

S-Match

7 years of research and exploitation



UNIVERSITY
OF TRENTO - Italy



Know Dive.



The Paper



ESWS 2004: Heraklion, Greece

S-Match: an algorithm and an implementation of semantic matching

Fausto Giunchiglia, Pavel Shvaiko, Mikalai Yatskevich

Dept. of Information and Communication Technology
University of Trento,
38050 Povo, Trento, Italy
{fausto, pavel, yatskevi}@dit.unitn.it

Abstract. We think of *Match* as an operator which takes two graph-like structures (e.g., conceptual hierarchies or ontologies) and produces a mapping between those nodes of the two graphs that correspond semantically to each other. Semantic matching is a novel approach where semantic correspondences are discovered by computing, and returning as a result, the semantic information implicitly or explicitly codified in the labels of nodes and arcs. In this paper we present an algorithm implementing semantic matching, and we discuss its implementation within the *S-Match* system. We also test *S-Match* against three state of the art matching systems. The results, though preliminary, look promising, in particular for what concerns precision and recall.

1 Introduction

We think of *Match* as an operator that takes two graph-like structures (e.g., conceptual hierarchies, database schemas or ontologies) and produces mappings among the nodes of the two graphs that correspond semantically to each other. *Match* is a critical operator in many well-known application domains, such as schema/ontology integration, data warehouses, and XML message mapping. More recently, new application domains have emerged, such as catalog matching, where the match operator is used to map entries of catalogs among business partners; or web service coordination, where match is used to identify dependencies among data sources.

We concentrate on *semantic matching*, as introduced in [6], based on the ideas and system described in [2]. The key intuition behind semantic matching is that we should calculate mappings by computing the semantic relations holding between the concepts (and not labels!) assigned to nodes. Thus, for instance, two concepts can be equivalent, one can be more general than the other, and so on. We classify all previous approaches under the heading of *syntactic matching*. These approaches, though implicitly or explicitly exploiting the semantic information codified in graphs, differ substantially from our approach in that, instead of computing semantic relations between nodes, they compute syntactic "similarity" coefficients between labels, in the [0,1] range. Some examples of previous solutions are [12], [1], [15], [18], [5], [10]; see [6] for an in-depth discussion about syntactic and semantic matching.

In this paper we propose and analyze in detail an algorithm and a system implementing semantic matching. Our approach is based on two key notions, the notion of



Fausto Giunchiglia



Pavel Shvaiko



Mikalai Yatskevich

The Team



S-Match

Fausto Giunchiglia
Pavel Shvaiko
Mikalai Yatskevich
Aliaksandr Autayeu



Fausto Giunchiglia
coordinator



Pavel Shvaiko



Mikalai Yatskevich



Aliaksandr Autayeu

Minimal Mappings

Fausto Giunchiglia
Vincenzo Maltese
Aliaksandr Autayeu

Structure Preserving Semantic Matching

Fausto Giunchiglia
Juan Pane
Lorenzino Vaccari
Gaia Treçarichi
Mikalai Yatskevich



Gaia Treçarichi



Juan Pane



Lorenzino Vaccari

Background Knowledge Datasets

Fausto Giunchiglia
Vincenzo Maltese
Feroz Farazi
Biswanath Dutta



Feroz Farazi



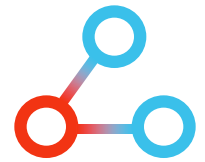
Biswanath Dutta



Vincenzo Maltese



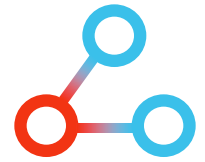
Overview



- Introduction to S-Match
- Lightweight Ontologies
- Matching Tools
 - S-Match
 - Structure Preserving Semantic Matching (SPSM)
 - MinSMatch for minimal mappings
- Evaluations
- Enhancements: NLP, BK
- Open Source Framework
- Exploitation
- Future activities



Living with heterogeneity [KER-03]



- The semantic web will be: huge, dynamic and heterogeneous. These are **not bugs**, these are **features**
- We must learn to live with them and master them
- Often information resources expressed in different ways must be reconciled before being used. Mismatch between formalized knowledge can occur when:
 - different languages are used
 - different **terminologies** are used
 - different **modeling** is used



On reducing heterogeneity [KER-03]



