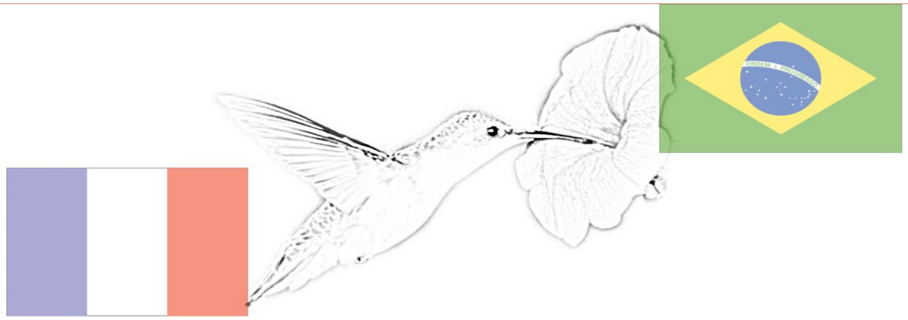


COLIBRI
Colóquio
Franco-Brasileiro

Using Semantics in Peer Data Management Systems

Carlos Eduardo Pires (cesp@cin.ufpe.br)
Damires Souza (damires@ifpb.edu.br)
Ana Carolina Salgado (acs@cin.ufpe.br)
Zoubida Kedad (zoubida.kedad@prism.uvsq.fr)
Mokrane Bouzeghoub (mokrane.bouzeghoub@prism.uvsq.fr)

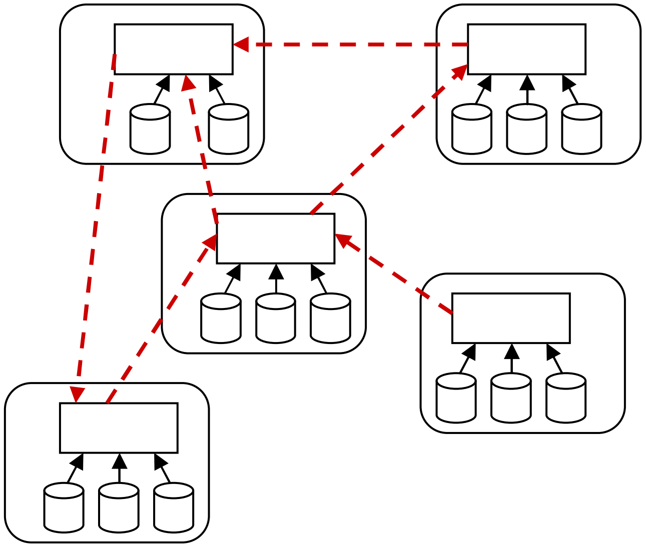
Centro de Informática U·F·P·E
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA PARAÍBA
UNIVERSITE DE VERSAILLES SAINT-QUENTIN-EN-YVELINES
Laboratoire PRISM

Outline

- Motivation
- SPEED Project
  - ✓ Peer Clustering
  - ✓ Query Reformulation
- Further Work
- Cooperation Status

Peer Data Management Systems (PDMS)

- Peers represent ***autonomous and heterogeneous*** data sources
- ***Sharing*** structured and semi-structured ***data***
- Data are represented through ***exported schemas***
- ***Lack of a unique global schema***
- Schema ***mappings***



Peer Data Management Systems (PDMS)

- A PDMS consist of a set of peers
  - ✓ ***Schema matching techniques*** are used to establish schema mappings: ***correspondences*** between schema elements
    - Schema mappings are defined between pairs of ***semantic neighbor peers***
  - ✓ ***Queries*** submitted at a peer are answered with data residing at that peer and with data that is reached through mappings over the semantic neighbors.

