INFORMATION INTEGRATION

FATIHA SAÏS

Slides: https://www.lri.fr/~sais/D2K/course1.pdf

UNIVERSITÉ PARIS SACLAY

D&K- DATA & KNOWLEDGE MASTER

COURSE PLANNING

- 02/12/2019, 13h30 - 16h30 (F. Saïs)

  Part 1- Semantic data integration – Data Linking and Identity Problem

- 09/19/2019, 13h30 - 16h33 (F. Saïs)

  Part 1- Cont. + Lab exercises on data linking and Web of data

- 16/12/2018, 13h30 - 16h30 (N. Pernelle)

  Part 2- Semantic data integration – Ontology Alignment and Knowledge discovery + Presentation of the projects (for Course Grading)

- 06/01/2020, 13h30 - 16h30 (F. Saïs, N. Pernelle)

  Lab session on projects

DI in Linked Data

COURSE PLANNING

- **02/12/2019, 13h30 - 16h30 (F. Saïs)**

  Part 1- Semantic data integration – Data Linking and Identity Problem

- **09/19/2019, 13h30 - 16h33 (F. Saïs)**

  Part 1- Cont. + Lab exercises on data linking and Web of data

- **16/12/2018, 13h30 - 16h30 (N. Pernelle)**

  Part 2- Semantic data integration – Ontology Alignment and Knowledge discovery + Presentation of the projects (for Course Grading)

- **06/01/2020, 13h30 - 16h30 (F. Saïs, N. Pernelle)**

  Lab session on projects

- **13/01/2020, 13h30 - 16h30 (S. Cohen-Boulakia)**

  Part 3- Querying and navigating through real biological databases, levels of heterogeneity, major kinds of data integration architecture to integrate bio data

- **20/01/2020, 13h30 - 16h30 (L. Ibanescu)**

  Part 4- Ontology modelling and semantic annotation

- **03/02/2020, 13h30 - 16h30 (All professors): Project evaluation**

DI in Linked Data

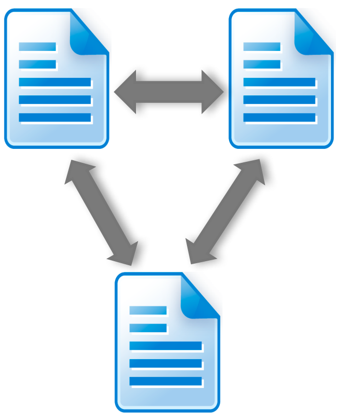
DI in Life Sciences

OUTLINE

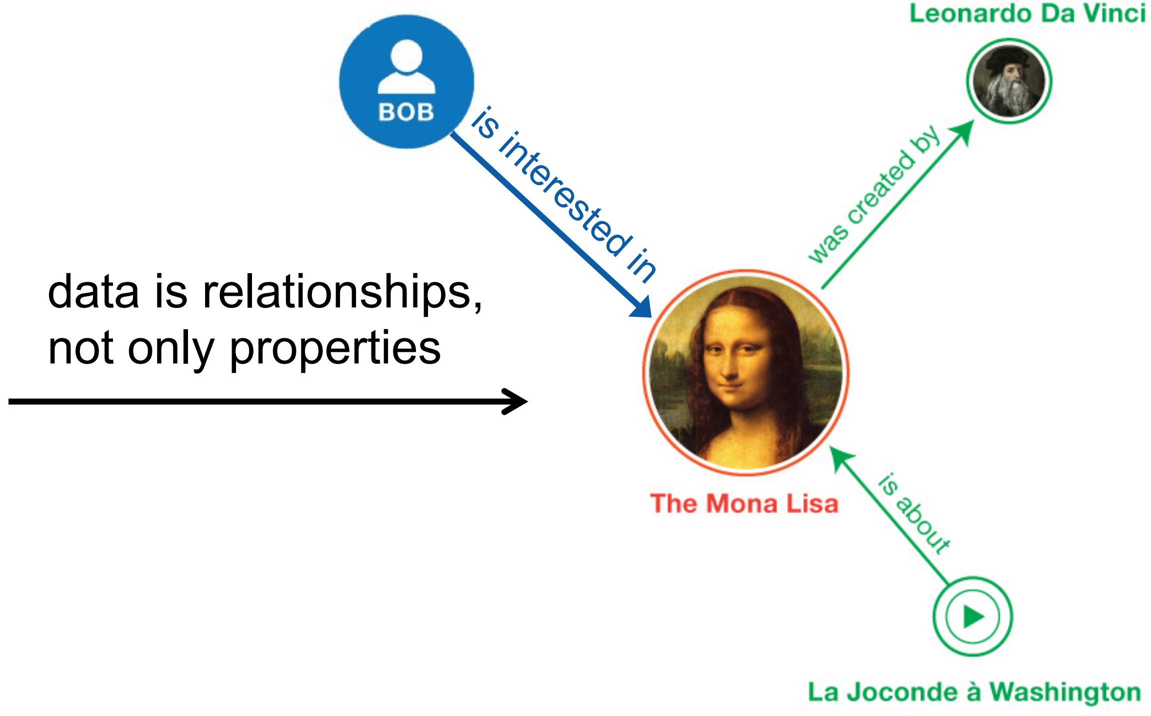
- **Introduction**
  - Linked Data
  - Knowledge graphs
  - Knowledge graph refinement
- **Data Linking**
- **Identity Problem**
- **Conclusion**

FROM THE WWW TO THE WEB OF DATA

- applying the principles of the WWW to data



data is relationships,
not only properties



LINKED DATA PRINCIPLES

Tim Berners Lee, 2006

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

→ so people can look up the data

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

→ to provide data for interested parties

③ **Include links to other resources**

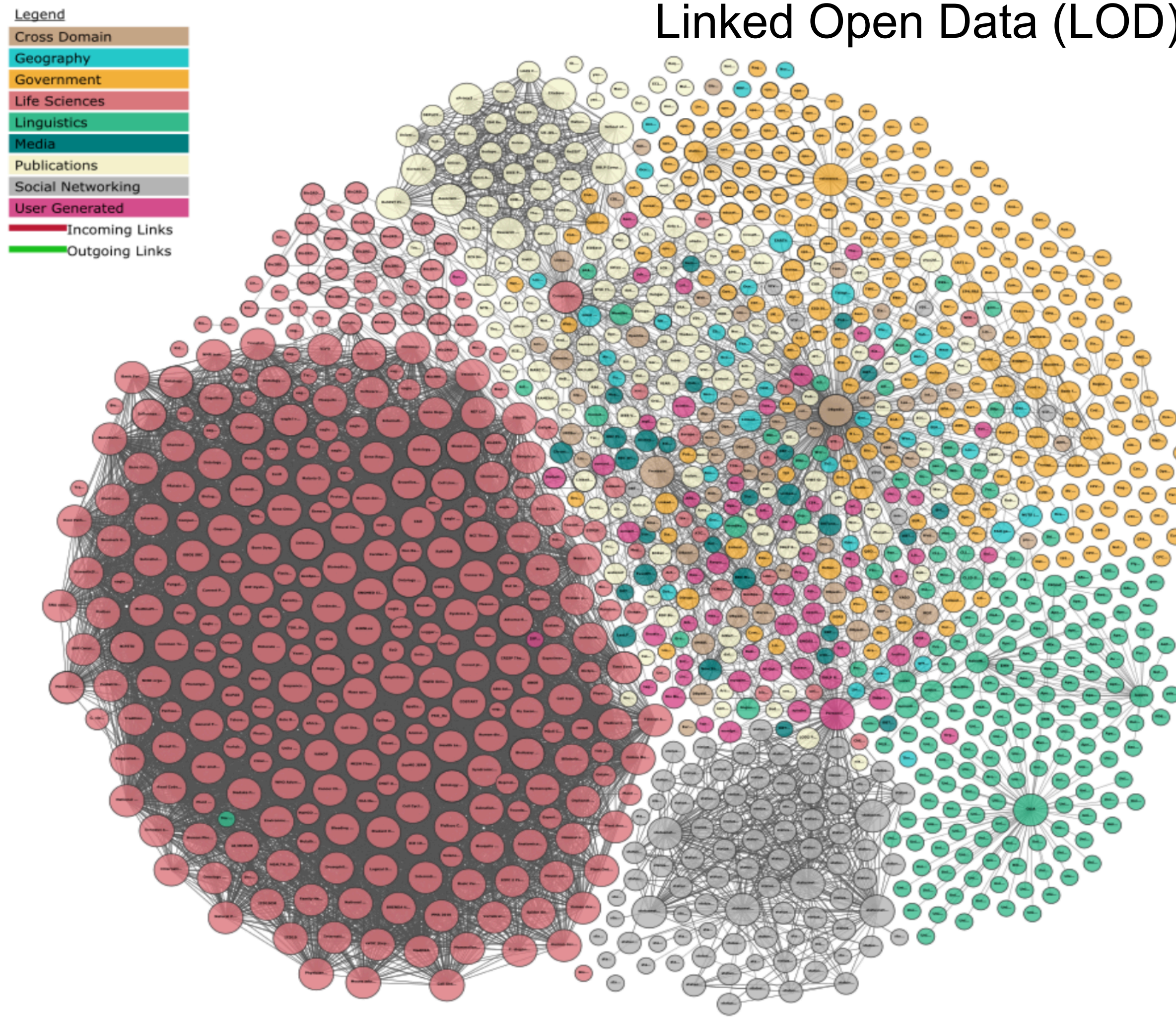
→so people can discover more information

→bridging disciplines and domains

➔ Unlock the potential of isolated repositories (islands)

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

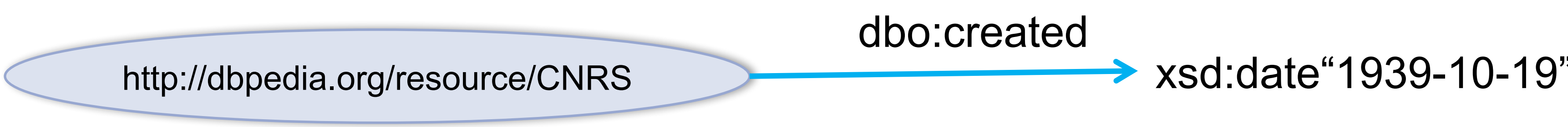
"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **RDF**: a data model for declaring metadata that describe resources on the Web
- **Resources**: Web pages, video or music files, PDF files, Web services, … identified by **URIs** (**Uniform Resource Identifiers**).

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **RDF**: a data model for declaring metadata that describe resources on the Web
- **Resources**: Web pages, video or music files, PDF files, Web services, … identified by **URIs** (**Uniform Resource Identifiers**).
- **Statements of < subject predicate object >**



Subject Predicate Object

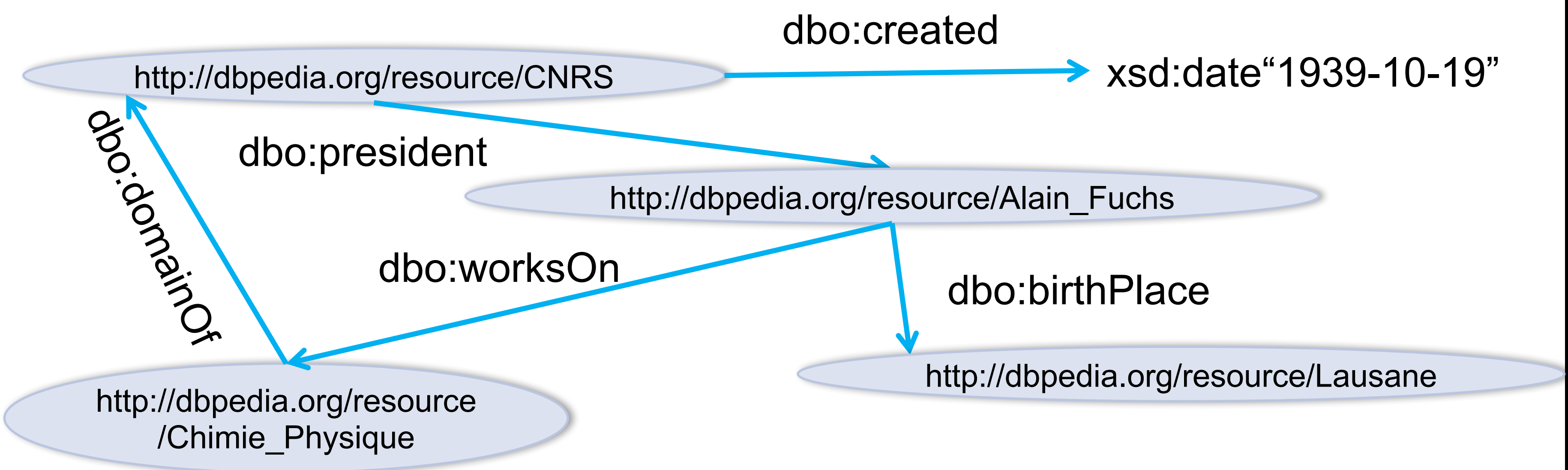
… is called a triple

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

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)

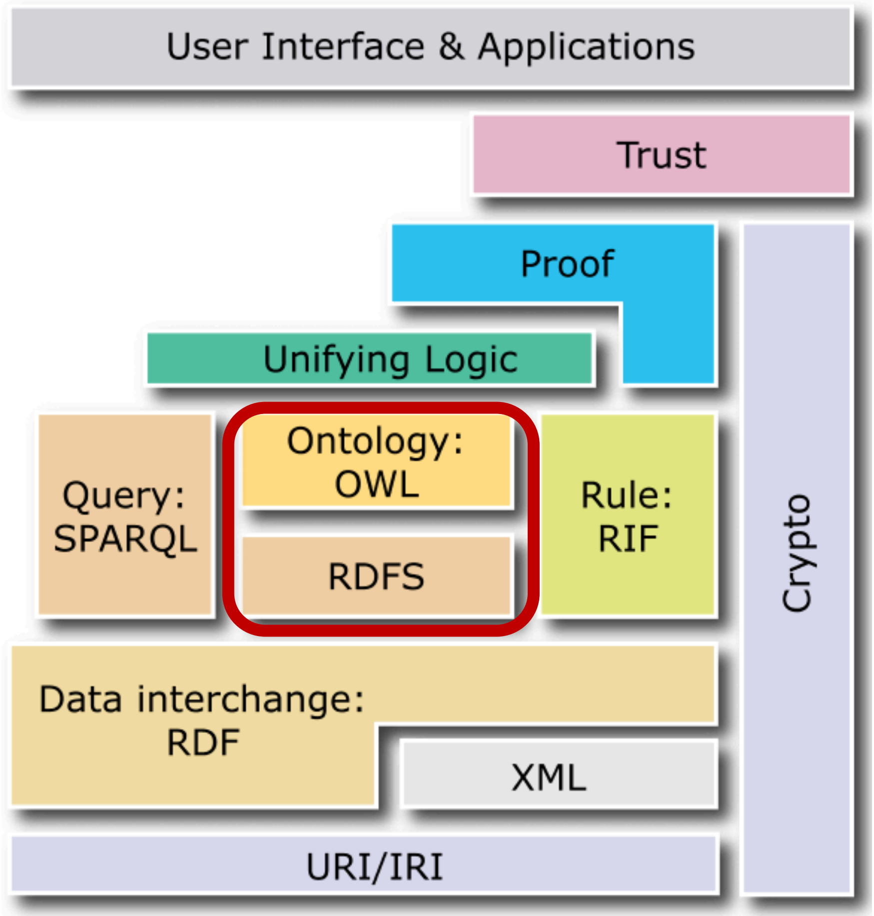
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema

- Lightweight ontologies

OWL – Web Ontology Language

- Expressive ontologies

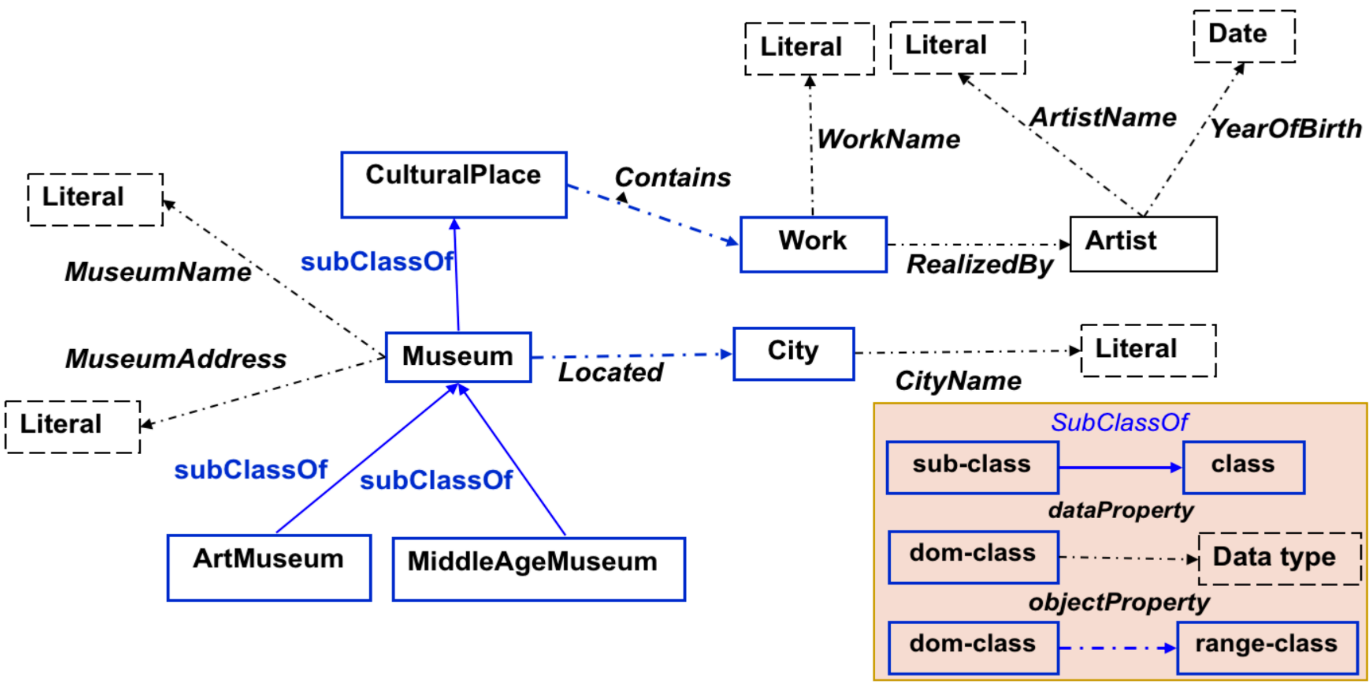


Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png

OWL – WEB ONTOLOGY LANGUAGE

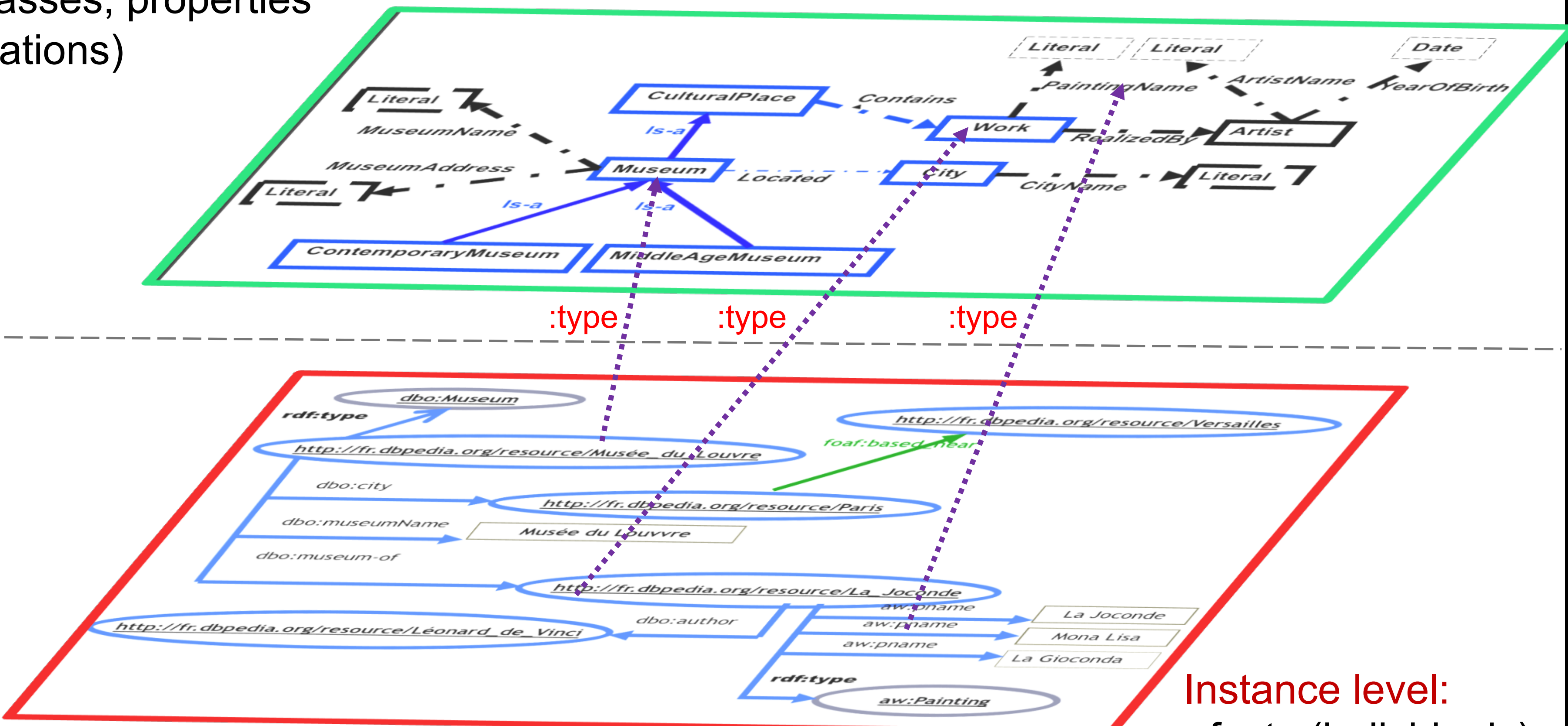
- **Classes:** concepts or collections of objects (individuals)
- **Properties:**
  - owl:DataTypeProperty (attribute)
  - owl:ObjectProperty (relation)
- **Individuals:** ground-level of the ontology (instances)

- **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)

OWL ONTOLOGY - REASONING

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

  Pompidou museum is an **Art Museum**

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

  Pompidou museum contains ***Hallucination partielle***

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

- **Infer that:**

➔ Pompidou museum is a CulturalPlace

< Pompidou_museum rdf:type CulturalPlace> .

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

➔ ***Hallucination partielle*** is a Work

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

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

KNOWLEDGE GRAPHS

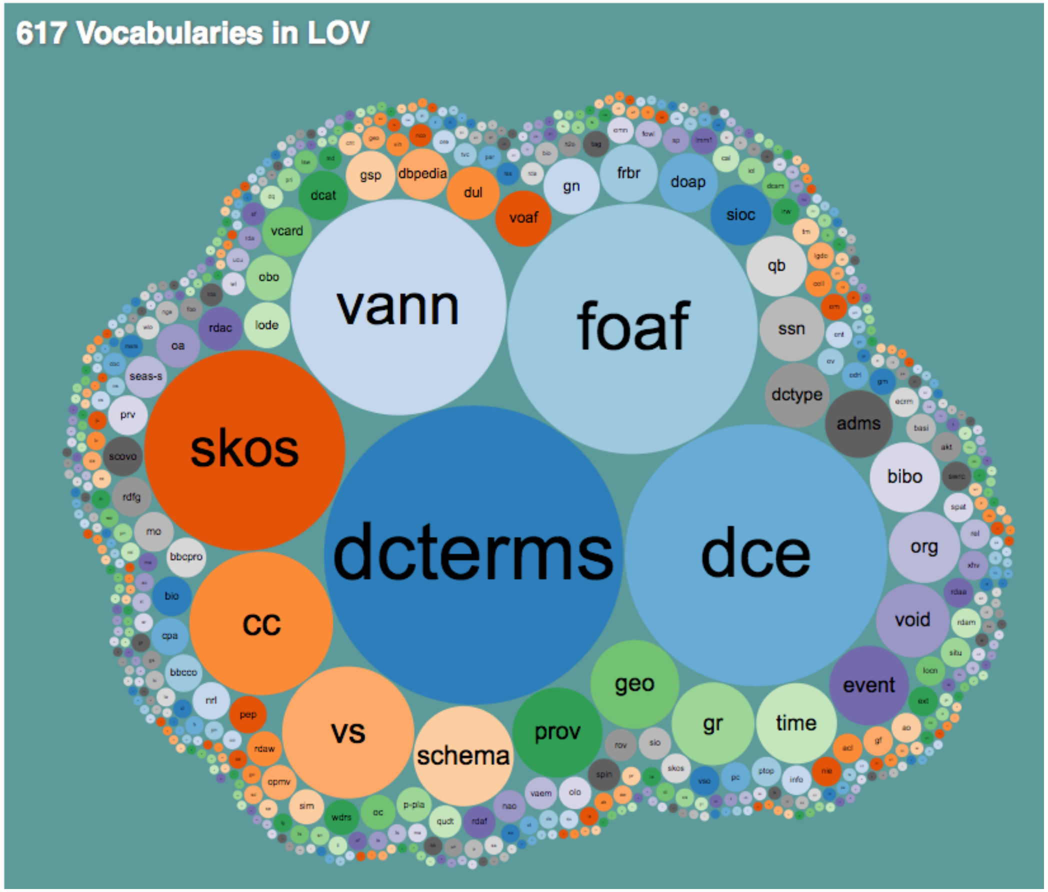
LINKED OPEN VOCABULARIES (LOV)

Linked Open Vocabularies

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

- Example:

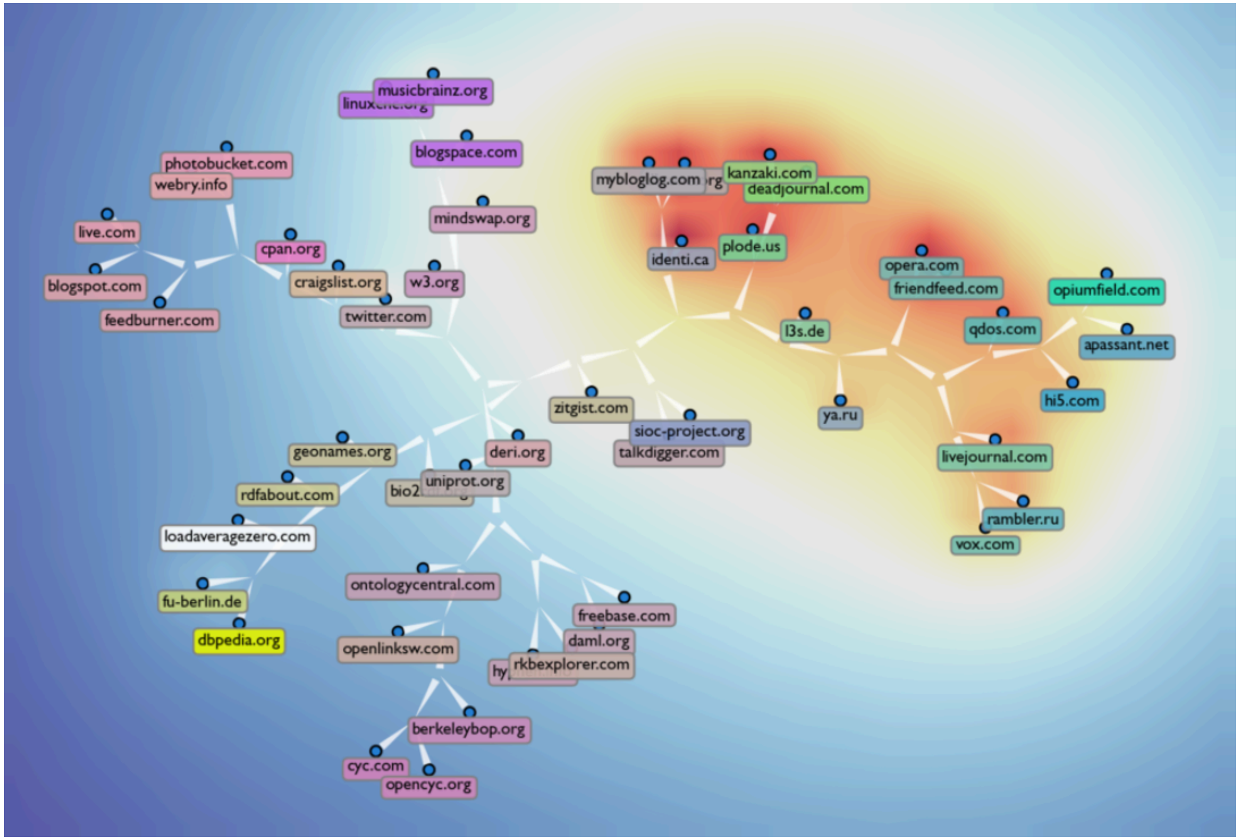
http://lov.okfn.org/dataset/lov/vocabs/foaf



ONTOLOGY PREDICATES SPREAD ON THE SEMANTIC WEB

- **RDFa (or [Resource Description Framework] in Attributes)**
- Top 50 web sites publishing Semantic Web data, clustered by predicates used.

