Coloring RDF Triples to Capture Provenance

G. Flouris[1], I. Fundulaki[1],
P. Pediaditis[1,2], Y. Theoharis[1,2],
V. Christophides[1,2]

[1:] *Institute of Computer Science, FORTH, Greece*
[2:] *Computer Science Department, University of Crete, Greece*

Presented by V. Papavasileiou, Institute of Computer Science, FORTH



Introduction

- Linked Data on the Web:
  - "The goal of […] is to enable people to *share structured data* on the Web as easily as they can share documents today."
  - "usefulness of data increases the more it is interlinked with other data"
  - Basic tenets
    - use the *RDF data model* to publish structured data on the Web
    - use typed *RDF links* to interlink the data sources
    - use *RDFS reasoning rules* to *infer implicit knowledge*

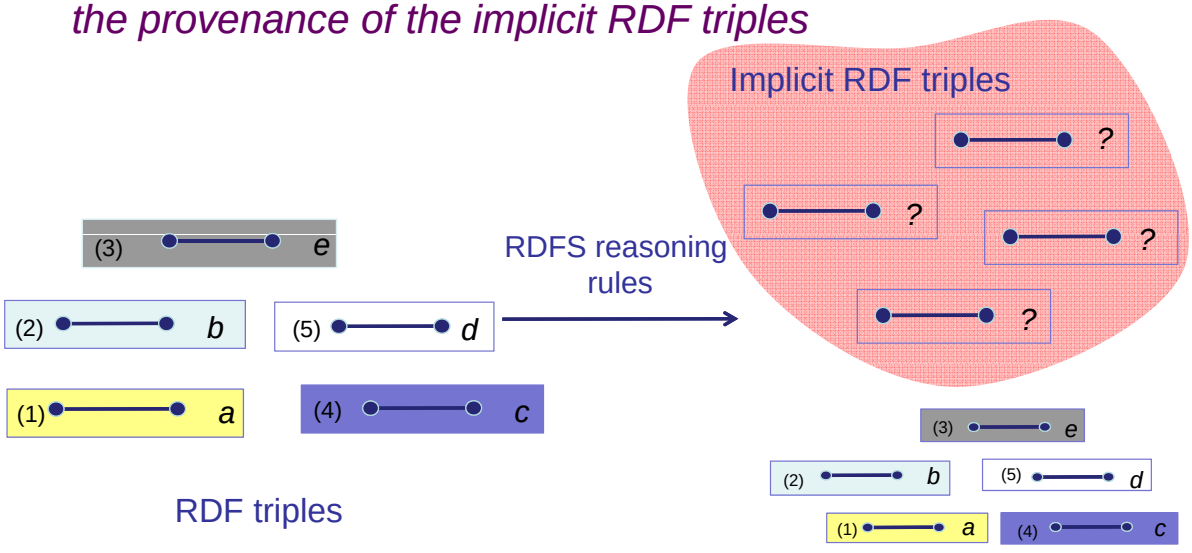


Provenance for RDF Data

- *Wikipedia*
  - " [provenance] means the *origin*, or the *source*, of something, or the history of the ownership or location of an object"
- *W3C Provenance Incubator Group Charter*
  - "With the arrival of massive amounts of Semantic Web Data […], *provenance* becomes an *important factor* in developing new Semantic Web applications."
  - "*reasoners* in the Semantic Web will need *explicit representations of provenance information* … "

Provenance for RDF Data: Problem Statement

Given a set of RDF triples whose *explicit provenance* is known, and the RDFS reasoning rules, we want to decide *the provenance of the implicit RDF triples*



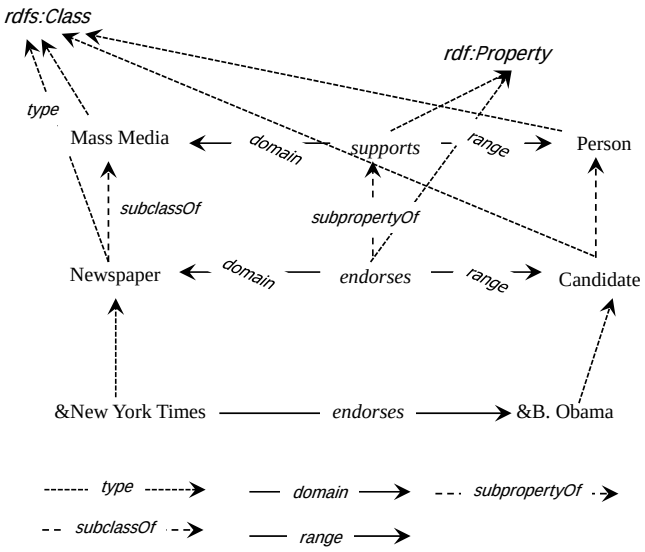
Provenance for RDF Data: Our approach

- *colors* to capture provenance of *explicit* and *implicit* RDF triples (*data and schema*)
- *quadruples* to represent provenance information
- *commutative semi-group structure* to reason over the provenance of *implicit triples*
  - *set of colors*
  - *binary operation "+"*
- *querying* and *updating* RDF *quadruples*

