

Aggregated search in large RDF repositories

Youssef Barhoun, Haytham Elghazel, Mohand-Saïd Hacid, Rafiqul Haque, Thanh-Huy Le

November 24, 2015



INSA



UNIVERSITÉ
LUMIÈRE
LYON 2



Tutelles du LIRIS et Ressources Humaines

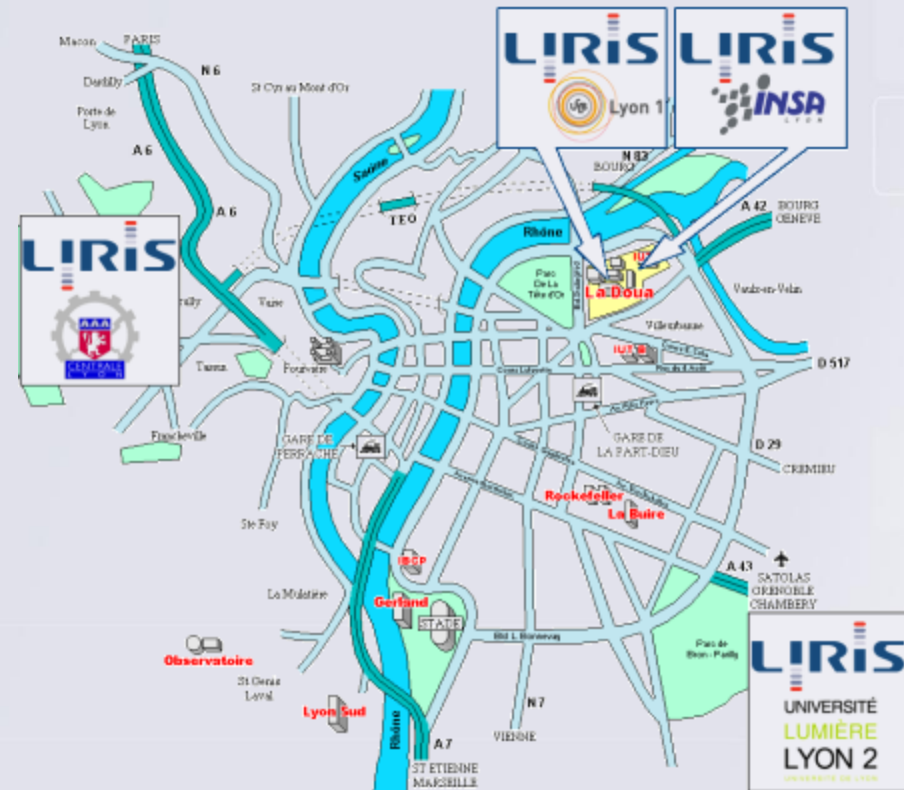
■ 5 tutelles (148 permanents)

- CNRS (15)
- INSA de Lyon (46)
- Université Lyon 1 (65)
- Université Lyon 2 (9)
- ECL (8)

- Université Lyon 3 (3)
- INRIA (1)
- Hors tutelle (1)

■ Sur 3 campus et 5 bâtiments : 327

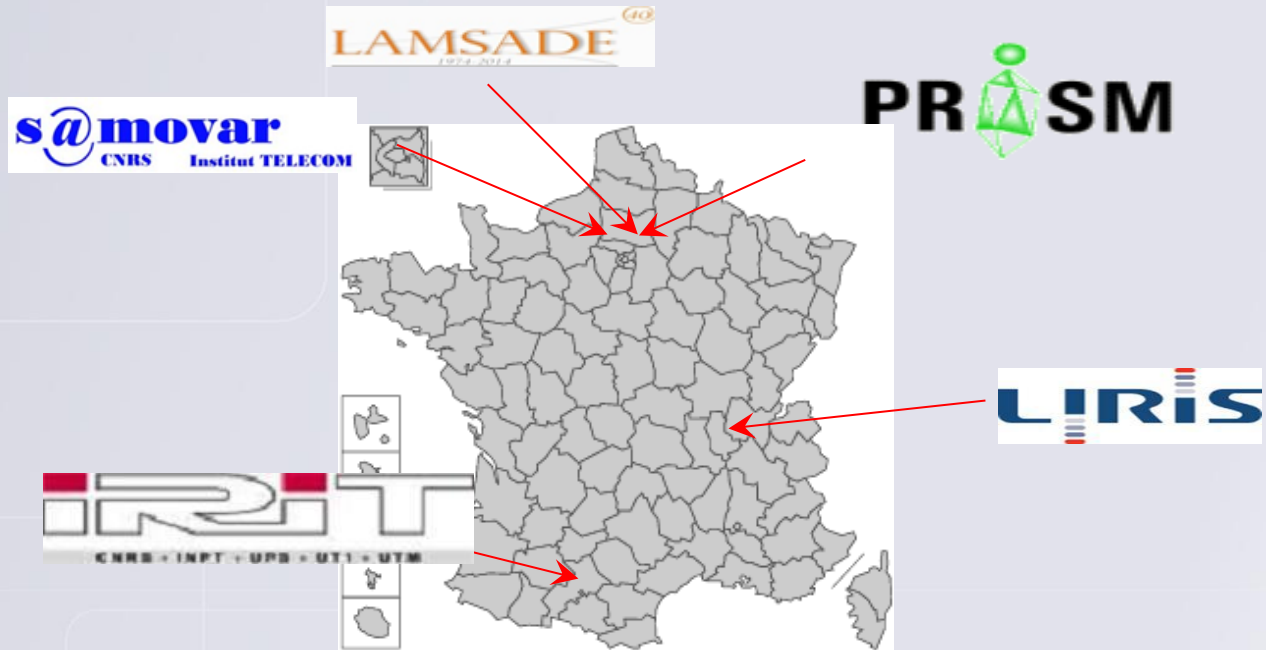
- Villeurbanne (291)
- Bron (16)
- Ecully (20)



La structure du laboratoire 6 pôles scientifiques et 14 équipes



<http://www.irit.fr/CAIR>



Motivation (1/4)

A user looking for informations about a given movie actor?



Brad Pitt

[Brad Pitt : des photos, des vidéos et...](#)
Brad Pitt : Né le 18 décembre 1963 dans l'OC... toutes les actu et les news de l'acteur...
cineday.orange.fr/star/brad-pitt/

[Brad Pitt](#)
Brad Pitt : Brad Pitt est né le 18...
www.cinefil.com/star/brad-pitt

[Brad Pitt, Leonardo DiCaprio, Miley Cyrus...](#)
Avant de devenir de grands acteurs pour la plupart, ils doivent passer par un certain nombre...
d'auditions Films, séries... À quoi cela ressemble ? La rédaction de meltyBuzz vous le fait découvi...
www.meltybuzz.fr/ - il y a 4 jours

[Fury, avec Brad Pitt, voit sa date de sortie avancée](#)
Le film de guerre de David Ayer sortira le 17 octobre au lieu du 23 novembre comme initialement...
prévu. Cette nouvelle stratégie permettrait au film de résister aux deux blockbusters Interstellar et...
www.lefigaro.fr/ - il y a 8 jours

[Angelina...elle interdit à Brad Pitt de tourner des scènes d'amour avec d'autre...](#)
Lorsqu'ils se sont rencontrés sur le plateau du film Mr...Angelina Jolie et Brad Pitt n'ont pas su...
résister à leur folle alchimie...Désormais heureuse en amour... Lire la suite de la news sur Public.fr
www.public.fr - il y a 8 jours

[Brad Pitt - AlloCiné](#)
Brad Pitt (William Bradley Pitt), Acteur, Producteur, Producteur exécutif. Découvrez sa biographie,
sa carrière en détail et toute son actualité...
www.allocine.fr/personne/fichepersonne_gen_cpersone=12302.h...

[Brad Pitt - Biographie et filmographie](#)
Dans quels films a joué Brad Pitt ? Découvrez les photos, la biographie de Brad Pitt...
www.linternaute.com/biographie/brad-pitt/

- ✓ Consider many documents
- ✓ Extract relevant information
- ✓ Syntesize/configure the final answer

Traditional search engines



News



WIKIPÉDIA
L'encyclopédie libre

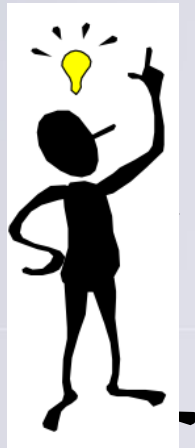


Images



Reviews

Motivation (2/4)



Brad Pitt

Aggregated Search Engine



Brad Pitt

Acteur

Brad Pitt est un acteur et producteur de cinéma américain né le 18 décembre 1963 à Shawnee, dans l'Oklahoma. Repéré dans une publicité pour Levi's, Brad Pitt sort de l'anonymat grâce à un petit rôle dans le film Thelma et Louise de Ridley Scott. Wikipédia

Naissance : 18 décembre 1963 (50 ans), Shawnee, Oklahoma, États-Unis

Taille : 1,80 m

Compagne : Angelina Jolie (2005–)

Épouse : Jennifer Aniston (m. 2000–2005)

Enfants : Shiloh Jolie-Pitt, Vivienne Marcheline Jolie-Pitt, plus...

Films

Voir d'autres éléments (plus de 45)



World War L'Etrange Mr. et Mrs. Seven Le Stratège



News



WIKIPÉDIA
L'encyclopédie libre



Images



Reviews

Motivation (3/4)

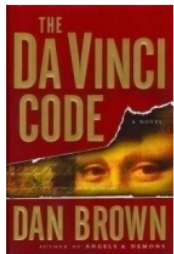
Opinion Analysis on Blog Articles



Tom Hanks, who is my favorite movie star act the leading role.



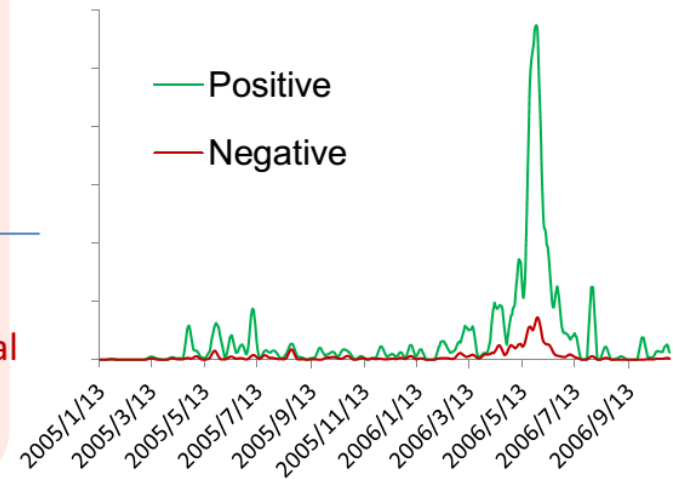
protesting... will lose your faith by watching the movie.



a good book to past time.

... so sick of people making such a big deal about a fiction **book**

Query="Da Vinci Code"



What did people like/dislike about "Da Vinci Code"?

<http://www.sigir2011.org/PDF/keynote-chengxiang-zhai.pdf>

Motivation (4/4)

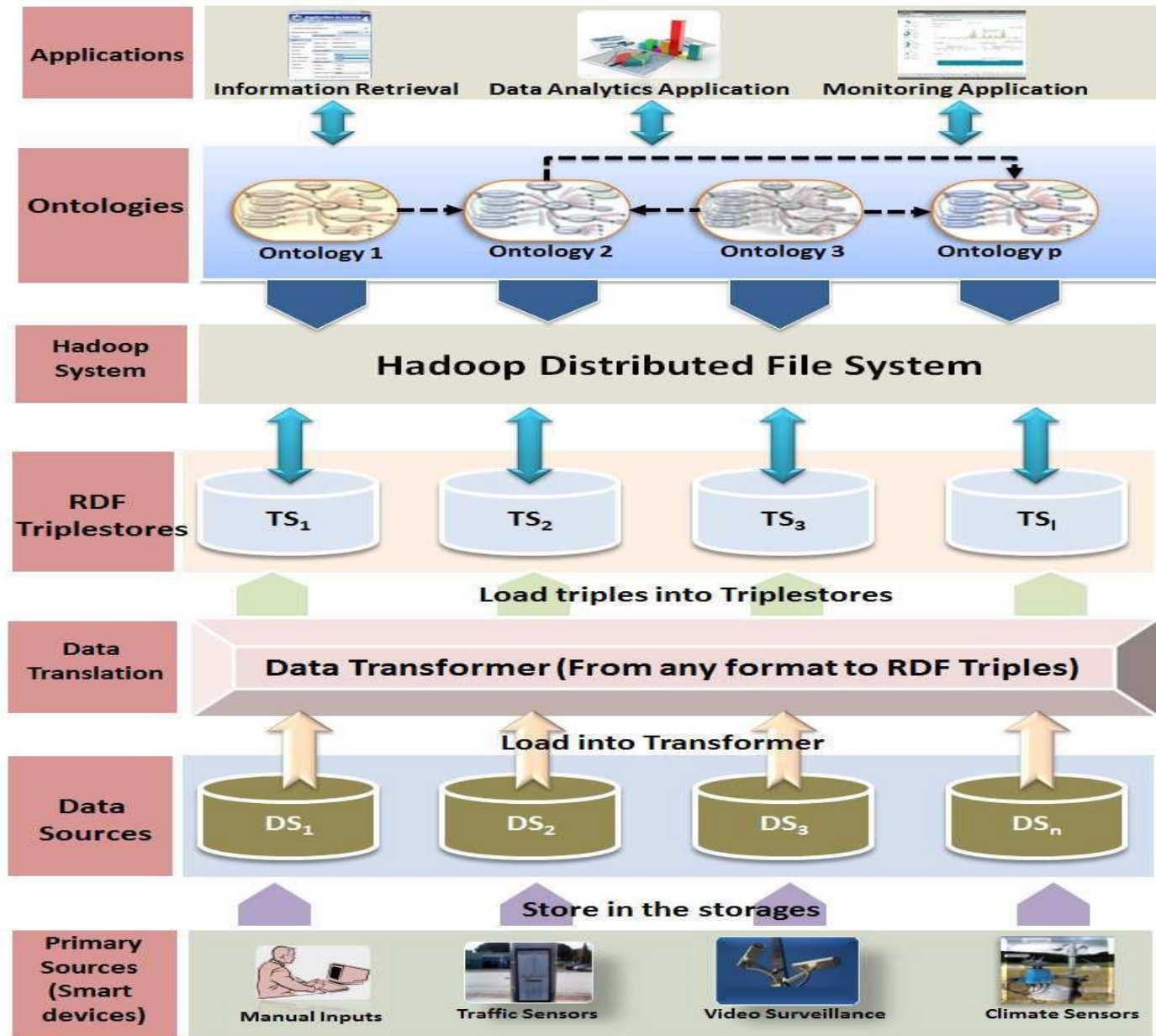
Query: phone number of people who authored a **Web service** related paper at the World Wide Web Conference 2011 (WWW'11)