FOAF



OUTLINE

- **Introduction**
  - Linked Data
  - Knowledge graphs
  - Knowledge graph refinement
- **Data Linking**
- **Identity Problem**
- **Conclusion**

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

2008

yago select knowledge

2012

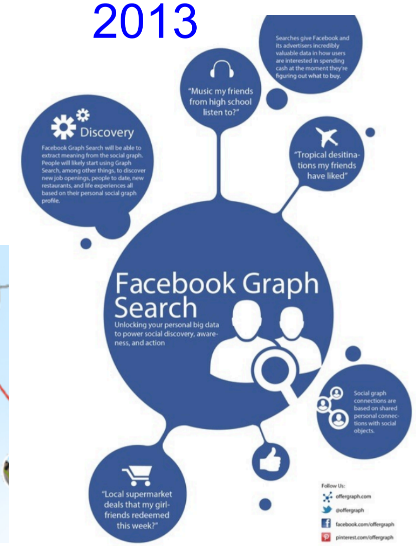
WIKIDATA

2007

Freebase

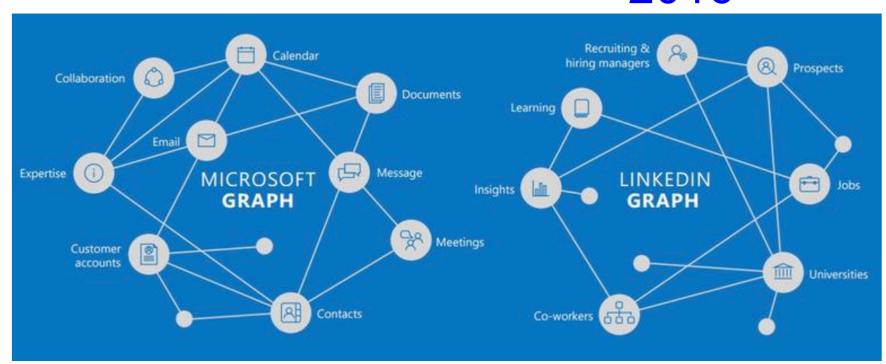
2012

2013



2015

2016



2013

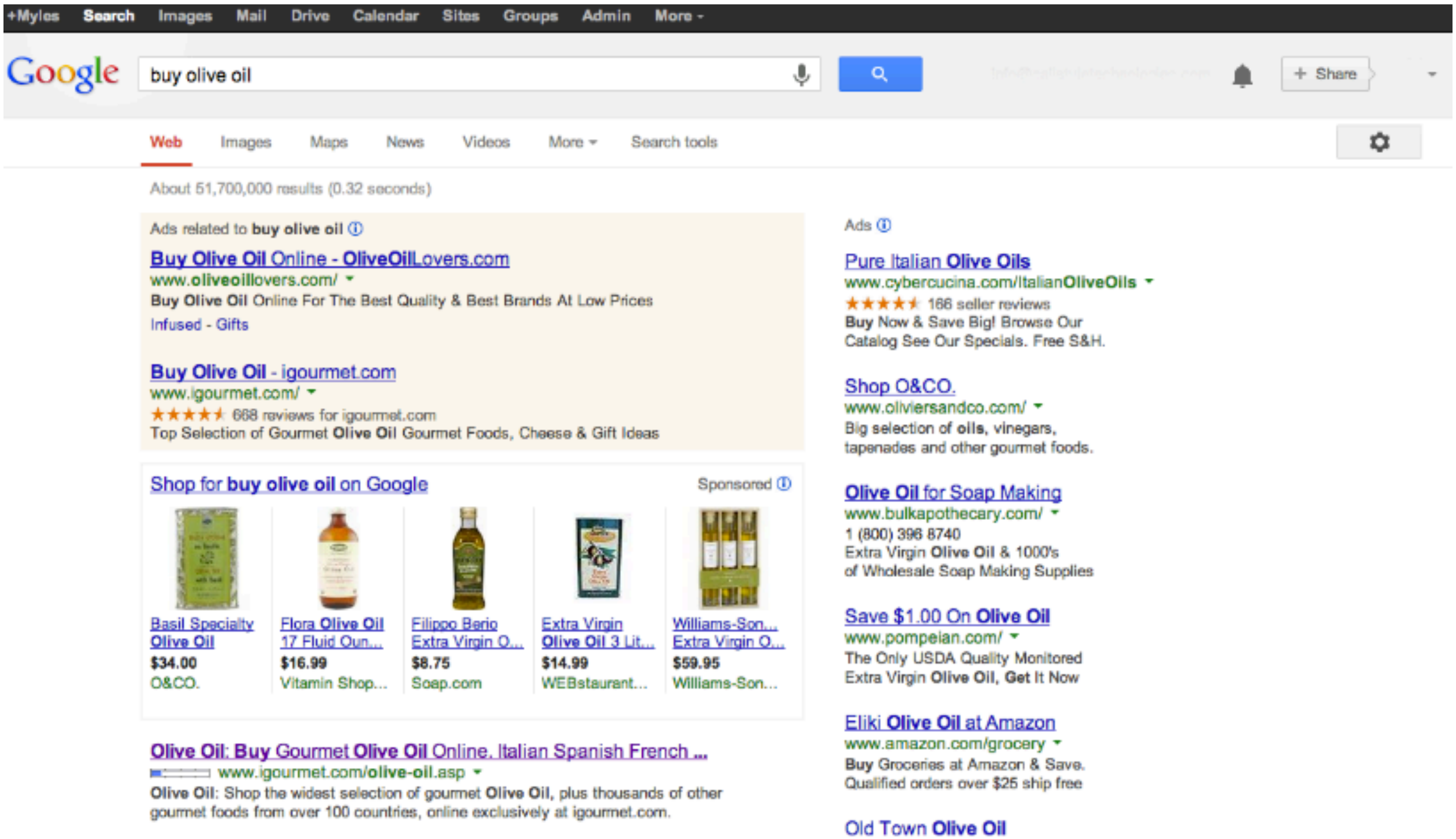
Yahoo!

Yahoo's new SERP designs mobile and knowledge graph

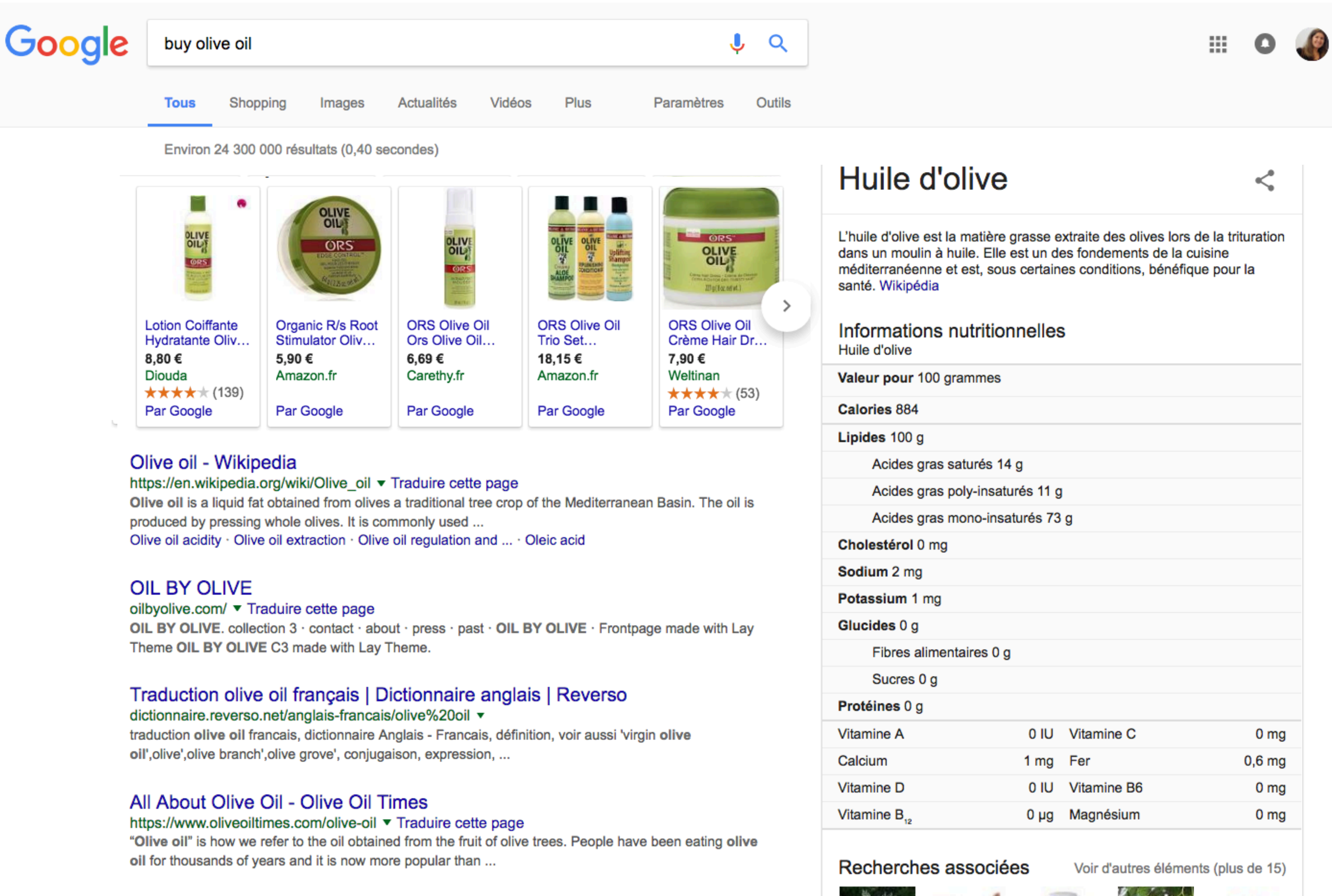
Academic side

Commercial side

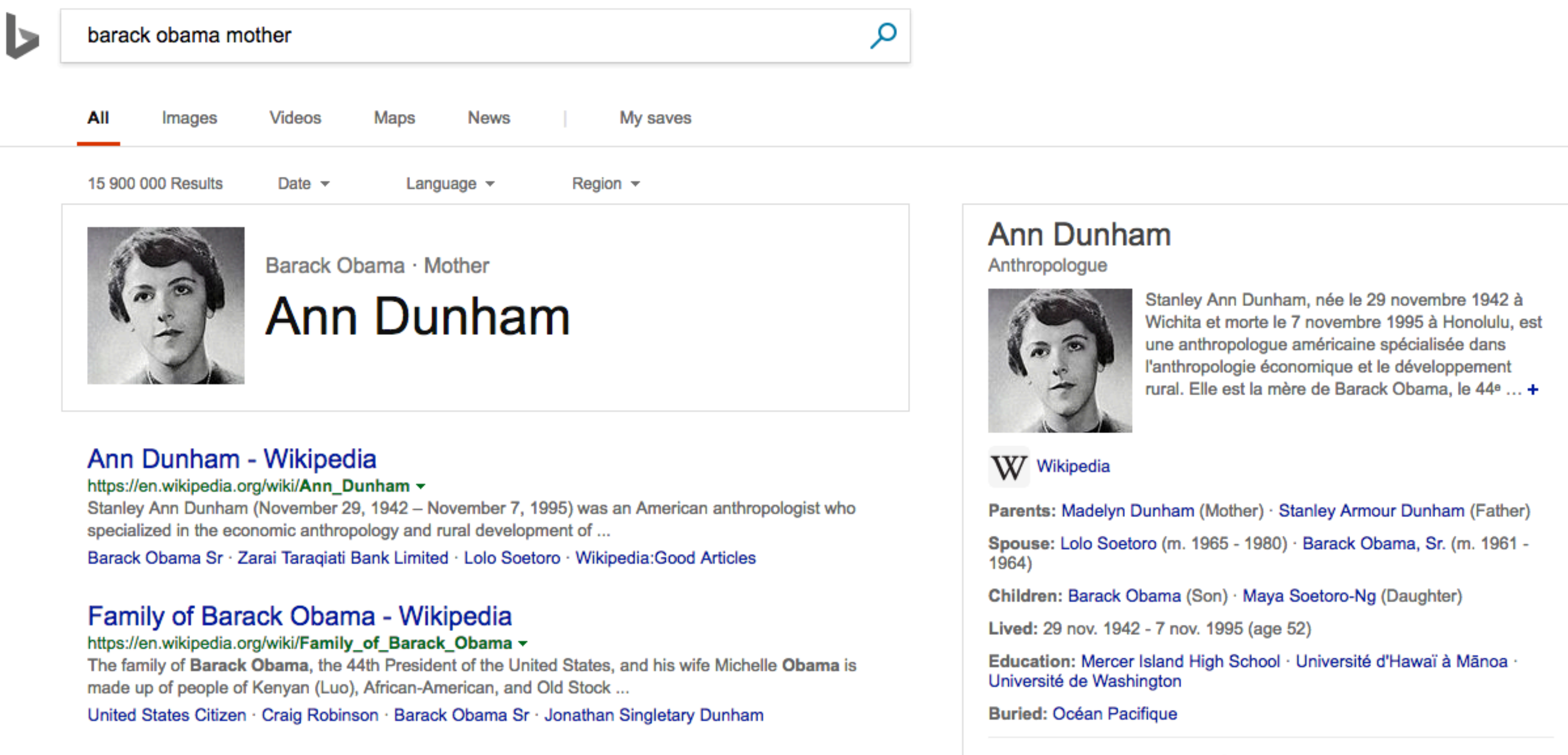
WEB SEARCH WITHOUT KNOWLEDGE GRAPHS



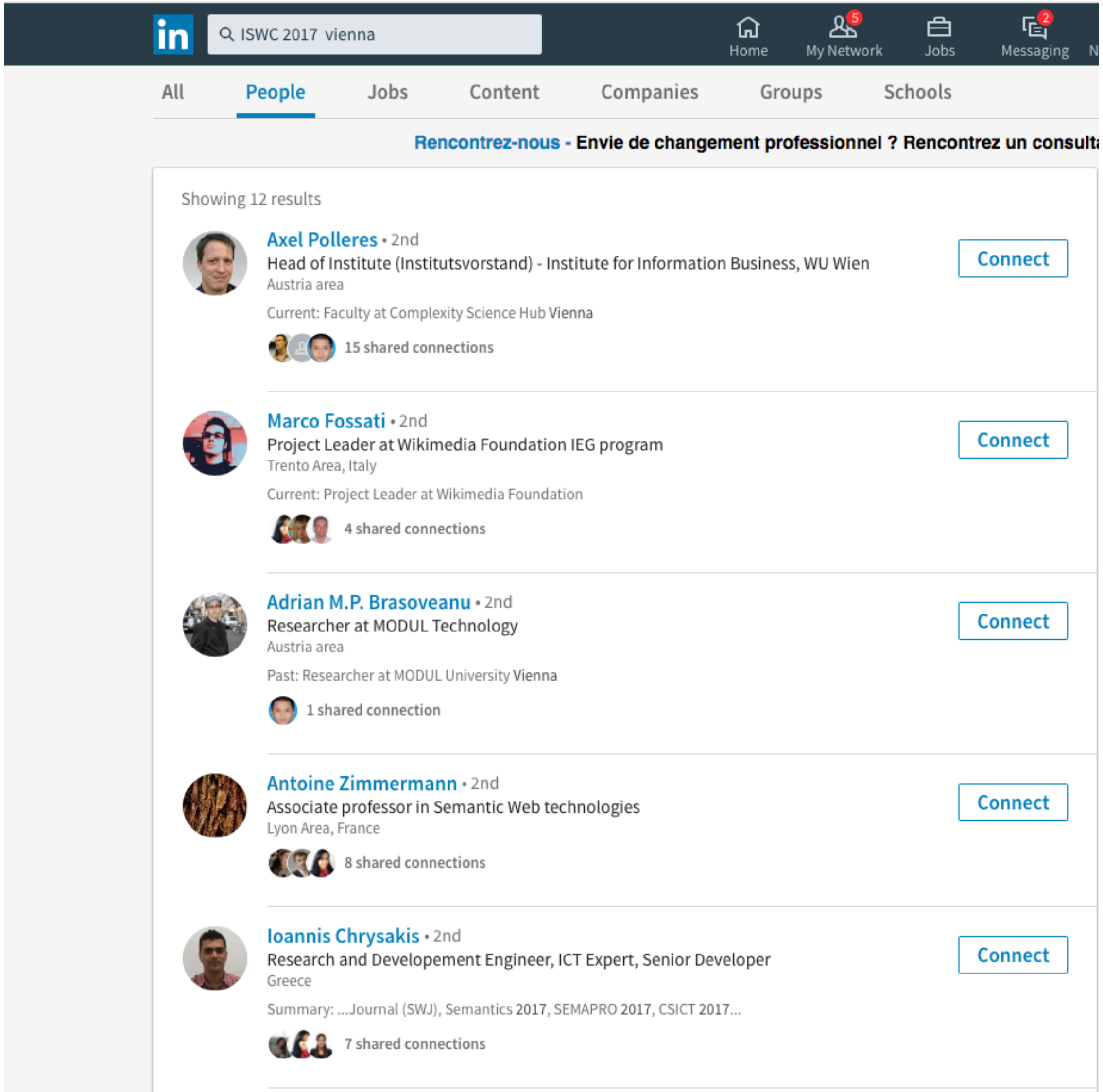
WEB SEARCH WITH KNOWLEDGE GRAPHS



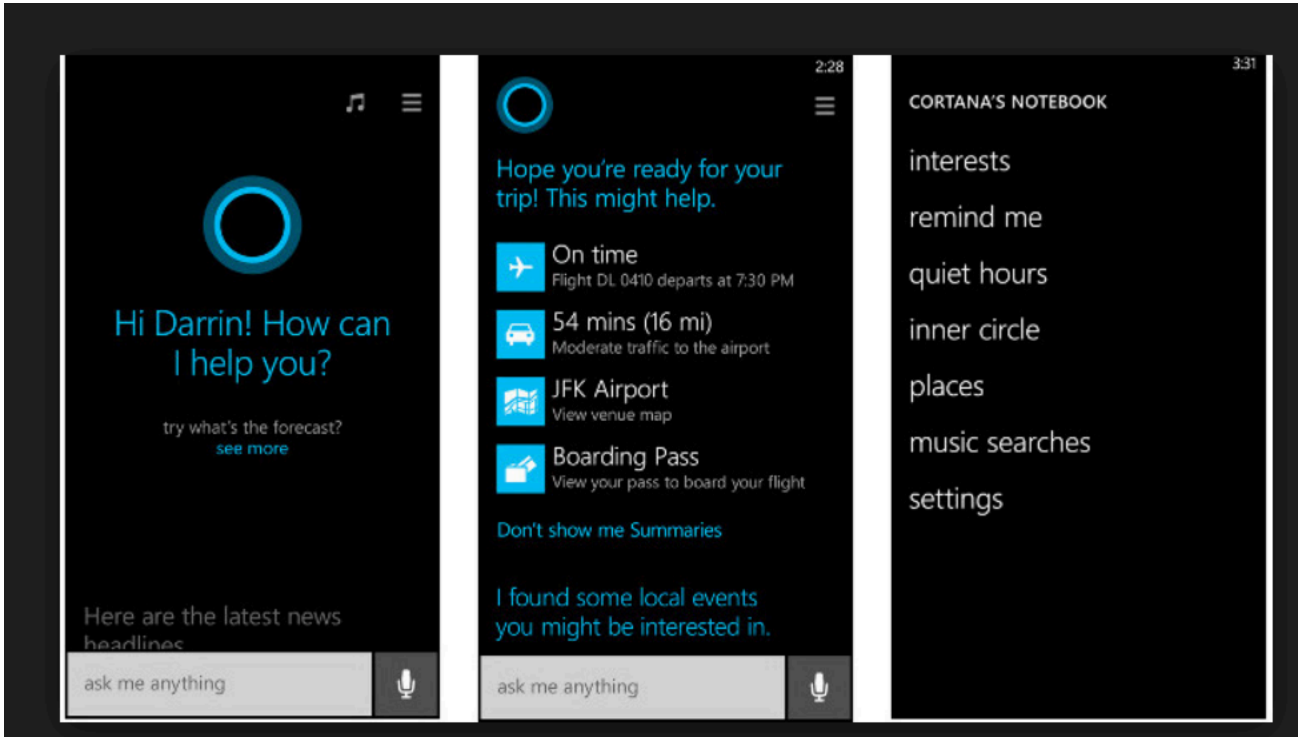
QUESTION ANSWERING WITH KNOWLEDGE GRAPHS



CONNECTING EVENTS AND PEOPLE 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: A DEFINITION …

The **Knowledge Graph** is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. Knowledge Graph display was added to Google's search engine in 2012, starting in the United States, having been announced on May 16, 2012.[1] It uses a graph database to provide structured and detailed information about the topic in addition to a list of links to other sites. The goal is that users would be able to use this information to resolve their query without having to navigate to other sites and assemble the information themselves.[2] The short summary provided in the knowledge graph is often used as a spoken answer in Google Assistant searches.[3]

Wikipedia (en)

This is not a formal definition!

KNOWLEDGE GRAPH: A DEFINITION …

[L. Ehrlinger and W. Wöß, SEMANTiCS'2016]

Definition	Source	
"A knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains."	Paulheim [16]	→ Populated Ontology
"Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities."	Journal of Web Semantics [12]	→ RDF Graph
"Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets."	Semantic Web Company [3]	→ Populated Ontology
"We define a Knowledge Graph as an RDF graph. An RDF graph consists of a set of RDF triples where each RDF triple (s, p, o) is an ordered set of the following RDF terms: a subject $s \in U \cup B$, a predicate $p \in U$, and an object $U \cup B \cup L$. An RDF term is either a URI $u \in U$, a blank node $b \in B$, or a literal $l \in L$."	Färber et al. [7]	→ RDF Graph
"[...] systems exist, [...], which use a variety of techniques to extract new knowledge, in the form of facts, from the web. These facts are interrelated, and hence, recently this extracted knowledge has been referred to as a knowledge graph."	Pujara et al. [17]	→ Extracted RDF Graph

[3] A. Blumauer. From Taxonomies over Ontologies to Knowledge Graphs, July 2014. https://blog.semanticweb.at/2014/07/15/from-taxonomies-over-ontologiesto-knowledge-graphs [August, 2016].

[7] M. Farber, B. Ell, C. Menne, A. Rettinger, and F. Bartscherer. Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web Journal, 2016. http://www.semantic-web-journal. net/content/linked-data-quality-dbpedia-freebaseopencyc-wikidata-and-yago [August, 2016] (revised version, under review).

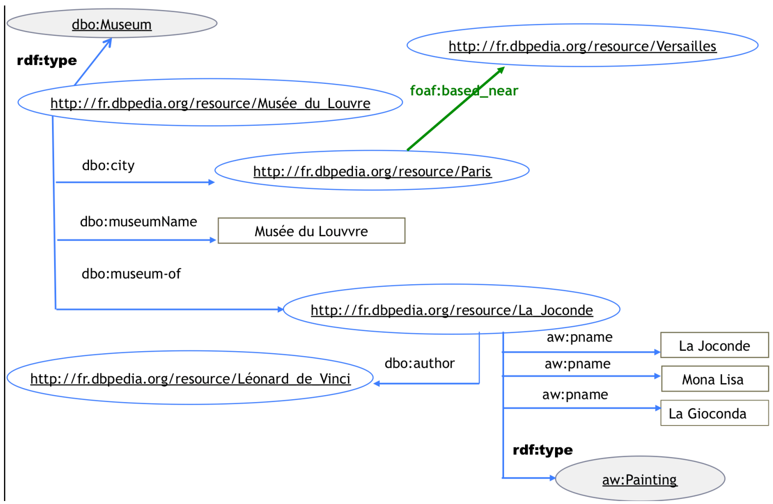
[12] M. Kroetsch and G. Weikum. Journal of Web Semantics: Special Issue on Knowledge Graphs. http://www.websemanticsjournal.org/index.php/ps/announcement/view/19 [August, 2016].