Resource Description Framework (RDF)

An RDF triple is of the form *(subject, predicate, object)* and represented as: *s* — *p* → *o*

subject	*predicate*	*object*
&New York Times	*endorses*	&B. Obama
Newspaper	*type*	*Class*
&New York Times	*type*	Newspaper
Mass Media	*type*	*Class*
Newspaper	*subclassOf*	*Mass Media*
Candidate	*type*	*Class*
&B. Obama	*type*	Candidate
endorses	*type*	*Property*
endorses	*range*	Candidate
endorses	*domain*	Newspaper
Person	*type*	*Class*
Candidate	*subclassOf*	*Person*
supports	*type*	*Property*
supports	*range*	Person
supports	*domain*	Mass Media
endorses	*subpropertyOf*	*supports*

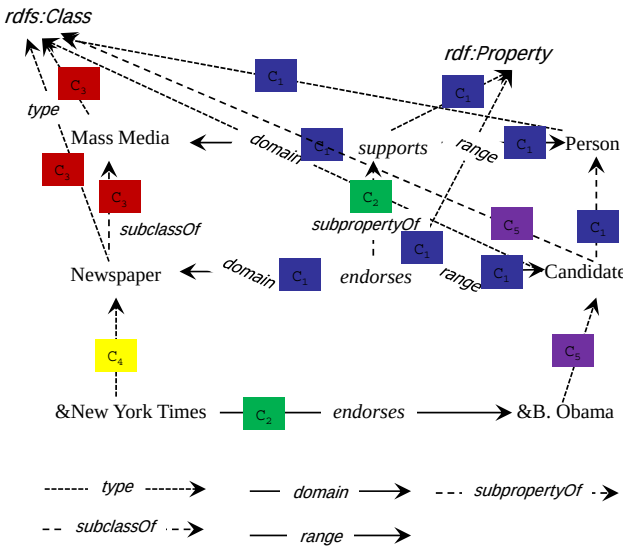


RDF and RDFS built-in properties

From Triples to Quadruples: Coloring RDF triples to capture provenance

- We extend a triple with a fourth column for colors *(subject, predicate, object, color)*

subject	*predicate*	*object*	*color*
&New York Times	*endorses*	&B. Obama	c_2
Newspaper	*type*	*Class*	c_3
&New York Times	*type*	Newspaper	c_4
Mass Media	*type*	*Class*	c_3
Newspaper	*subclassOf*	*Mass Media*	c_3
Candidate	*type*	*Class*	c_5
&B. Obama	*type*	Candidate	c_5
endorses	*type*	*Property*	c_1
endorses	*range*	Candidate	c_1
endorses	*domain*	Newspaper	c_1
Person	*type*	*Class*	c_1
Candidate	*subclassOf*	*Person*	c_1
supports	*type*	*Property*	c_1
supports	*range*	Person	c_1
supports	*domain*	Mass Media	c_1
endorses	*subpropertyOf*	*supports*	c_2



RDF and RDFS built-in properties

Talk Outline

- ✓ Introduction
- ✓ Provenance for RDF Data
  - ✓ Problem statement
  - ✓ Our approach
- ✓ RDF Data Model and RDF Schema Language
- Computing the provenance of implicit RDF triples
- Querying Colored RDF triples
- Updating Colored RDF triples
- Related Work
- Conclusions & Future Work

Computing the provenance of implicit RDF triples

- RDFS inference rules for triples
  - *subclassOf* and *subpropertyOf* are *reflexive* and *transitive*

(A_1, *subclassOf*, A_2) (A_2, *subclassOf*, A_3) ➔ (A_1, *subclassOf*, A_3)

  - "*transitivity*" of *type* relation (class and property instantiations)

(&r, *type*, A_1) (A_1, *subclassOf*, A_2) ➔ (&r, *type*, A_2)

Computing the provenance of implicit RDF triples

- RDFS inference rules for *quadruples*
  - *subclassOf* and *subpropertyOf* are *reflexive* and *transitive*