List of papers and their topics

```
1 SELECT DISTINCT ?author ?phone WHERE {  
2 <http://www.org/conference/wsc/www2011/proceedings>  
3 wsc : hasPart ?pub  
4 ?pub swc : hasTopic ?topic .  
5 ?topic rdfs : label ?topicLabel .  
6 FILTER regex ( str (?topicLabel) , "web services" , "i" ) .  
7 ?pub swrc : author ?author .  
8 {?author owl:sameAs ?authAlt } UNION {?authAlt owl:sameAs  
?author}  
9 ?authAlt foaf : phone ?phone
```

The phone numbers are provided by the authors.

The names of the paper topics are provided by the sources authoritative for the URIs used to represent the topics;



Objectives

- **Aggregation:** composing relevant pieces of information, each piece partially contributes to the answer but together they form a complete response.
- **Queries:** look for objects that do not exist as such in the sources, but are built by assembling fragments.
- **Applications:** analytical tasks (opinion analysis, trend analysis, product comparison, risk analysis, event summarization, Web services engineering).
- **Existing systems:** Bibliometric systems (list of publications of an individual + analytical information (rate of citation for each publication, indicators like h-index, the list of co-authors)).

Challenges

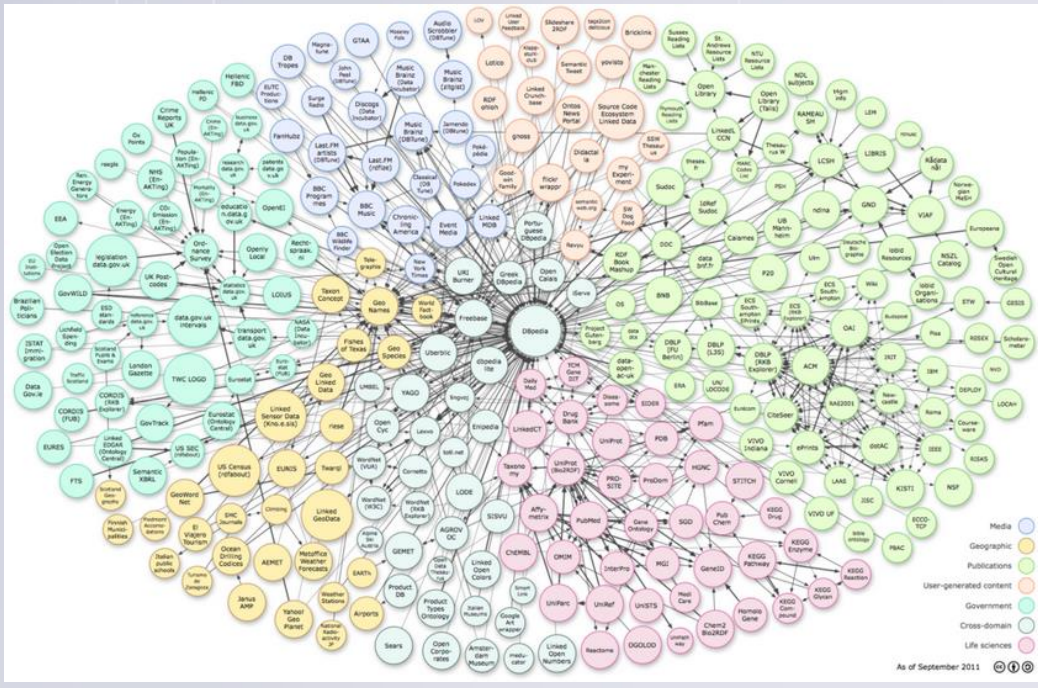
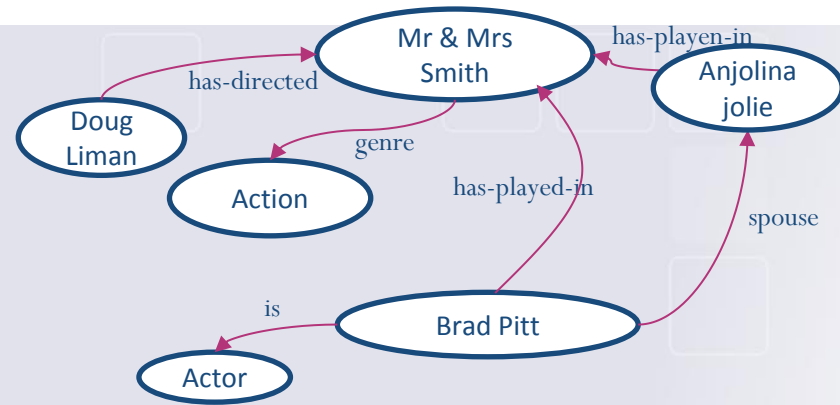
- **Semantic:** the *interpretation* of the query (the problems are related to the "vocabulary mismatch", the capture of the intent of the user) and the qualification of the results with regards to the initial (user) query.
- **Computational:** the *combinatorial* problem induced by the choice of fragments and multiple ways to aggregate them.

RDF - Resource Description Framework

(<Sujet>, <Prédicat>, <Objet>)

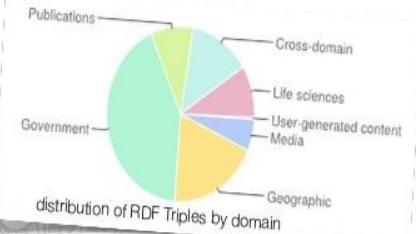


RDF Triple



Linking Open Data

Some statistics (as of 09/2011)



Domain	Number of datasets	Triples	%	(Out-)Links	%
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,328	7.11 %
Government	49	13,315,009,400	42.09 %	19,343,519	3.84 %
Publications	87	2,950,720,693	9.33 %	139,925,218	27.76 %
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	38.06 %
User-generated content	20	134,127,413	0.42 %	3,449,143	0.68 %
	295	31,634,213,770		503,998,829	

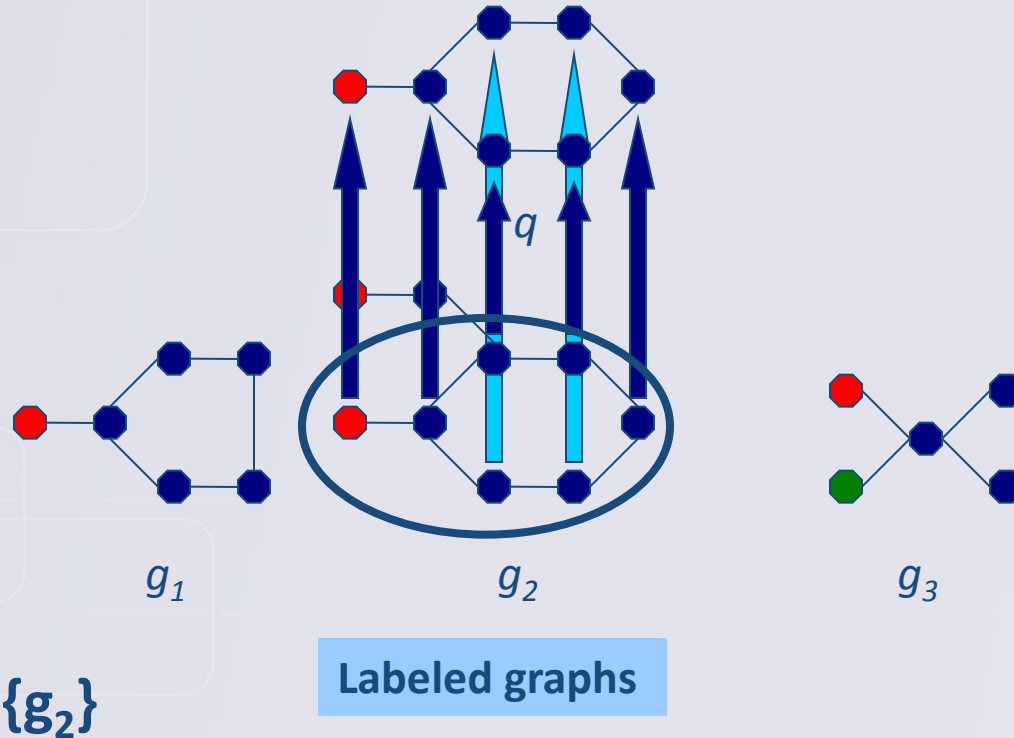
<http://www.linkeddata.org/>

Problem

■ Query

■ Data set

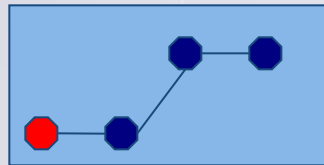
■ Answer



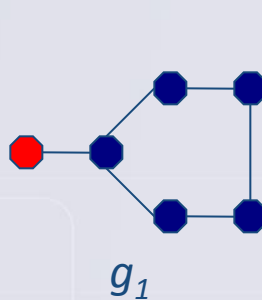
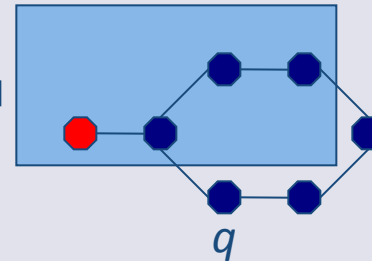
Exact matching

Approach: Filtering + Verification

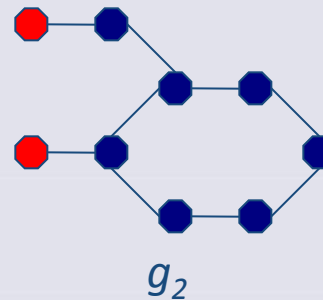
Pattern A:



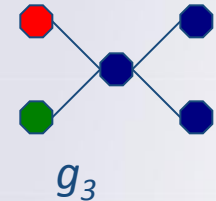
ID-List: $\{g_1, g_2\}$



g_1



g_2



g_3

Filtering:

Candidate (Pattern A)

Candidate (Pattern A)

Not relevant

Verification:

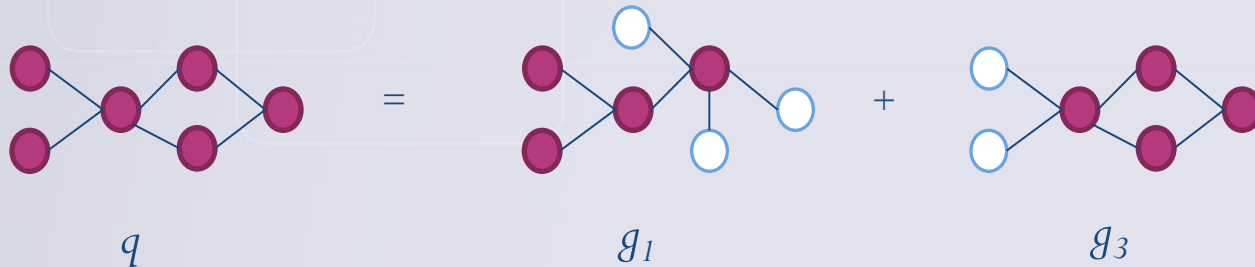
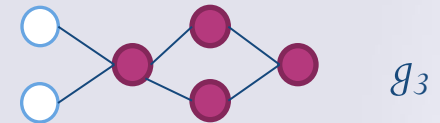
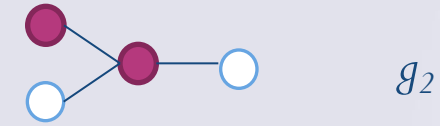
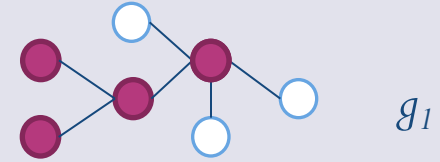
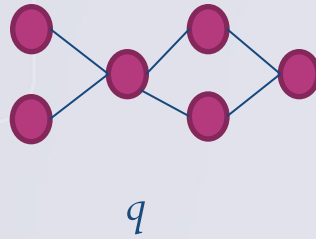
Not relevant

