



**HAL**  
open science

## Data Discovery in RDF Graphs

Ioana Manolescu

► **To cite this version:**

Ioana Manolescu. Data Discovery in RDF Graphs. DEXA 2017 - 28th International Conference on Database and Expert System Applications, Aug 2017, Lyon, France. pp.1-63. hal-01657144

**HAL Id: hal-01657144**

**<https://inria.hal.science/hal-01657144>**

Submitted on 6 Dec 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Data Discovery in RDF Graphs

Ioana Manolescu

INRIA and Ecole Polytechnique, France

`ioana.manolescu@inria.fr`

`http://pages.saclay.inria.fr/Ioana.Manolescu`

DEXA Conference, Aug 29, 2017

# Outline

- 1 Background: **semantic** RDF graphs
- 2 **Summarizing** semantic-rich RDF graphs [ČGM15a, ČGM15b, ČGM17a]
  - Joint work with Šejla Čebirić (Inria) and François Goasdoué (U. Rennes 1 and Inria)
- 3 Finding **insights** in RDF graphs [DMS17]
  - Joint work with Yanlei Diao and Shu Shang (Ecole Polytechnique and Inria)

# Part I

## Background: RDF graphs

# Big Data needs semantics

AI Magazine, Spring 2015



**DATA.GOV**

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

DATA CATALOG

# / Datasets

Filter by location:

93 datasets found for "Natural Disaster"

Order by: Relevance

**Natural Disaster**

**FEMA Disaster Declarations Summary**

Federal Emergency Management Agency Department of Homeland Security – FEMA Disaster Declarations Summary is a governmental dataset describing all Federally Declared Disasters. This information begins with the first disaster declaration.

**Northeast Crop Disaster Assistance Program**

Department of Agriculture – USDA's Farm Service Agency's (FSA) Northeast Crop Disaster Assistance Program (NAP) provides financial assistance to producers of nonperennial crops when they...

**Child Nutrition Programs Disaster Response Menu**

Department of Agriculture – This menu option provides an overview of what State agencies, Schools, Food Authorities (FA) participating in the National School Lunch and School Breakfast...

**DATA.GOV**

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

DATA CATALOG

# / Datasets

Filter by location:

243 datasets found for "Earthquakes"

Order by: Relevance

**Earthquakes**

**Earthquake Feeds**

US Geological Survey Department of the Interior – Near real-time earthquake information for a variety of time windows in a variety of formats.

**Earthquake Locations**

State of North Dakota – This layer has been compiled from real-time sources depicting the locations of earthquakes that have been confirmed to have occurred within the state of North Dakota.

**Earthquake Damage - General**

National Climatic and Atmospheric Administration, Department of Commerce – An earthquake is the sudden or breaking of the ground produced by sudden displacement of rock in the Earth's crust. Earthquakes result from crustal stress...

# Do we really need the semantics?

**Yes. All the time.**

**Application knowledge / constraints:**

- Every Senator is an ElectedOfficial which is a Person
- (On Wikipedia) being BornInAPlace means being a Person

# Do we really need the semantics?

**Yes. All the time.**

**Application knowledge / constraints:**

- Every Senator is an ElectedOfficial which is a Person
- (On Wikipedia) being BornInAPlace means being a Person

**Without the semantics, we may miss query answers**

Data	Constraints	Query
John is a <u>Senator</u>	Every <u>Senator</u> is a <u>Person</u>	Who is a <u>Person</u> ?

# Do we really need the semantics?

**Yes. All the time.**

**Application knowledge / constraints:**

- Every Senator is an ElectedOfficial which is a Person
- (On Wikipedia) being BornInAPlace means being a Person

**Semantic constraints are a compact way of encoding information**

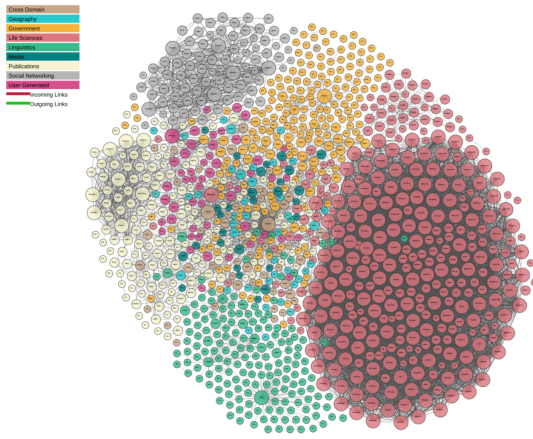
“Every ElectedOfficial is a Person” stated only once even if thousands of ElectedOfficials.



# Semantics for Web data

Data and metadata on the Web is often structured in **graphs**, e.g., **RDF** (W3C's Resource Description Framework)

- Famous application: the Linked Open Data cloud (2017)



# The Resource Description Framework (RDF)

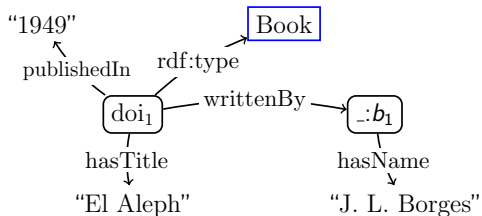
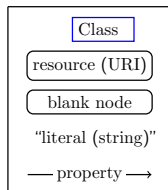
**RDF graph:** set of triples

Assertion	Triple	Relational notation	Intuition
Class	s rdf:type o	$o(s)$	"s is an o"
Property	s p o	$p(s, o)$	"The p of s is o"