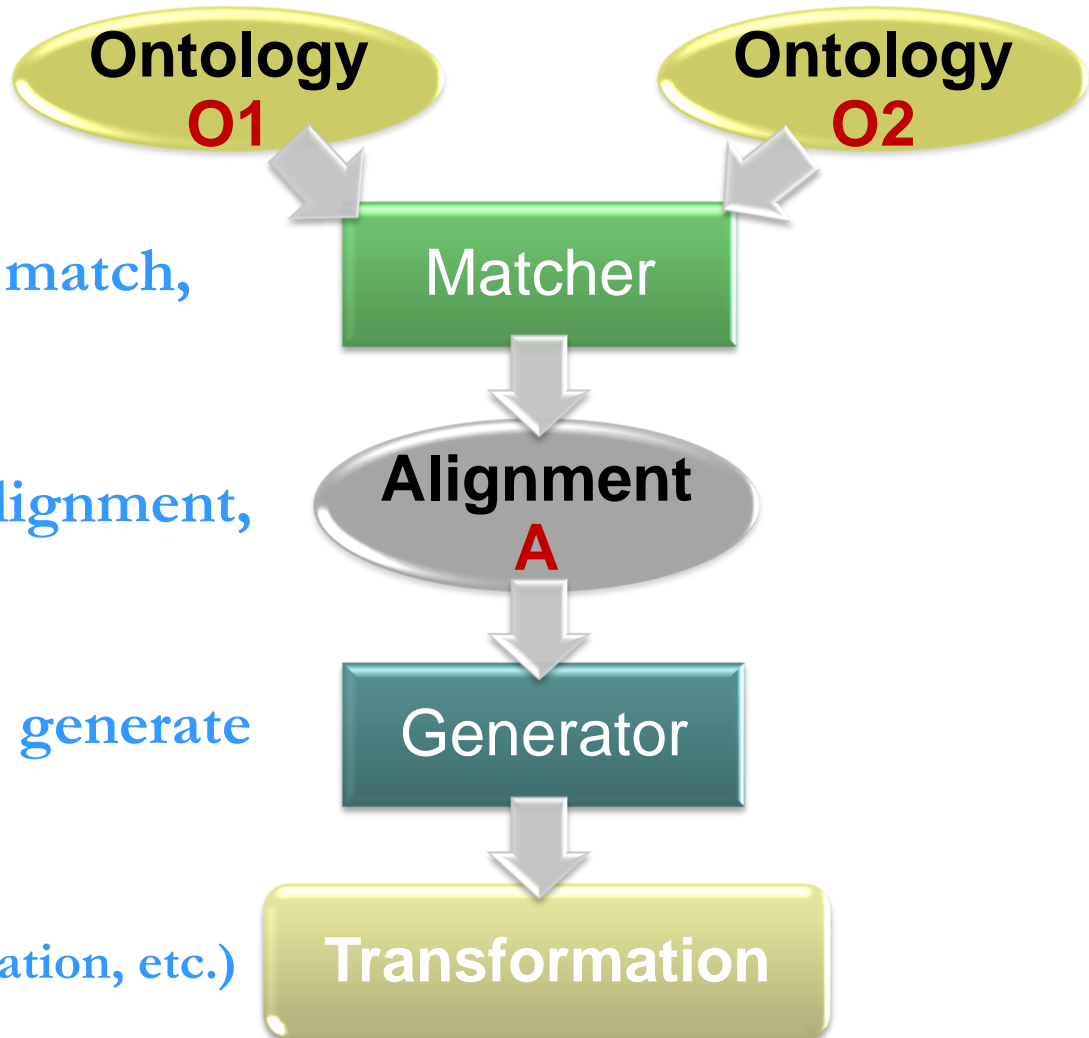
Reconciliation can be performed in 2 steps:

(i) match,

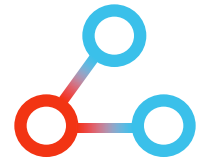
thereby determine an alignment,

(ii) generate

a processor (for transformation, etc.)



2004: what made the difference?



o- About 30+ matching systems existed in 2004

- o Cupid, COMA, Rondo, NOM, OLA, Prompt, Anchor-Prompt, CtxMatch, ...
- o now 100+ systems exist

o- [0..1] vs. { =, < , > , ⊥ }

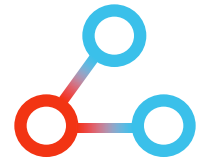
- o Most systems were computing and aggregating various **similarity measures in [0 1]** to produce alignments
- o We computed **logical relations**: equivalence, subsumption, ...

o- Heuristics vs. soundness and completeness

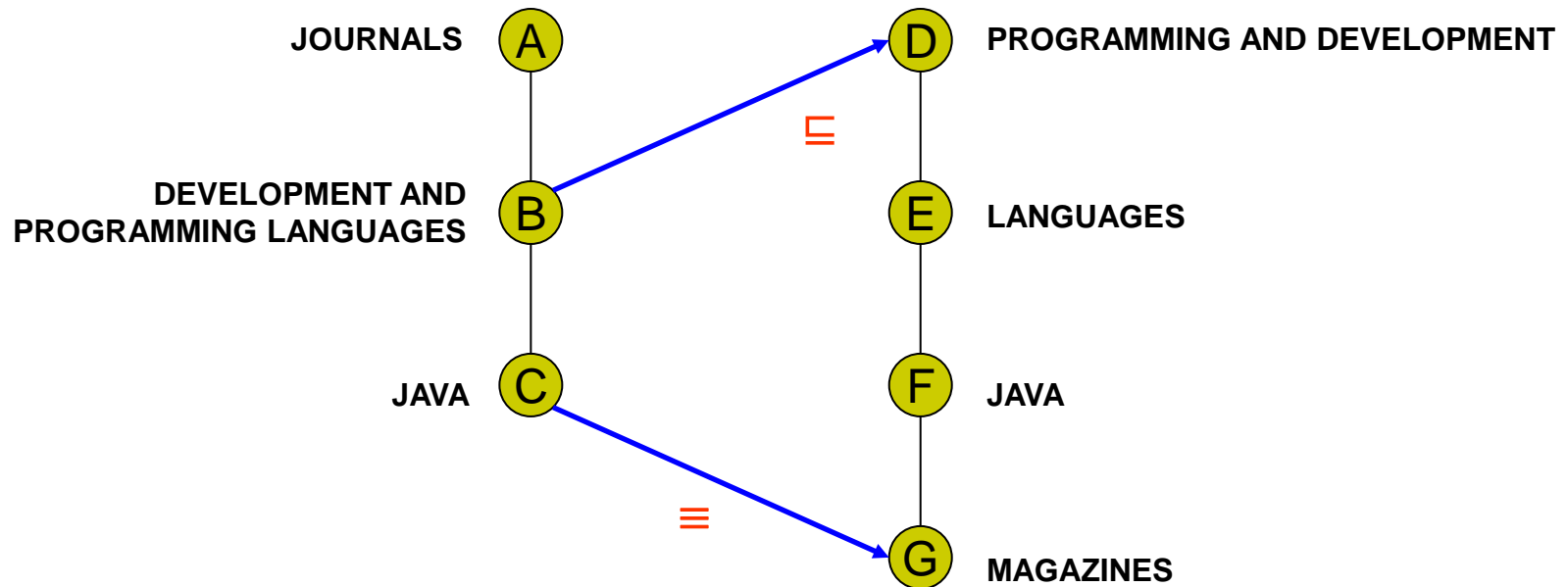
- o Most systems were using matching **heuristics** that sometimes worked well, sometimes not so well. We followed this path as well, but...
- o One step of the matching process was **sound and complete**