[16] H. Paulheim. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web Journal, (Preprint):1–20, 2016.

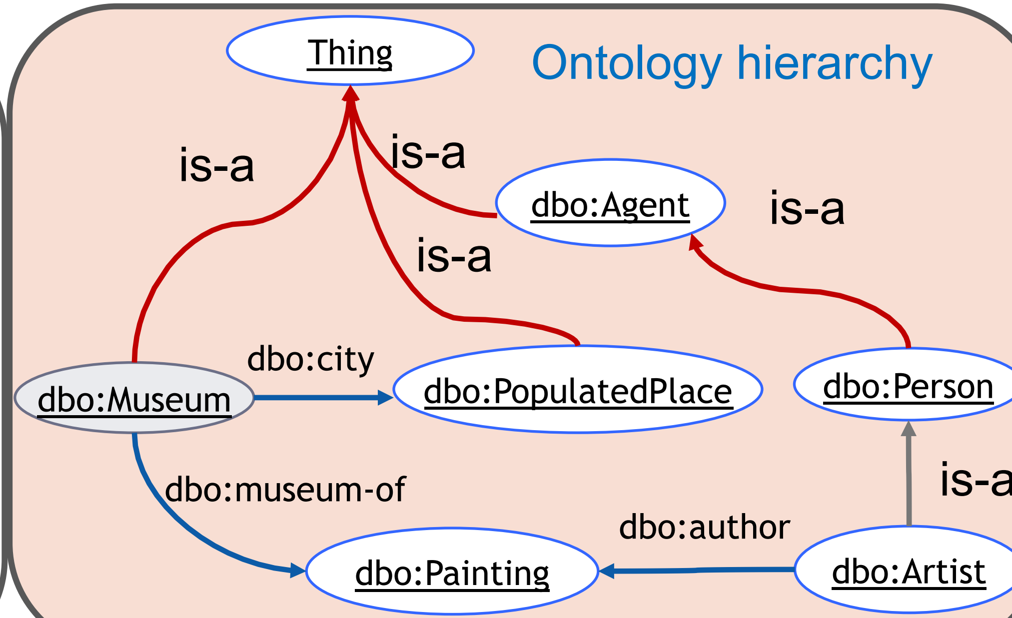
[17] J. Pujara, H. Miao, L. Getoor, and W. Cohen. Knowledge Graph Identification. In Proceedings of the 12th International Semantic Web Conference - Part I, ISWC '13, pages 542–557, New York, USA, 2013. Springer.

KNOWLEDGE GRAPH (KG)

RDF Graphs



Ontology hierarchy



Querying (SPARQL)

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

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

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

Ontology axioms and rules

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

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

KNOWLEDGE GRAPH 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

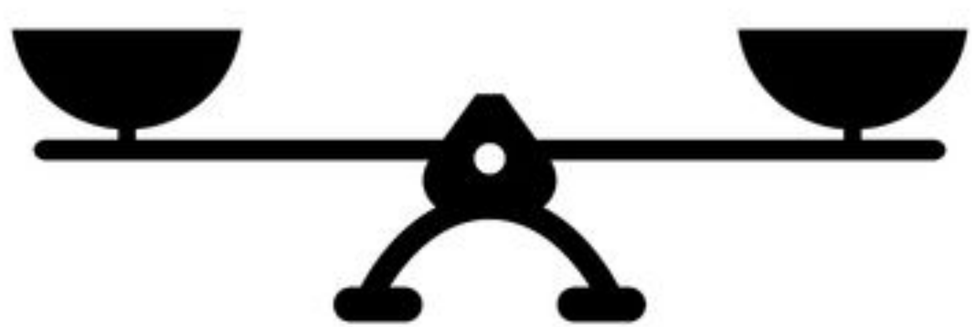
Donald John Trump (born June 14, 1946) is an American businessman, author, television producer, politician, and the Republican Party nominee for President of the United States in the 2016 election. He is the chairman and president of The Trump Organization, which is the principal holding company for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

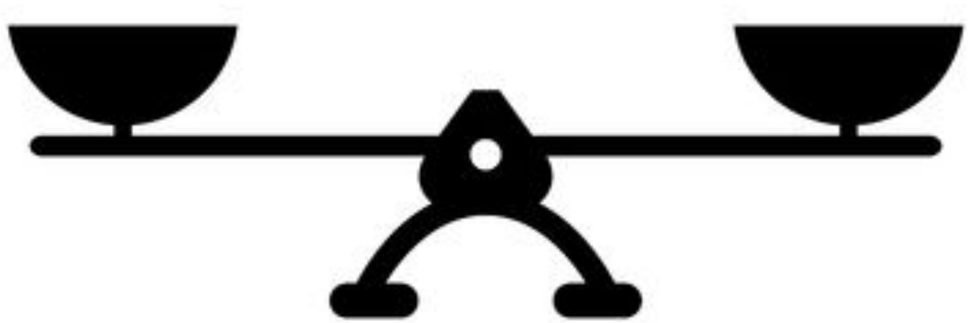
KNOWLEDGE GRAPH REFINEMENT

Completeness Correctness



KNOWLEDGE GRAPH REFINEMENT

Completeness Correctness



Data Linking
Ontology Alignment
Key discovery
Missing values prediction
…

Link Invalidation
Contextual identity
Error detection
…

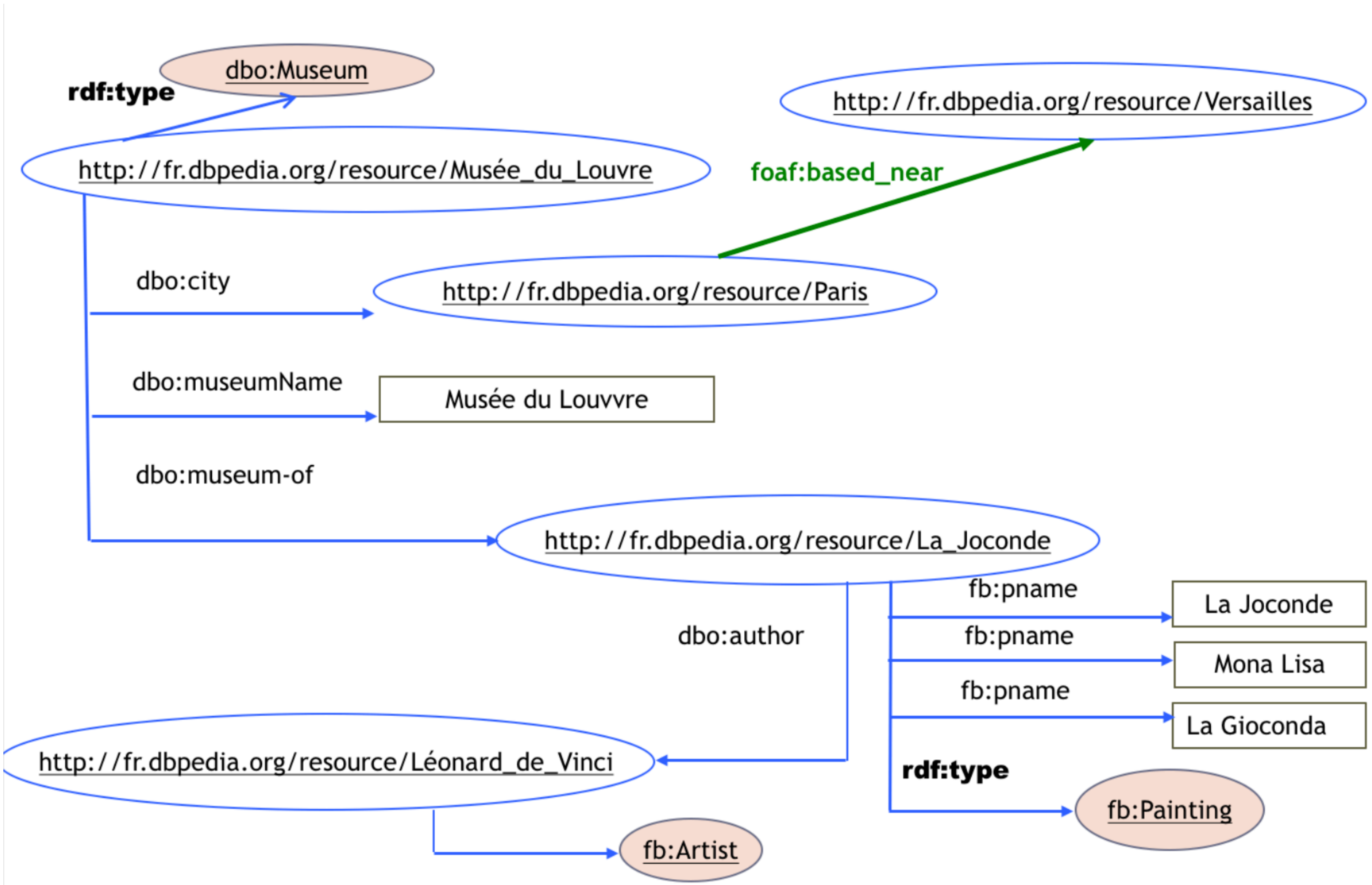
OUTLINE

- **Introduction**
  - Linked Data
  - Knowledge graphs
  - Knowledge graph refinement
- **Data Linking**
- **Identity Problem**
- **Conclusion**

1. DATA LINKING

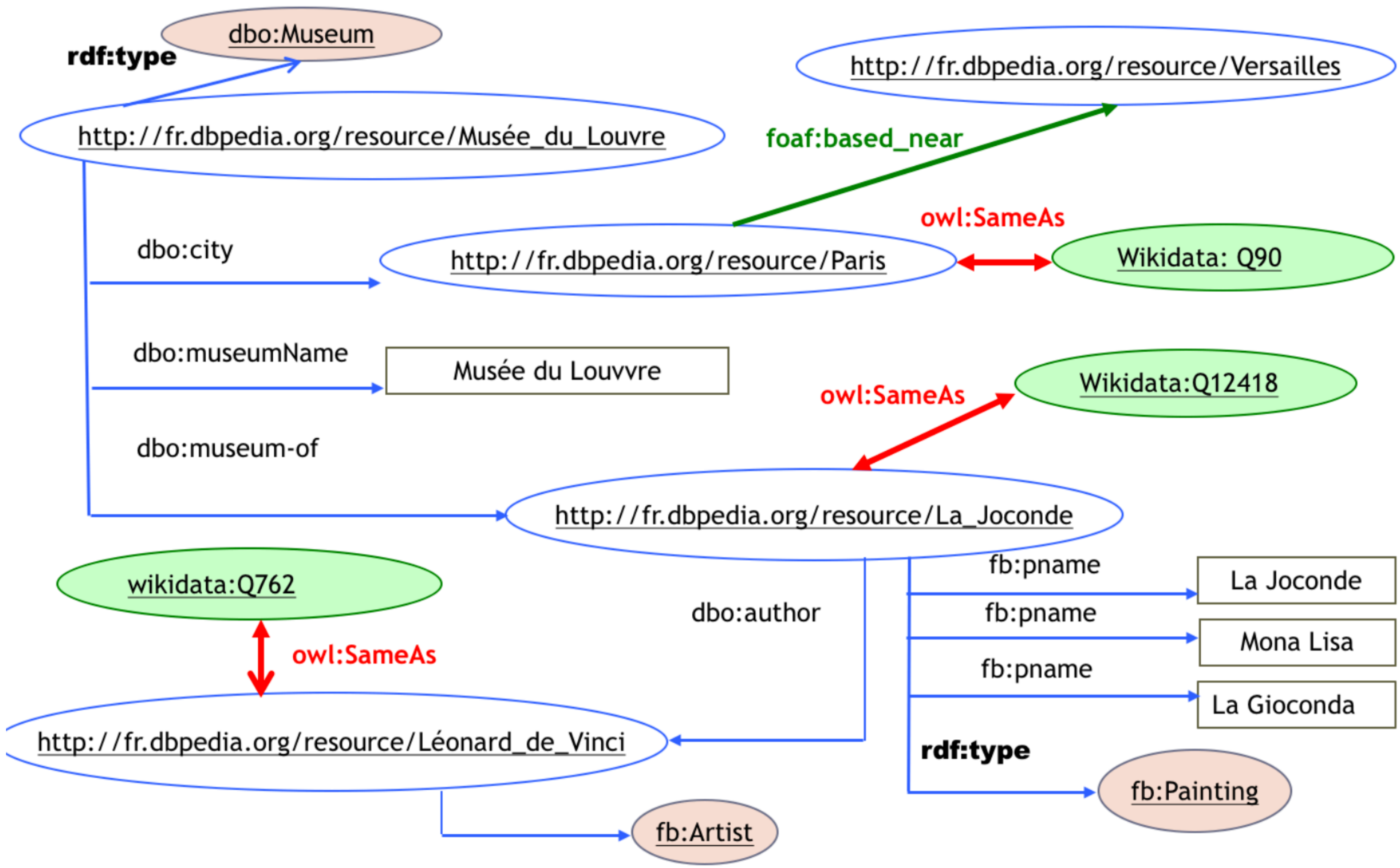
DATA LINKING

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



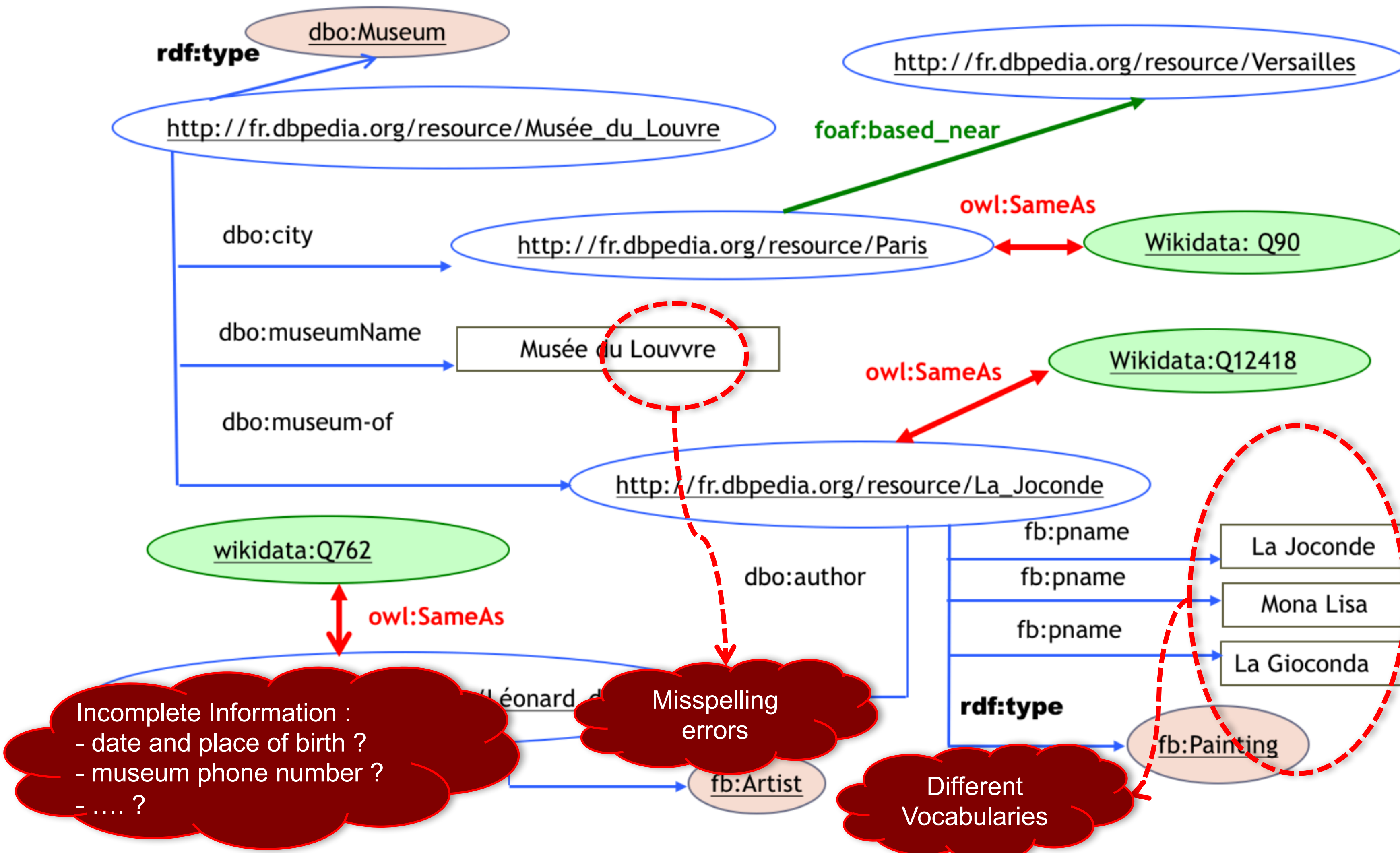
DATA LINKING

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



DATA LINKING: DIFFICULTIES

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



IDENTITY LINK DETECTION PROBLEM

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

- **Definition (Link Discovery)**
  - Given two sets U_1 and U_2 of resources
  - Find a partition of U_1 x U_2 such that :
    - S = {(u1,u2) $\in$ u1 × u2: owl:sameAs(s,t)} and
    - D = {(u1,u2) $\in$ u1 × u2: owl:differentFrom(s,t)}

- A method is said **total** when $(S \cup D) = (U_1 \times U_2)$
- A method is said **partial** when $(S \cup D) \subset (U_1 \times U_2)$
- **Naïve complexity** $\in O(U_1 \times U_2)$, i.e. $O(n^2)$

SOME OF HISTORY …

Problem which exists since the data exists … and under different terminologies: *record linkage, entity resolution, data cleaning, object coreference, duplicate detection, ….*

Automatic Linkage of Vital Records*

[NKAJ, Science 1959]

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

The term *record linkage* has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (*1*). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.

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

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

cord and t be sign ring e of files occurred with frequencies of about 10 percent of all record linkages involving live births and 25 percent of all li

ASIDE: DETECTING IDENTITY LINKS

Ironically "Identity link detection" has many duplicates

- Record linkage
- Duplicate detection
- Coreference resolution
- Reference reconciliation
- Entity resolution
- Fuzzy match
- Object consolidation
- Object identification
- Deduplication
- Entity clustering
- Approximate match
- Identity uncertainty
- Merge/purge
- Household matching
- Hardening soft databases
- Householding
- Reference matching
- Doubles

Lise Getoor, VLDB'12 tutorial

DATA LINKING IS MORE COMPLEX FOR GRAPHS THAN TABLES (WHY?)

	Databases	Semantic Web
Schema/Ontologies	**Same schema**	**Possibly different ontologies in the same dataset**
Multiple types	**Single relation**	**Several classes**
Open World Assumption	**NO**	**YES**
UNA-Unique Name Assumption	**Yes**	**May be no**
Data volume	**XX Thousands**	**XX Millions/Billions** **(e.g., DBpedia has 1.5 billion triples)**
Multiple values for a property	**NO**	**YES** P1 hasAuthor “Michel Chein” P1 hasAuthor “Marie-Christine Rousset”

- Can **propagate** similarity decisions ➔ more **expensive** but **better performance**
- Can be **generic** and use **domain knowledge**, e.g. ontology axioms

DATA LINKING APPROACHES: DIFFERENT CONTEXTS

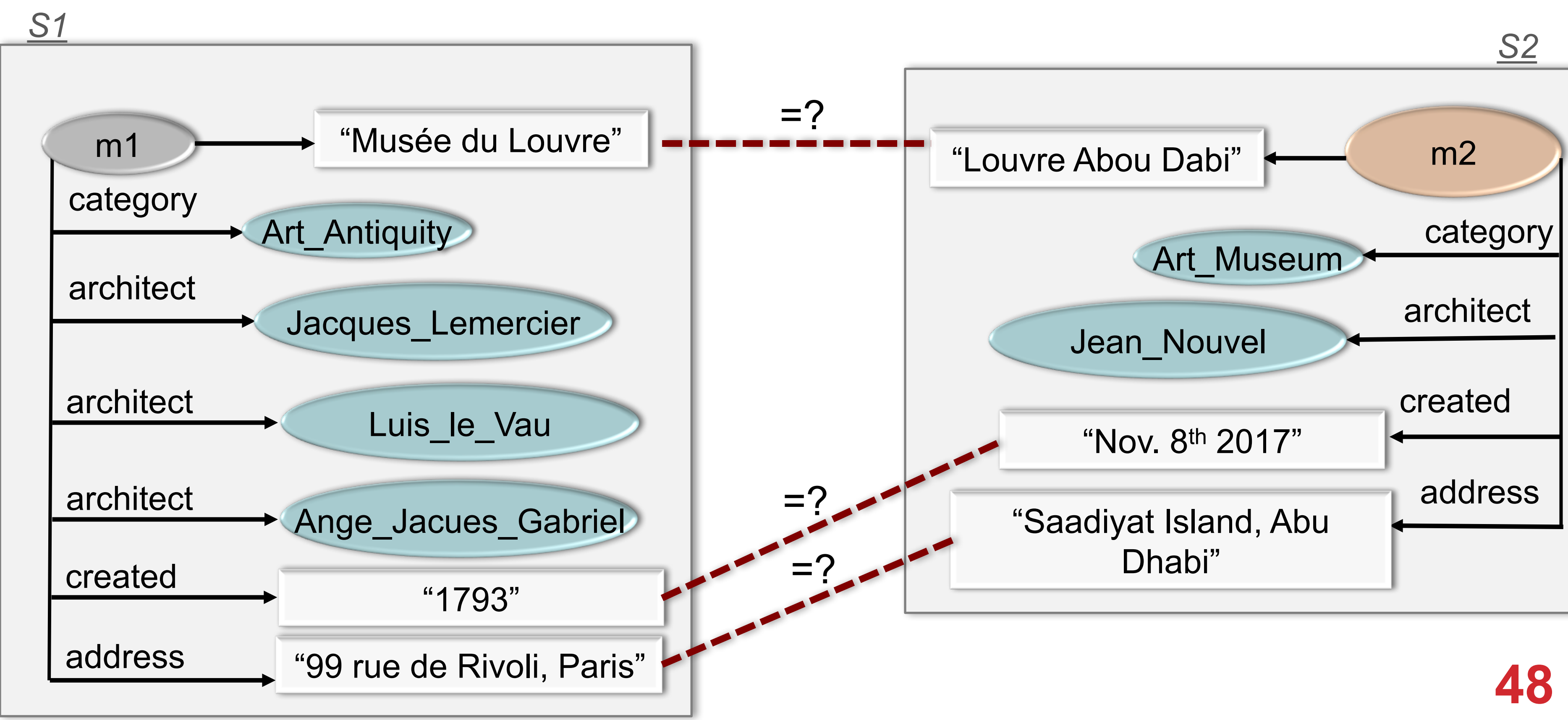
- Datasets conforming to the same ontology
- Datasets conforming to different ontologies
- Datasets without ontologies

DATA LINKING APPROACHES

- **Instance-based approaches**: consider only data type properties (attributes)
- **Graph-based approaches**: consider data type properties (attributes) as well as object properties (relations) to propagate similarity scores/linking decisions (collective data linking)
- **Supervised approaches**: need an expert to build samples of linked data to train models (manual and interactive approaches)
- **Rule-based approaches**: need knowledge to be declared in the ontology or in other format given by an expert

