
# BACKGROUND KNOWLEDGE IN ONTOLOGY MATCHING

**Pavel Shvaiko**

joint work with
Fausto Giunchiglia and Mikalai Yatskevich

INFINT 2007

Bertinoro Workshop on Information Integration

October 1, Italy



# Outline

- Introduction
- Lack of background knowledge
- Conclusions and future directions

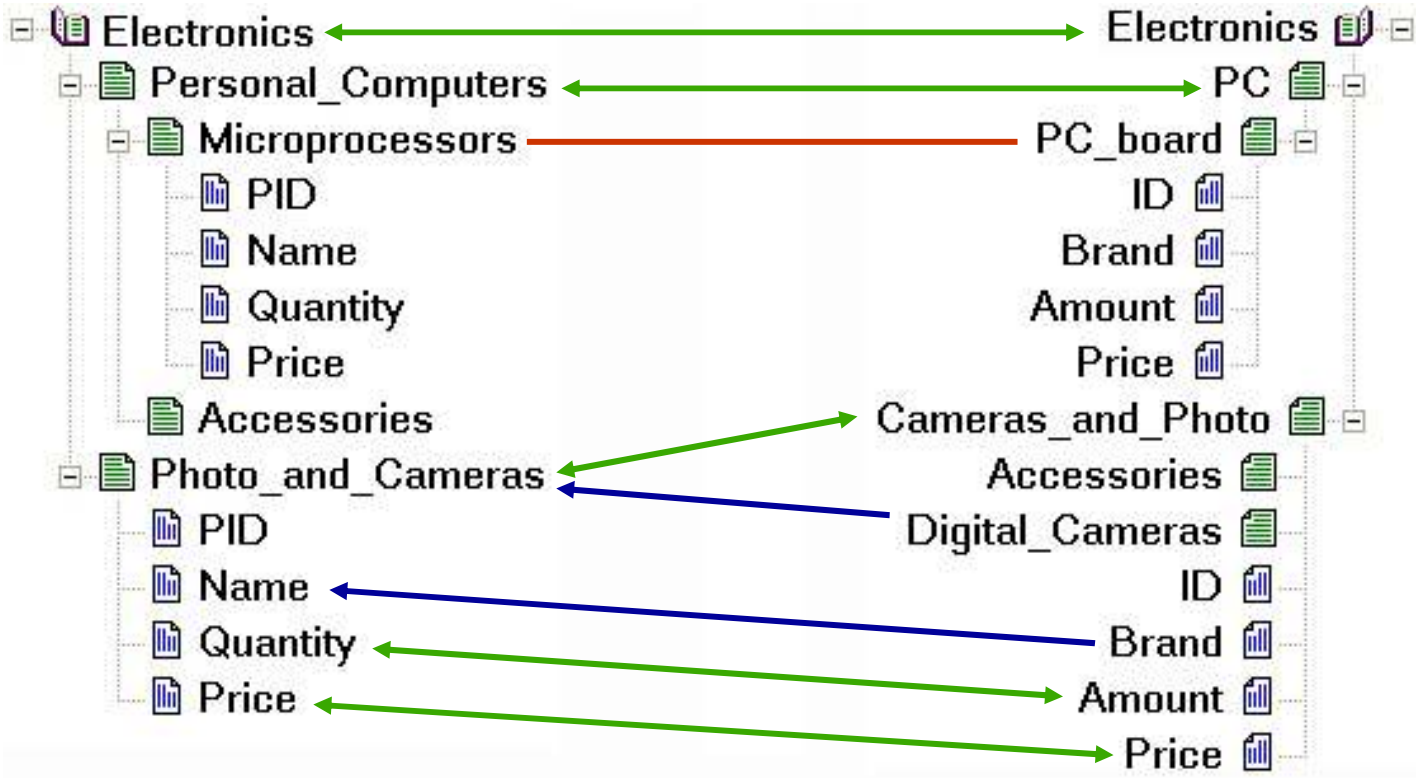


# Matching operation

**Matching** operation takes as input ontologies, each consisting of a set of discrete entities (e.g., tables, XML elements, classes, properties) and determines as output the correspondences (e.g., equivalence, subsumption) holding between these entities



