Foundations of RDF Databases (Overview)

Claudio Gutierrez

Department of Computer Science
Universidad de Chile

UPM - Madrid - Enero 2009

Joint Work With

- Renzo Angles
- Marcelo Arenas
- Carlos Hurtado
- Sergio Muñoz
- Jorge Pérez

Inspired by…

To the memory of

Alberto Mendelzon,

database theoretician

and Web enthusiast

Agenda

1. RDF and Databases
2. RDF and Database models
3. RDF Query Language
   – Requirements and Domains
   – Manifold Views
4. SPARQL
   – Syntax and Semantics
   – Complexity
   – Expressive Power

Agenda

1. RDF and Databases
2. RDF and Database models
3. RDF Query Language
   – Requirements and Domains
   – Manifold Views
4. SPARQL

Disclaimer

A particular view on the subject

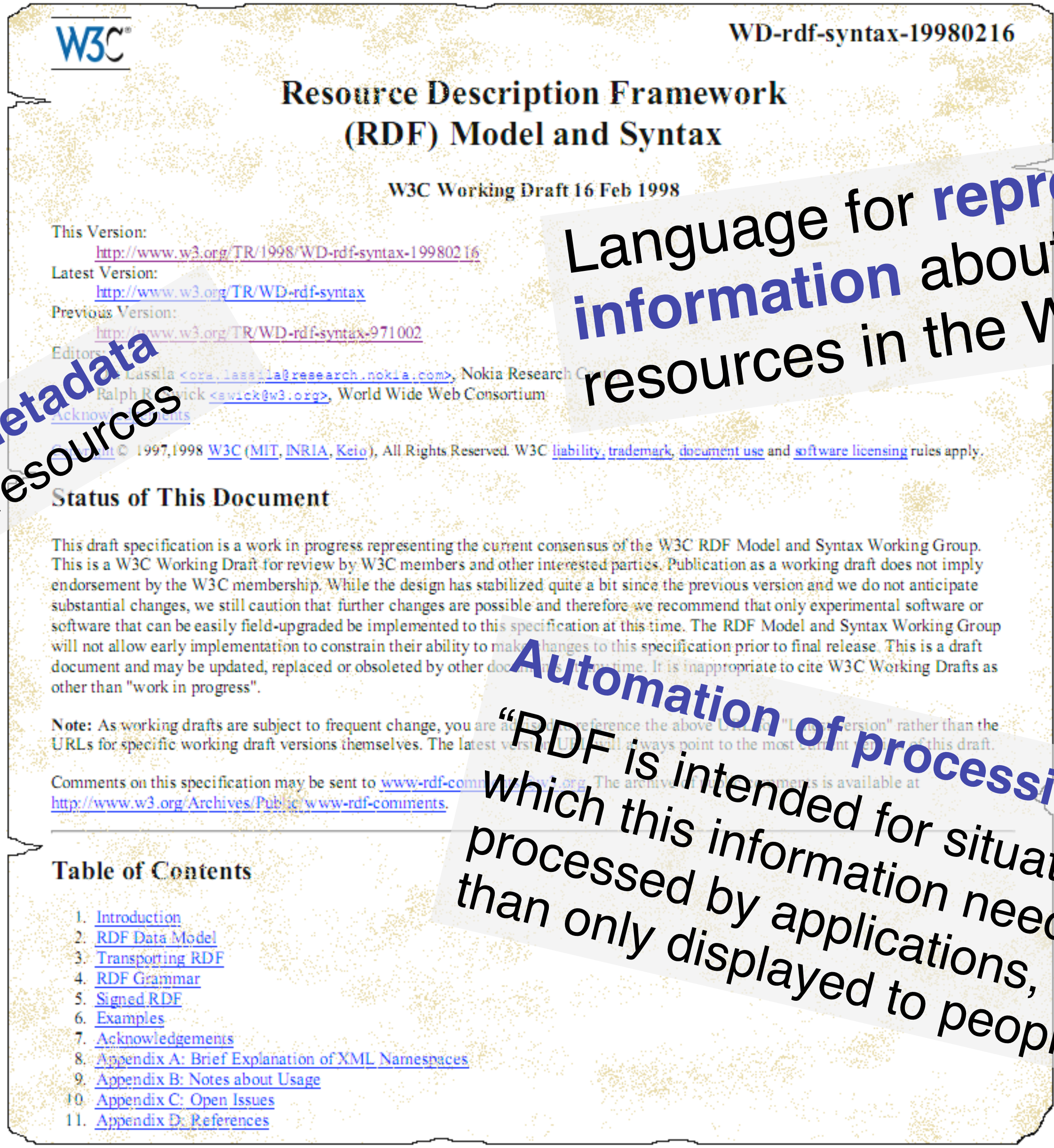
Not a survey!

The base of the Semantic Web is RDF

“The **Semantic Web** is the representation of **data** on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. **It is based on the Resource Description Framework (RDF)**”

http://www.w3.org/2001/sw/

RDF Recommendation (1999)



WD-rdf-syntax-19980216

Resource Description Framework (RDF) Model and Syntax

W3C Working Draft 16 Feb 1998

This Version:
http://www.w3.org/TR/1998/WD-rdf-syntax-19980216
Latest Version:
http://www.w3.org/TR/WD-rdf-syntax
Previous Version:
http://www.w3.org/TR/WD-rdf-syntax-971002
Editors:
Ora Lassila <ora.lassila@research.nokia.com>, Nokia Research Center
Ralph R. Swick <swick@w3.org>, World Wide Web Consortium
Acknowledgments

Status of This Document

This draft specification is a work in progress representing the current consensus of the W3C RDF Model and Syntax Working Group. This is a W3C Working Draft for review by W3C members and other interested parties. Publication as a working draft does not imply endorsement by the W3C membership. While the design has stabilized quite a bit since the previous version and we do not anticipate substantial changes, we still caution that further changes are possible and therefore we recommend that only experimental software or software that can be easily field-upgraded be implemented to this specification at this time. The RDF Model and Syntax Working Group will not allow early implementation to constrain their ability to make changes to this specification prior to final release. This is a draft document and may be updated, replaced or obsoleted by other documents at any time. It is inappropriate to cite W3C Working Drafts as other than "work in progress".

Note: As working drafts are subject to frequent change, you are advised to reference the above URL for "Latest version" rather than the URLs for specific working draft versions themselves. The latest version URL will always point to the most current version of this draft.

Comments on this specification may be sent to www-rdf-comments@w3.org. The archive of public comments is available at http://www.w3.org/Archives/Public/www-rdf-comments.

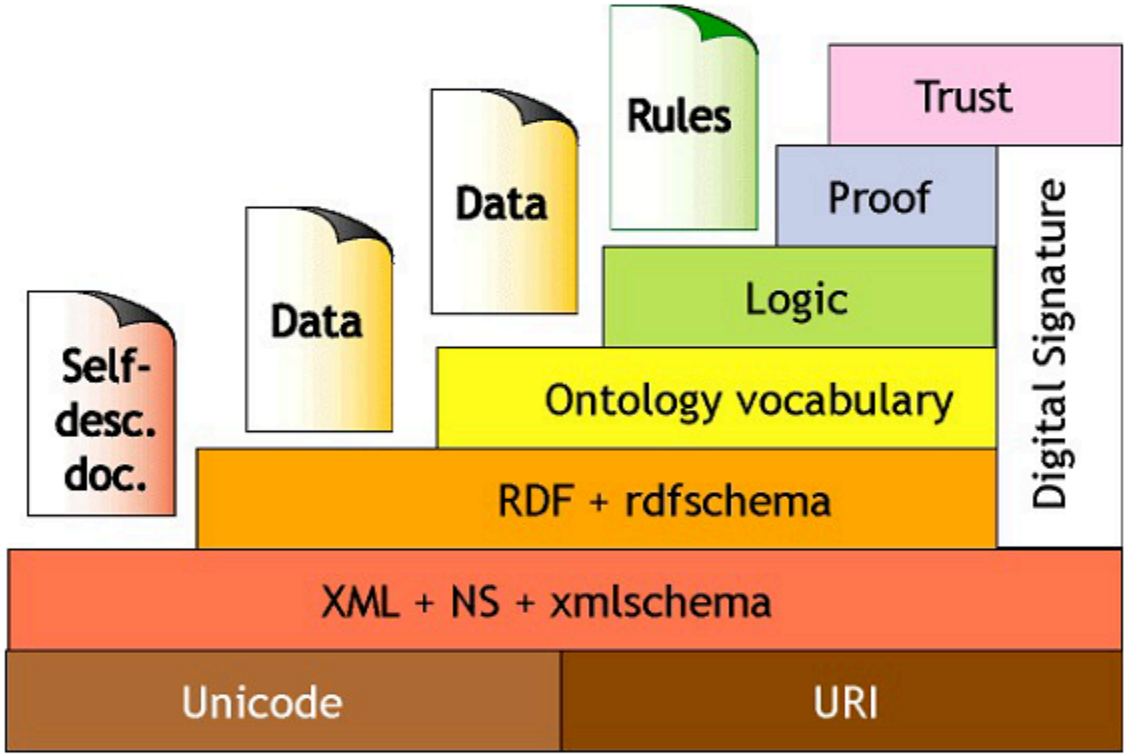
Table of Contents

Language for **representing information** about resources in the Web

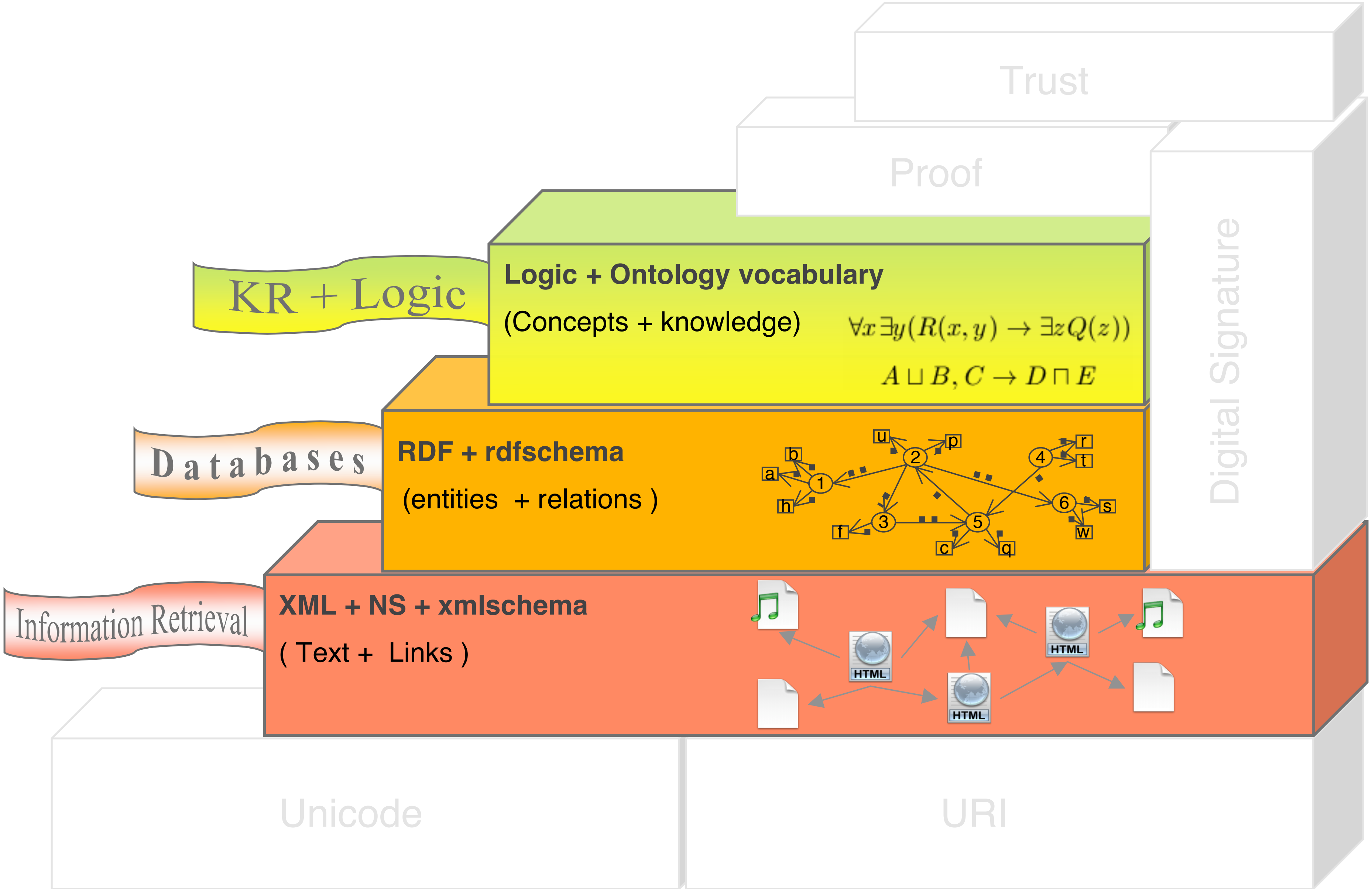
Particularly **metadata** about Web resources

Automation of processing: "RDF is intended for situations in which this information needs to be processed by applications, rather than only displayed to people"