(A_1, *subclassOf*, A_2, C_1) (A_2, *subclassOf*, A_3, C_2) ➔ (A_1, *subclassOf*, A_3, *?*)

  - "*transitivity*" of *type* relation (class and property instantiations)

(&r, *type*, A_1, C_1) (A_1, *subclassOf*, A_2, C_2) ➔ (&r, *type*, A_2, *?*)

Computing the provenance of implicit RDF triples

- Commutative semi-group (*C*, "+") as the provenance model
  - set *C* of colors , $C \subseteq$ URIs, distinguished from the set of class and property names
  - binary operation "+" to compose colors
- Provenance model captures *how provenance:*
  - *which explicit triples contribute to an implicit triple?*

Revisiting RDFS Inference Rules

(A_1, *subclassOf*, A_2, c_1) (A_2, *subclassOf*, A_3, c_2) ➔ (A_1, *subclassOf*, A_3, c_1 + c_2)

(&r, *type*, A_1, c_1) (A_1, *subclassOf*, A_2, c_2) ➔ (&r, *type*, A_2, c_1 + c_2)

Computing the provenance of implicit RDF triples

- Commutative semi-group (*C*, "+") as the provenance model
  - set *C* of colors , $C \subseteq$ URIs, distinguished from the set of class and property names
  - binary operation "+" to compose colors
  - Properties
    1. $c_1 + c_2 = c_2 + c_1$ (*commutativity*)
    2. $c_1 + (c_2 + c_3) = (c_1 + c_2) + c_3$ (*associativity*)
    3. $c + c = c$ (*idempotence*)
  - How provenance:
    - not interested how many times an explicit triple is considered: *idempotence*
    - order is unimportant: *commutative* and *associative*

Computing the provenance of implicit RDF triples

- From how provenance alone, we can maintain a *materialized view*

Explicit Triples: (A_1, *subclassOf*, A_2, c_1) (A_2, *subclassOf*, A_3, c_2) (A_3, *subclassOf*, A_4, c_5)

Materialized Implicit Triples: (A_1, *subclassOf*, A_4, $c_5 + c_2 + c_1$) (A_1, *subclassOf*, A_3, $c_2 + c_1$)

Operation: *delete* (A_3, *subclassOf*, A_4, c_5)

Two ways of maintaining the materialized view

1. *re-compute* the implicit triples
2. *incremental maintenance*: remove the implicit triples to which the explicit triple has contributed

we can maintain the view looking at the provenance of the implicit and explicit triple

(A_1, *subclassOf*, A_4, $c_5 + c_2 + c_1$)

Querying Colored RDF Triples

- Queries on the *subclassOf*, *subpropertyOf* and *type* RDF/S properties

subject	*predicate*	*object*	*color*
&New York Times	*endorses*	&B. Obama	c_2
Newspaper	*type*	*Class*	c_3
&New York Times	*type*	Newspaper	c_4
Mass Media	*type*	*Class*	c_3
Newspaper	*subclassOf*	*Mass Media*	c_3
Candidate	*type*	*Class*	c_5
&B. Obama	*type*	Candidate	c_5
endorses	*type*	*Property*	c_1
endorses	*range*	Candidate	c_1
endorses	*domain*	Newspaper	c_1
Person	*type*	*Class*	c_1
Candidate	*subclassOf*	*Person*	c_1
supports	*type*	*Property*	c_1
supports	*range*	Person	c_1
supports	*domain*	Mass Media	c_1
endorses	*subpropertyOf*	*supports*	c_2

Q1: *return quadruples defining the subclasses of Mass Media*

(?x, *subclassOf*, Mass Media, ?c)

(Newspaper, *subclassOf*, Mass Media, c_3)

Querying Colored RDF Triples

- Queries on the *subclassOf*, *subpropertyOf* and *type* RDF/S properties

subject	*predicate*	*object*	*color*
&New York Times	*endorses*	&B. Obama	c_2
Newspaper	*type*	*Class*	c_3
&New York Times	*type*	Newspaper	c_4
Mass Media	*type*	*Class*	c_3
Newspaper	*subclassOf*	*Mass Media*	c_3
Candidate	*type*	*Class*	c_5
&B. Obama	*type*	Candidate	c_5
endorses	*type*	*Property*	c_1
endorses	*range*	Candidate	c_1
endorses	*domain*	Newspaper	c_1
Person	*type*	*Class*	c_1
Candidate	*subclassOf*	*Person*	c_1
supports	*type*	*Property*	c_1
supports	*range*	Person	c_1
supports	*domain*	Mass Media	c_1
endorses	*subpropertyOf*	*supports*	c_2

Q2: *return quadruples defining the instances of Mass Media*

(?x, *type*, Mass Media, ?c)

