



**Stefano. Spaccapietra@epfl.ch**

---

**Swiss Federal Institute of Technology Lausanne  
(EPFL)**



# Emergent Semantics



**IFIP 2.6 WG Databases**

**<http://wise.vub.ac.be/ifipwg26>**

**Stefano Spaccapietra (chair)**

# IT has changed

---

- From Information Management (20th Century)  
to Information **Exchange** (21st Century)
  - ◆ From DBMS to Web Data Services
- From Centralization  
to **Decentralization**
  - ◆ From central control to self-organization

# The New Communication Issue

---

- From getting the line
- To getting the message

bonjour



hello

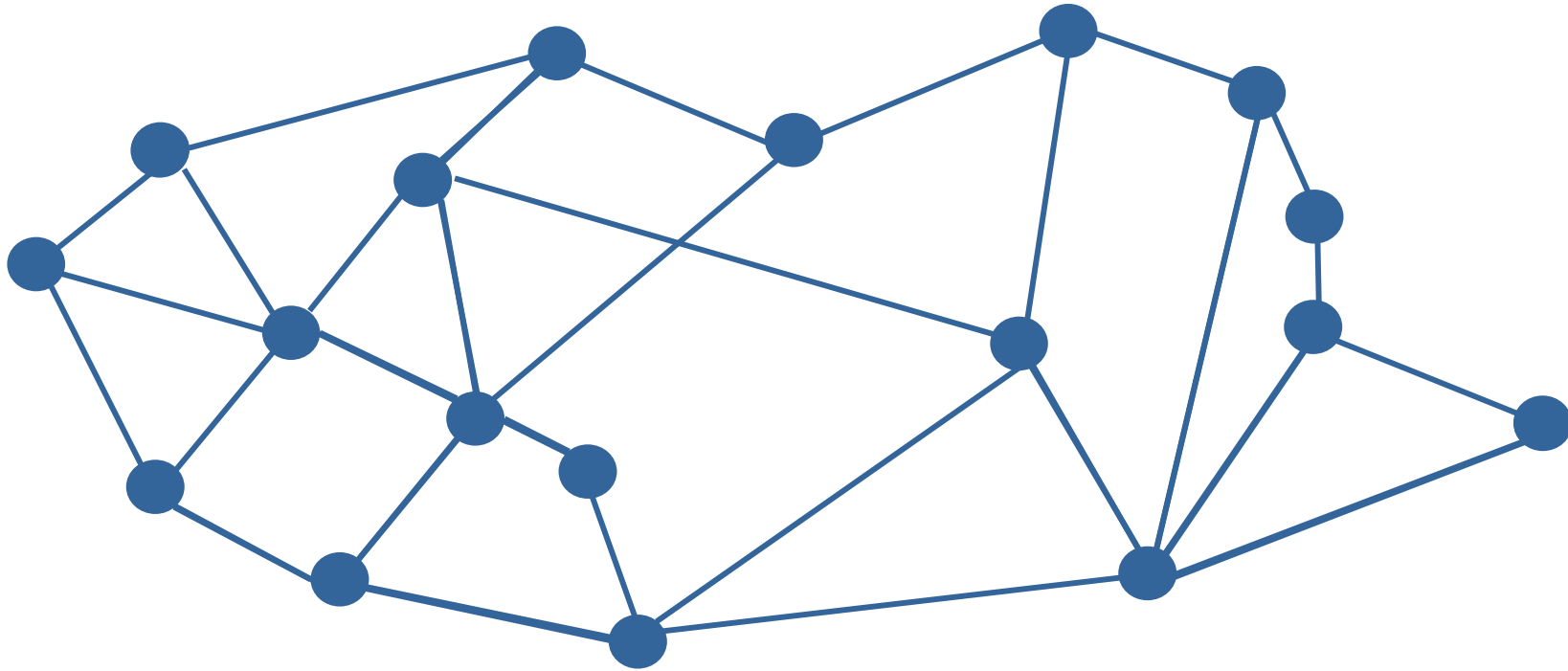


こんにちは



# New Challenge: Global Semantic Interoperability

---



How to obtain semantic interoperability among heterogeneous data sources **without** relying on pre-existing, global semantic repositories?