Layers of the Semantic Web

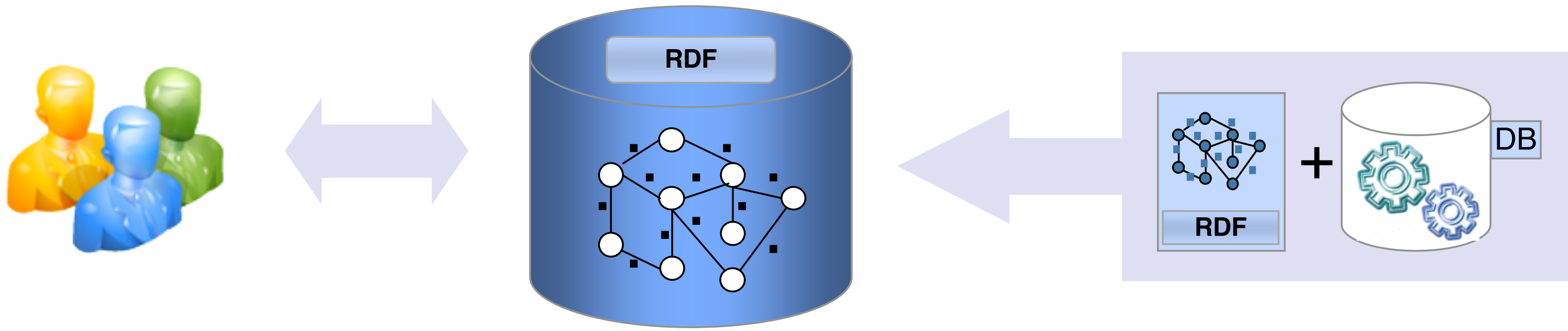


A Data Processing perspective



The Database Approach

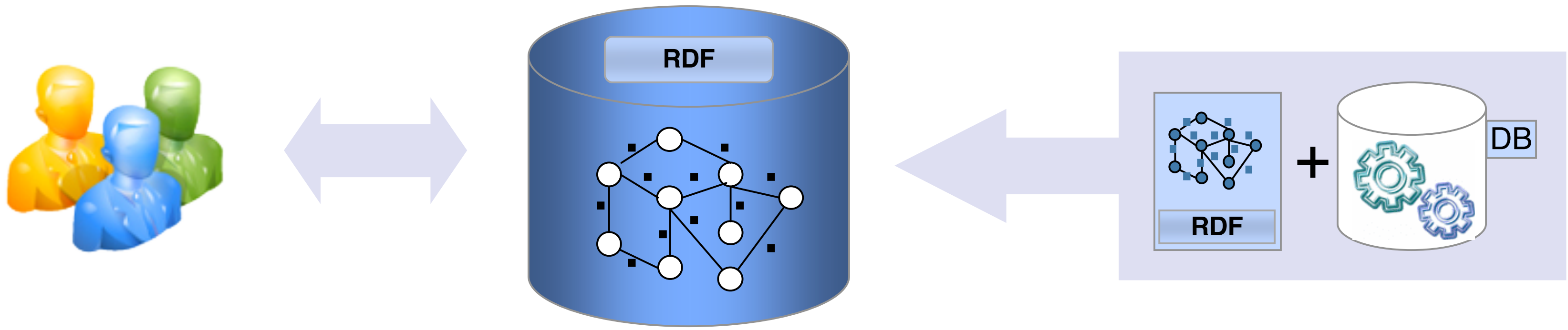
- Manage huge volumes of data with logical precision
- Separate modeling from implementation levels



The Database Approach

- Manage huge volumes of data with logical precision
- Separate modeling from implementation levels

As opposed to AI: DB primary concern is scalability. Then expressive power

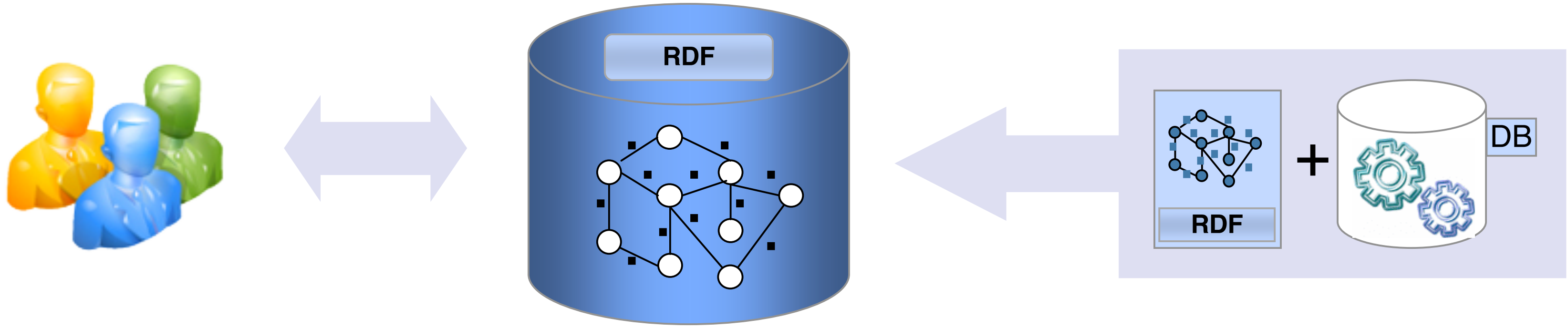


The Database Approach

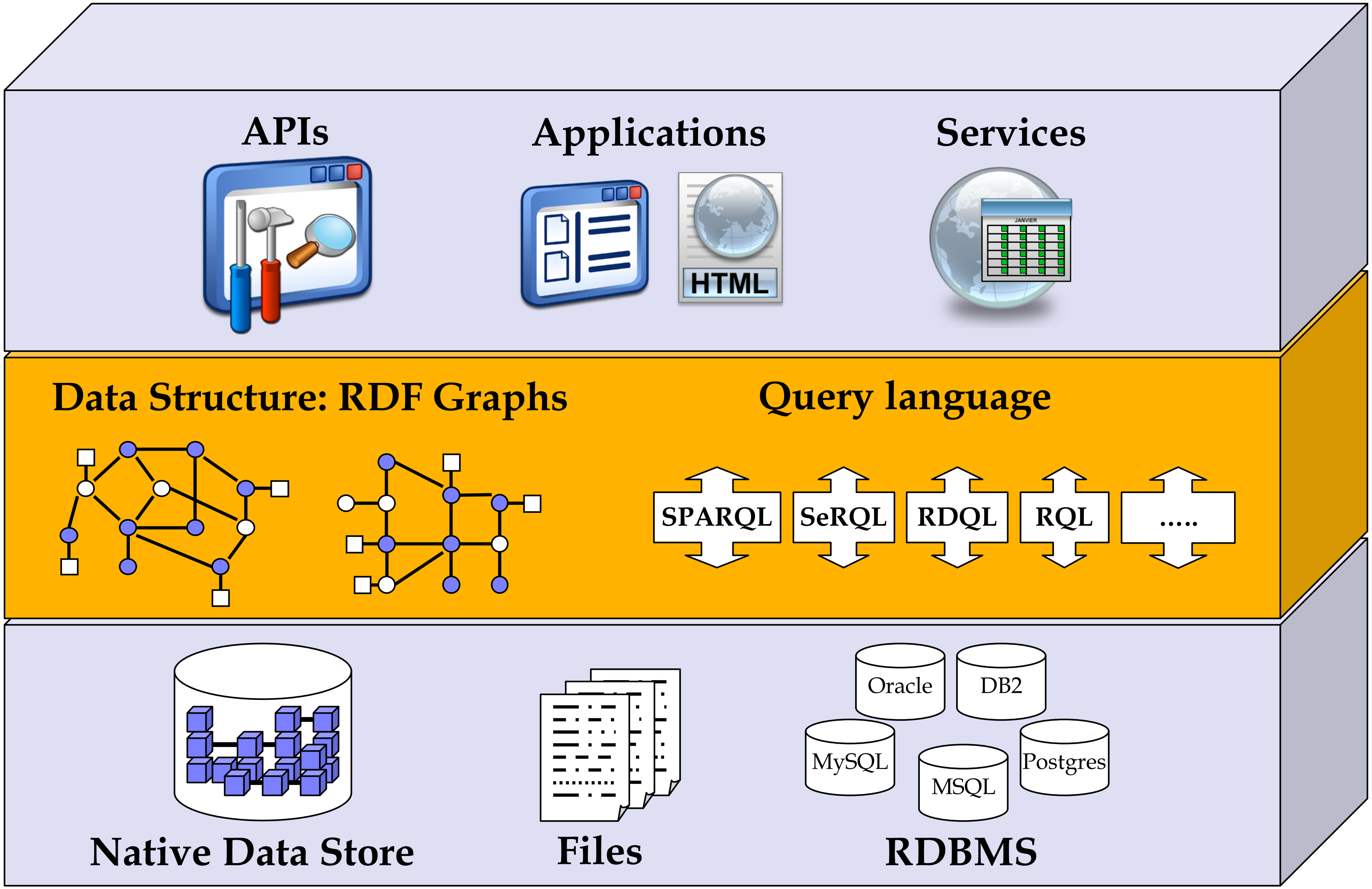
- Manage huge volumes of data with logical precision
- Separate modeling from implementation levels

As opposed to AI: DB primary concern is scalability. Then expressive power

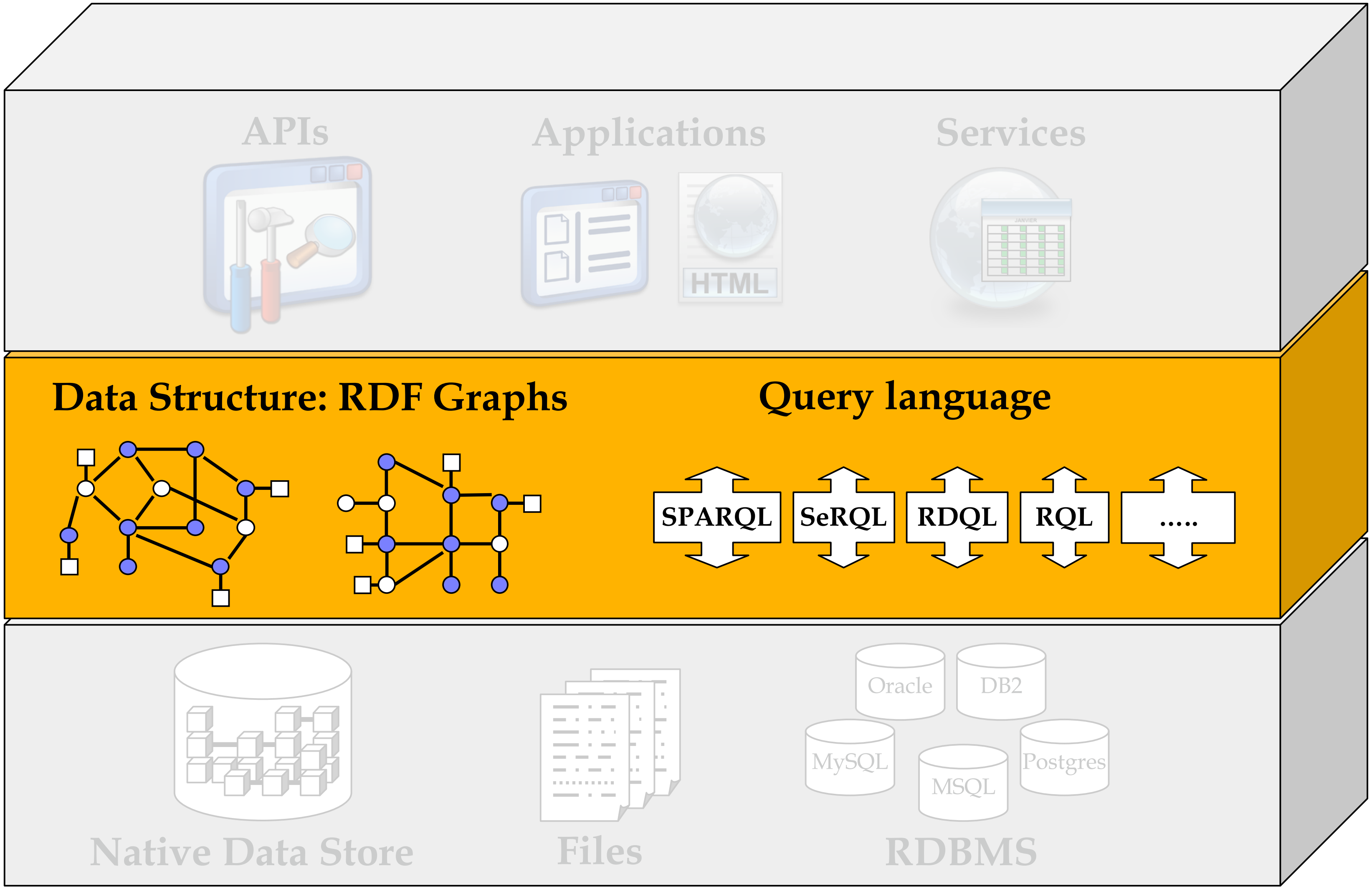
As opposed to IR: DB primary concern is precision. Then scalability (recall).



RDF Database Technology



This Course: Database Modeling Level



This Course: Database Modeling Level

Hence leaving out:

- Visualization, APIs, Services, etc.
- Indexing, storing, transactions, etc.

Hence leaving out:

- Visualization, APIs, Services, etc.
- Indexing, storing, transactions, etc.

But also leaving out:

Updating / Constraints / Temporality / Optimization / Aggregation / Flexibility / etc. / etc.

