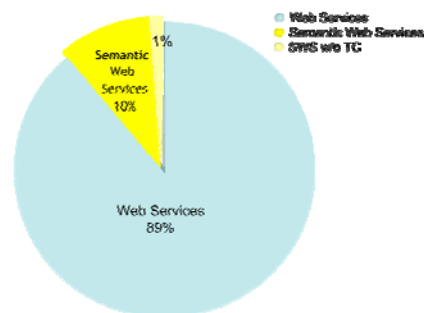
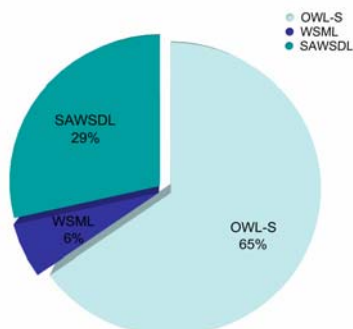




## 3<sup>rd</sup> International Semantic Service Selection Contest – Performance Evaluation of Semantic Service Matchmakers –

Matthias Klusch (DFKI, Germany)  
 Alain Leger (France Telecom Research, France)  
 David Martin (SRI International, USA)  
 Massimo Paolucci (NTT DoCoMo Research Europe, Germany)  
 Abraham Bernstein (University of Zurich, Switzerland)  