Solution

Existing approaches

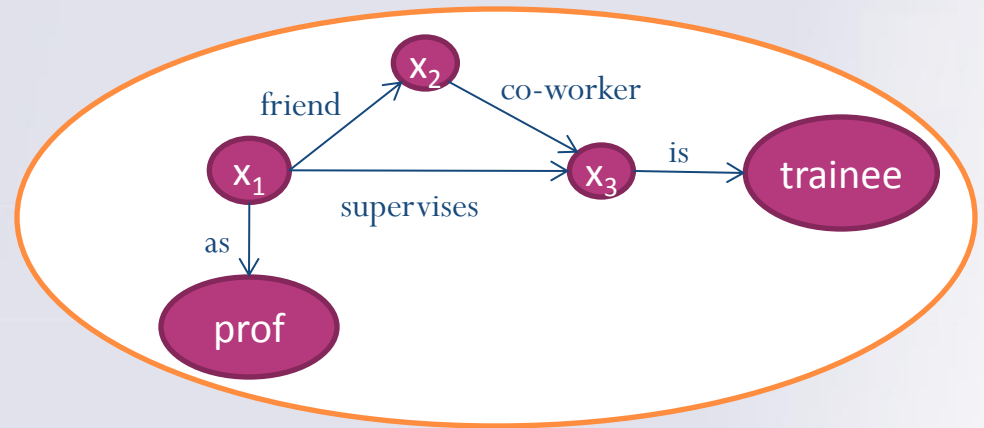
- **Graph mining:** mined patterns - paths
GraphGrep[ICPR'02], **trees** TreePi[ICDE'06] and
QuickSI[VLDB'08], **sub-graphs** FGIndex[SIGMOD'07] and
gIndex[SIGMOD'04]
- **Filtering** : index structure to speed up the search of
patterns in the query gIndex[SIGMOD'04],
FGIndex[SIGMOD'07], QuickSI[VLDB'08]
- **Verification** : graph isomorphism algorithms (improved
versions of Ullmann's algorithm) QuickSI[VLDB'08],
TreePi[ICDE'06], FGIndex[SIGMOD'07]

Aggregated Search



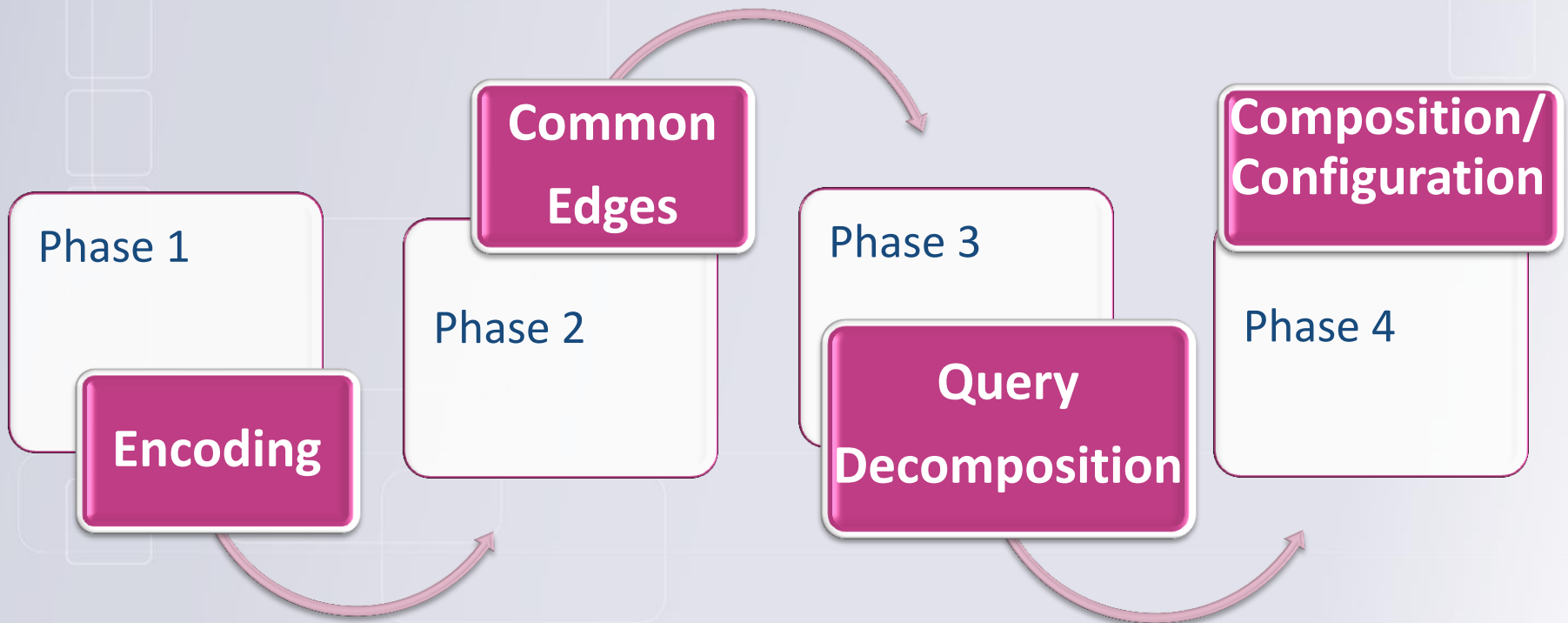
Queries

- Variables
- Constants (known resources)

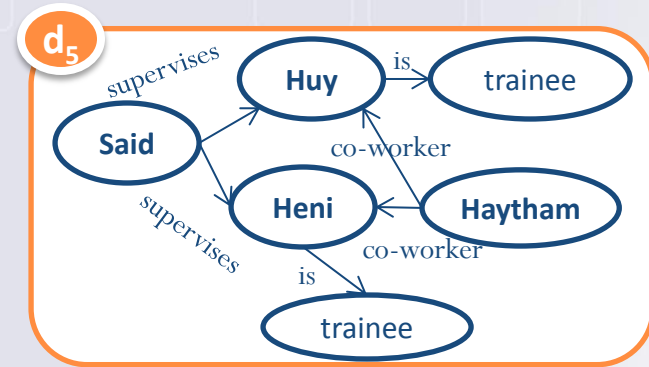
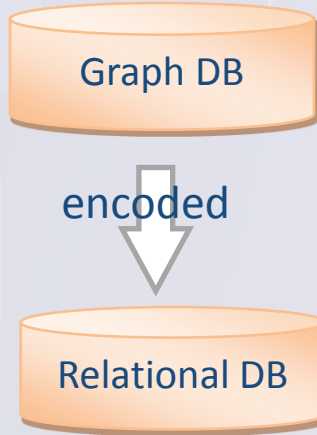
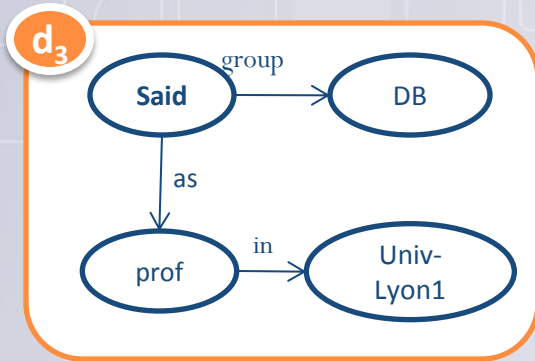


(SQL-based) Aggregated Search

General processing model



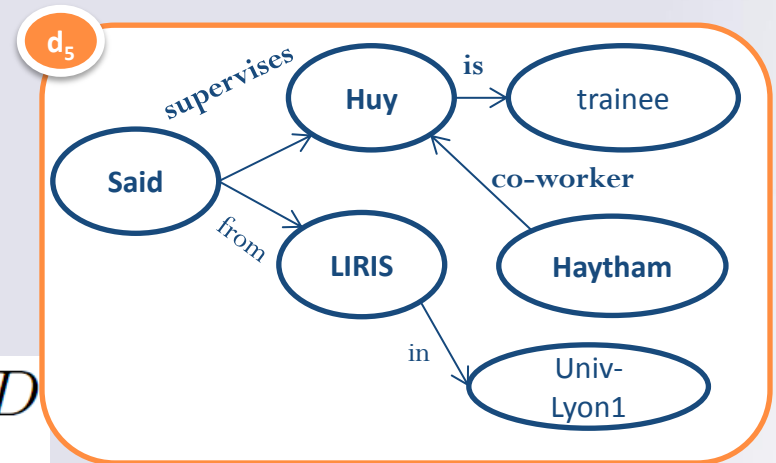
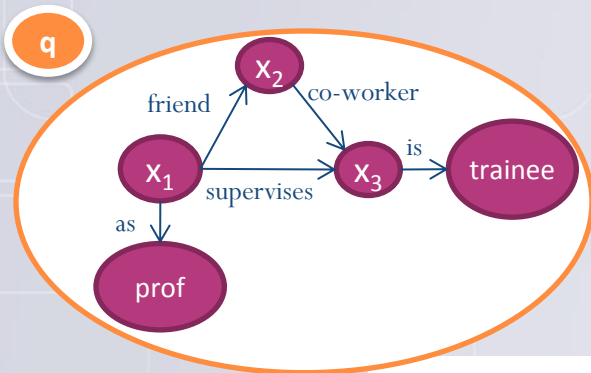
Phase 1: Edge-Edge Encoding Schemes (Sakr and Al-Naymat [1])



graphID	edgeID	eLabel	sVID	sVLabel	dVID	dVLabel
3	1	group	1	Said	2	DB
3	2	as	1	Said	3	prof
3	3	in	3	prof	4	Univ-Lyon1
5	4	supervises	5	Said	6	Huy
5	5	supervises	5	Said	7	Heni
5	6	is	6	Huy	8	trainee
5	7	is	7	Heni	9	trainee
5	8	co-worker	10	Haytham	6	Huy
5	9	co-worker	10	Haytham	7	Heni

Phase 2: Common edges search

- Discover all edges that belong to both the query graph q and a graph database D .
- Whether all edges in q are present in D .

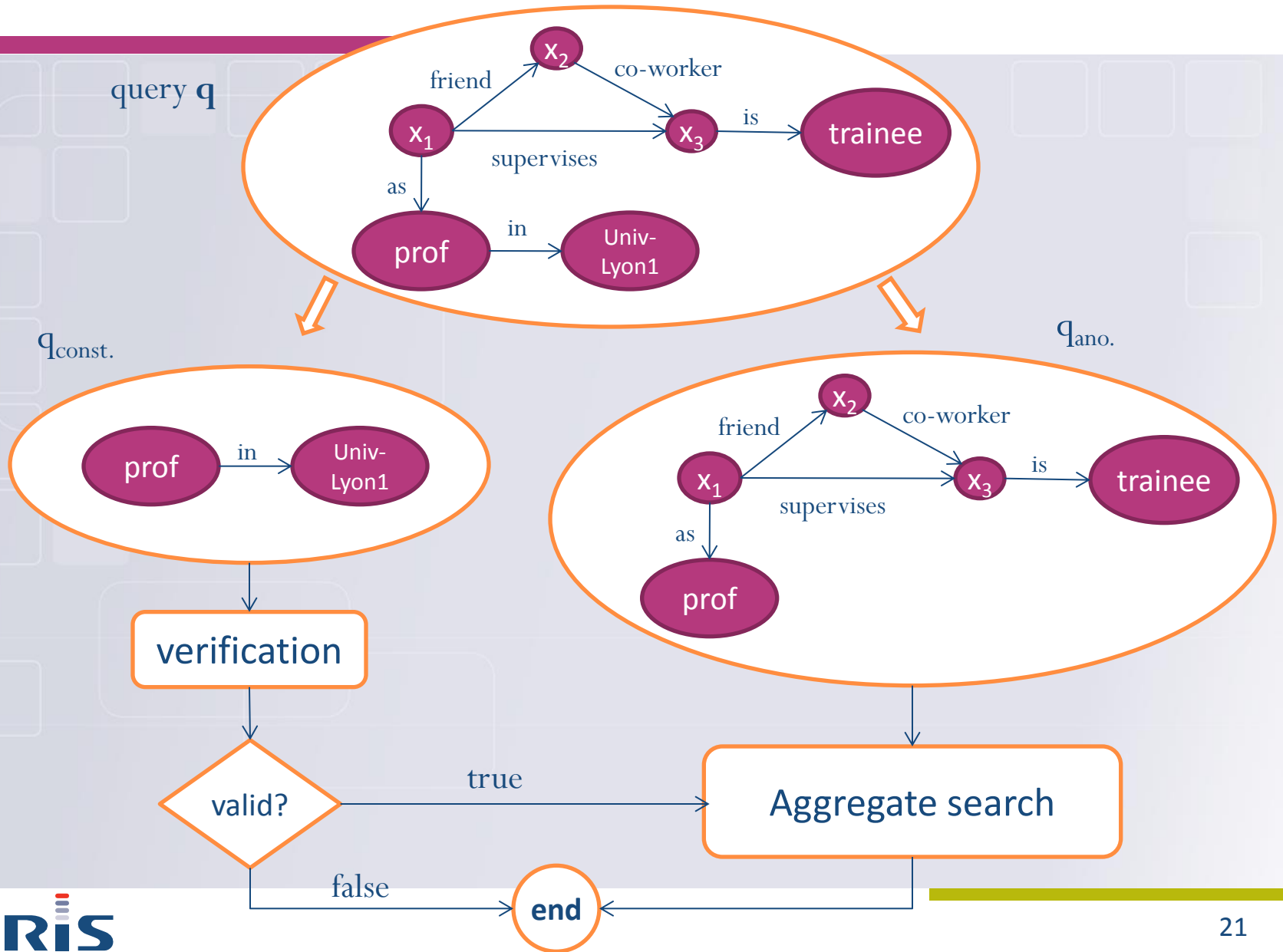


$$q \bowtie_{q.eLabel=D.eLabel} D$$

graphIDQ	graphID	edgeIDQ	edgeID	eLabel	sVLabel	dVLabel
1	5	1	4	supervises	Said	Huy
1	5	2	5	is	Huy	trainee
1	5	3	6	co-worker	Haytham	Huy

commonedges table

Phase 3: Query Decomposition



Phase 4: Evaluation & Configuration/Composition

$v = v_1$

Search(q, v, AV, C, R)

