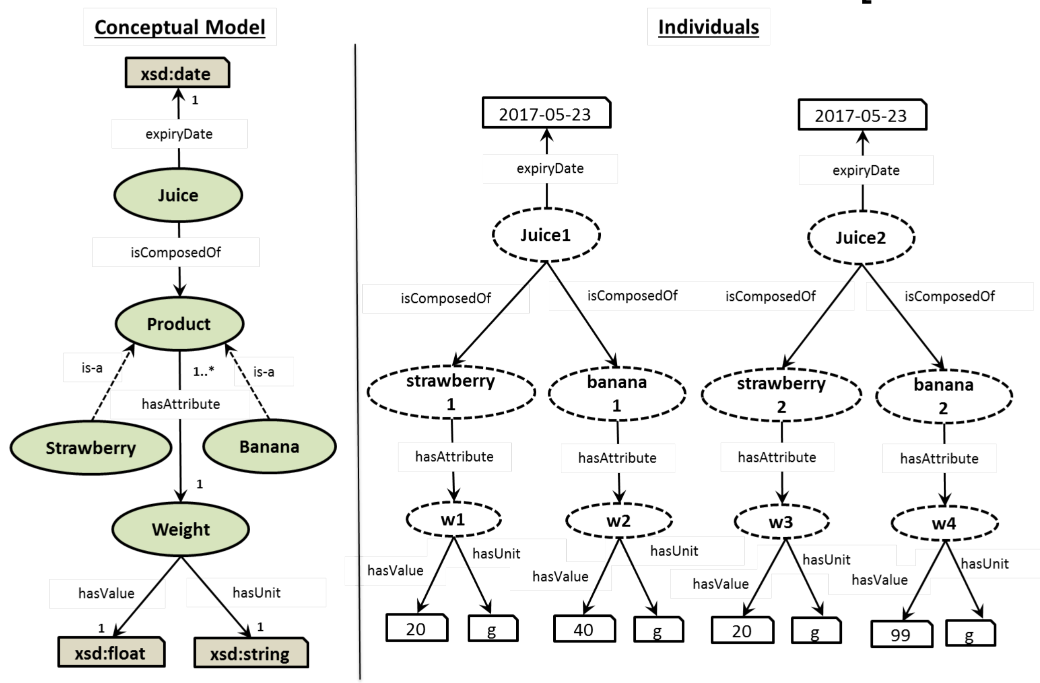
[Raad et al., 2017]



CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Contexts are defined as a **sub-ontology** of the domain ontology



Contextual Identity Link Example

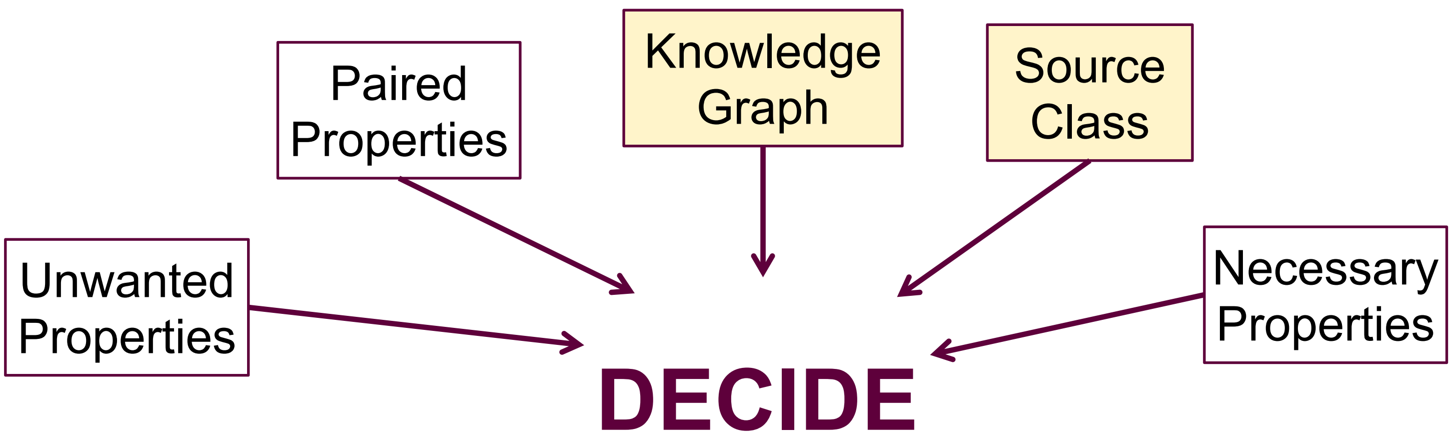
Π_a**(Juice) = {** (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf $^{-1}$}),
(Strawberry, {rdf:Type}, {hasAttribute, isComposedOf $^{-1}$}),
(Weight, {rdf:Type, hasValue, hasUnit}, {hasAttribute^{-1}}) **}**

identiConTo$_{<\Pi_a(\text{Juice})>}$**(juice1, juice2)**

CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

It automatically detects and adds these contextual identity links in the knowledge graph