# The Resource Description Framework (RDF)

**RDF graph:** set of triples

Assertion	Triple	Relational notation	Intuition
Class	$s \text{ rdf:type } o$	$o(s)$	"s is an o"
Property	$s \text{ p } o$	$p(s, o)$	"The p of s is o"

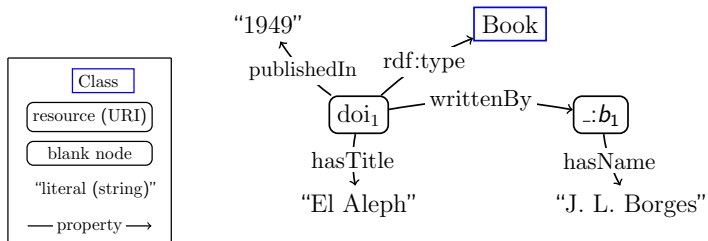


# The Resource Description Framework (RDF)

Assertion	Triple	Relational notation	Intuition
Class	<code>s rdf:type o</code>	$o(s)$	"s is an o"
Property	<code>s p o</code>	$p(s, o)$	"The p of s is o"

# The Resource Description Framework (RDF)

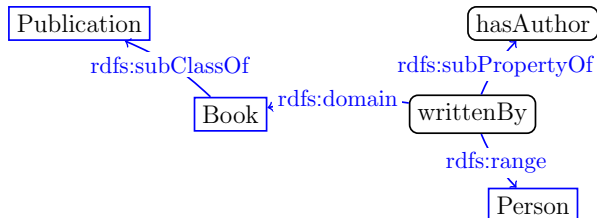
Assertion	Triple	Relational notation	Intuition
Class	$s \text{ rdf:type } o$	$o(s)$	"s is an o"
Property	$s \text{ p } o$	$p(s, o)$	"The p of s is o"



# RDF Schema (RDFS)

Declare **deductive constraints** between classes and properties

Constraint	Triple	OWA interpretation
Subclass	$c_1$ rdfs:subClassOf $c_2$	$c_1 \subseteq c_2$
Subproperty	$p_1$ rdfs:subPropertyOf $p_2$	$p_1 \subseteq p_2$
Domain typing	$p$ rdfs:domain $c$	$\Pi_{\text{domain}}(p) \subseteq c$
Range typing	$p$ rdfs:range $c$	$\Pi_{\text{range}}(p) \subseteq c$

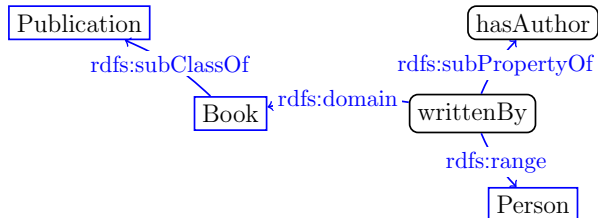


“Any  $c_1$  is also a  $c_2$ ”

# RDF Schema (RDFS)

Declare deductive constraints between classes and properties

Constraint	Triple	OWA interpretation
Subclass	$c_1$ rdfs:subClassOf $c_2$	$c_1 \subseteq c_2$
Subproperty	$p_1$ rdfs:subPropertyOf $p_2$	$p_1 \subseteq p_2$
Domain typing	$p$ rdfs:domain $c$	$\Pi_{\text{domain}}(p) \subseteq c$
Range typing	$p$ rdfs:range $c$	$\Pi_{\text{range}}(p) \subseteq c$

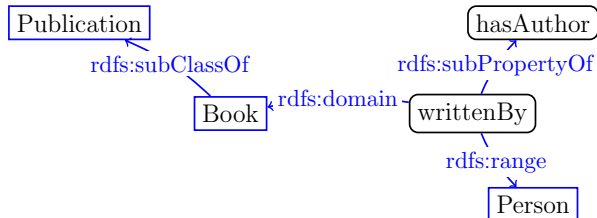


“If two resources are related by  $p_1$ , they are also related by  $p_2$ ”

# RDF Schema (RDFS)

Declare deductive constraints between classes and properties

Constraint	Triple	OWA interpretation
Subclass	$c_1$ rdfs:subClassOf $c_2$	$c_1 \subseteq c_2$
Subproperty	$p_1$ rdfs:subPropertyOf $p_2$	$p_1 \subseteq p_2$
Domain typing	$p$ rdfs:domain $c$	$\Pi_{\text{domain}}(p) \subseteq c$
Range typing	$p$ rdfs:range $c$	$\Pi_{\text{range}}(p) \subseteq c$



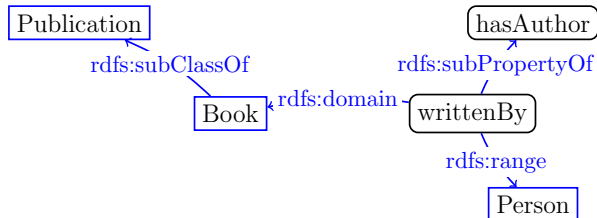
“Anyone having  $p$  is a  $c$ ”



# RDF Schema (RDFS)

Declare deductive constraints between classes and properties

Constraint	Triple	OWA interpretation
Subclass	$c_1$ rdfs:subClassOf $c_2$	$c_1 \subseteq c_2$
Subproperty	$p_1$ rdfs:subPropertyOf $p_2$	$p_1 \subseteq p_2$
Domain typing	$p$ rdfs:domain $c$	$\Pi_{\text{domain}}(p) \subseteq c$
Range typing	$p$ rdfs:range $c$	$\Pi_{\text{range}}(p) \subseteq c$



“Anyone who is a value of  $p$  is a  $c$ ”