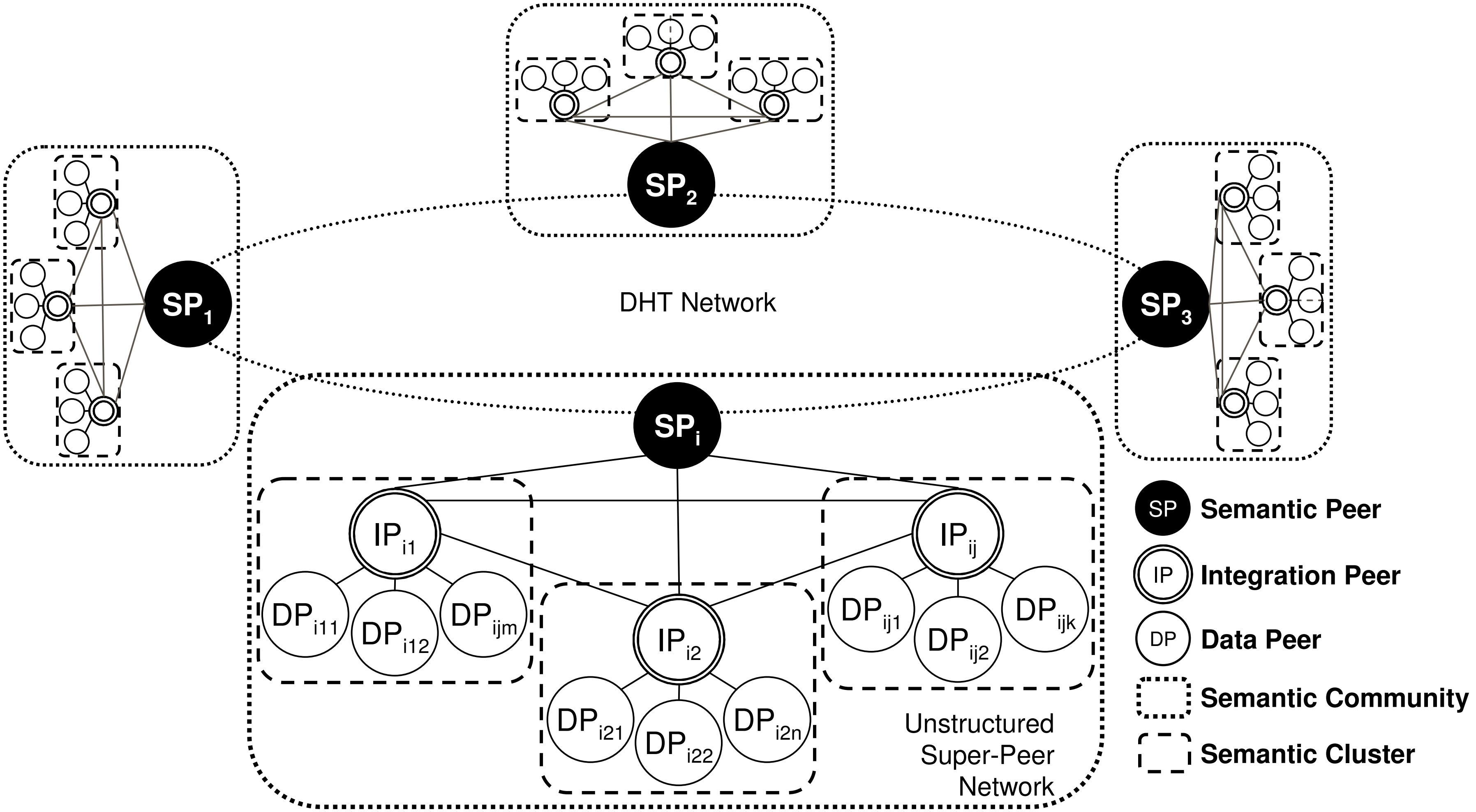
Data Management in PDMS

- ➢ A ***challenging problem***
  - ✓ Excessive number of peers, their autonomous nature, and the heterogeneity of their schemas
- ➢ ***Semantic knowledge*** in the form of ***ontologies*** has proven to be a helpful support
  - ✓ Ontologies can be used to represent the semantic ***content of data sources*** as well as ***to unify the semantic relationships*** between their schemas.

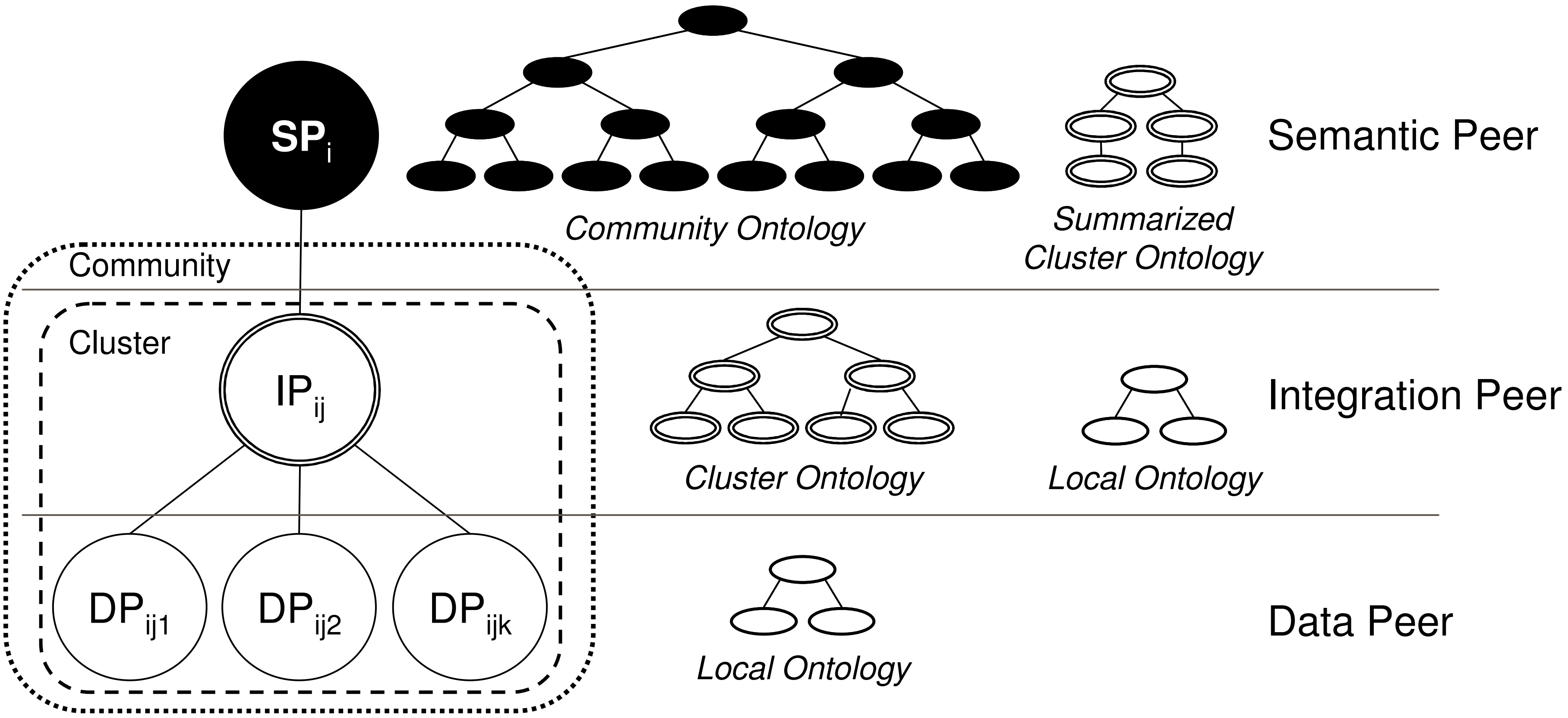
Goal of this Research Project

- To exploit the benefits provided by ***semantics*** through ***ontologies*** and ***contextual information*** to enhance data management issues in PDMS
- We propose ***semantic-based approaches*** to support:
  - ✓ Peer clustering
  - ✓ Schema summarization
  - ✓ Schema matching
  - ✓ Query reformulation

