A Graph Model for RDF

Based on a Diploma Thesis by J.Hayes, Universidad de Chile, 2004

Shima Dastgheib
Mehdi Allahyari

Advanced Database Management Systems
Spring 2012

Introduction

- A graph is a generalization of the simple concept of a collection of *nodes*, connected pair-wise by *edges*.
- very common to represent structures of any sort as graphs, because many practical questions can be reduced to graph problems.
- first contributions to graph theory is Leonhard Euler's discussion of the Seven Bridges of K¨onigsberg.

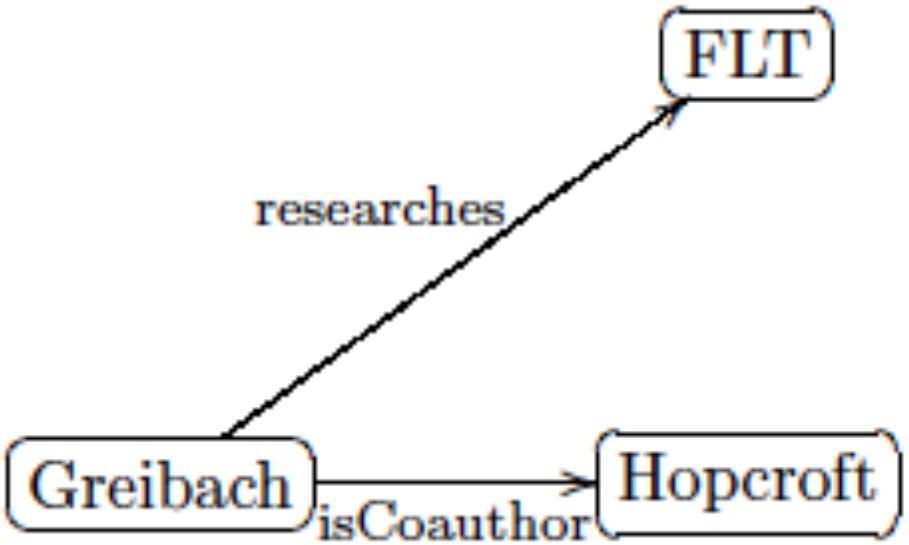
Web and RDF

- Web was built principally for human consumption, but due to its enormous size
- to make use of software agents for organizing, searching, and processing its content.
- Although the data displayed on the Web is *machine-readable*, it is not *machine understandable*, fundamental requirement for meaningful processing of it.

RDF

- A commonly accepted solution:
  - enrichment of human-targeted Web resources (Web pages, etc.) with machine-intelligible information, also referred to as *metadata annotation*.
- The RDF provides a simple triple syntax to express such annotations:
- a resource (the subject) is described by a property (the predicate) and its property value (the object).

RDF



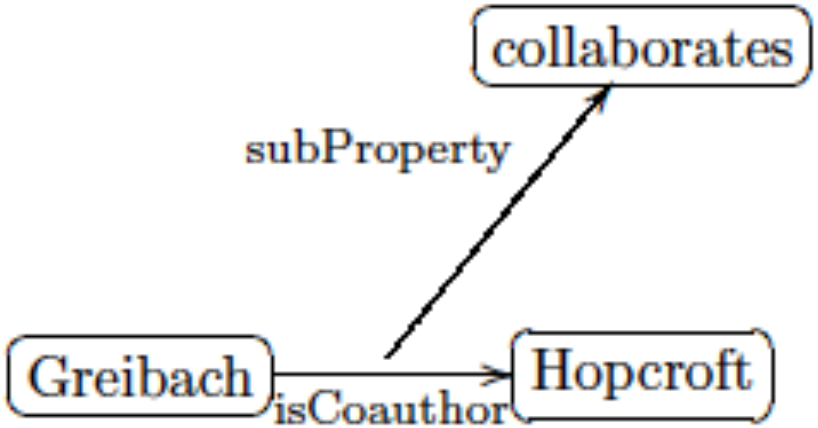
- directed labeled graphs can be employed to represent RDF.

RDF Graph Model (pros and cons)

- various purposes for this:
  - data can be conveniently visualized
  - Results for problems stated for graphs in general apply equally to RDF graphs.
    - Whether an RDF graph contains a certain type of pattern.
  - Programming libraries providing graph data structures and algorithms are available to facilitate the implementation of applications using RDF.