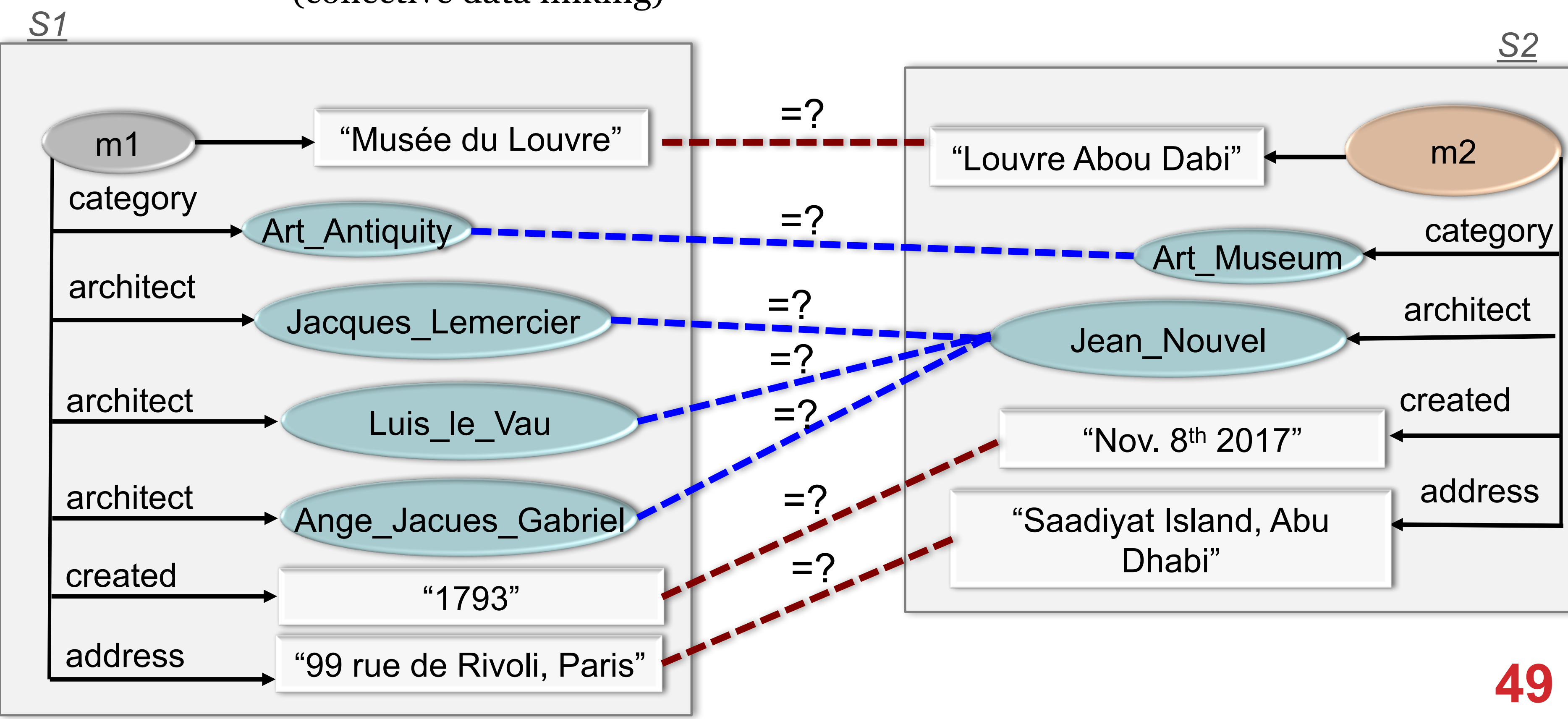
DATA LINKING APPROACHES

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



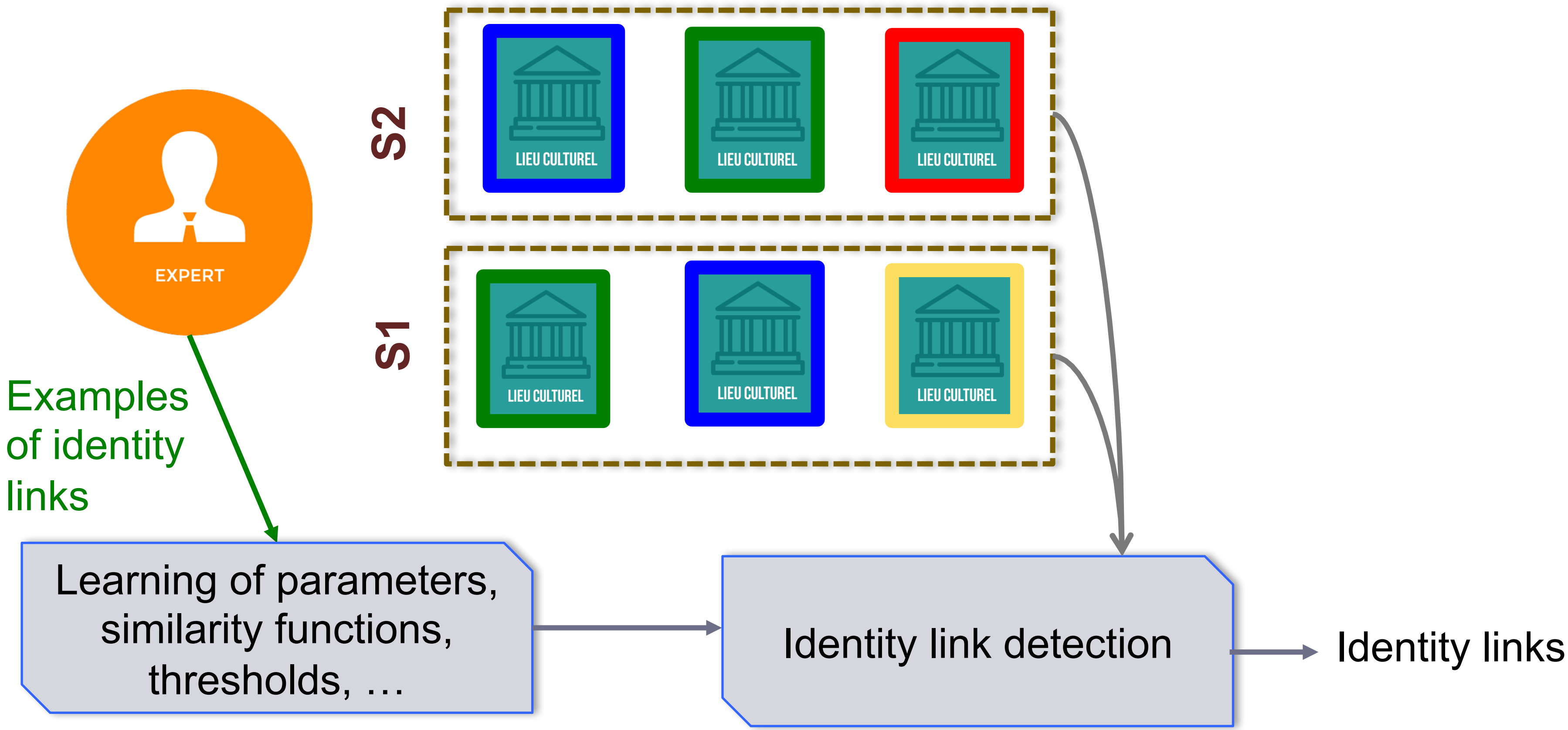
DATA LINKING APPROACHES

- **Graph-based approaches**:
  - consider data type properties (attributes) as well as
  - object properties (relations) to propagate similarity scores/linking decisions (collective data linking)



DATA LINKING APPROACHES

- **Supervised approaches**: need an expert to build samples of identity links to train models (manual and interactive approaches)



DATA LINKING APPROACHES

- **Rule-based approaches: need knowledge to be declared in the ontology or in other format given by an expert**

- **homepage(w1, y) ∧ homepage(w2, y) ➔ sameAs(w1, w2)**
  - *sameAs(Restaurant11, Restaurant21)*
  - *sameAs(Restaurant12, Restaurant22)*
  - *sameAs(Restaurant13, Restaurant23)*

	…	homepage
Restaurant11		www.kitchenbar.com
Restaurant12		www.jardin.fr
Restaurant13		www.gladys.fr
Restaurant14		…

SameAS

SameAS

SameAS

homepage	…	
www.kitchenbar.com		**Restaurant21**
www.jardin.fr		**Restaurant22**
www.gladys.fr		**Restaurant23**
…		**Restaurant24**

DATA LINKING APPROACHES: EVALUATION

- Effectiveness: evaluation of linking results in terms of recall and precision
  - Recall = (#correct-links-sys) /(#correct-links-groundtruth)
  - Precision = (#correct-links-sys) /(#links-sys)
  - F-measure (F1) = (2 x Recall x Precision) / (Recall +Precision)

- Efficiency: in terms of time and space (i.e. minimize the linking search space and the interaction actions with an expert/user).
- Robustness: override errors/mistakes in the data
- Use of benchmarks, like those of OAEI (Ontology Alignment Evaluation Initiative) or Lance

SIMILARITY MEASURES

For more details: William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. **A comparison of string distance metrics for name-matching tasks**. In *Proceedings of the 2003 International Conference on Information Integration on the Web* (IIWEB'03), Subbarao Kambhampati and Craig A. Knoblock (Eds.). AAAI Press 73-78.

SIMILARITY MEASURES

Need of normalization and similarity measures when comparing entities

- Use normalization methods for data property (attribute) values:
  - lemmatization (e.g. canaux → canal),
  - Stop words elimination (e.g. the, this, and, at, …),
  - Enforce common abbreviations (e.g. D&K →Data and Knowledge),
  - Part of ETL tools, commonly using field segmentation and dictionaries.

- Use similarity measures between two values
  - Basic problem: given two property values S and T quantify their 'similarity' in [0..1].
  - Problem challenging for strings

SIMILARITY MEASURES

- **Token based (e.g. Jaccard, TF/IDF cosinus) :**

  The similarty depends on the set of tokens that appear in both S and T.

- **Edit based (e.g. Levenstein, Jaro, Jaro-Winkler) :**

  The similarty depends on the smallest sequence of edit operations which transform S into T.

- **Hybrids (e.g. N-Grams, Jaro-Winkler/TF-IDF, Soundex)**

SIMILARITY MEASURES: TOKEN BASED

- **Jaccard measure:** **Jaccard(S,T) = |S∩T| / |S∪T|**

  Jaccard(« rue de la vieille pierre », « 11 rue vieille pierre ») =3/6

- **Cosinus (based on TF-IDF)**

  Widely used in traditional information retrieval (IR) approaches

  - Intuition: a term that is rare in the data is important and a term that is frequent in the string (value) is important.
  - Term frequency(TF): # of times a 'term' appears in the string compared with the size of the string.
  - Document frequency (IDF): the inverse of (# strings that contain the 'term'/ # of strings in the corpus)

SIMILARITY MEASURES: TOKEN BASED

Cosinus computation based on TF-IDF

- Compute for each value the set of terms represented in a vector of terms
- Compute for each term its weight TF-IDF:

$$V(w,S) = V'(w,S) / \sqrt{\sum_{w'} V'(w',S)^2}$$

With $V'(w, S) = \log(TF_{w,S} + 1).\log(IDFw)$

Let s, t be two values, S, T the sets of terms resp. and

V(w, S), V(w, T) the weights of the term w in S and T, resp.

$$Cosinus(s,t) = \sum_{w \in S \cap T} V(w,S) * V(w,T)$$

Example :

Low weights for "Corporation", high weights for "AT&T", "IBM"

Cosinus("AT&T", "AT&T Corporation") high

Cosinus("AT&T Corporation", "IBM Corporation") Low

SIMILARITY MEASURES: TOKEN BASED

Advantages:

- Efficient computation
- Word order is not significant

Disadvantages:

- Sensitive to spelling errors (Fathia, Sais)
- Sensitive to abbreviations (Univ. vs University)
- Sometimes order in words is meaningful (*Laurent Simon* vs *Simon Laurent*)

SIMILARITY MEASURES: EDIT BASED

- **Edit-based measure: "Levenstein" distance**
  - Character operations:
    - I (Insert), D(delete), R(replace), S (substitution).
  - Unit costs

  - Given two strings s,t edit(s, t):
    - Minimum cost sequence of operations to transform *s to t*.
    - Example: edit('Error', 'Eror')=1, edit('great', 'grate')=2

SIMILARITY MEASURES: EDIT BASED

- Levenstein(“William Cohen”, “Willliam Cohon”)

s	W	I	L	L	I	A	M	_	C	O	H	E		N
t	W	I	L	L	L	I	A	M	_	C	O	H	O	N
op	C	C	C	C	**I**	C	C	C	C	C	C	C	**S**	C
cost	0	0	0	0	1	1	1	1	1	1	1	1	2	2

matching

SIMILARITY MEASURES: EDIT BASED

- **Jaro**
  - For (S, T), the character *c* is common for (S, T):

    if (S_i=c), (T_j=c), and |i-j| < min(|S|,|T|) / 2.
  - The character *c and d* are ***transpositions*** if *c and d* are common for S and T and appear in different orders in S and T.

$$Jaro(S,T) = \frac{1}{3}\left(\frac{m}{|S|} + \frac{m}{|T|} + \frac{m-t}{m}\right)$$

  - Example: Jaro(Texas, Texhas) $= \frac{1}{3}\left(\frac{5}{5} + \frac{5}{6} + \frac{5-2}{5}\right)$ = 0,81

SIMILARITY MEASURES: EDIT BASED

Jaro-Winkler

- An extension of Jaro by considering the size of the longest prefix between S and T.

$$Jaro-Winkler(S,T) = Jaro(S,T) + \left(\frac{\max(P,4)}{10} * (1 - Jaro(S,T) \right)$$

- Example : Jaro-Winkler(Texas, Texhas) $= 0{,}81 + \left(\frac{4}{10} * (1 - 0{,}81) \right)$

  $= 0{,}88$

- Runtime efficiency
- Showed to be relevant for the comparison of person names [Cohen03].

SIMILARITY MEASURES: EDIT BASED

Advantages:

- Robustness when spelling errors exist
- Word order is significant

Disadvantages:

- High runtime
- Sometimes order in words is not meaningful (Univ. Paris Saclay and Paris Saclay University)

INSTANCE-BASED DATA LINKING APPROACHES

FRAMEWORK SILK [Volz et al'09]

- Provides a Link Specification Language(LSL)
- Allows specifying linking conditions between two datasets

- The linking conditions may be expressed in terms of:
  - Elementary similarity measures (e.g., Jaccard, Jaro) and
  - Aggregation functions (e.g. max, average) of the similarity scores

SIMILARITY MEASURES IN SILK

[Volz et al'09]

Metric	Description
jaroSimilarity	String similarity based on Jaro distance metric
jaroWinklerSimilarity	String similarity based on Jaro-Winkler metric
qGramSimilarity	String similarity based on q-grams
stringEquality	Returns 1 when strings are equal, 0 otherwise
numSimilarity	Percentual numeric similarity
dateSimilarity	Similarity between two date values
uriEquality	Returns 1 if two URIs are equal, 0 otherwise
taxonomicSimilarity	Metric based on the taxonomic distance of two concepts

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<Silk>
  <Prefixes>
    <Prefix id="rdfs" namespace="http://www.w3.org/2000/01/rdf-schema#" />
    <Prefix id="dbpedia" namespace="http://dbpedia.org/ontology/" />
    <Prefix id="gn" namespace="http://www.geonames.org/ontology#" />
  </Prefixes>

  <DataSources>
    <DataSource id="dbpedia">
      <Param name="endpointURI" value="http://demo_sparql_server1/sparql" />
      <Param name="graph" value="http://dbpedia.org" />
    </DataSource>

    <DataSource id="geonames">
      <Param name="endpointURI" value="http://demo_sparql_server2/sparql" />
      <Param name="graph" value="http://sws.geonames.org/" />
    </DataSource>
  </DataSources>
```

Prefixes

SPARQL endpoints

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<Interlinks>
    <Interlink id="cities">
      <LinkType>owl:sameAs</LinkType>
      <SourceDataset dataSource="dbpedia" var="a">
        <RestrictTo>
          ?a rdf:type dbpedia:City
        </RestrictTo>
      </SourceDataset>
      <TargetDataset dataSource="geonames" var="b">
        <RestrictTo>
          ?b rdf:type gn:P
        </RestrictTo>
      </TargetDataset>
```

Link types

Entities to be linked

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<LinkageRule>
        <Aggregate type="average">
          <Compare metric="levenshteinDistance" threshold="1">
            <Input path="?a/rdfs:label" />
            <Input path="?b/gn:name" />
          </Compare>
          <Compare metric="num" threshold="1000" >
            <Input path="?a/dbpedia:populationTotal" />
            <Input path="?b/gn:population" />
          </Compare>
        </Aggregate>
      </LinkageRule>

      <Filter limit="1" />
```

Aggregation function

Similarity measures

EXAMPLE OF LSL SPECIFICATION

[Volz et al'09]

```
<Outputs>
        <Output type="file" minConfidence="0.95">
          <Param name="file" value="accepted_links.nt" />
          <Param name="format" value="ntriples" />
        </Output>
        <Output type="file" maxConfidence="0.95">
          <Param name="file" value="verify_links.nt" />
          <Param name="format" value="alignment" />
        </Output>
      </Outputs>
    </Interlink>
  </Interlinks>

