Pragmatic Semantics for the Web of Data

AImWD -- Montpellier 2013
Stefan Schlobach
(based on work of and using slides from Christophe Gueret, Kathrin Denthler and Wouter Beek)
VU Amsterdam

Postulates

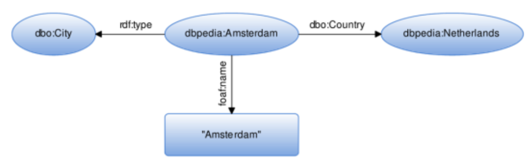
- The Web of Data requires semantics
- The Web of Data is not a database
- The Web of Data is a complex system
- Semantics for a database are not (always) suitable for complex systems
- We need new semantic paradigms
  - Voila: Pragmatic Semantics

Part1

CLASSICAL SEMANTICS FOR THE WEB OF DATA

Linked Data

- Graph/facts based knowledge representation
- Connect resources to properties / other resources
- Web-based: resources have a URI
  - Try http://dbpedia.org/resource/Amsterdam



Model theory for Semantic Web Languages: RDF, RDFS, OWL

- **Ontology and Data**: set of formulas S
- **Model**: formal structure satisfying all formulas in S
- Entailment: formula f entailed by S iff f in true in **all** models of S
- If contradiction, no models…
- No models, everything is entailed.

Part2

THE WEB OF DATA AS A COMPLEX SYSTEM

What is the problem?

- Frank and Christophe publish some open data
- Roi wants to combine and enrich it

Kennissen	Staad
Christophe	Amsterdam
Peter	Barcelona
David	Parijs

Ville	Pays
Barcelone	Espagne
Paris	France
Amsterdam	Pays-Bas

Marvel icons: mermer, DeviantArt

What is the problem?

Kennissen	Staad
Christophe	Amsterdam
Peter	Barcelona
David	Parijs

+

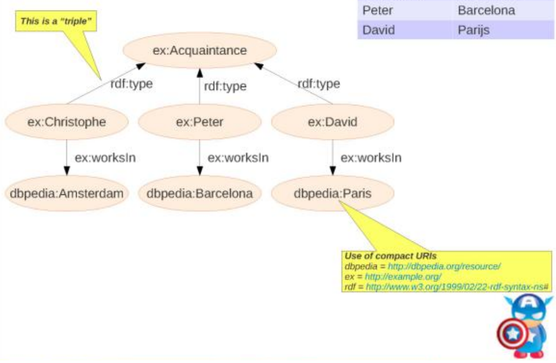
Ville	Pays
Barcelone	Espagne
Paris	France
Amsterdam	Pays-Bas

= ?

- Data integration issue
  - “Kennissen”, “Staad”, “Ville”, “Pays” ?
  - “Paris” = “Parijs” ?
  - “Amsterdam” = “Amsterdam” ?
- Lot of work, must be done again on updates

