université PARIS-SACLAY

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

ETIS

david
données et algorithmes
pour une ville intelligente et durable

Approaches Towards Unified Models for Integrating Web Knowledge Bases

Maria Koutraki

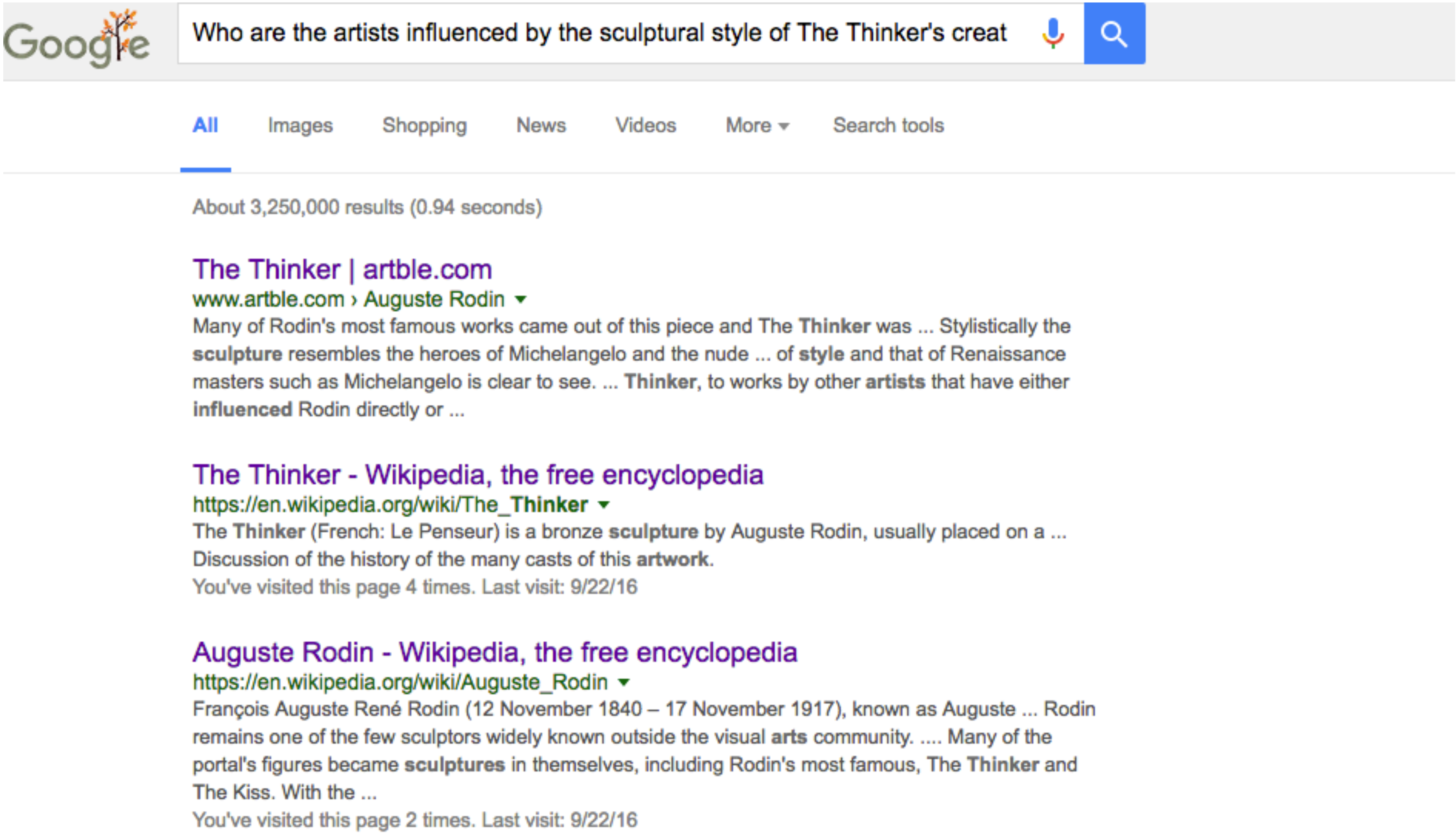
Joint work with: Nicoleta Preda, Dan Vodislav

Paris, 26/10/2016

Who are the artists influenced by the sculptural style of “The Thinker” 's creator?

Motivation – Unstructured Data

Q1: Who are the artists influenced by the sculptural style of The Thinker's creator?



Who are the artists influenced by the sculptural style of The Thinker's creat

About 3,250,000 results (0.94 seconds)

The Thinker | artble.com

www.artble.com › Auguste Rodin

Many of Rodin's most famous works came out of this piece and The **Thinker** was ... Stylistically the **sculpture** resembles the heroes of Michelangelo and the nude ... of **style** and that of Renaissance masters such as Michelangelo is clear to see. ... **Thinker**, to works by other **artists** that have either **influenced** Rodin directly or ...

The Thinker - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/The_**Thinker**

The **Thinker** (French: Le Penseur) is a bronze **sculpture** by Auguste Rodin, usually placed on a ... Discussion of the history of the many casts of this **artwork**.

You've visited this page 4 times. Last visit: 9/22/16

Auguste Rodin - Wikipedia, the free encyclopedia

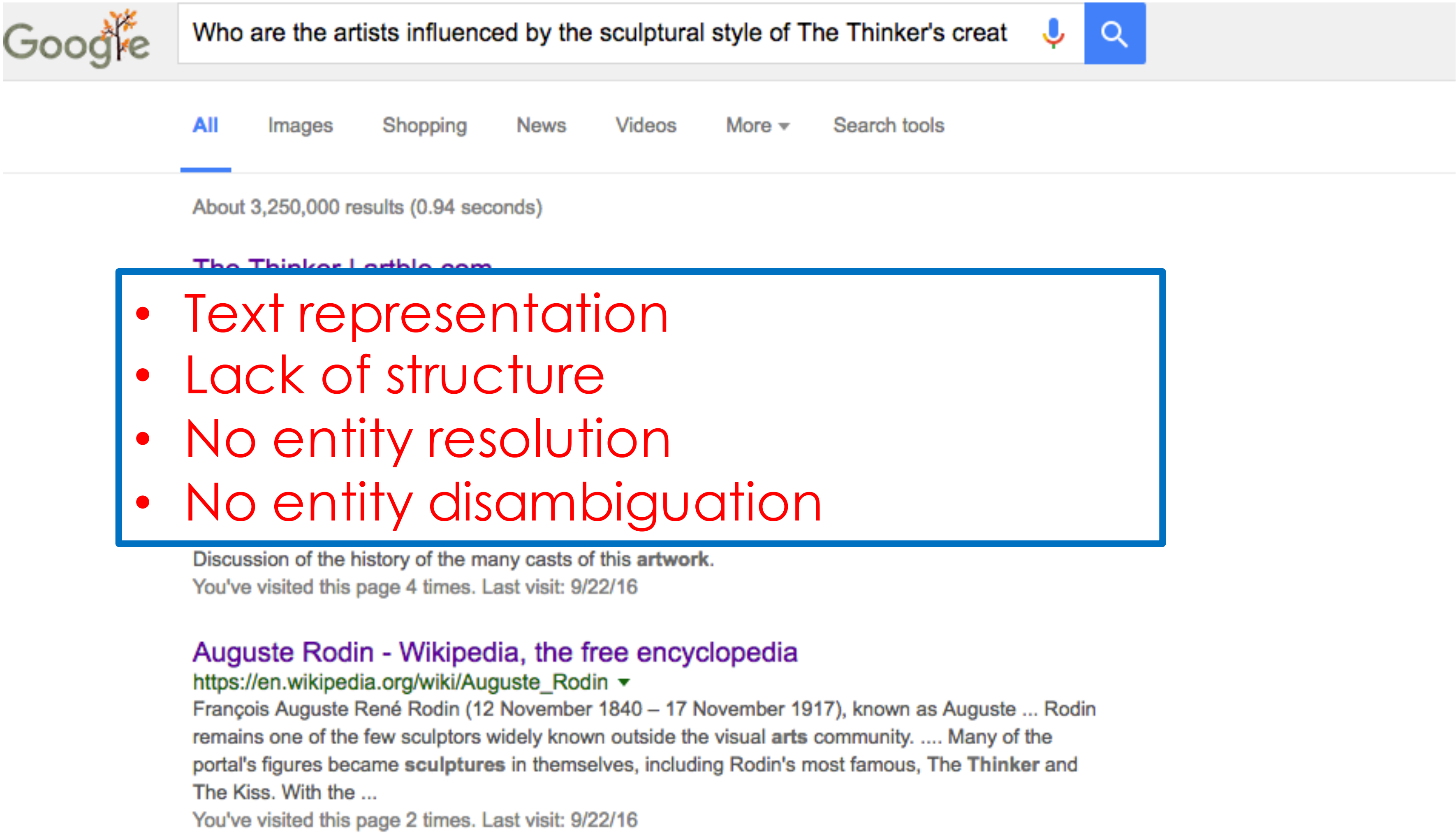
https://en.wikipedia.org/wiki/Auguste_Rodin

François Auguste René Rodin (12 November 1840 – 17 November 1917), known as Auguste ... Rodin remains one of the few sculptors widely known outside the visual **arts** community. .... Many of the portal's figures became **sculptures** in themselves, including Rodin's most famous, The **Thinker** and The Kiss. With the ...

You've visited this page 2 times. Last visit: 9/22/16

Motivation – Unstructured Data

Q1: Who are the artists influenced by the sculptural style of The Thinker's creator?



Google
Who are the artists influenced by the sculptural style of The Thinker's creat
All
Images
Shopping
News
Videos
More ▾
Search tools

About 3,250,000 results (0.94 seconds)

The Thinker | artble.com

Discussion of the history of the many casts of this **artwork**.
You've visited this page 4 times. Last visit: 9/22/16

Auguste Rodin - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Auguste_Rodin ▾
François Auguste René Rodin (12 November 1840 – 17 November 1917), known as Auguste ... Rodin remains one of the few sculptors widely known outside the visual **arts** community. .... Many of the portal's figures became **sculptures** in themselves, including Rodin's most famous, The **Thinker** and The Kiss. With the ...
You've visited this page 2 times. Last visit: 9/22/16

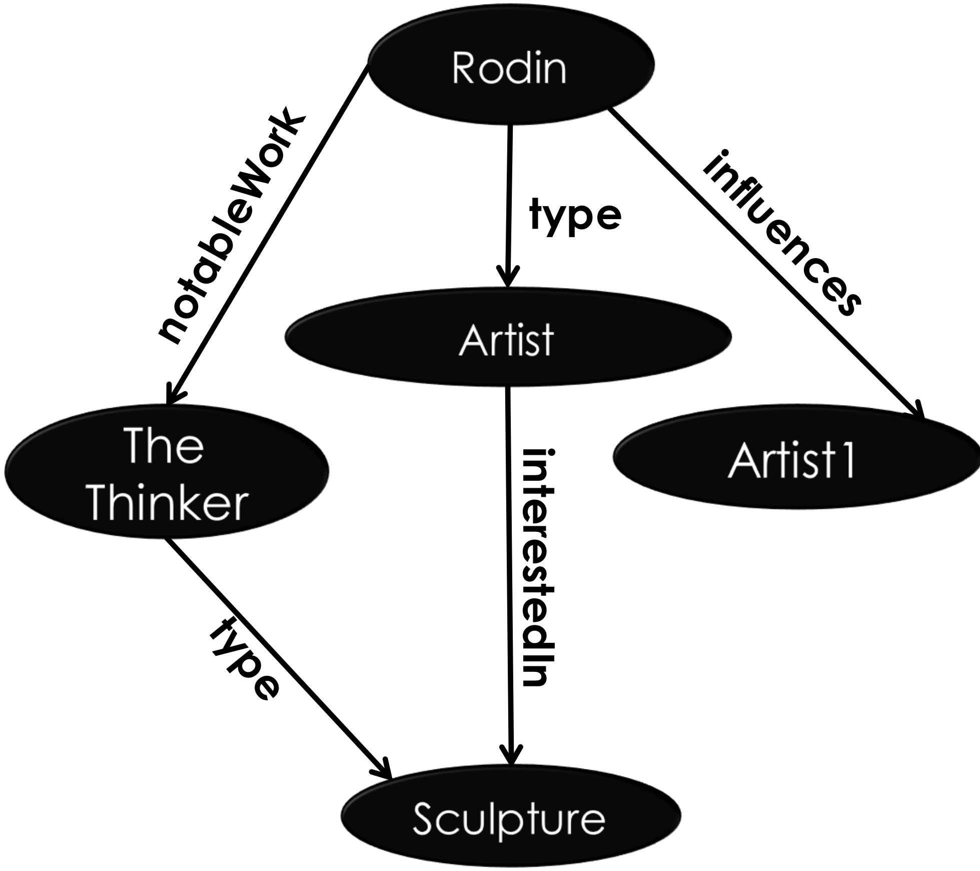
- Text representation
- Lack of structure
- No entity resolution
- No entity disambiguation

Motivation – Structured Data

What is structured data?

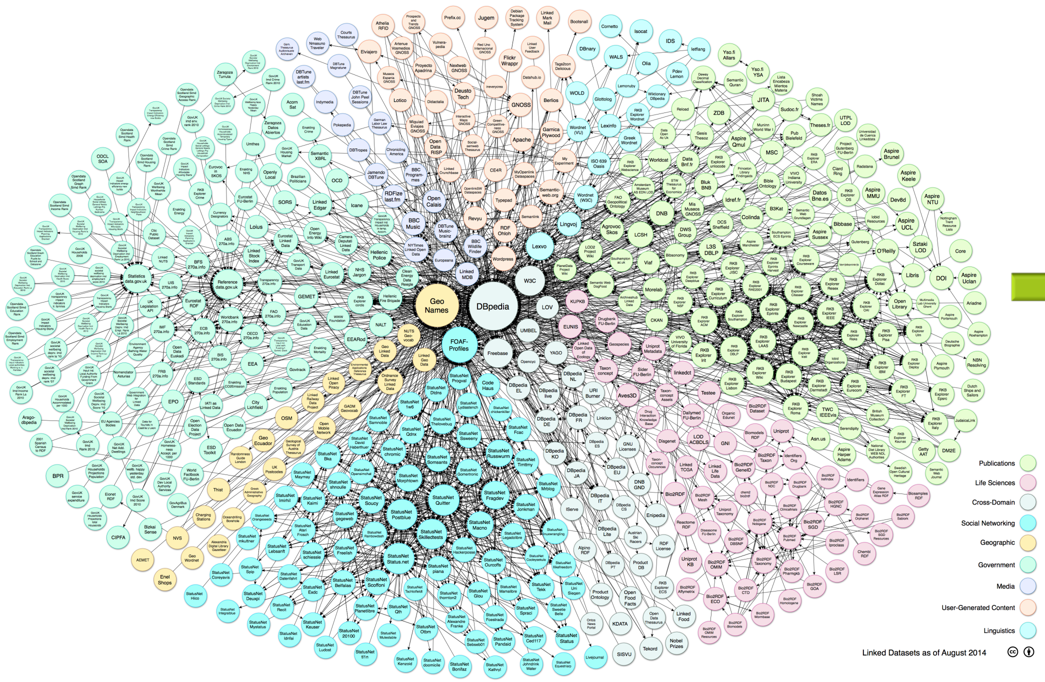
- •RDF – Resource Description Framework
- •W3C standard for describing web resources
- •Triple = statement of the form **(subject, property, object)**

Subject	Property	Object
Rodin	type	Artist
Artist	interestedIn	Sculpture
Rodin	notableWork	The Thinker
The Thinker	type	Sculpture
Rodin	influences	Artist1