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

For each pair of instances (i_1, i_2) of the source class
set of the most specific global contexts in which (i_1, i_2) are identical

LIONES*: CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Transformation of Micro-organisms

Digestion Process

PO²
A Process and Observation Ontology

RDF

PO²
A Process and Observation Ontology
CellExtraDry

- **Classes:** ≈ 4 700
- **Individuals:** ≈ 415 000
- **Statements:** ≈ 1 700 000

RDF

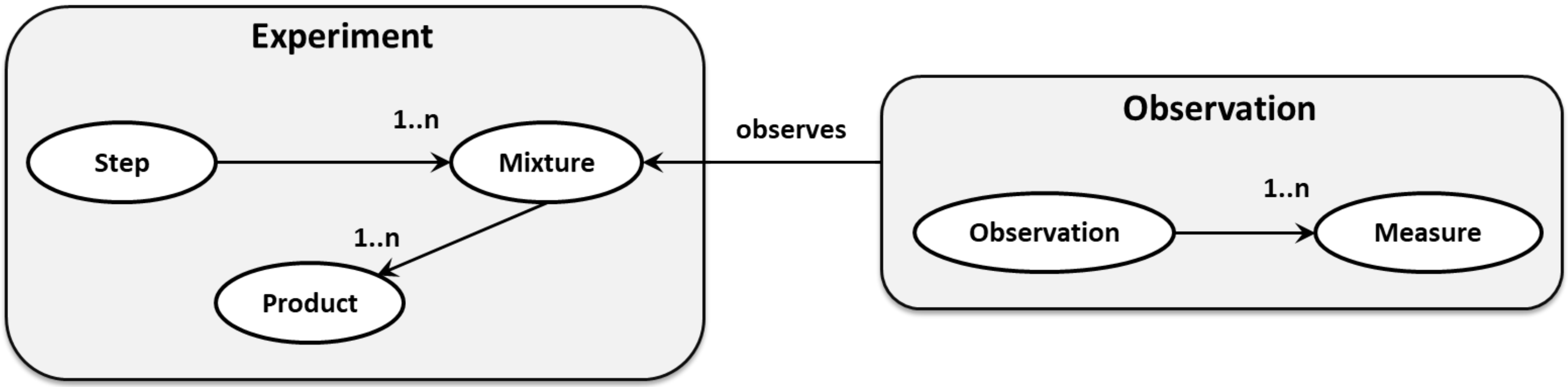
PO²
A Process and Observation Ontology
Carredas

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

* LIONES project funded by CDS Paris Saclay (2015-2018)

LIONES: CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]



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

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

with $m_1 \simeq m_2$

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

LIONES: 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 has evaluated the plausibility of the best **20 rules** (in termes of error rate and support)

LIONES: 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 has evaluated the plausibility of the best **20 rules** (in termes of error rate and support)