# Open-world assumption and RDF entailment

**RDF data model** based on the **open-world assumption**.

Deductive constraints lead to **implicit triples**:  
part of the graph even though not explicitly present

# Open-world assumption and RDF entailment

**RDF data model** based on the **open-world assumption**.

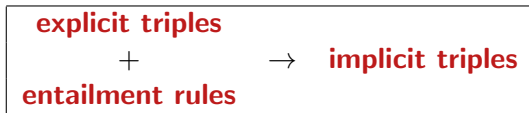
Deductive constraints lead to **implicit triples**:  
part of the graph even though not explicitly present

<b>explicit triples</b>			
	+		
		→	<b>implicit triples</b>
<b>entailment rules</b>			

# Open-world assumption and RDF entailment

**RDF data model** based on the **open-world assumption**.

Deductive constraints lead to **implicit triples**:  
part of the graph even though not explicitly present



Exhaustive application of entailment leads to **saturation (closure)**

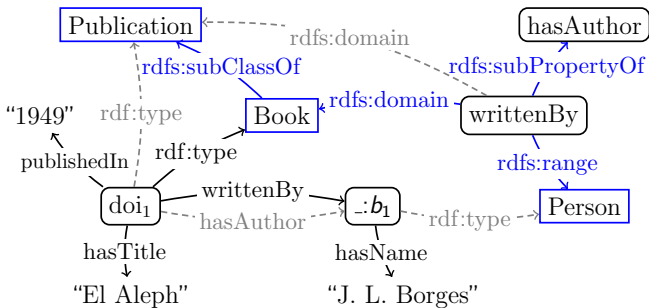
# The semantics of an RDF graph $G$ is its saturation $G^\infty$

## Sample instance entailment rules from schema and instance triples

$$\frac{c_1 \text{ rdfs:subClassOf } c_2 \wedge s \text{ rdf:type } c_1}{s \text{ rdf:type } c_2} \vdash_{\text{RDF}}$$

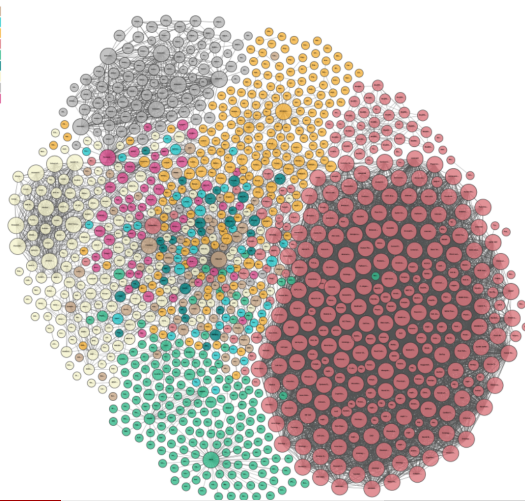
$$\frac{p_1 \text{ rdfs:subPropertyOf } p_2 \wedge s \text{ p}_1 \text{ o}}{s \text{ p}_2 \text{ o}} \vdash_{\text{RDF}}$$

$$\frac{p \text{ rdfs:domain } c \wedge s \text{ p o}}{s \text{ rdf:type } c} \vdash_{\text{RDF}}$$

$$\frac{p \text{ rdfs:range } c \wedge s \text{ p o}}{o \text{ rdf:type } c} \vdash_{\text{RDF}}$$


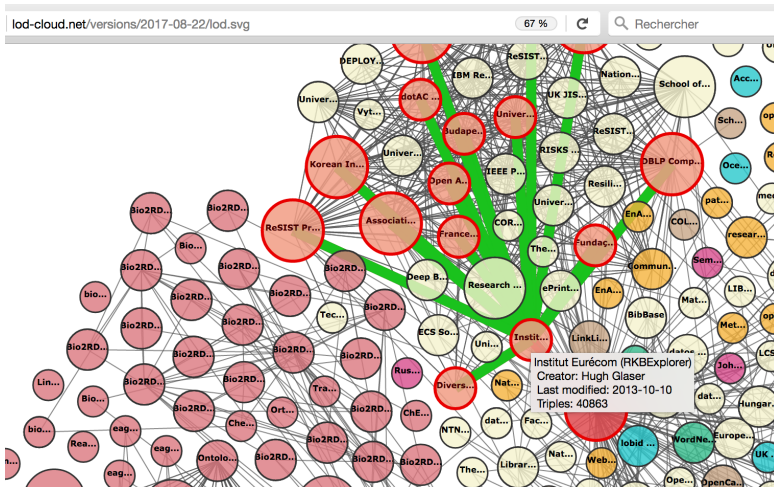
# RDF graph discovery

An RDF graph can be large and complex, lack a fixed schema, include many heterogeneous values...



# RDF graph discovery

An RDF graph can be large and complex, lack a fixed schema, include many heterogeneous values...



# RDF graph discovery

Two approaches:

- ① RDF summarization: compactly representing the explicit and implicit structure of a graph
- ② Insight discovery in RDF graphs: automatically identify aggregation queries with interesting results



## Part II

# RDF summarization

# RDF summaries

## Problem

RDF graph  $G$  is large, heterogeneous, partially implicit.  
How to compactly represent all its structure?