SPEED – An Ontology-based PDMS



Types of Ontologies

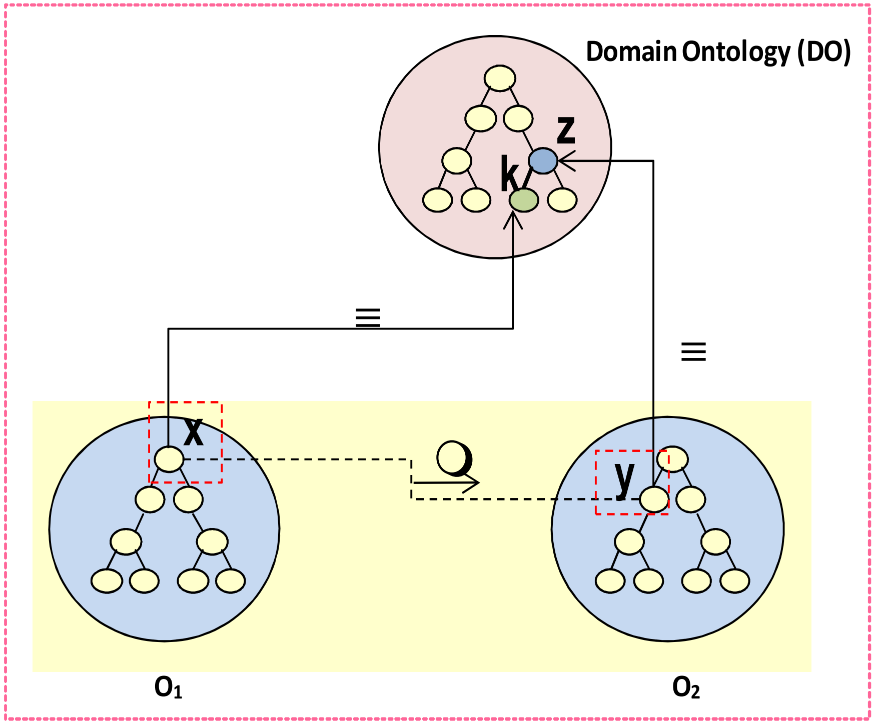


SemMatch – A Semantic Ontology Matcher

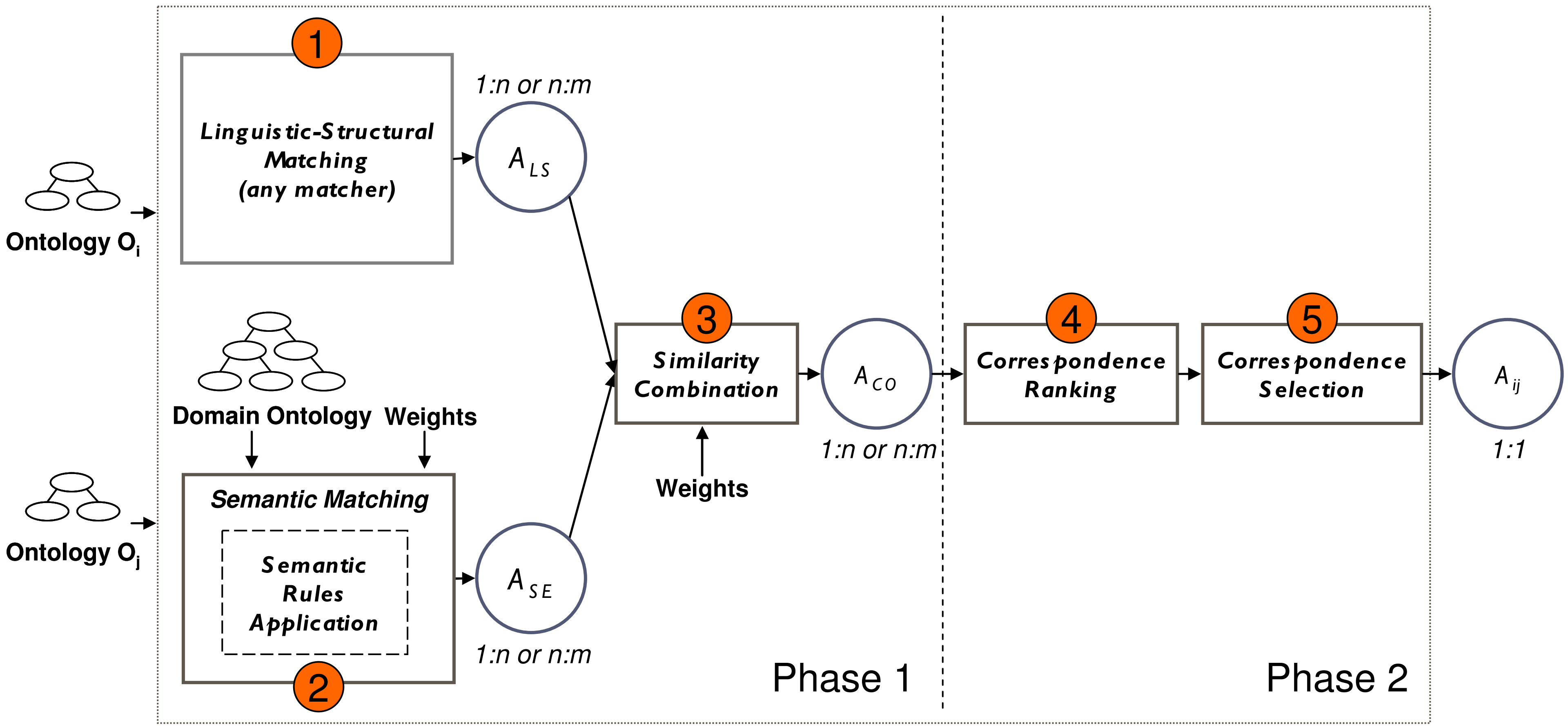
- ***Domain Ontologies – DO*** are used as background knowledge to identify seven types of ***semantic correspondences:***