 Ulrich Küster (University of Jena, Germany)

## Public Semantic Services in the Web 2009



Sousuo 24-09-09: 410 w/o test collections  
 3508 w/ test collections

Public semantic service retrieval test collections:

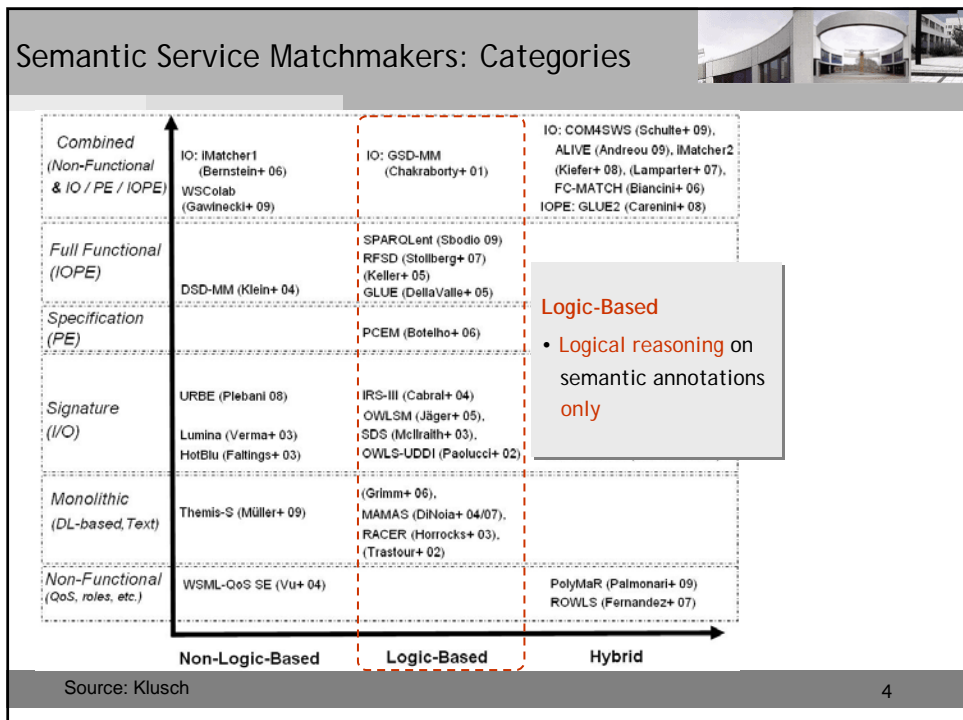
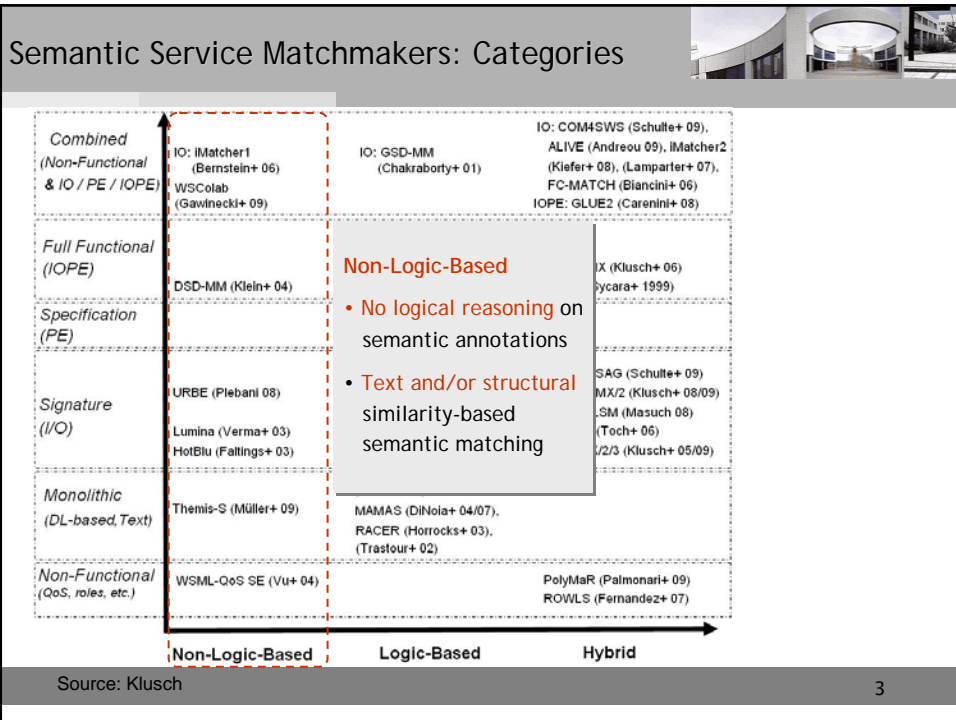
- OWL-S: OWLS-TC2 (semwebcentral.org), TC (ce.sharif.edu)
- SAWSDL: SAWSDL-TC1 (semwebcentral.org)
- None for WSML yet.