$AV = \{v_1, v_2, \dots, v_n\}$
set of variables in
descending order of
degree

C : set of common
edges

R : final answer (set)

set of label candidates of v
 $S = \{a_1, a_2, \dots, a_n\}$

Verification

no

yes

$S = \{a_1, a_2, \dots, a_m\}, m \leq n$

For each a_i in S

$q_v = \text{query_generator}(q, a_i);$

$R \leftarrow \{q\}$

$v = \text{nextVertex}(AV);$

yes

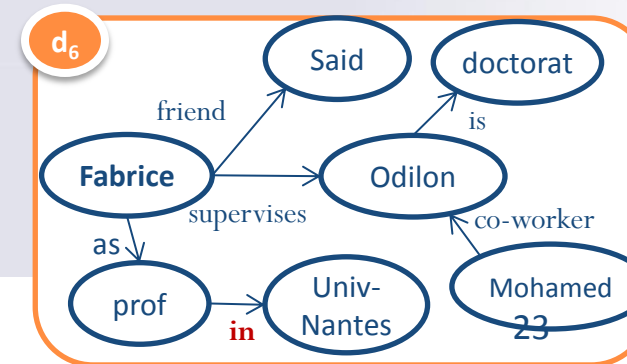
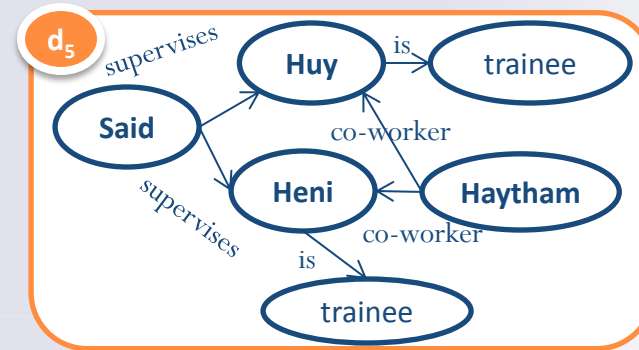
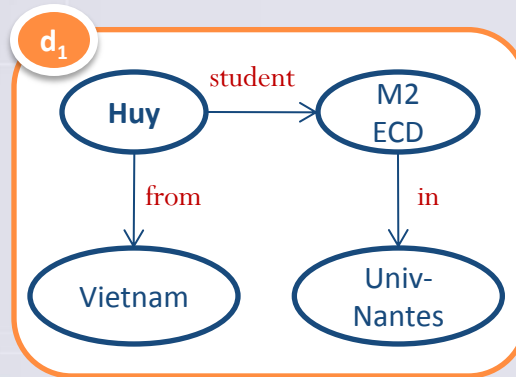
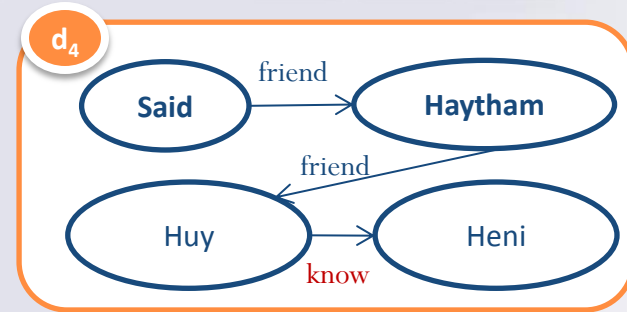
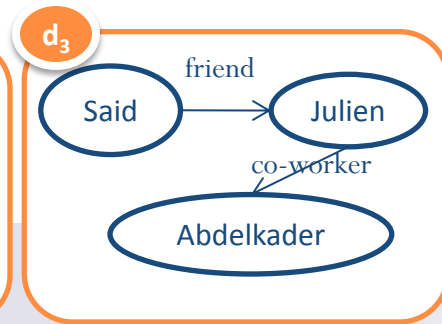
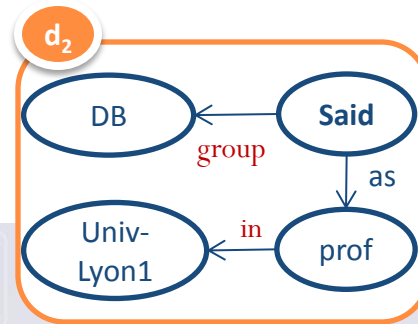
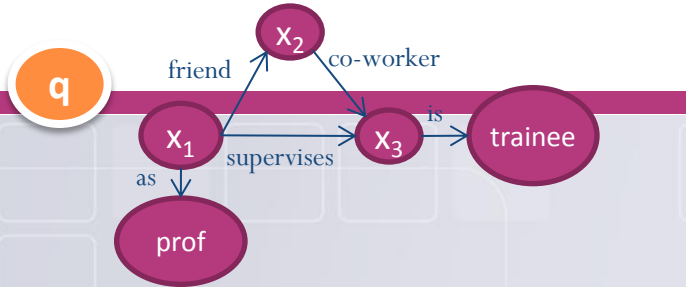
$v = \text{NULL}$

no

Search(q_v, v, AV, C, R)

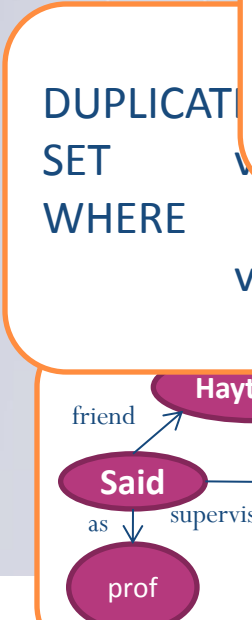
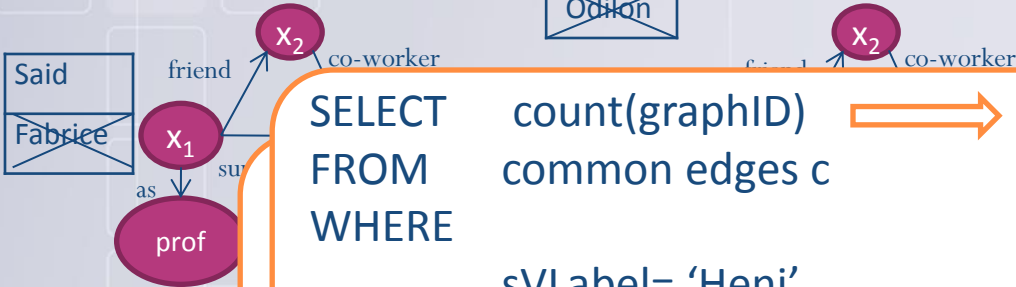
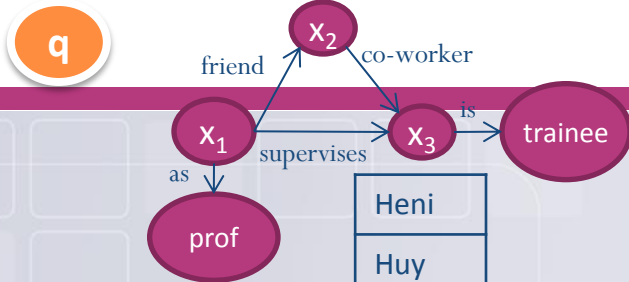
End

Common edges search



Evaluation & Configuration/Composition

1. Label - candidates
2. Validation
3. Query generator
4. Next variable
5. Recursion



```

SELECT count(graphID)
FROM common edges c
WHERE
  sVLabel= 'Heni'
  AND dVLabel = 'trainee'
  AND eLabel = 'is';
  
```

1

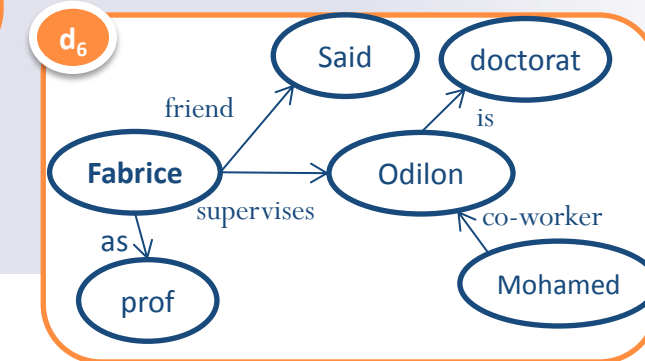
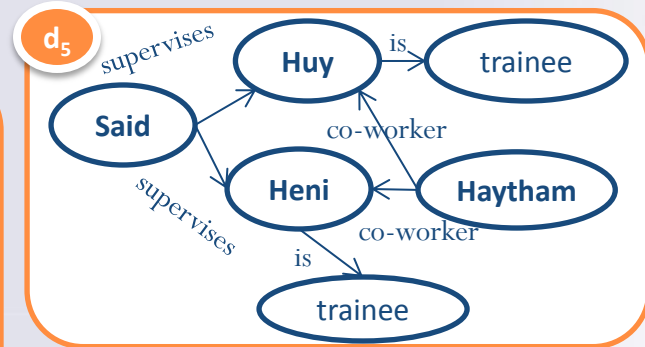
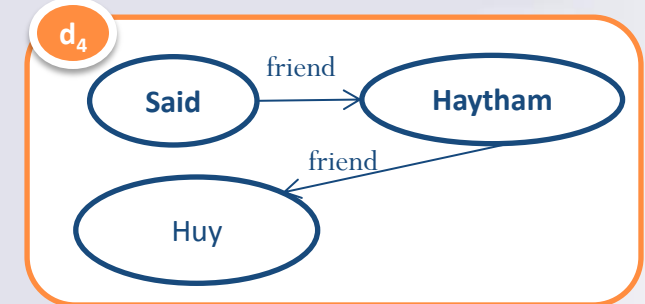
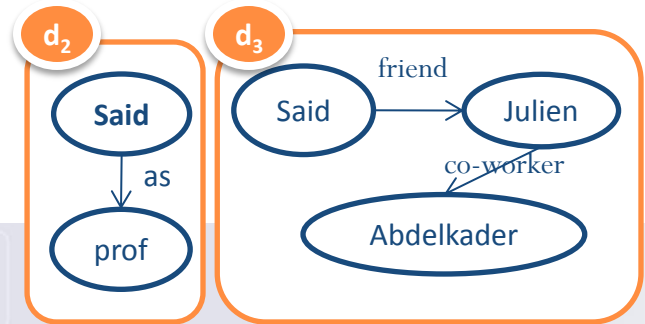
```

SELECT count(graphID)
FROM common edges c
WHERE
  sVLabel= 'Odilon'
  AND dVLabel = 'trainee'
  AND eLabel = 'is';
  