- *isEquivalentTo* :$O_1{:}x \xrightarrow{\equiv} O_2{:}y$
- *isSubConceptOf* : $O_1{:}x \xrightarrow{\sqsubseteq} O_2{:}y$
- *isSuperConceptOf* : $O_1{:}x \xrightarrow{\sqsupseteq} O_2{:}y$
- *isPartOf :* $O_1{:}x \xrightarrow{\rhd} O_2{:}y$
- *isWholeOf* : $O_1{:}x \xrightarrow{\lhd} O_2{:}y$
- *isCloseTo:* $O_1{:}x \xrightarrow{\approx} O_2{:}y$
- *isDisjointWith* $O_1{:}x \xrightarrow{\perp} O_2{:}y$

where x and y are elements belonging to the ontologies O_1 and O_2.



SemMatch – A Semantic Ontology Matcher



Ontology O_i

Ontology O_j

1 Linguistic-Structural Matching (any matcher)

1:n or n:m

A_{LS}

Domain Ontology

Weights

2 Semantic Matching

Semantic Rules Application

A_{SE}

1:n or n:m

3 Similarity Combination

Weights

A_{CO}

1:n or n:m

4 Correspondence Ranking

5 Correspondence Selection

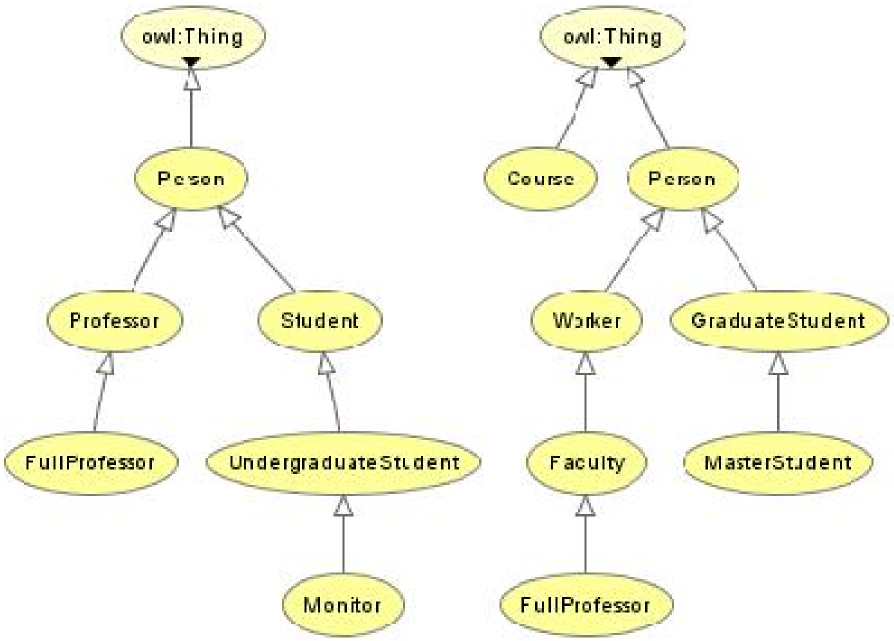
A_{ij}

1:1

Phase 1

Phase 2

Global Similarity Measure



Ontology O_i **Ontology O_j**

Alignment A_{ij}

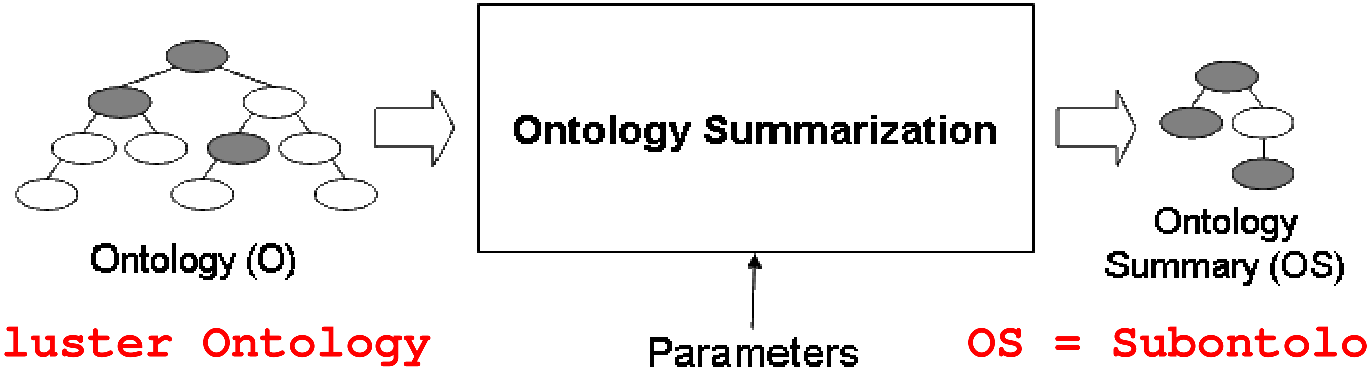
(1, Person, Person, isEquivalentTo, 1.0)
(2, FullProfessor, FullProfessor, isEquivalentTo, 1.0)
(3, UndergraduateStudent, Course, isPartOf, 0.3)
(4, Student, Person, isSubConceptOf, 0.8)
(5, Professor, Faculty, isSubConceptOf, 0.3)