</Silk>
```

Linking threshold

Possible links

KNOFUSS (INSTANCE-BASED, UNSUPERVISED)

[Nikolov et al'12]

- Learns linking rules using genetic algorithms:

$$Sim(i1, i2) = f_{ag}(w_{11}sim_{11}(V11,V21), \ldots w_{mn}sim_{mn}(V1m,V2n))$$

  - F_{ag} : aggregation function for the similarity scores
  - sim_{ij}: similarity measure between values V1i and V2j
  - w_{ij}: weights in [0..1]

- **Assumptions:**
  - Unique name assumption (UNA), i.e., two different URIs refer to two different entities.
  - Good coverage rate between the two datasets
  - Normalized similarity scores in [0..1]

KNOFUSS (INSTANCE-BASED, UNSUPERVISED)

[Nikolov et al'12]

Test case	Similarity function	Threshold
Person1	max(tokenized-jaro-winkler(soc_sec_id;soc_sec_id); monge-elkan(phone_number;phone_number))	≥0.87
Person2	max(jaro(phone_number;phone_number); jaro-winkler(soc_sec_id;soc_sec_id))	≥0.88
Restaurants (OAEI)	avg(0.22*tokenized-smith-waterman(phone_number;phone_number); 0.78*tokenized-smith-waterman(name;name))	≥0.91
Restaurants (fixed)	avg(0.35*tokenized-monge-elkan(phone_number;phone_number); 0.65*tokenized-smith-waterman(name;name))	≥0.88

Examples of linking rules learned on the OAEI'10 benchmark

Dataset	KnoFuss+GA	ObjectCoref	ASMOV	CODI	LN2R	RiMOM	FBEM
Person1	**1.00**	**1.00**	**1.00**	0.91	**1.00**	**1.00**	N/A
Person2	**0.99**	0.95	0.35	0.36	0.94	0.97	0.79
Restaurant (OAEI)	0.78	0.73	0.70	0.72	0.75	**0.81**	N/A
Restaurant (fixed)	**0.98**	0.89	N/A	N/A	N/A	N/A	0.96

Results in term of F-Measure on OAEI'10

LN2R: A Logical and Numerical Method for Reference Reconciliaton

[Saïs et al' 07, Saïs et al'09]

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

- A combination of two methods:
  - L2R, a Logical method for reference reconciliation: applies logical rules to infer sure owl:sameAs and owl:differentFrom links
  - N2R, a Numerical method for reference reconciliation: computes similarity scores for each pair of references
- **Assumptions**
  - The datasets are conforming to the same ontology
  - The ontology contains axioms

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

Ontology axioms

- Disjunction axioms between classes, DISJOINT(C, D)
- Functional properties axioms, PF(P)
- Inverse functional properties axioms, PFI(P)
- A set of properties that is functional or inverse functional axioms

Assumptions on the data

- Unique Name Assumption, UNA(src1)
- Local Unique Name Assumption, LUNA(R)

Example:

Authored(p, a1), Authored(p, a2), Authored(p, a3) …., Authored(p, an)
➔ (a1 $\neq$a2), (a1 $\neq$ a3), (a2 $\neq$ a3) , …

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

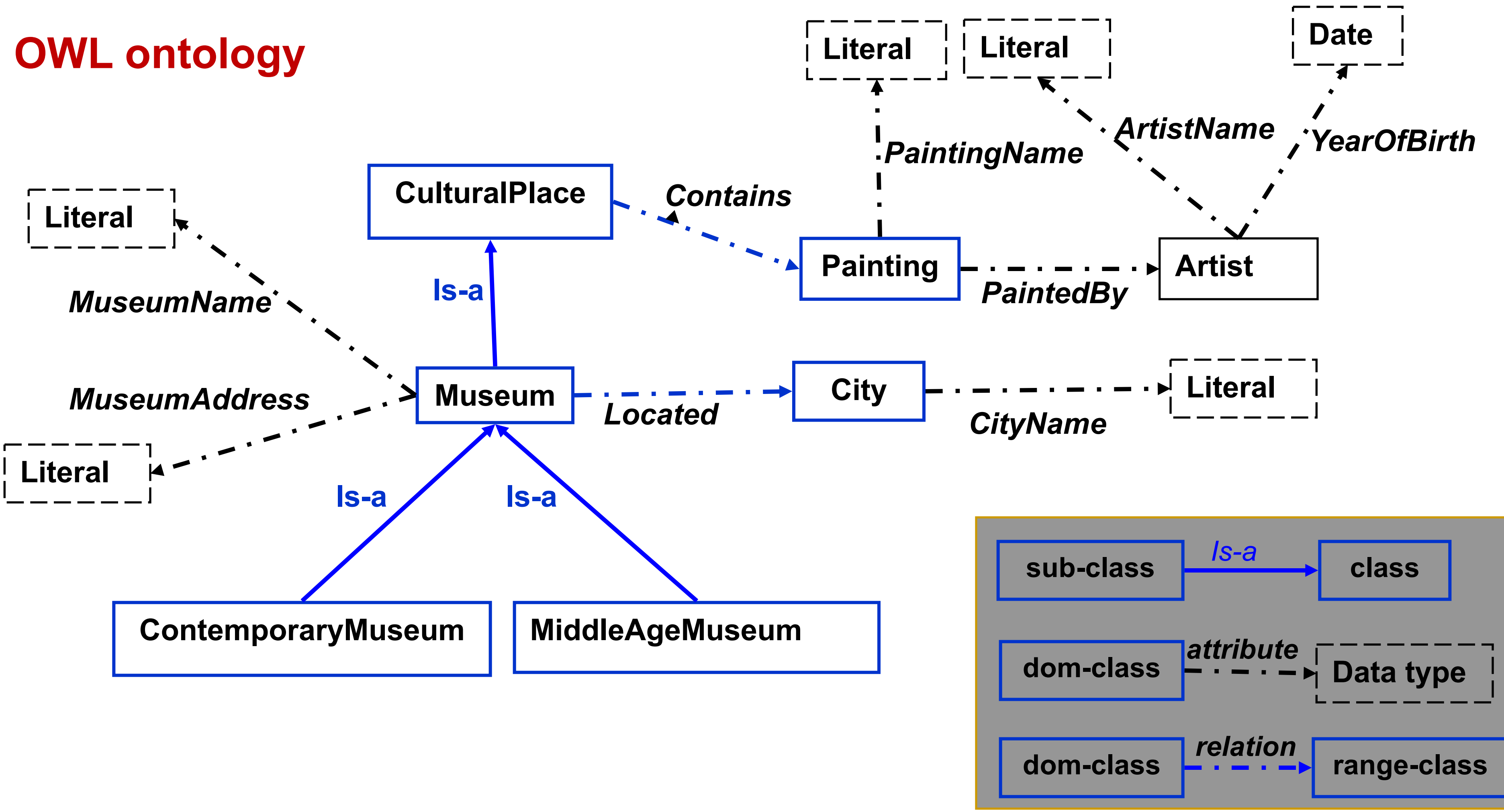
[Saïs et al' 07, Saïs et al'09]

- A combination of two methods:
  - L2R, a Logical method for reference reconciliation: applies logical rules to infer sure owl:sameAs and owl:differentFrom links
  - N2R, a Numerical method for reference reconciliation: computes similarity scores for each pair of references
- **Assumptions**
  - The datasets are conforming to the same ontology
  - The ontology contains axioms

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

OWL ontology

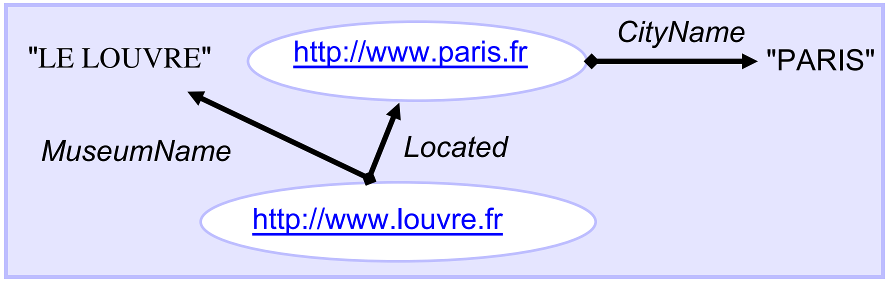


LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

RDF datasets

- **RDF Graphs:**



- **RDF Facts:**

Desc(http://www.louvre.fr)= {

Museum(http://www.louvre.fr),
Located(http://www.louvre.fr,http://www.paris.fr),
MuseumName(http://www.louvre.fr,"LE LOUVRE")}

Desc(http://www.paris.fr)= {

Located(http://www.louvre.fr,http://www.paris.fr),
CityName(http://www.paris.fr,"PARIS")}

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

Ontology axioms:

- Disjunction axioms between classes, DISJOINT(C, D)
- Functional properties axioms, PF(P)
- Inverse functional properties axioms, PFI(P)
- A set of properties that is functional or inverse functional axioms

Assumptions on the data

- Unique Name Assumption, UNA(src1)
- Local Unique Name Assumption, LUNA(R)

Example:

Authored(p, a1), Authored(p, a2), Authored(p, a3) …., Authored(p, an)
➔ (a1 $\neq$a2), (a1 $\neq$ a3), (a2 $\neq$ a3) , …

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

- Disjunction axioms between classes DISJOINT(C, D), its logical semantics:

$$\forall X \quad C(X) \Rightarrow \neg D(X)$$

- Functional properties axioms, PF(P), its logical semantics:

$$\forall X, Y, Z \quad P(X,Y) \wedge P(X, Z) \Rightarrow Y=Z$$

- Inverse functional properties axioms, its logical semantics:

$$\forall X, Y, Z \quad P(Y,X) \wedge P(Z, X) \Rightarrow Y=Z$$

LN2R
(GRAPH BASED, UNSUPERVISED AND INFORMED)

[Saïs et al' 07, Saïs et al'09]

SWRL rules are used to generalize:

- Functionality axioms to a set of properties (relations and attributes) *{P1,…, Pn}*, PF(*P1,…, Pn)*, *its logical semantics:*

$$\forall X_1,\dots,X_n,Y,Z \bigwedge_{\forall i \in [1..n]} (P_i(X_i,Y) \wedge P_i(X_i,Z)) \Rightarrow Y=Z$$

- Inverse functionality axioms to a set of properties (relations and attributes)*{P1,…, Pn}*, PF(*P1,…, Pn)*, *its logical semantics:*

$$\forall X_1,\dots,X_n,Y,Z \bigwedge_{\forall i \in [1..n]} (P_i(Y,X_i) \wedge P_i(Z,X_i)) \Rightarrow Y=Z$$

L2R: A LOGICAL METHOD FOR REFERENCE RECONCILIATION

L2R: AUTOMATIC GENERATION OF INFERENCE RULES

Translation of UNA(src1)

R1:src1(X) $\wedge$ src1(Y) $\wedge$ (X $\neq$ Y) $\Rightarrow$ $\neg$Reconcile(X,Y) ; ...

Translation of LUNA(R)

R11(R) : R(Z, X) $\wedge$ R(Z, Y) $\wedge$ (X $\neq$ Y) $\Rightarrow$ $\neg$Reconcile(X,Y) ; ...

Translation of DISJOINT(C, D):

R5(C, D) : C(X) $\wedge$ D(Y) $\Rightarrow$ $\neg$ Reconcile (X, Y)

Translation of PF(R):

R6.1(R): Reconcile(X, Y) $\wedge$ R(X, Z) $\wedge$ R(Y, W) $\Rightarrow$ Reconcile (Z, W)

R6.1(Located): Reconcile(X, Y) $\wedge$ Located (X, Z)$\wedge$Located (Y, W) $\Rightarrow$ Reconcile (Z, W)

Translation of PF(A):

R6.2(A): Reconcile(X, Y) $\wedge$ A(X, Z) $\wedge$ A(Y, W) $\Rightarrow$SynVals(Z, W)

R6.2(MuseumName):Reconcile(X,Y) $\wedge$ MuseumName (X, Z) $\wedge$ MuseumName (Y,W) $\Rightarrow$SynVals(Z, W)

L2R: INFERENCE ALGORITHM

- Apply until saturation the resolution principle [Robinson'65], by following the unit strategy

Resolution rule : $\frac{C_1 : (L_1), C_2 : (L_2 \vee C)}{C_{1,2} : (C_\sigma)}$ Avec $L_{1\sigma} = \neg L_{2\sigma}$

- R ∪ F: Horn clauses without functions, where :
  - R: rules in the form of horn clauses
  - F: unit clauses fully instantiated,
    - Reference descriptions: RDF facts (class-facts, relation-facts and attribute-facts).
    - Facts that express the reference origin: src1(i) and src2(j)
    - Facts that express the synonymy and not synonymy between values: SynVals(v1, v2) or ¬ SynVals(v1, v2)
- Computation of the set SatUnit(R ∪ F)

L2R: ALGORITHM PROPERTIES

- Termination of the algorithm: guaranteed thanks to the absence of function symbols in the knowledge base
- Completeness: for the deduction of all the unit clauses fully instantiated, *Reconcile* and *SynVals*.

Theorem : *Let R be a set of un Horn clauses without functions. Let F be a set of unit clauses fully instantiated. If $R \cup F$ is satisfiable, then:*

$$\forall\, p(\mathbf{a}),\quad (R \cup F \models p(\mathbf{a})) \Rightarrow (p(\mathbf{a}) \in SatUnit(R \cup F))$$

With $p(\mathbf{a})$, a unit clause fully instantiated and $SatUnit(R \cup F)$ is the set of inferred clauses by applying the unit resolution until saturation on $R \cup F$.

L2R: EXAMPLE OF AXIOMS

Disjunction : {DISJOINT(MiddleAgeMuseum,ContemporaryMuseum), DISJOINT(Painting, Artist), DISJOINT(CulturalPlace, City), DISJOINT(CulturalPlace,Painting)}.

Functional properties: {PF(Located), PF(PaintedBy), PF(ArtistName), PF(YearOfBirth), PF(PaintingName), PF(CityName), PF(MuseumName), PF(MuseumAddress)}.

Inverse functional properties:

{PFI(PaintingName, PaintedBy), PFI(Contains), PFI(ArtistName), PFI(MuseumName), PFI(MuseumAddress), PFI(CityName)}.

L2R: EXAMPLE OF DATASETS

S1

CulturalPlace(S1_m1); Museum(S1_m2);
MiddeleAgeMuseum(S1_m3), Painting(S1_p1);
Painting(S1_p2); Painting(S1_p3) Artist(S1_a1);
Artist(S1_a2); City(S1_c1);
MuseumName(S1_m1,"musee du LOUVRE");
Contains(S1_m1,S1_p1);
MuseumName(S1_m2,"musee des arts premiers");

MuseumAddress(S1_m2, "quai branly");
Located(S1_m2,S1_c1); CityName(S1_c1,"Paris");
PaintingName(S1_p1, "La Joconde");
PaintedBy(S1_p1,S1_a1);

ArtistName(S1_a1, "Leonard De Vinci");

PaintingName(S1_p2,"La Cene");

PaintedBy(S1_p2, S1_a1);

S2

Museum(S2_m1); Museum(S2_m2);
Painting(S2_p1); ContemporaryMuseum(S2_m4)
Painting(S2_p2);Painting(S2_p3); Artist(S2_a1);
City(S2_c1); MuseumName(S2_m1,"Le LOUVRE");
Located(S2_m1,S2_c1); Contains(S2_m1,S2_p2);
Contains(S2_m1, S2_p1);
MuseumName(S2_m2,"Musée du quai Branly");
MuseumAddress(S2_m2, "37 quai branly, portail Debilly"); Contains(S2_m1,S2_p3);
Located(S2_m2,S2_c1);
CityName(S2_c1, "Ville de paris");
PaintingName(S2_p2, "Vierge aux rochers");
PaintedBy(S2_p2,S2_a1);
ArtistName(S2_a1,"De Vinci");
PaintingName(S2_p3, "Sainte Anne, la vierge et l'enfant jesus"); PaintingName(S2_p1, "la Joconde");

The UNA is stated in the two sources S1 and S2.

L2R: RUNNING EXAMPLE DE

Instantiated rules

R1, R2

R5(CulturalPlace, Painting)
R5(Artist, Painting)
R5(MiddleAgeMuseum, ContemporaryMuseum)
…

Fact set

scr1(S1_m2), scr1(S1_p1), scr1(S1_p2), scr2(S2_m1),
scr2(S2_p1), scr2(S2_p2),
CulturalPlace(S1_m1), Painting(S2_p1)
Artist(S1_a1), Painting(S2_p2)
MiddeleAgeMuseum(S1_m3),ContemporaryMuseum(S2_m4)
…

REC

NREC

¬Reconcile(S1_m1,S1_m2), ¬Reconcile(S1_p1,S1_p2),
¬Reconcile(S2_m1,S2_p1), ¬Reconcile(S2_p1, S2_p2)
¬Reconcile(S1_m1, S2_p1),
¬Reconcile(S1_a1, S2_p1)
¬Reconcile(S1_m3, S2_m4)

SynVals("La Joconde"," la joconde")

L2R: RUNNING EXAMPLE DE

Instantiated rules

…
R7.2 (PaintingName)

Fact set

…
PaintingName(S1_p1,"La joconde"),
PaintingName(S2_p1," La Joconde")

REC

Reconcile(S2_p1, S1_p1)

NREC

¬Reconcile(S1_m1,S1_m2), ¬Reconcile(S1_p1,S1_p2),

¬Reconcile(S2_m1,S2_p1), ¬Reconcile(S2_p1, S2_p2)

¬Reconcile(S1_m1, S2_p1),
¬Reconcile(S1_a1, S2_p1)
¬Reconcile(S1_m3, S2_m4)

SynVals("La Joconde"," la joconde")

L2R: RUNNING EXAMPLE DE

Instantiated rules

…
R7.1(Contains)
R4. “UNA”
R6.2(MuseumName)
R6.1(Located),
R6.2(CityName)

Fact set

…
Contains(S1_m1, S1_p1), Contains(S2_m1, S2_p1)
src1(S1 m1), src2 (S2 m1), scr2 (S2 m2),
MuseumName(S1 m1, ‘musee du LOUV RE”)
MuseumName(S2 m1, “LE LOUV RE”)
Located(S1 m1, S1 c1), Located(S2 m1, S2 c1)

REC

Reconcile(S2_p1, S1_p1)
Reconcile(S1_m1, S2_m1)
Reconcile(S1_c1, S2_c1)

SynVals(“La Joconde”,” la joconde”)
SynVals(“musee du LOUVRE”, “LE LOUVRE”)
SynVals(”ville de Paris”,”Paris”)

NREC

¬Reconcile(S1_m1,S1_m2), ¬Reconcile(S1_p1,S1_p2),
¬Reconcile(S2_m1,S2_p1), ¬Reconcile(S2_p1, S2_p2)
¬Reconcile(S1_m1, S2_p1),
¬Reconcile(S1_a1, S2_p1)
¬Reconcile(S1_m3, S2_m4)
¬Reconcile(S2_m2, S1_m1)

L2R EXPERIMENTS

TWO DATASETS: ON TOURISM AND SCIENTIFIC PUBLICATIONS DOMAINS

FT_HOTELS (data of Mappy)

- A set of seven data sources where UNA is fulfilled:data linking problem for 21 pairs of data sources
- The sources contain in total 28 934 references that describe hotels in Europe.

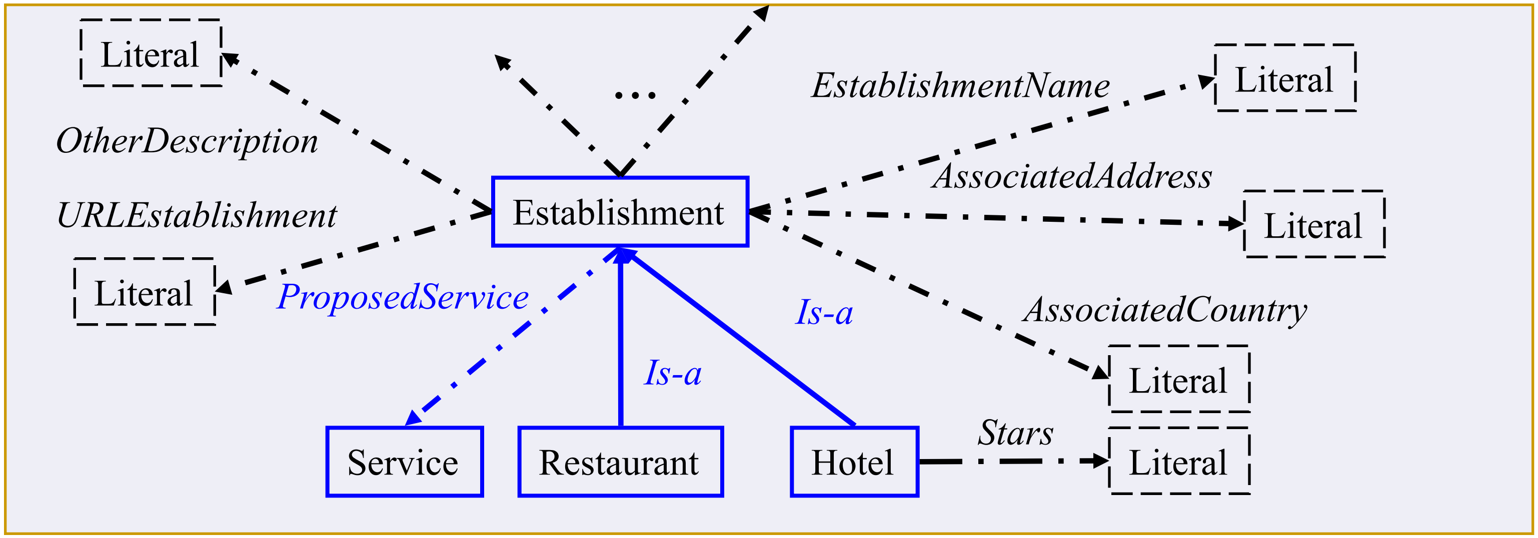
➔Integration of different data sources problem

Cora (a benchmark)

- A collection (in RDF) of 1295 paper citations of 112 different research, 1292 conferences and 3521 authors.
- UNA is not fulfilled.

➔ Data cleaning problem

L2R EXPERIMENTS: ONTOLOGY FOR FT_HOTELS



- ✓ DISJOINT(*Hotel*, *Service*)
- ✓ All the properties are functional (PF), except *OtherService*, *OtherDescription*
- ✓ One inverse funtional axiom that combines two attributes
  PFI(*EstablishmentName,AssociatedAddress)*
- ✓ UNA is declared.

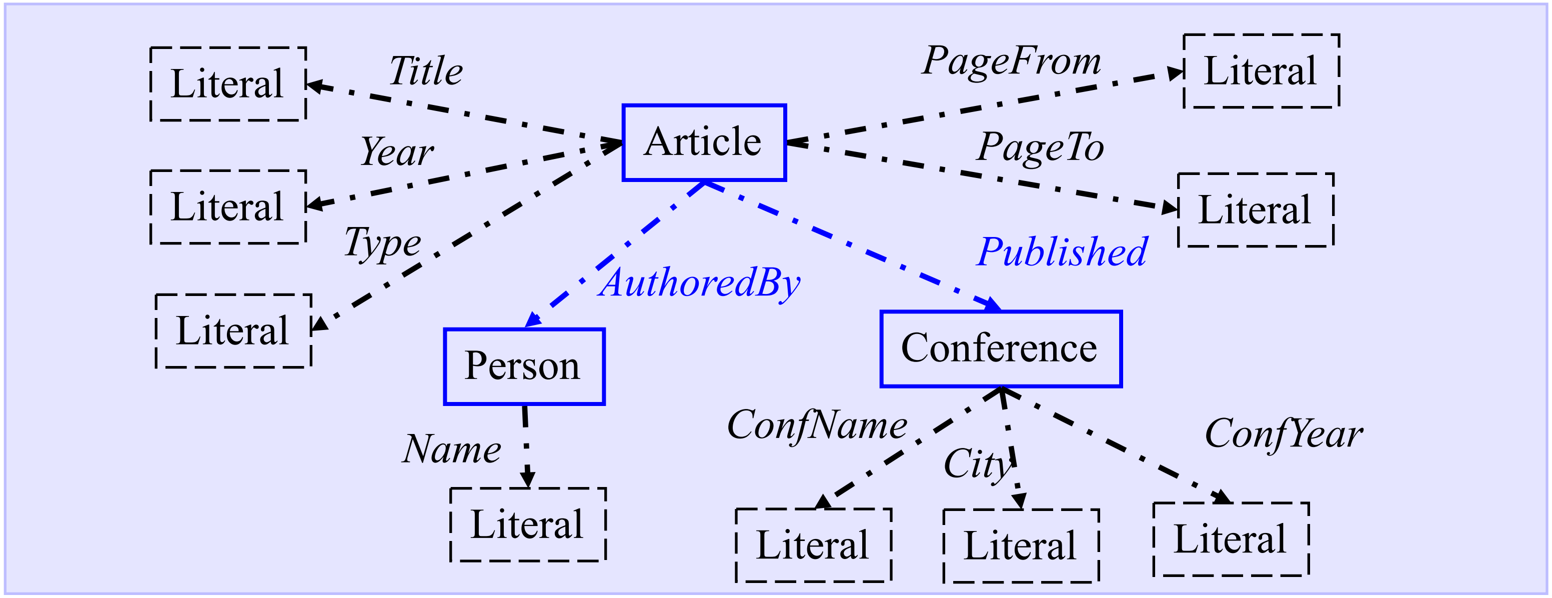
L2R EXPERIMENTS: FT_HOTELS

	First ontology	The enriched ontology (with Ddisj)
Recall (REC)	54%	54%
Recall (NREC)	8.2%	75.9%
Recall	8.3%	75.9%

The validation has been done manually on a pair of data sources which contain resp. 404 and 1392 reference of hotel.

The ontology enrichment led to an important increase of the recall.

L2R EXPERIMENTS: ONTOLOGY OF CORA



- ✓ DISJOINT(*Article,conference*), DISJOINT(*Article,Person*), DISJOINT(*Person,Conference*)