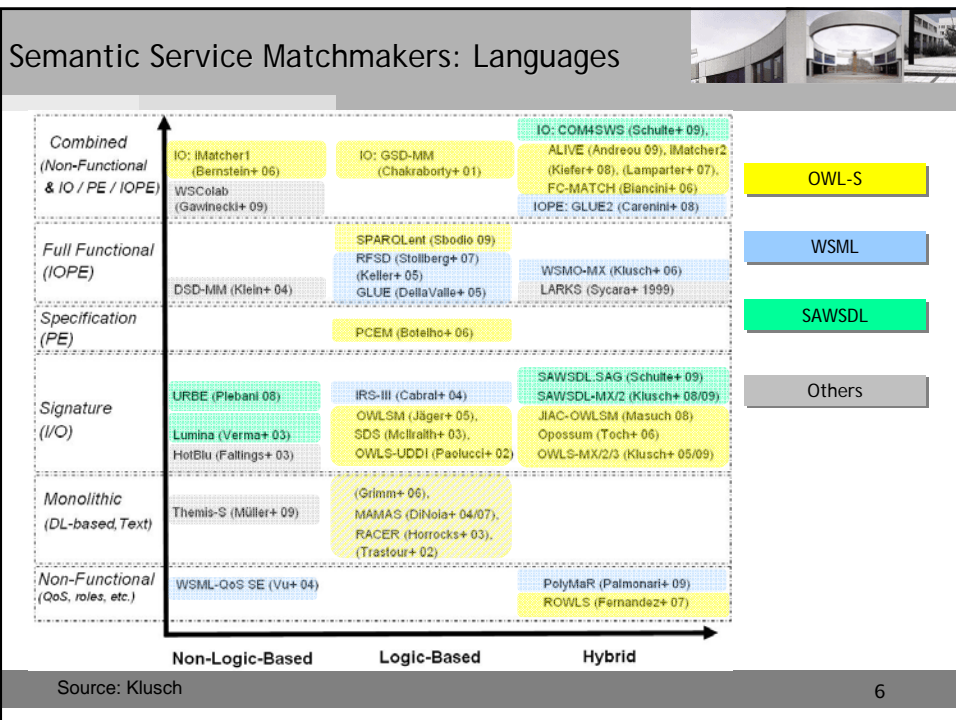
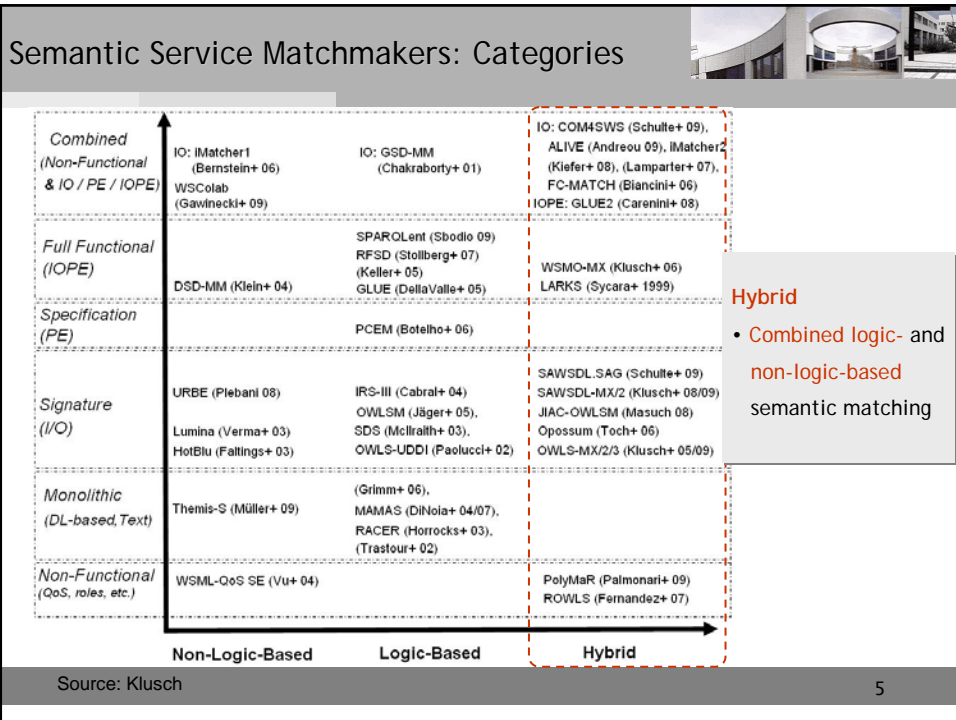