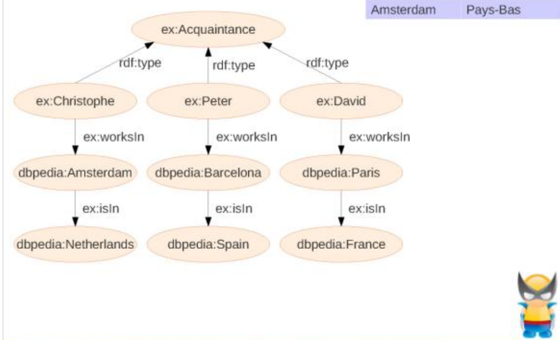
Frank publishes his data

Kennissen	Staad
Christophe	Amsterdam
Peter	Barcelona
David	Parijs

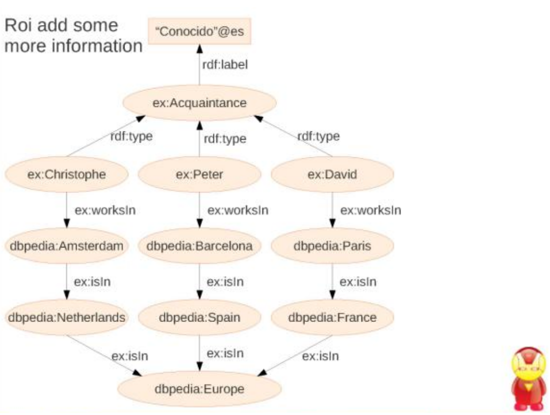


Christophe re-use part of Frank's data to publish his data

Ville	Pays
Barcelone	Espagne
Paris	France
Amsterdam	Pays-Bas



Roi add some more information



Since 2006, people are creating linked data

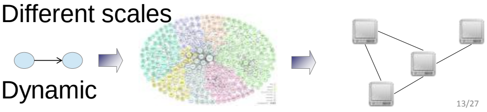
But publication and interpretation
are distributed processes.

The Web of Data is a Complex System.
Not a database.
It is a Marketplace of ideas.

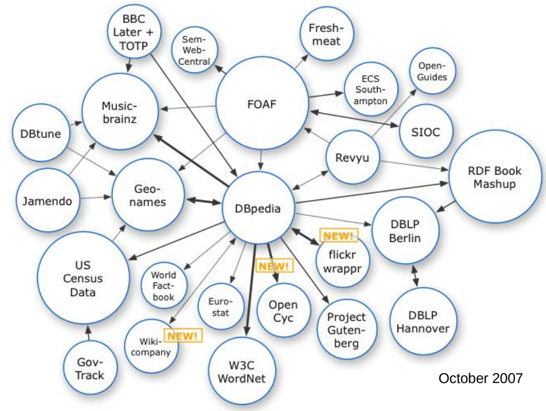
Key observations

- The Web of Data is more than the sum of its triples – it's a Complex System
- Different actors

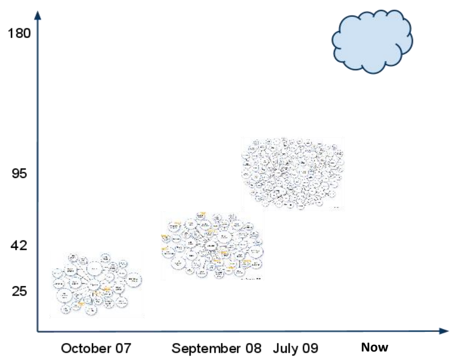
- Different scales
- Dynamic



13/27



Evolution of the Web of Data



The WoD is a complex system!

- Countless extremely heterogeneous datasets
  - general-purposed datasets, such as DBpedia
  - domain-oriented datasets, such as Bio2RDF
  - government data, music data, geological data, social network data, etc.

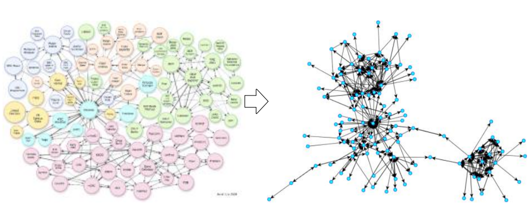
Hundrets of billions of RDF triples
  - Billions of links within the datasets
  - More than Million links between the datasets

Embedded rich semantics in the data
  - data points are typed
  - links are typed
  - **links is what makes the statements useful**

Information has impact on different scales

A new way of seeing the WoD

Consider the WoD as network

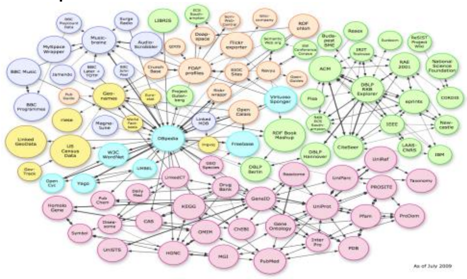


Relevant (Network) Properties of WoD

- Average path length
- Degree distribution
- Strongly connected components
- Degree centrality
- Between centrality
- Closeness centrality