```

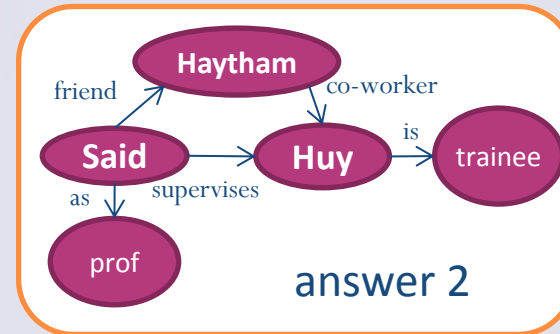
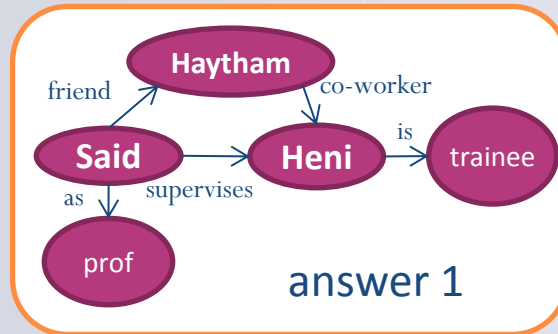
0

answer 1

answer 2

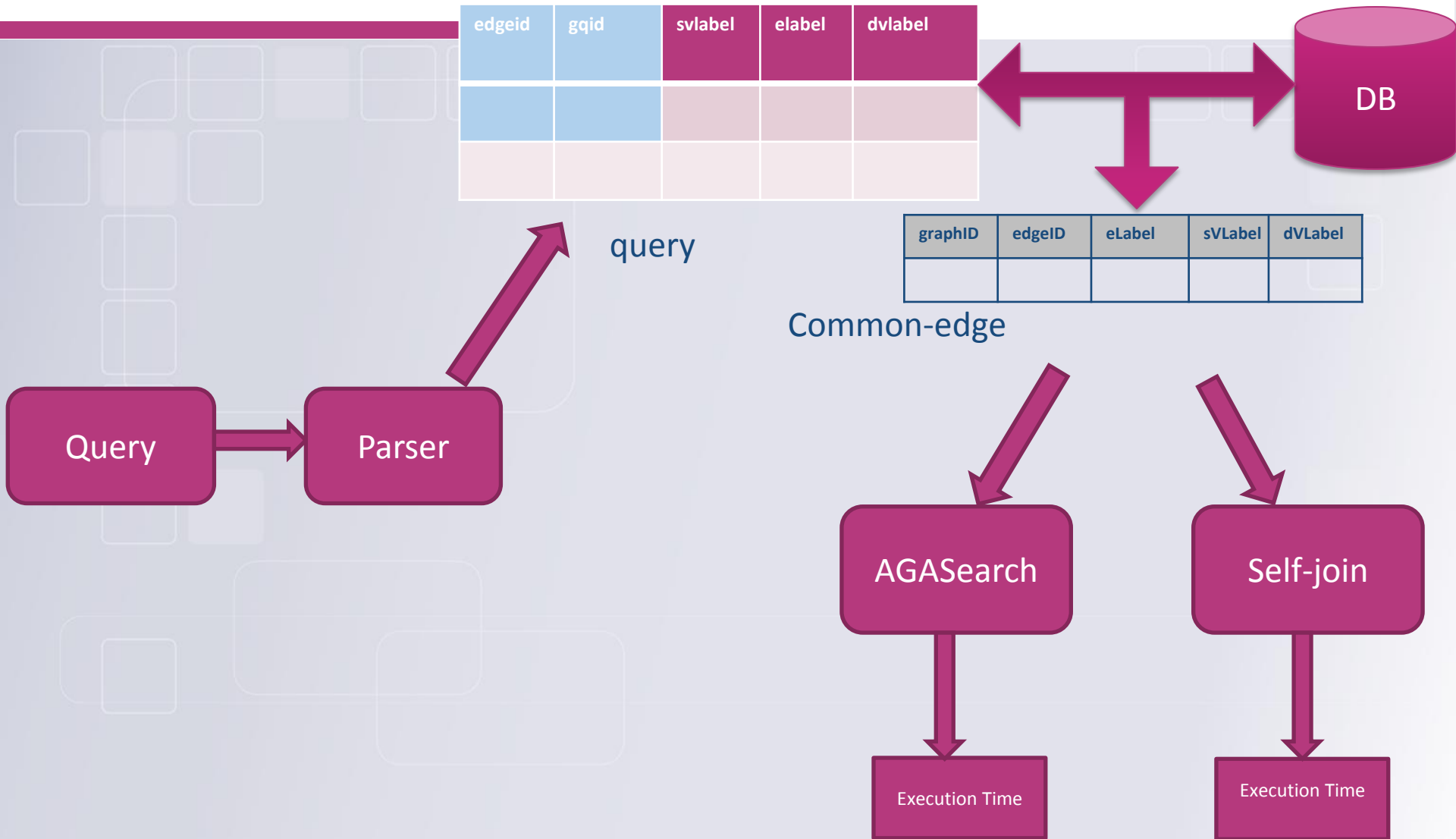


Final answer set



#	x_1	x_2	x_3
1	Said	Haytham	Heni
2	Said	Haytham	Huy

Architecture



Benchmark : Data Generation

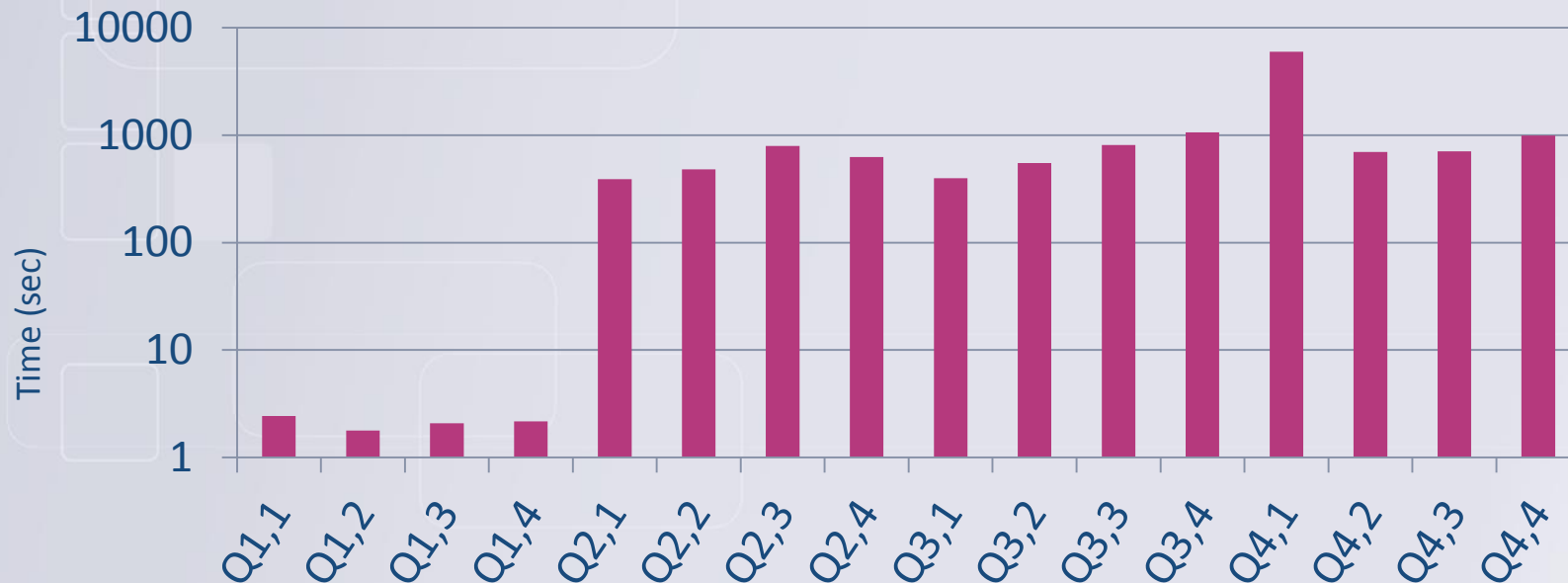
The Berlin SPARQL Benchmark (BSBM) [2]

Data Sets	Size	#Triples	Execution Time	Execution Time (self-join)
Edge2edge100	25,6 Mb	105124	7,56 min	2,215 s
Edge2edge100_ind	91,5 Mb	105124	3,963 min	1,117 s
Edge2edge10M	2,4 Gb	10036982	>35 min	>15 min
Edge2edge10M-ind	7,9 Gb	10036982	>20 min	20,642 s
Edge2edge100M	45,4 Gb	189905757	>40 min	>30 min

[2] Bizer C. & Schultz A. (2009). *The berlin sparql benchmark*.

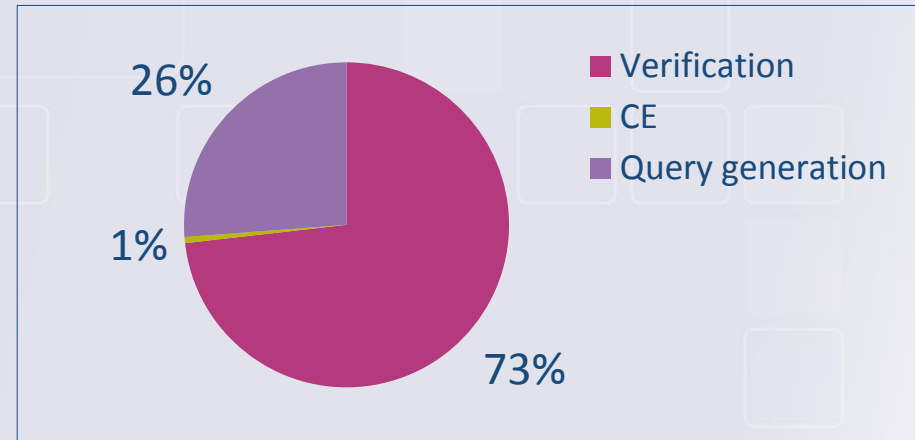
Experiments

- Benchmark Berlin (BSBM)
- Data Set 105 118 triples
- Queries: 16 queries ($Q_{i,j}$ with i :#variables, j :#constantes)
- Environnement: RAM (8 GB), Processor (3.2 GHz)

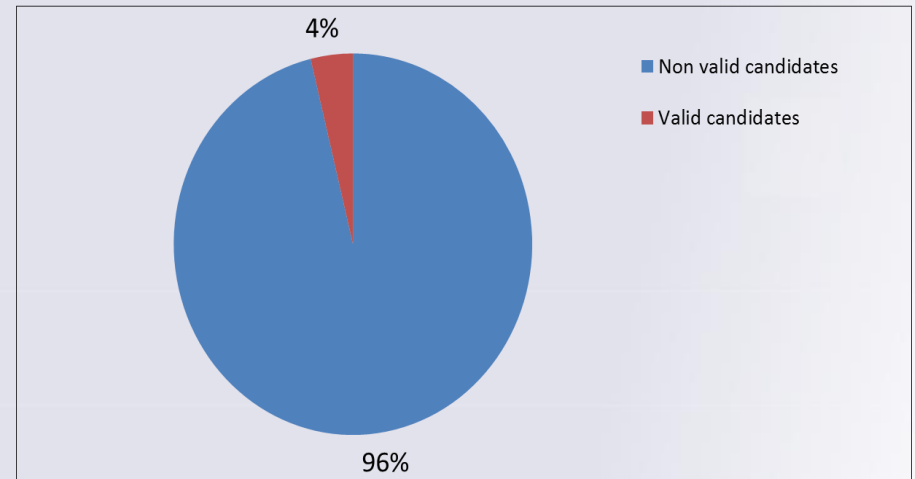


Cost of each phase

Requête	#ValidCandidates	#NonValidCandidates
Q1,1	100	0
Q1,2	200	89
Q1,3	300	89
Q1,4	400	183
Q2,1	12879	12579
Q2,2	21380	21277
Q2,3	32059	31950
Q2,4	32170	31951
Q3,1	13079	12579
Q3,2	21502	21372
Q3,3	32181	31951
Q3,4	42740	42529
Q4,1	162879	157757
Q4,2	22880	22745
Q4,3	22856	22739
Q4,4	44238	44101

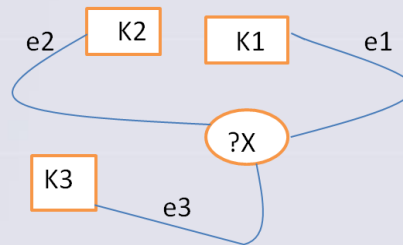
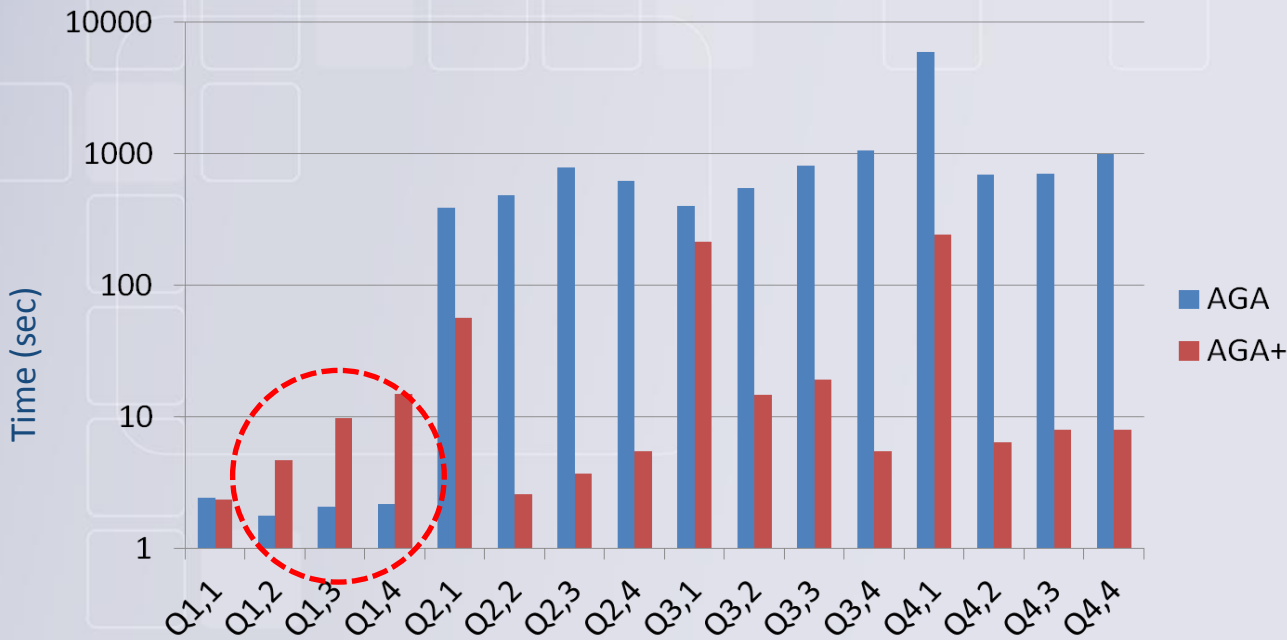