Agenda

1. RDF and Databases
2. RDF and Database models
3. RDF Query Language
   - Requirements and Domains
   - Manifold Views
4. SPARQL

Query Language

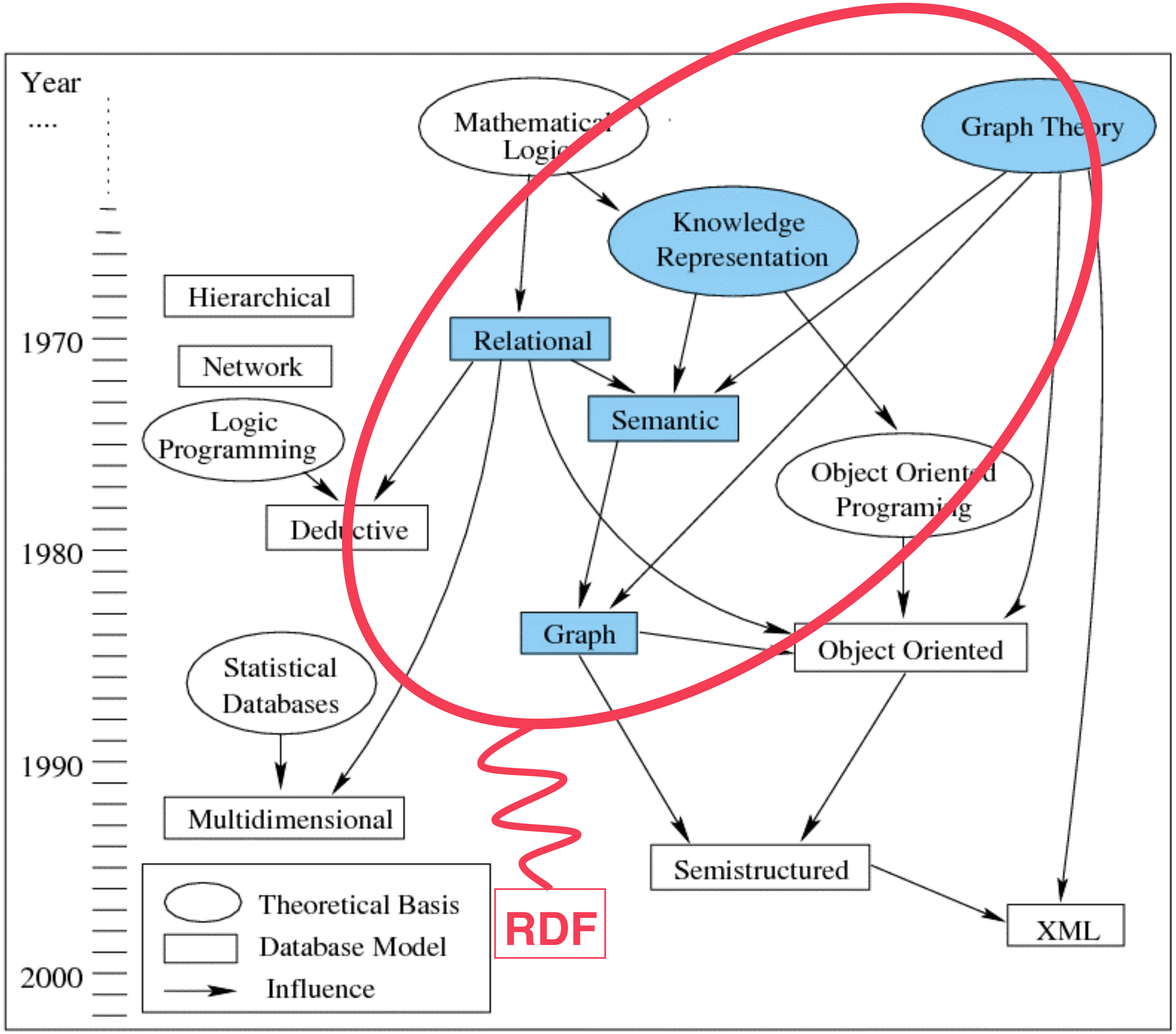
Integrity constraints

Data structures

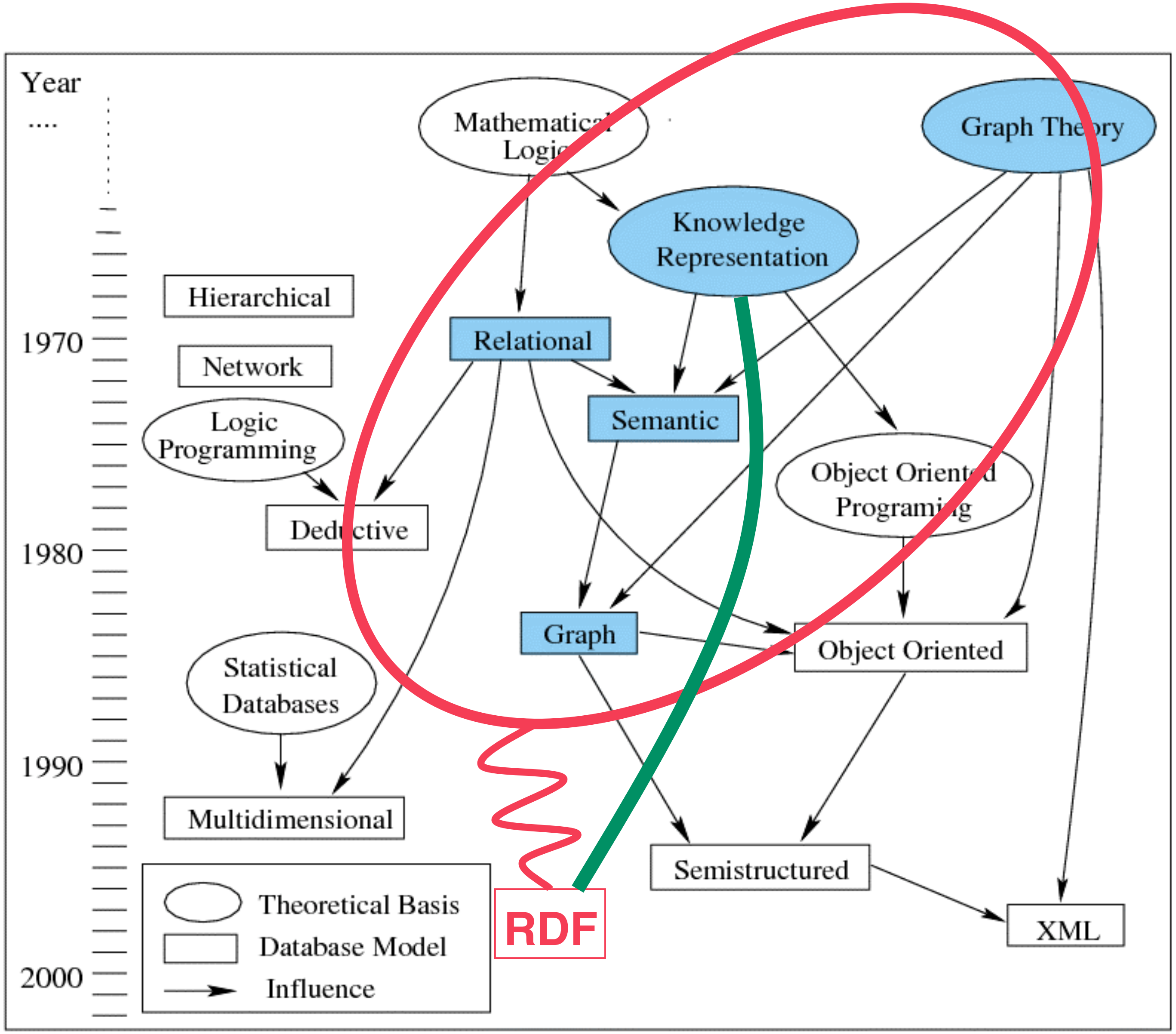
Query Language

Data structures

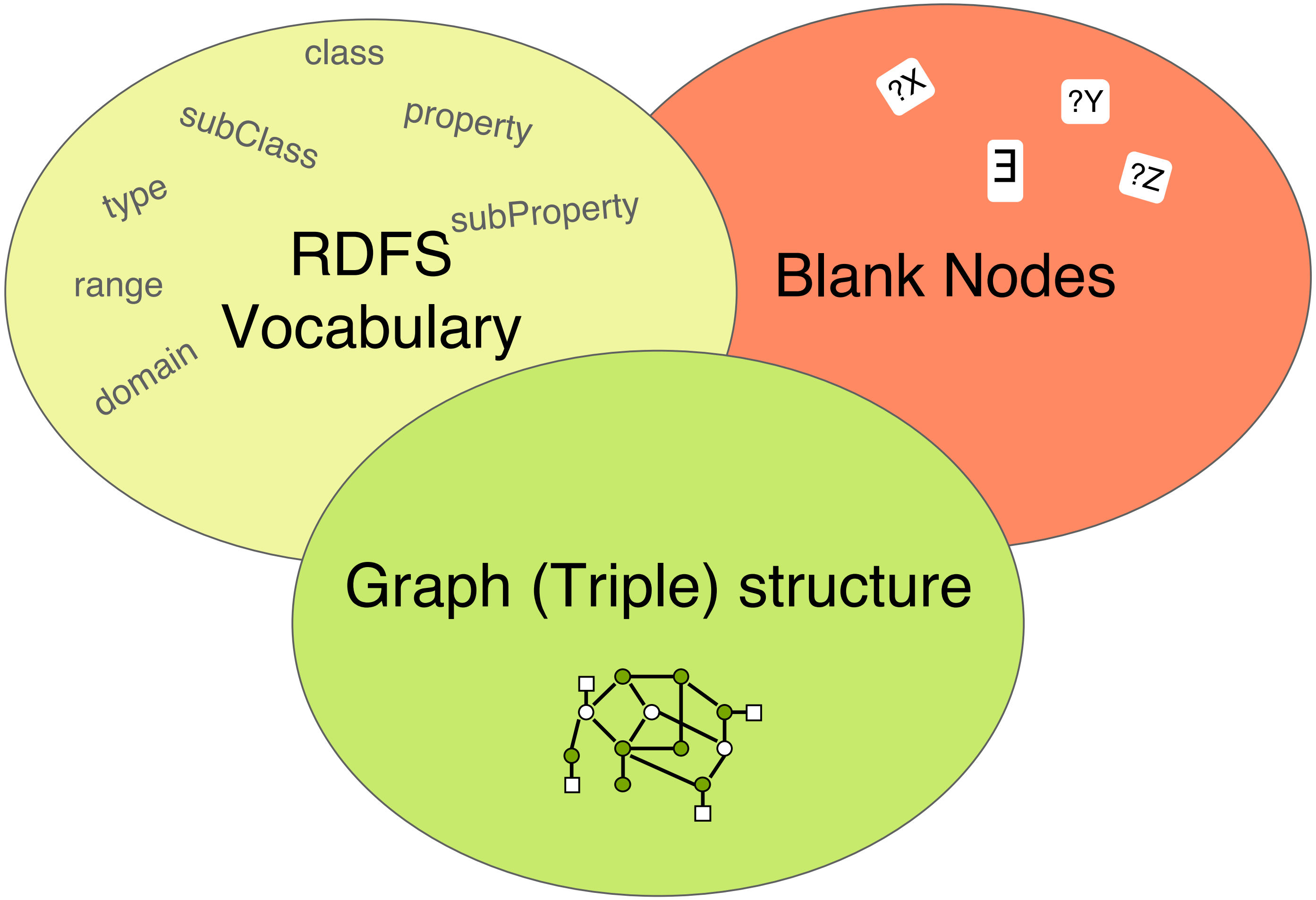
Evolution of Database Models



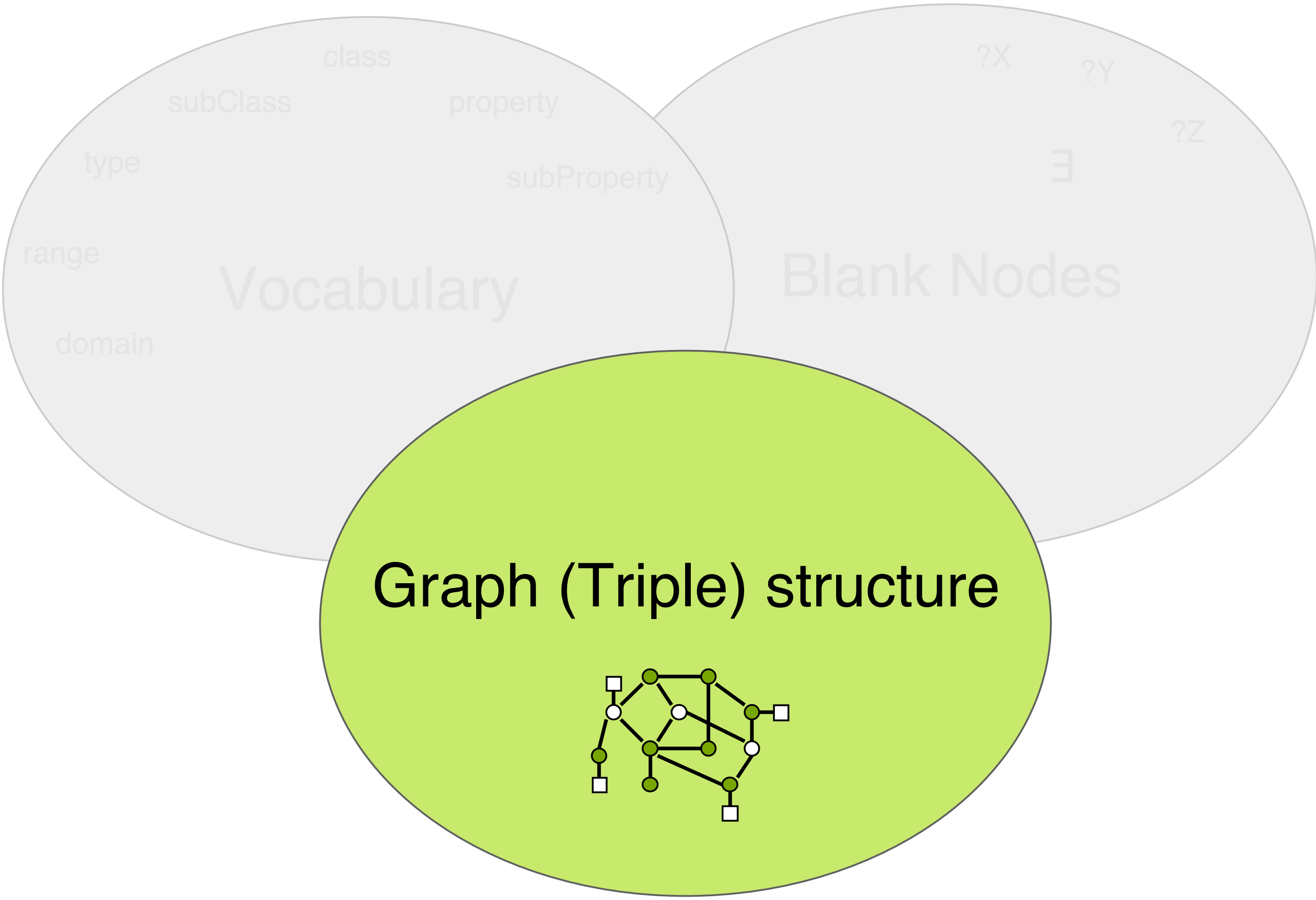
Evolution of Database Models



RDF Data Structure: three main blocks

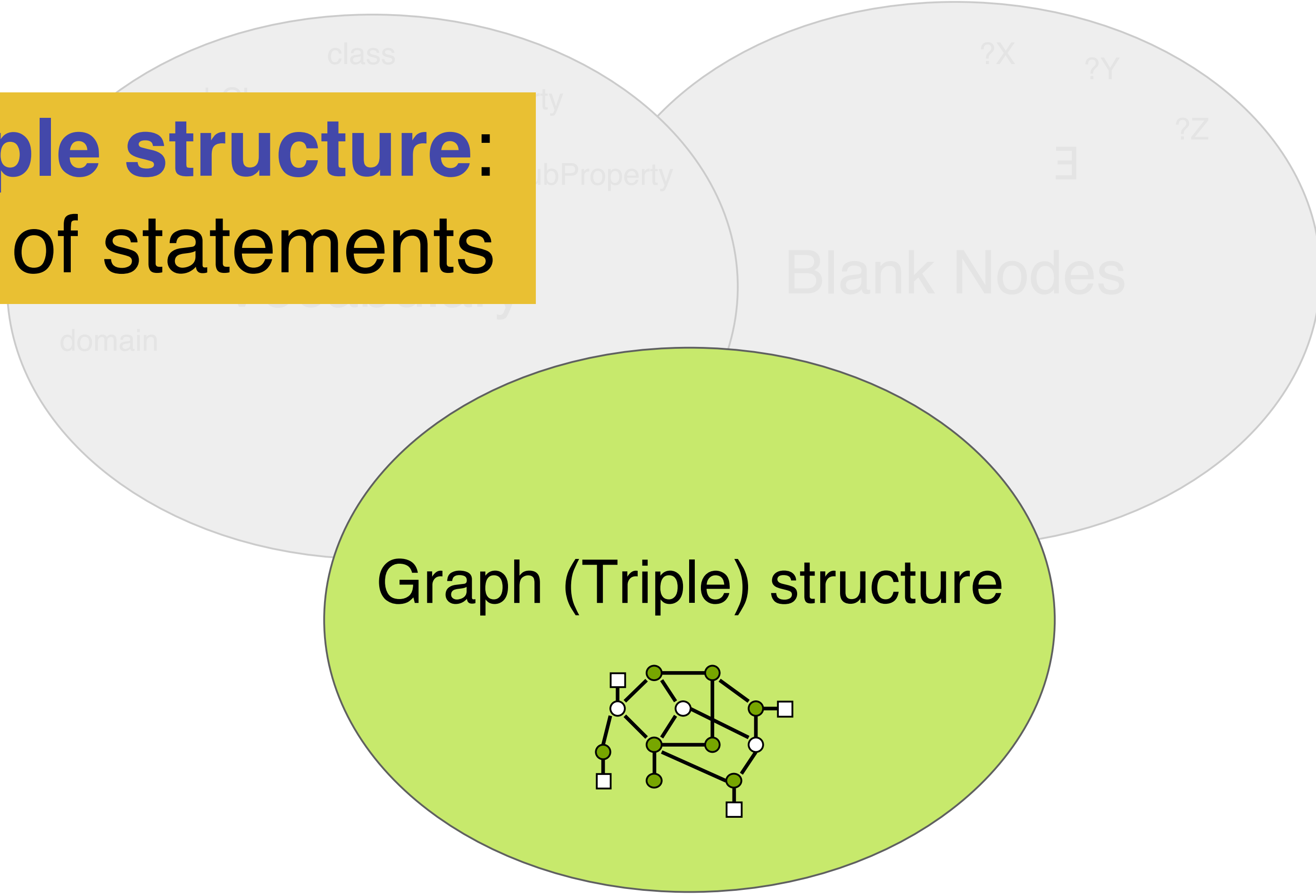


RDF Data Structure: the core



RDF Data Structure: the core

Triple structure:
set of statements



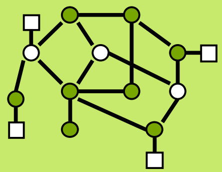
Triple structure: set of statements

Graph structure: linked network of statements.

class

domain

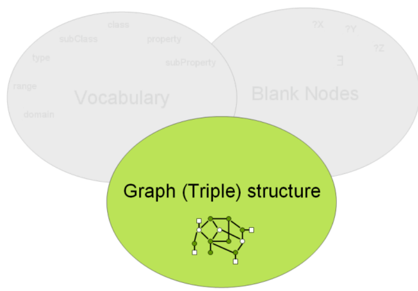
Graph (Triple) structure



RDF Data Structure: Relational Tables (Triple) view

- Triples as tuples
- Set of triples as Tables

Subject	Predicate	Object



RDF Data Structure: Relational Tables (Triple) view

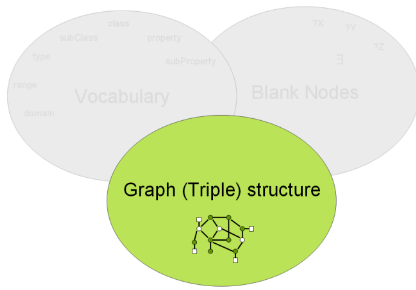
- Triples as tuples
- Tables of triples

Subject	Predicate	Object

Advantages:

+ Well studied and well understood

+ Reuse relational technologies



RDF Data Structure: Relational Tables (Triple) view

- Triples as tuples
- Tables of triples

Subject	Predicate	Object