Seekda 24-09-09

Source: Klusch

2







### Track 1: OWL-S Service Matchmakers

1. **JIAC-OWLSM** (TU Berlin, Germany)
2. **Opossum** (Technion, Israel)
3. **OWLS-MX 2.0** (DFKI, Germany)
4. **OWLS-MX 3.0** (DFKI, Germany)
5. **OWLS-iMatcher** (U Zurich, Switzerland)
6. **SPARQLent** (Hewlett-Packard EIC, Italy)
7. **ALIVE** (U Bath, UK)

### Track 2: SAWSDL Service Matchmakers

1. **URBE** (Politecnico di Milano, Italy)
2. **SAWSDL-MX2** (DFKI, Germany)
3. **COM4SWS** (TU Darmstadt, Germany)
4. **SAWSDL-iMatcher3/1** (U Zurich, Switzerland)



### Track 3: Initial Cross-Evaluation -

Matchmakers for different formats tested over same collection

SAWSDL: **SAWSDL-MX1**, **SAWSDL-MX2**, **SAWSDL-iMatcher3/1**

OCML-LISP: **IRS-III** (Open U, UK)

Natural Language Text and Tagging: **Themis-S** (U Muenster, Germany)

**WSColab** (U Modena & Reggio Emilia, Italy)

## S3 Contest 2009: Evaluation Setting



- **Service retrieval test collections**

- Track1: OWLS-TC 3.0 (1007serv, 29req, 24ont), WSDL 1.1, binary & graded relevance  
20-10-2009: 10.076 downloads (since April 2005) @semwebcentral.org
- Track2: SAWSDL-TC 1.0 (894serv, 26req, 24ont), WSDL 1.1, binary relevance  
20-10-2009: 234 downloads (since July 2008) @semwebcentral.org
- Track3: JGD50-SAWSDL, JGD50-OCML-LISP, JGD50-NL-Tags

- **Standard retrieval performance measures**

- Binary relevance: Macro-averaged recall/precision, Average precision
- Graded relevance: Q, nDCG (averaged cumulative gain)
- Average query response time: Elapsed time (secs) per query execution

- **Evaluation tool**

- SME<sup>2</sup> v2.1 @semwebcentral.org

Source: Klusch

9

## Evaluation Tool SME<sup>2</sup> v2.1



<http://projects.semwebcentral.org/projects/sme2/>, 20-10-2009: 834 D/L (since 4/2008)

Source: Klusch

10

## Track 1: OWL-S Matchmakers in Brief



- **JIAC-OWLSM**

- Selection: *Hybrid; Signature (I/O)*
  - **Logic-based match**: Logical I/O concept subsumption relation as numeric score
  - **Non-logic-based match**: *Integrated* string matching of I/O concept names `string.equal()`, `string.contains()`
  - **Ranking**: Linear weighted aggregation of logical and string matching scores
- Dev: Nils Masuch (TU Berlin, Germany)

- **Opossum** Selection: *Hybrid; Signature (I/O)*

- **Logic-based match**: Logical I/O concept relationship
- **Non-logic-based match**: Numerical score from logic-based match, shortest path distance, concept depth/avg. Ontology depth, subsequent **ranking**
- Dev: Eran Toch (CMU, USA); Avigdor Gal, Dov Dori (Technion, IL), Iris Reinhartz-Berger (Haifa U, IL)

Source: Klusch

11

## Track 1: OWL-S Matchmakers in Brief



- **OWLS-MX3**

- Selection: *Hybrid, adaptive; Signature (I/O)*
  - **Logic-based match** (cf. OWLS-MX2); **Non-logic-based match**: Text similarity-based (cf. OWLS-MX2), Ontology-based structural match - Separated filters
  - **Adaptive (offline)**: SVM relevance classifier for aggregation of matching degrees with subsequent **ranking**
- Dev: Matthias Klusch, Patrick Kapahnke (DFKI, Germany)

- **OWLS-iMatcher2**

- Selection: *Hybrid; Signature (I/O), Service Name*
  - **Logic-based**: Logical unfolding of I/O concepts (Pellet)
  - **Non-logic-based**: Text similarities of unfolded service signatures and names
  - **Ranking**: Text similarity
- Dev: Christoph Kiefer, Avi Bernstein (U Zurich, Switzerland)

Source: Klusch

12

## Track 1: OWL-S Matchmakers in Brief



### • SPARQLent

- Selection: *Logic-Based; Signature (I/O); Specification (PE)*
  - **Logic-based match:** P/E described in SPARQL, I/O represented as additional constraints; I/O concept match w/ RDF entailment rules for RDF-encoded OWL
  - **Ranking:** ?
- Dev: Marco Luca Sbordio (Hewlett-Packard EIC, Italy)

### • ALIVE

- Selection: *Hybrid semantic; Signature (I/O), Service description tag*
  - **Logic-based match:** Logical I/O concept subsumption
  - **Non-logic-based match:** Additional text similarity match of text annotations
  - **Ranking:** Logic-based degree followed by text similarity-based ranking
- Dev: Dimitris Andreou (U Bath, UK)

Source: Klusch

13

## Performance Evaluation (Binary Relevance)



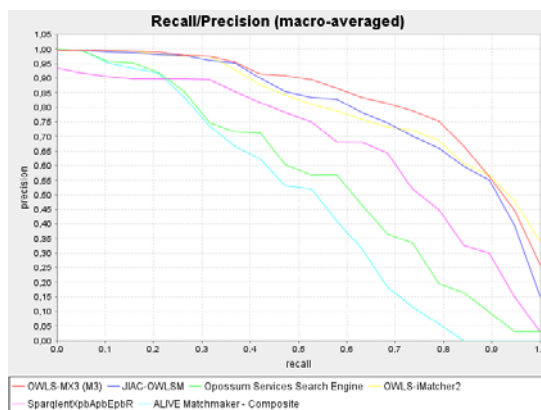
### Average Precision:

1. <b>OWLS-MX3</b>		<b>.861</b>
2. OWLS-iMatcher2		.846
3. JIAC-OWLSM		.814
4. SPARQLent		.718
5. OPOSSUM		.57
6. ALIVE		.5

### Avg Query Response Time (sec):

1. <b>OPOSSUM</b>		<b>.08</b>
2. ALIVE		.26
3. SPARQLent		.8
4. OWLS-iMatcher2		2.38
5. OWLS-MX3		4.37
6. JIAC-OWLSM		4.44

### Macro-averaged Recall/Precision:




Source: Klusch

14

## Performance Evaluation (Graded Relevance)



Precision	Q	nDCG	Average Precision (Binary Relevance):	
1. OWLS-MX3 	.86	.92	1. OWLS-MX3	.861
2. JIAC-OWLSM	.79	.89	2. OWLS-iMatcher2	.846
3. OWLS-iMatcher2	.83	.88	3. JIAC-OWLSM	.814
4. SPARQLent	.67	.82	4. SPARQLent	.718
5. OPOSSUM	.51	.71	5. OPOSSUM	.57
6. ALIVE	.42	.64	6. ALIVE	.5

Source: Klusch

15

## Track 2: SAWSDL Matchmakers in Brief



### • URBE

- Selection: *Non-logic-based; Signature (I/O)*
  - **Non-logic-based match:** Bipartite graph-matching of service operations; Ontology-based structural I/O concept similarity (worst-case path length in given reference ontology); Text similarity (WordNet) for property-class and XSD data type matching
  - **Ranking:** Weighted aggregation of structural and text matching scores
- Dev: Pierluigi Plebani (Politecnico di Milano, Italy)

### • COM4SWS

- Selection: *Hybrid; Signature (I/O)*
  - **Hybrid match:** Clustering (FarthestFirst, *syntactic* distance) of services in VSM (dim = #SAWSDL attributes); *logic-based* mutual (subclasses of) concept coverage
  - **Ranking:** Based on numeric results of bipartite graph-matching
- Dev: Stefan Schulte et al. (TU Darmstadt, Germany)

Source: Klusch

16



## Track 2: SAWSDL Matchmakers in Brief



### • SAWSDL-MX2

- Selection: *Hybrid, adaptive; Signature*

- **Logic-based match:** Logical I/O concept subsumption
- **Non-logic-based match:** Text similarity; Structural similarity of WSDL groundings
- **Adaptive (offline):** SVM classifier [TS = **10% SAWSDL-TC**] w/ **ranking**

Dev: Patrick Kapahnke, Matthias Klusch (DFKI, Germany)

### • SAWSDL-iMatcher3/1

- Selection: *Hybrid semantic, adaptive; Combined (I/O, Non-functional: Service name)*

- **Logic-based match:** Similarity based on I/O concept subsumption
- **Non-logic-based match:** Text similarity of service names
- **Adaptive (offline):** Linear regression model [TS = **full SAWSDL-TC**] w/ **ranking**

- Dev: Dengping Wei, Avi Bernstein (U Zurich, Switzerland)

Source: Klusch

17

## Performance Evaluation (Binary Relevance)



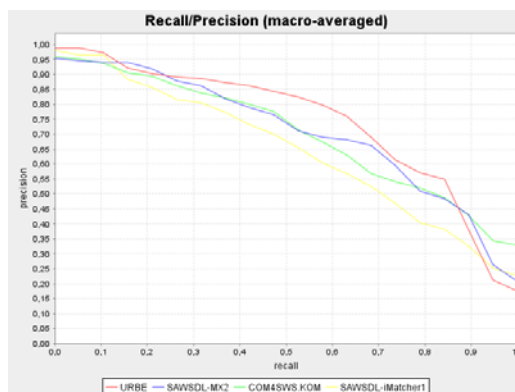
### Average Precision:

1. URBE	.727
2. COM4SWS	.681*
3. SAWSDL-MX2	.679
4. SAWSDL-iMatcher3/1	.635

### Avg Query Response Time (sec):

1. SAWSDL-iMatcher3/1	.75
2. COM4SWS	6.14**
3. SAWSDL-MX2	7.9
4. URBE	19.96

### Macro-averaged Recall/Precision:



\*\* W/o logic-based classification of service ontologies (building of matchmaker ontology) performed belatedly by COM4SWS at first query: else 62.29s ! \* COM4SWS precision: Variant w/o clustering (worse).