# Example: two XML schemas




Equivalence  Generality  Disjointness



# Outline

- Introduction
- Lack of background knowledge
- Conclusions and future directions



# Semantic matching in a nutshell

**Semantic matching: given two graphs *G1* and *G2*, for any node $n1_i \in G1$, find the strongest semantic relation *R'* holding with node $n2_j \in G2$**

**Computed *R's*, listed in the decreasing binding strength order:**

- **equivalence { = }**
- **more general/specific { ⊒ , ⊑ }**
- **disjointness { ⊥ }**
- **I don't know {idk}**

**We compute semantic relations by analyzing the *meaning* (*concepts*, *not labels*) which is codified in the elements and the structures of ontologies**

**Technically, labels at nodes written in natural language are translated into propositional logical formulas which explicitly codify the labels' intended meaning. This allows us to codify the matching problem into a propositional validity problem, which can then be efficiently resolved using sound and complete state of the art satisfiability (SAT) solvers**



# Problem of low recall (incompletness) - I

**Facts**

- Matching (usually) has two components: element level matching and structure level matching
- Contrarily to many other systems, the semantic matching structure level algorithm is correct and complete
- Still, the quality of results is not very good

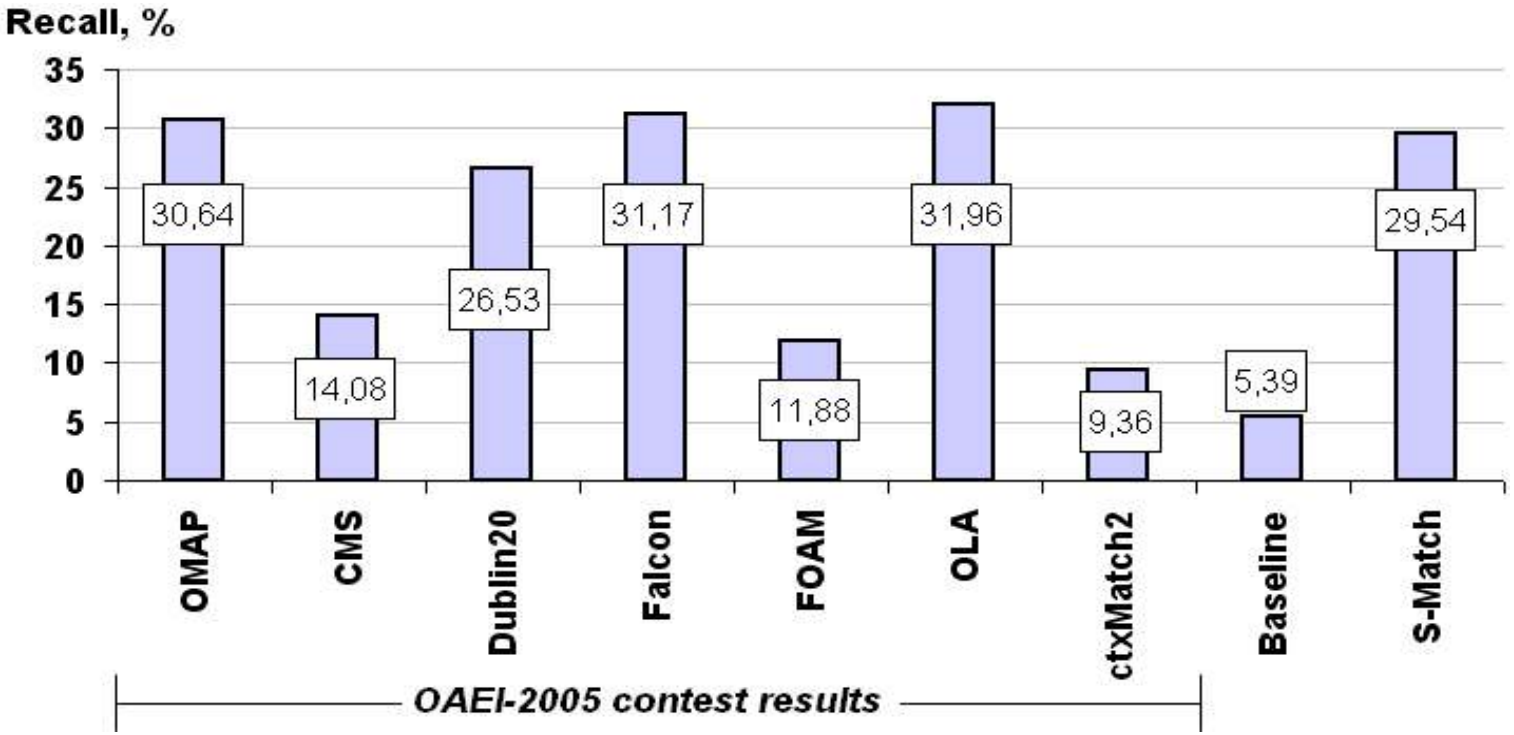
**Why? ... the problem of lack of knowledge**