What is Semantic Matching [KER-03]

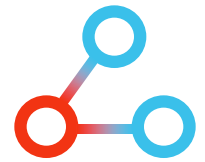


- o An operation that identifies semantically similar nodes in two graph-like structures



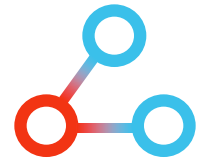
- o Applications: catalog integration, peer to peer information sharing, resource discovery, query answering, ...

The Key Idea [KER-03, ESWS-04]



- Take as input two **graph-like structures**, e.g., ontologies
- Return as output **logic relations**, e.g., equivalence, subsumption, which are supposed to hold between the nodes of the graphs
- Entities of the input ontologies are translated into **propositional formulas** which explicitly express the concept descriptions as encoded in the **ontology structure and in external resources**, such as WordNet
- Translation of the matching problem into a **propositional validity problem**
- Propositional validity problem, efficiently resolved using **sound and complete** propositional satisfiability (SAT) solvers

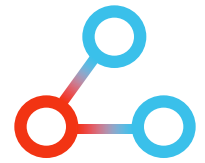
S-Match Algo Key Steps [ESWS-04]



- Given two trees (lightweight ontologies) T1 and T2 :
 1. For all **labels** in T1 and T2 compute concepts at labels (analysis of labels in isolation; from natural language to propositional logic)
 2. For all **nodes** in T1 and T2 compute concepts at nodes (take into account structure of the trees)
 3. For all pairs of labels in T1 and T2 compute relations between atomic **concepts at labels** (build Theory)
 4. For all pairs of nodes in T1 and T2 compute relations between **concepts at nodes** (run SAT)
- Steps 1, 2: **preprocessing phase** (once for all)
- Steps 3, 4: **matching phase** (run-time)

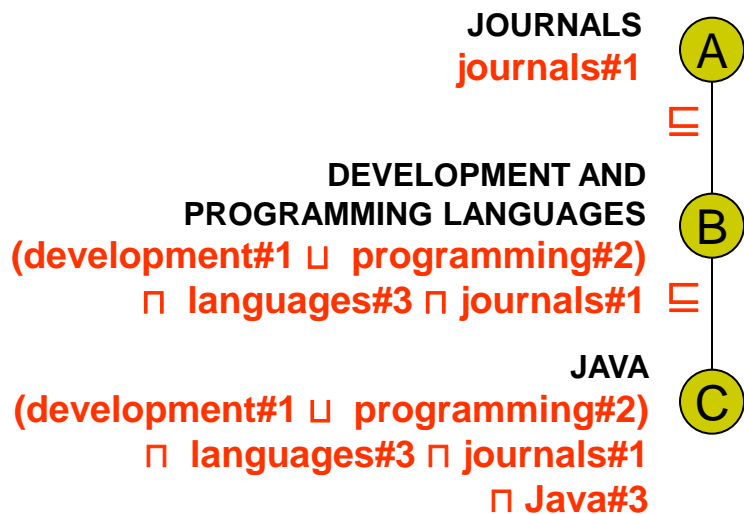
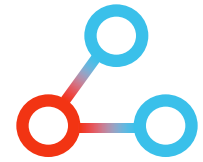


Lightweight Ontologies [JODS-05]



- Lightweight ontologies are tree structures where concepts at nodes are connected with subsumption in DL
- Many of the schemas in the world can be translated into lightweight ontologies
 - User classifications (file systems, email folder structures)
 - Web directories and business catalogues
 - Library classifications (thesauri, subject headings)
- **With the translation:**
 - Node labels are formulas in propositional Description Logic (DL)
 - Concepts are taken from WordNet senses (or other dictionaries)
 - Tree structures: each node formula is subsumed by parent node formula

Lightweight Ontologies (cont)



Matching Tools



o- S-Match: the basic semantic matching tool

- o It returns the set of semantic correspondences between two **lightweight ontologies**
- o Output: \perp , \exists , \sqsubseteq , \equiv

o- SPSM: Structure Preserving Semantic Matching

- o Only one correspondence per node is returned
- o It matches leaf nodes to leaf nodes and internal nodes to internal nodes
- o Used **to compare function definitions**

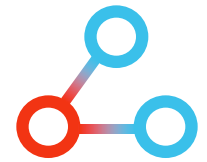
o- MinSMatch: to compute minimal mappings

- o It returns **the minimal set** of semantic correspondences between two lightweight ontologies. It always exists and it is unique
- o It computes **the set of maximum size** (containing the maximum number of minimal and redundant links) from the propagation of the links in the minimal set