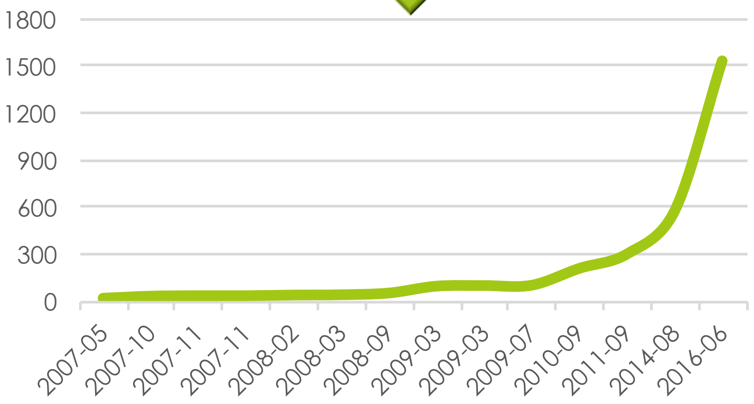
Motivation – Structured Data

Linked Open Data Cloud



Domains

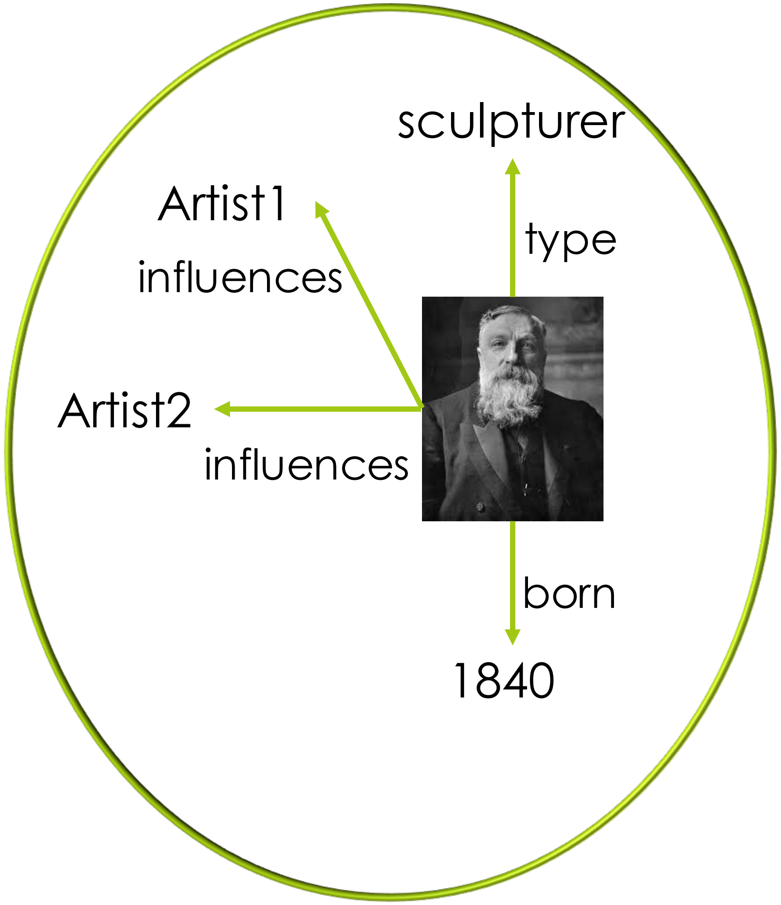
Topic	%
Government	18.05%
Publications	9.47%
Life Sciences	8.19%
User-generated content	4.73%
Cross-domain	4.04%
Media	2.17%
Geographic	2.07%
Social Web	51.28%



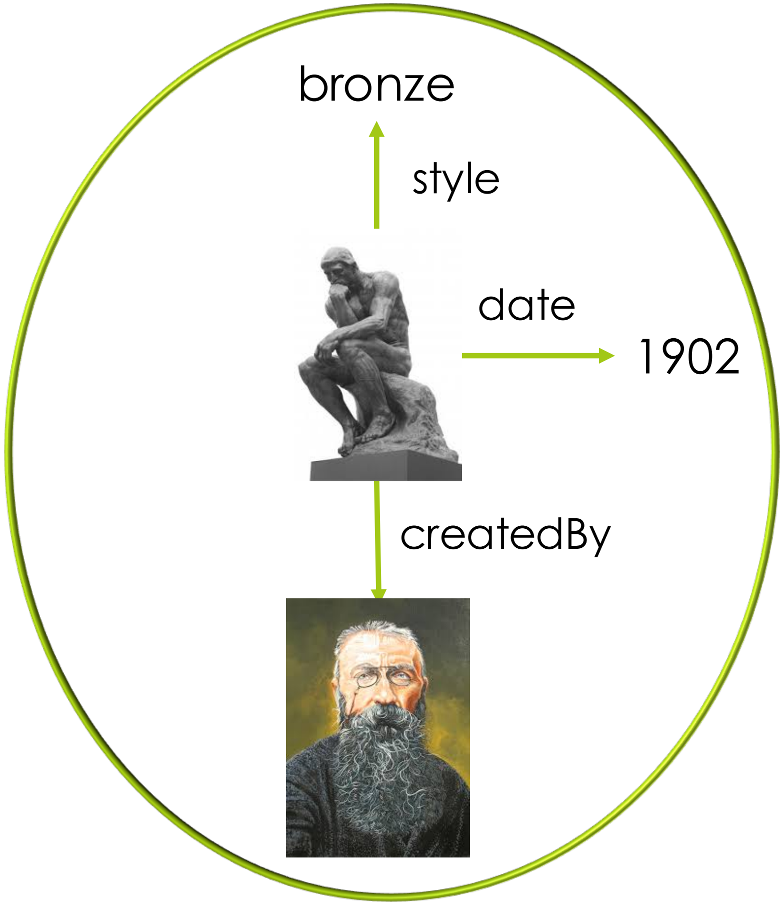
- Exponential increase of datasets and triples
- > 30 billion triples
- Automatically constructed KBs

Motivation – Structured Data

Q1: Who are the artists influenced by the sculptural style of The Thinker's creator?



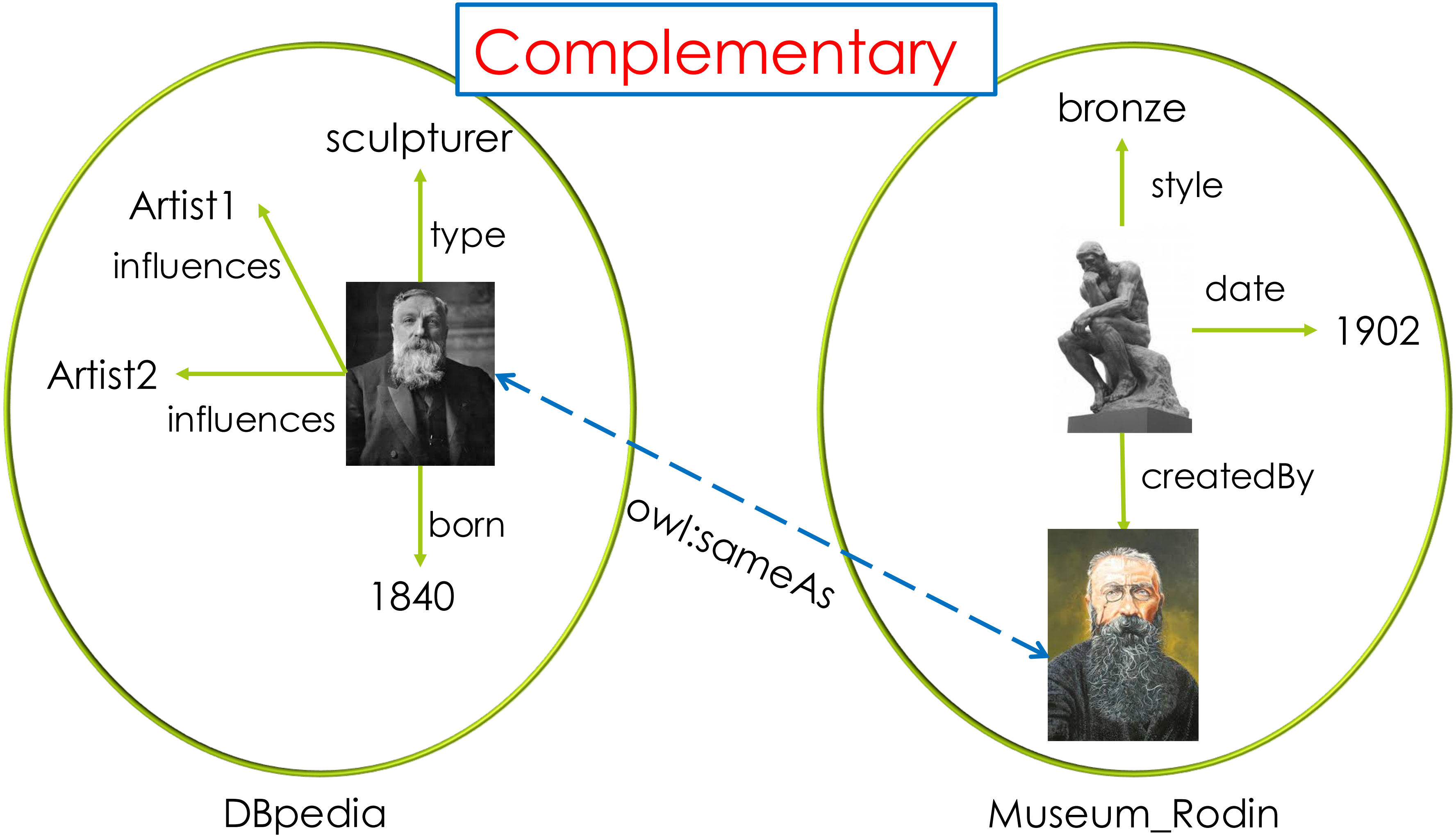
DBpedia



Museum_Rodin

Motivation – Structured Data

Q1: Who are the artists influenced by the sculptural style of The Thinker's creator?



Complementary

sculpturer

type

Artist1

influences

Artist2

influences

born

1840

bronze

style

date

1902

createdBy

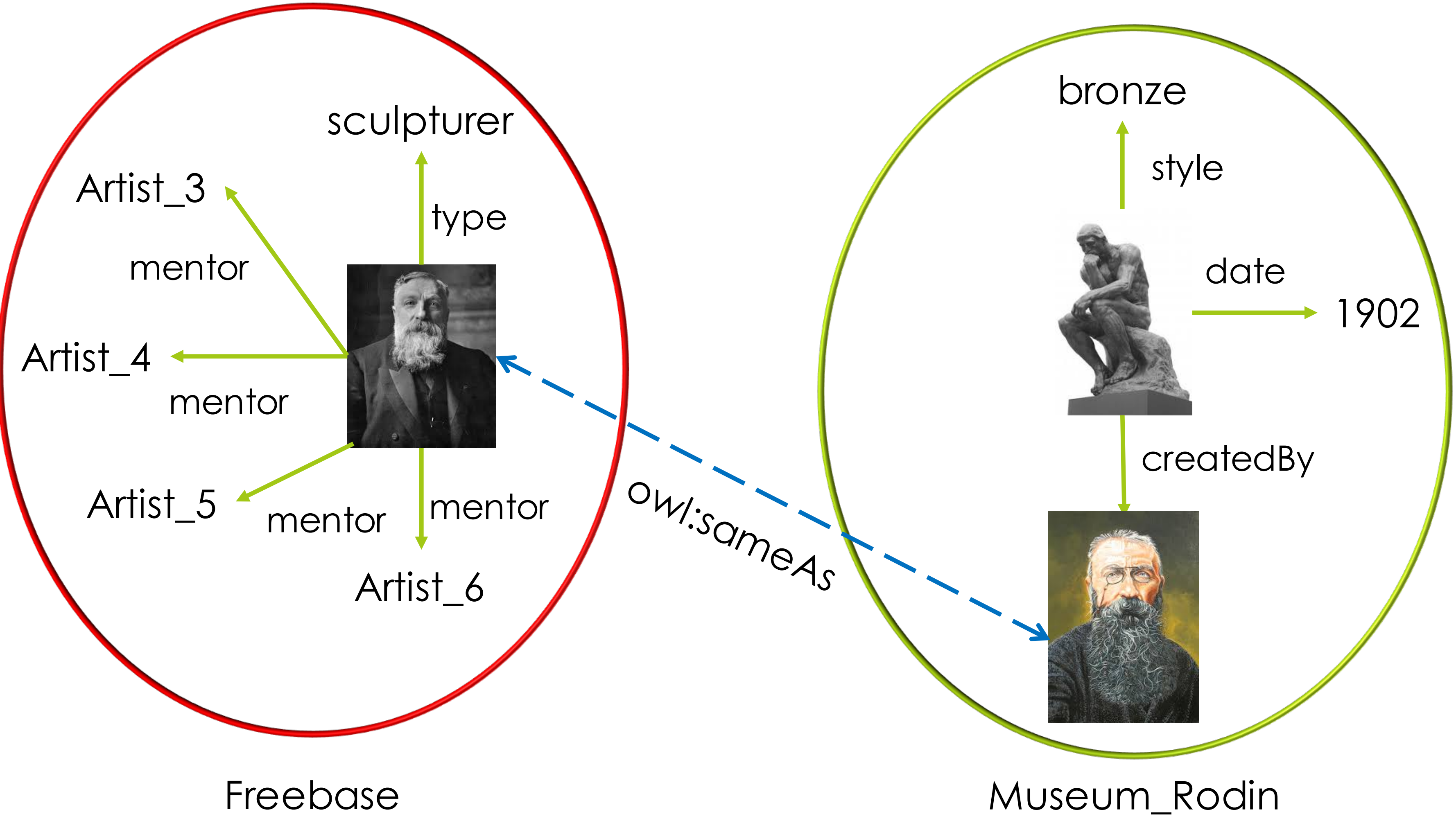
owl:sameAs

DBpedia

Museum_Rodin

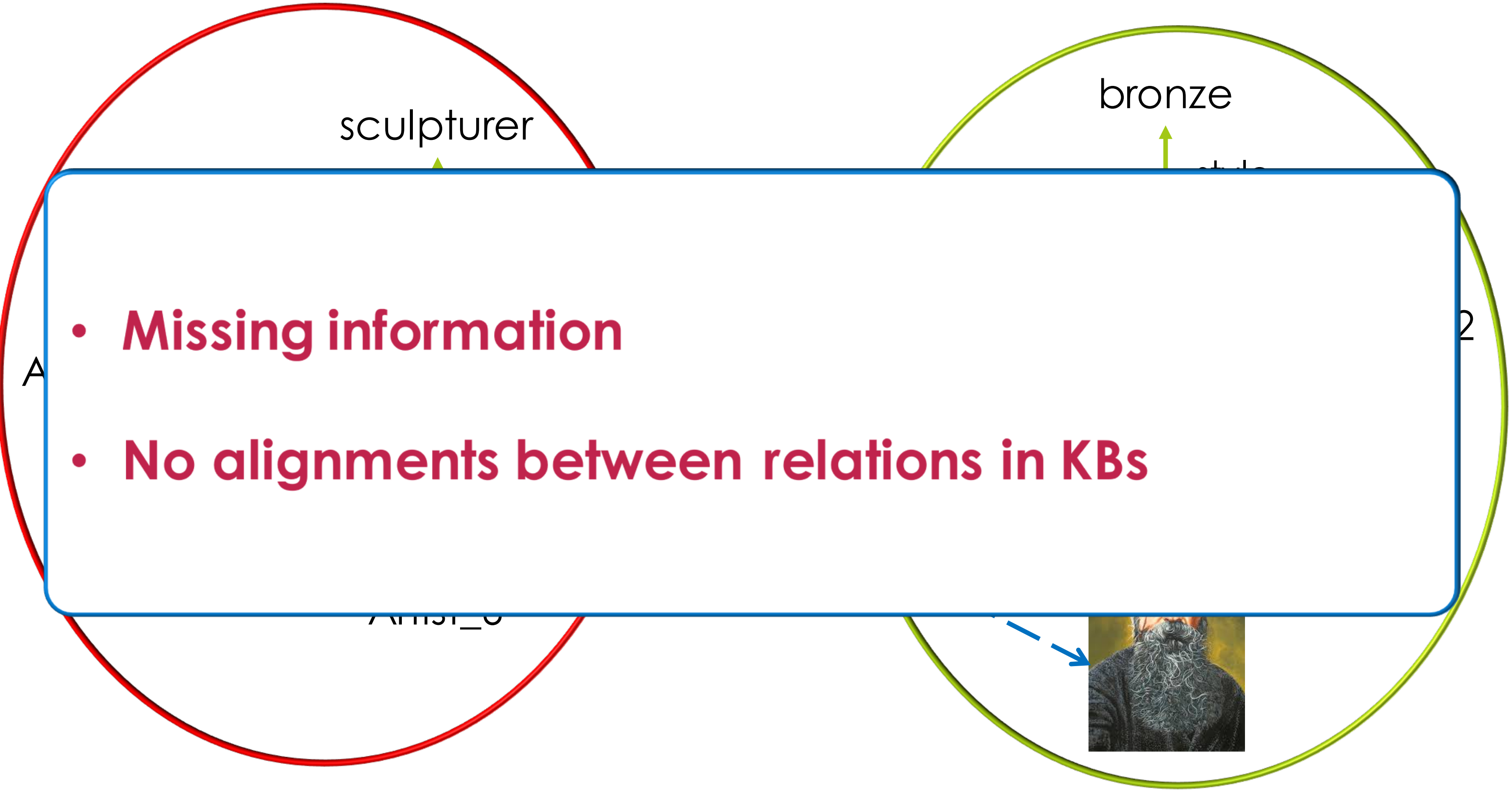
Motivation – Structured Data

Q1: Who are the artists influenced by the sculptural style of The Thinker's creator?