(&New York Times, *type*, Mass Media, c_4 + c_3)

Querying Colored RDF Triples

- Queries on the *subclassOf*, *subpropertyOf* and *type* RDF/S properties

subject	*predicate*	*object*	*color*
&New York Times	*endorses*	&B. Obama	c_2
Newspaper	*type*	*Class*	c_3
&New York Times	*type*	Newspaper	c_4
Mass Media	*type*	*Class*	c_3
Newspaper	*subclassOf*	*Mass Media*	c_3
Candidate	*type*	*Class*	c_5
&B. Obama	*type*	Candidate	c_5
endorses	*type*	*Property*	c_1
endorses	*range*	Candidate	c_1
endorses	*domain*	Newspaper	c_1
Person	*type*	*Class*	c_1
Candidate	*subclassOf*	*Person*	c_1
supports	*type*	*Property*	c_1
supports	*range*	Person	c_1
supports	*domain*	Mass Media	c_1
endorses	*subpropertyOf*	*supports*	c_2

Q3: *return quadruples defining the instances of any class colored* c_5

(?x, *type*, ?y, c_5)

(Candidate, *type*, *Class*, c_5)

(&B. Obama, *type*, Candidate, c_5)

Querying Colored RDF Triples

- Queries on the *subclassOf*, *subpropertyOf* and *type* RDF/S properties

subject	*predicate*	*object*	*color*
&New York Times	*endorses*	&B. Obama	c_2
Newspaper	*type*	*Class*	c_3
&New York Times	*type*	Newspaper	c_4
Mass Media	*type*	*Class*	c_3
Newspaper	*subclassOf*	*Mass Media*	c_3
Candidate	*type*	*Class*	c_5
&B. Obama	*type*	Candidate	c_5
endorses	*type*	*Property*	c_1
endorses	*range*	Candidate	c_1
endorses	*domain*	Newspaper	c_1
Person	*type*	*Class*	c_1
Candidate	*subclassOf*	*Person*	c_1
supports	*type*	*Property*	c_1
supports	*range*	Person	c_1
supports	*domain*	Mass Media	c_1
endorses	*subpropertyOf*	*supports*	c_2

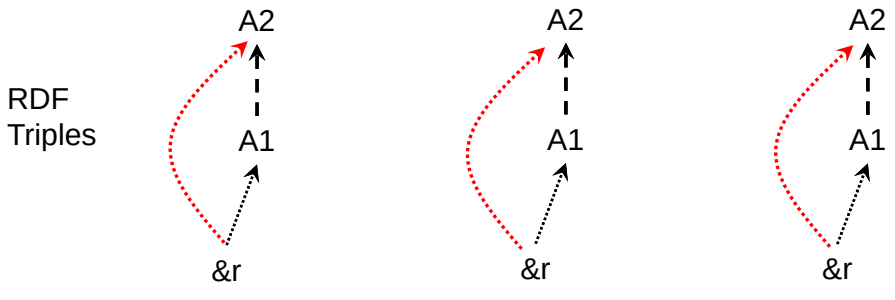
Q4: *return quadruples defining the instances of any class colored* $c_1 + c_5$

(?x, *type*, ?y, $c_1 + c_5$)

(&B. Obama, *type*, Person, $c_1 + c_5$)

Updating Colored RDF Triples

- Update Semantics [P. Gardenfors. The dynamics of belief systems: Foundations versus coherence theories. *Revue Internationale de Philosophie*, 44:24–46, 1992]
  - Foundational Semantics
    - Implicit information depends on the existence of explicit information and is *lost* when the explicit information upon which it depends is lost
  - Coherence Semantics
    - Implicit information is *first class citizen*, and is *retained* when the explicit information upon which it depends is lost