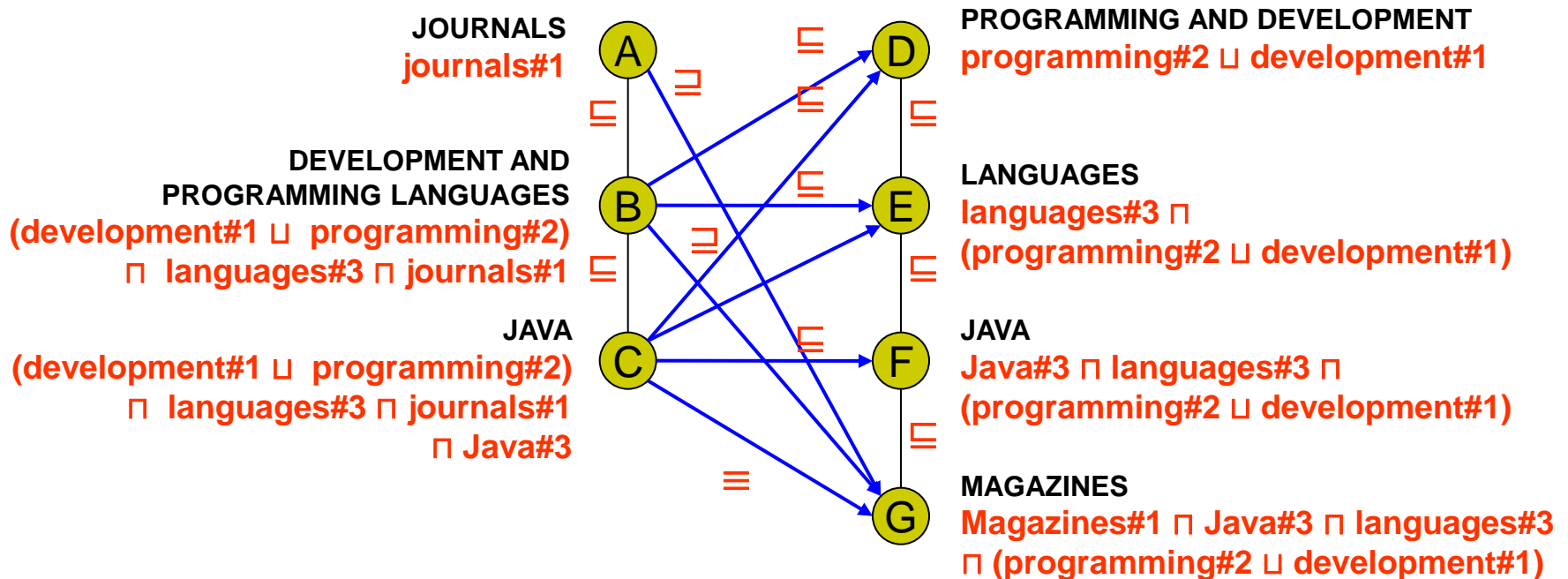
o- S-Match GUI



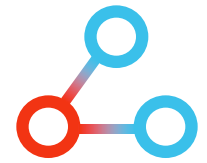
S-Match [ESWS-04]



- o An alignment is a set of mapping elements $\langle \text{source}, \text{target}, R \rangle$
 - o $R \in \{ \perp, \equiv, \sqsubseteq, \supseteq \}$ partially ordered
 - o For each pair of nodes a call to a SAT solver verifies if a given semantic relation holds between the two, given the available **background knowledge**
 - o Visualization and usability problems (e.g. validation and maintenance)



SPSM [ODBASE-08a]



o- SPSM: Structure Preserving Semantic Matching

o Example with two web services:

- o **Get_Wine**(Region, Country, Color, Price, Number_of _bottles)
- o **Get_Wine**(**Region**(Country, Area), Colour, Cost, Year, Quantity)
- o SPSM (T1,T2) = 0.62 + set of mapping elements

o Uses **abstraction** operations to preserve structures, namely it computes one-to-one correspondence, such that:

- o Functions are matched to functions
- o Variables are matched to variables

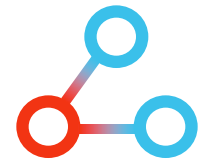
o Outputs a **global similarity measure** and a **set mapping elements**.