Source: Klusch

18

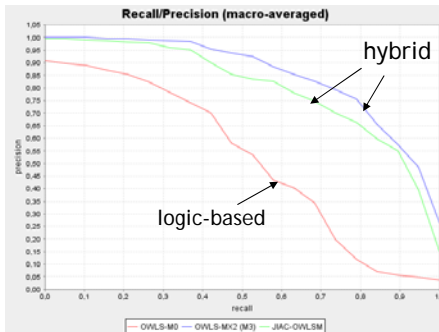
## Some Lessons Learned



### Logic-based vs. Hybrid semantic selection

1. *Integration of logic-based reasoning with text similarity may significantly improve precision at the cost of higher avg query response time.*

Example: Track 1 entries



		AP	AQRT
Logic-based	OWLS-M0	.74	2.66s
Hybrid	OWLS-MX2	.878	3.69s
	JIAC-OWLSM	.814	4.44s
Logic-based	SAWSDL-M0	.419	2.24s
Hybrid	SAWSDL-MX1	.556	2.83s

ALIVE variants (2 logic-based, 1 hybrid) with insignificant differences in precision.

Source: Klusch

19

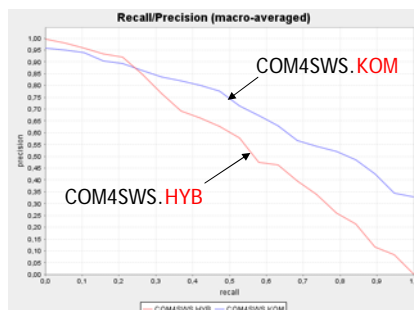
## Some Lessons Learned (2)



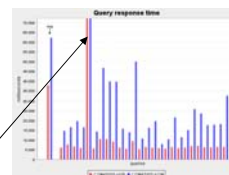
2. *Hybrid semantic matching can be less precise than mere logic-based matching in case of syntactic pre-filtering of services (two-phase vs. integrative hybrid).*

Example: COM4SWS

Hybrid variant \*.HYB prunes search space of *subsequent* logic-based only variant (\*.KOM) by cluster-based prefiltering of services, hence better query response time but at cost of precision



	AP	AQRT
COM4SWS.HYB	.559	6.14s*
COM4SWS.KOM	.681	19.24s*



\* Without its late logical services classification at first query only

Source: Klusch

20

## Some Lessons Learned (3)



### 3. *Adaptive hybrid semantic matchmakers can be competitive wrt both flexibility and performance.*

- >> Adaptive entries performed at least as good as fixed variants of entries in terms of precision (sometimes better: adaptive OWLS-MX3 in Track 1)
- >> Performance results vary depending on used training set:
  - All adaptive entries are off-line trained over different (sub-)sets of test collections
- >> More flexible: Adaptive aggregation renders matchmaking independent from adding or modifications of any test collection or matching filters
- >> All adaptive S3 entries are learning off-line:
  - OWLS-MX3 (SVM), SAWSDL-MX2 (SVM), SAWSDL-iMatcher3/1 (Regression)

Source: Klusch

21

## Some Lessons Learned (4)



### 4. *Majority of semantic service selection bases on signature (I/O) matching*

- First S3 entry featuring PE-matching this year (SPARQLent), plus ongoing work elsewhere (e.g. iSeM 1.0).
- Problem: No test collection including service PEs available!

### 5. *Query response times of matchmakers largely differ*

- Entries that use RDF triple stores and relational databases perform much faster than those with in-memory storage of logic-based reasoners.
- Non-logic-based semantic selection with text index-based retrieval fastest.