Alignment A_{ji}

(1, Person, Person, isEquivalentTo, 1.0)
(2, FullProfessor, FullProfessor, isEquivalentTo, 1.0)
(3, Course, UndergraduateStudent, isWholeOf, 0.3)
(4, Worker, Person, isSubConceptOf, 0.8)
(5, GraduateStudent, UndergraduateStudent, isDisjointWith, 0.0)
(6, Faculty, Professor, isSuperConceptOf, 0.8)
(7, MasterStudent, Student, isSubConceptOf, 0,8)

$$Weighted\ \ Average(O_i, O_j) = \frac{(1.0+1.0+0.3+0.8+0.8)+(1.0+1.0+0.3+0.8+0.0+0.8+0.8)}{|6|+|7|} = 0.66$$

Ontology Summarization



Ontology (O)

Ontology Summarization

Parameters

Ontology Summary (OS)

`Cluster Ontology`

`OS = Subontology(O)`

- ➢ Main use in Peer Clustering
  - ✓ Resume cluster ontologies (**semantic index**)
- ➢ A summary does not represent a cluster ontology in its entirety
  - ✓ **Improve ontology matching**

Relevance Measures

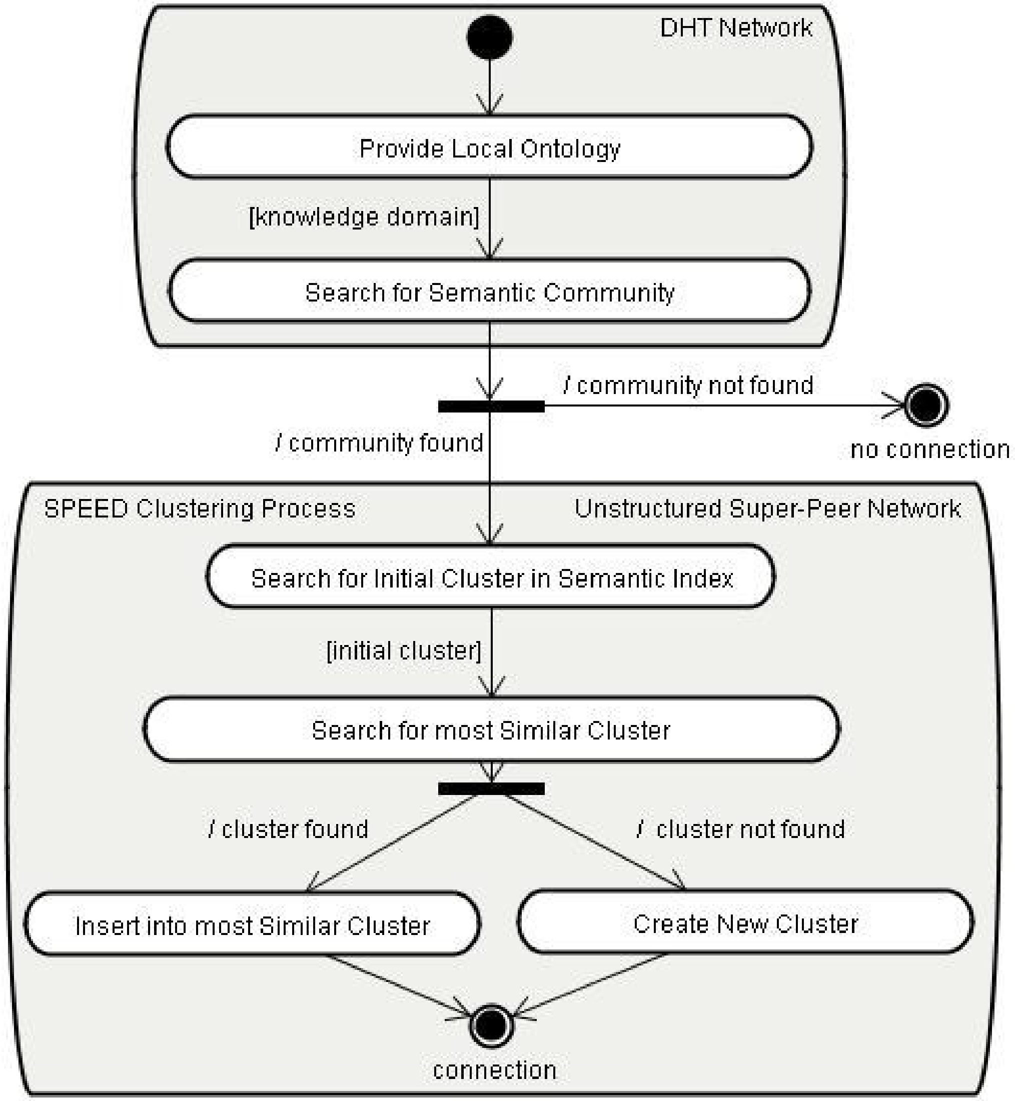
- ***Centrality:*** relationships (number and type) of a concept with other concepts in an ontology *O*

$$centrality(c_n) = \frac{nr \times \left(\frac{n_s \times w_s}{max_s} + \frac{n_{ud} \times w_{ud}}{max_{ud}} \right)}{|C| - 1}$$

- ***Frequency:*** occurrences of a concept in local ontologies $O_1,\ldots,O_n$ that compose *O*