Verification time > 70% of the total time



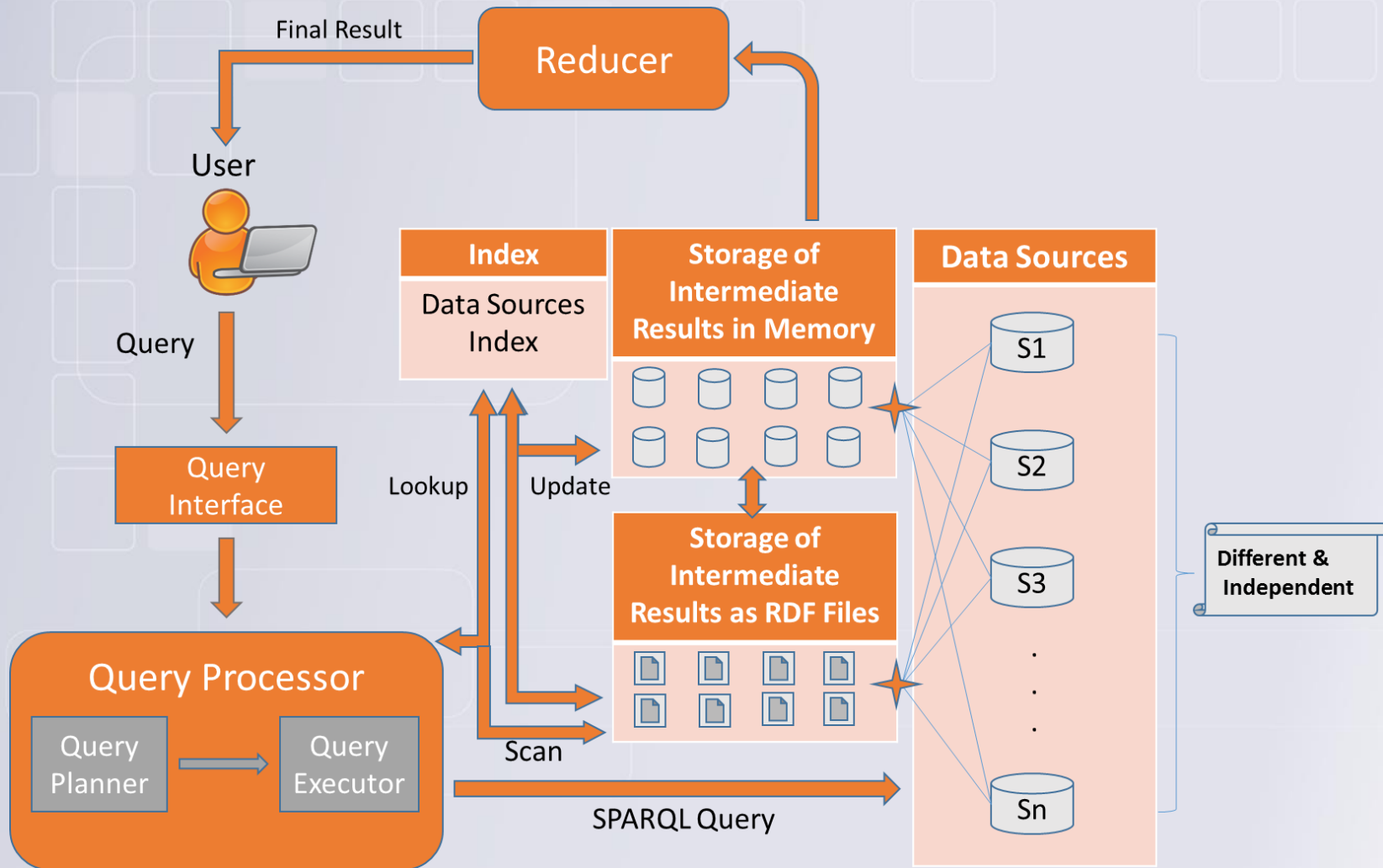
More than 95% of the selected candidates are not relevant

Results



	AGA	AGA+
Q1,1	2,42	2,37
Q1,2	1,78	4,66
Q1,3	2,08	9,73
Q1,4	2,16	15,01
Q2,1	388,71	56,39
Q2,2	481,64	2,60
Q2,3	792,38	3,70
Q2,4	626,54	5,47
Q3,1	398,54	214,17
Q3,2	549,11	14,76
Q3,3	808,67	19,14
Q3,4	1 065,16	5,50
Q4,1	5 974,79	241,55
Q4,2	698,15	6,37
Q4,3	705,66	7,95
Q4,4	995,51	8,01

iseeker



Example

A SPARQL Query Which Topic is Politics:

```
SELECT ?president ?party ?page
WHERE
{
  ?president <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/President> .
  ?president <http://dbpedia.org/ontology/nationality> <http://dbpedia.org/resource/United_States> .
  ?president <http://dbpedia.org/ontology/party> ?party .
  ?x <http://data.nytimes.com/elements/topicPage> ?page .
}
```

Three SPARQL EndPoints:

<http://dbpedia.org/sparql>

<http://ibm.rkbexplorer.com/sparql/>

<http://factforge.net/sparql/>

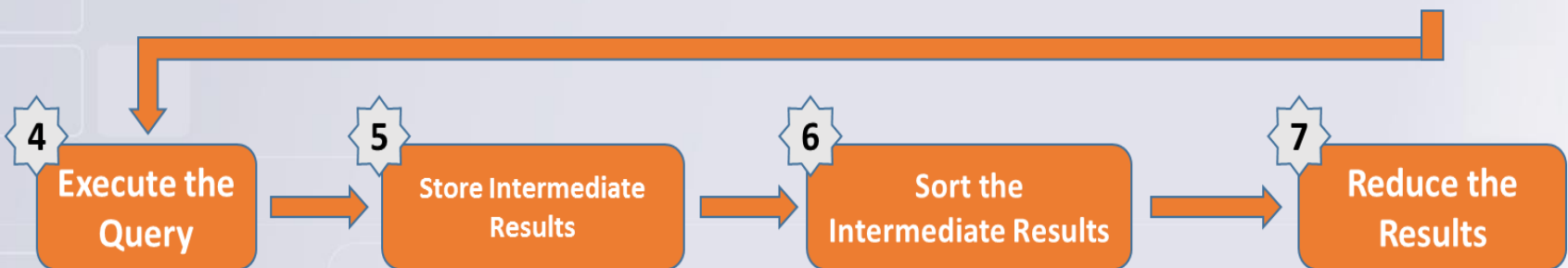




Q1: ?president <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/President> .
 Q2: ?president <http://dbpedia.org/ontology/nationality> <http://dbpedia.org/resource/United_States> .
 Q3: ?president <http://dbpedia.org/ontology/party> ?party .
 Q4: ?x <http://data.nytimes.com/elements/topicPage> ?page .

- Scan the index table to locate the endpoints where the data blocs reside.
 - Initially the index table is empty.
 - The index table may not contain information about the triple pattern that was never executed.

- The executor allows to choose which triple pattern (subquery) should be sent to which source(s).
- Choose for each subquery the source(s) that can provide results



- Send queries to the endpoints of sources.

- Fetch the results from the remote server.
- Store intermediate results in memory or in RDF files.
- Update the index table.

- Sort the intermediate results and remove redundant triple patterns.

- Combine the results sorted in the previous step and store them into RDF Files.

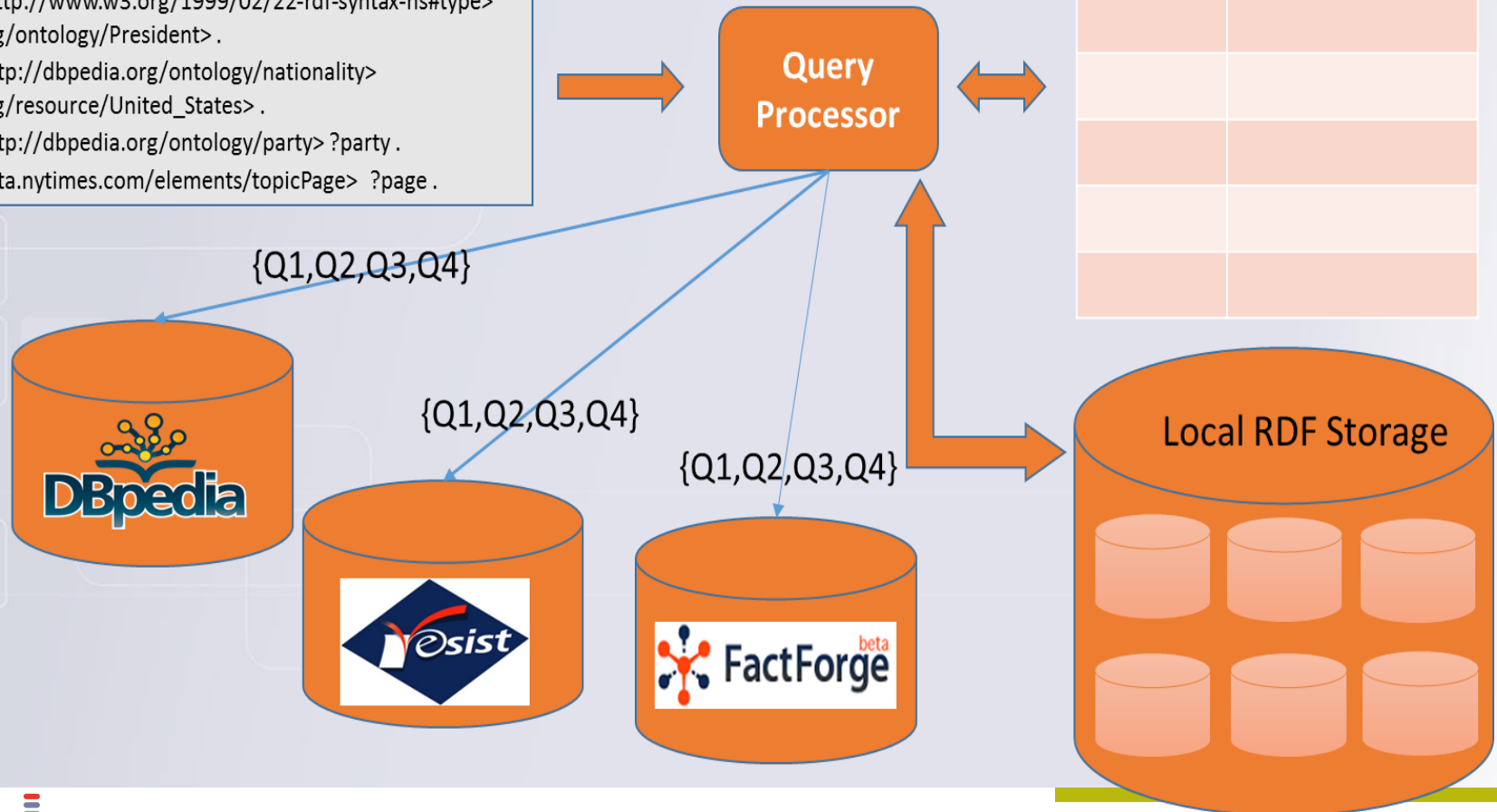
Note:

- The index table and the local RDF storage are built incrementally.
- Each query execution may result in new entries in both of them.

Index

Key	Value

Q1: ?president <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/President> .
Q2: ?president <http://dbpedia.org/ontology/nationality> <http://dbpedia.org/resource/United_States> .
Q3: ?president <http://dbpedia.org/ontology/party> ?party .
Q4: ?x <http://data.nytimes.com/elements/topicPage> ?page .