## Existing solutions

**Partial** representation (frequent patterns, statistics etc.)  
e.g., [NM11, LYL13]

**Potentially not compact** e.g., [GW97, CFKP15]  
Only for **explicit data**, e.g., [CDT13, ZDYZ14]

# A summary of DBLP data

150M triples





# RDF summaries

We define

- ① **RDF node equivalence relation**  $\equiv$ : equivalence relation such that class and property nodes are only equivalent to themselves
- ② **RDF summary**  $G_{/\equiv}$  of an RDF graph  $G$ : the **quotient** of  $G$  through  $\equiv$

Recall: quotient of a directed graph  $G$  by  $\equiv$

$G = (V, E)$ ,  $\equiv$  equivalence relation on  $V$

- $G_{/\equiv}$  nodes: one for  $\equiv$  equivalence class of  $V$
- $G_{/\equiv}$  edges:  $n_{/\equiv}^1 \xrightarrow{a} n_{/\equiv}^2$  iff  $\exists n_1 \xrightarrow{a} n_2 \in G$  such that  $n_1$  represented by  $n_{/\equiv}^1$ ,  $n_2$  represented by  $n_{/\equiv}^2$

## Why do we need a special RDF equivalence?

Why not use any node equivalence? E.g., forward and backward bisimilarity  $\sim_{fb}$  [HHK95]

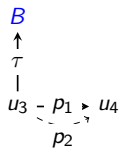
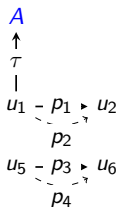
# Why do we need a special RDF equivalence?

Why not use any node equivalence? E.g., forward and backward bisimilarity  $\sim_{fb}$  [HHK95]

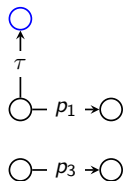
Sample graph  $G$  and its quotient through  $\sim_{fb}$

$p_1 \sim_{sp} p_2$

$p_3 \sim_{sp} p_4$



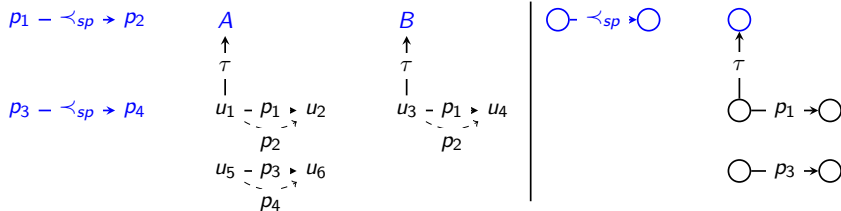
$\bigcirc \sim_{sp} \bigcirc$



# Why do we need a special RDF equivalence?

Why not use any node equivalence? E.g., forward and backward bisimilarity  $\sim_{fb}$  [HHK95]

Sample graph  $G$  and its quotient through  $\sim_{fb}$



Loss of class and (some) property names



# Why do we need a special RDF equivalence?

Why not use any graph node equivalence? E.g., forward and backward bisimilarity  $\sim_{fb}$

Sample graph  $G$  and its quotient through  $\sim_{fb}$

$p_1 - \prec_{sp} \rightarrow p_2$

$A$

$\uparrow$

$\tau$

$\downarrow$

$p_3 - \prec_{sp} \rightarrow p_4$

$u_1 - p_1 \rightarrow u_2$

$p_2$

$u_5 - p_3 \rightarrow u_6$

$p_4$

$B$

$\uparrow$

$\tau$

$\downarrow$

$u_3 - p_1 \rightarrow u_4$

$p_2$

$\bigcirc - \prec_{sp} \rightarrow \bigcirc$

$\bigcirc$

$\uparrow$

$\tau$

$\downarrow$

$\bigcirc - p_1 \rightarrow \bigcirc$

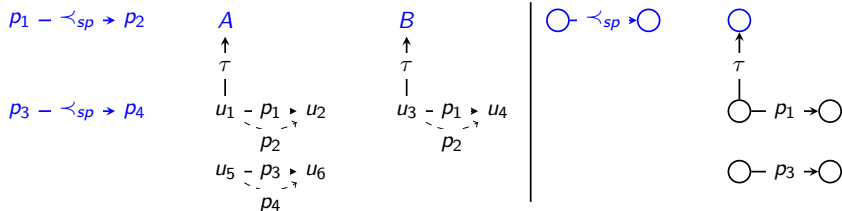
$\bigcirc - p_3 \rightarrow \bigcirc$

Loss of schema triples

# Why do we need a special RDF equivalence?

Why not use any graph node equivalence? E.g., forward and backward bisimilarity  $\sim_{fb}$

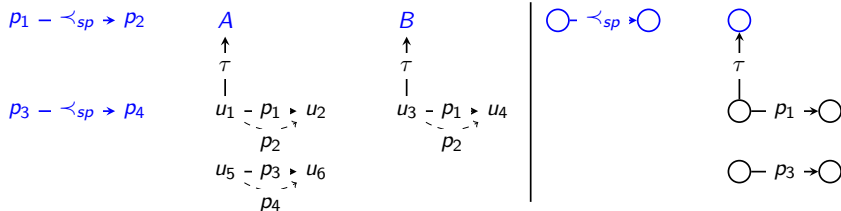
Sample graph  $G$  and its quotient through  $\sim_{fb}$



# Why do we need a special RDF equivalence?

Why not use any graph node equivalence? E.g., forward and backward bisimilarity  $\sim_{fb}$

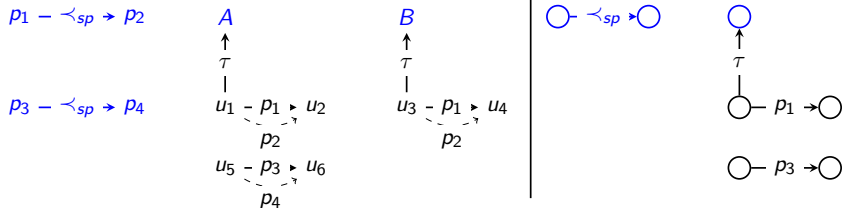
Sample graph  $G$  and its quotient through  $\sim_{fb}$