o Node matching is done with S-Match

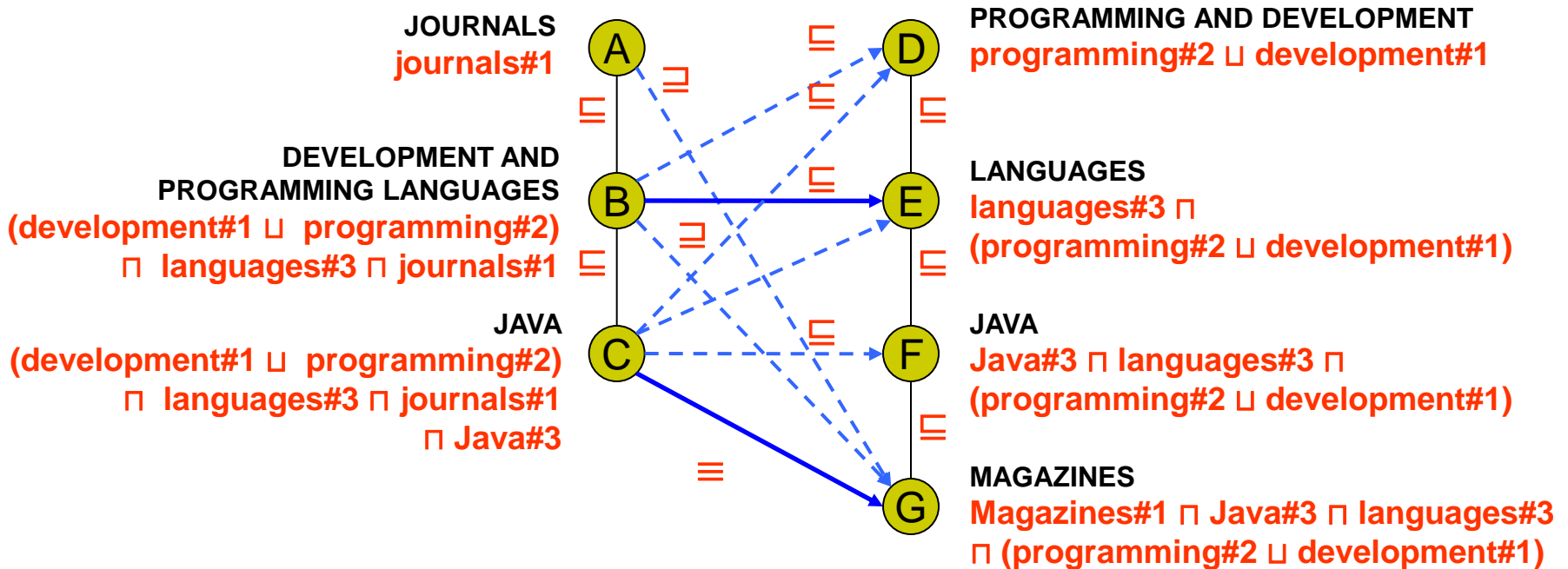
o A global similarity measure is computed using Tree edit distance



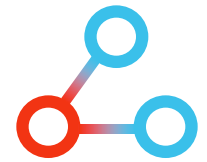
MinSMatch [ODBASE-10]



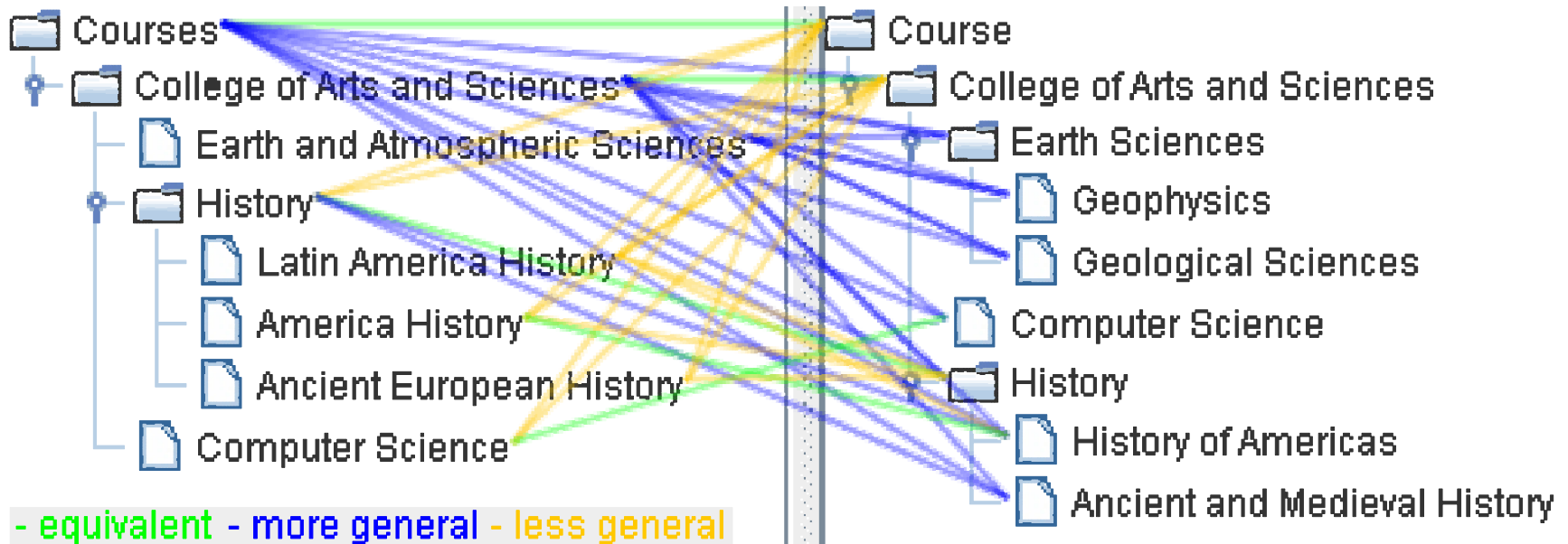
- Based on a set of **redundancy patterns** the **minimal mapping** is that minimal subset of correspondences such as all the others can be efficiently computed from them
- The **minimal mapping** **always exists** and it is **unique**



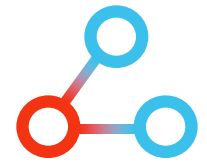
S-Match GUI [SWJ-10]



Traditional visualization: crowded already with only 34x39 nodes



S-Match GUI [SWJ-10]



New GUI

- node-links
- ellipsis
- hints
- path-to-root
- links table
- editing
- synchronized navigation

Source Target Mapping Edit View Options Help

Mapping: Config: s-match.properties

\test-data\cw\result-minimal-cw.txt

\test-data\cw\c.xml

\test-data\cw\w.xml

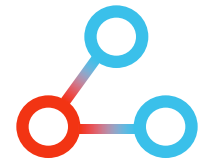
Source	Relation	Target
Earth and Atmospheric Sciences	more general	Earth Sciences
Economics	equivalent	Economics
English	equivalent	English

