REFERRING EXPRESSION (RULE) DISCOVERY FOR DATA LINKING

FATIHA SAÏS, joint work with,

Armita Khajeh Nassiri, Nathalie Pernelle and Gianluca Quercini

SÉMINAIRE RÉSIDENTIEL INRAE
SEMANTIC LINKED DATA

11-14 OCT. 2021- DOMAINE DU LAZARET - SÈTE

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

cnrs dépasser les frontières

université PARIS-SACLAY

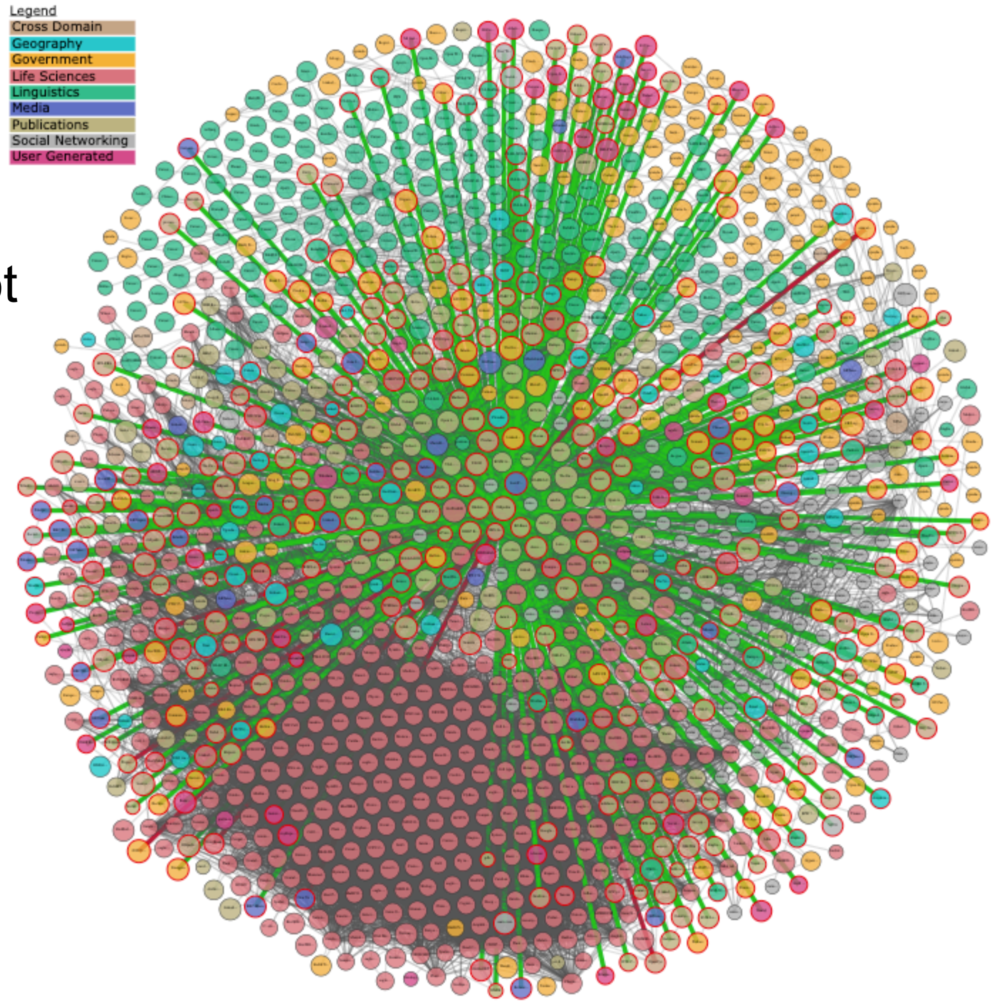
WEB OF DATA

Knowledge Graphs publicly available

- over 1 250 sources in LOD
- more than 650 k graphs in lod-a-lot
- over **100B** triples
- about **500M** links: most are **sameAs** links

LOD – Linked Data Cloud

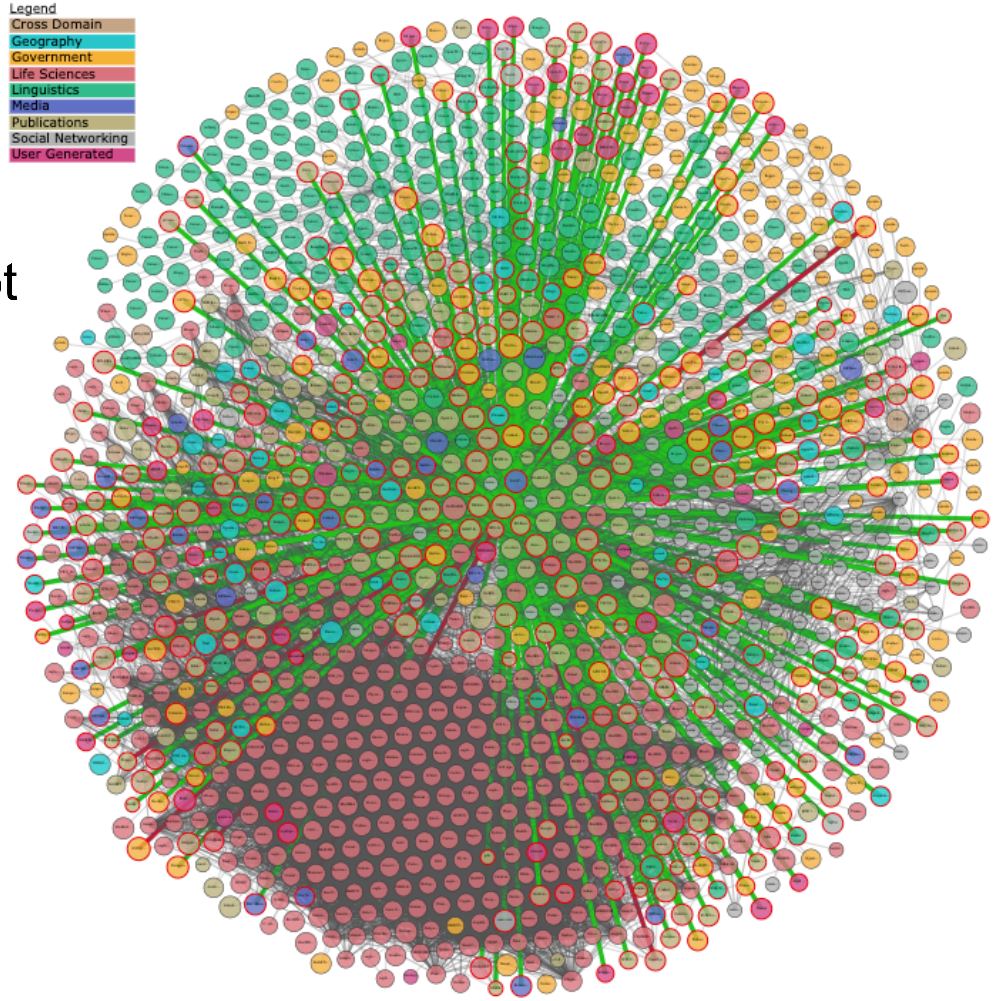
"Linking Open Data cloud diagram 2020, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"



WEB OF DATA

LOD – Linked Data Cloud

"Linking Open Data cloud diagram 2020, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"