Advantage

\+ Well studie
well und

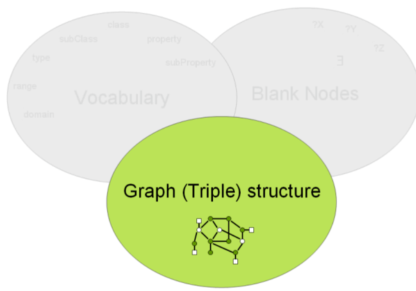
\+ Reuse relat
technolog

Problems (Questions):

\- Why yet another syntax for the relational model?

\- Was this the intended objective of RDF?

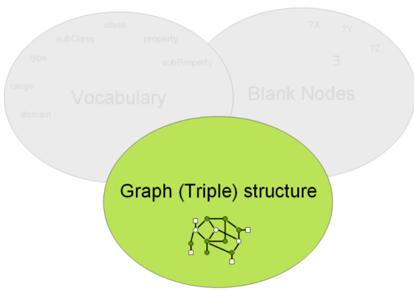
\- Expressive power limitations



RDF Data Structure: Graph Database Model view

Graph Database Models:

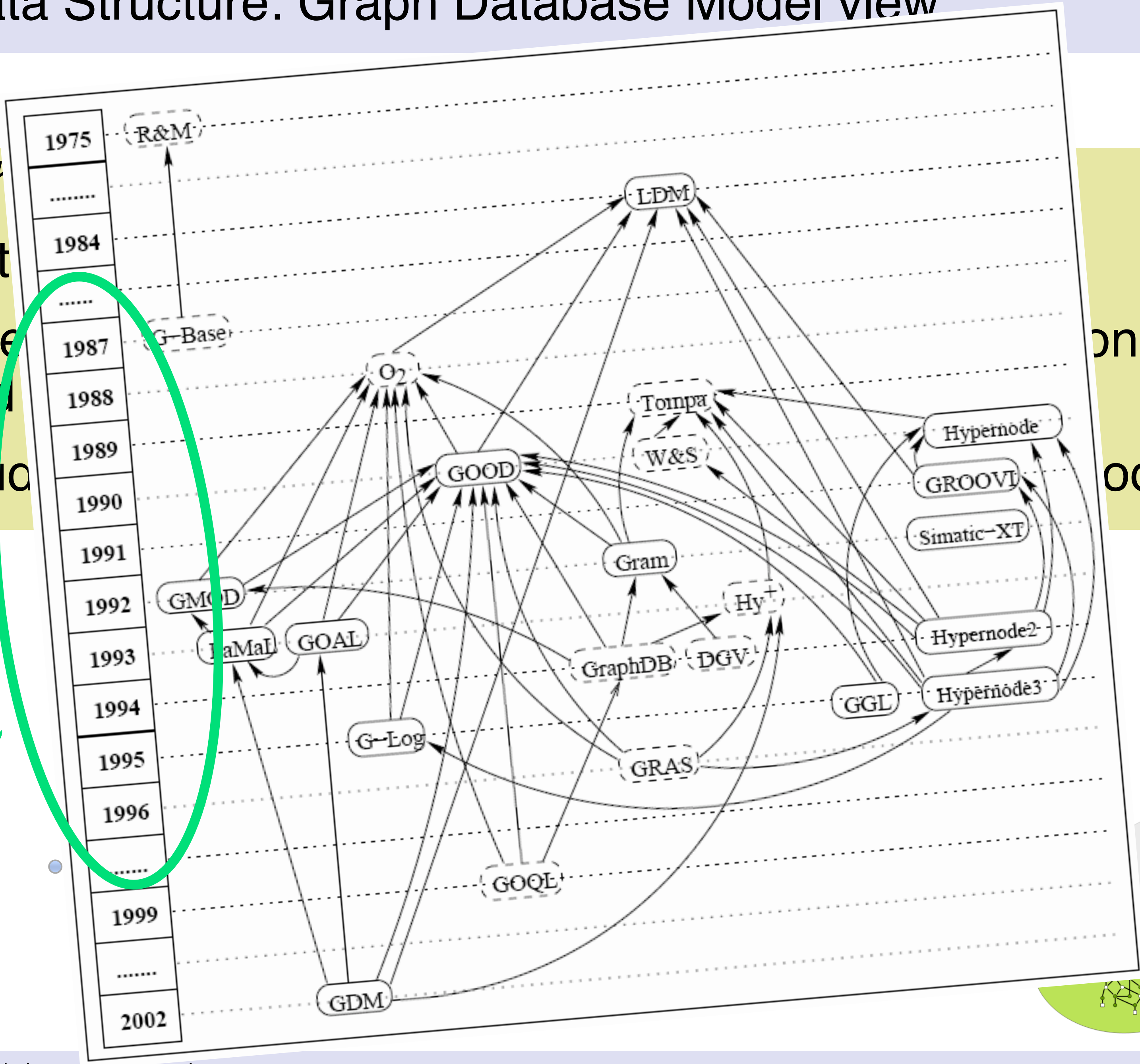
- Data and/or schema are represented by graphs
- Query language able to capture main graph operations and properties
- Studied by DB community, but still not well understood



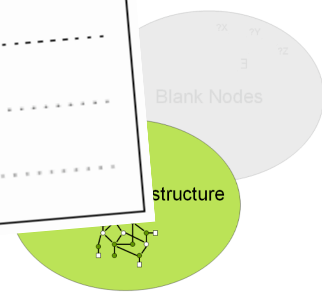
RDF Data Structure: Graph Database Model view

Graph

- Dat
- Que ... ons and
- Stu ... od



Golden age



RDF Data Structure: Graph query languages

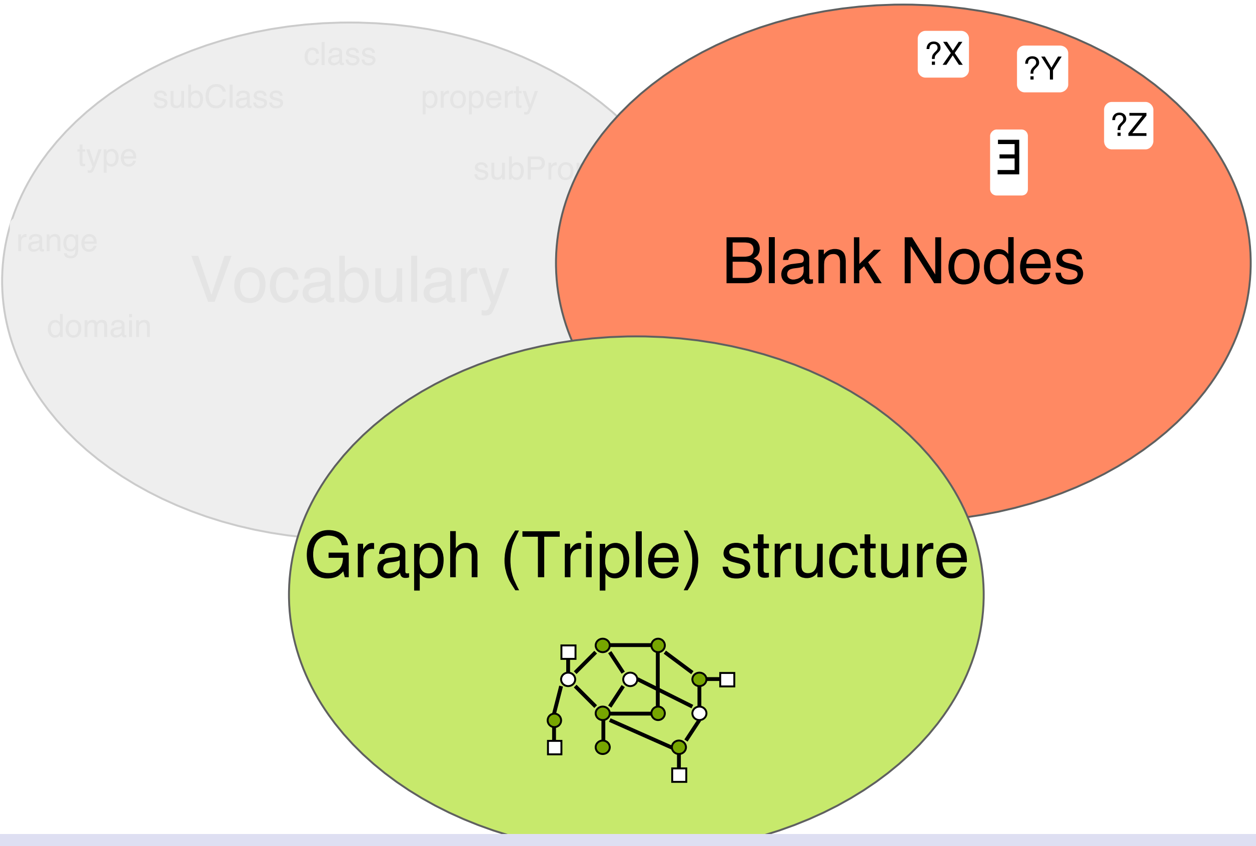
PROPERTY		Neighbo rhoods	Adjacent Edges	Degree of a Node	Path	Fixed-length path	Distance	Diameter
Graph Query Language	G							
	G+							
	Graph Log							
	Gram							
	Graph DB							
	Lorel							
	F-G							

RDF Data Structure: Graph query languages

PROPERTY		Neighborhoods	Adjacent Edges	Degree of Node	Path	Fixed-length path	Distance	Diameter
Graph Query Language	G							
	G+							
	Graph Log							
	Gram							
	Graph DB							
	Lorel							
	F-G							

Green light for graph features!