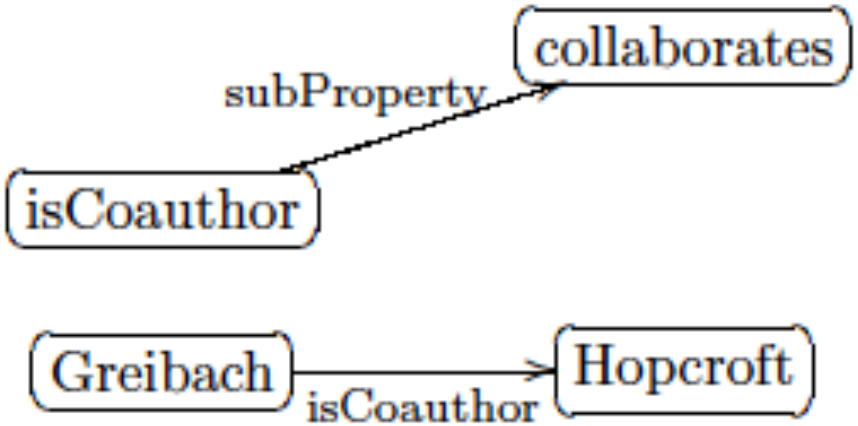
RDF Graph Model (pros and cons)

- graph representation has certain limitations:
  - RDF permits properties to be described just like other resources.
  - Example: <isCoauthor subProperty collaborates>



RDF Graph Model (pros and cons)

- somewhat strange: one of the edges connects an edge label with a node.
- The definition of graphs, however, implies that nodes and edges are distinct sets.
- Another way:



RDF Graph Model (pros and cons)

- avoids the non-standard edges of the previous example. Edges connect only nodes, but the labels of edges and nodes intersect.
- The disadvantage :
  - obtained graph does not truly represent the connectivity of the RDF data.
    - property isCoauthor is related to collaborates
- Solution: Bipartite graph for representing RDF

RDF Concept

- Formal Definition:
  - “uris” be the set of URIs, “blanks” the set of blank node identifiers, and “lits” the set of possible literal values of whatever datatype.

$$(s, p, o) \in (\text{uris} \cup \text{blanks}) \times (\text{uris}) \times (\text{uris} \cup \text{blanks} \cup \text{lits})$$

RDF Graph

- RDF Graph T is a set of RDF statements:
  - univ(**T**): set of all values occurring in all triples of **T**.
  - vocab(**T**): set of all values of the universe that are not blank nodes
- V be a set of URIs and literal values:

$$\mathrm{RDFG(V)} := \{\, T \;\; : T \text{ is RDF Graph and } \mathrm{vocab}(T) \subseteq V \,\}$$

- set of all RDF Graphs with a vocabulary included in **V**

RDF Data

- RDF statements are triples consisting of subject , predicate and object .
- URI references may occur as any part of a triple.
- Any collection of RDF data is an RDF Graph.
  - convincing for intuitive understanding
  - not compatible with the definition of a graph in a mathematical sense

Definition of Graph:

- A graph is a pair G = (N,E), where N is a set whose elements are called nodes, and E is a set of unordered pairs {u, v}, u, v ∈ N

Definition 2 (Bipartite Graph): A graph $G = (N, E)$ is said to be *bipartite* if $N = U \cup V, U \cap V = \emptyset$ and for all $\{u, v\} \in E$ it holds that $u \in U$ and $v \in V$. A bipartite graph is *regular* if for every $v_1, v_2 \in V$ $\text{degree}(v_1) = \text{degree}(v_2)$.

Formal Definition of the Representation of an RDF Graph

Definition 11 (Directed Labeled Graphs (DLG)): Let V be a vocabulary, T be an RDF Graph with $\text{vocab}(T) \subseteq V$ and $\mathcal{G}_{\text{dir, label, multi}}$ the set of directed, edge- and node-labeled multigraphs. We then define a map

$$\delta : \text{RDFG(V)} \rightarrow \mathcal{G}_{\text{dir, label, multi}}$$

as follows: $\delta(T) = (N, E, l_N, l_E)$, where

$$N = \{n_x : x \in \text{subj}(T) \cup \text{obj}(T)\} \text{ with}$$

$$l_N(n_x) = \begin{cases} (x, d_x) & \text{if } x \text{ is literal } (d_x \text{ is datatype identifier}) \\ x & \text{else} \end{cases}$$