Knowledge Graphs publicly available

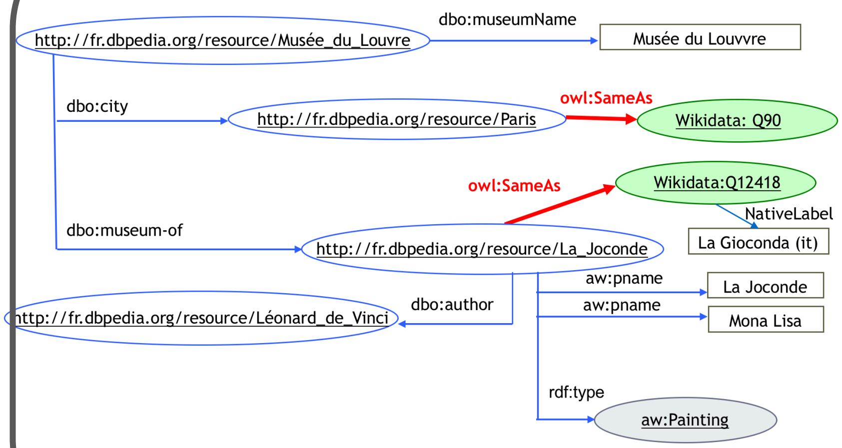
- over 1 250 sources in LOD
- more than 650 k graphs in lod-a-lot
- over **100B** triples
- about **500M** links: most are **sameAs** links

Application domains

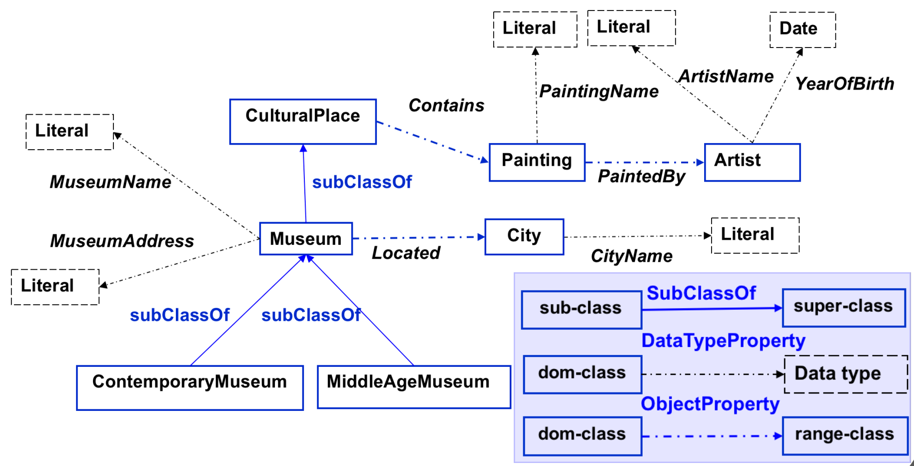
- **Cross-domain**: Dbpedia, yago, wikidata, …
- **Media & Music** : BBC, INA, MusicBrainz, …
- **Government**: US, UK, FR, DE, …
- **Geographic**: LinkedGeoData, IGN, …
- **Life sciences**: GO, SwissProt, Bio2RDF, …
- **Cultural heritage**: INA, BNF, Europeana, …
- **Law, Theology, Tourism**, …

KNOWLEDGE GRAPHS IN A WHOLE

RDF Graphs



OWL Ontology

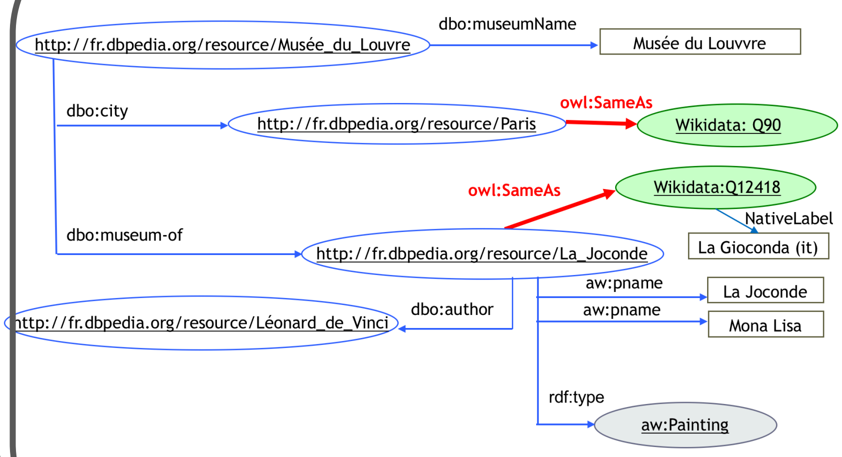


Ontology axioms and rules

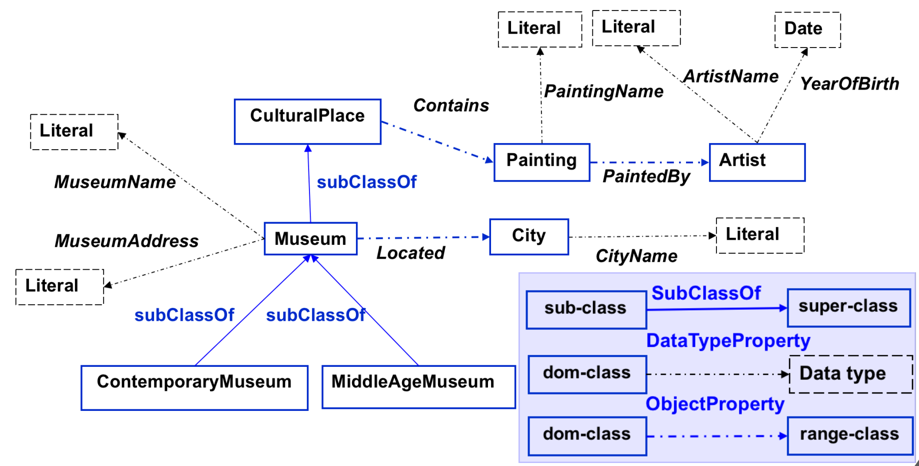
- Disjunction between classes/properties
- Subsumption (hierarchy)
- (inverse) Functionality of properties
- Symmetry
- Cardinalities
- Keys
- Logical rules
- ...

KNOWLEDGE GRAPHS IN A WHOLE

RDF Graphs



OWL Ontology



Querying (SPARQL)

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

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

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

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

Academic side

2007

2008

2012

2007

Commercial side

2012

2016

2018

2019

WHO IS DEVELOPING KNOWLEDGE GRAPHS?

Academic side

2007 DBpedia

2008 YAGO select knowledge

2012 WIKIDATA

2007 Freebase

NELL

Cyc

Commercial side

Tencent 腾讯, UniProt, USGS, 2012 Google, Bing, Alibaba.com, Baidu 百度, PubMed, 2016 facebook, Uber eats, airbnb, Antoni van Leeuwenhoek, Maastro, Deutsche Nationalbibliothek, europeana, BBC, The New York Times, Getty, Kungl. Biblioteket National Library of Sweden, Library of Congress, NXP, Renault, Bloomberg, Reuters, EPA, IOS Press, Politie, European Commission, Walmart, Siemens, Best Buy, 2018 ebay, Deloitte., Elsevier, Springer Nature, amazon.com, 2019, accenture

Targeted applications and services

- Information integration,
- recommendation,
- transparency,
- regularity compliance,
- multilingual support,
- conversational agents,
- …

Examples of use-cases

- **Web search**: "*things and not strings*" (e.g. GooglePanel)
- **Social network**: description of skills, jobs, schools, etc. (e.g. LinkedIn)
- **Commerce**: description of products, events, location…, behaviour patterns (e.g. Ebay ShopBot)
- **Finance**: emerging events detection, risk assessment (e.g. Bloomberg), …

DATA QUALITY IN KGS

COMPLETENESS?