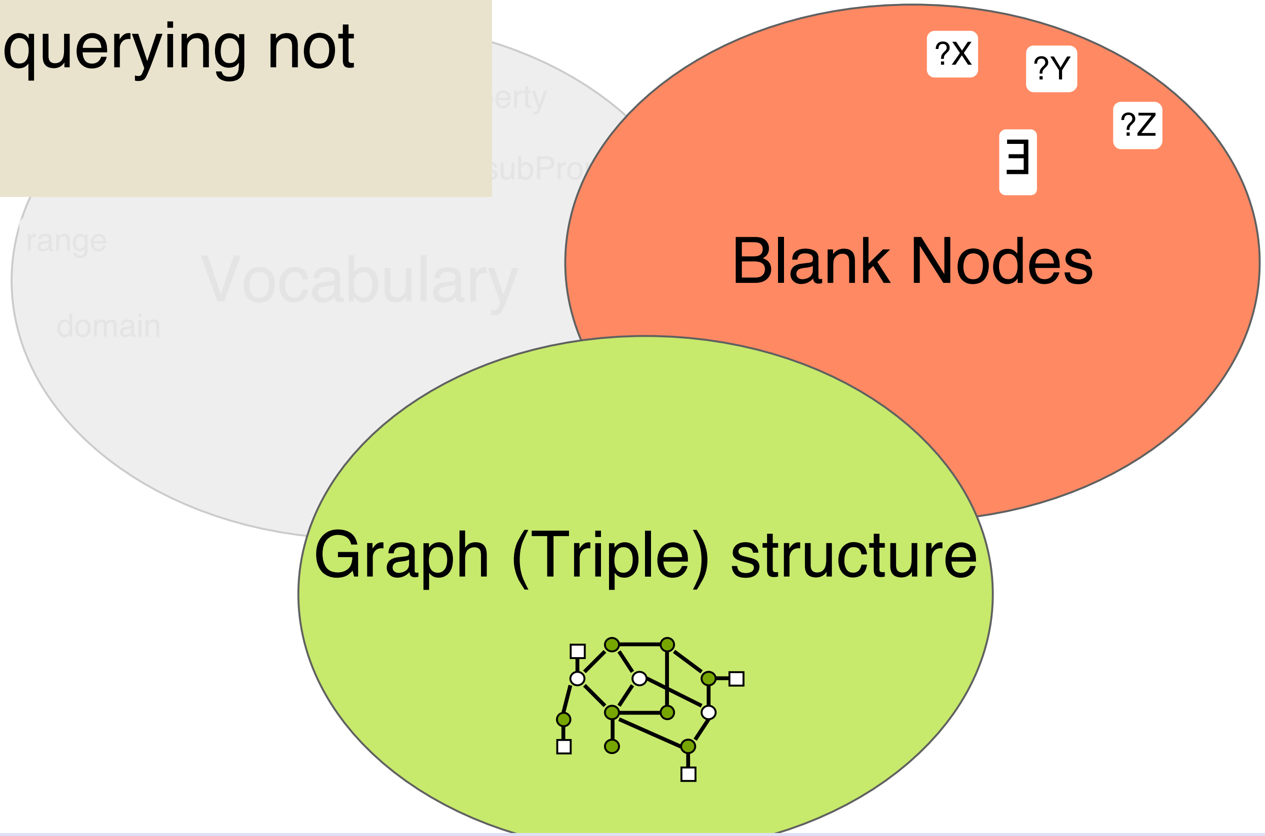
RDF Data Structure: Triple structure + Blank nodes



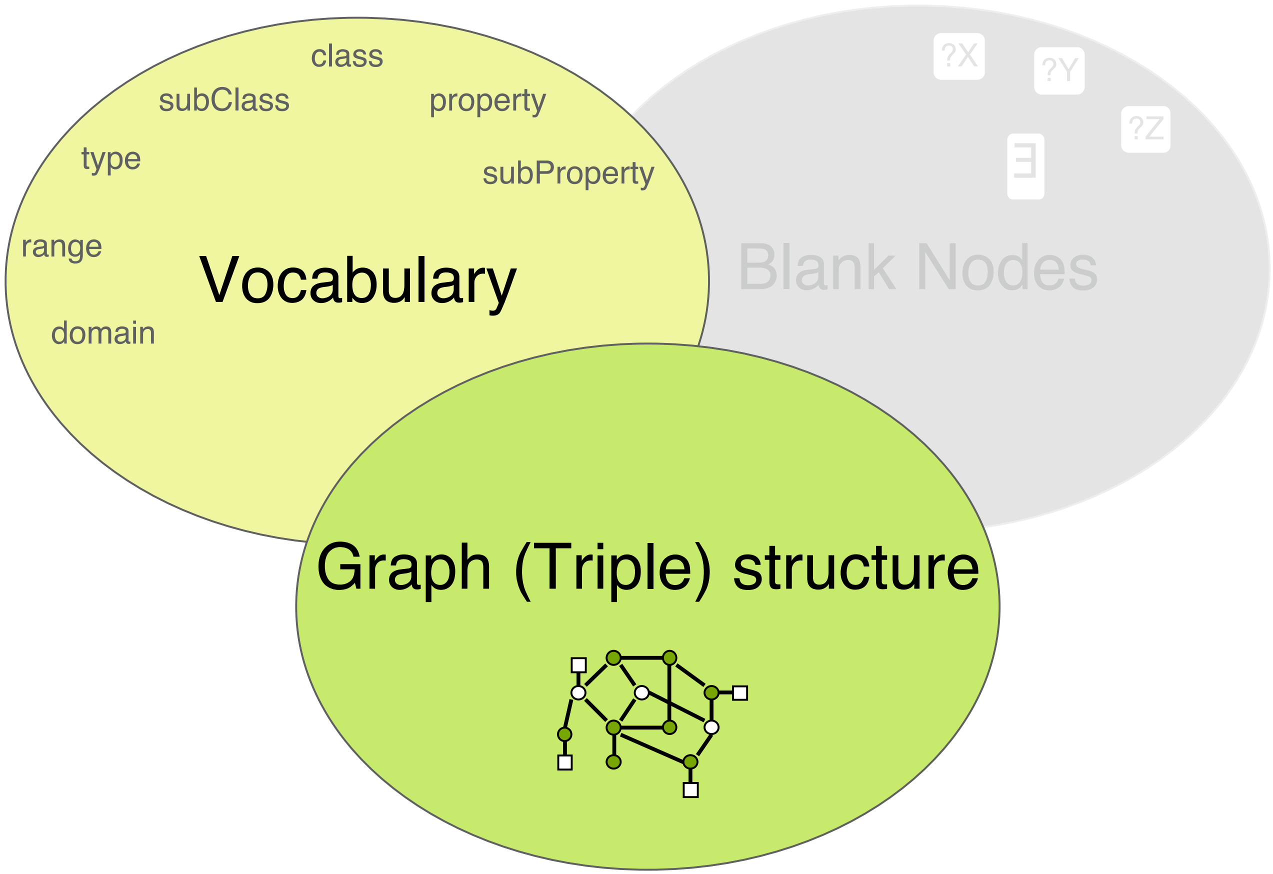
RDF Data Structure: Triple structure + Blank nodes

Complexity / Semantics issues:

- Deciding entailment becomes NP-complete.
- Deciding core is DP-complete
- Semantics of querying not simple

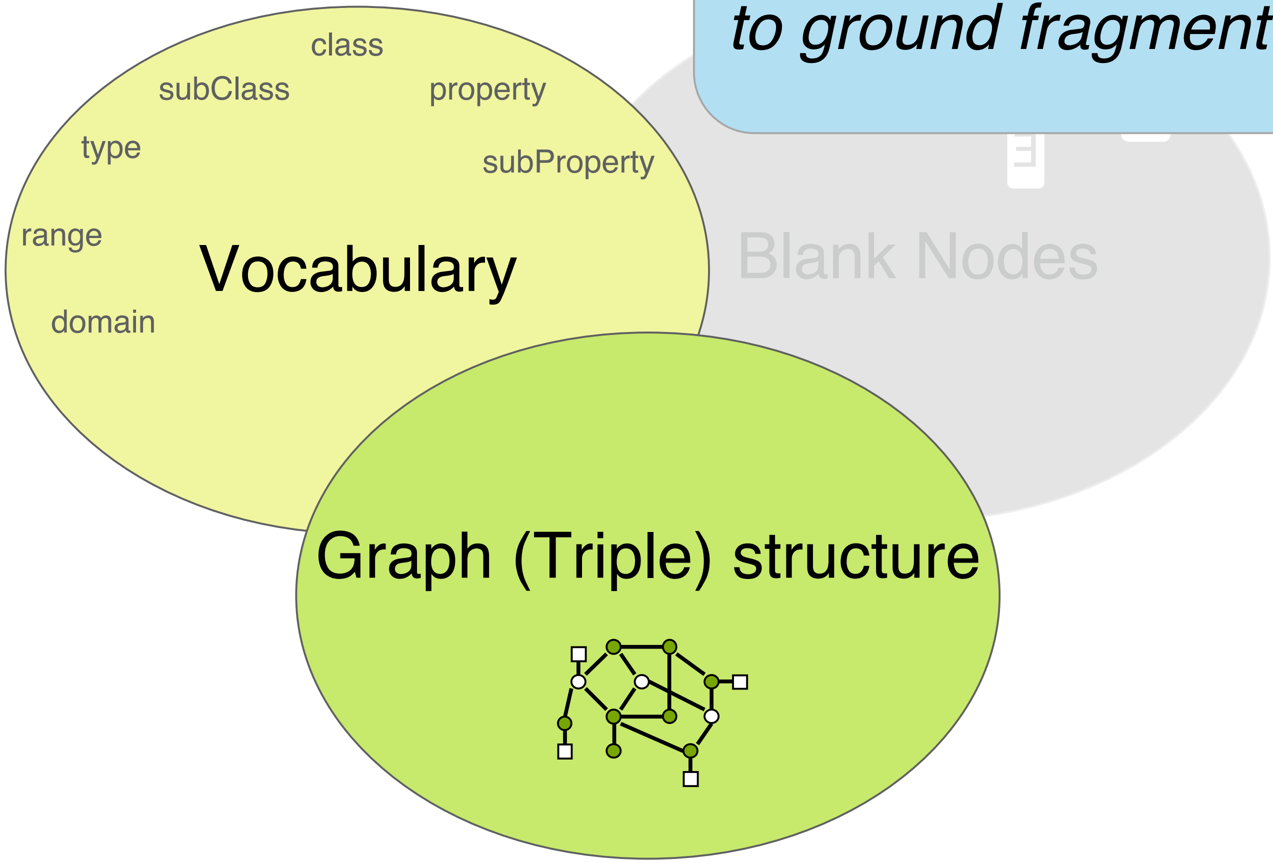


RDF Data Structure: Ground fragment



RDF Data Structure: Ground fragment

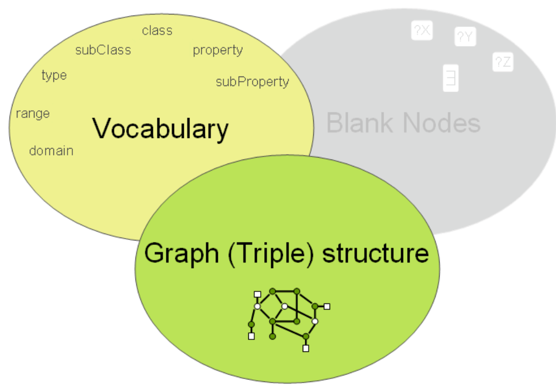
Good News: *Blank nodes can be treated orthogonally to ground fragment.*



More good news:

- Vocabulary can be *reduced* to

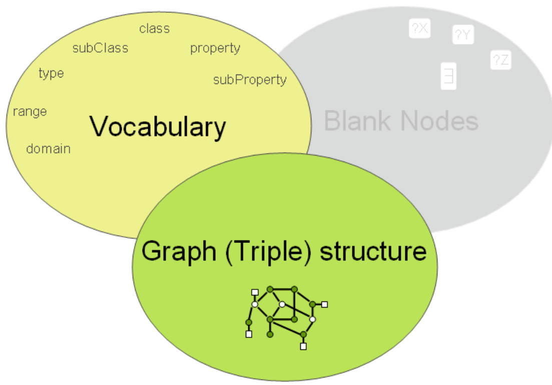
  { type, domain, range, subClassOf, subPropertyOf }



RDF Data Structure: Ground fragment

More good news:

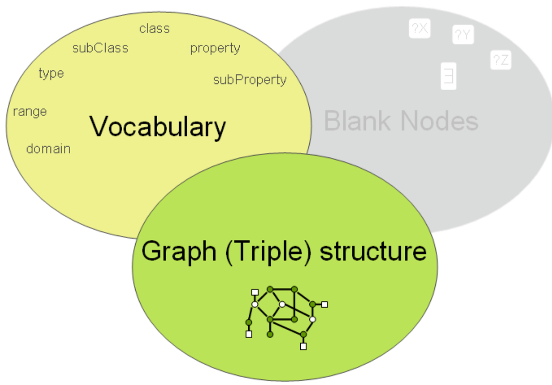
- Vocabulary can be *reduced* to
  { type, domain, range, subClassOf, subPropertyOf }
- Complex semantic rules and axioms can be *avoided*



RDF Data Structure: Ground fragment

More good news:

- Vocabulary can be *reduced* to
  { type, domain, range, subClassOf, subPropertyOf }
- Complex semantic rules and axioms can be *avoided*
- Structural (internal) constraints of the language can be *separated* from user-features.
  e.g. (Class, type, Resource)