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

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

WARM RULES – GRADUAL CAUSAL RULES DETECTION IN KNOWLEDGE GRAPHS

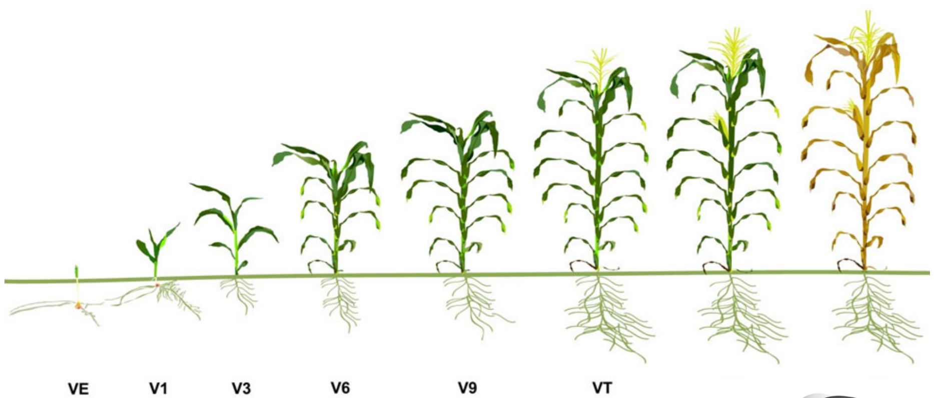
MIA/INRA

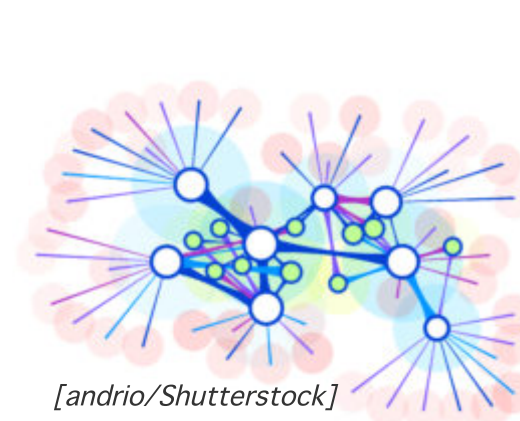
GQE

LRI/UPS

Application to plant development in climatic warming preoccupation

Phenology: the study of seasonal cycles of plants (timing and duration of flowering, fruiting, leaf out and leaf drop)





+

+

Sensor Measurement

Weather Forecast

Expert Knowledge and Observations

Domain-specific Knowledge Graphs

GRADUAL CAUSALITY RULE DISCOVERY IN RDF KGS

INSTITUT DATAIA
Science des données, Intelligence & Société

$$\underbrace{\mathrm{Pid1} \wedge \mathrm{Pid2} \wedge \ldots \wedge \mathrm{Pidn}}_{\text{Identity}} \wedge \underbrace{\mathrm{Pvar1} \wedge \mathrm{Pvar2} \wedge \ldots \wedge \mathrm{Pvarm}}_{\text{Variability/cause}} \Rightarrow \underbrace{\text{Peffect-var}}_{\text{Effect}}$$

Example:

Same-leafSize(X,Y) ∧ SameGroundPh(X,Y) ∧

Humidity(X,h1):t1 ∧ Humidity(Y,h2):t2 ∧

Temp(X,temp1):t1 ∧ Temp(Y,temp2):t2 ∧

(h1> h2) ∧ (temp1 < temp2) ∧ (t1< t2) ⇒ flowering-delay(Y)

CONCLUSION

- **Semantic Web standards, agronomic data/knowledge and many applications are there**

- **Promising applications are emerging for which reasoning on data is central:**
  - Information retrieval, decision-support, digital-assistants, …

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

DATA LINKING AND KNOWLEDGE DISCOVERY IN RDF DATA: METHODS AND SOME FEEDBACK FROM AGRONOMIC APPLICATIONS

FATIHA SAÏS

LAHDAK@LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

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

1ST DATAIA DAYS « LIFE SCIENCES & AI», DEC. 4TH 2019

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®

université PARIS-SACLAY

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.

KEY QUALITY MEASURES

Non-key probability: The probability that a set of properties contains instances sharing the same values for this set

$$\boxed{Pr_k = 1 - e^{-\frac{n(n-1)}{2p}}}$$

with

$$p = \Pi_{i=0}^{j} m_i$$

where

j : # of properties
m$_i$: # of distinct values

KEY QUALITY MEASURES

Non-key probability

- **Example:** 100 distinct wines

	Case1	Case2	Case3
WineName	15 distinct values	50 distinct values	90 distinct values
YearProduction	10 distinct values	50 distinct values	80 distinct values
Non-key probability	**1**	**0.87**	**0.49**

- **Intuition:** Higher is the non-key probability of *{wineName,yearProduction}* more the discovered key is **important**

EXPERIMENTS

Experimental data: 3 wine aroma datasets

- Different chemical based flavourings of wine
  - Concentration of each flavour in a wine

	# Instances	# Flavours
D1 (2011 – 2012)	63	19
D2 (2012 – 2013)	59	19
D3 (2013 – 2014)	44	19

Goal: Verify the interest of keys in numerical data

- Evaluate the **impact of quantiles** in the results
- Evaluate the **quality measures**

DATA PREPROCESSING

Quantiles

- Use the non-key probability to define the number of quantiles

5, 10, 12 quantiles

- Setting it at less than 5 => no keys are obtained
- Using 5 to 12 quantiles ensured a significantly high probability