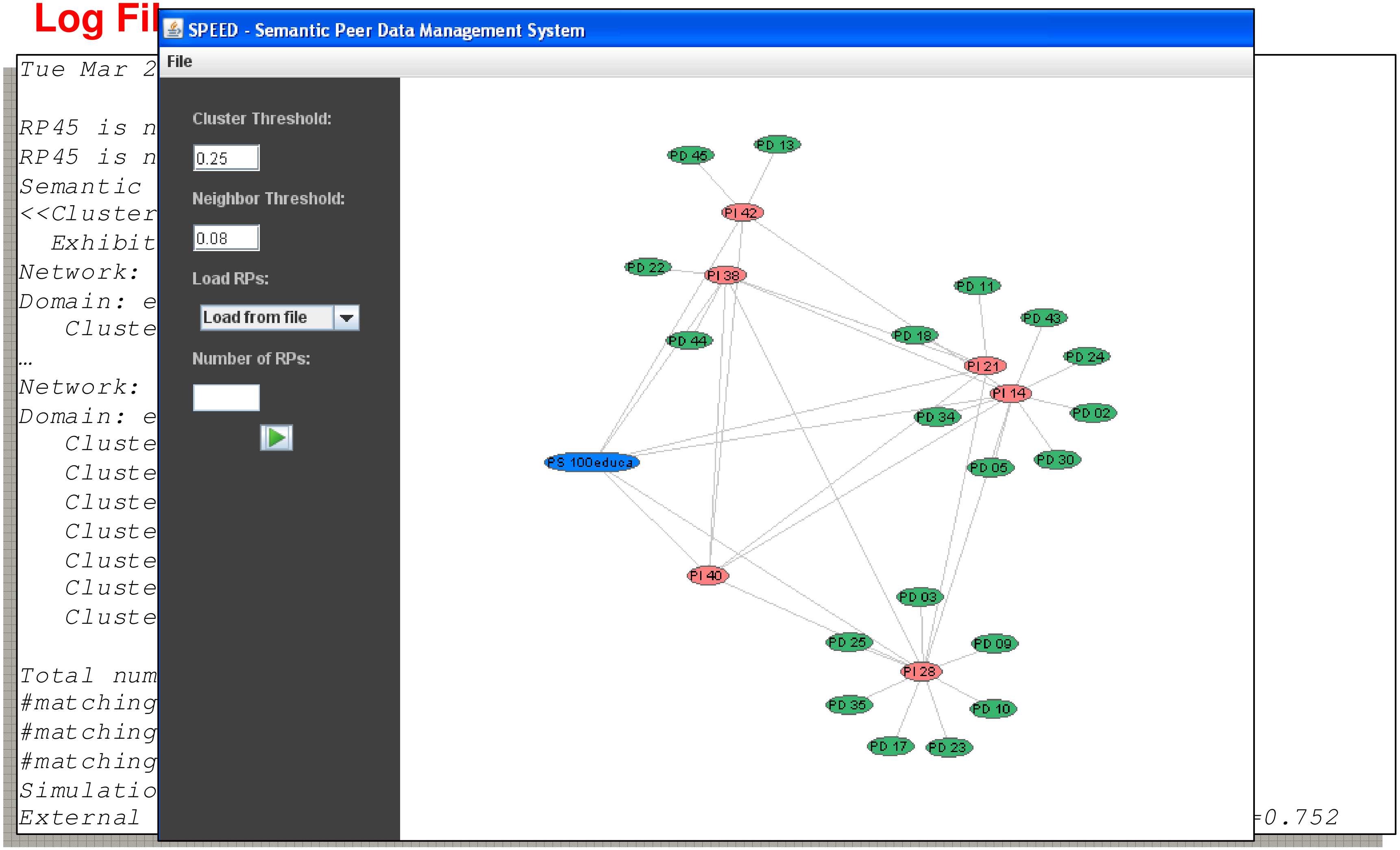
$$frequency(c_n) = \frac{|correspondences(c_n)|}{|O_1,\ldots,O_n|}$$

Ontology-based Peer Clustering



PDMS Simulator

Log Fil



```
Tue Mar 2

RP45 is n
RP45 is n
Semantic
<<Cluster
  Exhibit
Network:
Domain: e
   Cluste
…
Network:
Domain: e
   Cluste
   Cluste
   Cluste
   Cluste
   Cluste
   Cluste
   Cluste

Total num
#matching
#matching
#matching
Simulatio
External                                                               =0.752
```

SPEED - Semantic Peer Data Management System

File

Cluster Threshold: 0.25

Neighbor Threshold: 0.08

Load RPs: Load from file

Number of RPs:

Query Reformulation

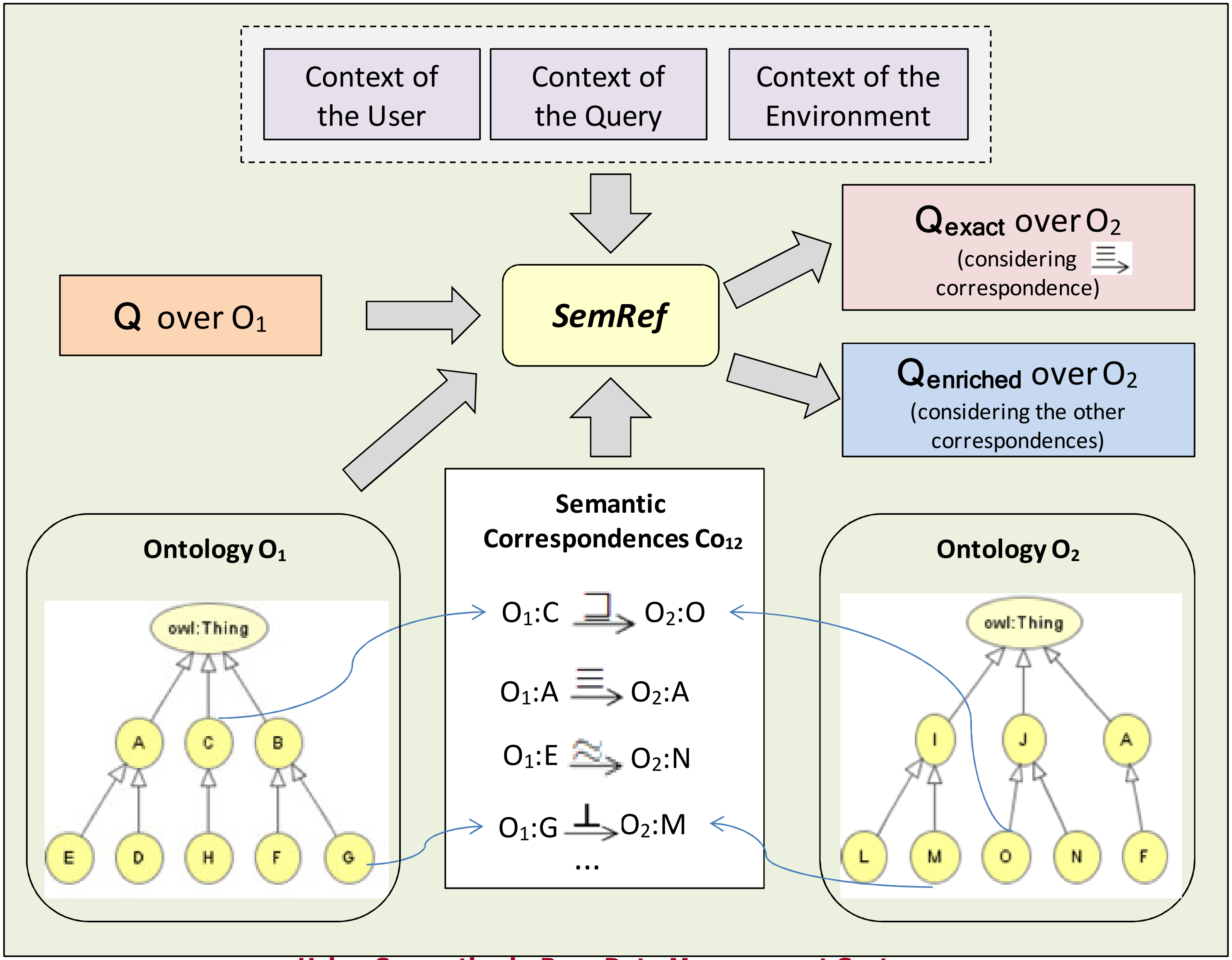
How to ***reformulate*** queries among the peers in such a way that the resulting ***set of answers*** expresses, as close as possible, what the ***users*** intended to obtain at query submission time, considering the ***dynamicity*** of the environment.

- Users’ preferences, query semantics and the current status of the environment are taken into account at query reformulation time: ***contextual information***
- The original query should be adapted to bridge the gap between the two sets of concepts: ***query enrichment***

The *SemRef* Approach - Using Context

- ***Users Context*** (preferences):
  - ✓ Exact reformulation is the default option
  - ✓ Enriching variables: ***Approximate***, ***Specialize***, ***Generalize***, *and* ***Compose****.*
- ***Query Context:*** Query semantics + Query reformulation mode
  - ***Restricted***: the priority is to produce an exact reformulation, although if it results empty, then an enriched reformulation may be provided
  - ***Expanded***: exact and enriched reformulations are to be produced.
- ***Environment Context:*** *path_length (*number of subsequent reformulations) + submission peer's identification and its neighbors context .

The *SemRef* Approach



Context of the User

Context of the Query

Context of the Environment

Q over O_1

SemRef

Q_{exact} over O_2 (considering $\xrightarrow{\equiv}$ correspondence)

$Q_{enriched}$ over O_2 (considering the other correspondences)