$$E = \{\, e_{s,p,o} : (s, p, o) \in T \,\} \text{ with}$$

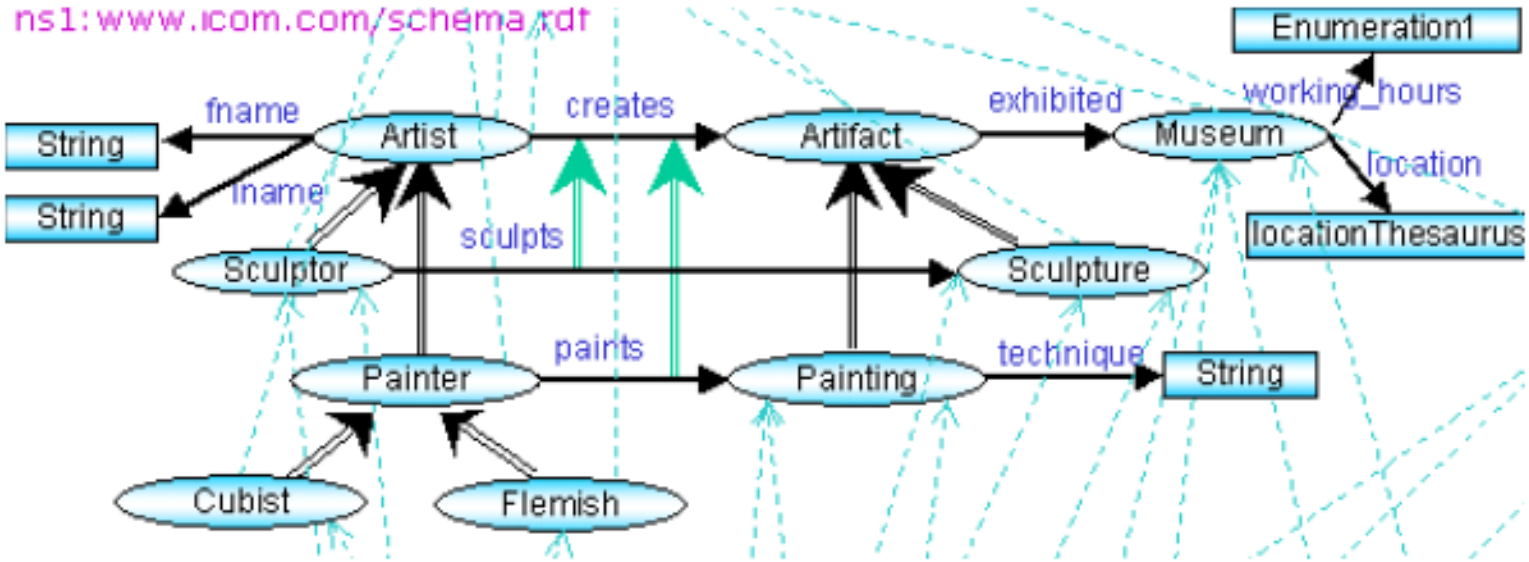
$$\text{from}(e_{s,p,o}) = n_s, \quad \text{to}(e_{s,p,o}) = n_o, \quad \text{and} \quad l_E(e_{s,p,o}) = p$$

□

Shortcomings of Directed Labeled Graphs

- in a given set of RDF data a URI reference may occur at the same time as the predicate of one statement and as the subject or object of others
- every reification of a statement lets the statement's property appear as the object(subject) of another statement.

Solution 1)


ns1:www.icom.com/schema.rdf
Enumeration1
working_hours
fname
creates
exhibited
String
Artist
Artifact
Museum
location
String
lname
locationThesaurus
sculpts
Sculptor
Sculpture
paints
technique
Painter
Painting
String
Cubist
Flemish