- ✓ All the properties are functional (PF), except *AuthoredBy*
- ✓ Two inverse funtional axioms that combine two attributes :

  PFI(*Title,Year,Type*), PFI(*ConfName, ConfYear*)
- ✓ LUNA(*AuthoredBy*).

L2R EXPERIMENTS: FT_HOTELS

	First ontology	First ontology + NSyn
Recall (REC)	52.7%	52.7%
Recall (NREC)	50.6%	94.9%
Recall	50.7%	94.4%

The results concern 1295 article reference and 1292 conference reference

For the references of Person, we obtained 4298 non reconciliations by exploiting LUNA on the relation *AuthoredBy*.

[Dong et al.'05] have obtained 97% of recall, computed only on REC, by using s supervised algorithm.

QUESTIONS?

N2R: A NUMERICAL METHOD FOR REFERENCE RECONCILIATION

[Saïs et al'09]

N2R: A NUMERICAL METHOD FOR REFERENCE RECONCILIATION

[Saïs et al'09]

- **N2R computes a similarity score for pair of references obtained from their common description.**
  - Uses known similarity measures, e.g. Jaccard, Jaro-Winkler.
  - Exploits ontology knowledge in a way to be coherent with L2R.
  - May consider the results of L2R: *Reconcile(i, i')*, *¬Reconcile(i, i')* , *SynVals(v, v')* and *¬SynVals(v, v')*.

N2R: COMMON DESCRIPTION

- Common attributes for a reference pair (i, i'):

$CAttr(i, i') = \{ a \mid \exists\, v, v' \in Val, \text{ st. } [a(i, v) \in Desc(i) \text{ and } a(i', v') \in Desc(i')] \}$

- Common relations for a reference pair (i, i'):

$CRel(i, i') = \{ r \mid \exists\, j, j' \in I, \text{ st. } [r(i, j) \in Desc(i) \text{ and } r(i', j') \in Desc(i')] \text{ or } [r(j, i) \in Desc(i) \text{ and } r(j', i') \in Desc(i')] \}$

- Set of values associated to a reference i:

$a+(i) = \{ v \mid \forall\, v, \text{ st. } a(i,v) \in Desc(i) \}$

- Set of references associated to a reference i:

$r+(i) = \{ j \mid \forall\, j, r(i, j) \in Desc(i) \}$

- Set of references to which a reference i is associated to a reference:
- $r-(i) = \{ j \mid \forall\, j, r(j, i) \in Desc(i) \}$

SIMILARITY DEPENDENCY MODELLING

[Saïs et al'09]

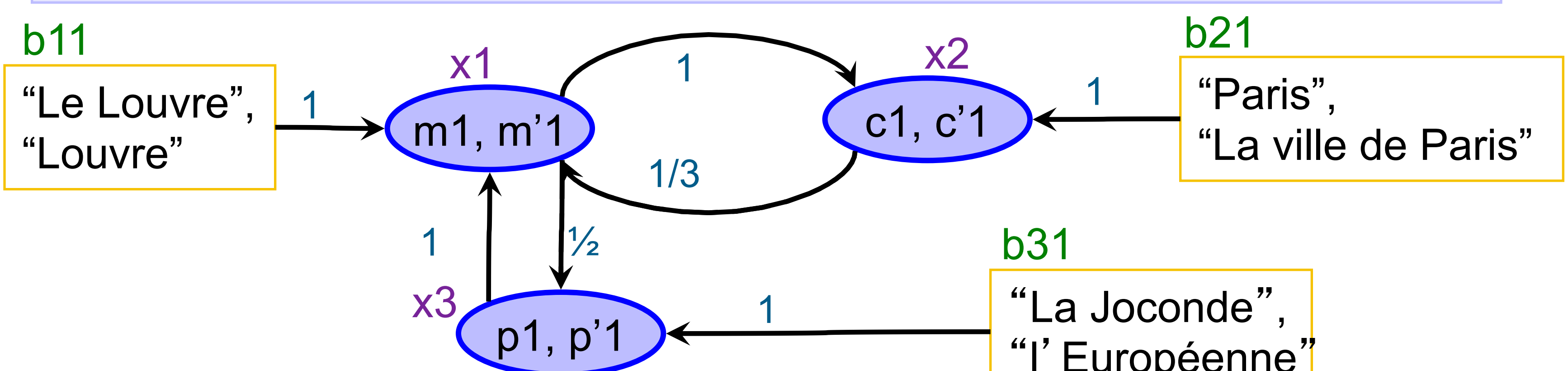
RDF facts in source S1:
Located(m1, c1), MuseumName(m1, "le Louvre")
Contains(m1, p1), CityName(c1, "Paris")
PaintingName(p1, "la Joconde")

RDF facts in source S2 :
Located(m'1, c'1), MuseumName(m'1, "Louvre")
Contains(m'1, p'1), CityName(c'1, "la Ville de Paris")
PaintingName(p'1, "l'Europèenne")

CAttr(m1, m'1) = {MuseumName} ,
CAttr(c1, c'1)= {CityName},CAttr(p1,p'1)={PaintingName}
CRel(m1, m'1)= {Located, Contains}
CRel(c1, c'1) = {Located }, CRel(p1,p'1) = {Contains}

MuseumName+(m1) = {"Le Louvre"},
MuseumName+(m'1) = {"Louvre"},
Located+(m1) = {c1}, Located+(m'1) = {c'1},
Located-(c1) = {m1} , Located-(c'1) = {m'1}, ….

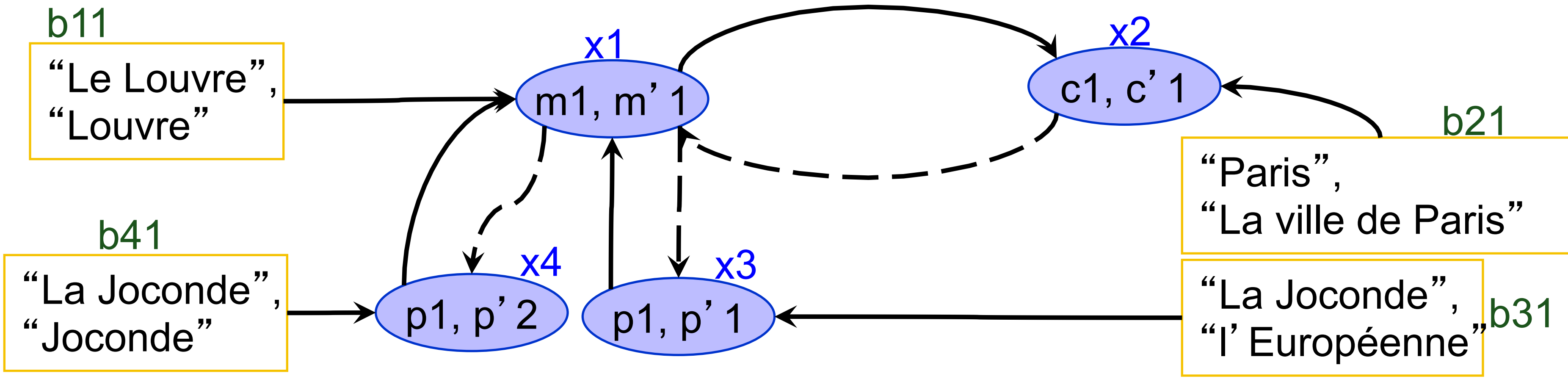
(c1, c'1) is functionally dependent on (m1, m'1)



➔ Equation system

N2R: ILLUSTRATION

[Saïs et al'09]



	x1	x2	x3	x4
Initialization	0.0	0.0	0.0	0.0
Iteration 1	0.8	0.3	0.1	0.7
Iteration 2	0.8	0.8	0.4	0.7
Iteration 3	0.8	0.8	0.4	0.7

x1 = max(max(b11, x3), x4), λ * x2)

x2 = max(b21, x1)

x3 = max(b31, λ* x1)

x4 = max(b41 , λ * x1)

λ= 1/(| CAttr | + | CRel |)   ε = 0.02

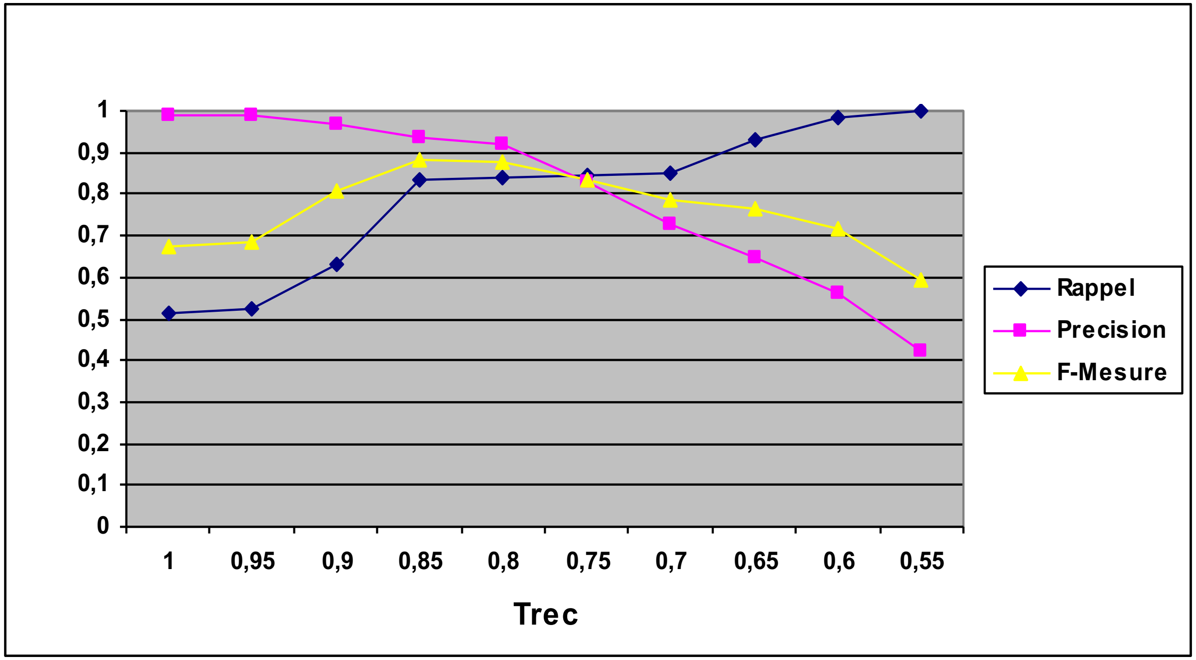
b11 = 0.8, b21 = 0.3, b31 = 0.1, b41 = 0.7

Solution: x1 = 0.8
x2 = 0.8
x3 = 0.4
x4 = 0.7

N2R EXPERIMENTS

N2R: RESULTS ON CORA

[Saïs et al'09]



Trec=1, all the reconciliations obtained by L2R are also obtained by N2R.

Trec=1 to Trec=0.85, the recall increases of 33 % while the precision decreases only of 6 %.

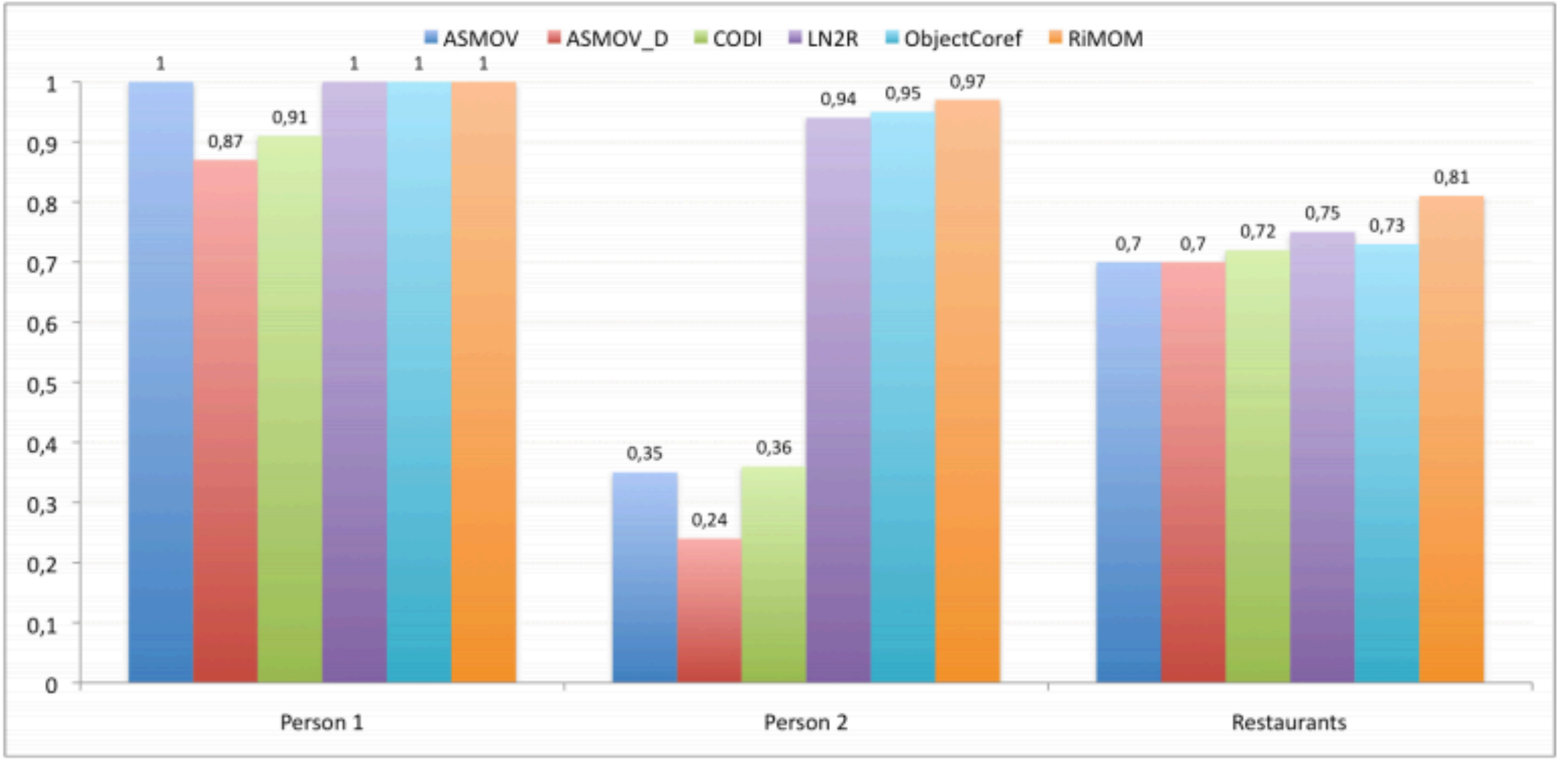
Trec = 0.85, the F-measure is of 88 %:

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

N2R: RESULTS IN OAEI[2] 2010

[Saïs et al'09]

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



[2] Ontology Alignment Evaluation Initiative

IMPORT BY QUERY

[Al Bakri et al 15]

A **knowledge-based** approach based on a backward-chaining algorithm that combines **:**

- Local reasoning (forward reasoning)
- External querying to bypass local data incompleteness (backward chaining)

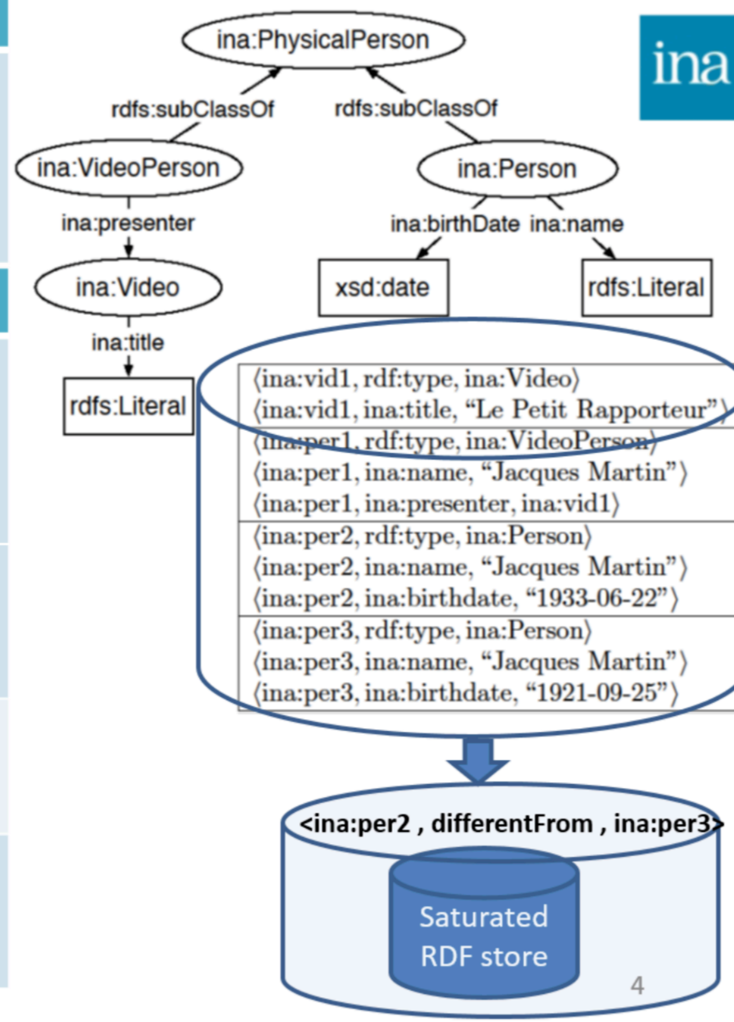
To infer a target owl:sameAs or contradict it.

Knowledge: (inverse) functional properties, composite keys, semantics of owl:sameAs (transitivity) and owl:differentFrom.

IMPORT BY QUERY

[Al Bakri et al 15]

	IF	THEN
R1	?p1 name ?name ?p1 birthdate ?d ?p2 name ?name ?p2 birthdate ?d	?p1 same_as ?p2
	IF	**THEN**
R2	?p1 name ?name ?p1 ina:presenter ?v1, ?v1 title ?t ?p2 name ?name ?p2 db:presenter ?t	?p1 same_as ?p2
R3	?p1 birthdate ?d1 ?p2 birthdate ?d2 ?d1 <> ?d2	?p1 differentFrom?p2
R4	?x1 same_as ?x2 ?x2 same_as ?x3	?x1 same_as ?x3
R5	?x1 same_as ?x2 ?x2 differentFrom ?x3	?x1 differentFrom ?x3

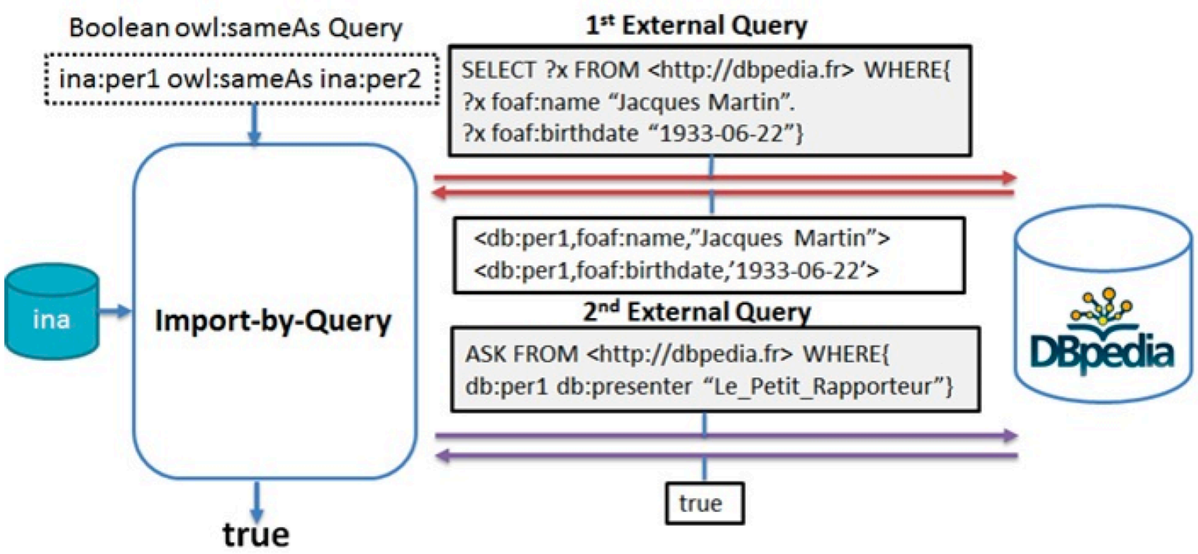


BUT <ina:per1, same_as, ina:per2> ? STILL UNKNOWN

IMPORT BY QUERY

[Al Bakri et al 15]

Builds on demand queries to some entry points of Linked Data
Alternates subquery rewriting steps based on backward chaining and external query evaluation (adaptation of Query-Subquery algorithm).



IMPORT BY QUERY- EXPERIMENTS

[Al Bakri et al 15]

1.5 million RDF facts, provided by a french national audiovisual institute (INA)
35 rules (built with the help of INA experts), 0.5 million external facts (DBPedia).

	IF	THEN
r7	$\langle ?x1$,`foaf:name`$, ?name1\rangle$, $\langle ?x2$, `skos:altLabel`$, ?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$
r8	$\langle ?x1$,`foaf:name`$, ?name1\rangle$, $\langle ?x2$, `skos:prefLabel`$, ?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$
r9	$\langle ?x1$,`rdfs:label`$, ?name1\rangle$, $\langle ?x2$, `skos:prefLabel`$, ?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$
r10	$\langle ?x1$,`rdfs:label`$, ?name1\rangle$, $\langle ?x2$, `skos:altLabel`$, ?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$
r11	$\langle ?x1$,`prop-fr:nom`$, ?name1\rangle$, $\langle ?x2$, `skos:prefLabel`$, ?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$
r12	$\langle ?x1$,`prop-fr:nom`$, ?name1\rangle$, $\langle ?x2$, `skos:altLabel`$, ?name2\rangle$, Similar($?name1, ?name2, 0.99$)	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$

	IF	THEN
r13	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$, $\langle ?x1$, `dbpedia:birthYear`$, ?Y1\rangle$, $\langle ?x2$, `ina:birthYear`$, ?Y1\rangle$ $\langle ?x1$, `dbpedia:deathYear`$, ?Y2\rangle$, $\langle ?x2$, `ina:deathYear`$, ?Y2\rangle$	$\langle ?x1$, `ina:sameAs`$, ?x2\rangle$
r14	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$, $\langle ?x1$, `dbpedia:birthYear`$, ?Y1\rangle$, $\langle ?x2$, `ina:birthYear`$, ?Y2\rangle$ notEqual($Y1, Y2$)	$\langle ?x1$, `ina:differentFrom`$, ?x2\rangle$
r15	$\langle ?x1$, `ina:sameNameDBp`$, ?x2\rangle$, $\langle ?x1$, `dbpedia:deathYear`$, ?Y1\rangle$, $\langle ?x2$, `ina:deathYear`$, ?Y2\rangle$ notEqual($Y1, Y2$)	$\langle ?x1$, `ina:differentFrom`$, ?x2\rangle$

IMPORT BY QUERY - EXPERIMENTS

[Al Bakri et al 15]

- External information can be useful to link Data

  2 links (108 differentFrom) with INA

  versus 4,884 links (resp.9,700) with DBPEDIA
- 100 % precision if the facts and rules are correct

  500 have been manually checked
- Reasoning allows to discover more links

  Silk only discovered 2% of the sameAs links discovered by the forward reasoner.
- Low number of imported facts

  Only 6,000 facts are needed (among 500,000 facts of the DBPedia extract)
- Efficient: 191s forward chaining, 7s per query (in average)

DATA LINKING: SUMMARY

- **Knowledge-based approaches** can take into account many kinds of knowledge:
  - ontology axioms, expert knowledge, assumption on datasets, referring expressions …
- Such approaches can easily be extended by new rules.

- **Logical approaches** infer *sure* identity links, can be used to infer differentFrom links.
- Can deal with large datasets:
  - forward chaining can be parallelized [Hogan et al. 12],
  - backward chaining can be used efficiently (minimization of the number of imported facts from external sources) [Al Bakri et al. 15].

DATA LINKING: SUMMARY

- Logical approaches are partial: they cannot decide for all pairs.

- Strong assumptions: data is clean, rules are certain (but even transitivity can lead to many wrong decisions!)

+ In **numerical approaches**, similarity scores can be propagated (equation system, probabilistic datalog).

+ Uncertainty can be modelled (similarity of literals, rules with exceptions, uncertain facts).

+- Similarity scores can be assigned to **more instance pairs**, but the decision is not guaranteed.

- The obtained scores are not so significant, **thresholds are difficult to fix**.

+- Linking rules are not always available but **can be discovered** from the data (e.g., key discovery approaches)

INSTANCE BASED ONTOLOGY MATCHING

ONTOLOGY MATCHING

- **Ontology alignment** [Shvaiko,Euzenat13]: computes a set A of mappings between elements (classes, properties) of two ontologies O1 and O2:

  f(O1,O2)=A

- The relations that are used to express a mapping can be: owl:equivalentClass, owl:equivalentProperty, rdfs:subClassOf, skos:closeMatch, skos:broader, etc.

- Example: A={(

  owl:equivalentClass(http://dbpedia.org/ontology/City, http://schema.org/City, *0.8*)}

KINDS OF INFORMATION

- **Terminology**: lexical information describing the ontology elements (i.e. labels, comments, ...)
  - *Example: Way* vs *Underground way*

- **Structure**: hierarchy of classes and properties (relations/attributes)
  - *Example: the sub-classes of Way are very similar to the sub-classes of Path*

- **Extension**: the existence of common instances!!

PARIS

[Suchanek et al. 12]

- Objective: instance-based ontology alignment and data linking (graph-based, unsupervised and probabilistic)
- Inputs: two populated RDFS ontologies with UNA (two different URI refer to two different entities)
- Principle:
  - Compute the similarities between literal values ("12 cm"="12")
  - Iterate (1) and (2) until a fix-point :
    1. Compute the probability that two instances are linked

$$P(i_1 = i_2)$$

    1. Compute the probabilities of subClassOf/subPropertyOf

$$P(C_i \subseteq C_j), P(P_i \subseteq P_j)$$

PARIS

[Suchanek et al. 12]

- Property functionality degree (computed from data)
    - *The more a property is functional the more the probability of* X=Y will be.
- **Local functionality**: $Fun(p,x) = 1 / \#y{:}p(x, y)$
- **Global functionality**: $Fun(p) = (\#x : \exists y{:}p(x,y)) / (\#x,y : p(x,y))$

- **Example**:

city(m1,Londres), city(m1,Orsay), city(m2,Tokyo)

Fun(city,m1)= ½ Fun(city,m2)=1

Fun(city)=2/3

➔ The same is done for **inverse functionality** (denoted fun^{-1})

PARIS

[Suchanek et al. 12]

Link probability computation

- **Positive evidence (P1)**: if there exists a property that is highly inverse functional which has range values that are equal with a high probability

$$P_1(x = x') = 1 - \prod_{\substack{r(x,y) \\ r(x',y')}} (1 - Fun^{-1}(p).P(y = y'))$$

isbn(x,isbn1), isbn(x',isbn2), P(isbn1=isbn2) = 1, fun^{-1}(isbn)=1 ...

P1(x=x') = 1 - ((1 - (1.1)) . ...) = 1-(0. ...) = 1

- **Negative evidence (P2)**: if there exists a property that is highly functional which has range values for the probability to be equal is very low.
- **Combination** : P(x=x') = P1(x=x').P2(x=x')

PARIS

[Suchanek et al. 12]

- The probabilities of the existence of a subsumption mapping between properties and between classes are also computed
- It is based on the proportion of common instances comparing to the number of instances of the general class

$$P(C \subseteq C') = \#(C \cap C') / \# C$$

$$P(p \subseteq p') = \#(p \cap p') / \# p$$

- To compute these probabilities, the probabilities of the existence of a sameAs link between instances are exploited.

PARIS - EXPERIMENTS

[Suchanek et al. 12]

Ontology	#Instances	#Classes	#Relations
Yago	2 795 289	292 206	67
Dbpedia	2 365 777	318	1 109

Linking or mapping if the probability >0.4

Instances			Classes		Relations	
Précision	Rappel	F-Mesure	Yago $\subseteq$ DBp Précision	DBp$\subseteq$ Yago Précision	Yago $\subseteq$ DBp Précision	DBp$\subseteq$ Yago Précision
90%	73%	81%	-	-	100%	92%
90%	73%	81%	94%	84%	100%	92%

<u>Instances</u>: DBPedia and Yago uses the URIs of Wikipedia (recall and precision possible)

<u>Classes/properties</u>: sampling + expert