Motivation – Structured Data

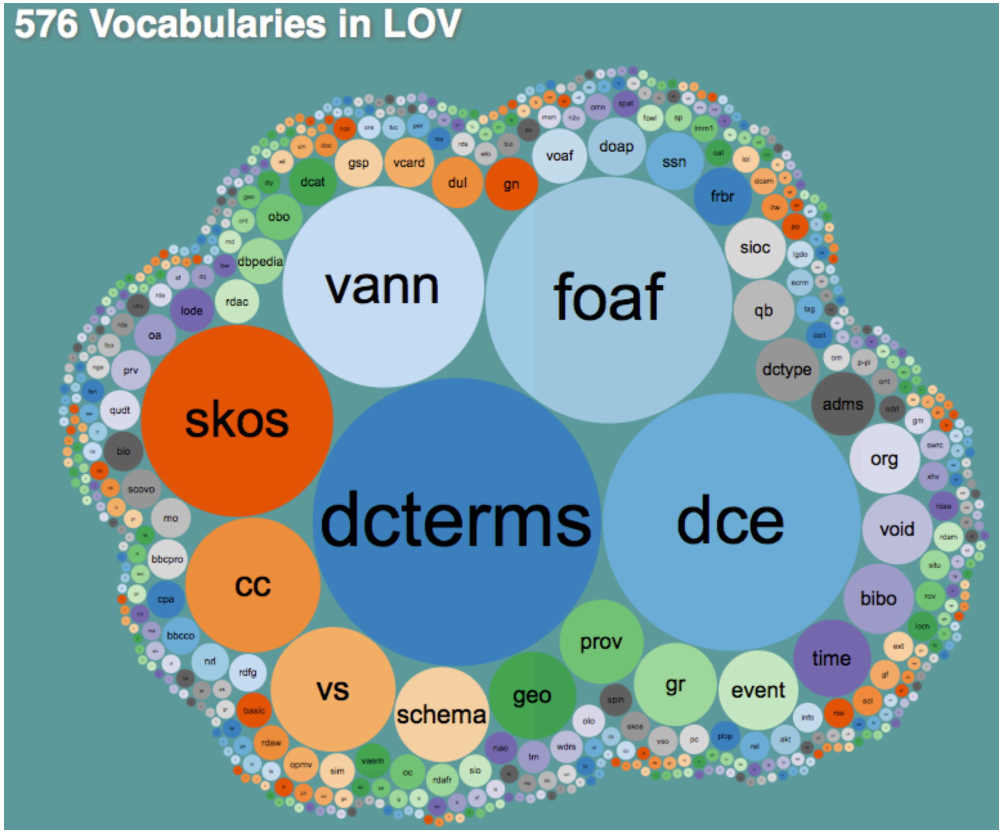
Q1: Who are the artists influenced by the sculptural style of The Thinker's creator?



- **Missing information**
- **No alignments between relations in KBs**

Motivation – Structured Data

Diverse schemas for representation in LOD



- ~576 schemas/vocabularies used for representation
- Diverse quality of schemas[1]
- Duplicate representation of similar concepts/classes and relations
- Lack of explicit alignment between classes/relations (with only up to 2%)[2]

[1] Aimilia Magkanaraki, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis: Benchmarking RDF Schemas for the Semantic Web. International Semantic Web Conference 2002: 132-146

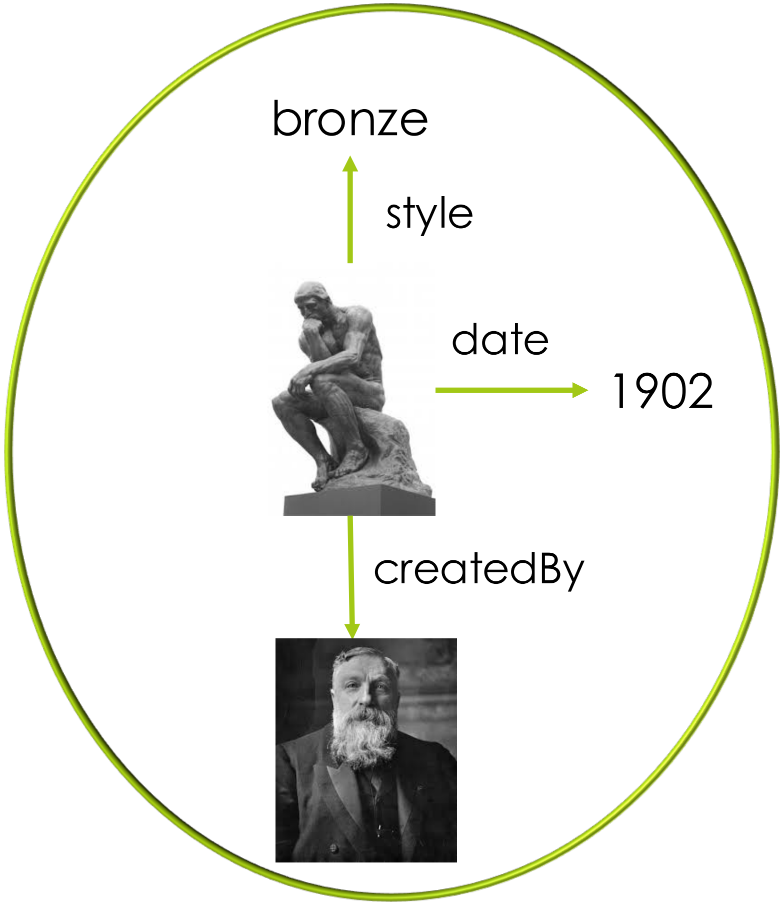
[2] Max Schmachtenberg, Christian Bizer, Heiko Paulheim: Adoption of the Linked Data Best Practices in Different Topical Domains. International Semantic Web Conference (1) 2014: 245-260

Motivation – Web services

Q2: Which are the museums that hold sculptures similar to The Thinker and have open exhibitions in Paris?

Motivation – Web services

Q2: Which are the museums that hold sculptures similar to The Thinker and have open exhibitions in Paris?



bronze

style

date

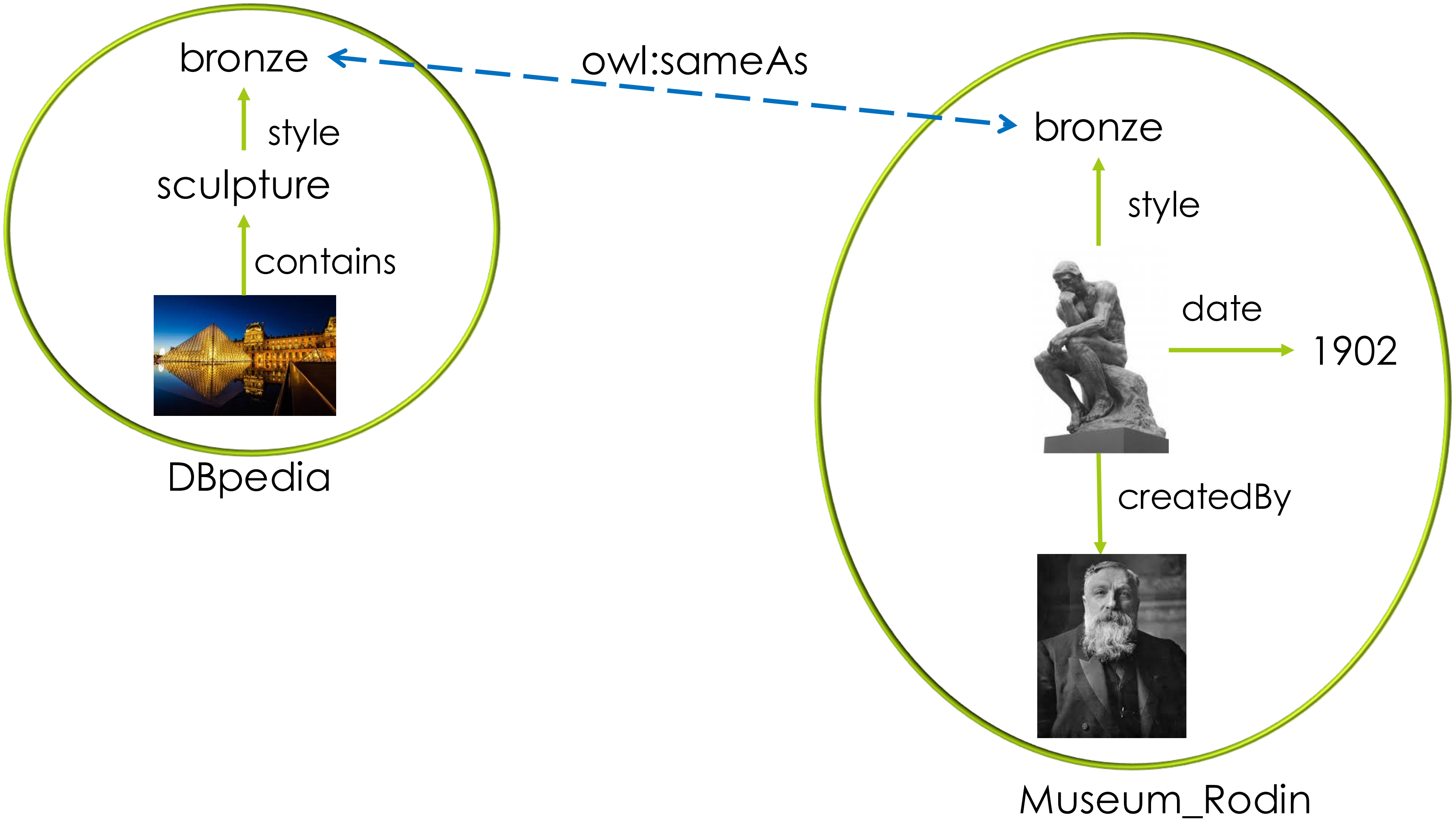
1902

createdBy

Museum_Rodin

Motivation – Web services

Q2: Which are the museums that hold sculptures similar to The Thinker and have open exhibitions in Paris?



bronze

style

sculpture

contains

DBpedia

owl:sameAs

bronze

style

date

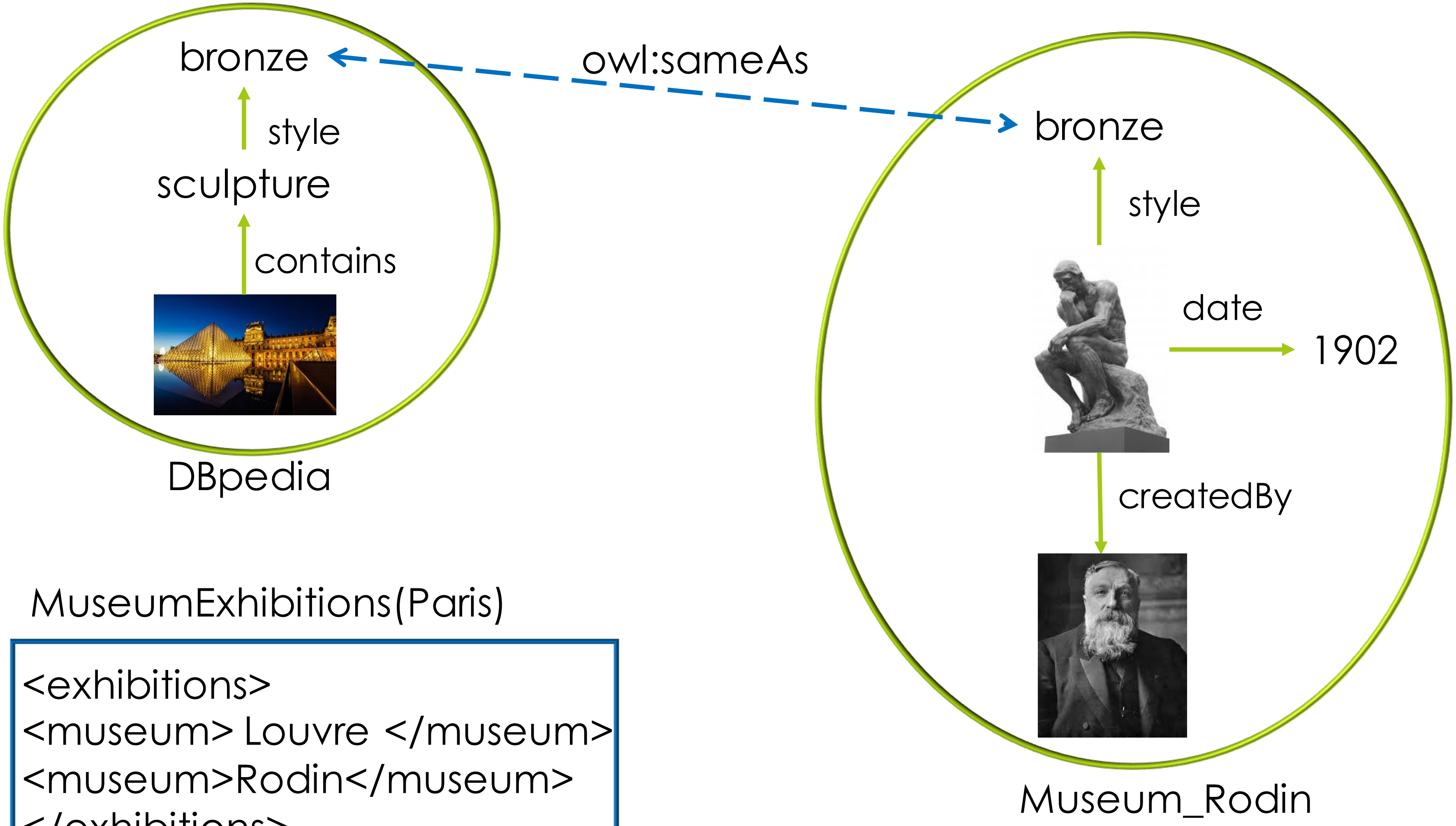
1902

createdBy

Museum_Rodin

Motivation – Web services

Q2: Which are the museums that hold sculptures similar to The Thinker and have open exhibitions in Paris?