# Problem of low recall (incompletness) - II

## Preliminary (analytical) evaluation

| Matching tasks | #nodes | max depth | #labels per tree |
|---|---|---|---|
| Google vs Looksmart | 706/1081 | 11/16 | 1048/1715 |
| Google vs Yahoo | 561/665 | 11/11 | 722/945 |
| Yahoo vs Looksmart | 74/140 | 8/10 | 101/222 |

Dataset
[P. Avesani et al., ISWC'05]






# On increasing the recall: an overview

## Multiple strategies

- Strengthen element level matchers
- Reuse of previous match results from the same domain of interest
  - PO = Purchase Order
- Use general knowledge sources (unlikely to help)
  - WWW
- Use, if available (!), domain specific sources of knowledge
  - FMA
  - Corpuses



# Iterative semantic matching (ISM)

## The idea

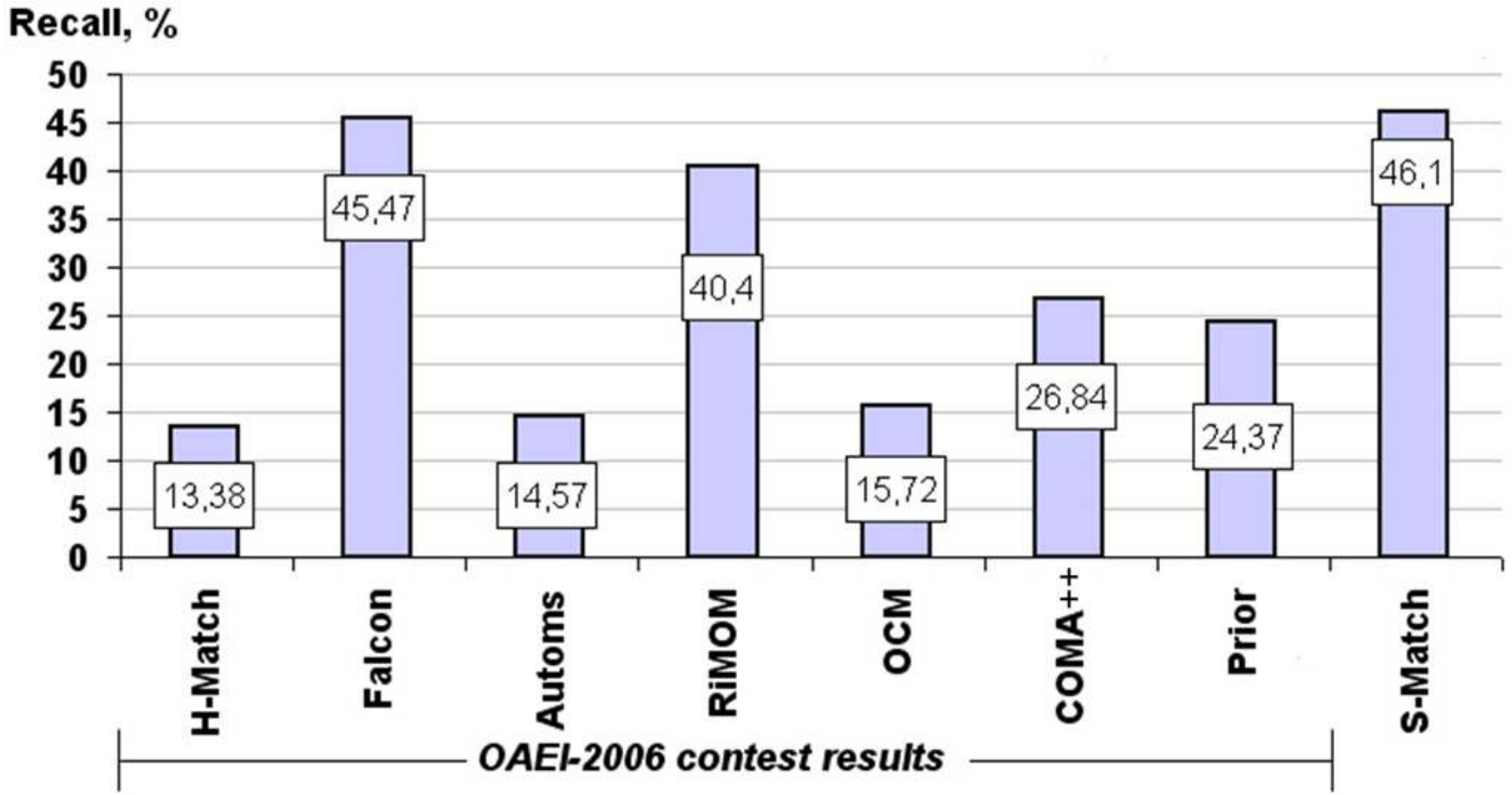
Repeat *element level matching* and *structure level matching* of the matching algorithm for some critical (hard) matching tasks