2011-01-10 21:14:36,015 INFO SMatchGUI - Reading properties ..\conf\SMatchGUI.properties

2011-01-10 21:14:36,837 INFO SMatchGUI - Creating MatchManager with config: ..\conf\s-match.properties



MinSMatch Evaluation [ODBASE-10]



Mapping sizes and percentage of reduction on standard datasets

Datasets (nodes)	Mapping of maximum size	Minimal Mapping size	Reduction (%)
#1 Cornell/Washington (34/39)	223	36	83.86
#2 Topia/Icon (542/999)	5491	243	95.57
#3 Web dir. Source/Target (2857/6628)	282648	30956	89.05
#4 EClass/UNSPSC (3358/5293)	39818	12754	67.97

Reduction in run time and calls to SAT

#	Run Time (ms)			Calls to logical reasoner (SAT)		
	S-Match	MinSMatch	Reduction (%)	S-Match	MinSMatch	Reduction (%)
1	472	397	15.88	3978	2273	42.86
2	141040	67125	52.40	1624374	616371	62.05
3	3593058	1847252	48.58	56808588	19246095	66.12
4	6440952	2642064	58.98	53321682	17961866	66.31



MinSMatch Evaluation [ODBASE-10]



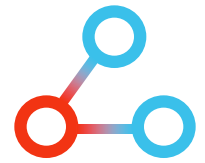
Mapping sizes and percentage of reduction on NALT and LCSH

Id	Source	Branch
A	NALT	Chemistry and Physics
B	NALT	Natural Resources, Earth and Environmental Sciences
C	LCSH	Chemical Elements
D	LCSH	Chemicals
E	LCSH	Management
F	LCSH	Natural resources

Branches	Mapping of maximum size	Minimal mapping size	Reduction (%)
A vs. C	17716	7541	57,43
A vs. D	139121	994	99,29
A vs. E	9579	1254	86,91
B vs. F	27191	1232	95,47



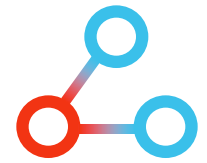
Improved NLP [ISWC-07, ECDL-10]



- Classifications, database schemas, APIs...
- Natural Language Metadata: labels, very short pieces of text
 - short context to no context
 - special syntax tools
 - biased toward nouns distribution of parts of speech
- Improved NLP: manual annotation + language analysis
 - tokenization
 - parts of speech tagging
 - lightweight parsing: simple NP-based grammar
- +18% in translation accuracy



Improved BK [ECAI-06, ISWC-10]



BK: Background Knowledge

WordNet

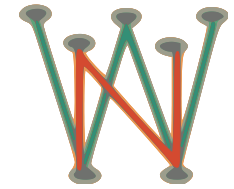
- <http://wordnet.princeton.edu>
- general, small, single language
- ~120K concepts, covers daily language

GeoWordNet

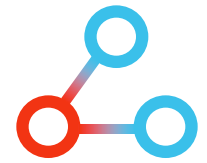
- <http://geowordnet.semanticmatching.org/>
- specific, huge, several languages
- ~3.6M+ entities, 7.2M+ relations, world places

Entitypedia

- <http://entitypedia.org/>
- general, huge, multilingual,
- covers world entities and domains, coming soon...



Open Source Framework [SM, SWJ-10]