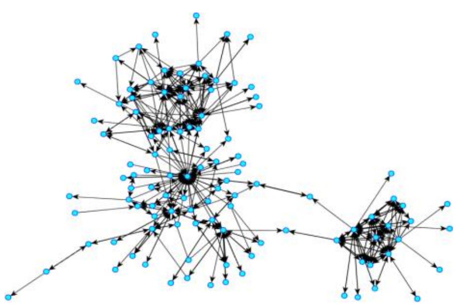
Scales of observation of the WoD

1. Graphs scale

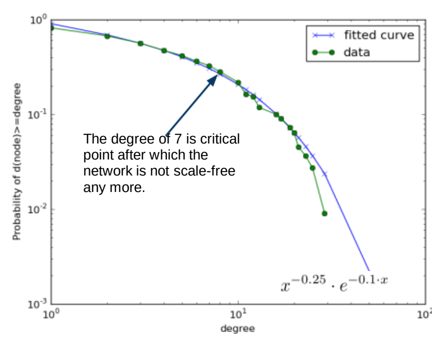


Graph-scale WoD network

- Each dataset is a node
- Edges are weighted, directed connections between the datasets
  - if there is at least one triple having a subject within dataset 1 and an object within dataset 2, then there is an edge between these two datasets.
  - the number of such triples is the weight of the edge.



- 110 nodes with 350 edges
- Average path length is 2.16
- 50 components



Top central nodes

Node	Value
DBpedia	0.332
DBLP Berlin	0.108
DBLP (RKB)	0.100
DBLP Hannover	0.097
FOAF profiles	0.075

Betweenness centrality

Node	Value
DBpedia	0.762
Geonames	0.614
Drug Bank	0.576
Linked MDB	0.544
Flickr wrappr	0.526

Closeness centrality

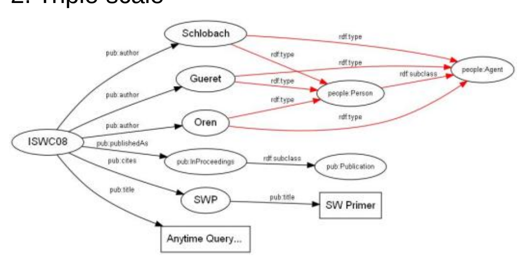
Node	Value
DBpedia	0.505
UniProt	0.266
DBLP (RKB)	0.266
ACM (RKB)	0.229
GeneID	0.211

Degree centrality

Every centrality has a specific meaning...

Scales of observation of the WoD

2. Triple scale



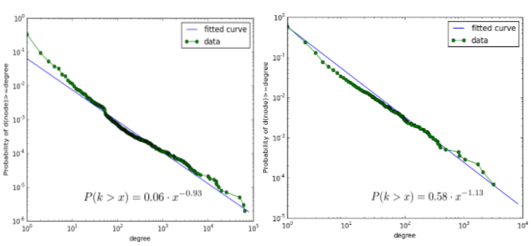
Triple-scale WoD network

- We took the 10 million triples from the dataset crawled from the WoD, provided by the billion triple challenge 2009
- This "BTC" network is defined as G=(V, (E, L)), where
  - V is a set of nodes, and each node is a URI or a literal
  - E is a set of edges
  - L is a set of labels, each label characterising a relation between nodes
- We applied a few strategies to aggregate data for comparison.

Triple-scale network and its aggregations
- BTC aggregated: triples are aggregated by the domain names
- BTC aggregated + filter: only domain names shared with the graph-scale network

Network	Nodes	Eges	Average path length	Components
BTC	605K	860K	2.15	602K
BTC aggregated	14K	31K	2.80	7K
BTC aggregated + filter	37	91	1.88	17

Power-law distribution

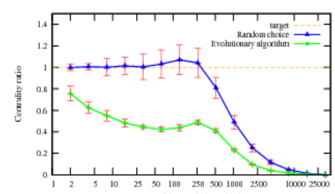


BTC BTC aggregated

Degree distribution

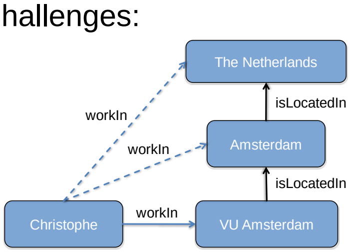
Monitoring and Improving the WoD

- Linked data is meant to be browsed, jumping from one resource to another
- The presence of Hubs is critical for the paths
- Create alternate paths to be used in case of failure



Guéret, Groth, van Harmelen, Schlobach, "*Finding the Achilles Heel of the Web of Data: using network analysis for link-recommendation*"

Challenges:



The links have explicit semantics, which brings implicit links deduced after the reasoning process

Challenges:

- Multi-relations links
  - FOAF (social networks + personal information)
  - SIOC (relations characterising blogs)
  - SWRC (describing research work)
  - …

  Different filtering produce different networks
  Centrality status of nodes changes *w.r.t* the networks

- Dynamics
  - Data will be continuously added and linked.