5h00 to compute the linking probabilities for instances in one iteration (2h for the classes and 20 minutes for the properties)

DATA LINKING: SUMMARY

Numerous and different approaches …

- Supervised approaches: needs samples of linked data

  → It can be avoided by using assumptions like (UNA)
- Graph-based approaches: decision propagation (good recall but highly time consuming)

DATA LINKING: SUMMARY

Numerous and different approaches …

- Supervised approaches: needs samples of linked data

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

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

- Logical approaches: good precision but partial

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

DATA LINKING: SUMMARY

Numerous and different approaches …

- Supervised approaches: needs samples of linked data

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

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

- Logical approaches: good precision but partial

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

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

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

OUTLINE

- **Introduction**
  - Linked Data
  - Knowledge graphs
  - Knowledge graph refinement
- **Data Linking**
- **Identity Problem**
- **Conclusion**

IDENTITY PROBLEM

OWL:SAMEAS

The standardized Semantic Web identity predicate

Indicates that two different names (IRIs) refer to the same real-world entity

Strict semantics:

1) Reflexive,
2) Symmetric,
3) Transitive,
4) $\forall X \forall Y$ owl:sameAs$(X, Y) \wedge p(X, Z) \Rightarrow p(Y, Z)$

Essential in a decentralized knowledge space like the Web of Data

IDENTITY IS COMPLEX …

"Lessons Learned: Managing Identity is Hard"

Jamie Taylor
in ISWC 2017

Google

Knowledge Graph

"Biggest Problem: Identity"

Alan Patterson
in ISWC 2018

ebay

Knowledge Graph

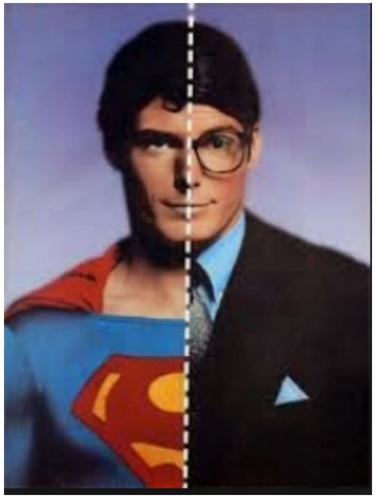
Source: Aaron Bradley
Twitter, October 26th, 2018

IDENTITY IS COMPLEX ...

From a Philosophical Point of View [Beek, 2018]

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

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



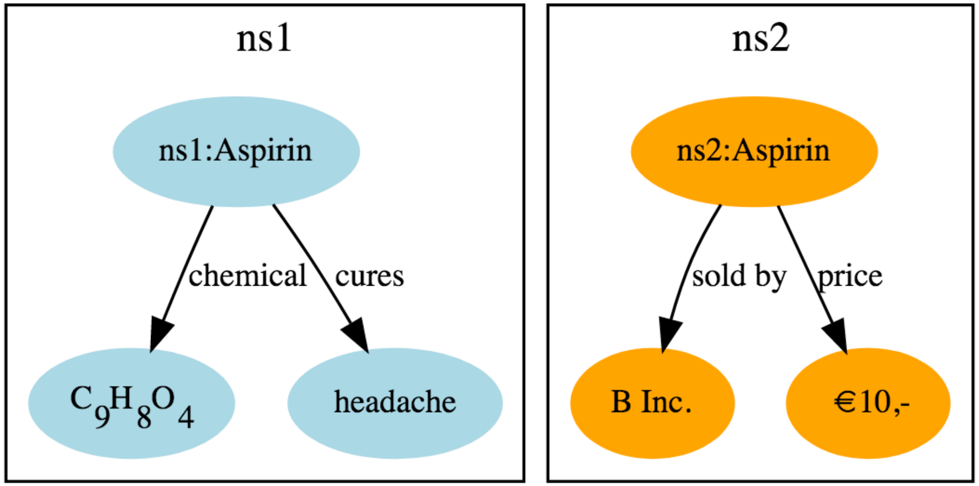
IDENTITY IS COMPLEX ...

From a Philosophical Point of View [Beek, 2018]

① Identity does not hold across modal contexts

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

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



IDENTITY IS COMPLEX …

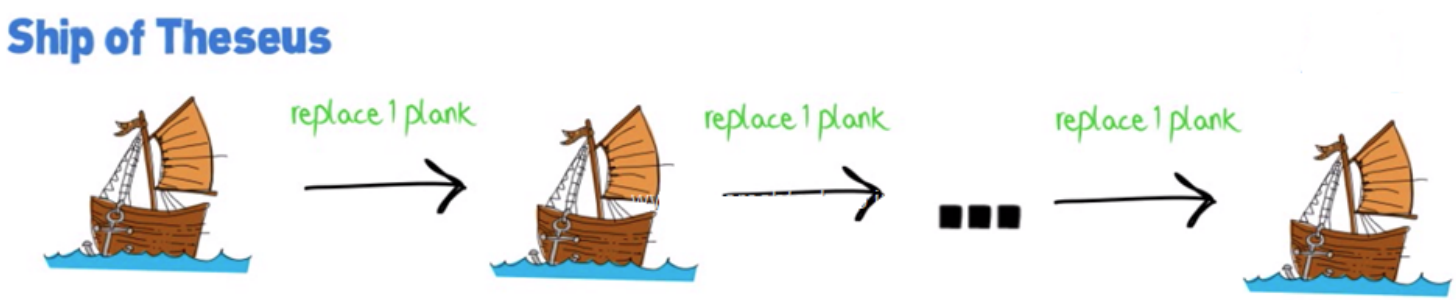
From a Philosophical Point of View [Beek, 2018]

① Identity does not hold across modal contexts

② Identity is context-dependent [Geach, 1967]

③ **Identity over time poses problems**

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



IDENTITY IS COMPLEX …

From an Operational Point of View

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

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

IDENTITY IS COMPLEX …

From an Operational Point of View

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

② **Modelers have different opinions about whether two objects are the same**

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

IDENTITY IS COMPLEX …

From an Operational Point of View

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

② Modelers have different opinions about whether two objects are the same

③ **Data linkage approaches are rarely 100% precise**

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

IDENTITY IS COMPLEX …

From an Operational Point of View

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

② Modelers have different opinions about whether two objects are the same

③ Data linkage approaches are rarely 100% precise

④ **Lack of alternative well-defined and standardized identity links**

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

THE 'SAMEAS PROBLEM'

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

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

THE 'SAMEAS PROBLEM'

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

However:

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

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

HOW TO LIMIT THIS ‘SAMEAS PROBLEM’?

- **Help users and applications identify IRIs referring to the same real-world entity, and distinguish between different entities**
  - Centralized Identity Management Services
  - Identity Observatories

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

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

IDENTITY MANAGMENT SERVICES

[BEEK, RAAD, ET AL. 2018]

SAMEAS.CC [Beek et al., 2018]

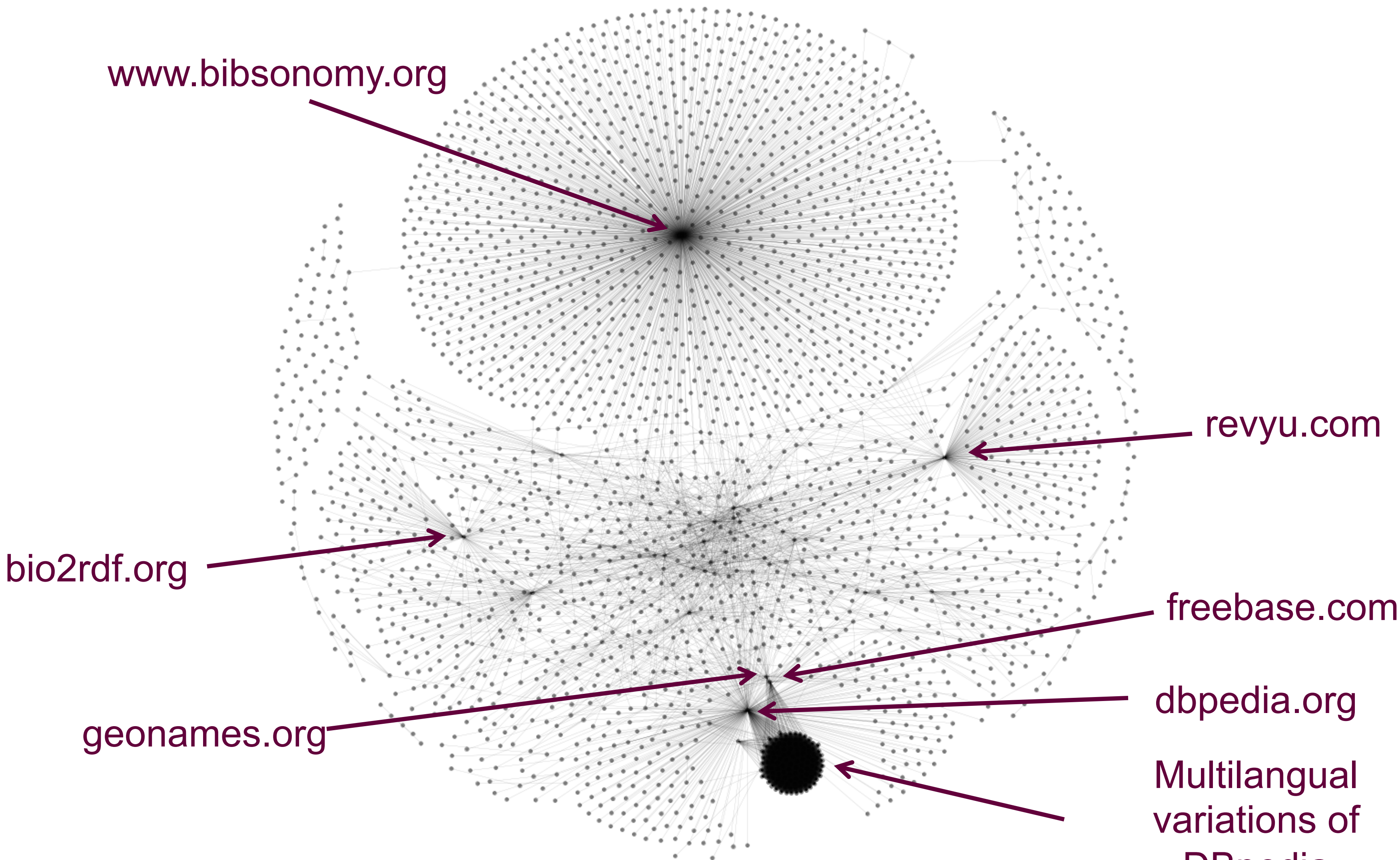
- Provides the largest collection of owl:sameAs triples
- Transitive closure of **558M distinct owl:sameAs** collected from the 2015 LOD Laundromat corpus
- Resulting in **49M equivalence classes**, that covers more than **179M terms**
- *Hosted at http://sameas.cc*

IDENTITY OBSERVATORIES

	sameas.org	LODsyndesis	sameas.cc
# Terms	203,953,936	65,315,931	179,739,567
# Statements	346,425,685	44,028,829	558,943,116
# owl:sameAs	Unknown	44,028,829	558,943,116
# Partitions	62,591,808	24,076,816	48,999,148
# Eq. Classes	Unknown	24,076,816	48,999,148

- **sameas.org:** Identity Bundles are not semantically interpretable (e.g. cannot be used by a DL reasoner to infer new facts)
- **LODsyndesis:** an order of magnitude smaller (link coverage)
- **sameas.cc**: static service with links from the 2015 LOD Cloud crawl

INTER-DATASET IDENTITY [Beek et al., 2018]



http://sameas.cc/explicit/img

IDENTITY OBSERVATORIES

Despite their technical limitations, identity observatories are more adopted in Linked Data applications

Not solely contributes in understanding the meaning of IRIs, but also there are many use-cases for such services:

- sameas.org and sameas.cc used as the basis of several **link invalidation** approaches [de Melo, 2013] [Cuzzola et al., 2015] [Valdestilhas et al., 2017] [Raad et al., 2018]
- **Query Answering** (under entailment) [Joshi et al., 2012]
- **Ontology Alignment**

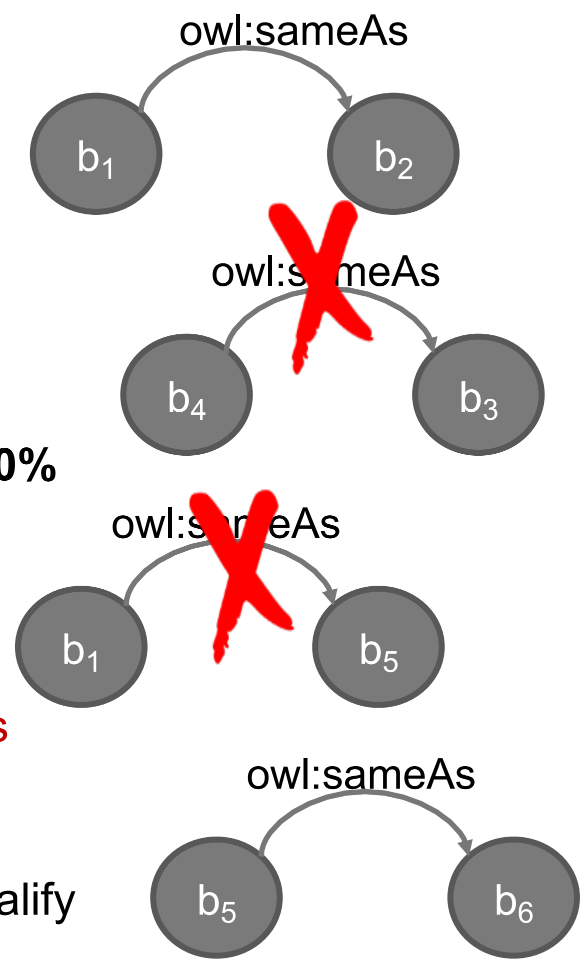
LINK INVALIDATION

IDENTITY CRISIS

- **owl:sameAs, indicates that two different descriptions refer to the same entity**
- **owl:sameAs semantics is too strict**
  - Reflexive, symmetric, transitive and
  - Property sharing:

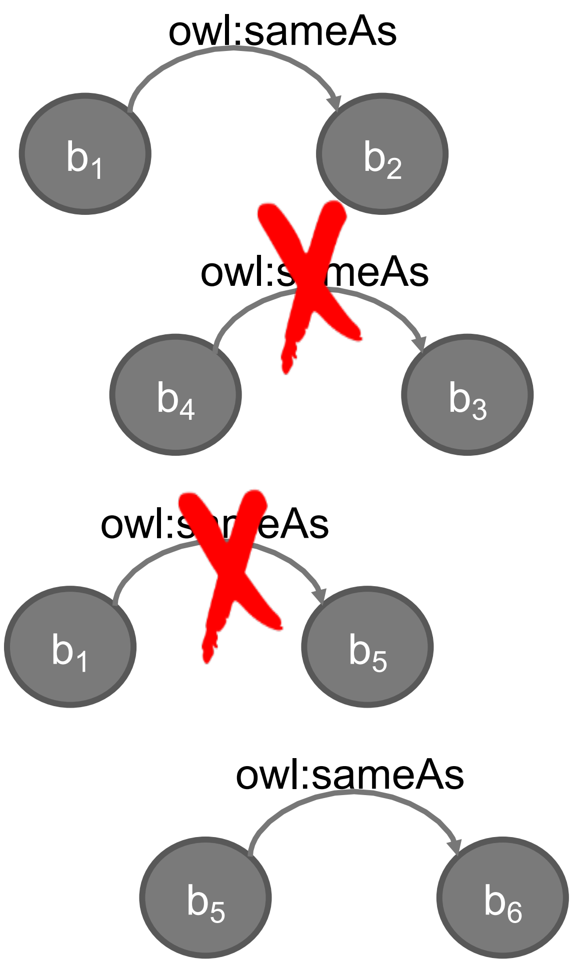
  $\forall X \forall Y$ owl:sameAs(X, Y)$\wedge$ p(X, Z) $\Rightarrow$ p(Y, Z)

- **Automatic data linking tools do not guarantee 100% precision, because of:**
  - Errors, missing information, data freshness, …
- **[Halpin et al. 2010]** showed that 37% of owl:sameAs links randomly selected among 250 identity links between books were incorrect.
- Problem: how to (semi)-automatically invalidate/requalify **owl:sameAs** links**?**



IDENTITY CRISIS: SOME SOLUTIONS

- [Halpin et al. 2010]: propose ontology of identity and invalidation of identity links by crowdsourcing.
- [de Melo 2013]: uses the Unique Name Assumption and the transitivity of links to detect inconsistencies in the data.
- [Papaleo et al. 2014, Papageorgiou et al. 2017]: exploit some ontology axioms to logically/numerically detect invalid identity links.
- [Raad et al. 2018] exploit identity graph topology and community detection to determine incorrect sameAs links.
- [Raad et al. 2017] computes contextual identity links for each pair of instances



IDENTITY PROBLEM: SOME SOLUTIONS

1. Erroneous identity link detection

2. Use of Alternate Links

3. Contextual identity link detection

1. ERRONEOUS IDENTITY LINK DETECTION

LOGICAL AND NUMERICAL APPROACH FOR LINK INVALIDATION

- Two **ontology-based methods** to detect invalid sameAs statements: a logical method and a numerical method

- We build 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).

- We analyse the descriptions provided in these contextual graphs to eventually detect inconsistencies or high dissimilarities.

LOGICAL METHOD

[Papaleo et al. 2014]

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

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.

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

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

… knowledge from expert or extracted.

LOGICAL METHOD

[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) ∧numOfPages(x,w_1) ∧ numOfPages(y,w_2) → SynVals(w_1,w_2)

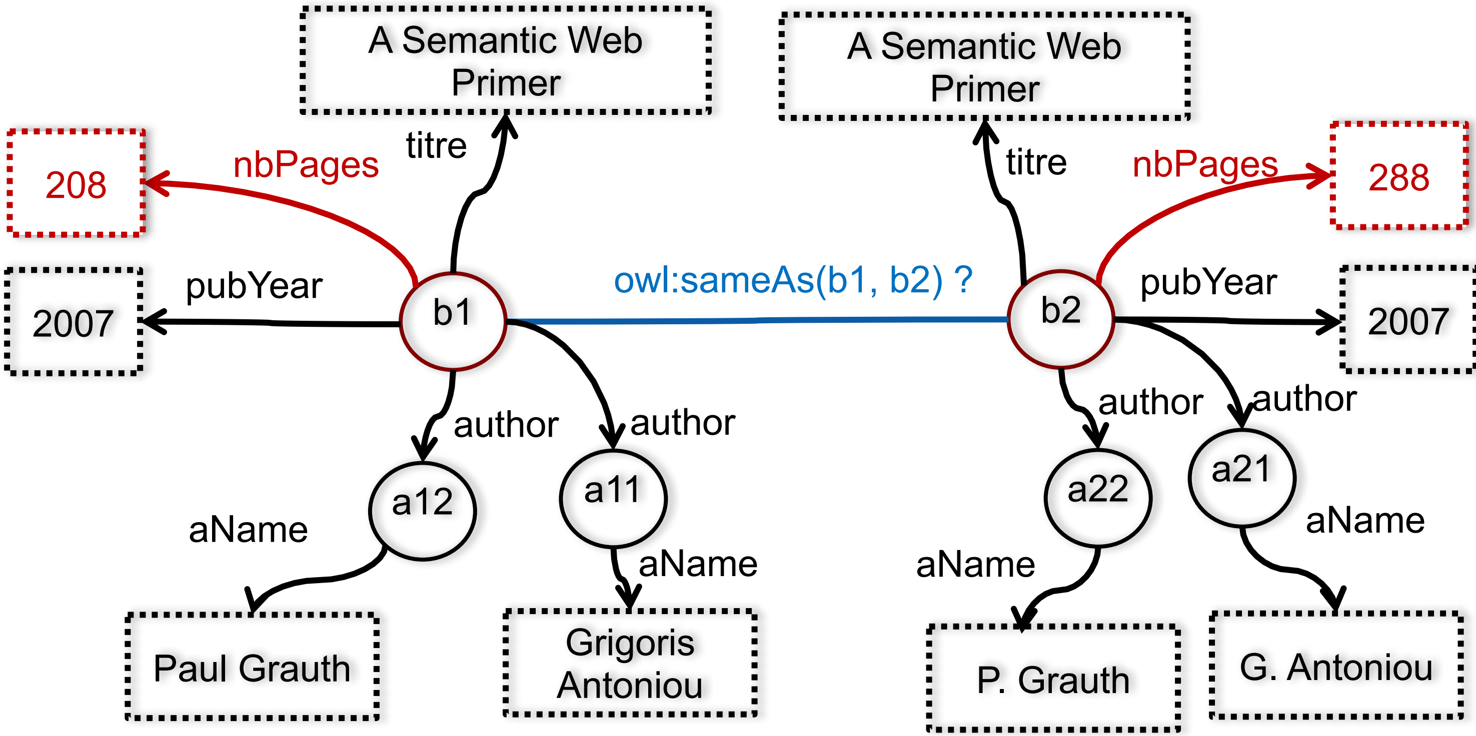
local complete properties

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

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

LOGICAL INVALIDAITON

[Papaleo et al. 2014]



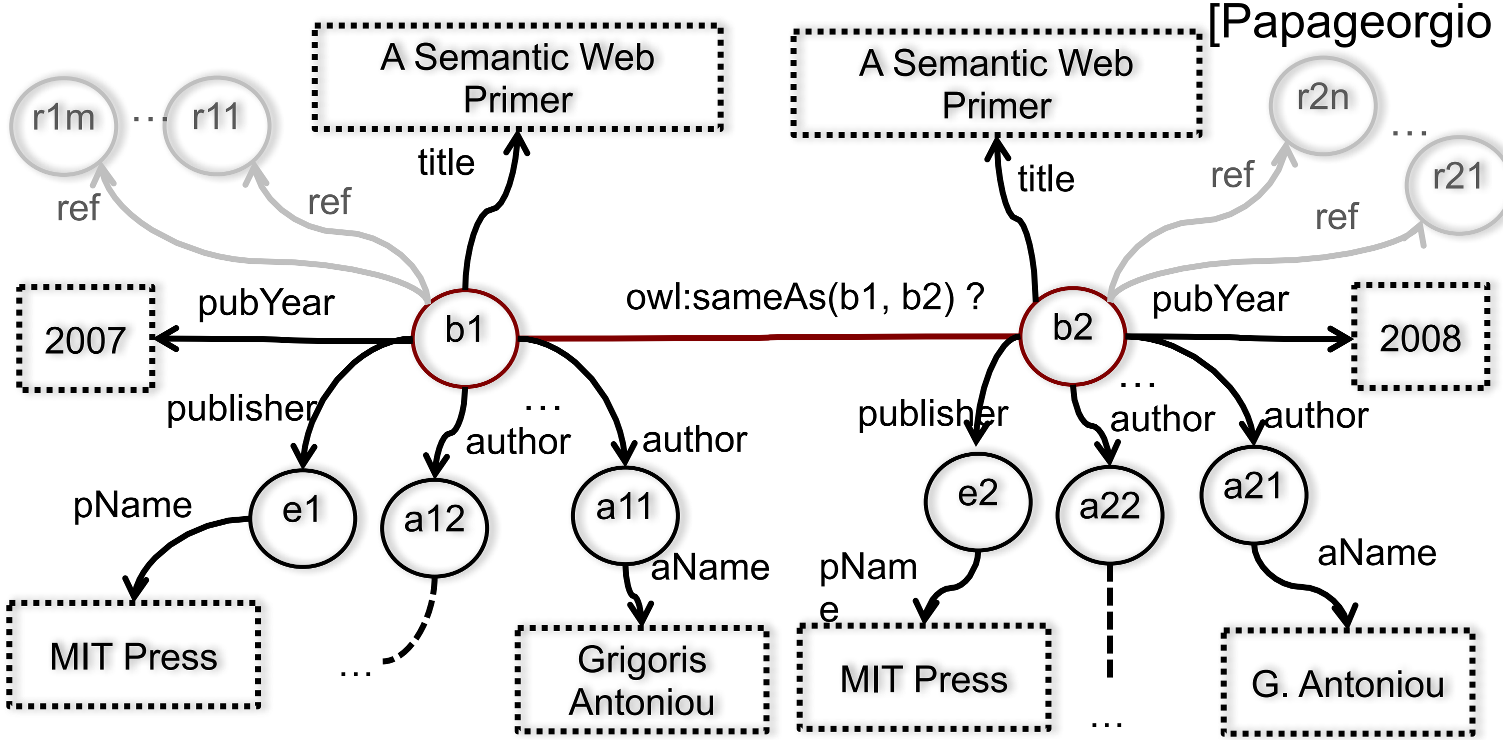
- If the property *nb-pages is declared as* functional then:

$$nbPages(b1, n1) \wedge nbPages(b2, n2) \wedge (b1=b2) \Rightarrow n1=n2.$$

➔ owl:sameAs(b1, b2) est faux.

NUMERICAL METHOD: EXAMPLE

[Papageorgio et al. 2017]



- P= {title, pubYear, publisher, author, aName, pName}
- Sim("A Semantic Web …``, "A Semantic Web …``) = 1, Sim(" 2007", "2008") = 0
- Sim("MIT Press``, "MIT Press``) = 1, ➔ CSim(e1, e2) = 1
- Sim("Grigoris Antoniou``, "G. Antoniou``) = 0.5, ... ➔ CSim(a11, a21) = 0.5, ….
  ➔ CSim(a12, a22) = 0.5

- CSim(b1, b2) = 0.62 if agregation function is average and
- CSim(b1, b2) = 0 if agregation function is minimum

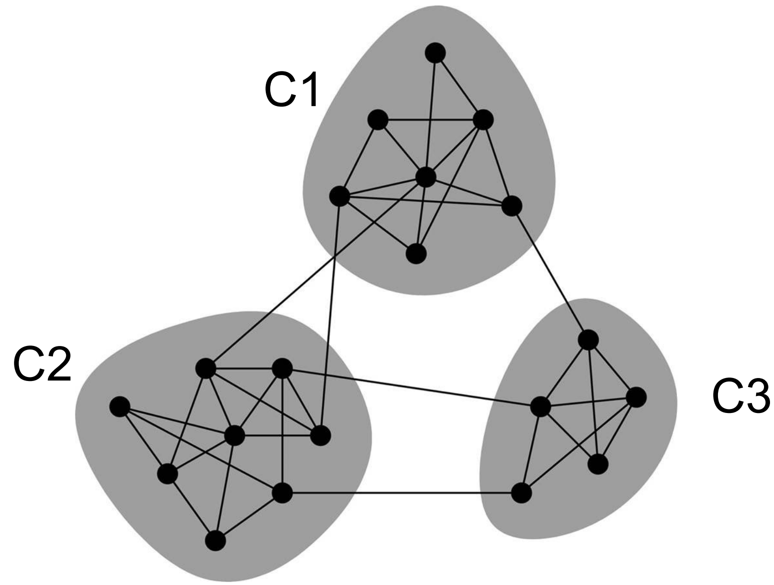
COMPARAISON LOGICAL/NUMERICAL

	Logical method [Papaleo, Pernelle and Saïs (2014)]			**Numerical method** (Agregation = average) **[Papageorgiou, Pernelle and Saïs 2017]**			
Datasets	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**thresh old**
Person1	0.69	0.98	**0.81**	1.0	0.98	**0.99**	0.3
Person2	0.5	**1.0**	**0.67**	0.994	**0.989**	**0.99**	0.2
Restaurant	0.63	0.97	**0.77**	0.97	1.0	**0.98**	0.4

- Average gain of 23% F-measure (significant increase in precision, comparable recall)

NETWORK BASED

[Raad *et al.*, 2018, under review]



- 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.*, 2018, under review]

Ranking of identity links

intra-community erroneousness degree

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

inter-community erroneousness degree

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

Correct link 0 — Error degree — 1 Erroneous link

NETWORK BASED

[Raad *et al.*, 2018, under review]

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)
- Identity network of **331M** edges and **179M** nodes: after removing symmetric and reflexive links.

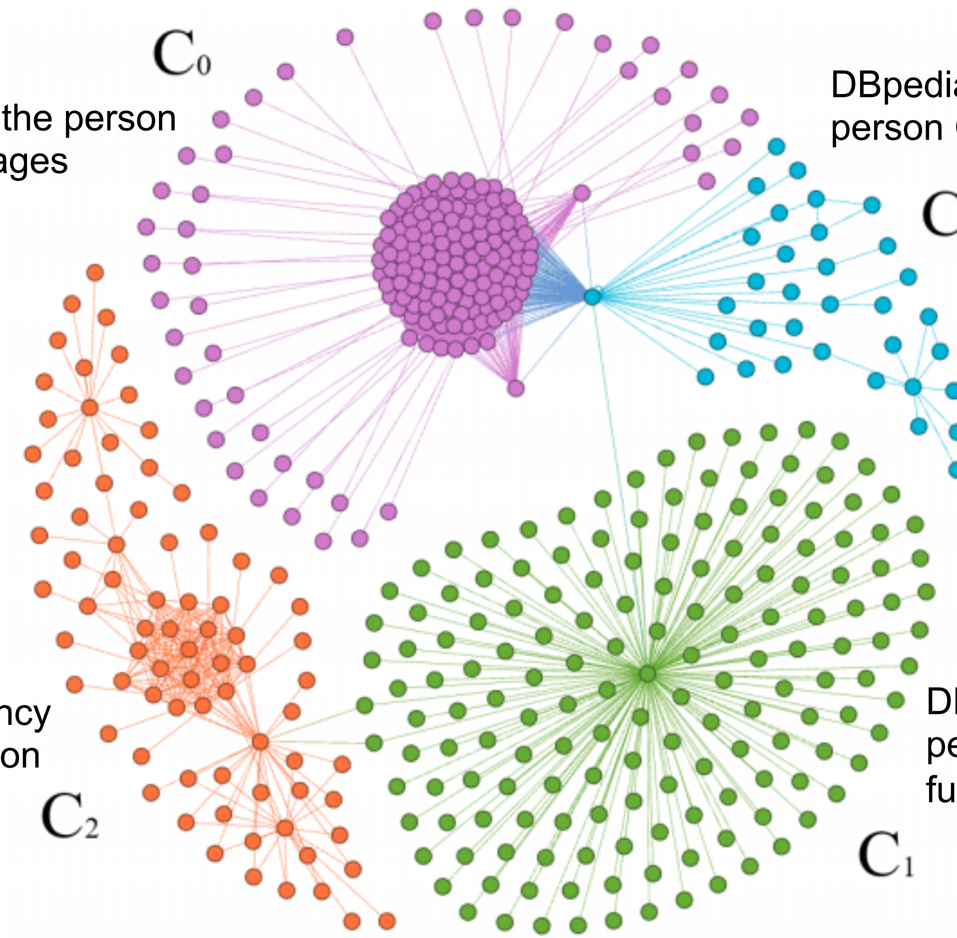
NETWORK BASED

[Raad *et al.*, 2018, under review]

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

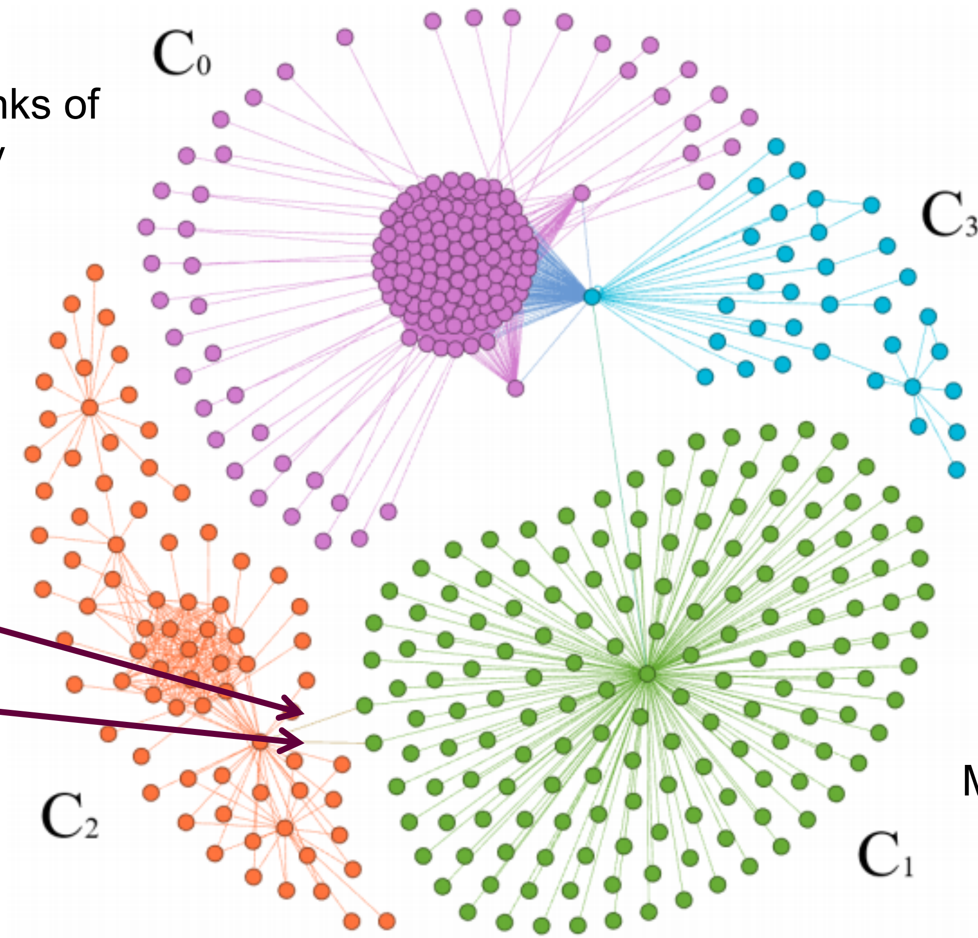
NETWORK BASED

[Raad *et al.*, 2018, under review]

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

NETWORK BASED

[Raad *et al.*, 2018, under review]

Precision on a randomly chosen set identity links from LOD

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

- **Scales up** to a graph of **28.3 billion** triples: **12 hours**
- **Validates correct owl:sameAs links**
  - 100% of owl:sameAs with an **erroneousness degree <0.4** are correct
- Can **invalidate a large set of owl: sameAs links** on the LOD:
  - **1.26M** owl:sameAs have an **erroneousness degree** in [0.99, 1]

ERRONEOUS LINK DETECTION: SUMMARY

Positive points

- Different approaches relaying on different kinds of information (constraints, axioms, content and network)
- Good scalability of the approaches: up to 28.3 Billion triples
- Evaluations on real data on the LOD

ERRONEOUS LINK DETECTION: SUMMARY

Positive points

- Different approaches relaying on different kinds of information (constraints, axioms, content and network)
- Good scalability of the approaches: up to 28.3 Billion triples
- Evaluations on real data on the LOD

Limitations

- **Qualitative evaluation** often missing or conducted on only insignificant number of links (**max= 200** over **331M**)
- Some **assumptions** can be assumed on only **few datasets** on the LOD: UNA and provenance information.
- **Ontology axioms** are not always **available**: how to ensure their **validity** in every dataset. Is the LocatedIn is functional for every museum?
- **Difference** relationships are rarely available: useful for inconsistency checking