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

Incomplete data
DBPedia: 1.7M person, 700K missing birth dates

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

DATA QUALITY IN KGS

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!

Errors

Yago: 9K cases of Childs born before their parents

RULE MINING - FOR DATA QUALITY IMPROVEMENT

- Error detection
- Fact checking

- Fact prediction
- Data linking
- …
-

OUTILINE

- **Rule mining** : techniques and main differences
- Referring expressions : RE-miner for data linking
- Conclusion

RULE MINING

A **horn rule** or implication : $B1 \wedge B2 \wedge \dots \wedge Bn \Rightarrow r(x, y)$

Body **Head**

Example:

$hasChild(p, c) \wedge isCitizenOf(p, s) \Rightarrow isCitizenOf(c, s)$

$motherOf(m, c) \Rightarrow \neg fatherOf(m, c)$

$worksAt(p, c) \Rightarrow affiliatedTo(p, c)$

RULE MINING

A **horn rule** or implication : $B1 \wedge B2 \wedge \ldots \wedge Bn \Rightarrow r(x, y)$

Body **Head**

Example:

$hasChild(p, c) \wedge isCitizenOf(p, s) \Rightarrow isCitizenOf(c, s)$ **Prediction**

$motherOf(m, c) \Rightarrow \neg fatherOf(m, c)$ **Error detection**

$worksAt(p, c) \Rightarrow affiliatedTo(p, c)$ **Ontology alignement**

RULE MINING

Knowledge bases are not complete

➡ So the rules are not necessarily always correct
➡ **measures** : **confidence** and **support**

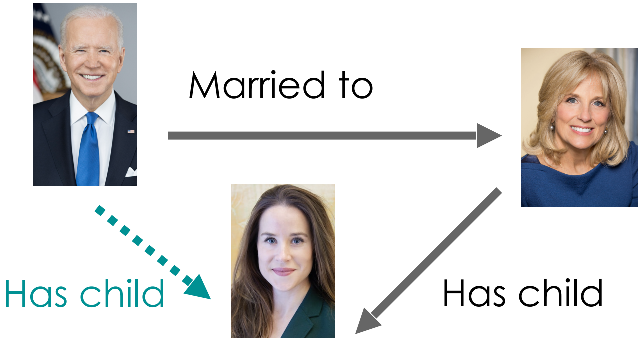
RULE: hasChild(X,Y) ∧ marriedTo(X,Z) → hasChild(Z,Y)

RULE MINING

Knowledge bases are not complete

➡ So the rules are not necessarily always correct
➡ **measures** : **confidence** and **support**

RULE: hasChild(X,Y) ∧ marriedTo(X,Z) → hasChild(Z,Y)



X= Joe, Y= Jill, Z= Ashley

Prediction: hasChild(Joe, Ashley)

Support: Number of true predictions of the rule in KB
Confidence: Number of true predictions / Number of total predictions

RULE MINING: EXISTING TECHNIQUES

1. **Generate and test Techniques, heuristic technique with backtracking (AMIE3,RUDIK)**
   - Consider a candidate rule
   - Compute quality measures for this rule
   - Refine the rule to generate more candidates and test

2. **Divide and Conquer Techniques (Tilde)**
   - Divide: search for a rule that is valid on a part of knowledge base
   - Conquer: recursively conquer the remaining examples by learning more rules
   - Combine the rules to form the final solution

RULE MINING: EXISTING TECHNIQUES

1. **Generate and test Techniques, heuristic technique with backtracking (AMIE3,RUDIK)**
   - Consider a candidate rule
   - Compute quality measures for this rule
   - Refine the rule to generate more candidates and test

Guarantees to find all rules that fulfill quality measures and the language bias

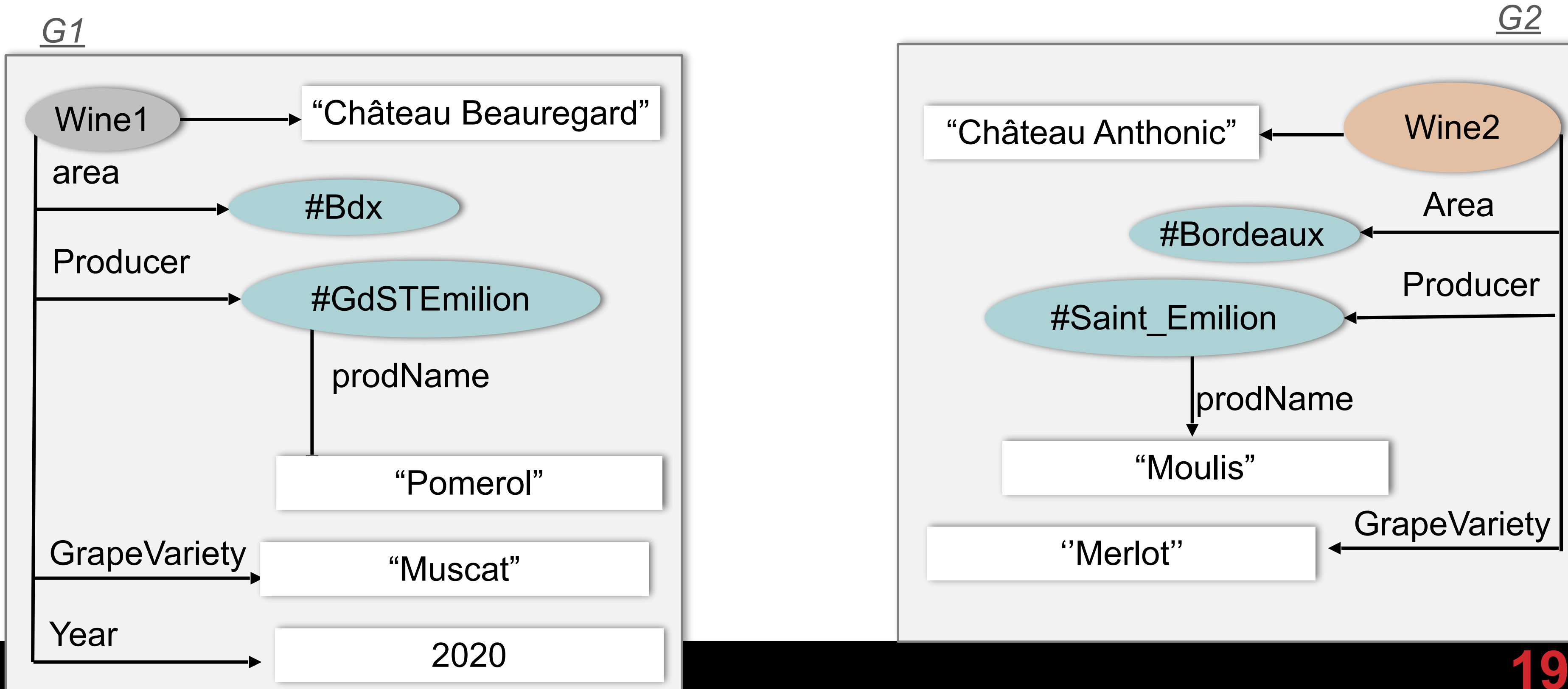
2. **Divide and Conquer Techniques (Tilde)**
   - Divide: search for a rule that is valid on a part of knowledge base
   - Conquer: recursively conquer the remaining examples by learning more rules
   - Combine the rules to form the final solution