Part3

FORMAL INTERACTIONS WITH THE WEB OF DATA

Interacting with Linked Data

TABLE 2 Tasks and traditional solving methods to make use of a set of triples T. In the table, $\vdash$ stands for logical entailment and $t < Q$ implies that t is an instance of Q.

TASK	FORMAL PROBLEM DEFINITION	TRADITIONAL SOLVING METHODS
QUERYING	GIVEN T AND A QUERY Q, RETURN THE SET OF TRIPLES $\{t \in T\}$ SUCH THAT $T \vdash t < Q$	LOOKUP AND JOIN
STORAGE	GIVEN T AND A TRIPLE t RETURN $T \cup t$	CENTRALIZED INDICES, DISTRIBUTED HASH-TABLES
ENTAILMENT	GIVEN T. DERIVE $t \notin T$ WITH $T \vdash t$	CENTRALIZED AND PARALLELIZED DEDUCTION (RULES).
CONSISTENCY	GIVEN T. CHECK WHETHER $T \vdash \perp$ (FALSE)	LOGICAL REASONING
MAPPING	GIVEN T AND A MAPPING CONDITION C. RETURN $S,O \in T \times T$ SUCH THAT $C(S,O)$ LIKELY HOLDS WITH RESPECT TO T	SIMILARITIES SEARCH BETWEEN RESOURCES AND CLASSES. INDUCTIVE REASONING.

- Common semantic paradigm
- Common goals:
  - Completeness: all the answers
  - Soundness: only exact answers

When solutions do not (quite) fit the problem ...

Motivation

- In the context of Web data ?
  - Issues with scale
  - Issues with lack of consistency
  - Issues with contextualised views over the World
- Revise the goals
  - As many answers as possible (or needed)
  - Answers as accurate as possible (or needed)

From logic to optimisation

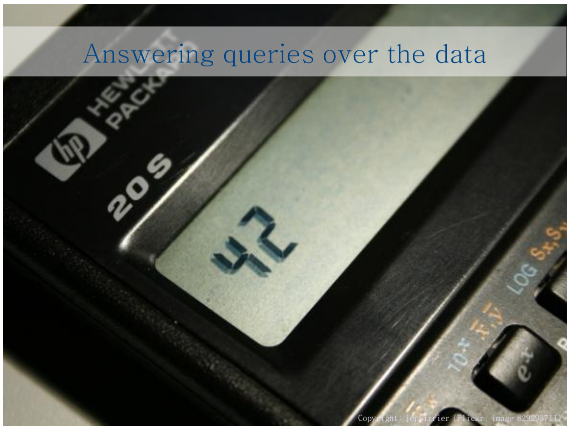
- Optimise towards the revised goals
- Need methods that cope with uncertainty, context, noise, scale, ...

TABLE 3 The interaction with the Semantic Web is made through different tasks which can be phrased either as logic or optimization problems.

TASK	LOGIC PROBLEM	OPTIMIZATION PROBLEM	RELATED WORK
QUERYING	CONSTRAINT SATISFACTION	CONSTRAINED OPTIMIZATION	ERDF [13]
STORAGE	CONSTRUCTION OF SETS	CLUSTERING	SWARMLINDA [20]
ENTAILMENT	LOGICAL DEDUCTION	MULTI-OBJECTIVE OPTIMIZATION	SWARMS [24]
CONSISTENCY	(UN)SATISFIABILITY CHECKING	CONSTRAINED OPTIMIZATION	-
MAPPING	LOGICAL DEDUCTION	CLASSIFICATION	PSO [36], GOSSIPING [31], EVOLUTIONARY STRATEGY [37]