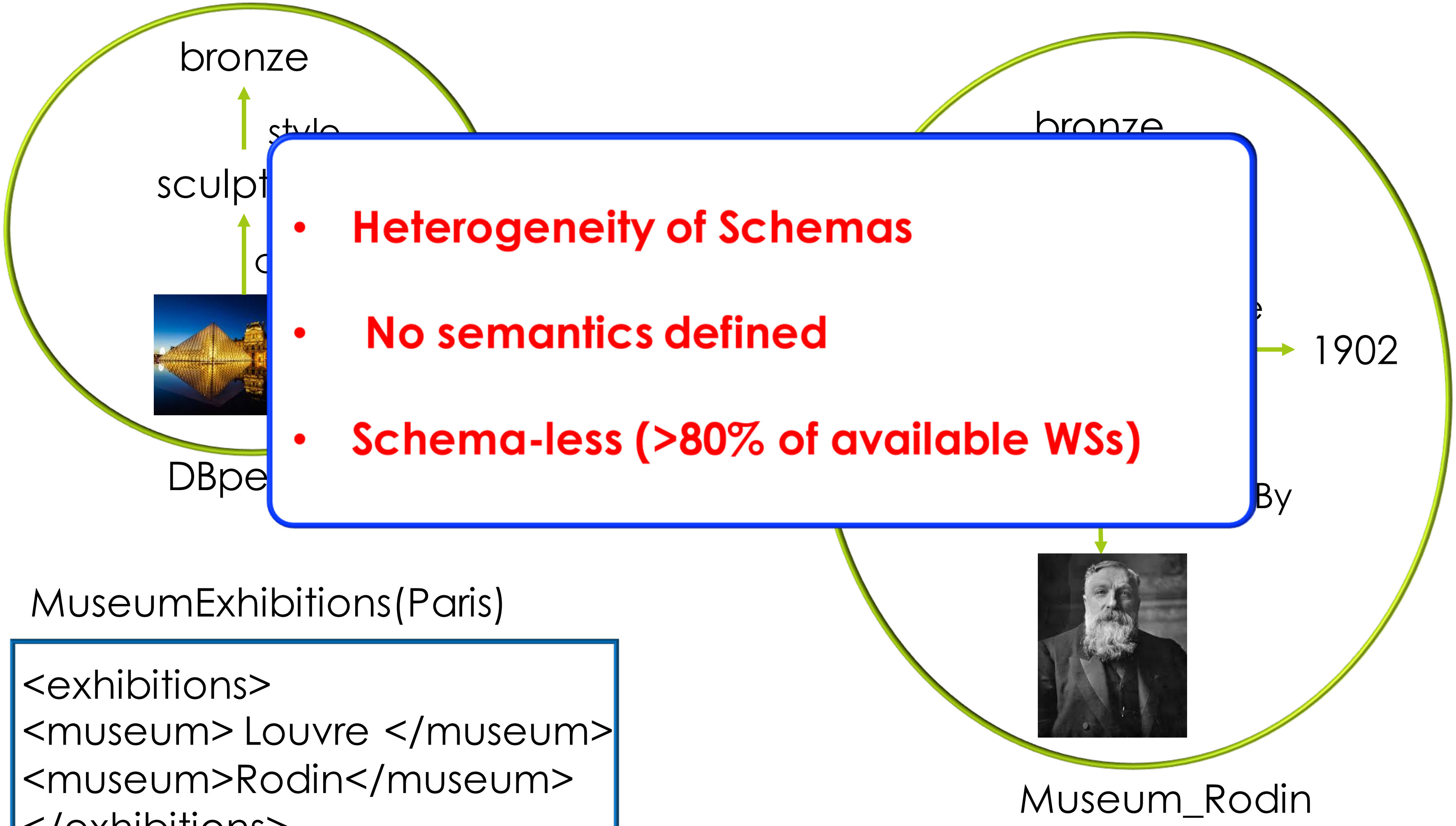


MuseumExhibitions(Paris)

```
<exhibitions>
<museum> Louvre </museum>
<museum>Rodin</museum>
</exhibitions>
```

Motivation – Web services

Q2: Which are the museums that hold sculptures similar to The Thinker and have open exhibitions in Paris?



- **Heterogeneity of Schemas**
- **No semantics defined**
- **Schema-less (>80% of available WSs)**

MuseumExhibitions(Paris)

```
<exhibitions>
<museum> Louvre </museum>
<museum>Rodin</museum>
</exhibitions>
```

Motivation – Web services

More than 12000 APIs* from various domains:

- Search (3200 APIs)

  Google tripadvisor

- Social (3000 APIs)

  facebook twitter foursquare

- Traveling (1200 APIs)

  Google Maps GeoNames Expedia Affiliate Network

- Music (1000 APIs)

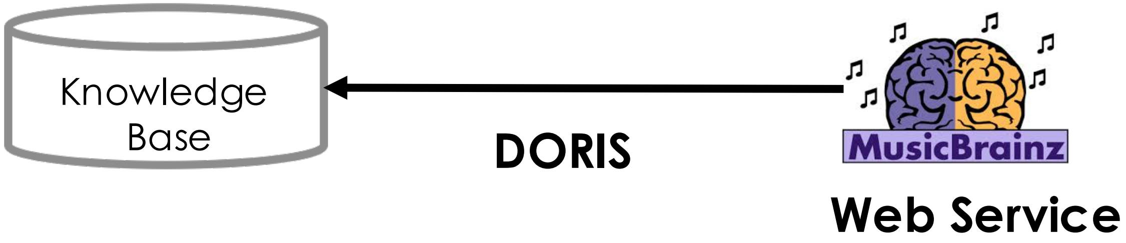
  MusicBrainz last.fm Discogs

- Financial (1200 APIs), Science (600 APIs), Weather (300 APIs)

*Source: ProgrammableWeb.com

Context & Objectives

- **PART I – DORIS:** *Deriving Intensional Description for Web Services*

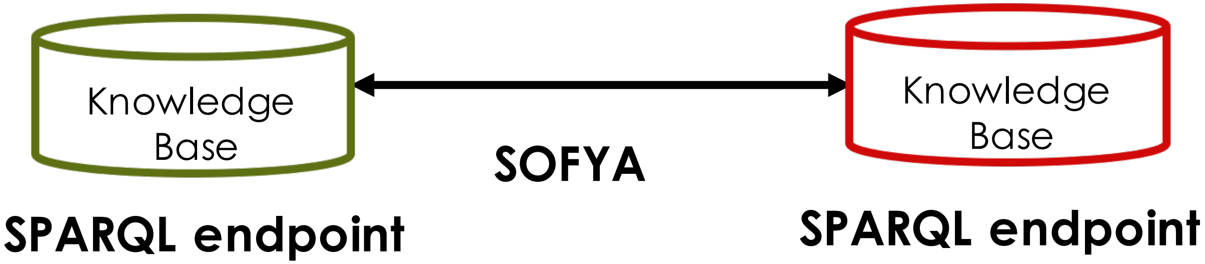


Knowledge Base

DORIS

Web Service

- **PART II – SOFYA:** *Online Relation Alignment on Linked Datasets*



Knowledge Base

SPARQL endpoint

SOFYA

Knowledge Base

SPARQL endpoint

Part I: Deriving Intensional Descriptions for Web Services

[CIKM’15, ISWC’15, BDA’15]

Web Services

What is a Web service?

- Way of publishing/exporting data
- A Web service (WS) is a function
- Consider WSs implementing REST: Interfaces to data sources
- Call a WS:
  - URL address of WS
  - Input value

Example: “get artworks by artist name” – exported by DORIS_museums

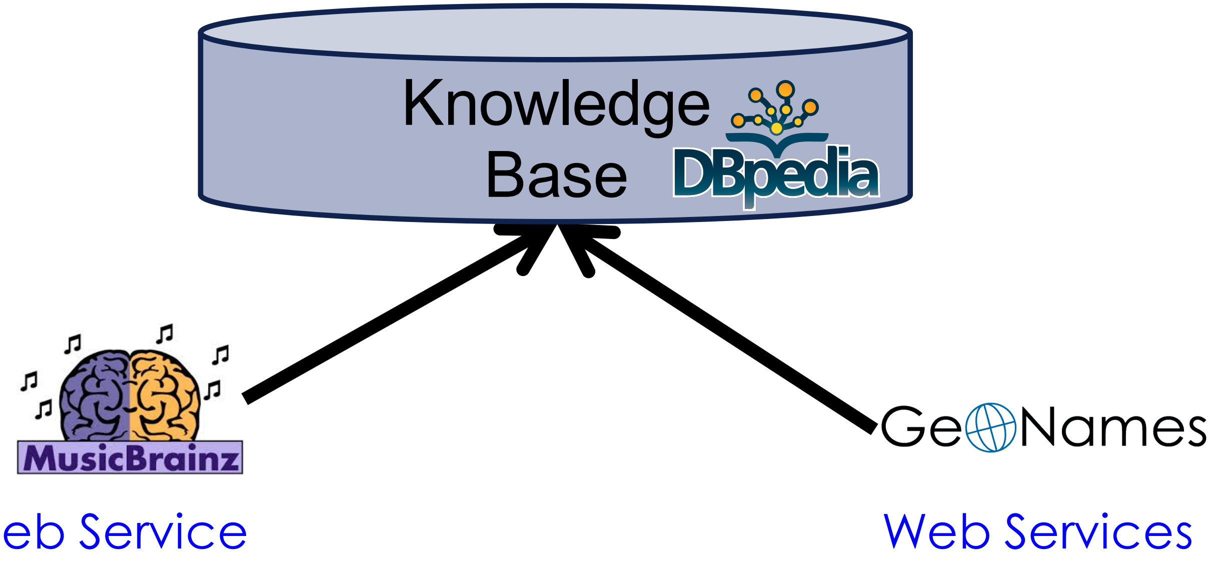
- call for input “Rodin”: http://doris_museums.com?artist=Rodin
- Output: XML document

Objective

Uniform access to Web services!

Local as view approach:

- We consider as target source a given Knowledge Base (RDF)
- Infer a mapping function (transform XML call results → RDF)
- Infer a description (parameterized query over the target KB)



Mapping function (σ)

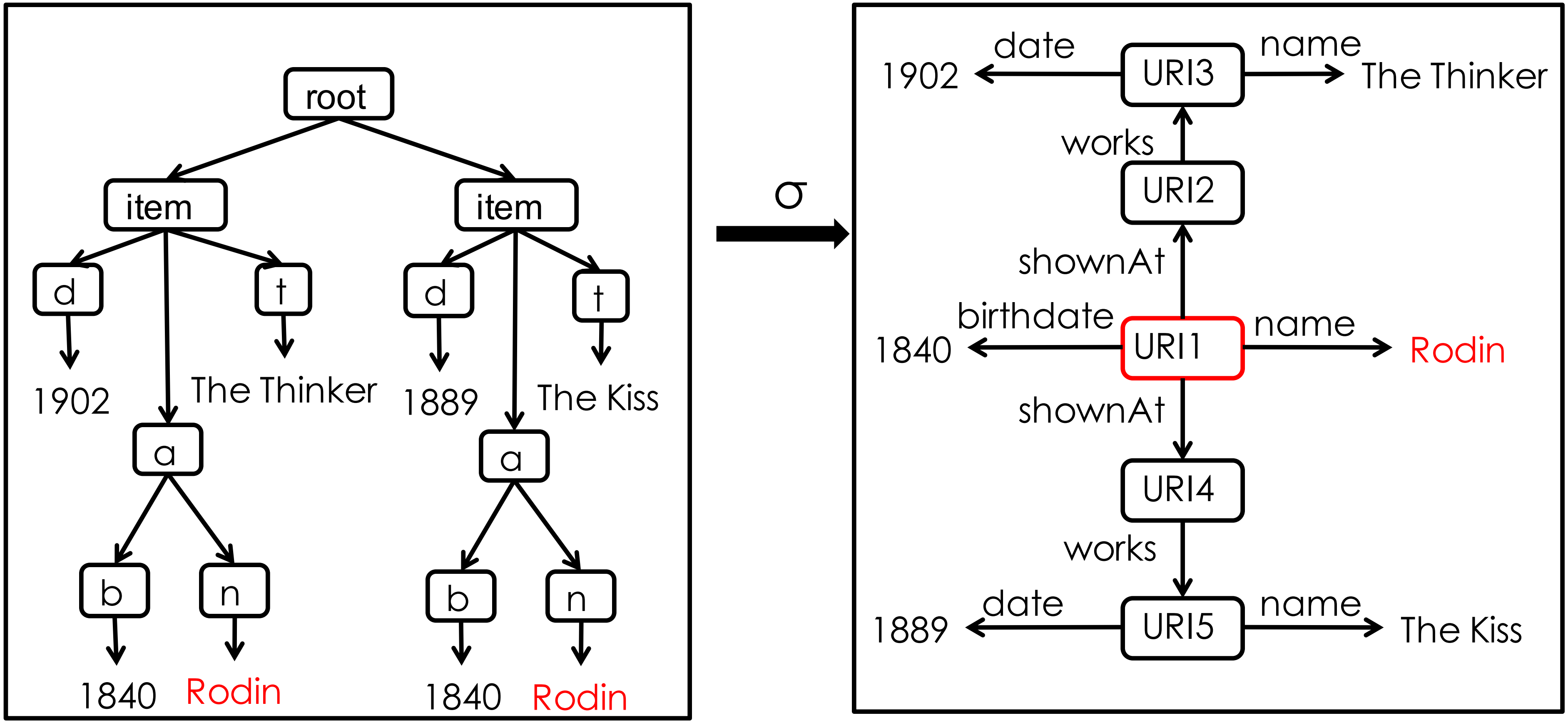
Web service: "get artworks by artist"

WS call result (XML)

KB fragment (RDF)

R: getArtWorksByArtist(Rodin)

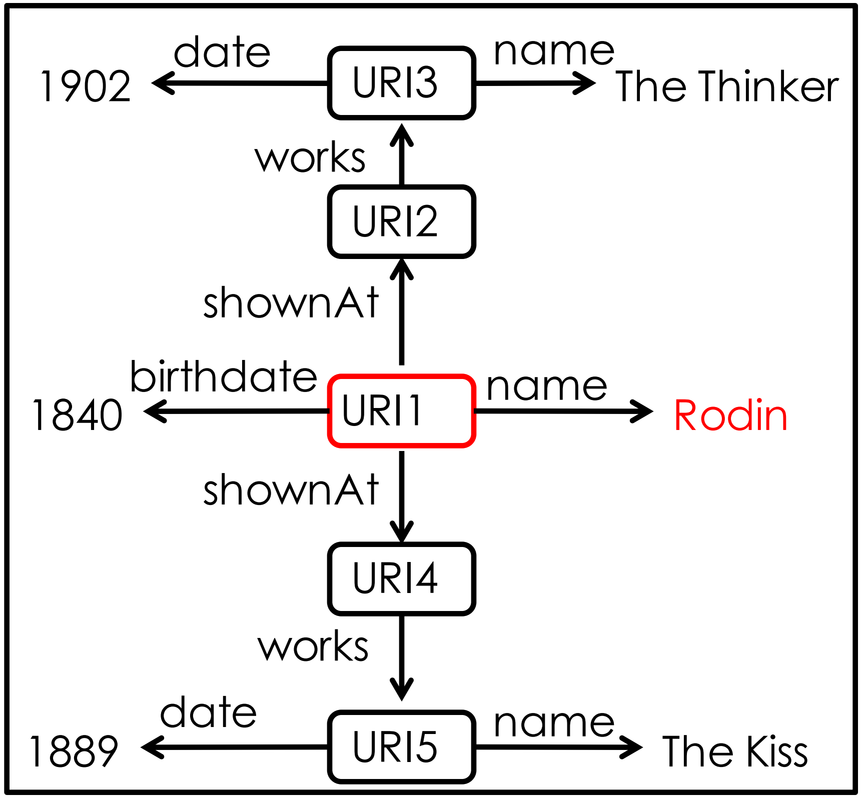
σ(R)



Parameterized Query

Schema of the parameterized query: the KB schema

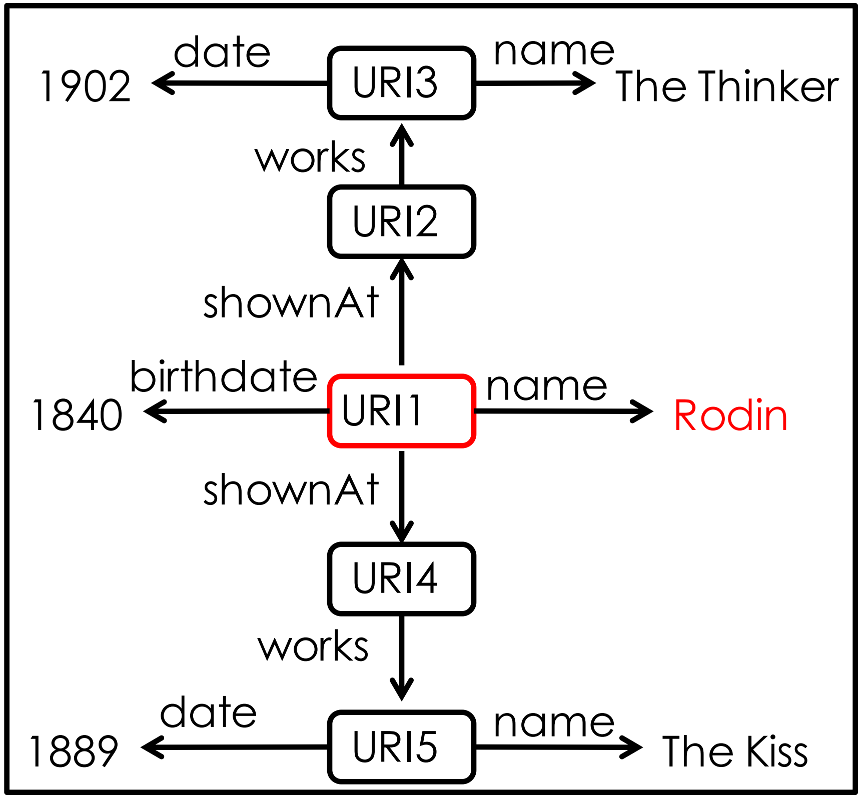
σ(getArtworksByArtist(Rodin))