OUTILINE

- Rule mining : techniques and main differences
- **Referring expressions : RE-miner for data linking**
- Conclusion

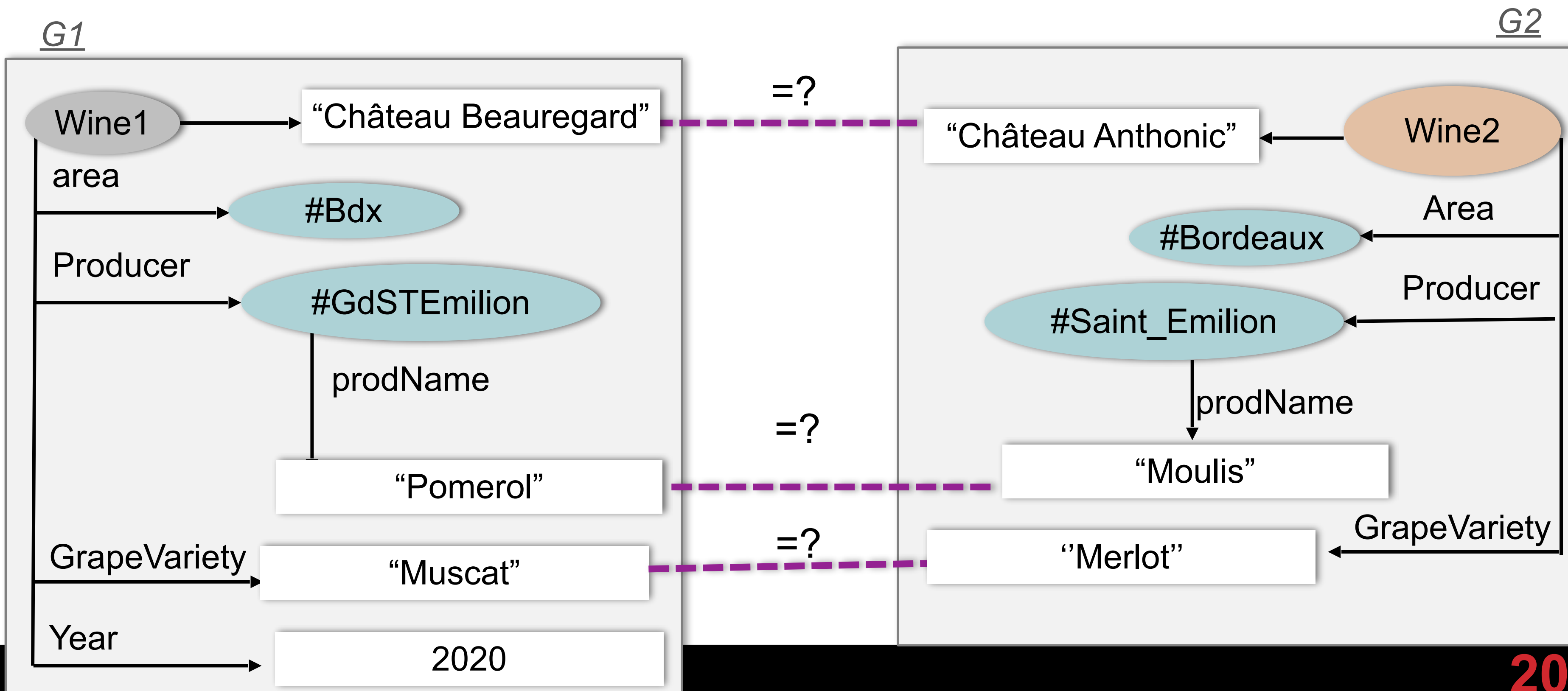
IDENTITY LINK DETECTION

- **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).
  - **Instance-based**: consider only data type properties (attributes)



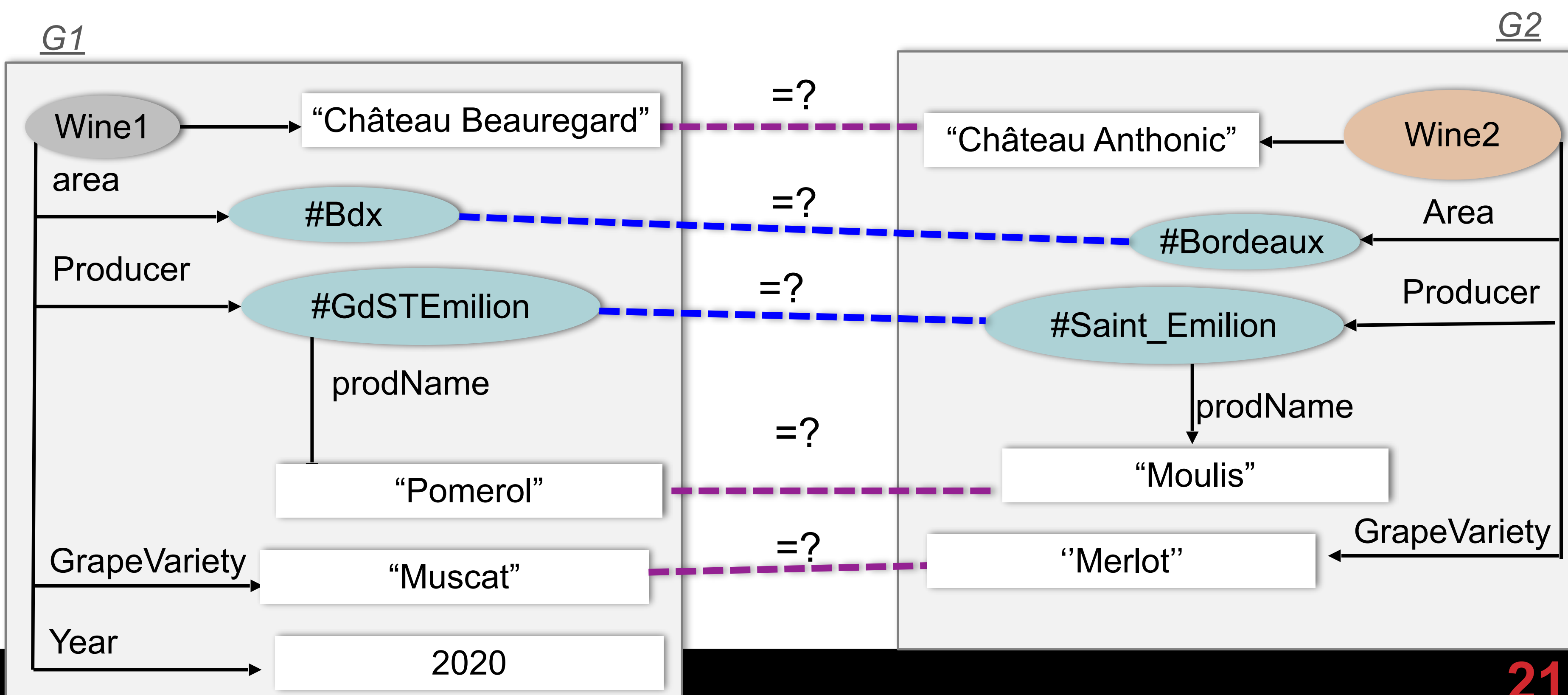
IDENTITY LINK DETECTION

- **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).
  - **Instance-based**: consider only data type properties (attributes)
    - **Similarity** on literal values