## ISM macro steps

- Discover *critical points* in the matching process
- Generate candidate *missing axiom(s)*
- Re-run SAT solver on a critical task taking into account the new axiom(s)
- If SAT returns **false**, save the newly discovered axiom(s) for future reuse



# OAEI-2006: web directories test case





# Outline

- Introduction
- Lack of background knowledge
- **Conclusions and future directions**



# Conclusions

The problem of missing **domain knowledge** is a major problem of all (!) matching systems

- This problem on the industrial size matching tasks is very hard
- We have investigated it by examples of light weight ontologies, such as Google and Yahoo
- Partial solution by applying semantic matching iteratively



# Future directions

- Iterative semantic matching
  - New element level matchers
- Interactive semantic matching
  - GUI
  - Cutomizing technology
- Extensive evaluation
  - Testing methodology
  - Industry-strength tasks



# References


- Project website - KNOWDIVE: http://www.dit.unitn.it/~knowdive/
- Ontology Matching website: http://www.OntologyMatching.org
- F. Giunchiglia, M. Yatskevich, P. Shvaiko: Semantic matching: algorithms and implementation. Journal on Data Semantics, IX, 2007.
- F. Giunchiglia, P. Shvaiko, M. Yatskevich: Discovering missing background knowledge in ontology matching. In Proceedings of *ECAI, 2006.*
- P. Avesani, F. Giunchiglia, M. Yatskevich: A large scale taxonomy mapping evaluation. In Proceedings of *ISWC, 2005.*
- J. Euzenat, P. Shvaiko : Ontology matching. Springer, 2007.
- E. Rahm, P. Bernstein. A survey of approaches to automatic schema matching. VLDB Journal, 2001.
- R. Gligorov, Z. Aleksovski, W. ten Kate, F. van Harmelen. Using google distance to weight approximate ontology. In Proceedings of WWW, 2007.
- S. Zhang, O. Bodenreider. Experience in aligning anatomical ontologies. International Journal on Semantic Web and Information Systems, 2007.
- J. Madhavan, P. Bernstein, A. Doan, A. Halevy. Corpus-based schema matching. In Proceedings of ICDE, 2005.
- H.-H. Do and E. Rahm. COMA – a system for flexible combination of schema matching approaches. In Proceedings of VLDB, 2002.