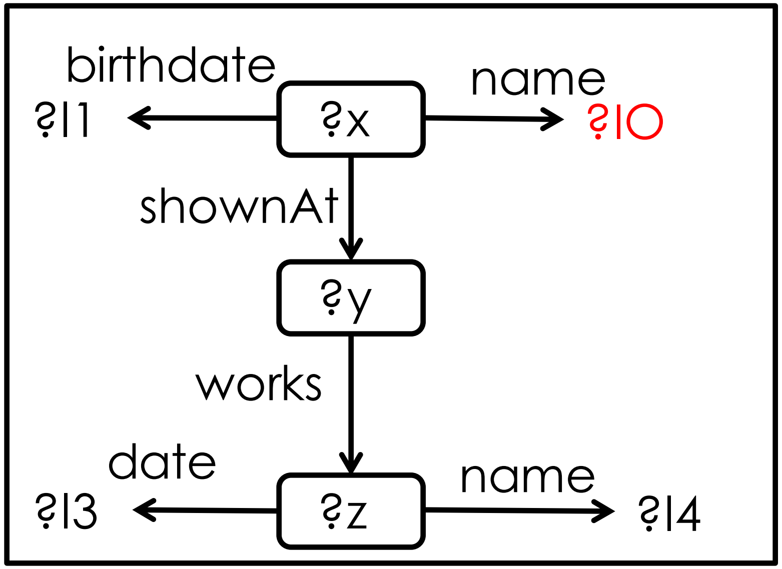
Parameterized Query

Schema of the parameterized query: the KB schema

σ(getArtworksByArtist(Rodin))



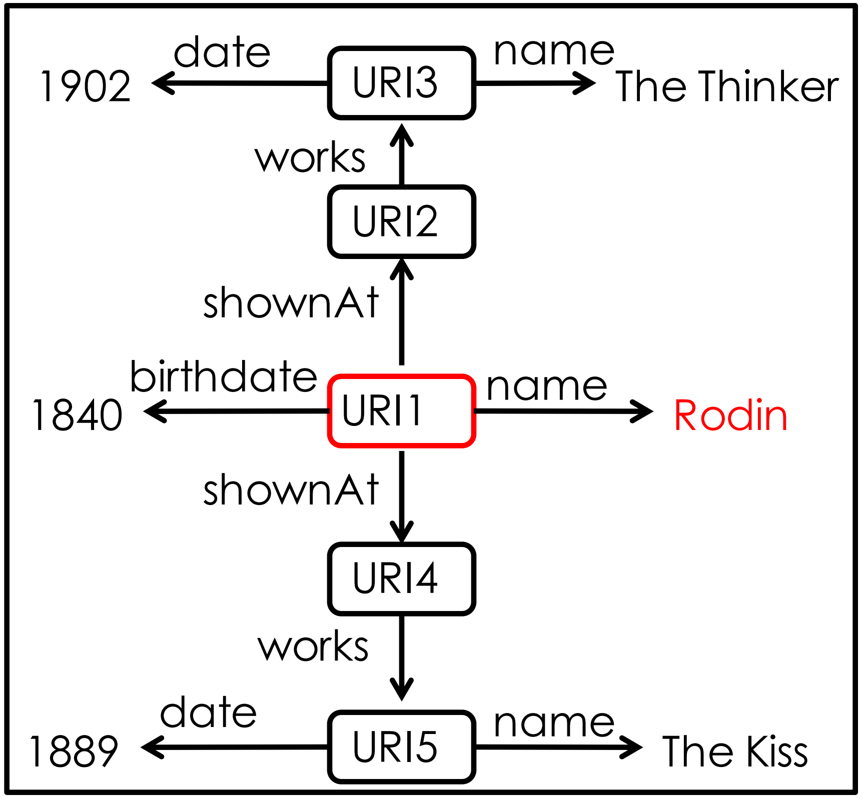
σ(getArtworksByArtist(?IO))



Parameterized Query

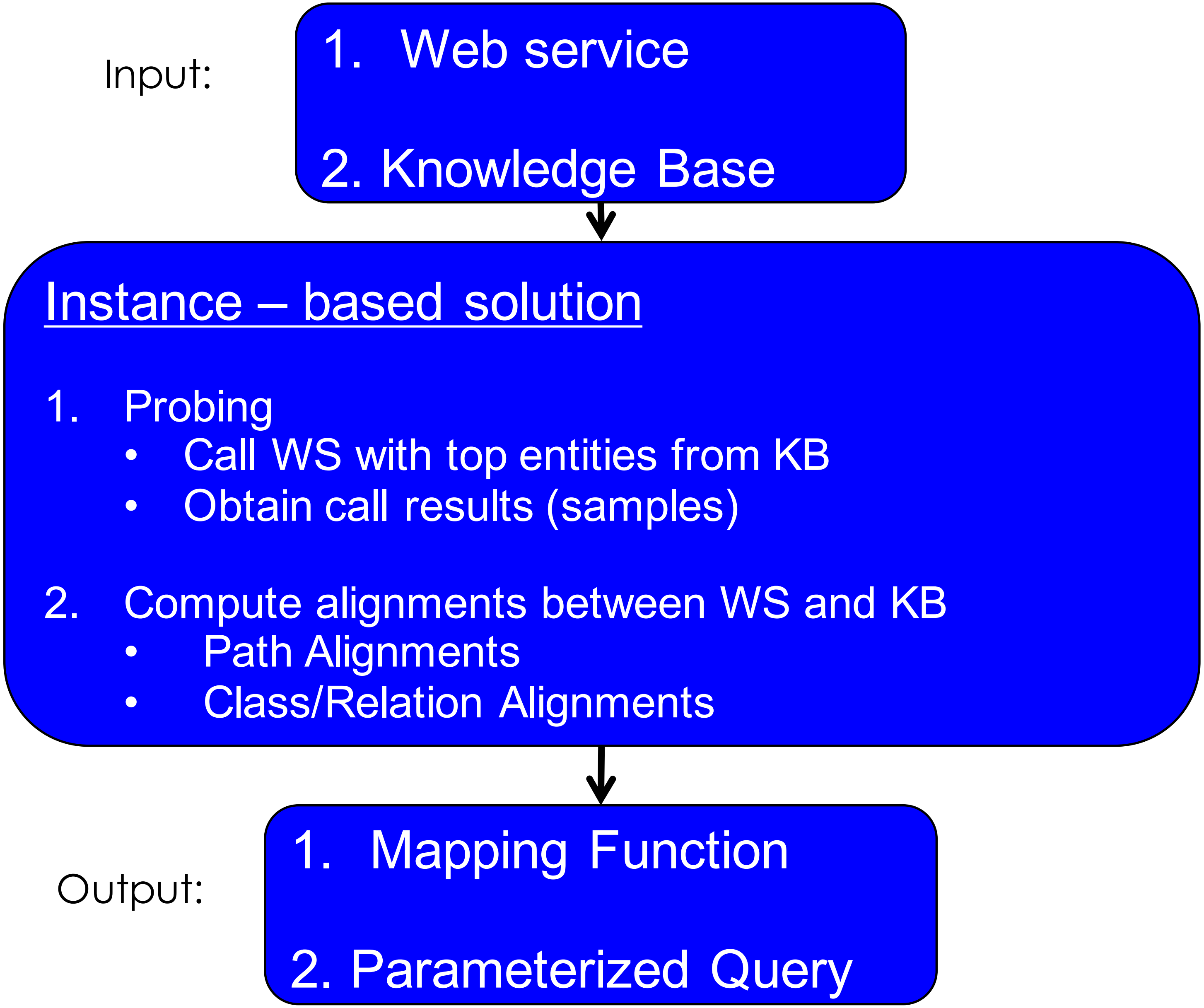
Schema of the parameterized query: the KB schema

σ(getArtworksByArtist(Rodin))



σ(getArworksByArtist(?IO,?I1, ?I3, ?I4))
←
name(?x, ?IO),
birthdate(?x,?I1),
shownAt(?x, ?y),
works(?y, ?z),
date(?z, ?I3),
name(?z, ?I4)

Overview – DORIS system



Input:

1. Web service
2. Knowledge Base

Instance – based solution

1. Probing
   - Call WS with top entities from KB
   - Obtain call results (samples)
2. Compute alignments between WS and KB
   - Path Alignments
   - Class/Relation Alignments

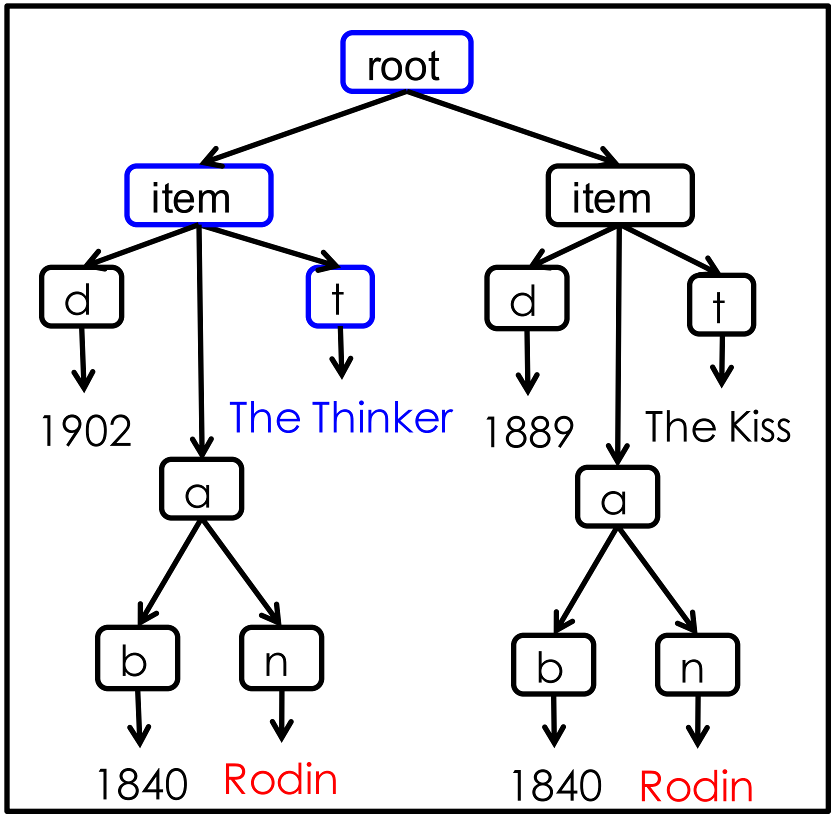
Output:

1. Mapping Function
2. Parameterized Query

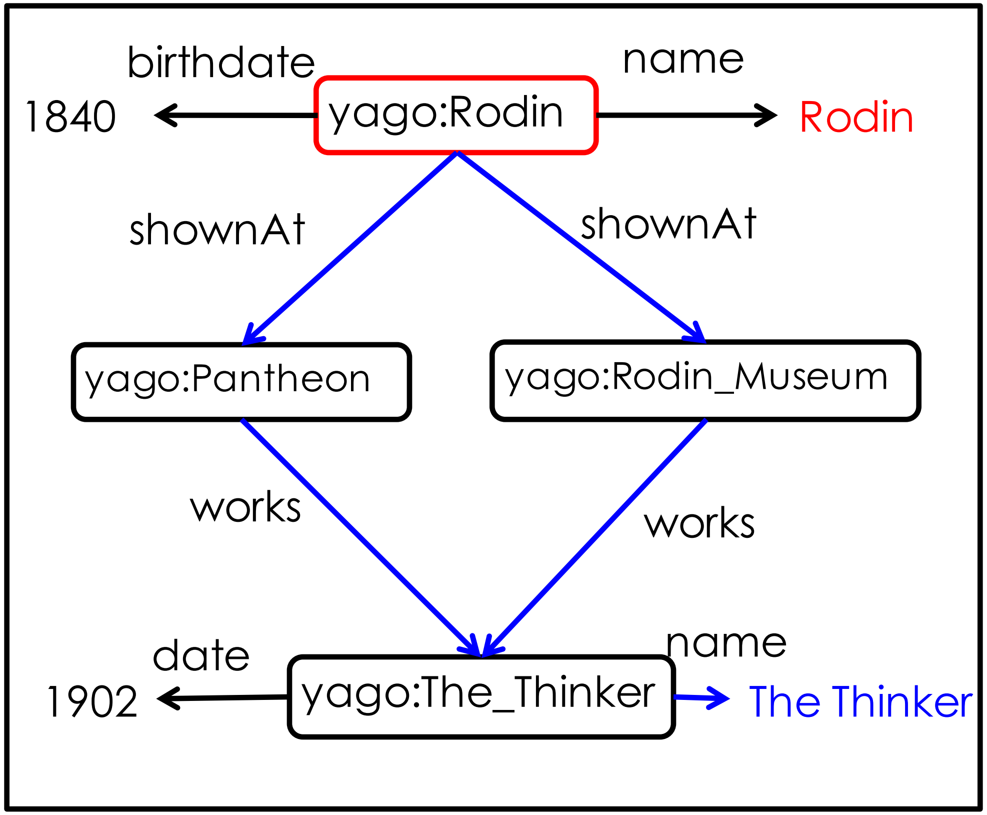
Path Alignments

- Relevant WS call result to an input entity (Rodin)
- Leaf nodes in call result encode attributes for input entity
- Linear XML paths in WS call result correspond to input entity – literal paths

getArtWorksByArtist(Rodin)

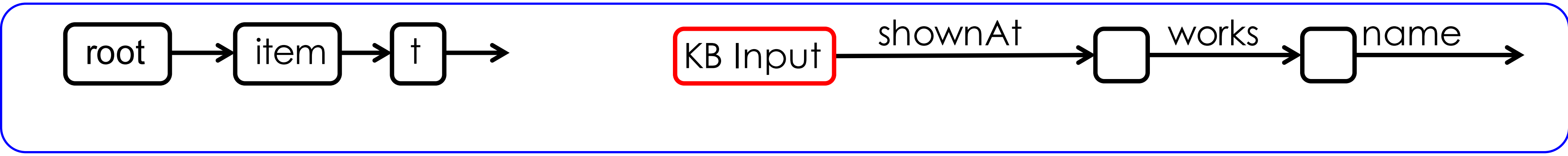


yago fragment (Rodin)