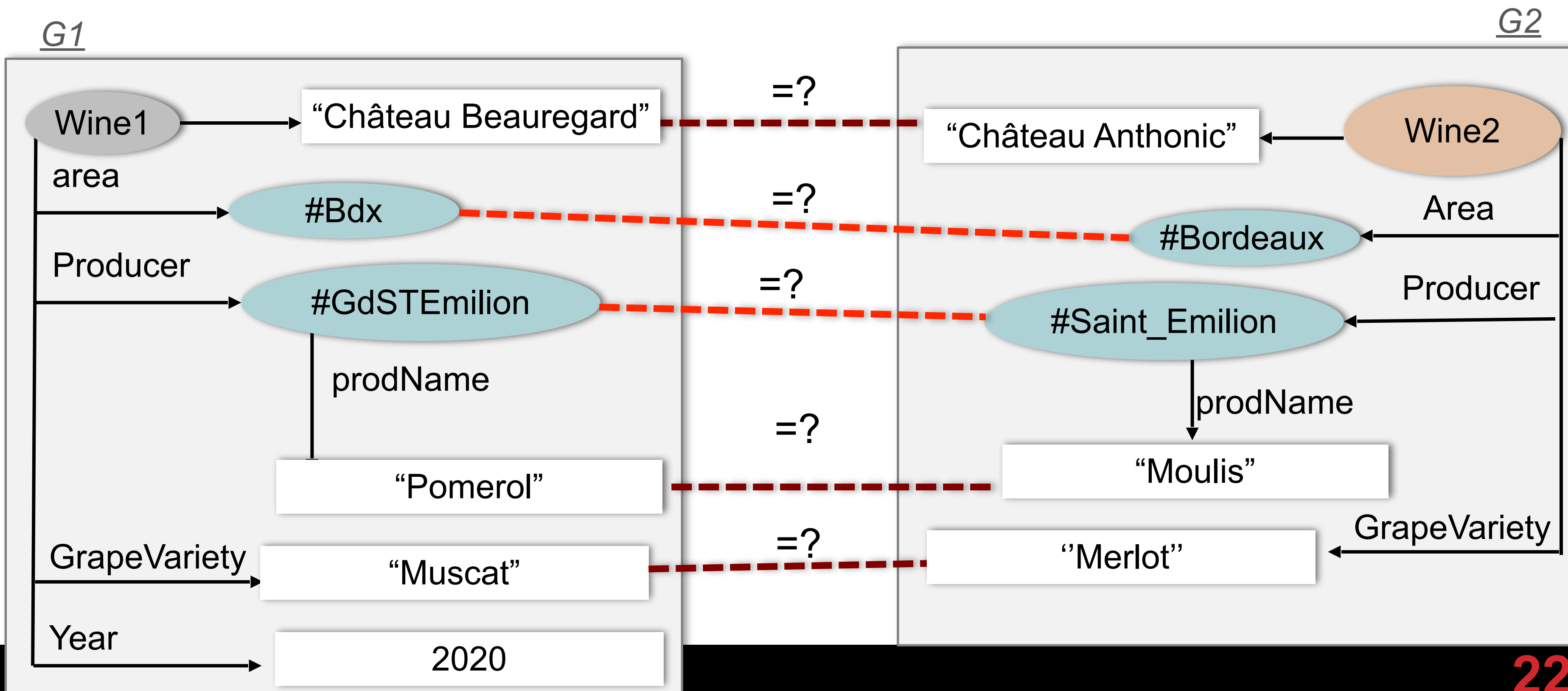
IDENTITY LINK DETECTION

- **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).
  - **Instance-based**: consider only data type properties (attributes)
  - **Graph-based**: + object properties (relations) and similarity propagation



IDENTITY LINK DETECTION

- **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).
  - **Instance-based**: consider only data type properties (attributes)
  - **Graph-based**: + object properties (relations) and similarity propagation
  - **Rule-based**: rules, area(X, Z), area(Y,Z), producer(X, W), producer(Y, W) ==> X=Y



REFERRING EXPRESSIONS DISCOVERY FOR DATA LINKING

PhD of Armita Khajeh Nassiri (2020-2023)

Co-supervised with N. Pernelle, G. Quercini

PSPC AIDA Project (2019-2023), collaboration with IBM France

WHAT IS A REFERRING EXPRESSION

Description that uniquely characterizes an instance in a given context.

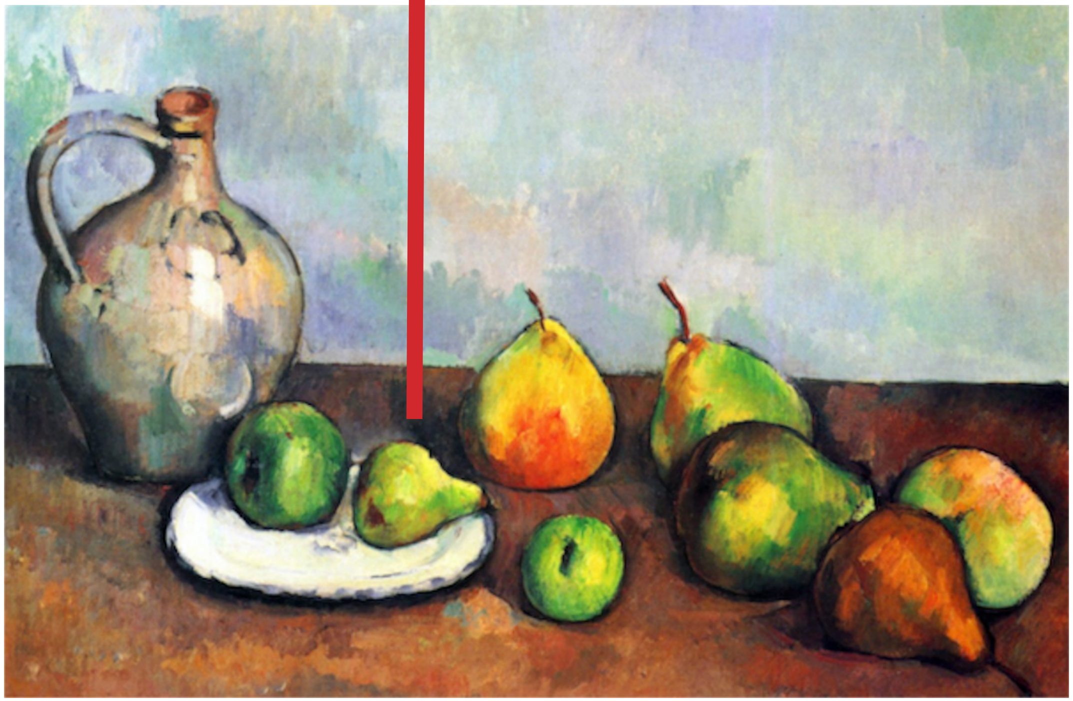
The **44th President** of **USA** → **Barack Obama**

WHAT IS A REFERRING EXPRESSION

Description that uniquely characterizes an instance in a given context.

The **44th President** of **USA** → **Barack Obama**

?



WHAT IS A REFERRING EXPRESSION

Description that uniquely characterizes an instance in a given context.

The **44th President** of **USA** → **Barack Obama**

?