# New Approach: Emergent Semantics

---

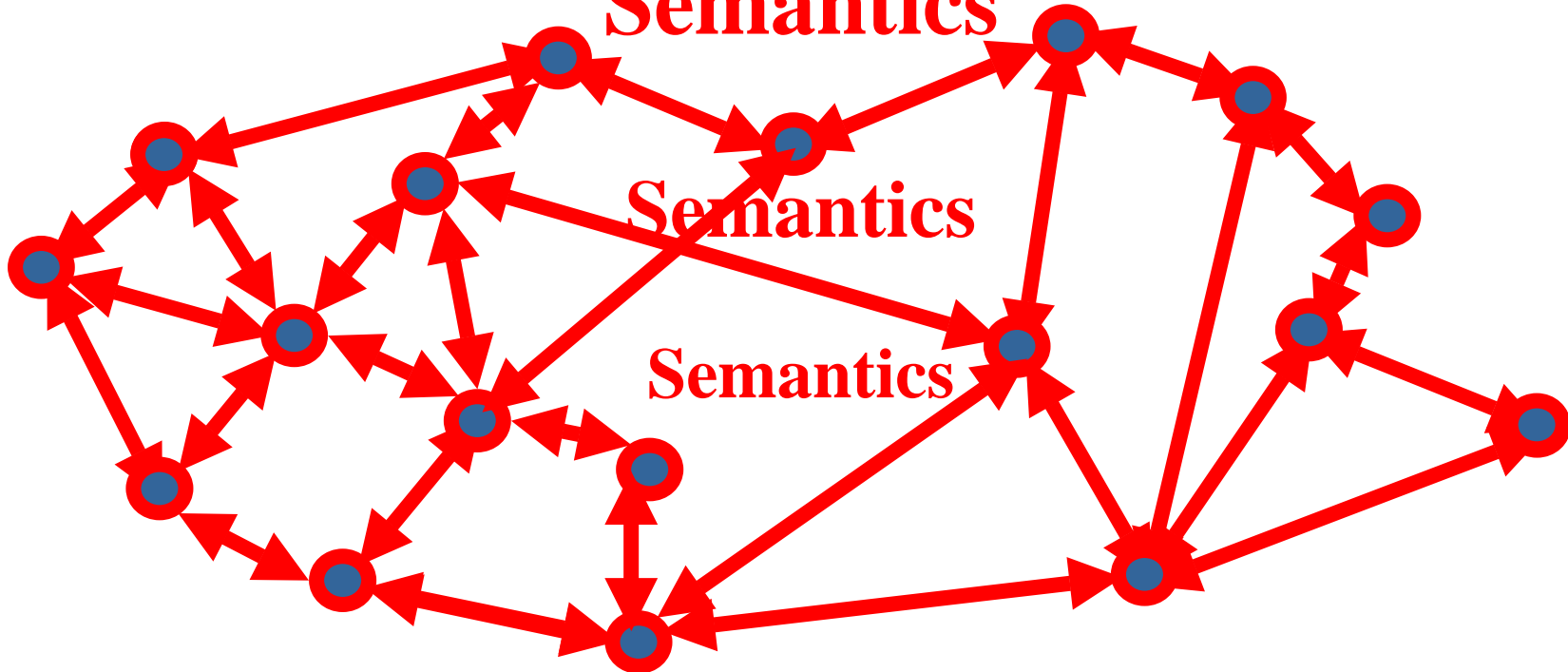
## Semantics

Semantics

Semantics

Semantics

Semantics



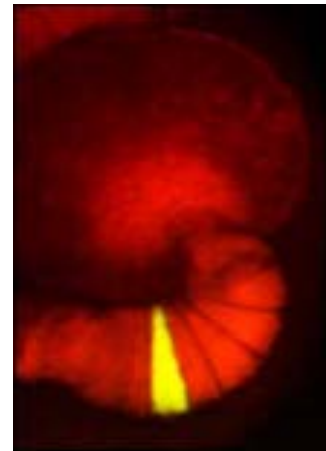
# Exchange = Syntactic & Semantic Agreement

- Shared Vocabulary
- Shared Reference Ontology



# Syntax (Shared Vocabulary) Isn't Enough

- Example: Searching Biological DBs
  - ◆ Without schema (like Google)
- Searching for data on "anglerfish"
  - Easy
  - Results will be precise
- Searching for "leech"
  - Organism leech
  - Authors: "Bleech", "Leechman", ...
  - Protein sequences: ...MNTS**LEECH**MPKGD...
- Searching for "257" ...