Path Alignments

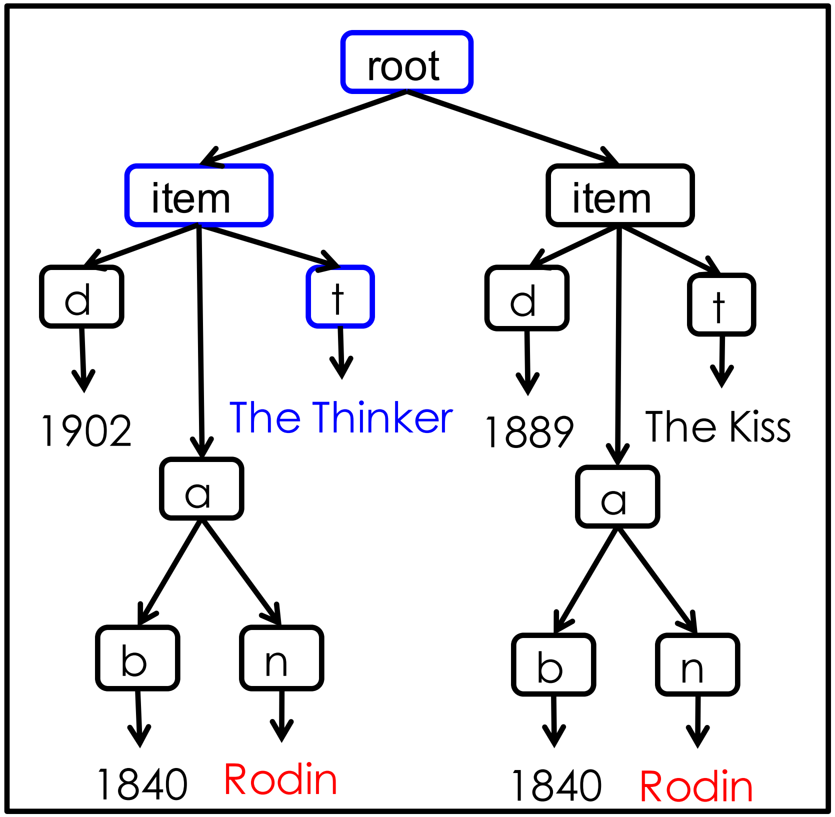
Path Pairs:



root → item → t →

KB Input → shownAt → works → name →

getArtWorksByArtist(Rodin)



root

item

d

1902

t

The Thinker

a

b

1840

n

Rodin

item

d

1889

t

The Kiss

a

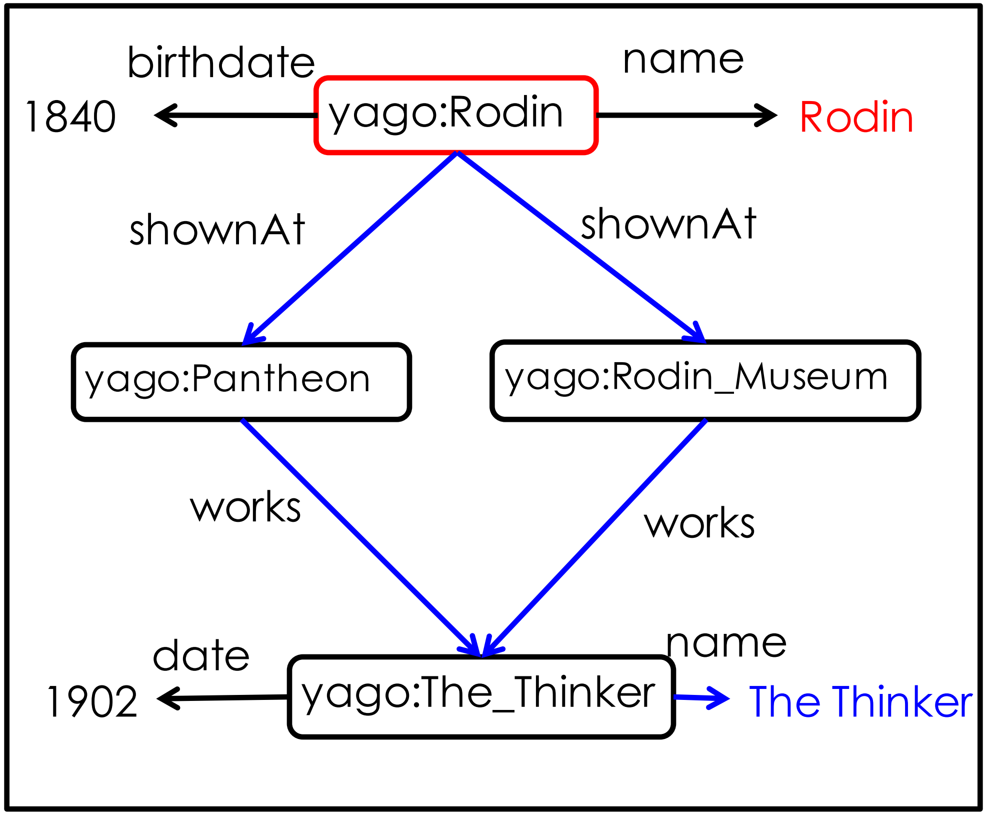
b

1840

n

Rodin

yago fragment (Rodin)



birthdate

1840

yago:Rodin

name

Rodin

shownAt

shownAt

yago:Pantheon

yago:Rodin_Museum

works

works

date

1902

yago:The_Thinker

name

The Thinker

Metrics for Path Alignments

1. Overlapping: align two paths if the results of the one overlap the results of the other over a threshold *a*.

$$Overlap_{conf}(p,p') = \frac{\#x : \exists y : p(x,y) \wedge p'(x,y)}{\#x} > \alpha$$

#x: number of samples

2. Inclusions: align two paths if the results of the one are included in the results of the other over a threshold *a*.

- Compute both ways inclusions: KB path ⇆ WS path

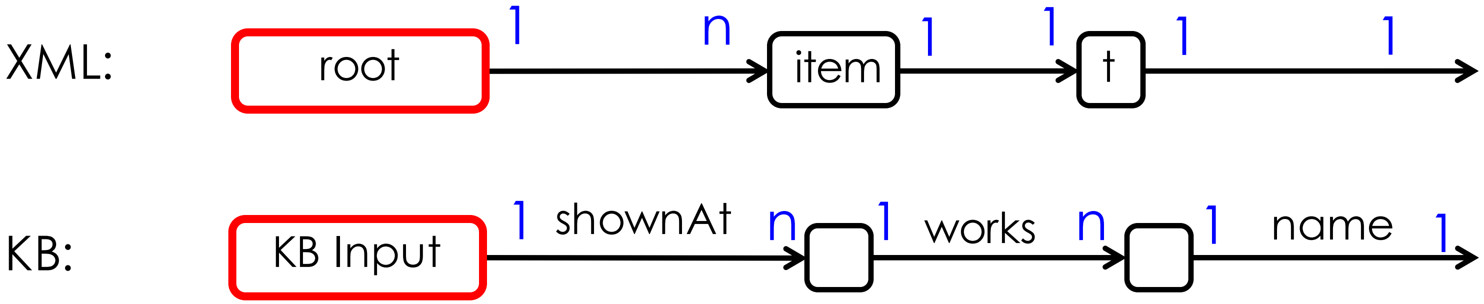
- *Partial completeness assumption: "a source knows either all or none of the p-attributes of some x"*

$$pca_{conf}(p,p') = \frac{\#(x,y) : \exists y : p(x,y) \wedge p'(x,y)}{\#(x,y) : \exists y' : p(x,y) \wedge p'(x,y')} > \alpha$$

Class & Relation Alignments

Problem: Identify XML nodes representing entities

- Idea: starting from the right-most side, align functional sub-paths (paths selecting one value)
- Assumption: the XML call result encode at least a function property per class of entities

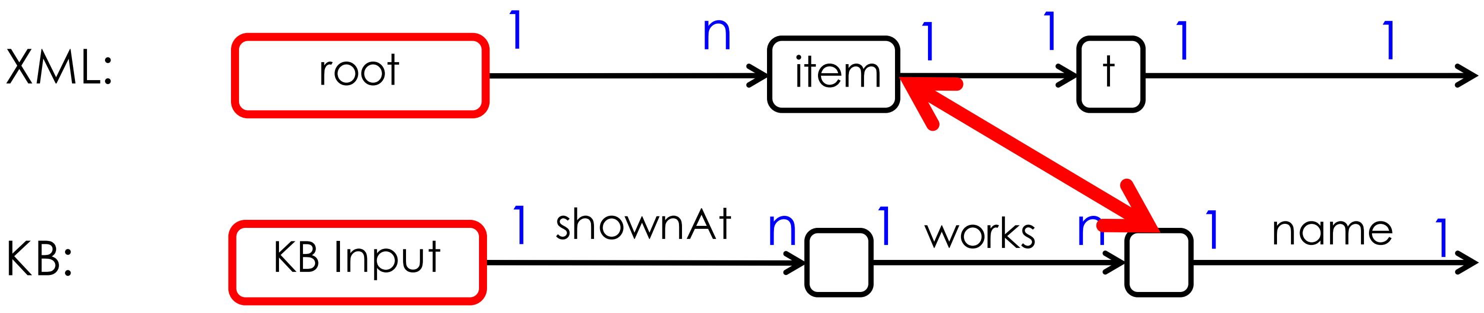


→ "item" nodes correspond to artworks

Class & Relation Alignments

Problem: Identify XML nodes representing entities

- Idea: starting from the right-most side, align functional sub-paths (paths selecting one value)
- Assumption: the XML call result encode at least a function property per class of entities



→ "item" nodes correspond to artworks

Class & Relation Alignments

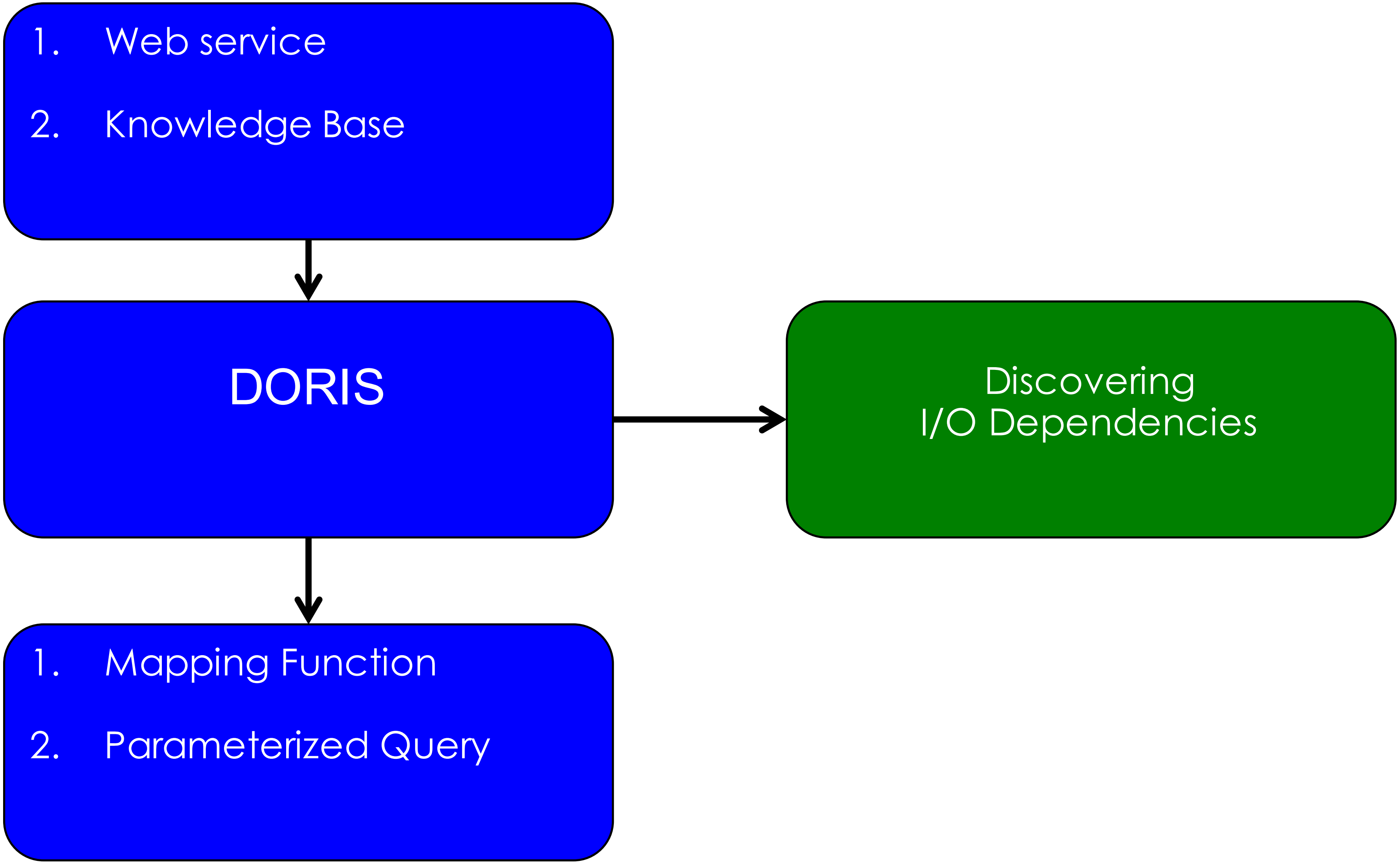
Compute Functionality

- KB: "A relation r(x,y) is called functional if for x there are not more than one y."

$$fun(r) = \frac{\#x : \exists y : r(x,y)}{\#(x,y) : r(x,y)} > \beta$$

- XML: "A path is functional if there are no two sibling nodes sharing the same label".

Overview



1. Web service
2. Knowledge Base

DORIS

Discovering
I/O Dependencies

1. Mapping Function
2. Parameterized Query

Discovering I/O Dependencies

Auguste Rodin

getArtworksByArworkID

getArtworksByArtist

ID_THE_THINKER

- 1.96 m
- Bronze

ID_THE_KISS

- 1.81 m
- Bronze

Auguste Rodin

- *The Thinker* **ID_THE_THINKER**
- *The Kiss* **ID_THE_KISS**

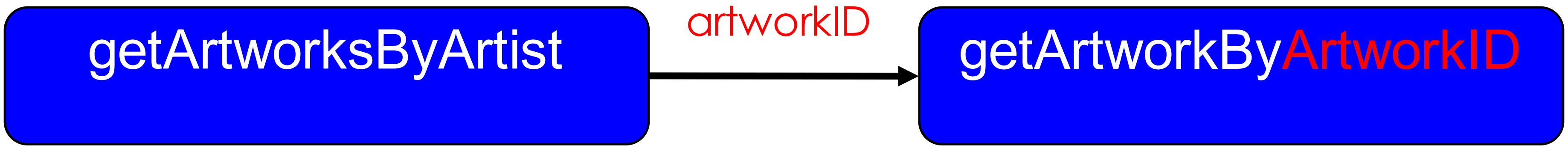
Join the output from the two calls

Discovering I/O Dependencies

Solution

- Discover "hidden" input types for Web services in the outputs of mapped (solved) Web services

Example:



Experimental Setup - Results

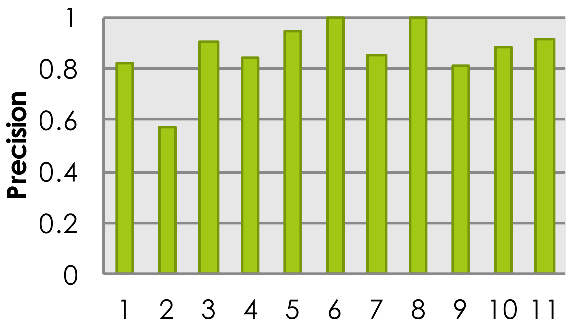
- 3 KB Tested (YAGO, DBpedia, BNF)
- > 50 Web Services (music, movies, books, geodata)
- → High Precision and Recall
- Summarization of Class/Relation alignment experiments:

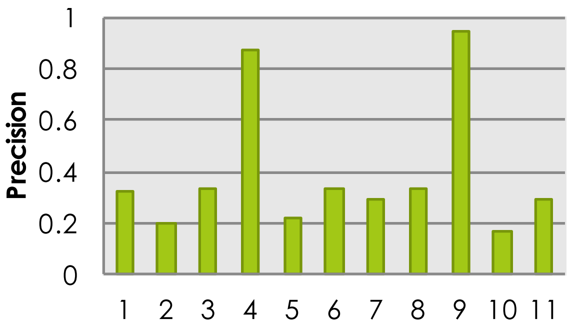
	Precision		Recall	
	Classes	Relations	Classes	Relations
YAGO	0.92	0.91	0.96	0.93
DBpedia	0.91	0.92	0.98	0.95
BNF *	1	1	1	1

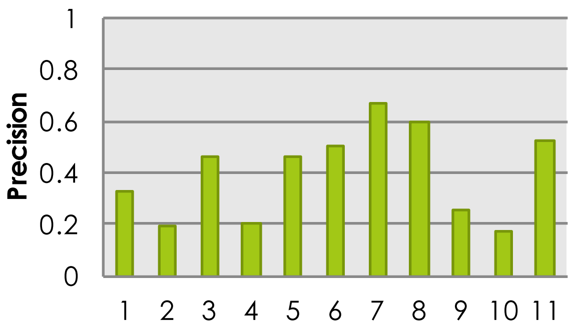
*Tested only with WSs from "Books" domain

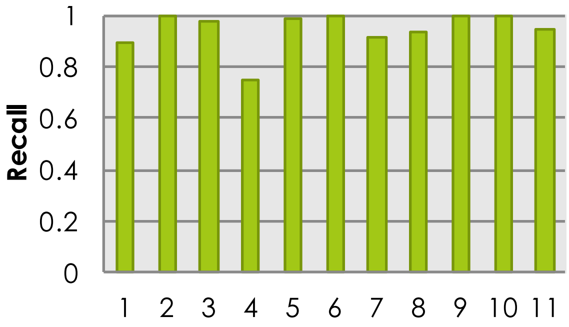
Evaluation Results

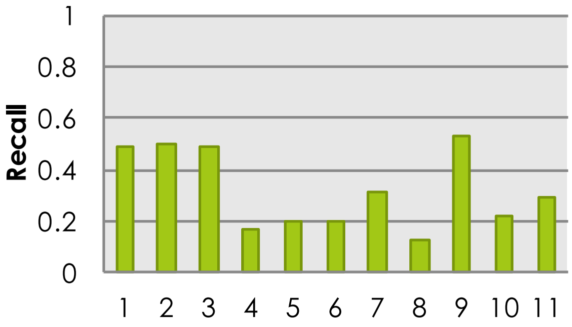
- Path Alignment
- Music Domain: 25 Web services

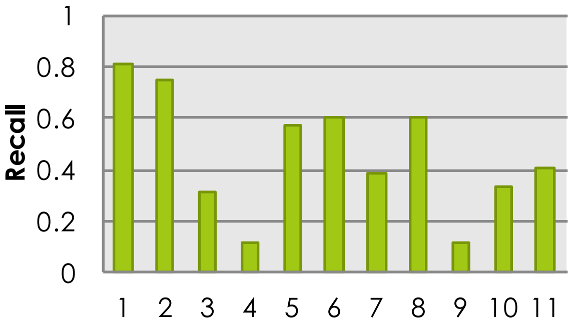

Overlap


KB → WS


WS → KB


Overlap


KB → WS


WS → KB

- More results : http://oasis.prism.uvsq.fr/doris/index.html

Conclusions - DORIS

- We proposed **DORIS**, a system that provides a formal description of the output of a Web service in terms of a global schema

- We provide a transformation function, as a script, to transform the output of the Web service in terms of a global schema.