# Ontology Matching @ ISWC'07+ASWC'07

http://om2007.OntologyMatching.org

OM-2007

# Ontology Alignment Evaluation Initiative
# OAEI–2007 campaign

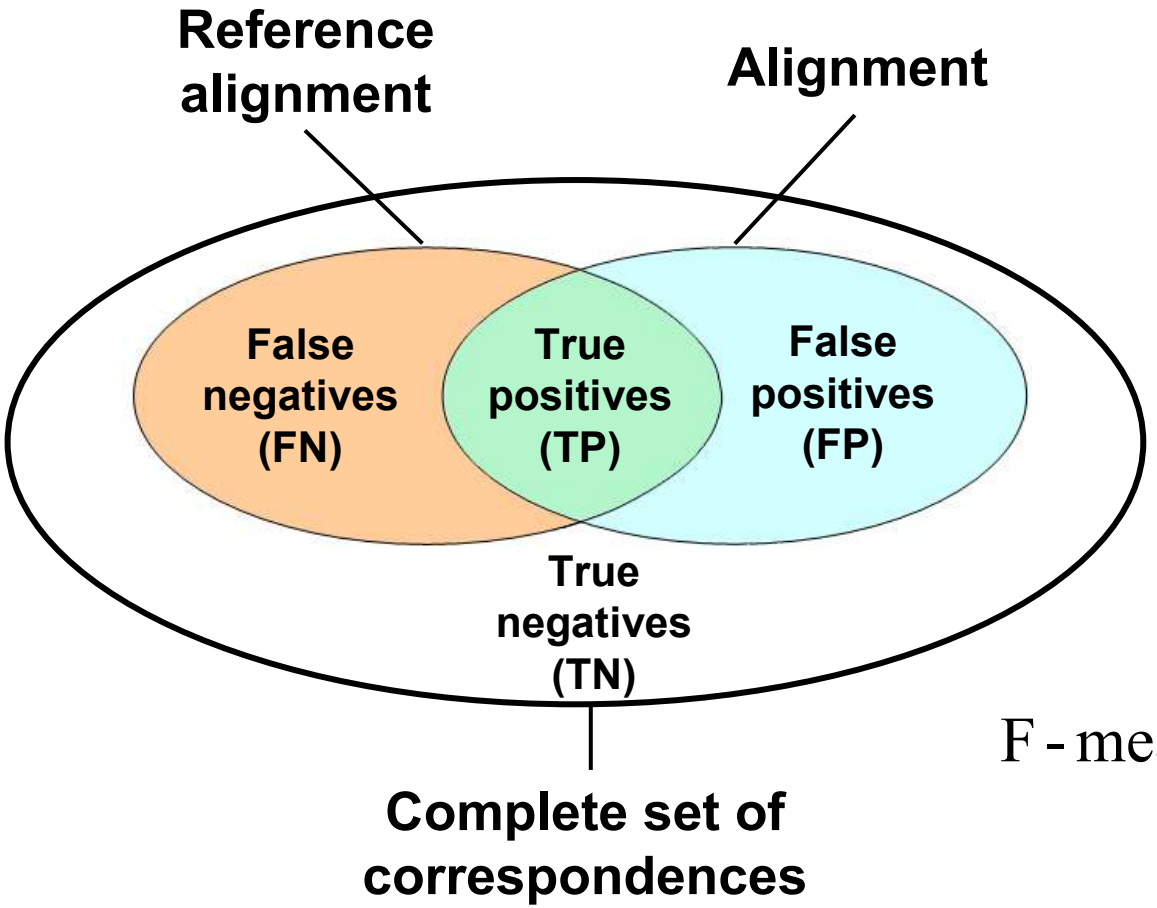
http://oaei.OntologyMatching.org/2007




# Thank you
# for your attention and interest!



# Evaluation (quality) measures




$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

$$\text{Recall} = \frac{|TP|}{|FN| + |TP|}$$

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Overall} = \text{Recall} \cdot \left( 2 - \frac{1}{\text{Precision}} \right)$$



# Discussion

- Missing background knowledge
  - Reuse
  - Asking agents on the web
- Interactive matching approaches
- Evaluation
  - Testing methodology
  - Industry-strength tasks
- Mapping generation



# Discussion