○ <http://semanticmatching.org/> since March 2010

○ SF.net community

○ Source Code

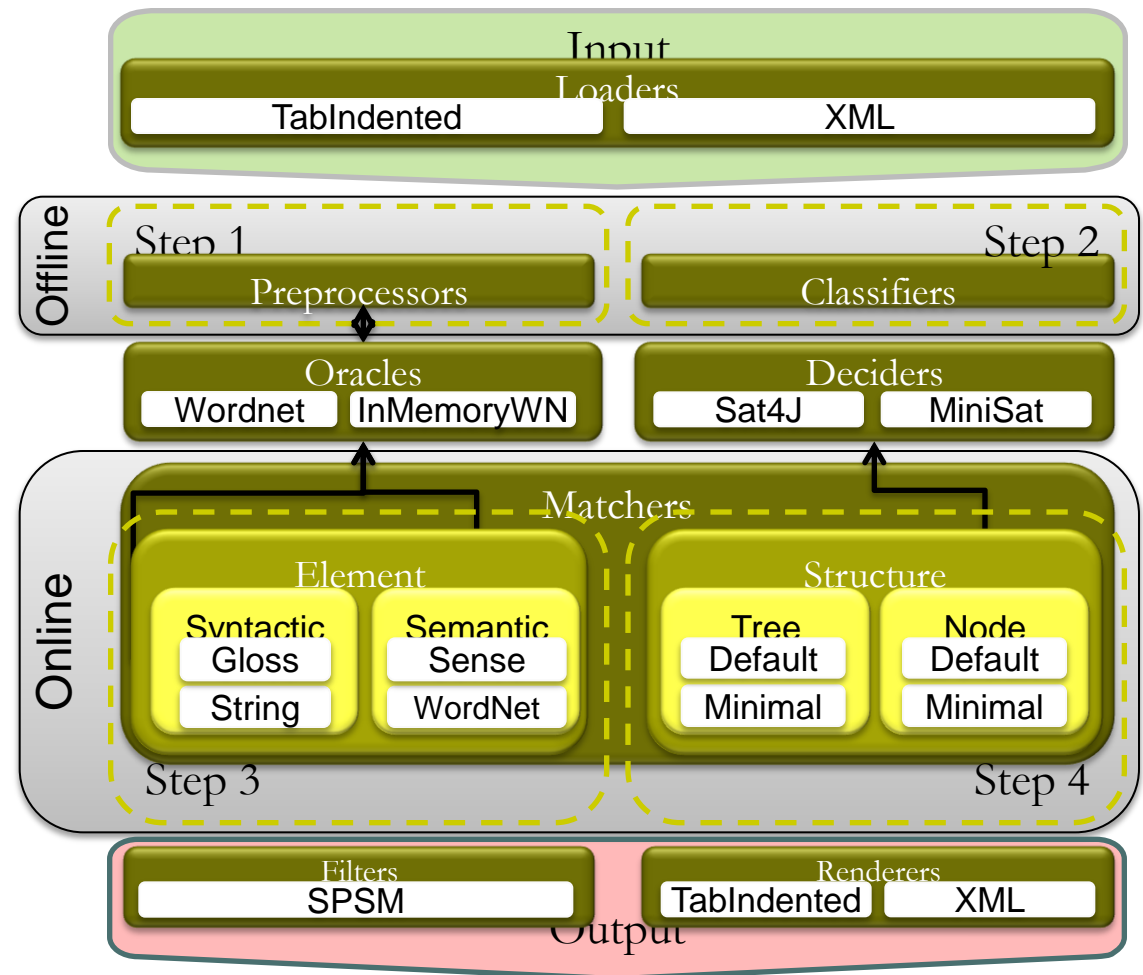
○ Documentation

○ Data sets

○ LGPL

○ CC-BY

○ almost 2000 dls



Exploitation



Semantic Geo-Catalog (SGC)

S-Match is used to match a user query to a faceted ontology in the geo-spatial domain



Experiments in the agriculture domain

S-Match to match AGROVOC with CABI



Interconcept

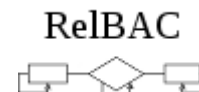
MinSMatch to match Knowledge Organization Systems in digital libraries



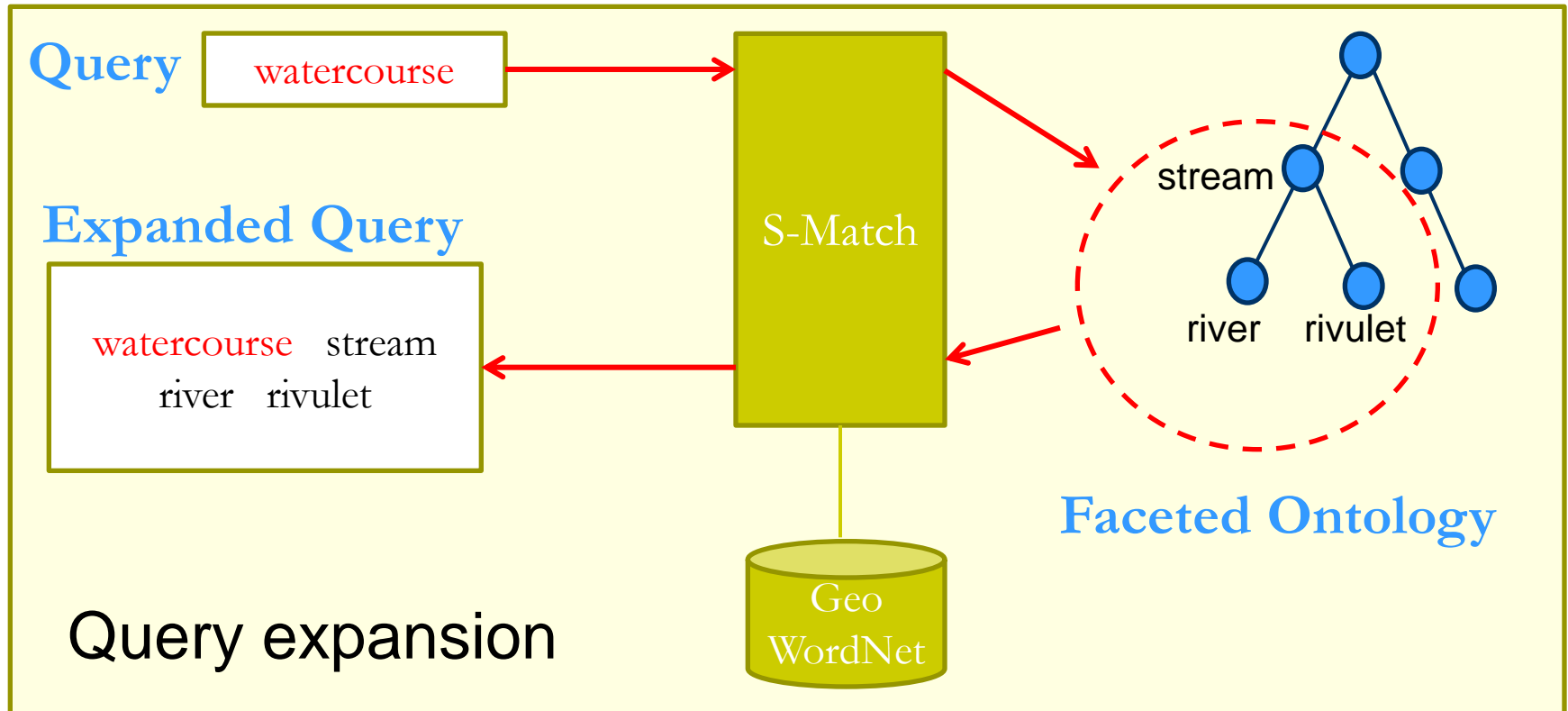
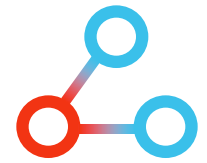
Open Knowledge

SPSM to match web services

And others ...