ERRONEOUS LINK DETECTION: SUMMARY

*"common metrics such as centrality, clustering, and degree **are insufficient for detecting quality** … Description Richness and Open SameAs Chain metrics look more promising, especially at detecting good and bad links, respectively, they report too **many false positives** for reference sets"* [Gueret et al., 2012]	*"Data linking algorithms (LIMES, SILK and DBpedia Extraction Framework) have a better consistency index than repositories such as sameas.org (13%) "* [Valdestilhas et al., 2017]	*"Due to the subjectivity of near-identity and similarity, we suggest that additional properties be used to describe the exact nature of the relationship"* [de Melo 2013]
Need for hybrid approaches	Need for more controlled link publication protocols	Need for alternate links

2. USE OF ALTERNATE LINKS

2. USE OF ALTERNATE LINKS

Use of weaker alternative links to express relatedness between resources/concepts.

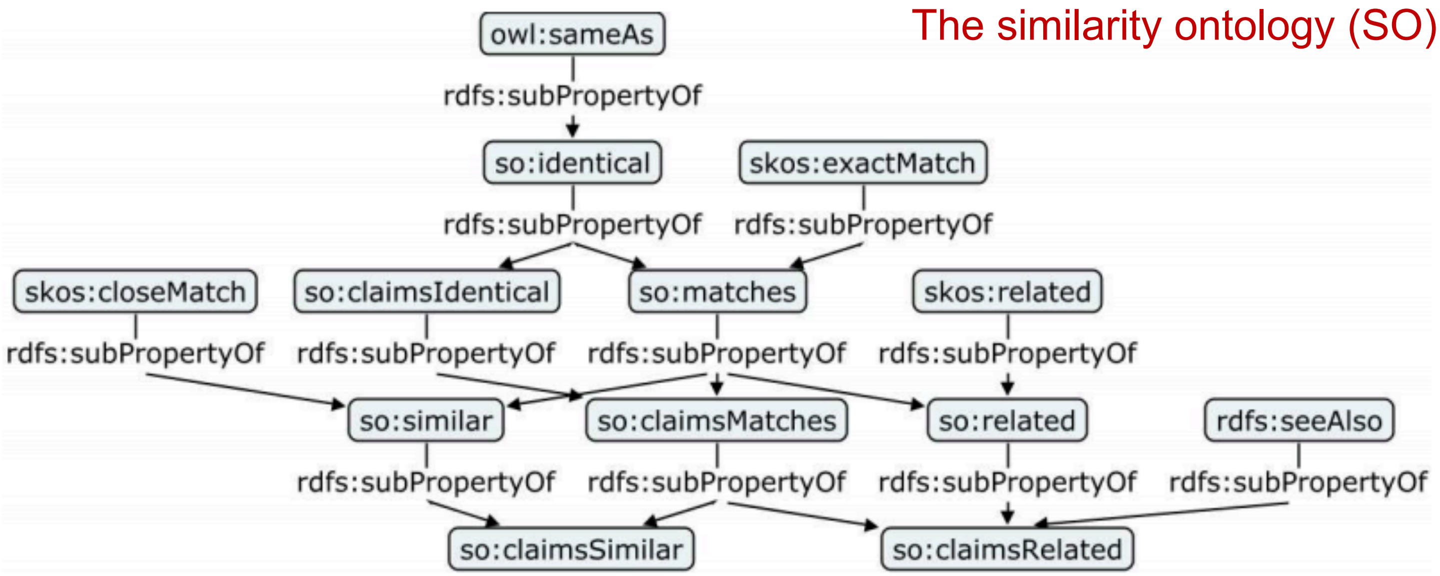
- UMBEL[1] vocabulary introduces umbel:isLike "*to assert a link between similar individuals who may be believed to be identical*"
- Vocab.org[2] introduces similarTo to be used when having two things that are not the owl:sameAs
- [de Melo, 2013] introduces lvont:nearlySameAs and lvont:somewhatSameAs, two predicates for expressing near-identity in the Lexvo.org[3]
- **Use of domain-specific identity relations:**
  - ex:sameBook to express identity between two books

[1] http://umbel.org
[2] http://vocab.org
[3] http://lexvo.org

2. USE OF ALTERNATE LINKS

- [Halpin et al., 2010] proposed a similarity ontology (SO) in which they hierarchically represent 13 different predicates including 8 new ones.

The similarity ontology (SO)



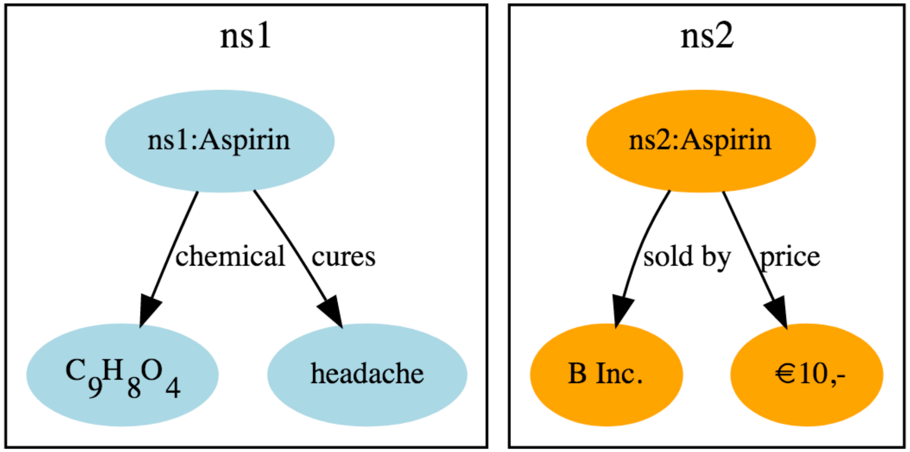
		Transitive	Non-transitive
Reflexive	Symmetric	*so:identical*	*so:similar*
	Non-Symmetric	*so:claimsIdentical*	*so:claimsSimilar*
Non-Reflexive	Symmetric	*so:matches*	*so:related*
	Non-Symmetric	*so:claimsMatches*	*so:claimsRelated*

Reflexivity, Symmetry and Transitivity properties for the 8 new predicates.

3. CONTEXTUAL IDENTITY LINKS

3. CONTEXTUAL IDENTITY LINKS

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



3. CONTEXTUAL IDENTITY LINKS

[Beek et al., 2016]

- Propose an approach that allows the **characterization o**f the **context** in which an **identity link is valid**

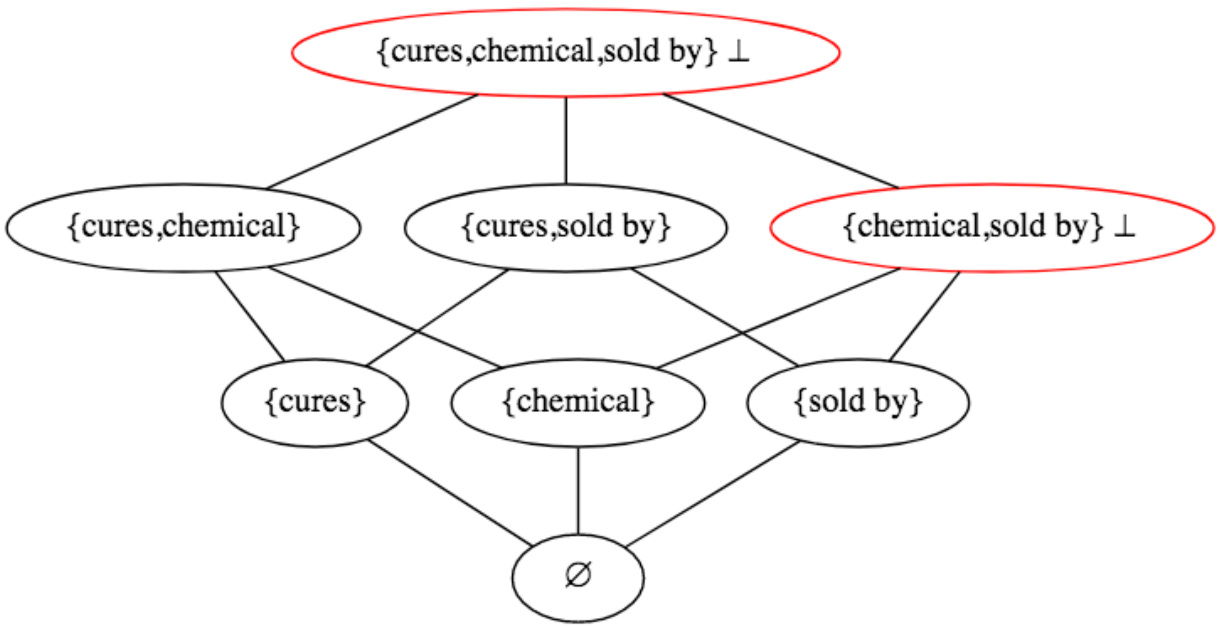
- A context is a **subset of properties** for which two individuals must have the **same values**

- Contextual identity link **preserves equivalence relation**, w.r.t. a subset of the properties

3. CONTEXTUAL IDENTITY LINKS

[Beek et al., 2016]

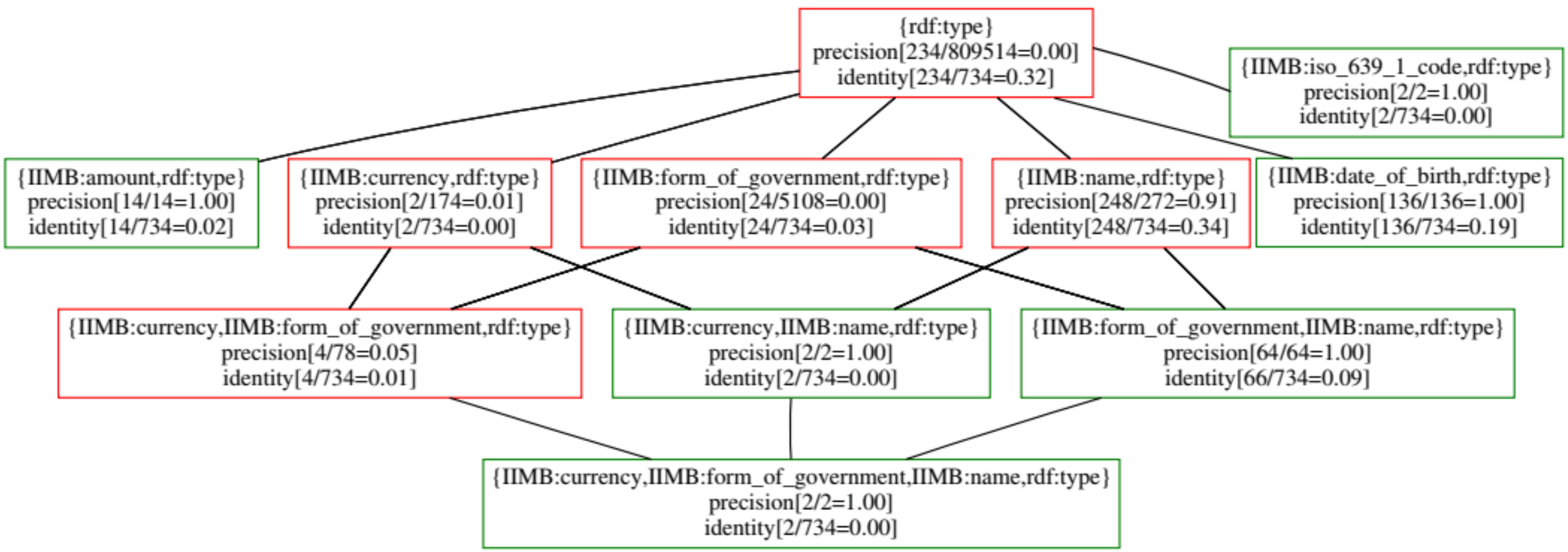
- All the **possible subsets** of properties organized in a **lattice** using the set inclusion relation.



3. CONTEXTUAL IDENTITY LINKS

[Beek et al., 2016]

- Evaluation on a dataset in the instance matching track of the OAEI2012 : a variant of the IIMB datasets.
- The obtained identity subrelations



3. CONTEXTUAL IDENTITY LINKS

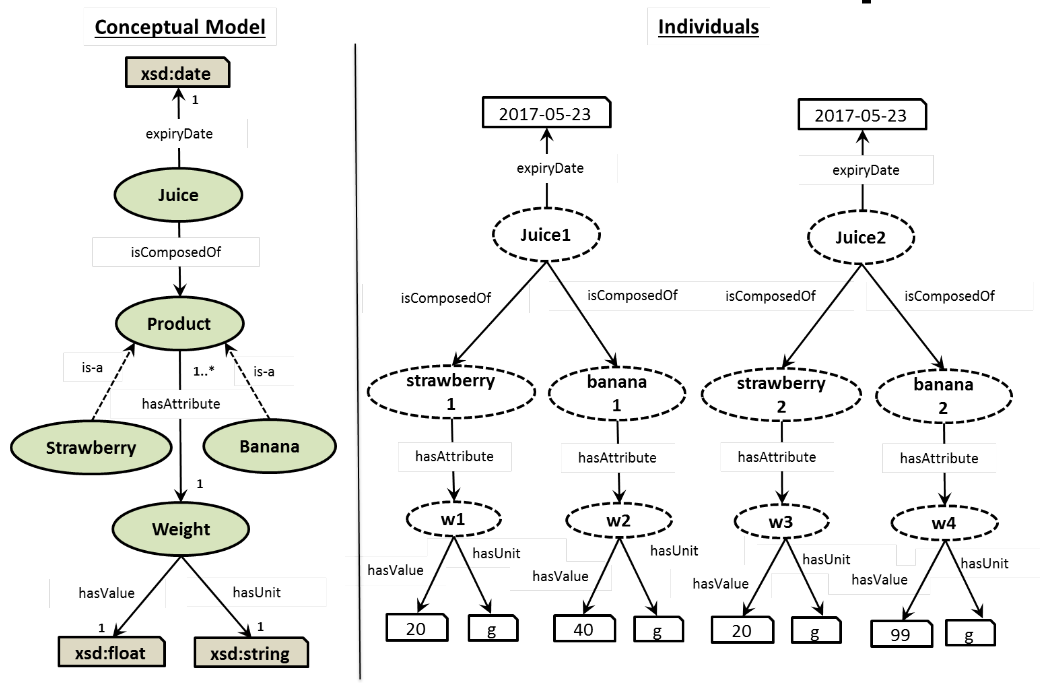
[Raad et al., 2017]

- New predicate *:identiConTo* 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
- All the possible contexts are organized in a lattice using an order relation.

3. CONTEXTUAL IDENTITY LINKS

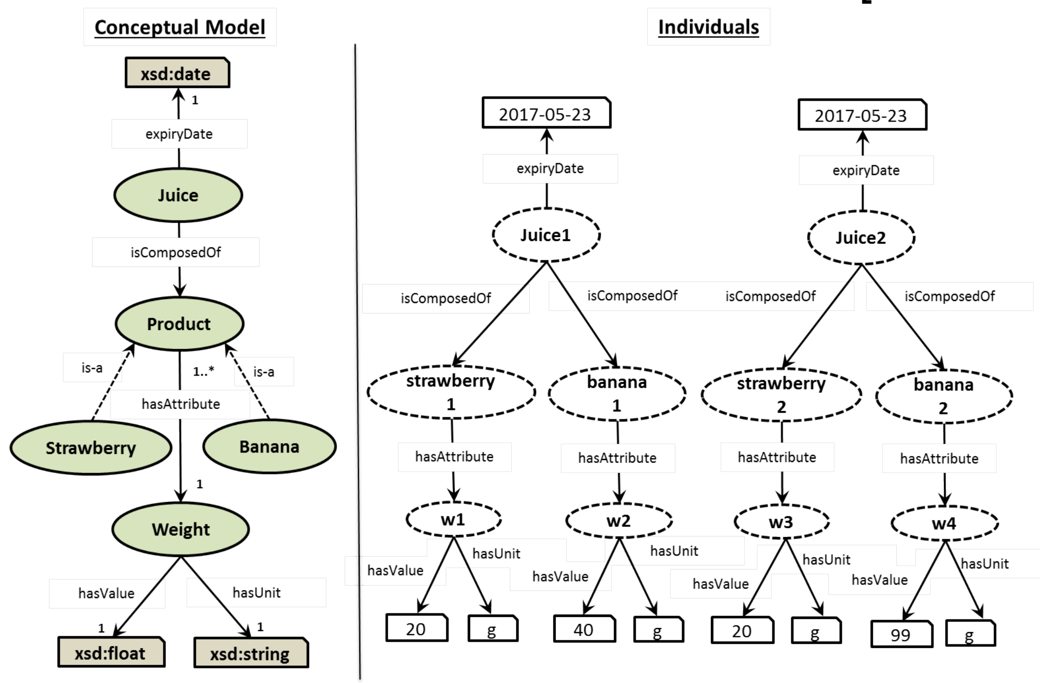
[Raad et al., 2017]



3. 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)**

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

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}}) **}**

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

Π_c**(Juice) = {** (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf $^{-1}$}),
(Strawberry, {rdf:Type}, {isComposedOf $^{-1}$}) **}**

3. CONTEXTUAL IDENTITY LINKS

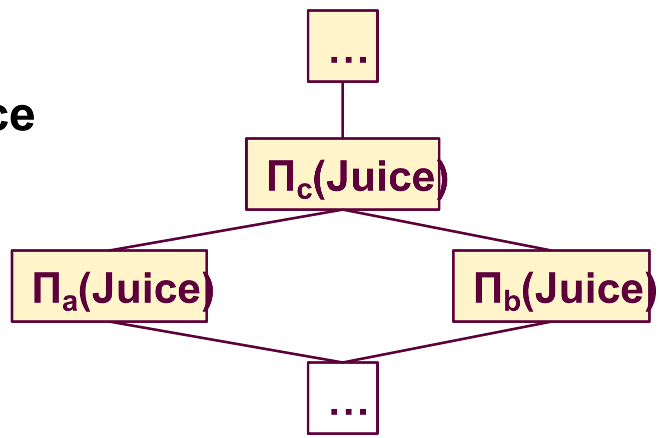
[Raad et al., 2017]

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}}) **}**

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

Π_c**(Juice) = {** (Juice, {rdf:Type, expiryDate}, {isComposedOf}),
(Banana, {rdf:Type}, {isComposedOf^{-1}}),
(Strawberry, {rdf:Type}, {isComposedOf^{-1}}) **}**

The **possible contexts** are organized in a **lattice** using an order relation.



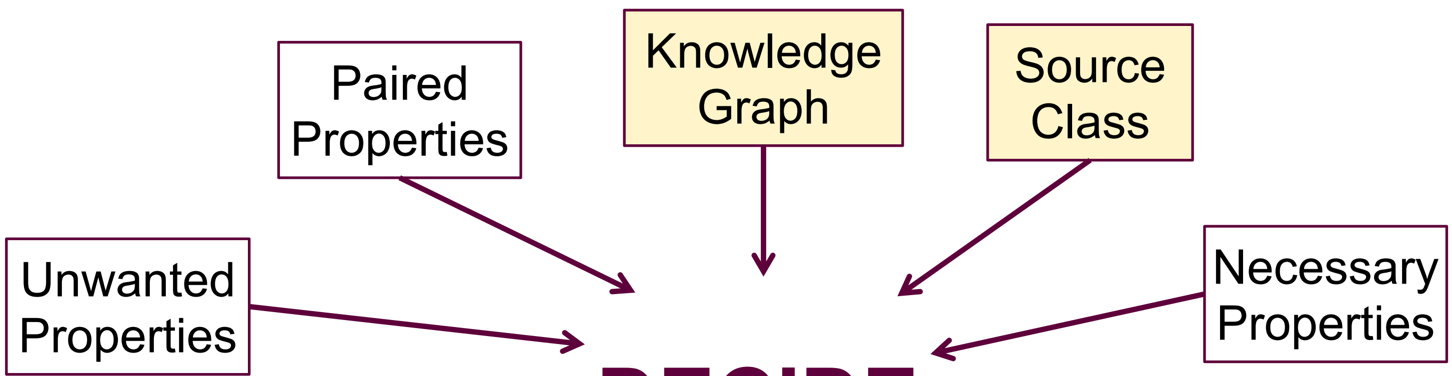
Π_a(Juice) ≤ Π_c(Juice) **Π_b(Juice) ≤ Π_c(Juice)**

each local context in Π_c(Juice) is less specific or equal to its corresponding local context in Π_a(Juice)

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

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



DECIDE

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

3. 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

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

PO²
A Process and Observation Ontology

CellExtraDry + Carredas

- 950 *classes*
- 1,5 million *triplets*
- 284 *processus de transformations*

Irrelevant properties
up = (observation, observes, *)

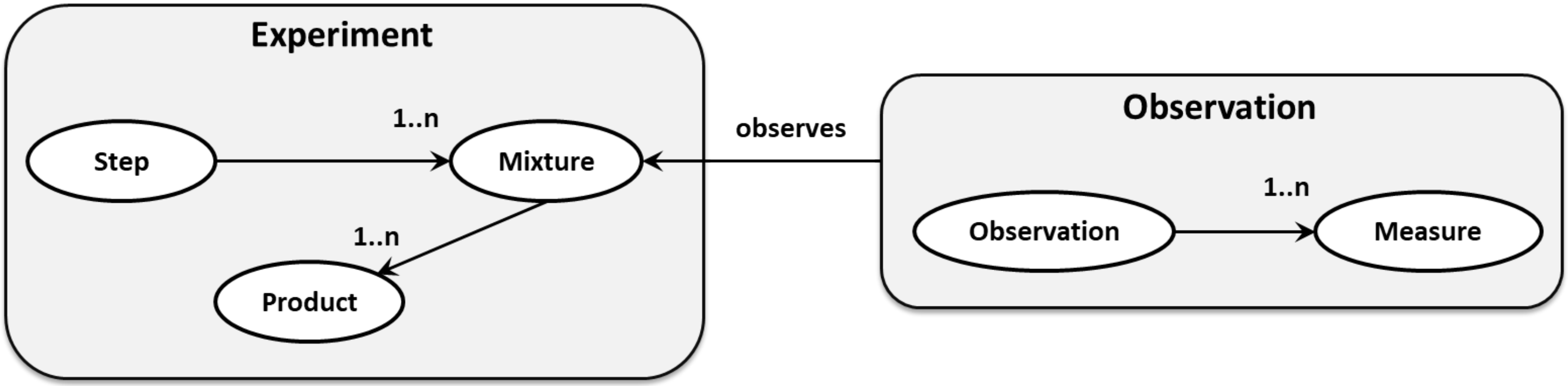
DECIDE

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

	Experiment 1	Experiment 2
	Mixture	**Step**
# Instances	1,187	581
# Possible pairs	703,891	168,490
# Distinct Global Contexts	2 232	718
# Contextual identity links	1, 279,376	348,017
# Contextual identity links per pair	1.81	2.06

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]



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

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

with $m_1 \simeq m_2$

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

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Detection of 38 844 rules

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

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

3. CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

Detection of 38 844 rules

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

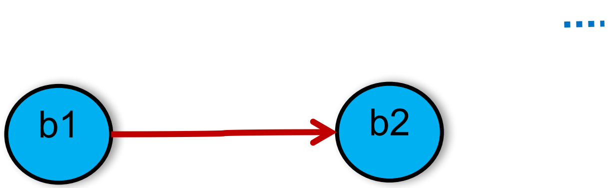
The domain experts has evaluated the plausibility of the best **20 rules** (in termes of error rate and support)

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

IDENTITY PROBLEM: SUMMARY

- Different kinds of identity relationship

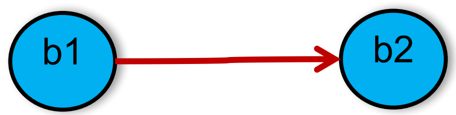


IDENTITY PROBLEM: SUMMARY

- Different kinds of identity relationship

owl:sameAs — Genuine identity

lvont:somewhatSameAs — Near/weak identity

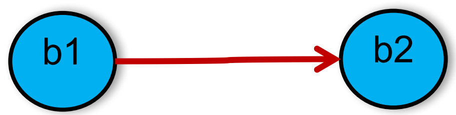


IDENTITY PROBLEM: SUMMARY

- Different kinds of identity relationship

owl:sameAs — Genuine identity

lvont:somewhatSameAs — Near/weak identity



:sameBook

:sameExpression — Subjective identity

:sameArtWork

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

IDENTITY PROBLEM: SUMMARY

- Different kinds of identity relationship
- Need of hybrid methods

Network Topology

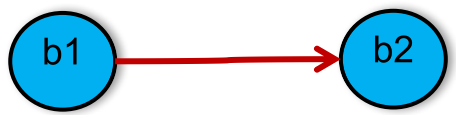
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

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

IDENTITY PROBLEM: SUMMARY

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

Network Topology

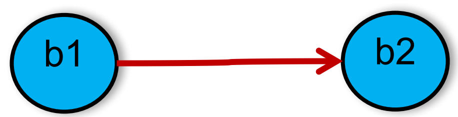
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

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

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

Link Properties: Transitivity, symmetry, …

Link added-value: Information gain, reachability, …

Link meta-data: availability, evolution

IDENTITY PROBLEM: SUMMARY

What is about the **distinctness** relation?

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

Network Topology

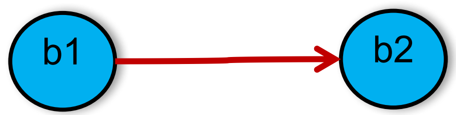
Source Reliability

Link Content

Ontology Axioms

owl:sameAs

lvont:somewhatSameAs



:sameBook

:sameExpression

:sameArtWork

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

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

Link Properties: Transitivity, symmetry, …

Link added-value: Information gain, reachability, …

Link meta-data: availability, evolution

REFERENCES (1)

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

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

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

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

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

G. de Melo. In AAAI 2013.

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

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

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

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

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

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

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

REFERENCES (2)

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

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

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

H. Paulheim. In WoDOOM 2014.

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

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

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

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

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

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

REFERENCES (3)

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

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

[Atencia et al.'12] Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking. Manuel Atencia, Jérôme David, François Scharffe. In EKAW 2012

[Cohen et al. 2003] A comparison of string distance metrics for name-matching tasks.

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg.

In IIWEB@AAAI 2003.

[Ferrara13] Evaluation of instance matching tools: The experience of OAEI.

Alfio Ferrara, Andriy Nikolov, Jan Noessner, François Scharffe. OM@ISWC 2013

[Hu et al. 2011] A Self-Training Approach for Resolving Object Coreference on the Semantic Web. Wei Hu, Jianfeng Chen, Yuzhong Qu. In WWW 2011

[Kang et al. 2008] Interactive Entity Resolution in Relational Data: A Visual Analytic Tool and Its Evaluation. Kang, Getoor, Shneiderman, Bilgic, Licamele,

In IEEE Trans. Vis. Comput. Graph2008

[P.N. Mendes et al'12] Sieve Linked Data Quality Assessment and Fusion

Pablo N. Mendes, Hannes Mühleisen, Christian Bizer

In the international workshop LWDM@EDBT 2012.

REFERENCES (4)

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

Douglas B. Lenat and Edward A. Feigenbaum

In Artificial Intelligence 47 (1991)

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

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

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

Nathalie Pernelle, Fatiha Saïs. and Danai Symeounidou.

In Journal of Web Semantics

[Saïs et al.07] L2R: a Logical method for Reference Reconciliation.

Fatiha Saïs, Nathalie Pernelle and Marie-Christine Rousset. In AAAI 2007.

[Saïs et al.09] Combining a Logical and a Numerical Method for Data Reconciliation.

Fatiha Saïs., Nathalie Pernelle and Marie-Christine Rousset.

In Journal of Data Semantics.

[Saïs et Thomopoulos'08] Reference Fusion and Flexible Querying.

Fatiha Saïs and Rallou Thomopoulos.

In OTM ODBASE 2008.

REFERENCES (5)

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

Pavel Shvaiko, Jérôme Euzenat. In TKDE 2013

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

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

[Soru et al. 2015] ROCKER: a refinement operator for key discovery.

Soru, Tommaso, Edgard Marx, and Axel-Cyrille Ngonga Ngomo.

In *WWW*, 2015.

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

Symeonidou, Danai, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs.

In ISWC 2014.

[Symeonidou et al. 2017*] VICKEY: Mining Conditional Keys on RDF datasets* .

Danai Symeonidou, Luis Galarraga, Nathalie Pernelle, Fatiha Saïs and Fabian Suchanek.
In ISWC 2017.

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

REFERENCES (6)

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

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

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

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

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

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