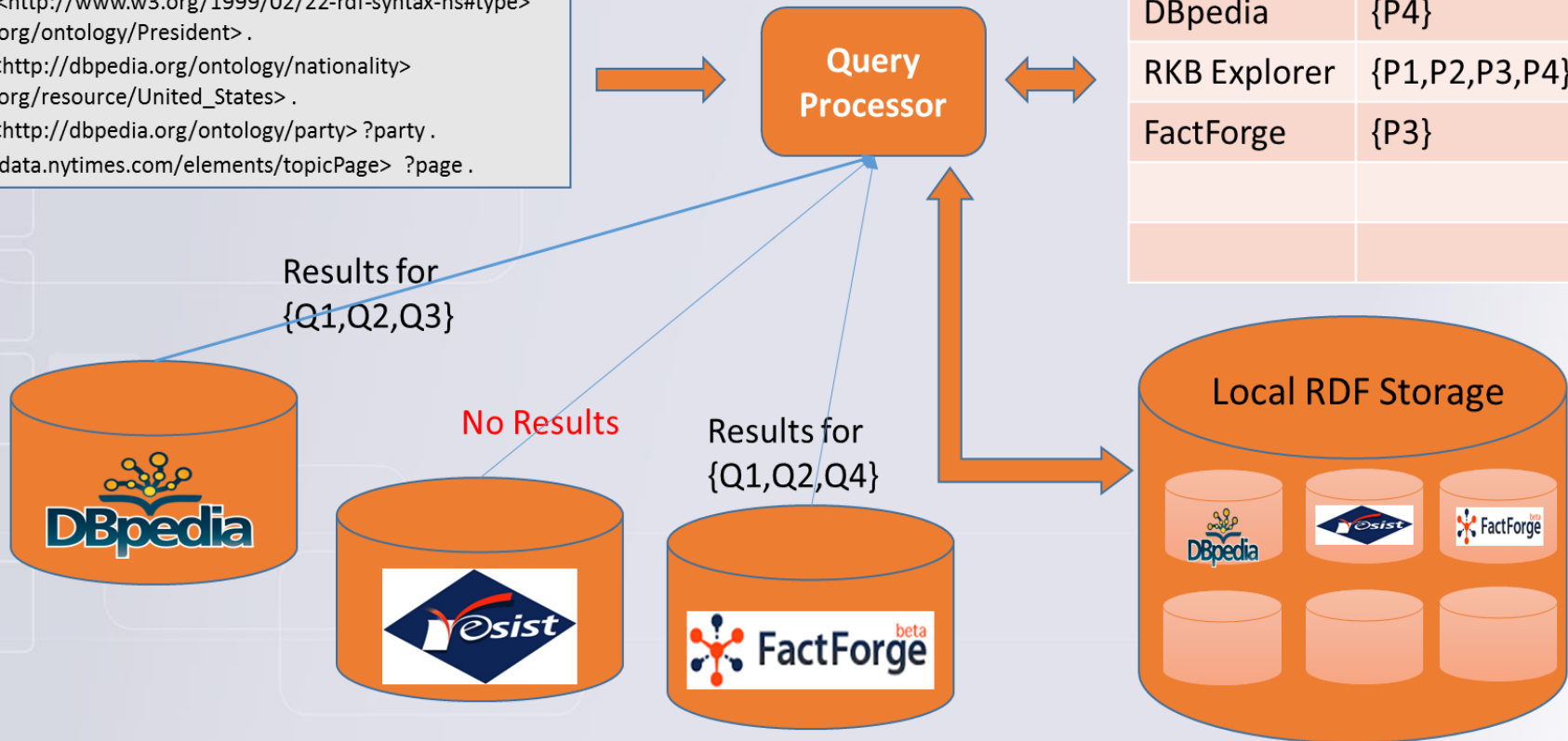
Note:

- The index table and the local storage are updated.

Index

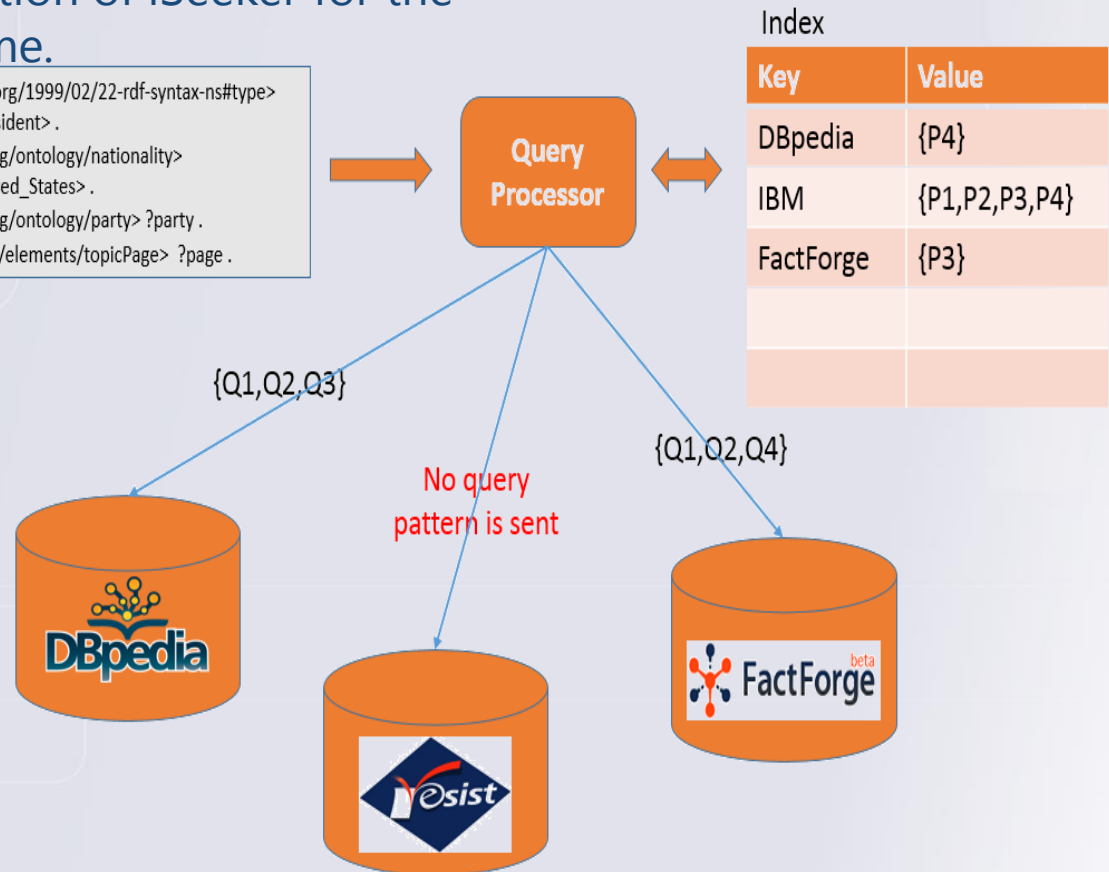
Key	Value
DBpedia	{P4}
RKB Explorer	{P1,P2,P3,P4}
FactForge	{P3}


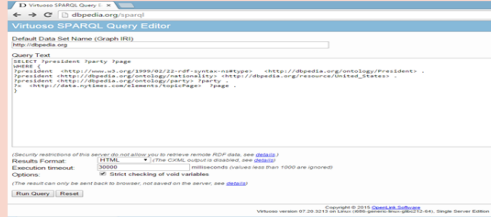

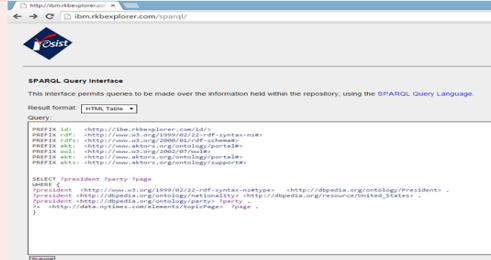

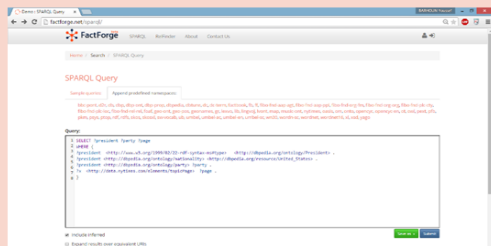
```
Q1: ?president <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/President> .
Q2: ?president <http://dbpedia.org/ontology/nationality>
<http://dbpedia.org/resource/United_States> .
Q3: ?president <http://dbpedia.org/ontology/party> ?party .
Q4: ?x <http://data.nytimes.com/elements/topicPage> ?page .
```






❖ The execution of iSeeker for the second time.

Q1: ?president <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
 <http://dbpedia.org/ontology/President> .
Q2: ?president <http://dbpedia.org/ontology/nationality>
 <http://dbpedia.org/resource/United_States> .
Q3: ?president <http://dbpedia.org/ontology/party> ?party .
Q4: ?x <http://data.nytimes.com/elements/topicPage> ?page .



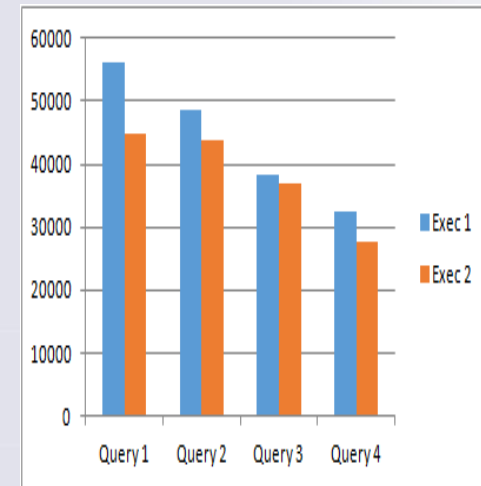
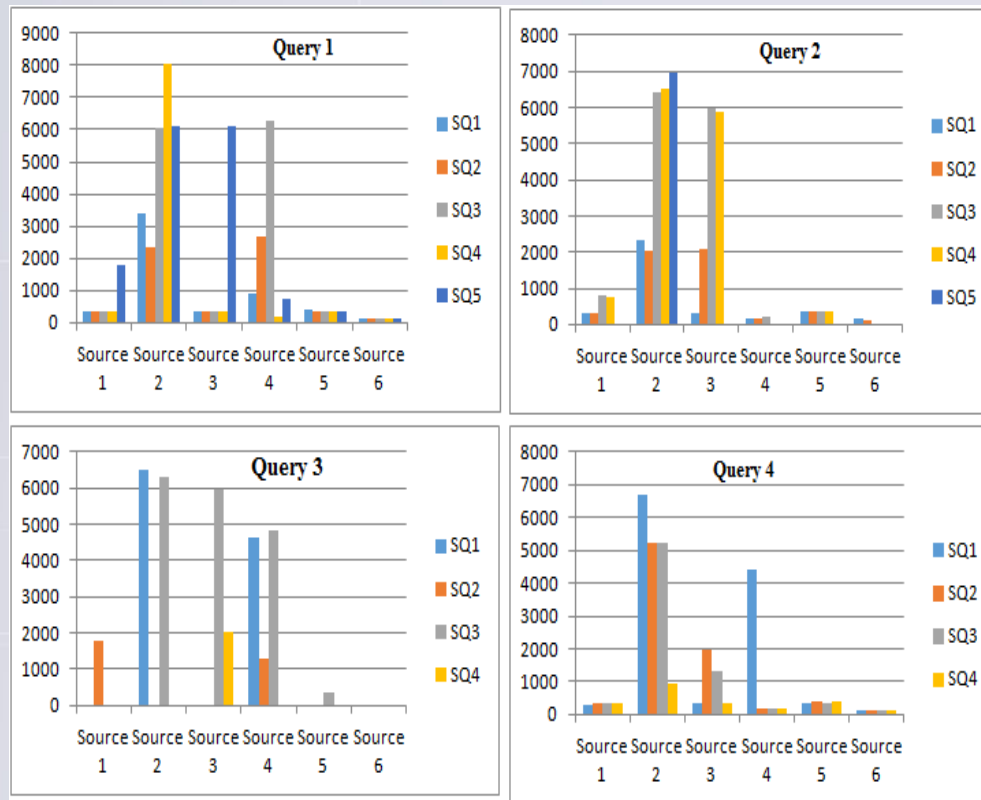
Endpoints	Screenshots of Execution	Results
		No Result
		No Result
		No Result

Endpoints	SubQueries	Number of Triples Results	Execution Times (Executor)	Size Of the RDF File
	Q1	2206	1 s 593 ms	307 KB
	Q2	10000	2 s 836 ms	1315 KB
	Q3	10000	4 s 934 ms	1426 KB
	Q4	0	0 s 187 ms	0 KB
	Q1	0	0 s 251 ms	0 KB
	Q2	0	0 s 234 ms	0 KB
	Q3	0	0 s 249 ms	0 KB
	Q4	0	0 s 235 ms	0 KB
	Q1	4505	2 s 644 ms	600KB
	Q2	0	0 s 297 ms	0 KB
	Q3	68838	33 s 532 ms	8889 KB
	Q4	36248	20 s 339 ms	6566 KB