- We proposed and algorithm that discovers I/O dependences between Web services of the same API

Part II: Online Relation Alignment on Linked Datasets

[EDBT'16]

Approach: Online Relation Alignment

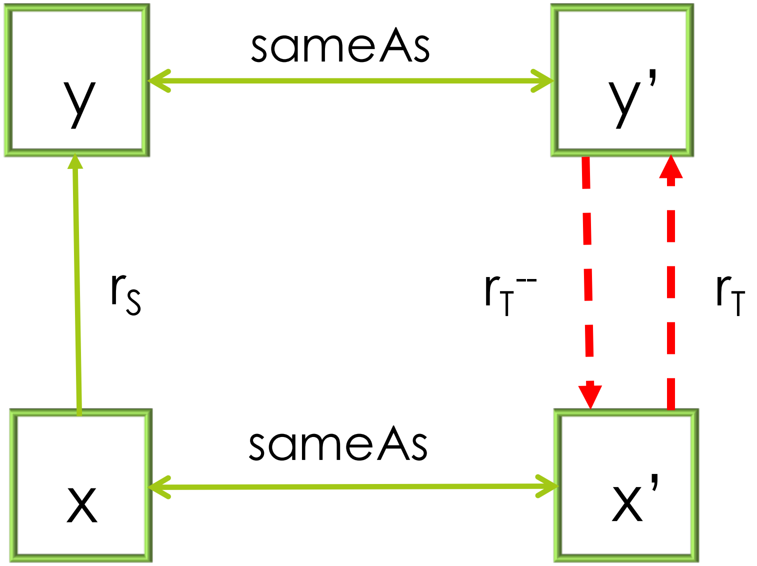
- Goal: Compute one-to-one relation alignments
  - Equivalence or subsumptions

- Align KBs published by SPARQL endpoints
- The entities of the two KBs are aligned via *sameAs* links

- Approach:
  - Instance-based
  - Supervised Model (features computed on KB instances)
  - Sample for a minimal set of entities to perform the alignment process

Approach: Outline

1



KB_S

SPARQL endpoint

KB_T

SPARQL endpoint

2

Candidates for alignment:

$r_S \subseteq r_{T1}$

$r_S \subseteq r_{T2}$

$r_S \subseteq r_{T3}$

…

3

Classify the alignments:

$r_S \subseteq r_{T1}$ (correct)

$r_S \subseteq r_{T2}$ (incorrect)

$r_S \subseteq r_{T3}$ (correct)

…

Approach: Features

..as matchers

Feature group
Inductive Logic Programming (ILP)
General Statistics (GS)
Lexical

Features – ILP: CWA & PCA

- Closed world assumption (cwa): for a relation r the KB contains all the facts.

$$cwa_{conf}(r_s \subseteq r_t) = \frac{overlap(r_s, r_t)}{|r_s|}$$

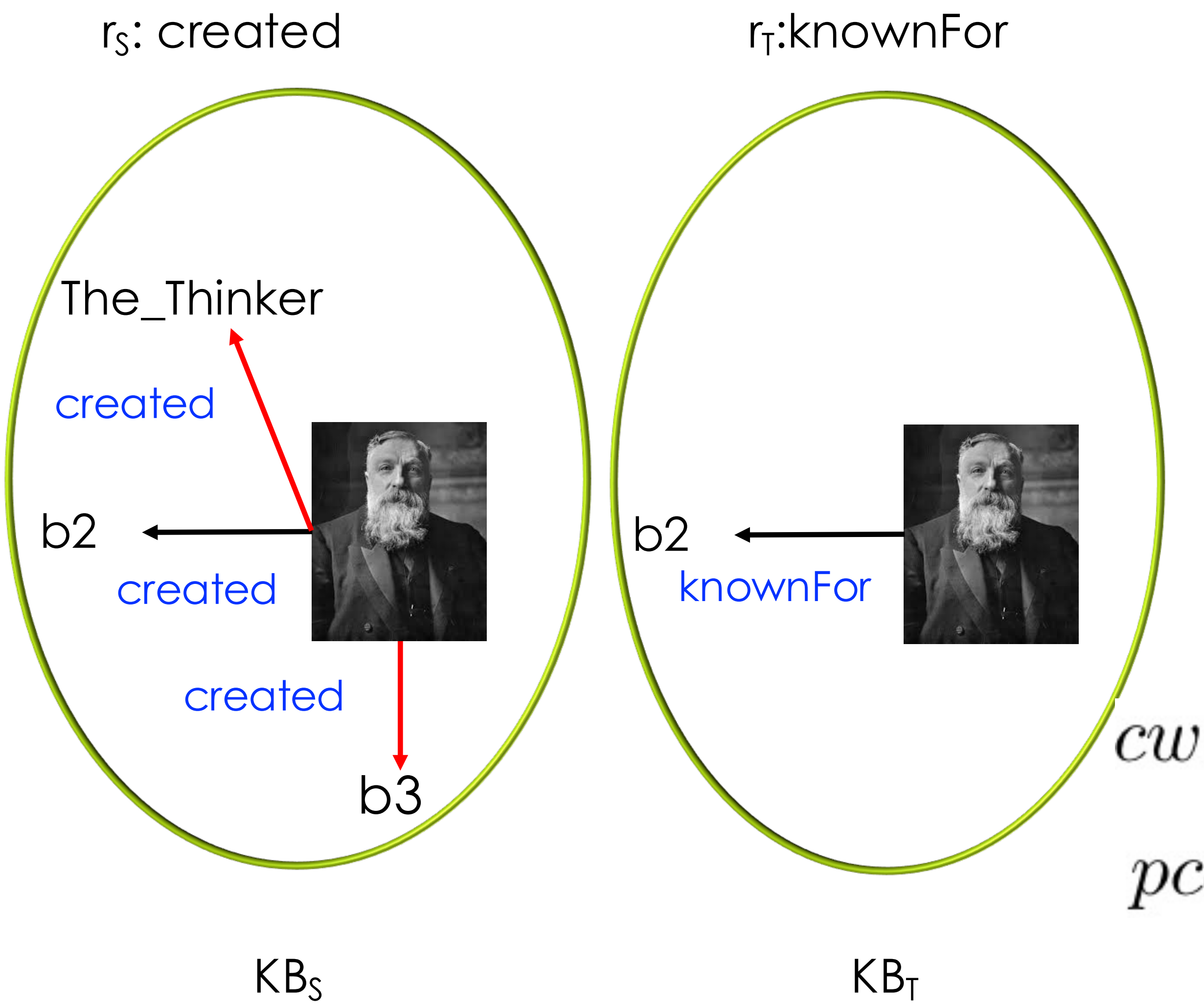
  - Good precision, bad recall
  - Absent data – counter examples

- Partial completeness assumption (pca): for a subject x and relation r, the KB contains ether all or none of the facts.

$$pca_{conf}(r_s \subseteq r_t) = \frac{overlap(r_s, r_t)}{overlap(r_s, r_t) + counter(r_s, r_t)}$$

Features – ILP: CWA & PCA

Example 1



$$overlap(r_s, r_t) = 1$$

$$|r_s| = 3$$

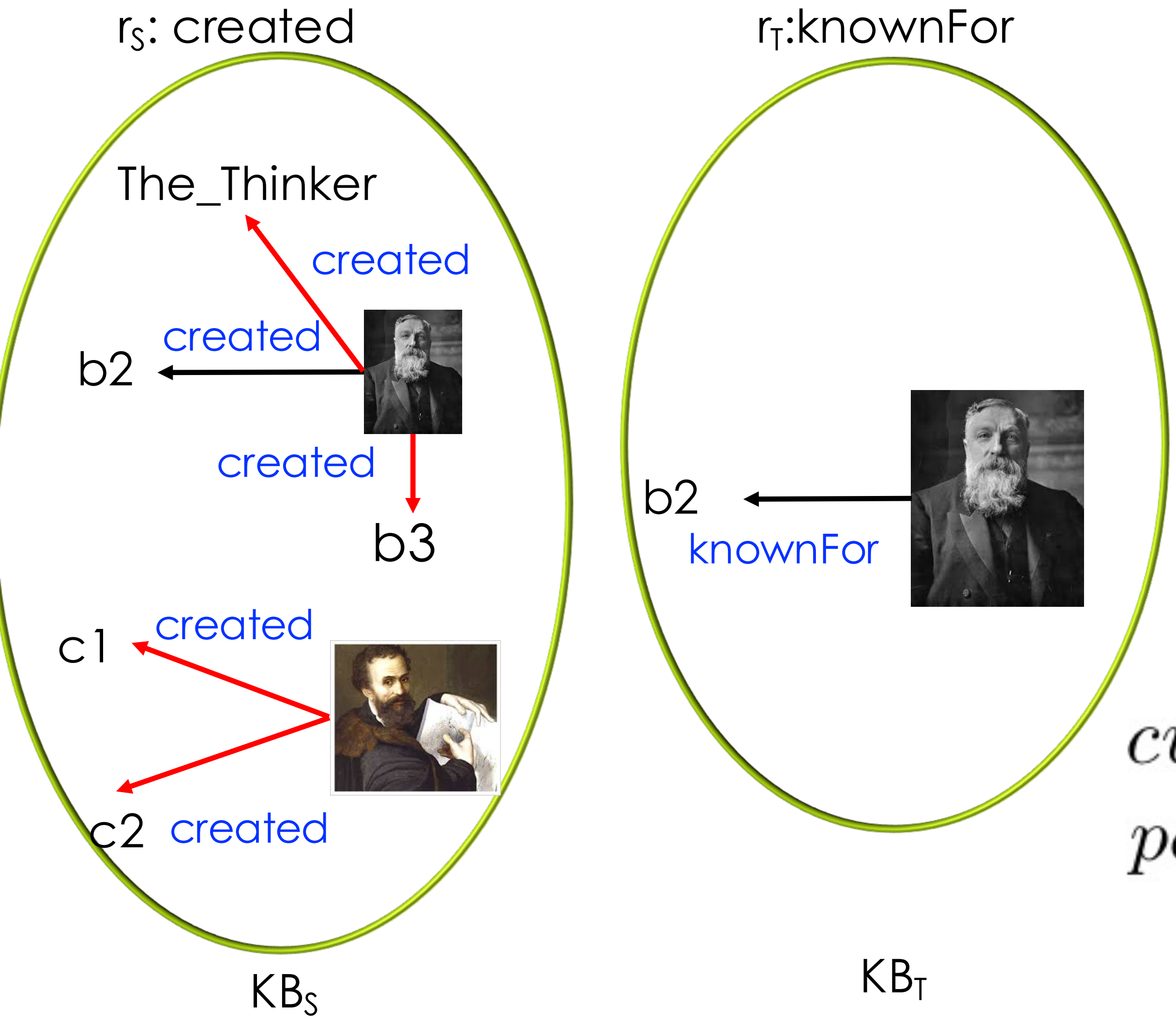
$$counter(r_s, r_t) = 2$$

$$cwa_{conf}(r_s, r_t) = 0.33$$

$$pca_{conf}(r_s, r_t) = 0.33$$

Features – ILP: CWA & PCA

Example 2



$$overlap(r_s, r_t) = 1$$

$$\boxed{|r_s| = 5}$$

$$counter(r_s, r_t) = 2$$

$$cwa_{conf}(r_s, r_t) = 0.2$$

$$pca_{conf}(r_s, r_t) = 0.33$$

Features – Relation Functionality

- Functionality: “A relation r(x,y) is called functional if for x there are not more than one y.”

$$fun(r) = \frac{\#x : \exists y : r(x,y)}{\#(x,y) : r(x,y)}$$

$$r_s \subseteq r_t \Rightarrow fun(r_s) \geq fun(r_t)$$

- If **r**$_s$ is subsumed in **r**$_t$ the functionality should be higher
- Target relations should have better coverage of facts

Features - ILP: PIA

- Partial completeness assumption - pca
  - good performance for functional relations
  - Penalizes the non-functional relations

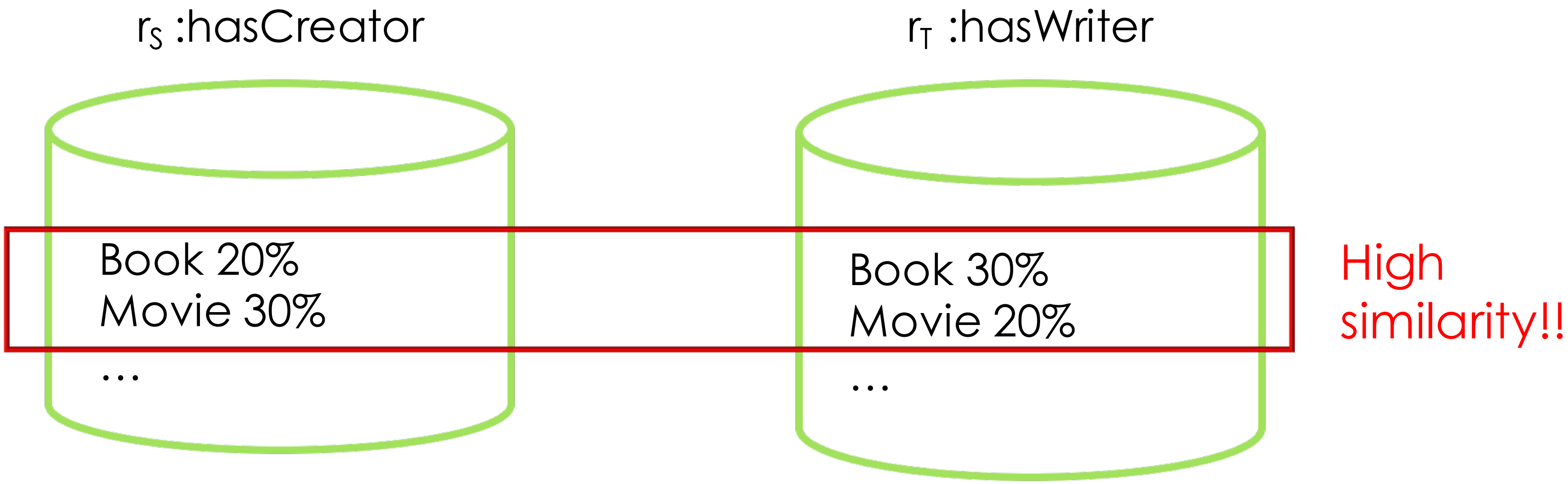
- Propose: Partial incompleteness assumption – pia

$$\frac{overlap(r_s, r_t)}{overlap(r_s, r_t) + (counter(r_s, r_t) \boxed{\times func(r_s)})}$$

  - The more important the counter example is the more should count!