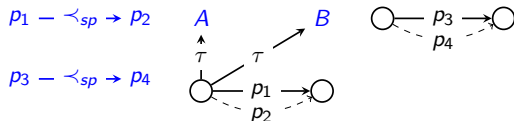
Loss of implicit triples

# Why do we need a special RDF equivalence?

## Sample graph G and its quotient through $\sim_{fb}$



## Quotient of the same graph through the RDF node equivalence $\equiv_{fb}$



## Formal summary properties

For any RDF equivalence relation  $\equiv$ :

Size limit	The summary is at most as large as the graph.
Schema preservation	The schema of $G_{/\equiv}$ is the schema of $G$ .
Representativeness	<p>Any conjunctive query <math>q</math> with answers on <math>G</math> also has answers on its summary:</p> $q(G^\infty) \neq \emptyset \Rightarrow q((G_{/\equiv})^\infty) \neq \emptyset$ <p>This enables <b>query pruning (for query answering) without saturating <math>G</math></b></p>

# Which equivalence relations to use?

## Equivalence notions previously studied

- Forward / backward / forward and backward simulation
- Forward / backward / forward and backward bisimulation

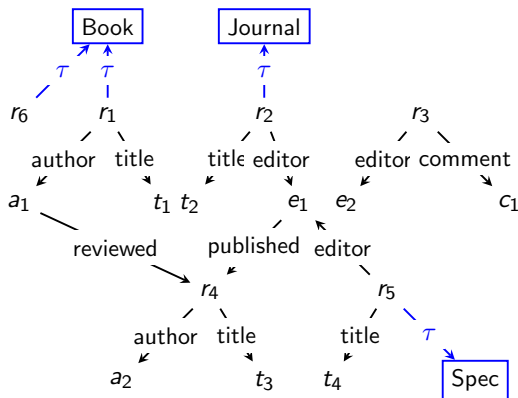
Adapted to semantic RDF graphs

## Novel equivalence notions we introduce (see next)

- Flexible similarity suited to heterogeneous graphs
- Based on **property cliques** and possibly on RDF types

# RDF node equivalence based on property cliques

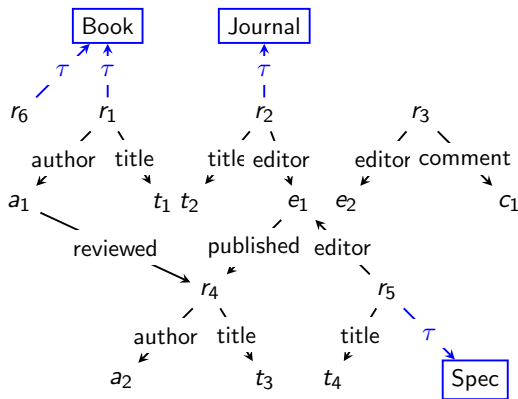
Intuition:  $a_1, a_2$  are similar;  $r_1, r_2, r_3, r_4, r_5$  are similar



# RDF node equivalence based on property cliques

**Output property cliques:**  $\{a, t, e, c\}$ ;  $\{r\}$ ;  $\{p\}$ ;  $\emptyset$

**Input property cliques:**  $\{a\}$ ;  $\{t\}$ ;  $\{e\}$ ;  $\{c\}$ ;  $\{r, p\}$ ;  $\emptyset$

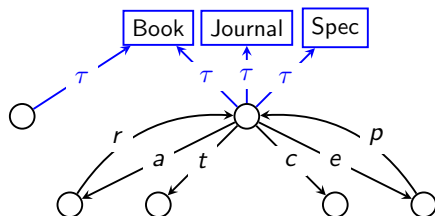




## Weak clique-based summaries

Two nodes are weakly equivalent ( $\equiv_W$ ) iff they have **the same input clique** **or** **the same output clique**.

Weak summary  $G_{/\equiv_W}$  of the sample RDF graph  $G$ :

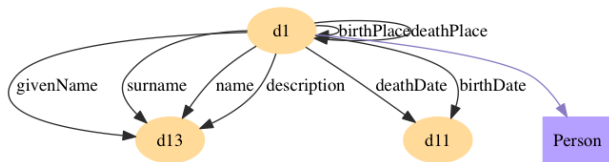


**Property:** In  $G_{/\equiv_W}$ , each data property appears exactly once  $\Rightarrow$  its nodes are “source of  $p$ , target of  $p$ ” for each  $p$  [ČGM15b].

## Weak clique-based summaries

**Property:**  $G_{/\equiv W}$  nodes are “source of  $p$ , target of  $p$ ” for each  $p$ .

**Detecting errors in the data:** why do the birthplace and deathplace loop?



Looking in the data, we find:

---

```

<http://dbpedia.org/resource/Kunitomo_Ikkansai> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .

```

---

```

<http://dbpedia.org/resource/Kunitomo_Ikkansai> <http://dbpedia.org/ontology/birthPlace>
<http://dbpedia.org/resource/Kunitomo_Ikkansai> .

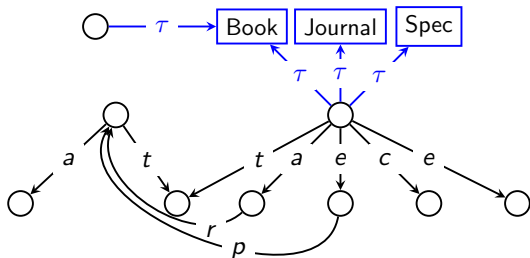
```

---

## Strong clique-based summaries

Two nodes are strongly equivalent ( $\equiv_S$ ) iff they have **the same input clique** **and** **the same output clique**.