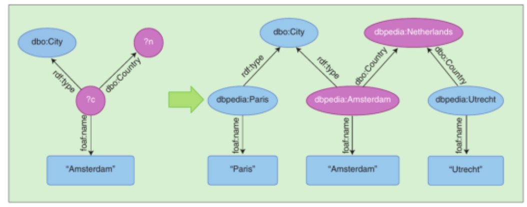
Nature inspired methods for interacting with complex systems

- Advantageous properties
  - Adaptation
  - Simplicity
  - Interactivity: Anytime, user in the loop
  - Scalability and robustness
  - Good for dealing with dynamic information
- Studied for different interaction types



The problem

- Match a graph pattern to the data
- Most common approach
  - Join partial results for each edge of the query

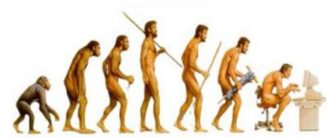


Solving approaches

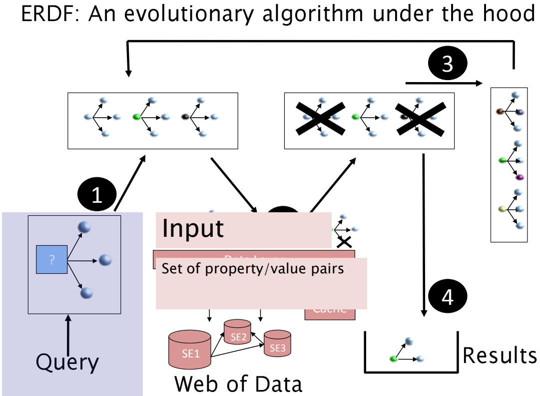
- Logic-based
  - Find all the answers matching all of the query pattern
- Optimisation
  - Find answers matching as much of the query as possible
- Important implications of the optimisation
  - Only some of the answers will be found
  - Some of the answers found will be partially true

Evolutionary Computing

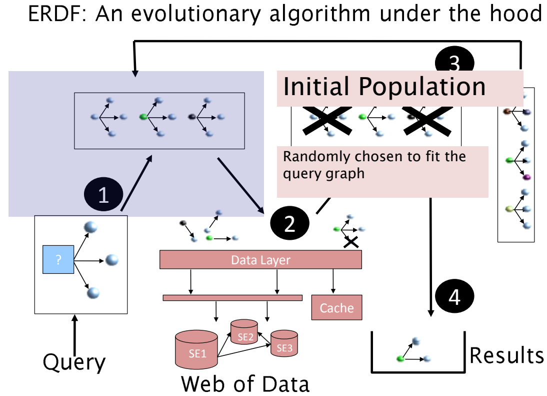
- Competition to survive in an environment with limited resources
- Inspired by theory of evolution (only best adapted can survive)



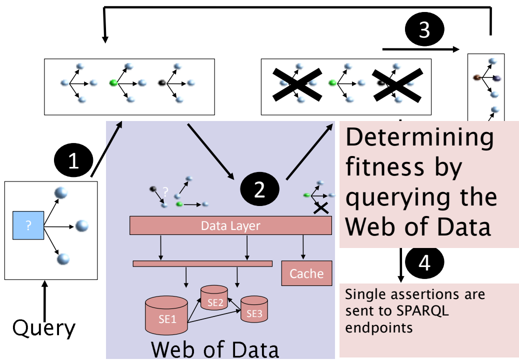

ERDF: An evolutionary algorithm under the hood



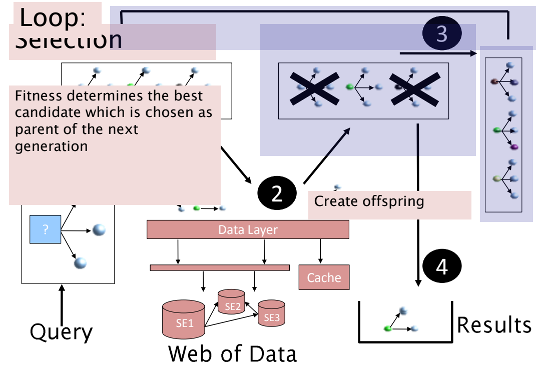
ERDF: An evolutionary algorithm under the hood