Features – GS: Type similarity

- Check the type distribution similarity between relations r_S and r_T.
- Example:



r_S :hasCreator

Book 20%
Movie 30%
…

r_T :hasWriter

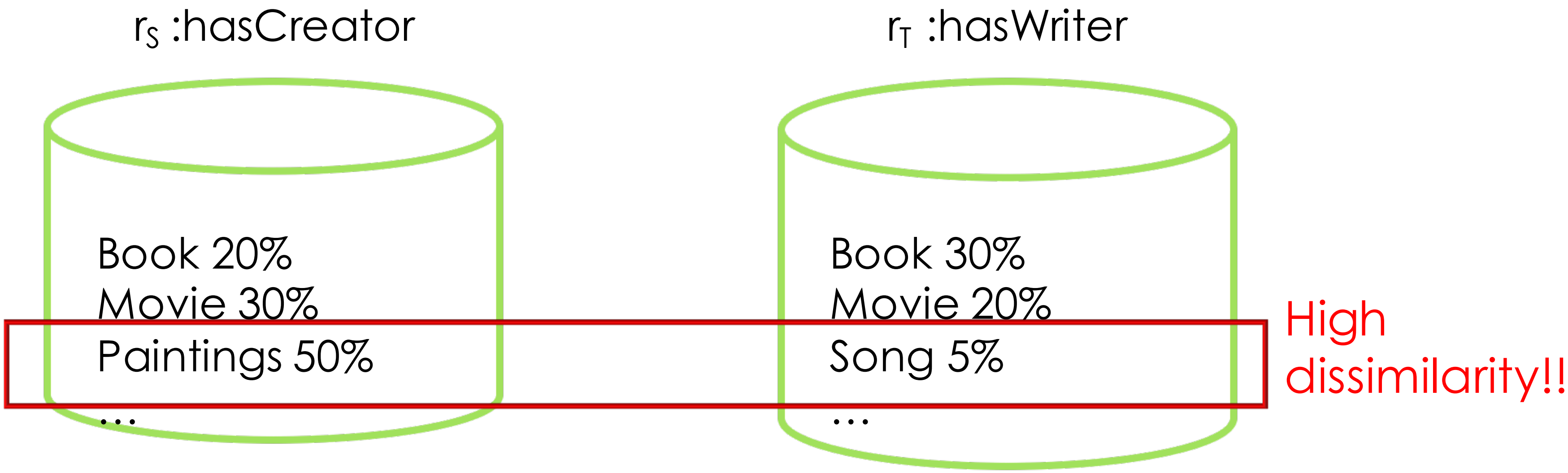
Book 30%
Movie 20%
…

High similarity!!

- Weighted Jaccard similarity metric to assess if the two relations have similar structure in terms of types.
- High similarity – Good indicator for *equivalence/subsumption* between relations

Features – GS: Type dissimilarity

- Check if type distribution in r_S contains type that do not exist in r_T.
- Example:



r_S :hasCreator

Book 20%
Movie 30%
Paintings 50%
…

r_T :hasWriter

Book 30%
Movie 20%
Song 5%
…

High dissimilarity!!

- For missing types and based on their ratio we can accurately assess that r_T does not subsume r_S.

Features – GS: Relevance likelihood

- Likelihood of ILP scores: depend on the datasets the matchers varies !!
- Compute the likelihood of specific ILP scores being indicators of subsumption for a relation pair!
  - pca likelihood
  - cwa likelihood
  - Joint pca & cwa likelihood
- Compute the likelihood of a relation alignment being correct given a specific ILP score.
- Probabilities are measured on the training set! Assign the scores on the test set

Approach: Efficiency Issues

- Challenges
  - Bandwidth
  - Time-out at SPARQL endpoints
- Approach
  - Reduce data transfers
  - Retrieve a subset of instances for a given relation
- Solution
  - Sample for a minimal subset of instances for the relation alignment
    - First-N
    - Random
    - Stratified

Experimental Setup

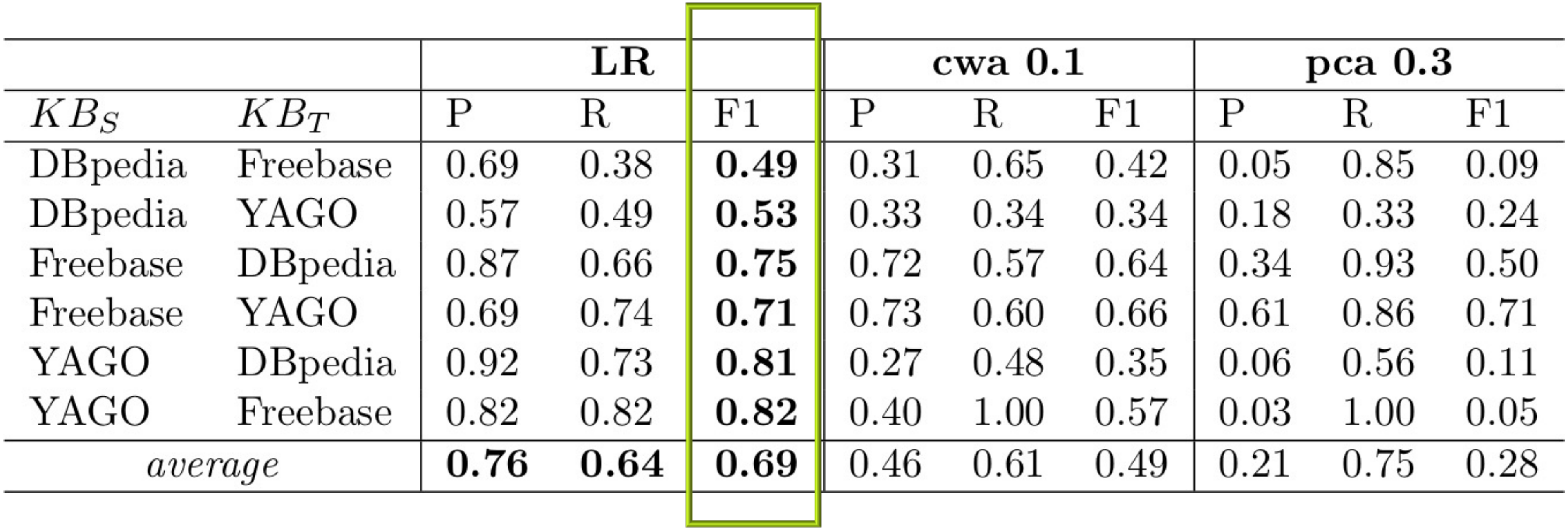
- 3 Knowledge Bases
  - YAGO, DBpedia, Freebase (e.g. YAGO → DBpedia)
- Relations

KB	YAGO	DBpedia	Freebase
#relations	36	563	1666

- Baselines
  - cwa (used in PARIS)
  - pca (used in ROSA)
- SOFYA: Logistic Regression (any other supervised model can be applied)

Evaluation Results: Performance

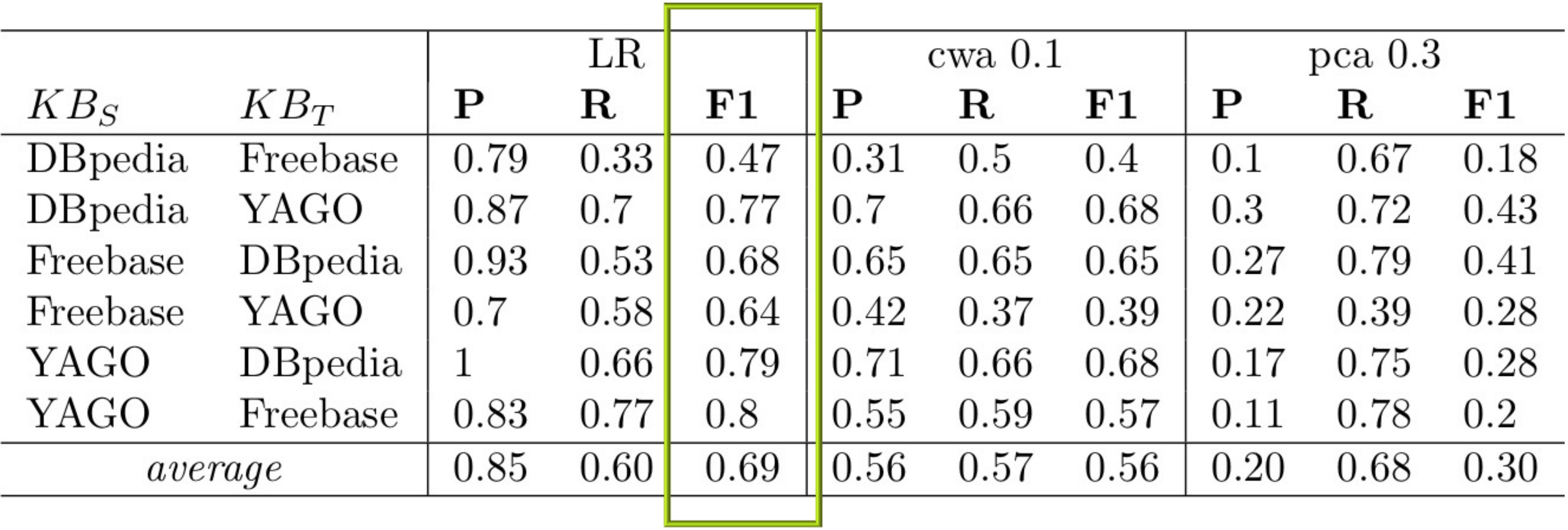
- Full Data: Comparison of the different models and competitors



		LR			cwa 0.1			pca 0.3		
KB_S	KB_T	P	R	F1	P	R	F1	P	R	F1
DBpedia	Freebase	0.69	0.38	**0.49**	0.31	0.65	0.42	0.05	0.85	0.09
DBpedia	YAGO	0.57	0.49	**0.53**	0.33	0.34	0.34	0.18	0.33	0.24
Freebase	DBpedia	0.87	0.66	**0.75**	0.72	0.57	0.64	0.34	0.93	0.50
Freebase	YAGO	0.69	0.74	**0.71**	0.73	0.60	0.66	0.61	0.86	0.71
YAGO	DBpedia	0.92	0.73	**0.81**	0.27	0.48	0.35	0.06	0.56	0.11
YAGO	Freebase	0.82	0.82	**0.82**	0.40	1.00	0.57	0.03	1.00	0.05
average		**0.76**	**0.64**	**0.69**	0.46	0.61	0.49	0.21	0.75	0.28

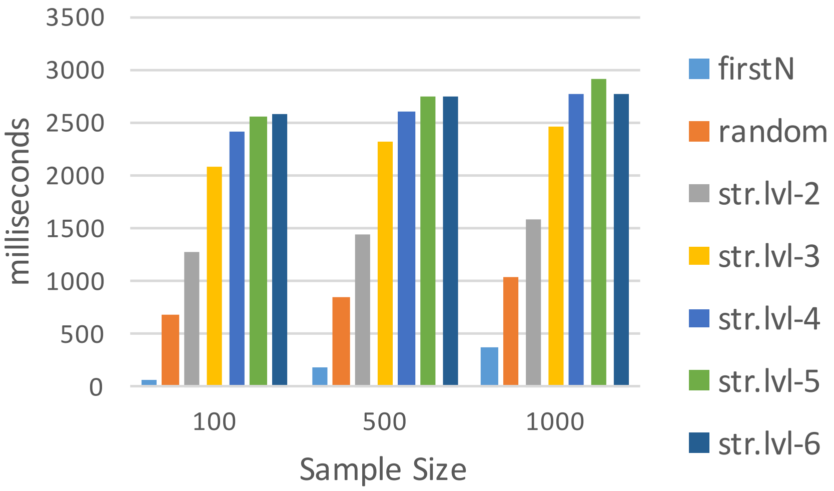
Evaluation Results: Performance

- Sampled Data: Individual results on sampling – Stratified Level 3 – 50 entity samples

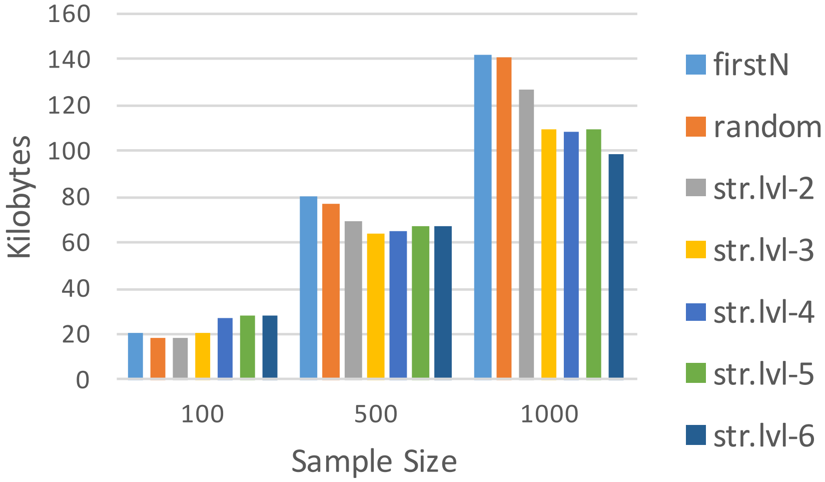


		LR			cwa 0.1			pca 0.3		
KB_S	KB_T	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
DBpedia	Freebase	0.79	0.33	0.47	0.31	0.5	0.4	0.1	0.67	0.18
DBpedia	YAGO	0.87	0.7	0.77	0.7	0.66	0.68	0.3	0.72	0.43
Freebase	DBpedia	0.93	0.53	0.68	0.65	0.65	0.65	0.27	0.79	0.41
Freebase	YAGO	0.7	0.58	0.64	0.42	0.37	0.39	0.22	0.39	0.28
YAGO	DBpedia	1	0.66	0.79	0.71	0.66	0.68	0.17	0.75	0.28
YAGO	Freebase	0.83	0.77	0.8	0.55	0.59	0.57	0.11	0.78	0.2
average		0.85	0.60	0.69	0.56	0.57	0.56	0.20	0.68	0.30

Evaluation Results: Efficiency



SPARQL Sampling time in milliseconds



Bandwidth usage in in kilobytes

Conclusions - SOFYA

- We proposed **SOFYA**, an instance-based relation alignment approach, discovering subsumptions of relations

- We propose supervised machine learning models, that combine a set of light-weight features to decide if the subsumption relationship is correct or incorrect

- Overcome main drawbacks of existing schema matching approaches, through efficient alignment algorithms

- Harness the complementarity of LOD sources through relation alignments at query time

Future/Ongoing work

- Automatic discovery of input types in DORIS

- Investigate for additional features in SOFYA

- Relation alignment for complex relations: 1-n relations in SOFYA

- Compute subsumption of relations starting from the super-relation in SOFYA

Publications (1/2)

- National conferences:
  - **Mapping Web Services to Knowledge Bases**, 2015, Bases de Données Avancées (**BDA**), Maria Koutraki, Dan Vodislav, Nicoleta Preda
  - **DORIS: Discovering Ontological Relations in Services**, 2015, Bases de Données Avancées (**BDA**), Maria Koutraki, Dan Vodislav, Nicoleta Preda
  - **Uniformly Querying Web Knowledge Bases,** 2016, **parisDB**, Maria Koutraki, Nicoleta Preda, Dan Vodislav

Publications (2/2)

- International conferences:
  - **Deriving Intensional Descriptions for Web Services**, 2015, International Conference on Information and Knowledge Management (**CIKM**), Maria Koutraki, Dan Vodislav, Nicoleta Preda
  - **DORIS: Discovering Ontological Relations in Services**, 2015, International Semantic Web Conference (**ISWC**), Maria Koutraki, Dan Vodislav, Nicoleta Preda
  - **SOFYA: Semantic on-the-fly Relation Alignment**, 2016, International Conference on Extending Database Technology (**EDBT**), Maria Koutraki, Nicoleta Preda, Dan Vodislav

Thank you all !

Questions ?