The **fruit** right to the **left of** the **big orange pear**

The **second closest fruit** to the **vase**

WHAT KIND OF RES DO WE DISCOVER ?

By **instantiating** keys, we will uniquely find each instance, hence a RE.

We have already used keys for data linking.

Hence, let’s find REs from **maximal non-key** sets of a class

Example: Non-key for book:

[hasPages, yearPublished, countryOfPublication]

WHAT KIND OF RES DO WE DISCOVER ?

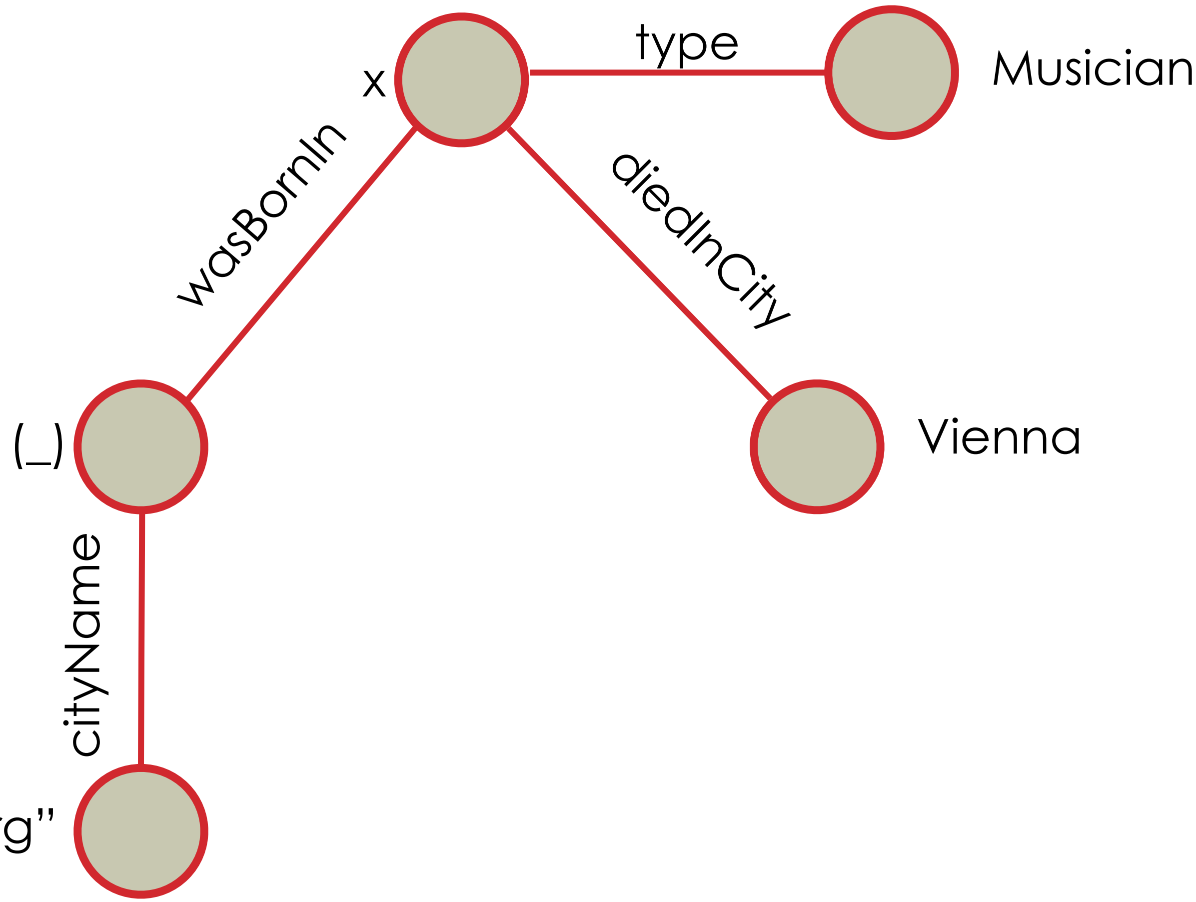
There can be many different unique expressions (REs) that identify an entity with different levels of **expressivity**.

We discover **minimal** REs that are valid in **one class** of a **knowledge graph**

A Referring expression for **u** (an instance of type **C** in knowledge graph G), is a **connected subgraph pattern** rooted by **x**

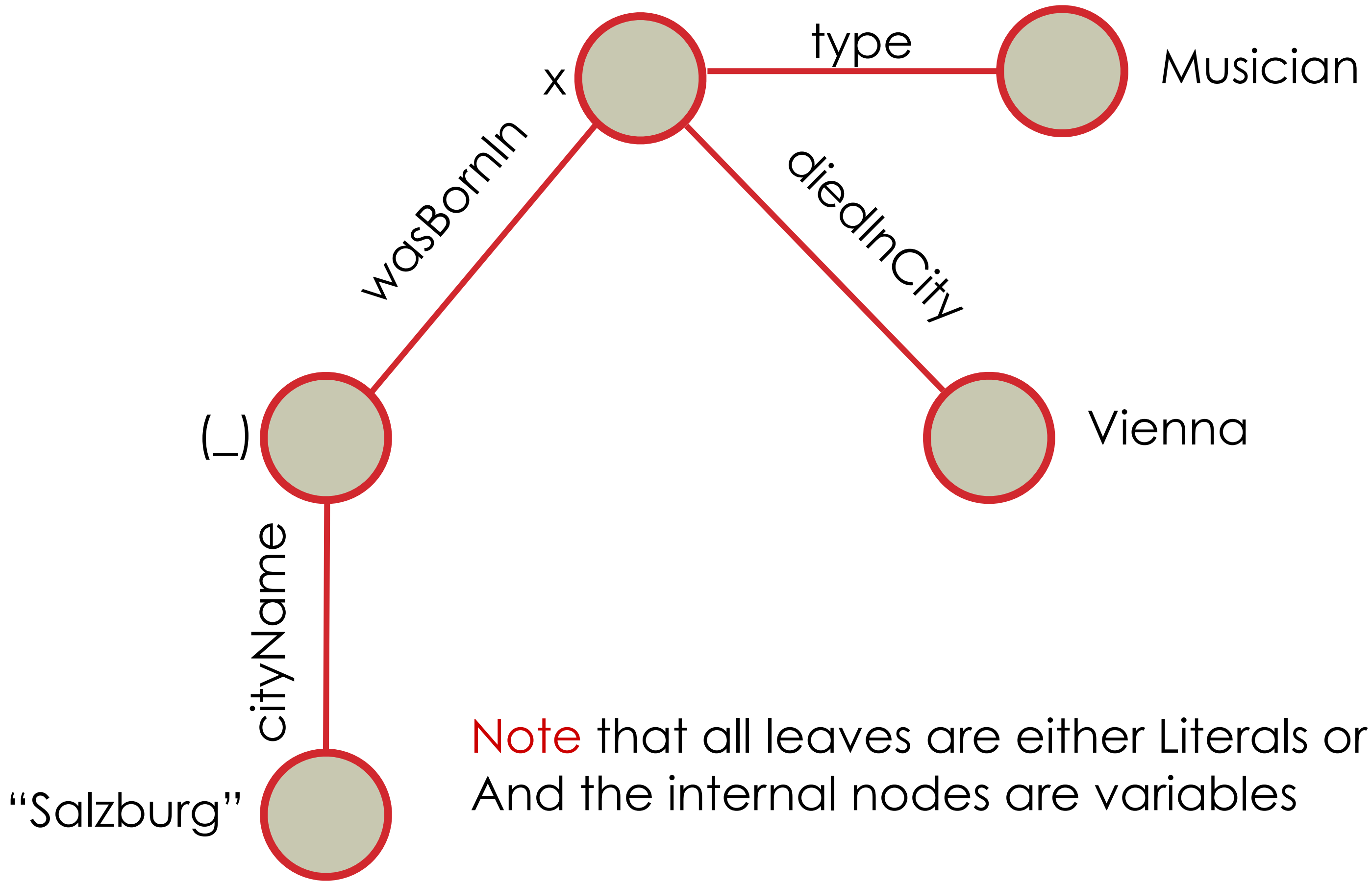
WHAT KIND OF RES DO WE DISCOVER ?