Ontology O_1

Semantic Correspondences Co_{12}

$O_1{:}C \xrightarrow{\sqsupseteq} O_2{:}O$

$O_1{:}A \xrightarrow{\equiv} O_2{:}A$

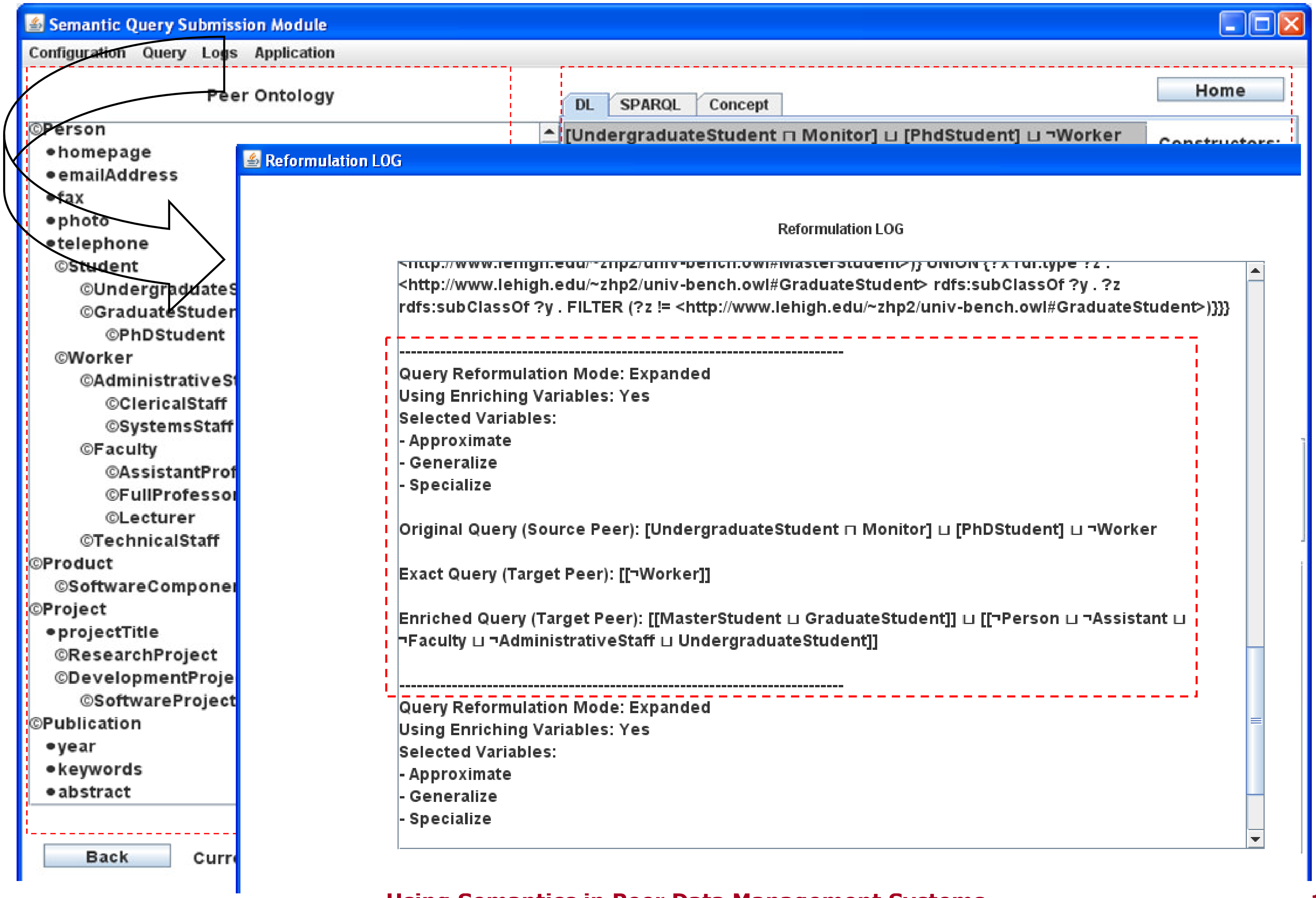
$O_1{:}E \xrightarrow{\approx} O_2{:}N$

$O_1{:}G \xrightarrow{\perp} O_2{:}M$

…

Ontology O_2

SemRef Module



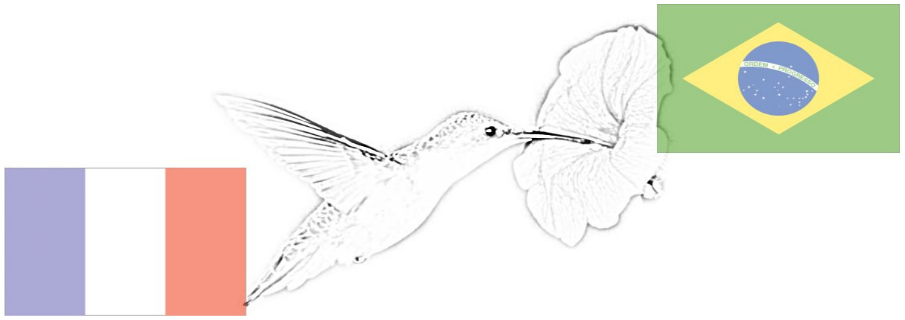
```
<http://www.lehigh.edu/~zhp2/univ-bench.owl#MasterStudent>}} UNION {?x rdf:type ?z .
<http://www.lehigh.edu/~zhp2/univ-bench.owl#GraduateStudent> rdfs:subClassOf ?y . ?z
rdfs:subClassOf ?y . FILTER (?z != <http://www.lehigh.edu/~zhp2/univ-bench.owl#GraduateStudent>)}}}

--------------------------------------------------------------------------------
Query Reformulation Mode: Expanded
Using Enriching Variables: Yes
Selected Variables:
- Approximate
- Generalize
- Specialize

Original Query (Source Peer): [UndergraduateStudent ⊓ Monitor] ⊔ [PhdStudent] ⊔ ¬Worker

Exact Query (Target Peer): [[¬Worker]]

Enriched Query (Target Peer): [[MasterStudent ⊔ GraduateStudent]] ⊔ [[¬Person ⊔ ¬Assistant ⊔
¬Faculty ⊔ ¬AdministrativeStaff ⊔ UndergraduateStudent]]

--------------------------------------------------------------------------------
Query Reformulation Mode: Expanded
Using Enriching Variables: Yes
Selected Variables:
- Approximate
- Generalize
- Specialize
```

Further Work

- Two relevant issues:
  - ✓ (i) ***the maintenance of semantic communities***
    - the evolution of cluster ontologies
  - ✓ (ii) ***query routing***
    - preserve the query semantics at the best possible level of approximation
    - enhance the selection of relevant semantic neighbors
    - personalize query results according to user’s profile
- Proposal of an Ontology Management Framework
  - ✓ Match, merge, translate and summarize

Cooperation Status

- CIn/UFPE and PRiSM/UVSQ
  - ✓ 90’s two PhD students
  - ✓ 2002 a PhD ‘sandwich’ and a scientific visit
  - ✓ Since then
    - Research visits
    - Cooperation project: STIC/Amsud (2008-2009)
      - France: Univ. de Versailles and Univ. Paul Cézanne (Aix-Marseille)
      - Brazil: UFPE and UFC
      - Uruguay: Universidad de la República
    - A sabatical year (2007-2008)
    - Another PhD ‘sandwich’ (2008)
    - Joint publications



COLIBRI
Colóquio
Franco-Brasileiro

Using Semantics in Peer Data Management Systems

Carlos Eduardo Pires (cesp@cin.ufpe.br)
Damires Souza (damires@ifpb.edu.br)
Ana Carolina Salgado (acs@cin.ufpe.br)
Zoubida Kedad (zoubida.kedad@prism.uvsq.fr)
Mokrane Bouzeghoub (mokrane.bouzeghoub@prism.uvsq.fr)

Centro de Informática U·F·P·E

INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PARAÍBA

UNIVERSITE DE VERSAILLES
SAINT-QUENTIN-EN-YVELINES

Laboratoire PRISM