Self-Organizing Ontologies

Karl Aberer
EPFL
Computer & Communication Systems
Institute for Core Computing Science
Distributed Information Systems Laboratory

karl.aberer@epfl.ch
lsirwww.epfl.ch

October 29, 2002
ODBASE 2002
Panel Statement
Why are Schemas Important?

• Example: Searching biological databases
  – Without schema (like Google, Gnutella)

• Searching for data on "anglerfish"
  – Results will be precise

• This seems easy, but the same for "leech"
  – Organism leech
  – Authors: "Bleech", "Leechman", ...
  – Protein sequences: ...MNTSLEECHMPKGD...

• Search for "257" ...

©2002, Karl Aberer, EPFL-I&C-IIF,
Laboratoire de systèmes d’informations répartis
Schema Heterogeneity

- Different databases – Different schemas
  - SwissProt: Find `<Species> leech </Species>`
  - EMBLChange: Find `<Organism> leech </Organism>`

- Standardization (global ontology) ?
  - Music files: clear scope, simple semantics
  - Scientific databases: different scope, distributed knowledge, little agreement, etc.
Translating Heterogeneous Schemas

- A non-expert may be able to relate
  - `<Organism> <-> <Species>`
  - `<Author> <-> <Authors>` etc.

- But what about
  - `<AaMutType> <-> <DnaMutType>`
  - `<FtKey> <-> <FtKey>`

  in Swisschange and EMBLChange?

- The answers can only be given by the experts ...
  ... sometimes only by the data owners!

**Approach:** ask them to provide their translations from some "known" schema to their "own" schema (local step)
Local Semantic Interoperability (Translation)

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)

<SP_entry>
  <ID>CBPH_L001</ID>
  <Authors>Roth</Authors>
  <Organism>
    Lophius americanus
    (American goosefish)
    (Anglerfish).
  </Organism>
  <Sequence>
    MKQICSIVLL ...
  </Sequence>
</SP_entry>

EMBLChange (own schema)

<T12>
  <SP_entry>
    <ID>$ec/ID</ID>
    <Organism>$ec/Species</Organism>
  </SP_entry>
</T12>

Computer-processable languages: XML, XQuery
Global Semantic Interoperability

SwissProt peers
authors, titles, organism, ...

other peers
authors, ...

EMBLChange peers
species, ...

A lab at MIT
organism
organism
organism
Swissprot site at Geneva

organism
organism
organism

Query posted at EPFL

species

A lab in Trondheim
EMBLChange site at Cambridge

Semantic Gossiping
How to Detect a Semantic Agreement?

SwissProt peers: authors, titles, organism, ...

other peers: authors, ...

EMBLChange peers: species, ...

Check what is preserved in cycles (semantic kernels)!
Research Questions

- Many fundamental problems
  - Complex data types and mappings
  - Partial agreements
    (e.g. agreement on schema but not on data)
  - Erroneous agreements

- Approach: algorithms and tools
  - to automatically generate, detect and use local semantic relationships (such as translations)
  - identify which are correct with a high probability
    (via semantic kernels)
  - control of global search (via semantic gossiping)
Conclusion

- View semantics as an emergent property of a network of semantic relationships

- CS focus on developing tools supporting the uncovering of global agreement (emergent ontologies)

- These tools are concerned with syntax and properties of large networks

- and therefore are application-independent