Mozart : a **musician** born in a city named **Salzburg** and died in **Vienna**.



WHAT KIND OF RES DO WE DISCOVER ?

Mozart : a **musician** born in a city named **Salzburg** and died in **Vienna**.



Note that all leaves are either Literals or URIs
And the internal nodes are variables

OVERVIEW OF RE-MINER ALGORITHM

Input: knowledge graph G, class C

Output: The set of minimal REs valid within class C

OVERVIEW OF RE-MINER ALGORITHM

Input: knowledge graph *G*, class *C*

Output: The set of minimal REs valid within class *C*

1- Find the **maximal non-key** *NK* set for *C* using **SAKey**

2- Group *NK* based on cardinality

OVERVIEW OF RE-MINER ALGORITHM

Input: knowledge graph *G*, class *C*

Output: The set of minimal REs valid within class *C*

1- Find the **maximal non-key** *NK* set for *C* using **SAKey**

2- Group *NK* based on cardinality

Example for class book:

NK= [{author, year}, {publisher, author, language}]

Grouped based on Cardinality:

Level1: [{author}, {year}, {publisher}, {language}]
Level2: [{author, year}, {publisher, author}, {publisher,language}, {author, language}]
Level3: [{publisher, author, language}]

OVERVIEW OF RE-MINER ALGORITHM

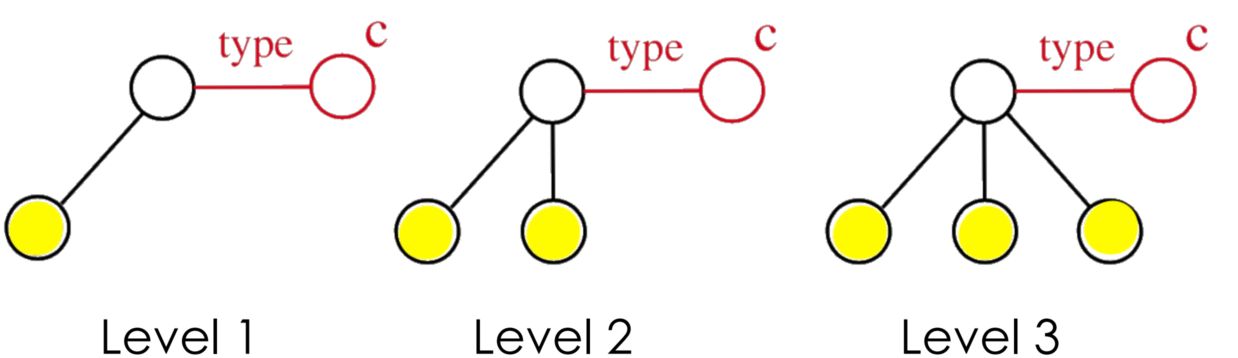
Input: knowledge graph *G*, class *C*

Output: The set of minimal REs valid within class *C*

1- Find the **maximal non-key** *NK* set for *C* using **SAKey**

2- Group *NK* based on cardinality

3- Starting from level 1, construct **candidate REs** and keep those that are valid



OVERVIEW OF RE-MINER ALGORITHM

Input: knowledge graph *G*, class *C*

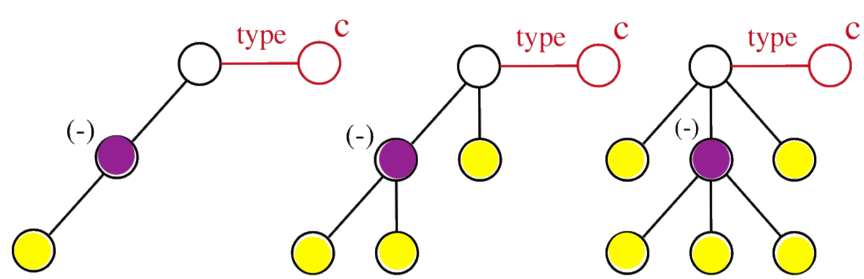
Output: The set of minimal REs valid within class *C*

1- Find the **maximal non-key** *NK* set for *C* using **SAKey**

2- Group *NK* based on cardinality

3- Starting from level 1, construct **candidate REs** and keep those that are valid

4- Increase the depth of subgraph following the same procedure for the class of the entity we aim to replace with an existential quantifier.



OVERVIEW OF RE-MINER ALGORITHM

Input: knowledge graph *G*, class *C*

Output: The set of minimal REs valid within class *C*

1- Find the **maximal non-key** *NK* set for *C* using **SAKey**

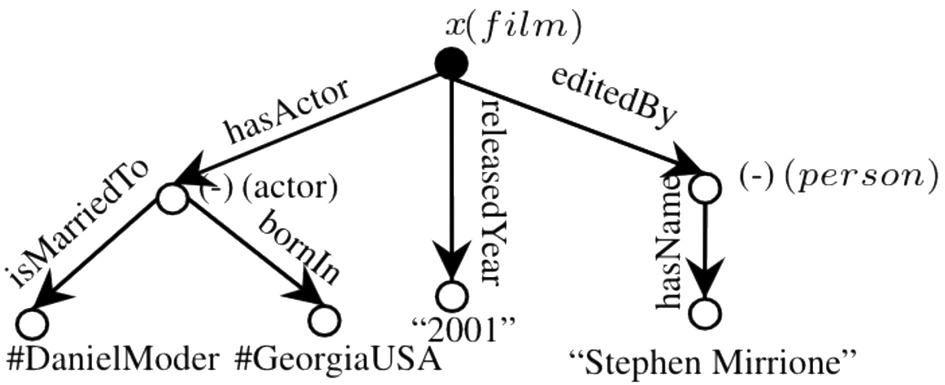
2- Group *NK* based on cardinality

3- Starting from level 1, construct **candidate REs** and keep those that are valid