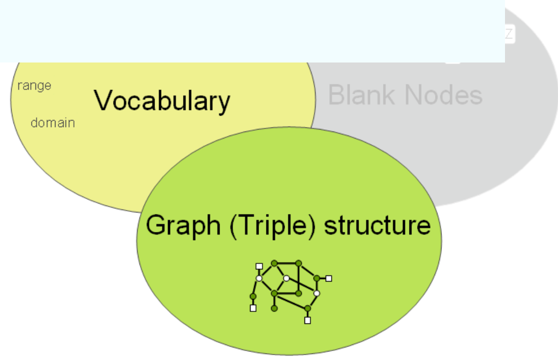


More good news:

- Vocabulary can be *reduced* to

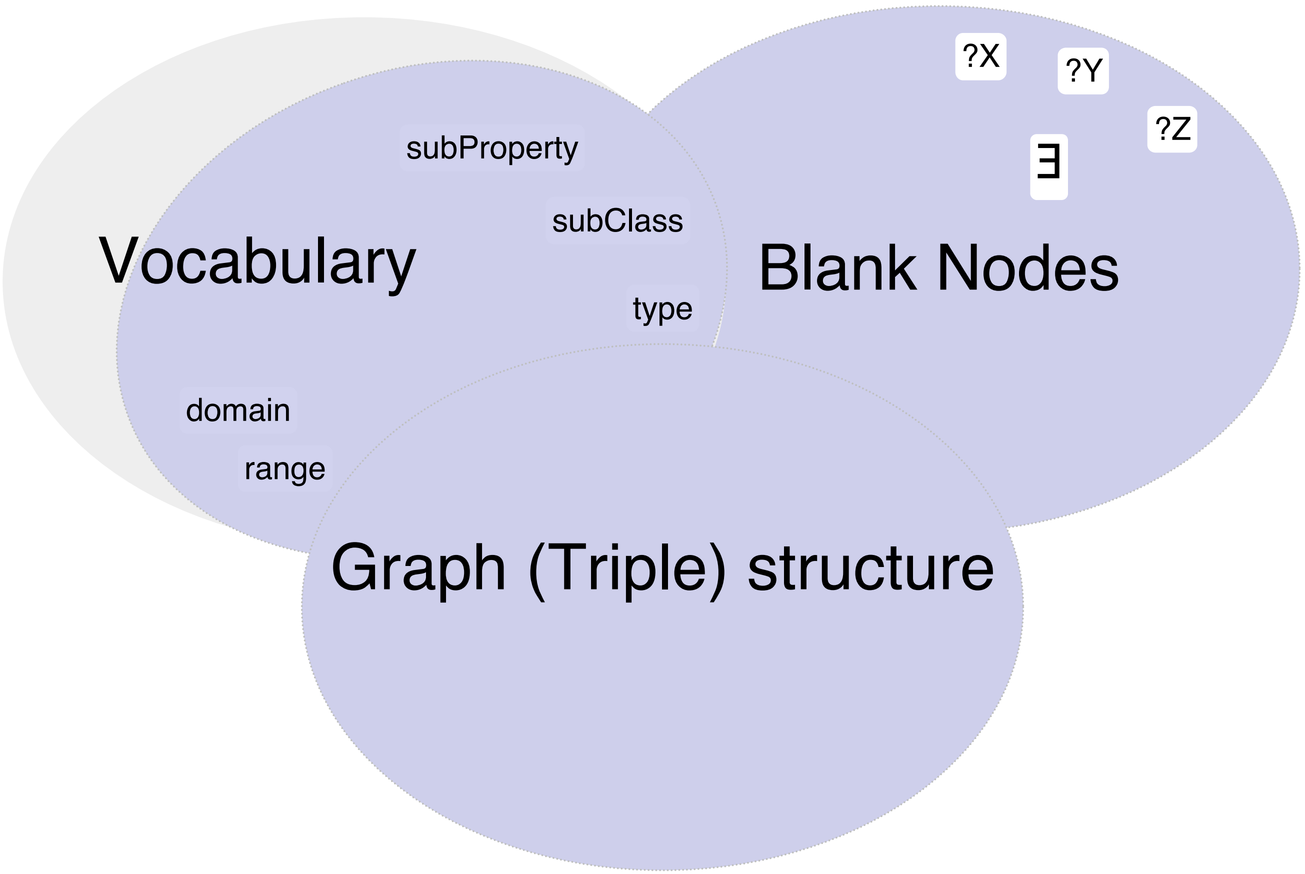
  { type, domain, range, subClassOf, subPropertyOf }
- Complex semantic rules and axioms can be *avoided*
- Structural (internal) constraints of the language can be *separated* from user-features.

  e.g. (Class, type, Resource)
- Features which do not add expressive power can be *avoided*, e.g. reflexivity of subClass and subProperty.

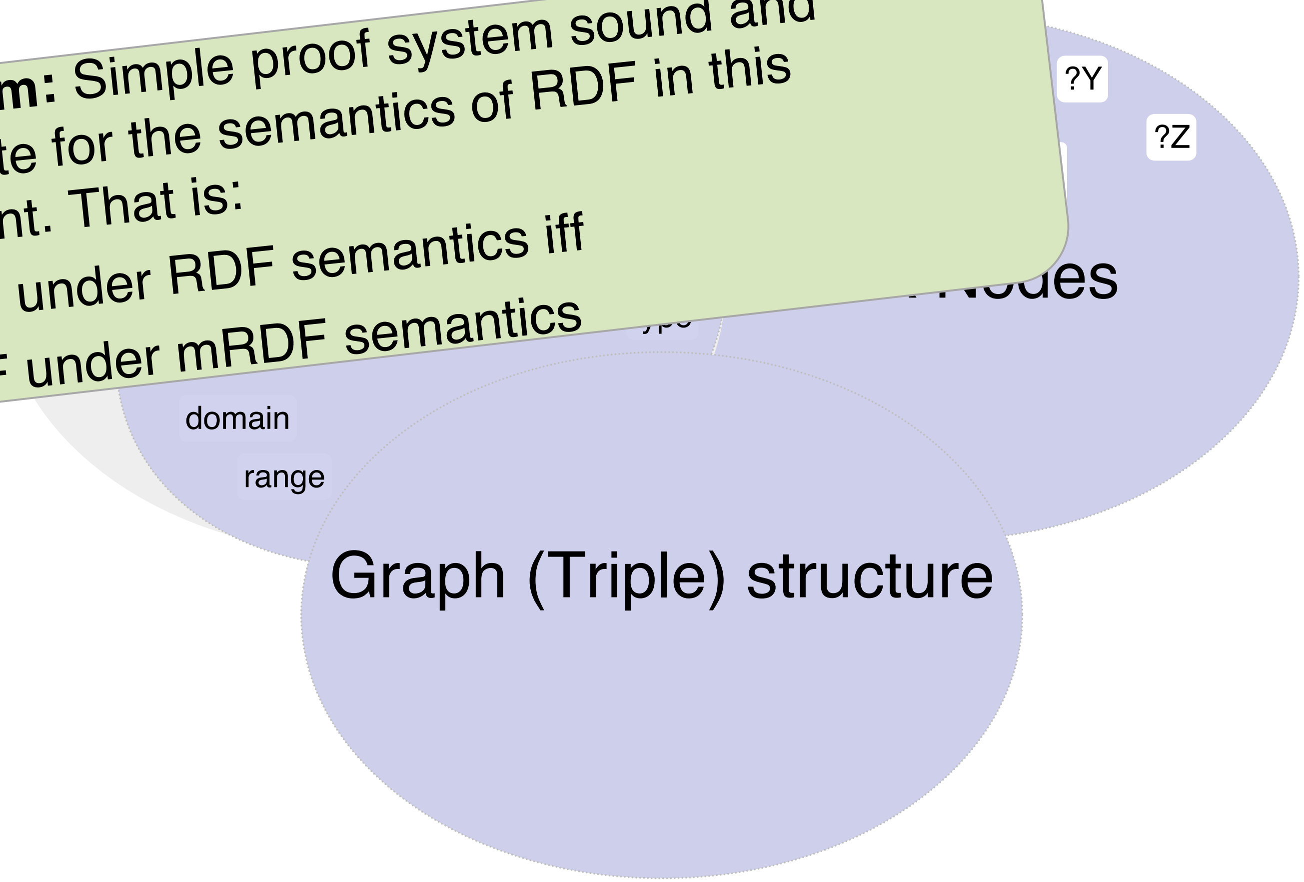


RDF Data Structure: A *minimal fragment*

{subClass, subProperty, type, domain, range}



{subClass, subProperty, type, domain, range}

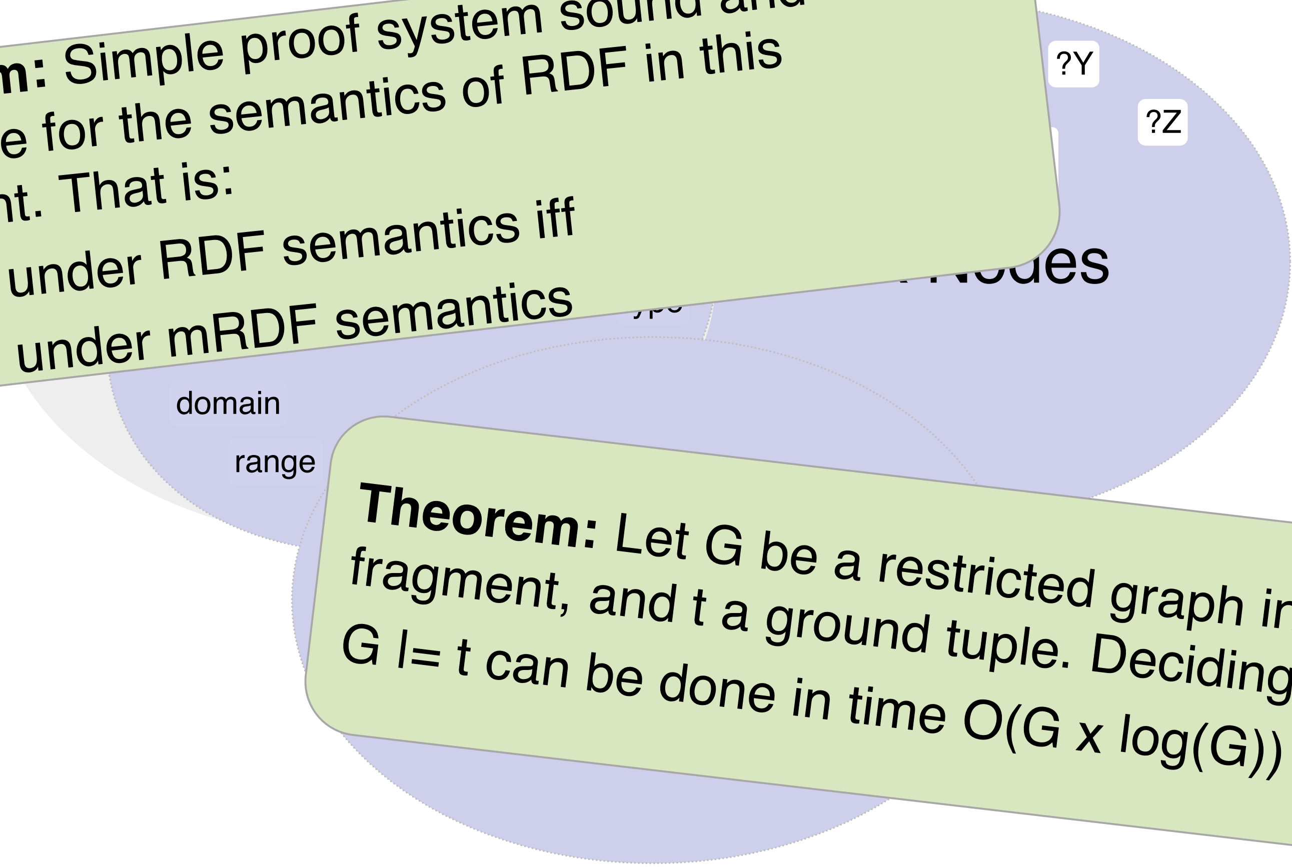


Theorem: Simple proof system sound and complete for the semantics of RDF in this fragment. That is:

G |= F under RDF semantics iff
G |= F under mRDF semantics

RDF Data Structure: A *minimal fragment*

{subClass, subProperty, type, domain, range}



Theorem: Simple proof system sound and complete for the semantics of RDF in this fragment. That is:

G |= F under RDF semantics iff

G |= F under mRDF semantics

Theorem: Let G be a restricted graph in the fragment, and t a ground tuple. Deciding if G |= t can be done in time O(G x log(G))

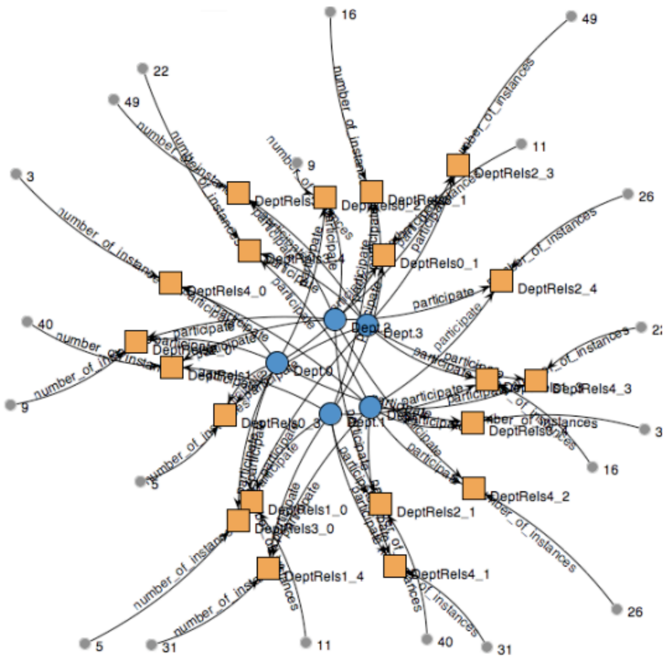
Agenda

1. RDF and Databases
2. RDF and Database models
3. RDF Query Language
   – Requirements and Domains
   – Manifold Views
4. SPARQL