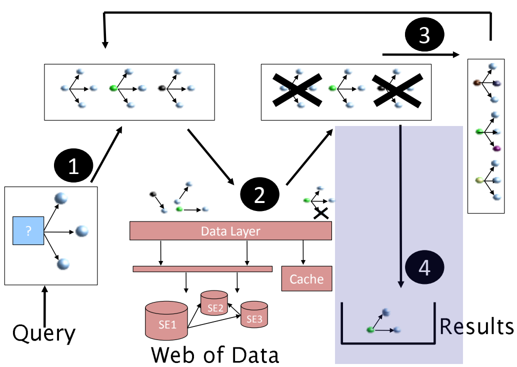
ERDF: An evolutionary algorithm under the hood



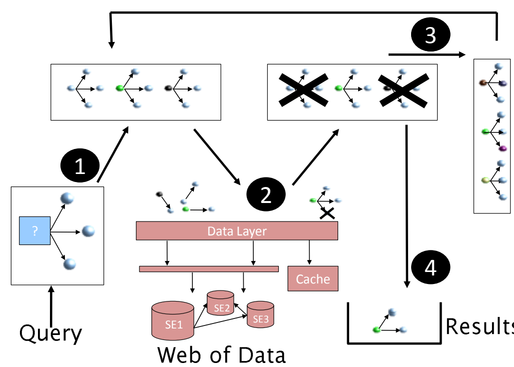
ERDF: An evolutionary algorithm under the hood



ERDF: An evolutionary algorithm under the hood



ERDF: An evolutionary algorithm under the hood



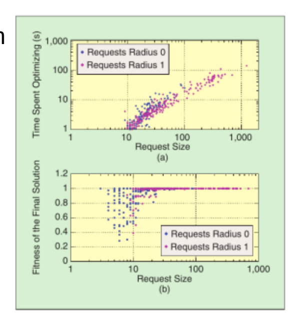
Properties of eRDF

- ✓ Scalable
- ✓ Lean
- ✓ Robust
- ✓ Anytime
- ✓ Approximate

- Arbitrary SPARQL endpoints
- Join-free, so scaling to more endpoints is comparably pain free

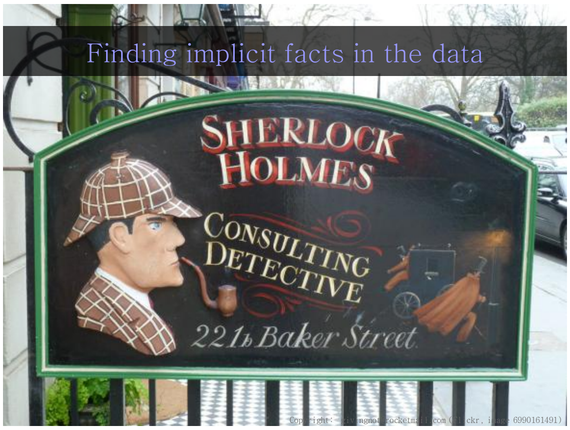
Some results

- Tested on queries with varied complexity
- Works best with more complex queries
- Find exact answers when there are some



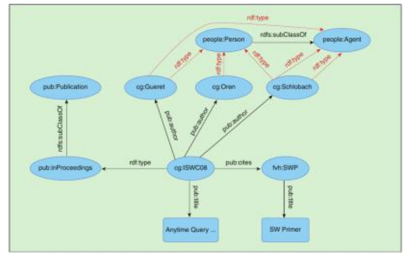
48/18

Finding implicit facts in the data



The problem

- Deduce new facts from others
- Most common approach
  - Centralise all the facts, batch process deductions



Solving approaches

- Logic-based
  - Find all the facts that can be derived from the data
- Optimisation
  - Find as many facts as possible while preserving consistency
- Important implications of the optimisation
  - Only some of the facts will be found
  - Unstable content

Collective Intelligence

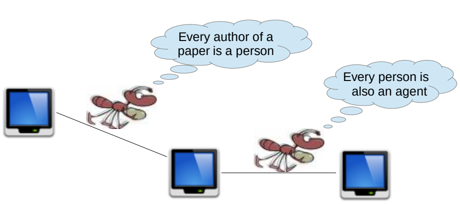
- Individuals showing intelligence when acting as a group. Notion of emerging behaviour.
- Swarms inspired by flocks of birds, social insects (ants,bees,...), schools of fish, ...