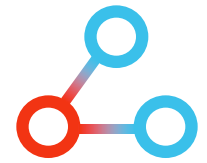


Semantic Geo-Catalog [ESWC-11]



- The query expansion component integrated with the geo-catalog
- The local dataset of the Province of Trento has been used to construct the faceted ontology and integrated with GeoWordNet

Semantic Matching: Theory and Practice



by Fausto Giunchiglia and Aliaksandr Autayeu

end 2011 - beginning 2012

Fundamentals

- Introduction to Semantic Matching
- Lightweight Ontologies
- Basic Algorithm
- Structure Preserving Semantic Matching
- Minimal Semantic Matching
- Non-Standard Uses of Matching

The Framework

- Introduction to the S-Match
- Input: Everything is a Tree
- Processing Natural Language Metadata
- Background Knowledge

... The Framework

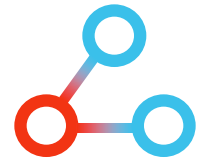
- ...
- Background Knowledge
- Element-level Matching
- Structure-level Matching
- Advanced Matching
- Output: Semantic Mappings
- Framework Extensions

Datasets and Evaluation

- Evaluation Issues and Methodology
- Datasets
- Evaluating Conversion into Lightweight Ontologies
- Evaluating Matching Techniques



Other Relevant Initiatives



○- OAEI: *Ontology Alignment Evaluation Initiative*

- since 2004, supported by
 - Pavel Shvaiko, Mikalai Yatskevich, Juan Pane
- <http://oaei.ontologymatching.org/>

○- *Ontology Matching Workshop at ISWC*

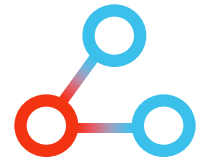
- since 2006, supported by
 - Pavel Shvaiko, Fausto Giunchiglia
- <http://om2011.ontologymatching.org/>

○- *Book on Ontology matching [OMB-07]*

- In 2007, by Pavel Shvaiko and others



References



- ❑ [KER-03] F. Giunchiglia, P. Shvaiko: **Semantic Matching**. *The Knowledge Engineering Review Journal*, 2003
- ❑ [ESWS-04] F. Giunchiglia, P. Shvaiko, M. Yatskevich: **S-Match: an Algorithm and an Implementation of Semantic Matching**. *ESWS 2004. (extended and updated version [JODS-07])*
- ❑ [JODS-05] F. Giunchiglia, M. Marchese, I. Zaihrayeu. **Encoding Classifications into Lightweight Ontologies**. *In Journal of Data Semantics (JoDS), VIII: Special Issue on Extended Papers from ESWC 2005*.
- ❑ [ISWC-05] P. Avesani, F. Giunchiglia, M. Yatskevich : **A Large Scale Taxonomy Mapping Evaluation**. ISWC 2005
- ❑ [ECAI-06] F. Giunchiglia, P. Shvaiko, M. Yatskevich: **Discovering Missing Background Knowledge in Ontology Matching** ECAI 2006
- ❑ [JODS-07] F. Giunchiglia, P. Shvaiko, M. Yatskevich. **Semantic Matching: algorithms and implementation**. *Journal of Data Semantics, 2007, v. IX, p.1-38*.
- ❑ [ISWC-07] I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, X. Huang. **From Web Directories to Ontologies: Natural Language Processing Challenges**. *ISWC 2007*.
- ❑ [OMB-07] J. Euzenat, P. Shvaiko. **Ontology Matching**. Springer, 2007.
- ❑ [ODBASE-08a] F. Giunchiglia, F. McNeill, M. Yatskevich, J. Pane, P. Besana, P. Shvaiko. **Approximate structure-preserving semantic matching**. *ODBASE 2008*.
- ❑ [KER-09] F. Giunchiglia, M. Yatskevich, P. Avesani, P. Shvaiko. **A large dataset for the evaluation of ontology matching**. *In Knowledge Eng. Review 24(2)(2009)*
- ❑ [ODBASE-10] V. Maltese, F. Giunchiglia, A. Autayeu. **Save up to 99% of your time in mapping validation**. *ODBASE 2010*.
- ❑ [ECDL-10] A. Autayeu, F. Giunchiglia, P. Andrews. **Lightweight Parsing of Classifications into Lightweight Ontologies**. *ECDL 2010*.
- ❑ [SWJ-10] F. Giunchiglia, A. Autayeu, J. Pane. **S-Match: an open source framework for matching lightweight ontologies**. *In Semantic Web Journal, S. I. on Semantic Web Tools And Systems, 2010*.
- ❑ [ESWC-10] F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta: **GeoWordNet: A Resource for Geo-spatial Applications**. ESWC 2010
- ❑ [ESWC-11] F. Farazi, V. Maltese, F. Giunchiglia, A. Ivanyukovich. **A faceted ontology for a semantic geo-catalogue**. *ESWC 2011*.
- ❑ [SM] F. Giunchiglia, A. Autayeu. <http://semanticmatching.org/>



Thank you for your time and interest!

Questions?

<http://semanticmatching.org/>