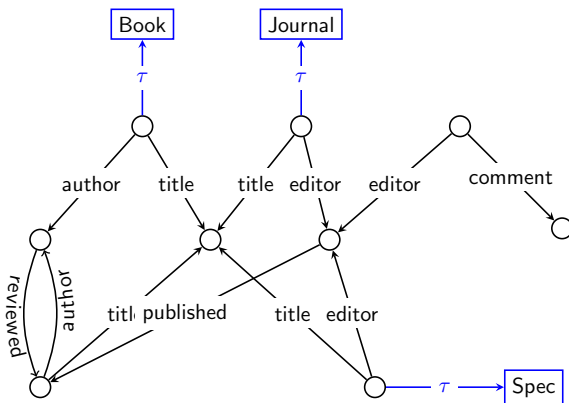
Strong summary  $G_{/\equiv_S}$  of the sample RDF graph  $G$ :



# Using types for summarization

Group nodes **first by their types**; then group untyped nodes by their property cliques.

Typed weak summary  $G_{\equiv_{TW}}$  of the sample RDF graph  $G$ :



On this example, this is also the typed strong summary  $G_{\equiv_{TS}}$ .

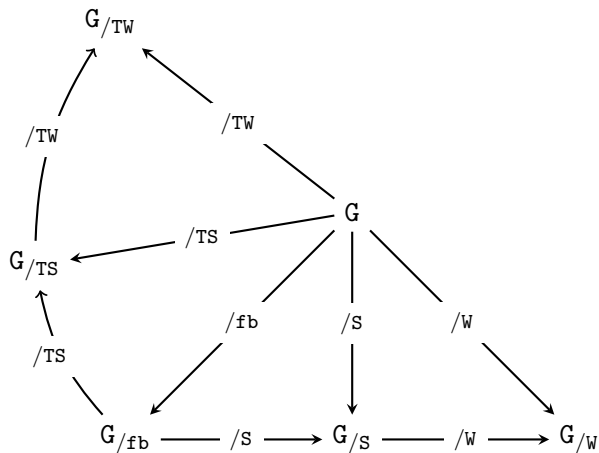
# RDF summaries outline

Summary	Weak?	Strong?	Types first?
$G/\equiv W$	✓		
$G/\equiv S$		✓	
$G/\equiv TW$	✓		✓
$G/\equiv TS$		✓	✓

## RDF summaries outline

Summary	Weak?	Strong?	FW bisim?	BW bisim?	Types first?
$G \equiv W$	✓				
$G \equiv S$		✓			
$G \equiv TW$	✓				✓
$G \equiv TS$		✓			✓
$G \equiv fw$			✓		
$G \equiv bw$				✓	
$G \equiv fb$			✓	✓	
$G \equiv fw, T$			✓		✓
$G \equiv bw, T$				✓	✓
$G \equiv fb, T$			✓	✓	✓

## Relations between RDF summaries [ČGM17b]



## Summary size comparison (more in [ČGM17b])

Graph G	G	Summary $G_{/\equiv}$	$ G_{/\equiv} $	$cf_{\equiv}$
DBLP	150,787,464	$G_{/W}$	<b>71</b>	2,123,767
DBLP	150,787,464	$G_{/S}$	<b>206</b>	731,978
DBLP	150,787,464	$G_{/fw}$	<b>262,695</b>	574
LUBM1M	1,227,868	$G_{/W}$	<b>161</b>	7,579
LUBM1M	1,227,868	$G_{/S}$	<b>207</b>	5,903
LUBM1M	1,227,868	$G_{/fw}$	<b>1982</b>	617
LUBM10M	11,990,183	$G_{/W}$	<b>162</b>	74,013
LUBM10M	11,990,183	$G_{/S}$	<b>206</b>	58,204
LUBM10M	11,990,183	$G_{/fw}$	<b>24,958</b>	480
LUBM10M	11,990,183	$G_{/bw}$	<b>6,162</b>	1,944
LUBM10M	11,990,183	$G_{/fb}$	<b>11,990,076</b>	1



# Summarizing $G^\infty$

Recall: With an RDF Schema, the semantics of  $G$  is  $G^\infty \Rightarrow$   
We really need  $(G^\infty)_{/\equiv}$ !

- 1 Saturate  $G$ , then summarize
- 2 Can we avoid saturating  $G$ ?...

# Summarizing $G^\infty$

Recall: With an RDF Schema, **the semantics of  $G$  is  $G^\infty$**   $\Rightarrow$   
 We really need  $(G^\infty)_{/\equiv}$ !

- ① Saturate  $G$ , then summarize
- ② Can we avoid saturating  $G$ ?...

## Shortcut theorem [ČGM17a]

For the summaries  $G_{/W}$ ,  $G_{/S}$ ,  $G_{/fw}$ ,  $G_{/bw}$ ,  $G_{/fb}$ :

$$(G^\infty)_{/\equiv} \text{ is the same as } ((G_{/\equiv})^\infty)_{/\equiv}$$

Also: **sufficient condition** for any  $\equiv$  to admit the shortcut.