# Semantics: Interrelating Different Worlds

---

- An agent may be able to relate
  - ◆ <Author>  $\Leftrightarrow$  <Authors>
- A human non-expert (or an intelligent agent equipped with **ontologies**) may be able to relate
  - ◆ <Organism>  $\Leftrightarrow$  <Species>
- Only data owners or local experts (e.g., in SwissProt and EMBLChange) can relate
  - ◆ <AaMutType>  $\Leftrightarrow$  <DnaMutType>
  - ◆ <FtKey>  $\Leftrightarrow$  <FtKey>
    - => manual mapping (translation)

# Organizational Issues

---

When organizing the source nodes

- Can we build a minimal agreement to start with?
  - ◆ "Local" agreements (minimal semantic distance)
  - ◆ Manually => trustability
  - ◆ (Semi-)Automated => how?
- Can we cluster nodes into local communities (based on semantic homogeneity)?
  - ◆ Semantic distance measure
  - ◆ Dynamic, incremental clustering

# Example of Local Semantic Interoperability

```
Q1=  
<ID>$sp/ID</ID>  
FOR $sp IN /SP_entry  
WHERE "anglerfish" IN $sp/organism
```

```
Q2=  
<ID>$sp/ID</ID>  
FOR $sp IN T12  
WHERE "anglerfish" IN $sp/organism
```

**SwissProt  
(known schema)**

**EMBLChange  
(own schema)**

```
<SP_entry>  
<ID>CBPH_LOPAM</ID>  
<Authors>Roth</Authors>  
<Organism>  
  Lophius americanus  
  (American goosefish)  
  (Anglerfish).  
</Organism>  
<Sequence>  
  MKQICSIVLL ...  
</Sequence>  
</SP_entry>
```

**T12** =  
<SP\_entry>  
<ID>\$ec/ID</ID>  
<Organism>  
 \$ec/Species  
</Organism>  
</SP\_Entry>  
FOR \$ec IN /EC\_entry

```
<EC_entry>  
<ID>  
  LAJAFGI_5; VRT  
</ID>  
<Species>  
  Lophius americanus  
  (anglerfish)  
</Species>  
<SQ>  
  Sequence 1 BP  
</SQ>  
</EC_entry>
```

(Computer-processable languages: XML, XQuery)

# Operational Issues

---

When querying:

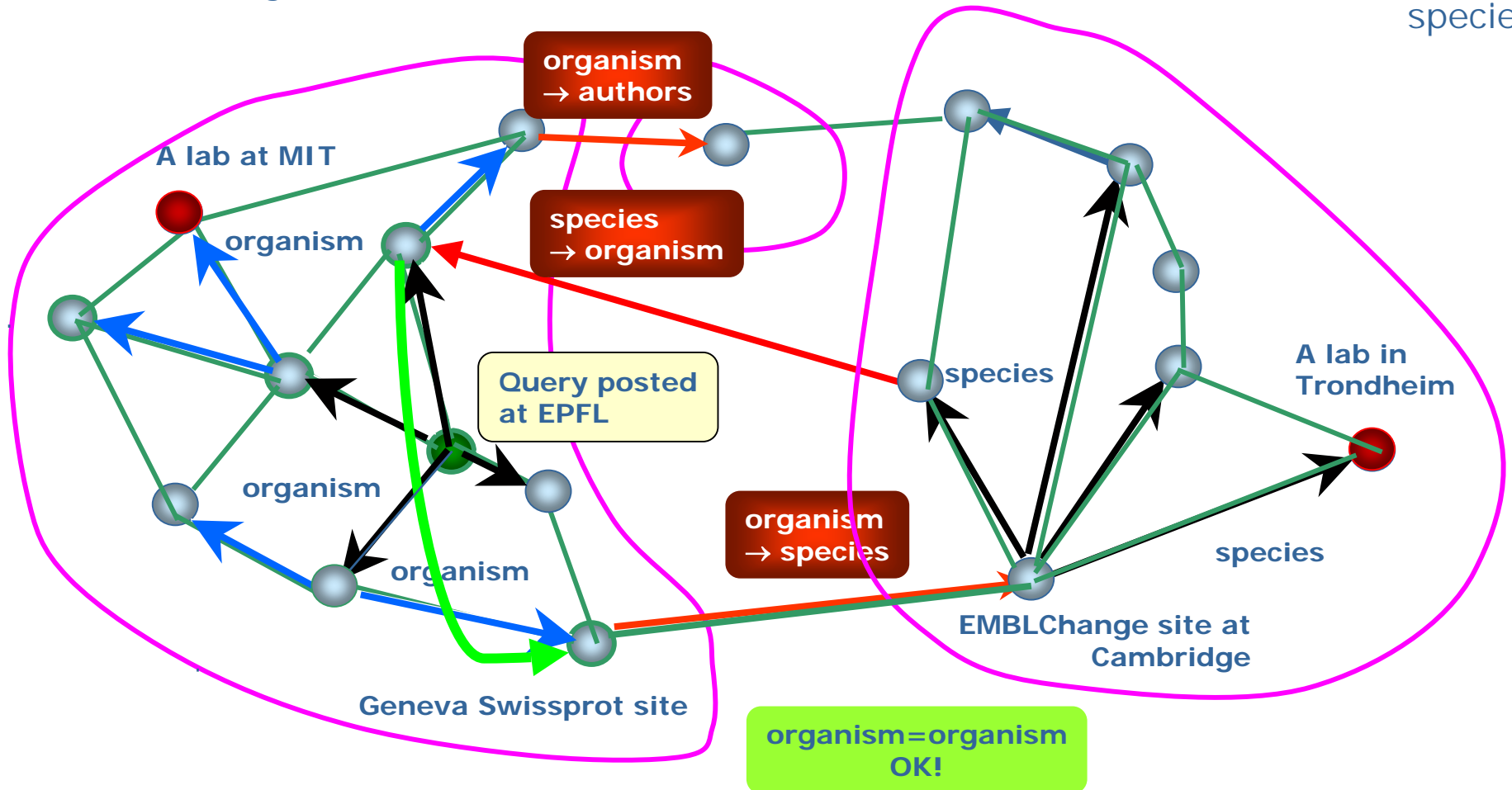
- Optimize the navigation strategy
  - ◆ Send query to nodes in the local community
    - => Low recall
  - ◆ Send query to the entire network (Query Flooding)
    - => Low precision, high network load
  - ◆ Send query to nodes most likely to be able to provide a result
    - => Classification strategy?
- Check query propagation
  - ◆ Is the query still relevant?
- Use cycles to check semantic agreement
- Learn from results to consolidate the semantic network

# Detecting Semantic Agreement

**SwissProt** peers  
authors, titles, organism, ...

other peers  
authors, ...

**EMBLChange** peers  
species, ...



Check what is preserved in cycles !

# Similarity Measures

---

## ■ Syntactic Similarity

- ◆ Similarity measure between an original and a transformed query.
- ◆ Iterative computation of information loss in selections / projections.

## ■ Semantic Similarities

- ◆ Use feedback to analyze the likelihood of the correctness of translations

# Semantic Similarity

---

- **Cycles Detection**

- ◆ Monitor query propagation to detect query cycles

- **Results Analysis**

- ◆ Variety of techniques
- ◆ E.g., use annotations (content-based retrieval techniques):
  - check if the annotation of returned documents produces a result similar to the annotation of documents known to correspond to the query

- => **Self-repairing semantic networks**

# Research Questions

---

- Many fundamental problems
  - ◆ Erroneous agreements
  - ◆ Agreement on schema but not on data
  - ◆ Complex data types and mappings
  - ◆ Overlapping data collections
  
- Algorithms and tools
  - ◆ To automatically generate, detect and use local translations
  - ◆ To identify which translations are correct (with a high probability)
  - ◆ To control the global search (via semantic gossiping)

# Complementary Interoperability Issues

---

- Information Acquisition
- Information Extraction
- Information Dissemination
- Information Monitoring

# Conclusion

---

- The "Semantics" problem is well known
- A general centralized solution is only a dream
- It could work in a limited context
- A decentralized approach seems more appropriate for unbounded contexts
- Let's hope it will work ...

**Thanks for your attention**