RDF Query language: Social Networks domain

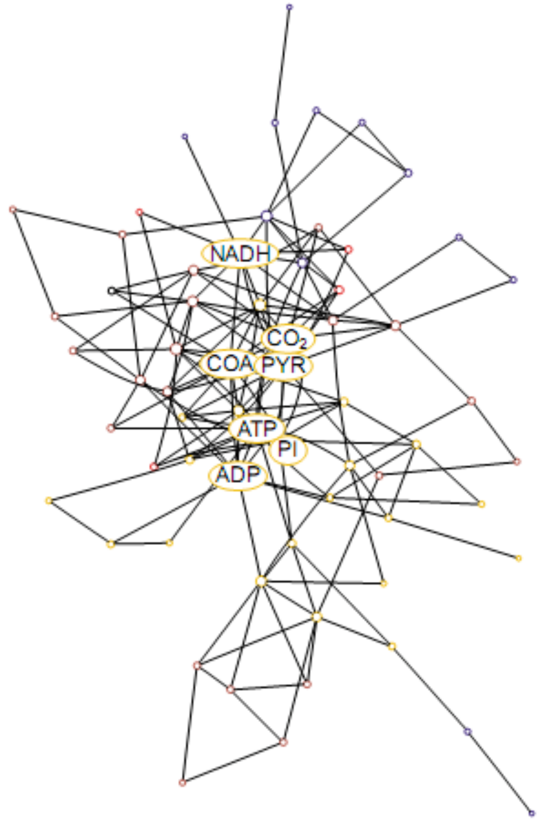
Chapter title	**Use Case (local)**
Looking for Social Structure	+ Directed to undirected binary relations + Remove relations
Attributes and Relations	+ Extract a subnetwork based on attributes + Group actors based on attributes + Selective grouping of actors based on attributes
Cohesive Subgroups	+ Extract the subnetwork induced by cliques of size n + Build a hierarchy of cliques
Frienship	+ Extract subnetwork by time
Affiliations	+ Two-mode network to one-mode network
Center and Periphery	+ Group multiple binary relations
Brokers and Bridges	+ Extract egonetwork of an actor + Remove relations between groups
Diffusion	+ Selective counting of neighbors + Operations between attributes + Change relation direction based on attributes
Prestige	+ Discretize an attribute
Ranking	+ Find triads by type
Genealogies and Citations	+ Loop removal

Subgraph family	**Use Case (Global)**
Paths and Cycles	+ Geodesics
Groups (k-neighbors, k-core, n-cliques, k-plex, etc.)	+ Detect cohesive subgroups + Egonetworks + Input Domain
Connected components	+ Connected components + Clustering + Bicomponents and brockers



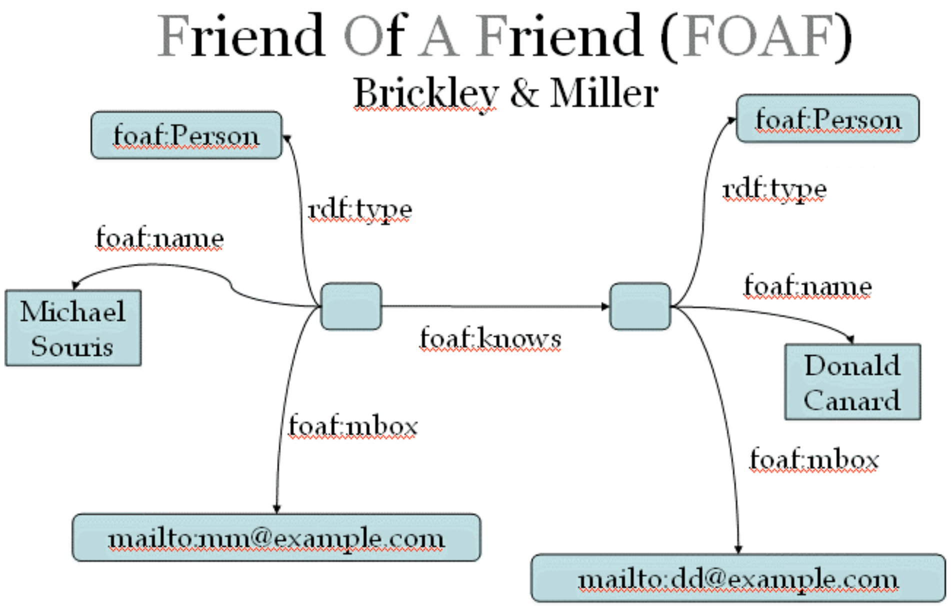
RDF Query Language: Biology domain

Use Case	Graph Query
Chemical structure associated with a node	Node matching
Find the difference in metabolisms between two microbes	Graph intersection, union, difference
To combine multiple protein interaction graphs	Majority graph query
To construct pathways from individual reactions	Graph composition
To connect pathways, metabolism of co-existing organisms	Graph composition
Identify “important” paths from nutrients to chemical outputs	Shortest path queries
Find all products ultimately derived from a particular reaction	Transitive Closure
Observe multiple products are co-regulated	Least common ancestor
To find biopathways graph motifs	Frequent subgraph recognition
Chemical info retrieval	Subgraph isomorphism
Kinaze enzyme	Subgraph homomorphism
Enzyme taxonomies	Subsumption testing
To find biopathways graph motifs	Frequent subgraph recognition



RDF Query Language: Web domain

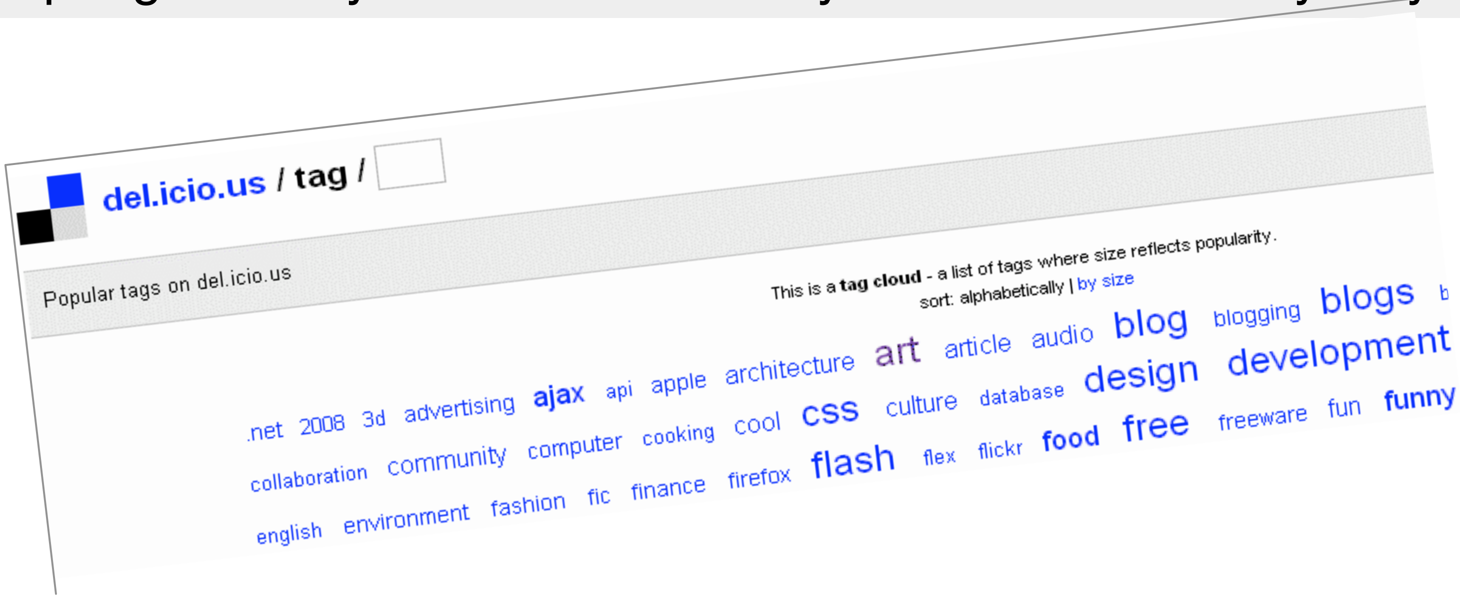
Use Case	Graph Query
What is/are the most cited paper/s?	Degree of a node
What is the influence of article D?	Paths
What is the Erdös distance between authos X and author Y?	Distance
Are suspects A and B related?	Paths
All relatives of degree one of Alice	Adjacency



RDF Query Language: Tagging domain

Tags

A tag is simply a word you use to describe a bookmark. Unlike folders, you make up tags when you need them and you can use as many as you like.



Minimalist design:

- Tags + Bundles (classes)
- No inheritance, no intersection, etc.
- Renaming

RDF Query Language: Standardization's view

- **SQL:** Great for finding data from tabular representations, can get complex when many tables are involved in a given query

SQL, XQuery and SPARQL: What's Wrong with this Picture?

Jim Melton (Oracle; XML Query Working Group, XML Coord. Group)

Sixth annual W3C Technical Plenary (March 2006)

RDF Query Language: Standardization's view

- **SQL:** Great for finding data from tabular representations, can get complex when many tables are involved in a given query
- **XQuery:** Great for finding data in tree representations, can get complex when many relationships have to be traversed

SQL, XQuery and SPARQL: What's Wrong with this Picture?

Jim Melton (Oracle; XML Query Working Group, XML Coord. Group)

Sixth annual W3C Technical Plenary (March 2006)

Standardization's view (Jim Melton, Oracle, 2006)

- **SQL:** Great for finding data from tabular representations, can get complex when many tables are involved in a given query
- **XQuery:** Great for finding data in tree representations, can get complex when many relationships have to be traversed
- **SPARQL**: Good pattern matching paradigm, especially when relationships have to be used to answer a query

SQL, XQuery and SPARQL: What's Wrong with this Picture?

Jim Melton (Oracle; XML Query Working Group, XML Coord. Group)

Sixth annual W3C Technical Plenary (March 2006)

Standardization's view (Jim Melton, Oracle, 2006)

- **SQL:** Great for finding data from tabular representations, can get complex when many tables are involved in a given query
- **XQuery:** Great for finding data in tree representations, can get complex when many relationships have to be traversed
- **SPARQL**: Good pattern m… …adigm, especially when relationshi… …be used to answer a query

SPARQL = Sympathy Queen?

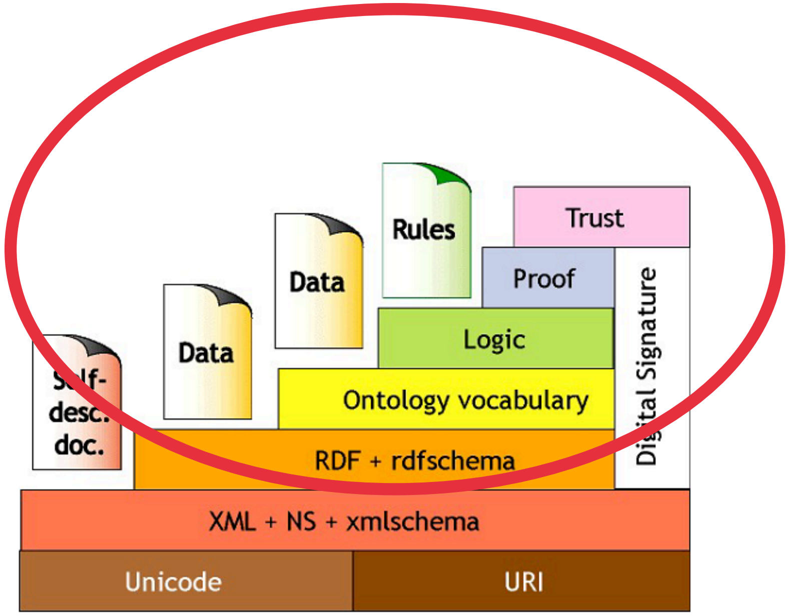
SQL, XQuery and SPARQL: What's Wrong with this Picture?