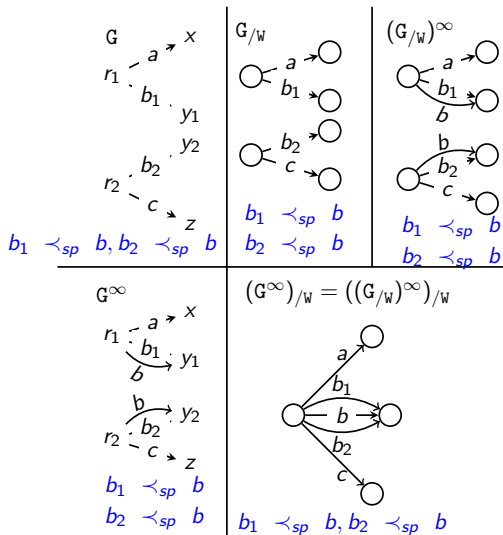
## Shortcut toward the summary of $G^\infty$

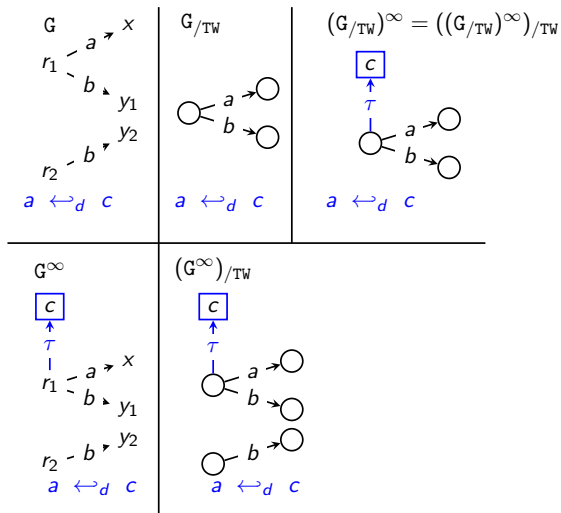
Direct  $G \rightarrow \mathbf{sat.} \rightarrow G^\infty \rightarrow \mathbf{summ.} \rightarrow (G^\infty)_\equiv$

Shortcut  $G \rightarrow \mathbf{summ.} \rightarrow G_\equiv \rightarrow \mathbf{sat.} \rightarrow (G_\equiv)^\infty \rightarrow \mathbf{summ.} \rightarrow ((G_\equiv)^\infty)_\equiv$

If  $G_\equiv$  is much smaller than  $G$ , **the shortcut may be faster!**

Up to 20 times in our experiments [ČGM17b]


Shortcut example:  $G_W$ 

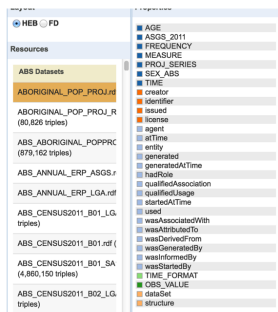
Shortcut counter-example:  $G_{TW}$ 

# Summary-enabled LOD cloud exploration

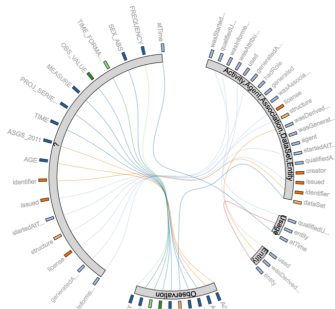
ILDA Inria team (E. Pietriga, H. Ozaygen)

Use summary to derive visualisation instead of the original graph  
(smaller, faster)

abs-linked-data : Australian Bureau of Statistics (ABS) Linked Data 



The screenshot shows a web interface for exploring Linked Data. On the left, there is a sidebar with a search bar containing 'HEB' and 'FD'. Below it, a 'Resources' section lists various datasets from the Australian Bureau of Statistics (ABS), including 'ABORIGINAL\_POP\_PROJ.rdf', 'ABS\_ABORIGINAL\_POPPRC', 'ABS\_ANNUAL\_ERP\_ASGS.r', 'ABS\_ANNUAL\_ERP\_LGA.rdf', 'ABS\_CENSUS2011\_B01\_LG', 'ABS\_CENSUS2011\_B01.rdf', 'ABS\_CENSUS2011\_B01\_SA', and 'ABS\_CENSUS2011\_B02\_LG'. On the right, a legend lists various properties and their corresponding colors: AGE (blue), ASGS\_2011 (dark blue), FREQUENCY (light blue), MEASURE (orange), PROJ\_SERIES (dark blue), SEX\_ABS (dark blue), TIME (light blue), creator (orange), identifier (orange), issued (orange), license (orange), agent (light blue), atTime (light blue), entity (light blue), generated (light blue), generatedAtTime (light blue), hadRole (light blue), qualifiedAssociation (light blue), qualifiedUsage (light blue), startedAtTime (light blue), used (light blue), wasAssociatedWith (light blue), wasAttributedTo (light blue), wasDerivedFrom (light blue), wasGeneratedBy (light blue), wasInformedBy (light blue), wasStartedBy (light blue), TIME\_FORMAT (light blue), OBS\_VALUE (green), dataSet (orange), and structure (orange).