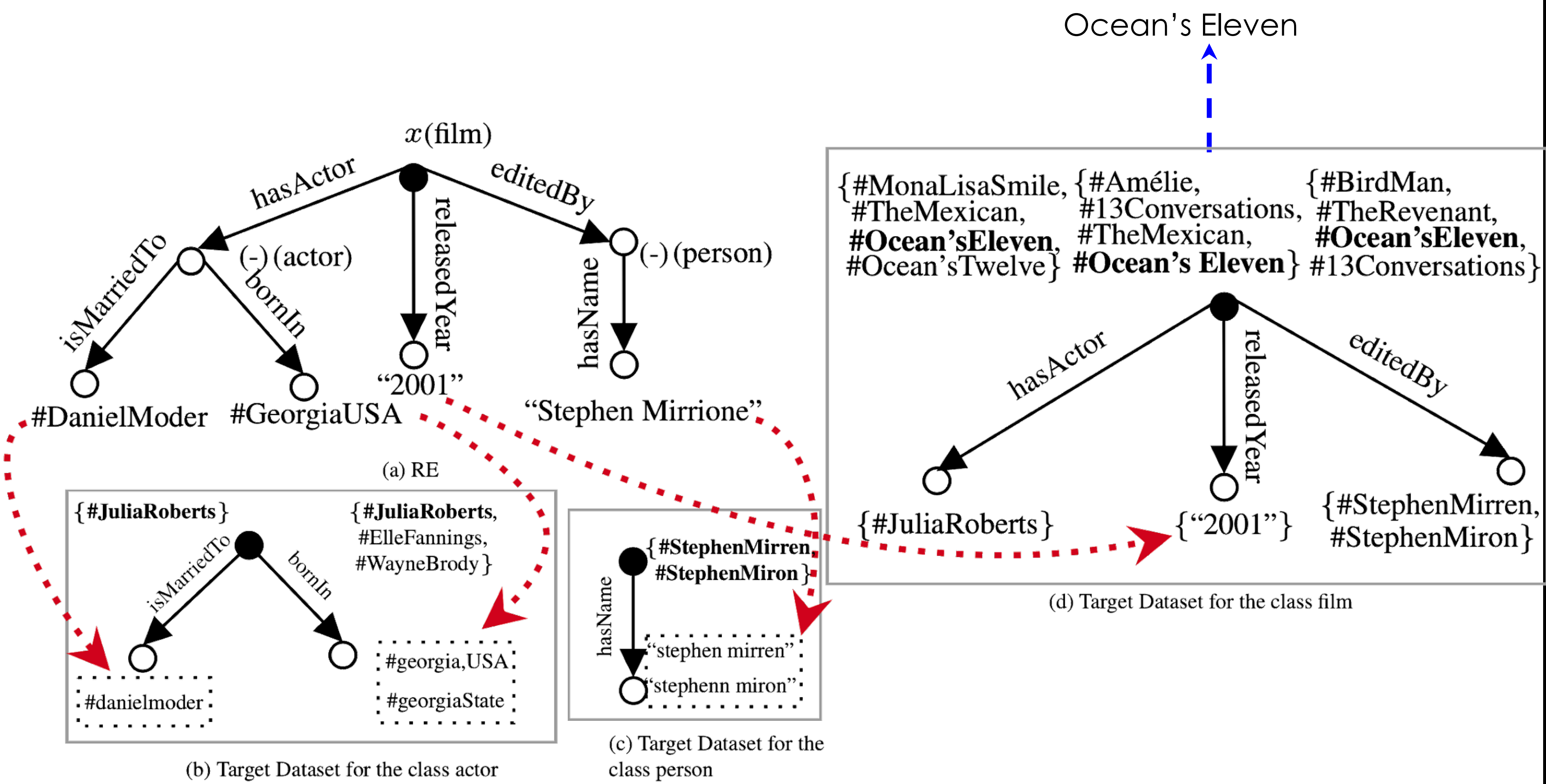
4- Increase the depth of subgraph following the same procedure for the class of the entity we aim to replace with an existential quantifier.

5- A **post-processing** step that recursively replaces **IRIs** in REs with an instantiation of minimal key properties to discover **extended REs.**

DATA LINKING WITH REs



DATA LINKING WITH REs



DATA LINKING WITH REs: experiments

Dataset for **10 classes** of **YAGO** and **DBpedia** used in VICKEY experimentations, we only consider mapped properties, and depth = 1.

class	#Triples	#Properties	#NKs	#REs	Run time
Actor	514.7 K	16	69	725.6 K	95.1 s
Album	381.1 K	5	2	212.1 K	14.7 s
Book	92.5 K	7	6	66.3 K	3.5 s
Film	533.5 K	9	7	690.9 K	102.3 s
Mountain	116.7 K	6	4	59.2 K	1.4 s
Museum	81.6 K	7	5	53.5 K	2.6 s
Organization	**2.2 M**	17	43	**68.3 M**	**3.48 h**
Scientist	335.6 K	18	92	309.9 K	64.0 s
University	131.8 K	9	9	161.8 K	17.7 s
City	1.1 M	17	29	1.2 M	109.7 s

DATA LINKING WITH REs: experiments

Dataset for **10 classes** of **YAGO** and **DBpedia** like the one used in VICKEY.

Hypothesis: If a description uniquely identifies an entity in one KG, it's likely that the same description identifies the same entity in the other KG.

For each RE of an entity in YAGO, if the description is fulfilled by only one entity in DBPedia, we will link the two.
(some Res are discarded)

DATA LINKING WITH REs: experiments

Linking results with keys, keys + conditional keys and REs
To compare literal values : only string equality !

Class	Recall			Precision			F1		
	Ks	Ks+CKs	RE	Ks	Ks+CKs	RE	Ks	Ks+CKs	RE
Actor	0.27	0.60	**0.66**	0.99	0.99	0.99	0.43	0.75	**0.79**
Album	0.00	0.15	**0.64**	1.00	0.99	0.98	0.00	0.26	**0.77**
Book	0.03	0.13	**0.77**	1.00	0.99	0.97	0.06	0.23	**0.86**
Film	0.04	0.39	**0.73**	0.99	0.98	0.94	0.08	0.55	**0.82**
Mountain	0.00	0.29	**0.77**	1.00	0.99	0.98	0.00	0.45	**0.86**
University	0.09	0.25	**0.65**	0.99	0.99	0.98	0.16	0.40	**0.78**

DATA LINKING WITH REs: experiments

Linking results with keys, keys + conditional keys and REs
To compare literal values : only string equality !

Class	Recall			Precision			F1		
	Ks	Ks+CKs	RE	Ks	Ks+CKs	RE	Ks	Ks+CKs	RE
Actor	0.27	0.60	**0.66**	0.99	0.99	0.99	0.43	0.75	**0.79**
Album	0.00	0.15	**0.64**	1.00	0.99	0.98	0.00	0.26	**0.77**
Book	0.03	0.13	**0.77**	1.00	0.99	0.97	0.06	0.23	**0.86**
Film	0.04	0.39	**0.73**	0.99	0.98	0.94	0.08	0.55	**0.82**
Mountain	0.00	0.29	**0.77**	1.00	0.99	0.98	0.00	0.45	**0.86**
University	0.09	0.25	**0.65**	0.99	0.99	0.98	0.16	0.40	**0.78**

Published **at ISWC'2020**, **ranked 1st** for SPIMBENCH at **IM-OAEI 2020**.

CONCLUSION

RULE MINING FUTURE CHALLENGES

- Deal with **numerical values**: discretisation, domain expert, combination of ML and symbolic AI

- Rule mining for **decision making** : AIDA with IBM

- Rule mining for **explanation** : causality to explain the impact of climate change on Maïs development

- Scalability (ex. AMIE3 timeout for NB atoms > 4)

DATA LINKING FUTURE CHALLENGES

- Multi-source and simultaneous schema/data linking
- Scalability
- Link invalidation

REFERRING EXPRESSION DISCOVERY FOR DATA LINKING

FATIHA SAÏS, joint work with,

Armita Khajeh Nassiri, Nathalie Pernelle and Gianluca Quercini

SÉMINAIRE RÉSIDENTIEL INRAE
SEMANTIC LINKED DATA

11-14 OCT. 2021- DOMAINE DU LAZARET - SÈTE