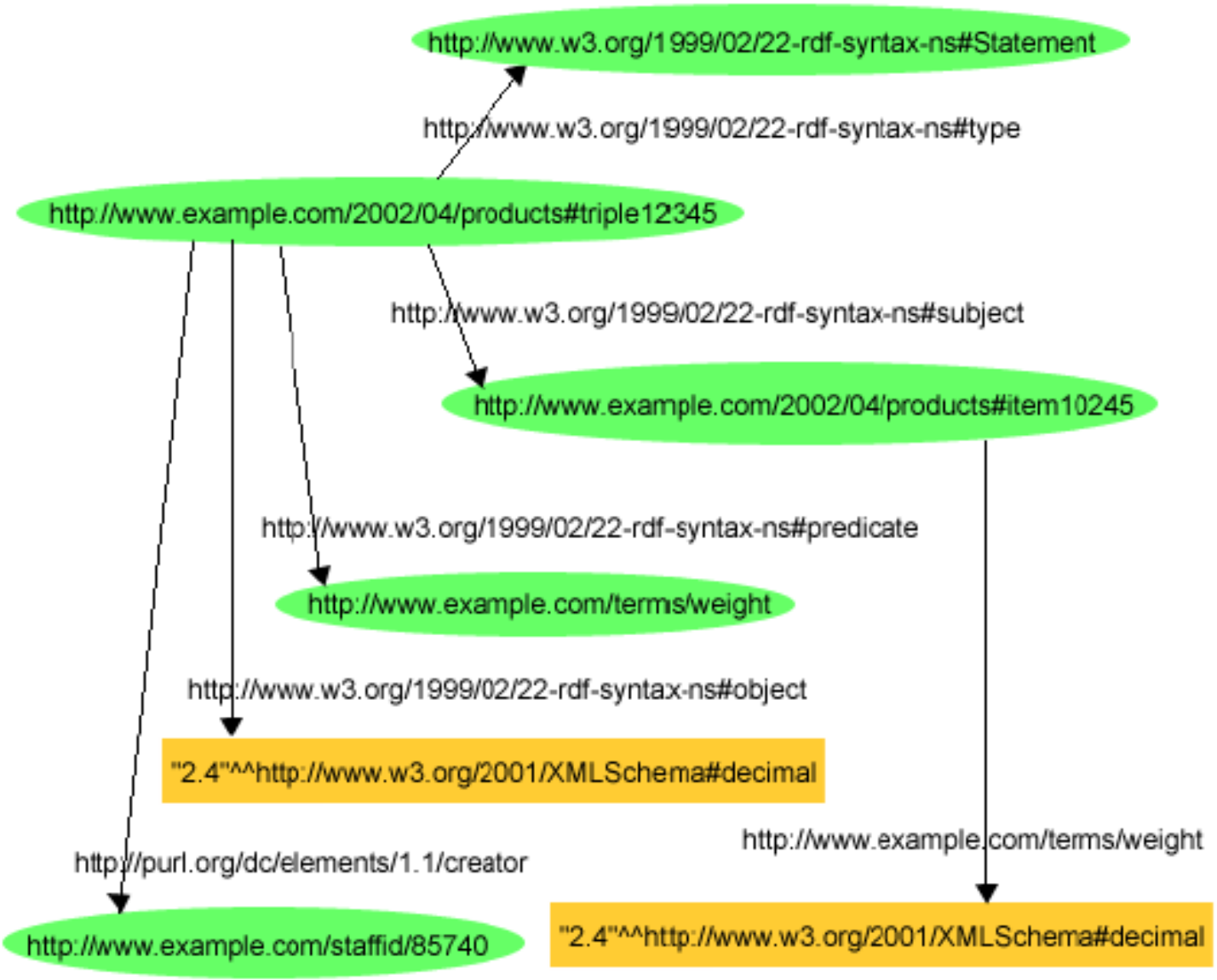
Issues

- Puzzling drawings
- Sets of arcs and nodes which intersect
  - does not correspond to the commonly accepted definition of graphs.
  - Reduces the task from graph representation to visualization for humans and gives

Solution 2)

The information resource *p* occurs multiple times in the graph:

- once for each usage as a predicate (as edge label)
- once for all uses as a subject or object (as node).


http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.example.com/2002/04/products#triple12345
http://www.w3.org/1999/02/22-rdf-syntax-ns#subject
http://www.example.com/2002/04/products#item10245
http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate
http://www.example.com/terms/weight
http://www.w3.org/1999/02/22-rdf-syntax-ns#object
"2.4"^^http://www.w3.org/2001/XMLSchema#decimal
http://purl.org/dc/elements/1.1/creator
http://www.example.com/terms/weight
http://www.example.com/staffid/85740
"2.4"^^http://www.w3.org/2001/XMLSchema#decimal

Issues

- Duplicating properties in the graph representation of an RDF Graph makes it unsuitable for the study of *connectivity*.
- Information about a property (its sub- and super-properties, its domain and range) are disconnected from the actual usage of the property. This might result in users drawing misleading conclusions;

From Binary to Ternary

- RDF triples establish ternary relations which cannot be truly represented by the binary edges of classic graphs.
- Labeling the edges neglects the fact that properties are information resources in their own right.
- a proposed approach based on ternary edges (hypergraphs)
  - Beyond the scope of this presentation

Reference

- J. Hayes, A Graph Model for RDF, Diploma Thesis, Technische Universitt Darmstadt/ Universidad de Chile, 2004.