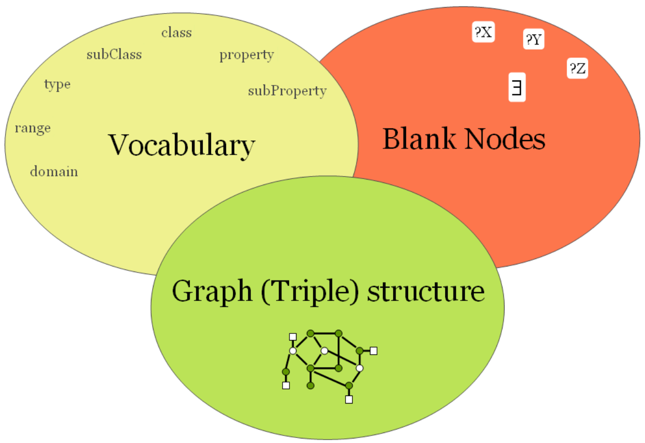
Jim Melton (Oracle; XML Query Working Group, XML Coord. Group)

Sixth annual W3C Technical Plenary (March 2006)

RDF Query Language: Logician's view

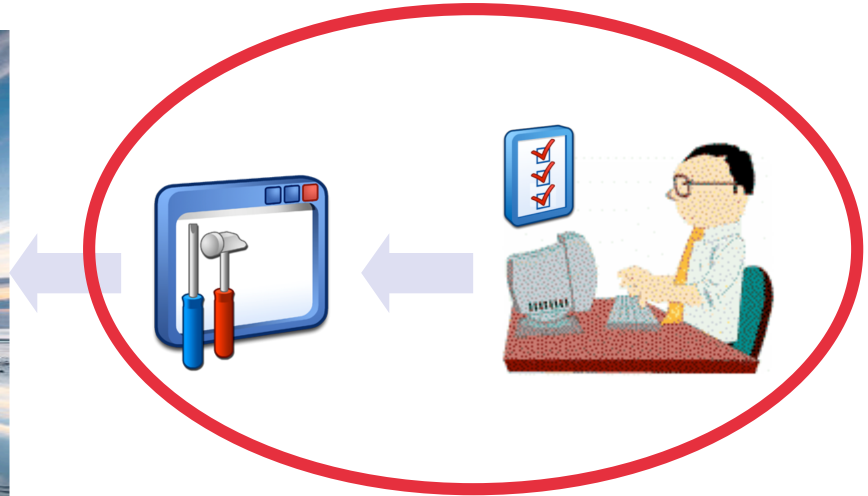
- RDF is the first level of a logical tower
- Emphasis in logic features of RDF model
- Keep an eye in extensions to more expressive logics
- Bad news: complexity issues





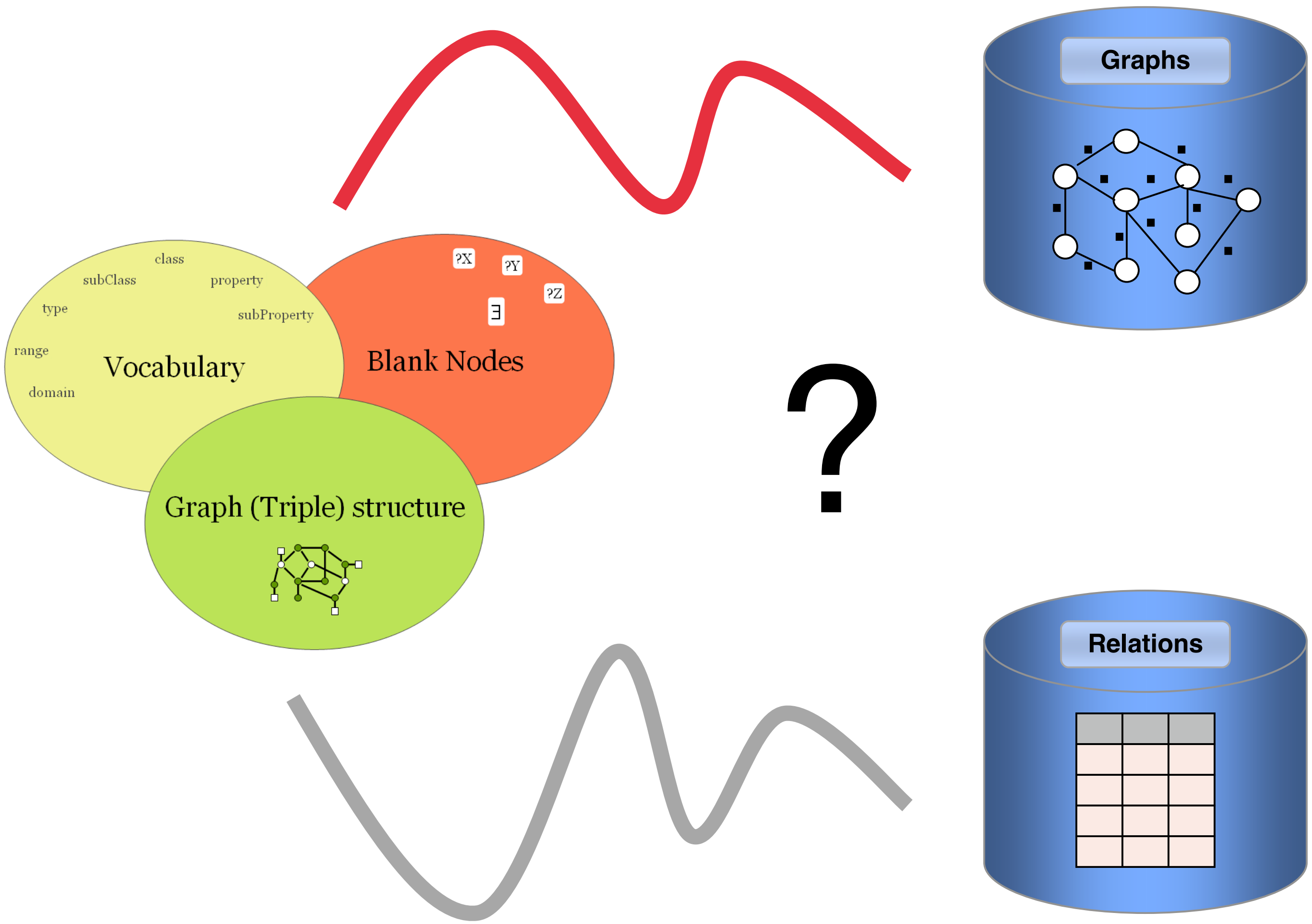
RDF Query Language: Developer's view

- How do we answer the most common queries?
- How do we cope with APIs and store developments?
- Design usually influenced by current programming and system tools.
- Not always concerned with scalability and long term.



RDF Query Language: Database theoretician's view

RDF as a graph data model?



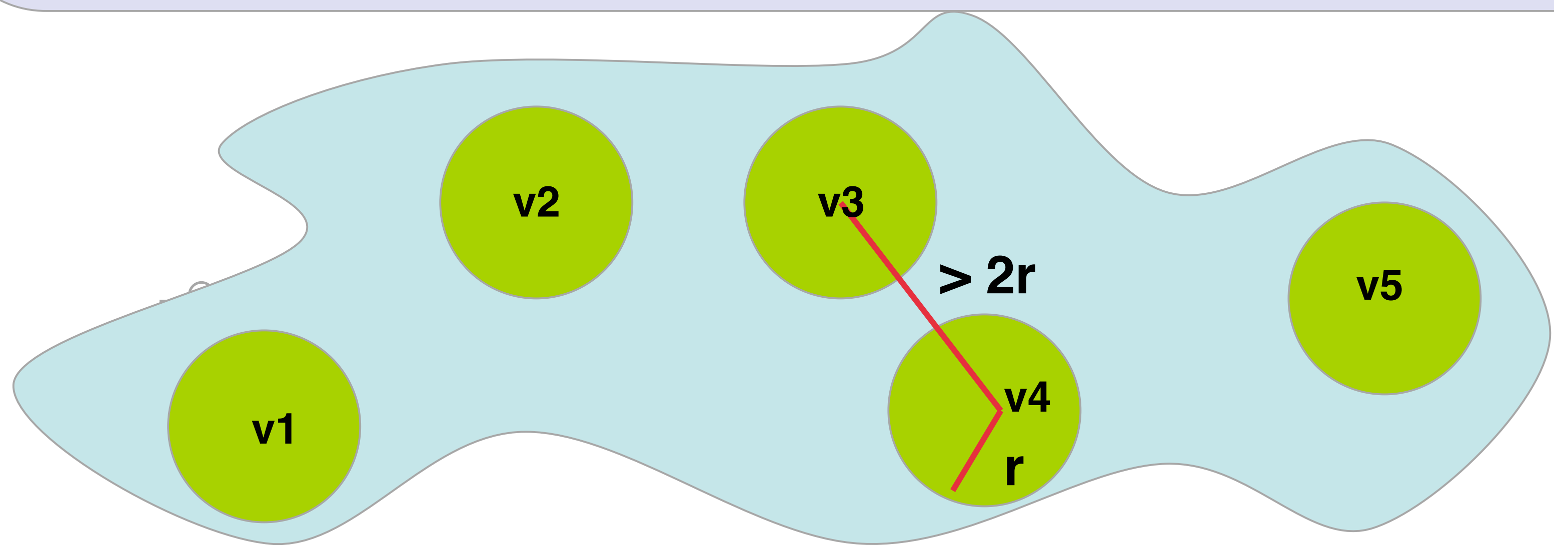
RDF as a relational model?

RDF Query Language: Database theoretician's view

Theorem (Gaifman). A property of graphs is expressible by a closed first order formula iff it is equivalent to a combination of properties of the form

$$\exists v_1, \ldots, v_s [\bigwedge_{1 \leq i \leq s} P(N(v_i, r)) \ \wedge \bigwedge_{1 \leq i < j \leq s} d(v_i, v_j) > 2r]$$

where $v_1, \ldots, v_s$ denote vertices and d(x,y) denotes distance



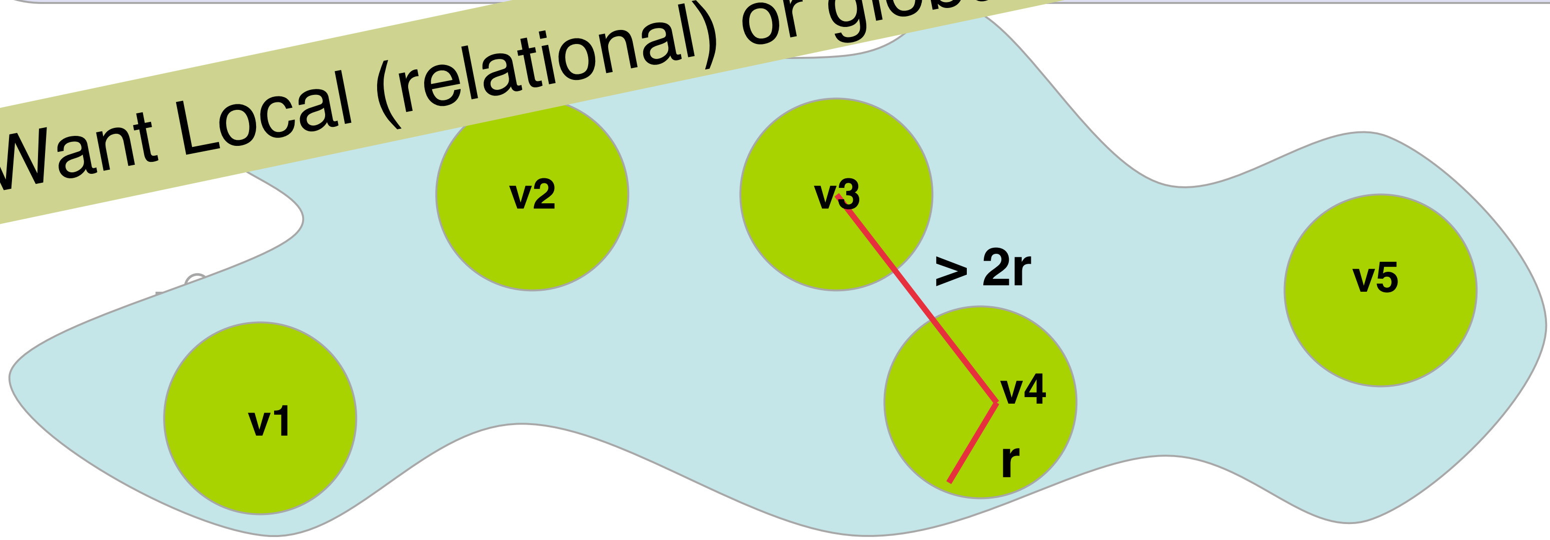
RDF Query Language: Database theoretician's view

Theorem (Gaifman). A property of graphs is expressible by a closed first order formula iff it is equivalent to a combination of properties of the form

$$\exists v_1, \ldots, v_s [\bigwedge_{1 \leq i \leq s} P(N(v_i, r)) \wedge \bigwedge_{1 \leq i < j \leq s} d(v_i, v_j) > 2r]$$

where $v_1, \ldots, v_s$ denote vertices and d(x,y) i...

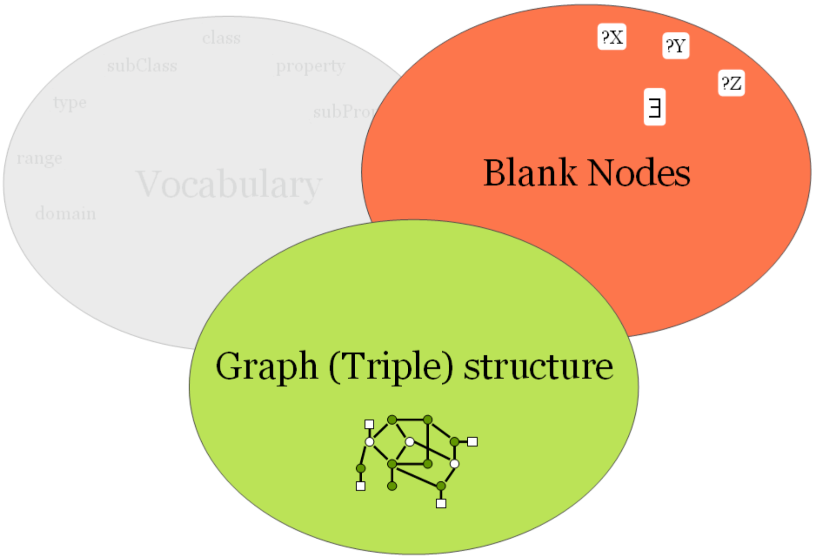
Want Local (relational) or global (graph) queries?



W3C Working Group's view

SPARQL (W3C Recommendation, 2008)

- Relational view of querying
- RDF = triples + blanks
- Pattern matching

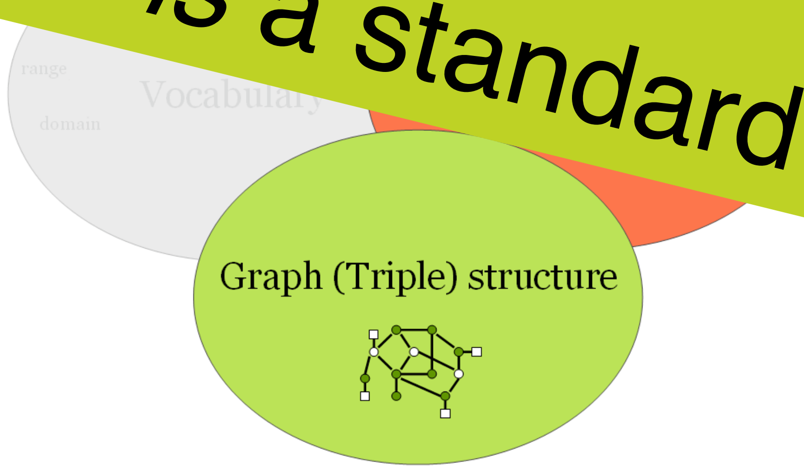


W3C Working Group's view

SPARQL (W3C Recommendation, 2008)

- Re... tional view of querying
- R... ...blanks
- F...

Good News:
there is a standard!



SPARQL Query (General Structure)