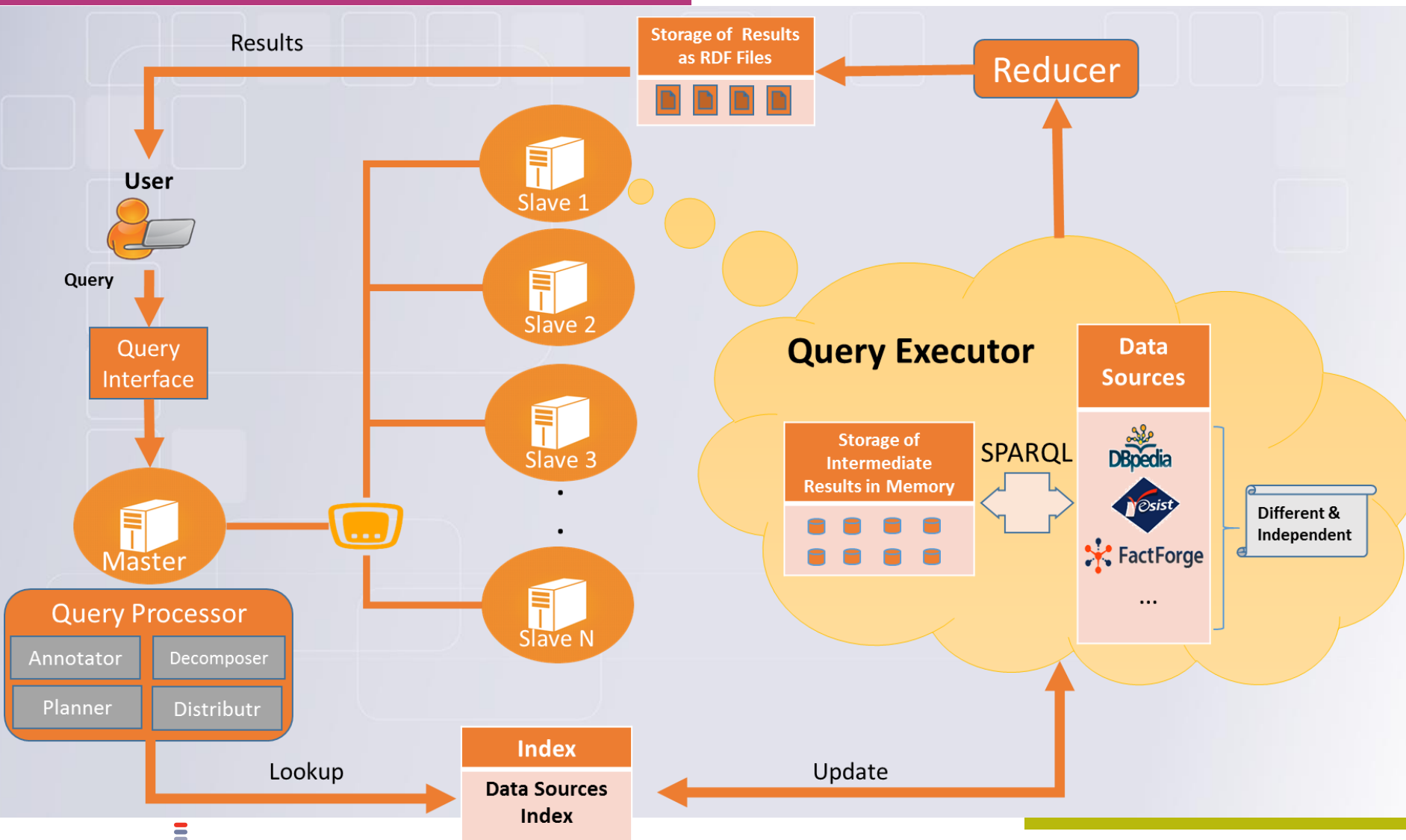
* Time Execution of Reducer when it's performing alone After Executor finished : 0 s 125 ms

SubQueries	Number of triples before Reducer	Number of triples after Reducer	Number of redundant triples removed
Q1	6711	4744	1967
Q2	10000	10000	0
Q3	78838	78313	525
Q4	36248	36248	0

We selected four Sparql queries and six endpoints recommended by Fedbench. The queries are available online(*), The selected queries cover three different topics including politics, movies, and geographical location.



* <https://code.google.com/p/fbench/>



Query Annotator (1/2)

```
SELECT DISTINCT ?drug ?enzyme ?reaction
```

```
Where {
```

```
  ?drug1 drugCategory antibiotics .
```

```
  ?drug2 drugCategory antiviralAgents .
```

```
  ?drug3 drugCategory antihypertensiveAgents .
```

```
  ?I1 interactionDrug2 ?drug1 .
```

```
  ?I2 interactionDrug1 ?drug .
```

```
  ?I1 interactionDrug1 ?drug .
```

```
  ?I2 interactionDrug2 ?drug2 .
```

```
  ?I3 interactionDrug2 ?drug3 .
```

```
  ?I3 interactionDug1 ?drug .
```

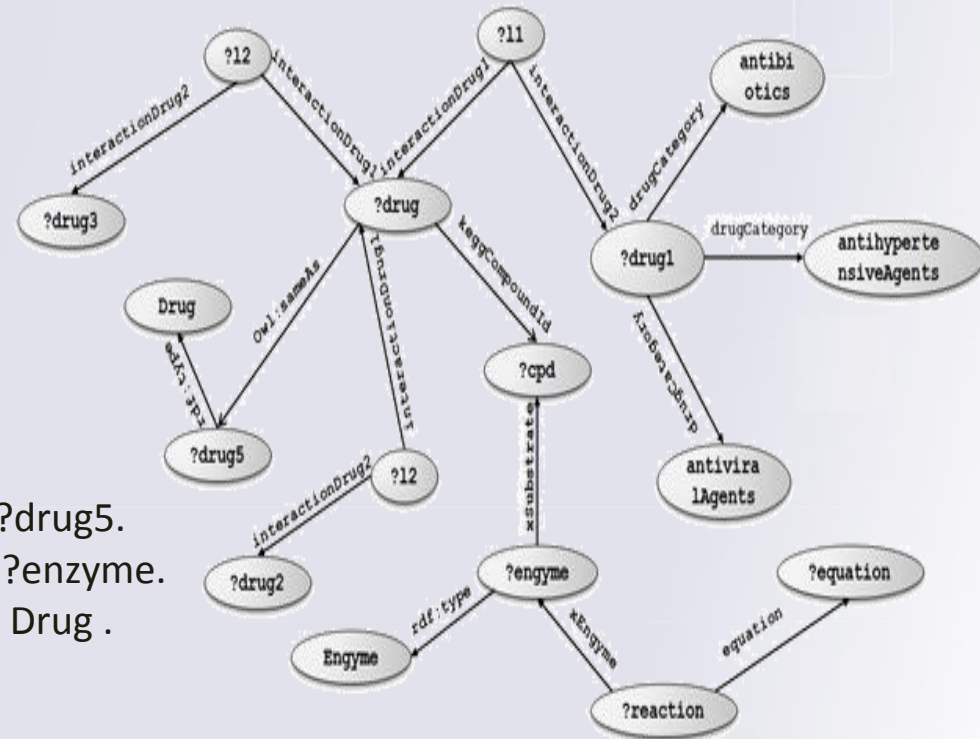
```
  ?drug keggCompoundId ?cpd .
```

```
  ?enzyme xSubstrate ?cpd . ?drug owl:sameAs ?drug5.
```

```
  ?enzyme rdf:type Enzyme . ?reaction xEnzyme ?enzyme.
```

```
  ?reaction equation ?equation . ?drug5 rdf:type Drug .
```

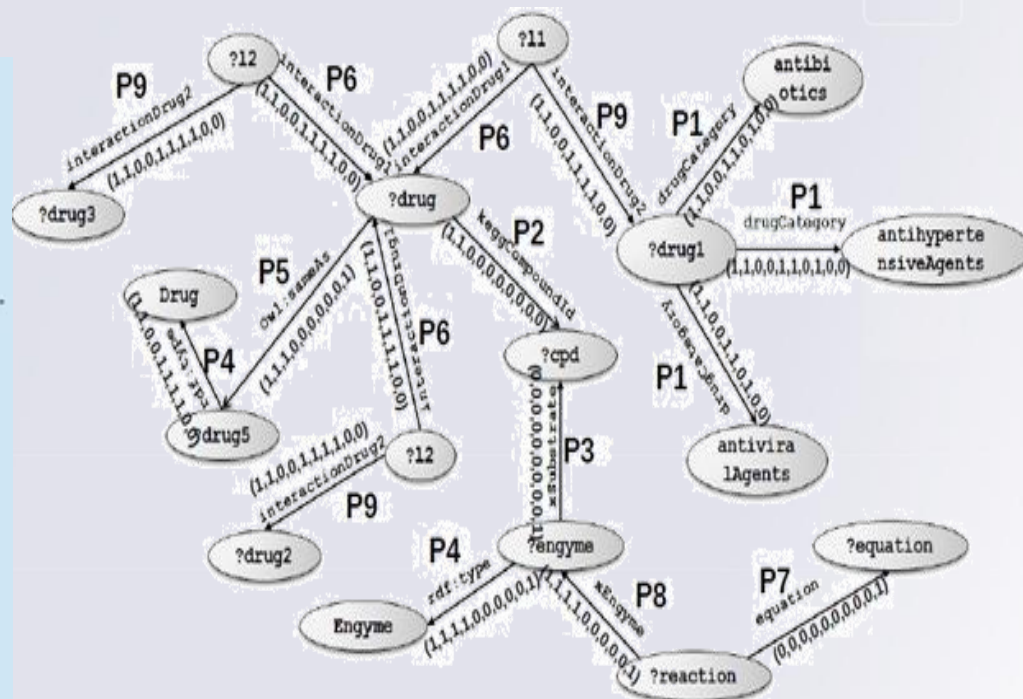
```
}
```



Query Annotator (2/2)

Fu-Berlin Drugbank (S1), Fu-Berlin DBCategory (S2), DBPedia (S3), ChEBI (S4), SIDER(S5), LinkedCT (S6), Pittsburg repository (S7), DailyMed (S8), KEGG (S9) and Bio2RDF(S10).

No	Predicates	Sources
01	DrugCategory (P1)	Fu-Berlin Drugbank, Fu-Berlin DBCategory, SIDER, LinkedCT, DailyMed (Total 5 sources)
02	KeggCompoundId (P2)	Fu-Berlin Drugbank, Fu-Berlin DBCategory (Total 2 sources)
03	Xsubstrate (P3)	KEGG (Total 1 source)
04	rdf:type (P4)	DBPedia, KEGG, ChEBI, Fu-Berlin Drugbank, Fu-Berlin, DBCategory (Total 6 sources)
05	owl:sameAs (P5)	Fu-Berlin Drugbank, Fu-Berlin DBCategory, KEGG, DBPedia (Total 4 sources)
06	InteractionDrug1 (P6)	Fu-Berlin Drugbank, Fu-Berlin DBCategory, SIDER, LinkedCT, DailyMed, Pittsburg Repository (Total 6 sources)
07	Equation (P7)	KEGG (Total 1 source)
08	Xenzyme (P8)	KEGG, Fu-Berlin Drugbank, Fu-Berlin DBCategory, ChEBI, DBPedia (Total 5 sources)
09	InteractionDrug2 (P9)	Fu-Berlin Drugbank, Fu-Berlin DBCategory, SIDER, LinkedCT, DailyMed, Pittsburg Repository (Total 6 sources)



Highly Connected Graph Clustering Algorithms

Matrix of First Iteration

Predicates	P1	P2	P3	P4	P5	P6	P7	P8	P9
P1	-	2	0	5	2	5	0	2	5
P2	-	-	0	2	2	2	0	2	2
P3	-	-	-	0	0	0	1	1	0
P4	-	-	-	-	3	6	1	5	6
P5	-	-	-	-	-	2	1	4	2
P6	-	-	-	-	-	-	0	2	6
P7	-	-	-	-	-	-	-	1	2
P8	-	-	-	-	-	-	-	-	0
P9	-	-	-	-	-	-	-	-	-

Matrix of Second Iteration

Predicates	P1	P2	P3	P5	C1(P6+P9+P4)	P7	P8
P1	-	2	0	2	5	0	2
P2	-	-	0	2	2	0	2
P3	-	-	-	0	0	1	1
P5	-	-	-	-	2	1	4
C1(P6+P9+P4)	-	-	-	-	-	0	2
P7	-	-	-	-	-	-	1
P8	-	-	-	-	-	-	-

Sixth and Final Iteration

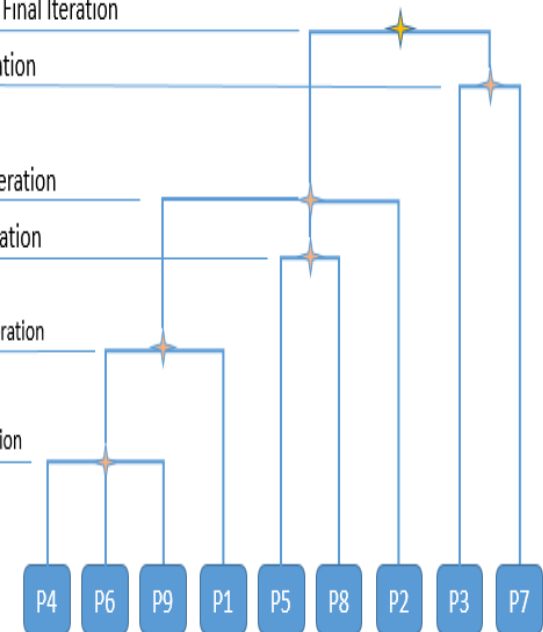
Fifth Iteration

Fourth Iteration

Third Iteration

Second Iteration

First Iteration



MERCI