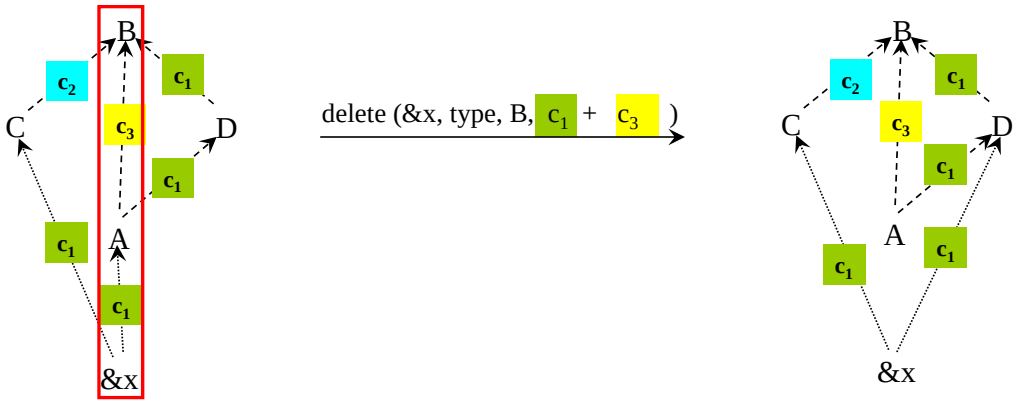


Updating Colored RDF Triples

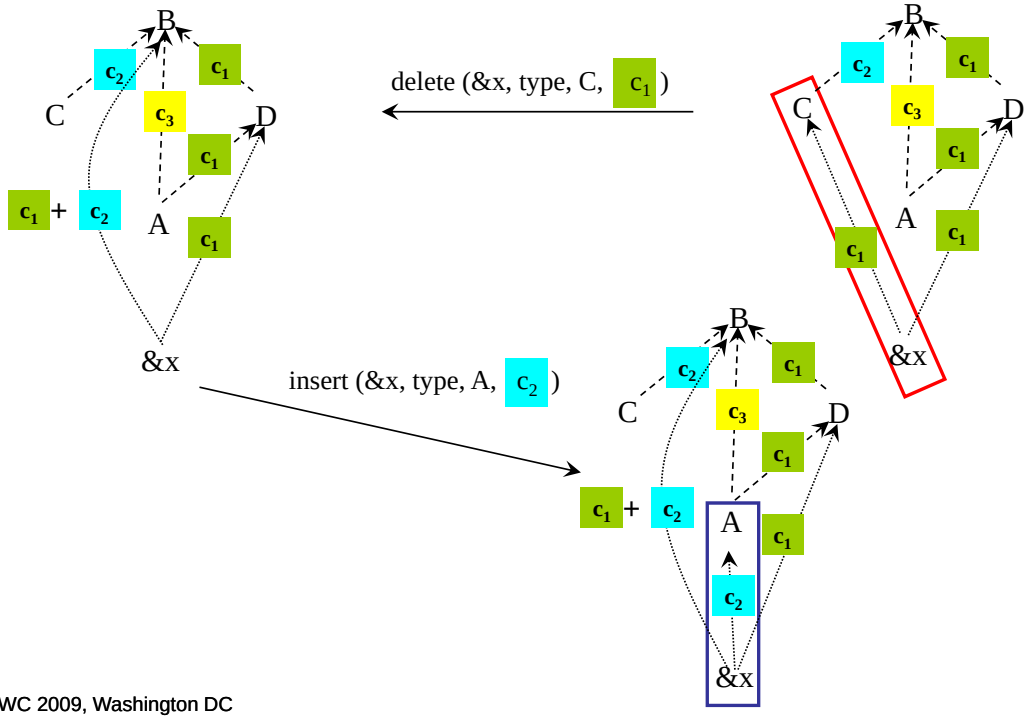
- We adopt *Coherence Semantics*
  - Explicit and Implicit (acquired through inference) knowledge *is a first class citizen* (foundational: explicit more important)
    - Explicit knowledge is just a representation of the knowledge, not *the* knowledge
  - *Updates on Explicit* and *Implicit knowledge* (foundational: explicit triples only)
  - *Better Information Preservation (Principle of Minimal Change)*:
    - implicit knowledge becomes explicit when in danger to be lost (unlike foundational)

Updating Colored RDF triples

- Language of atomic insert and delete operations!!



Updating Colored RDF triples



Related Work

- *In the RDF context*
  - Named Graphs [Carroll et al. 2005, Watkins et al. 2006],
  - Quadruples [Dumbill et al. 2003, MacGregor et al. 2003]
  - Both express provenance of *explicit* triples
    - We focus on the provenance of *implicit* triples (RDFS inference)

- *Relational & XML DBs*
  - [Green et al. 2007] considers *recursion* but not *updates*.
  - [Buneman et al. 2006] considers *updates* but not *recursion*.
  - [Geerts et al. 2006] considers neither *updates* nor *recursion*.
    - We consider both *recursion* and *updates* .



Conclusion & Future Work

- *Conclusions*
  - Employ colors to manage *Explicit Provenance Information* in the presence of RDFS *Inference*
  - Manage *(query, update)* provenance information

- *Future Work*
  - Study expressive provenance models for SPARQL expressions
  - Explore the ability to express the same provenance semantics *implicitly* (i.e. without storing the colors)
  - Study provenance ontology updates



Thank you!