VU University Amsterdam

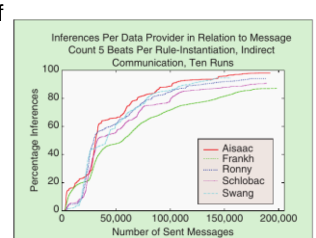
An optimisation approach: Swarms

- Swarm of micro-reasoners
  - Browse the graph, applying rules when possible
  - Deduced facts disappear after some time



Some results

- If they stay, most of the implicit facts are derived
- Ants need to follow each other to deal with precedence of rules
- Several ants per rule are needed



Related findings and approaches

- Storage optimisation using swarms (SwarmLinda from FU Berlin)
- Join optimisation with swarms (RCQ-ACS Erasmus Rotterdam)
- Emergent Semantics (eXascale Infolab Fribourg)
- Previous speaker (argumentation based semantics)

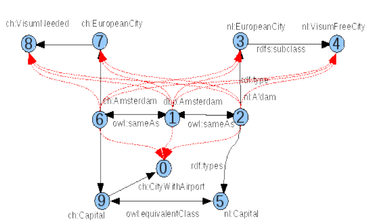
The day Semantics died…. ?

AImWD -- Montpellier 2013
Stefan Schlobach
(based on work of and using slides from Christophe Gueret, Kathrin Denthler and Wouter Beek)
VU Amsterdam

Part4

PRAGMATIC SEMANTICS FOR THE WEB OF DATA

There is meaning in the structure



Requirements

- Standard languages
- Standard semantics still valid (for simple data)
- Integrate structural properties
  - Popularity of nodes/triples
  - "Distance" between triples
  - Frequency of triples

Semantics not strict, but pragmatic

Intuitively: a statement twenty times made is more true than a statement once made

Approach

- Entailment defined through optimality over different (possibly competing) notions of truth
- Make as much information in the data explicit, and turn it into first-class semantics citizens (truth orderings)
- P*ragmatic entailment* is defined through multi-objective optimisation.
- Interoperability is then achieved by enriching an ontology with meta-information about semantic orderings, as well as agreement on the weighting of orderings.

Subset based truth orderings

– the size of the minimal entailing subontology
– ratio of sub-models in which a formula is satisfied versus the total number of sub-models
– ratio between sub-ontologies of O in which a formula holds holds versus the number of all sub-ontologies

Truth based on part of the given information

Graph-based truth orderings

- A shortest path ordering (diameter of the induced sub-graphs). Such a notion is a proxy for confidence of derivation. A
- A random-walk distance or edge-weights, induce orderings that are clustering-aware, with sub-ontologies entailing a formula have more cohesion than others.
- PageRank orderings can be used as proxies for popularity

Truth given on the structure of given information

Pragmatic Entailment

- A *pragmatic closure* C for an ontology O and orderings f1 to fn is then a set of formulas that is Pareto-optimal w.r.t. the optimisation problem max[f1 (C),…,fn (C)].

PraSem

- **Project title** : Pragmatic Semantics for the Web of Data
- **Acronym**: PraSem
- **Runtime:** Nov 2012-Oct 2016
- **Main researcher:** Wouter Beek
- **People involved:** Stefan Schlobach, Christophe Gueret, Kathrin Denthler, Pepijn Kroes, Frank van Harmelen, and hopefully more people soon.

Deal with Open World Assumption

Deal with incompleteness

Formalise approximations

Take home message

- The Web of Data requires semantics
- The Web of Data is not a database
- The Web of Data is a complex system
- Semantics for a database are not (always) suitable for complex systems
- We need new semantic paradigms
  – Voila: Pragmatic Semantics