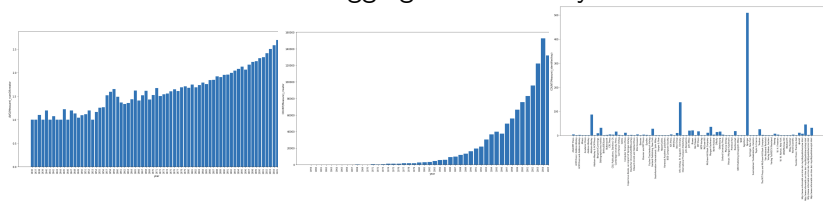
## Part III

# Finding insights in RDF graphs

# Insight in an RDF graph

We consider an insight to be **the result of an aggregation query over the RDF graph**

We focus one-dimensional aggregates  $\Rightarrow$  2D layout



An insight is **interesting** if a certain measure (e.g., variance) on its set of aggregation values is high

**Problem**

**Problem:** given a graph  $G$ , find the top- $k$  insights



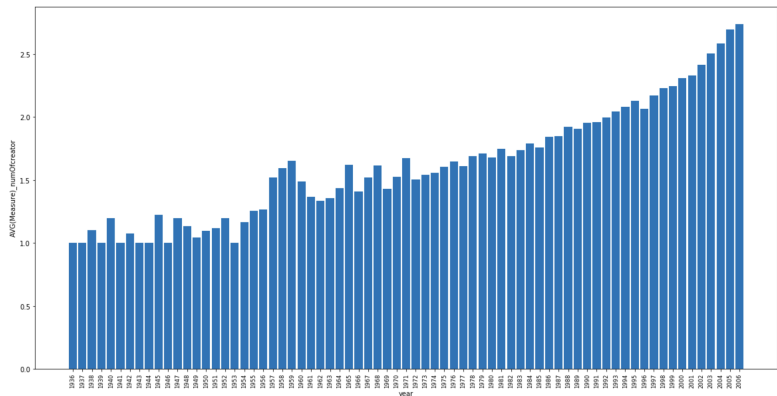
# Dagger approach

**Dagger:** Digging for Interesting Aggregates in RDF Graphs [DMS17] (ongoing)

1. **Candidate facts** Resources from  $G$ : of a certain type, or having certain property sets
2. **Candidate dimension** Properties of the candidate facts, with strong support and relatively few distinct values.  
Also: derived properties, e.g., authors count;
3. **Candidate measure** Another property of the candidate facts  
Also: automatic value typing
4. **Candidate aggregation function** Chosen depending on the measure type

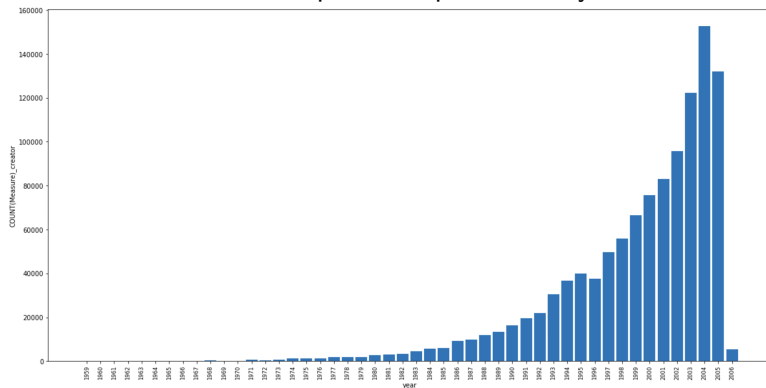
# Dagger-selected aggregate in DBLP data

Average number of authors of journal articles, per publication year



# Dagger-selected aggregate in DBLP data

Number of book authors, per book publication year





# Part IV

## Conclusion

# The need for RDF graph discovery tools

- RDF graphs can be **large and complex**, they lack a prescriptive schema
- Semantic rules lead to **implicit data**
- Toward helping users to discover RDF graphs:
  - ① **Structural quotient summaries** representing the complete graph structure; compact clique-based summaries; available at:  
<https://team.inria.fr/cedar/projects/rdfsummary/>
  - ② **Insight discovery**: interesting aggregate queries; project Web page:  
<https://team.inria.fr/cedar/projects/dagger/>
- Many follow-up directions: parallelization, more interestingness measures, extensions to ML.

# References

- [CDT13] Stéphane Campinas, Renaud Delbru, and Giovanni Tummarello. Efficiency and precision trade-offs in graph summary algorithms. In IDEAS, 2013.
- [CFKP15] Mariano P. Consens, Valeria Fionda, Shahan Khatchadourian, and Giuseppe Pirrò. S+EPPs: Construct and explore bisimulation summaries + optimize navigational queries (demo). PVLDB, 8(12), 2015.
- [ČGM15a] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Query-oriented summarization of RDF graphs. In BICOD, 2015.
- [ČGM15b] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Query-oriented summarization of RDF graphs (demonstration). PVLDB, 8(12), 2015.
- [ČGM17a] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. A framework for efficient representative summarization of RDF graphs. In International Semantic Web Conference (ISWC), 2017.
- [ČGM17b] Šejla Čebirić, François Goasdoué, and Ioana Manolescu. Query-Oriented Summarization of RDF Graphs. Research Report RR-8920, INRIA, 2017.
- [DMS17] Yanlei Diao, Ioana Manolescu, and Shu Shang. Dagger: Digging for interesting aggregates in RDF graphs. In International Semantic Web Conference (ISWC), 2017.

## References (cont.)

- [GW97] Roy Goldman and Jennifer Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In VLDB, 1997.
- [HHK95] Monika Rauch Henzinger, Thomas A. Henzinger, and Peter W. Kopke. Computing simulations on finite and infinite graphs. In FOCS, 1995.
- [LYL13] Shou-De Lin, Mi-Yen Yeh, and Cheng-Te Li. Sampling and summarization for social networks (tutorial), 2013.
- [NM11] Thomas Neumann and Guido Moerkotte. Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins. In ICDE, 2011.
- [ZDYZ14] Haiwei Zhang, Yuanyuan Duan, Xiaojie Yuan, and Ying Zhang. ASSG: adaptive structural summary for RDF graph data. In ISWC (Posters and Demonstrations), 2014.