### 6. *Graded relevance sets appear to enable higher precision*

- All S3 track 1 entries performed more precise over OWLS-TC3 with graded relevance sets (Discounted cumulative gain for cut-off n=100) i.p. for top positions of rankings
- Graded relevance sets will be included in upcoming SAWSDL-TC2

Source: Klusch

22

## Track 3: Cross-Evaluation



- **Specific Domain Test Collection: Jena Geography Dataset JGD**
  - Full set consists of 201 geoservices (WSDL, REST-based), 10 queries, graded relevance.
  - **Initial test set JGD50: Only 50 services, 9 queries.**
  - ! **Services provided by S3 organizers, semantic annotations by participants.**
  - ! **Each JGD50 service semantically annotated in different ways:**
    - >> JGD50-NL-Tags: **Monolithic text; Folksonomy-based tagging** -- for Themis-S, WSColab
    - >> JGD50-SAWSDL: **SAWSDL** -- for Track-2 entries
    - >> JGD50-OCML-LISP: **LISP syntax with OCML semantics** -- for IRS-III
- **Comparative performance evaluation over JGD50**
  - Retrieval performance (Q, nDCG; AQR); Evaluation Tool: SME2 v2.1
  - Efforts of service annotation: *N/A (no sufficient feedback from annotators)*

## Track 3: Selection Tools in Brief



For JGD50-NL-Tags: Services/queries summarized into text or tagged

- **Themis-S**
  - Selection: *Non-logic-based; Monolithic (Text)*
    - **Non-logic-based match:** Text similarity between bags of extracted (via WordNet) and weighted concepts in service/query text (docs) over enhanced Topic-based Vector Space Model (eTVSM) with respective **ranking**
  - Dev: Oliver Müller (U Münster, Germany)
- **WSColab**
  - Selection: *non-logic; Combined (tags for I/O, non-functional parameters/"behavior")*
    - **Non-logic-based:** Text similarity of tags (TFIDF/cosine)
    - **Ranking:** Tag text similarity (returns only matching services)
  - All results for WSColab averaged over five different query wordings.
  - Dev: Maciej Gawinecki (U Modena & Reggio Emilia, Italy)

## Track 3: Selection Tools in Brief



For JGD50-OCML-LISP: Services/goals described in OCML-LISP

- **IRS-III**

- Selection: *Logic-based; Signature*
  - Logic-based match: OCML rule-based relational matches between I/O concepts
  - Ranking: Number of I/O concept matches; returns only matching services
- Dev: Liliana Cabral+ (Open University, UK)

For JGD50-SAWSDL: Service descriptions in semantically annotated WSDL 1.1

- **SAWSDL-MX1** (hybrid; signature)
- **SAWSDL-MX2** (hybrid, adaptive; signature); Training Set = 20% of JGD50-SAWSDL
- **SAWSDL-iMatcher3/1** (hybrid, adaptive; combined); Training Set = SAWSDL-TC1


Evaluation over full JGD50-SAWSDL (Test set otherwise too small).

- URBE: NullPointerExceptions during JGD50-SAWSDL service parsing
- COM4SWS: Supports only WSDL 2.0 (No JGD50-SAWSDL with WSDL 2.0 available)


## Performance Evaluation (Binary Relevance)



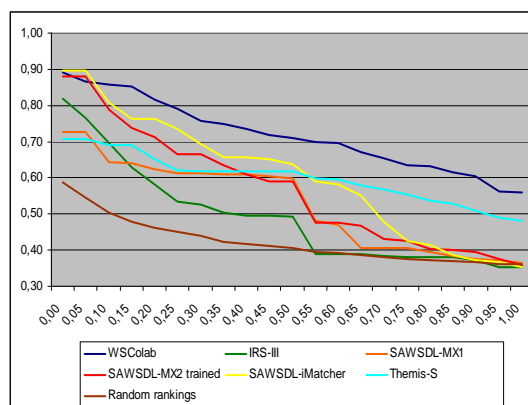
### Average Precision

1. **WScolab**  **0.54**
2. SAWSDL-iMatcher 0.53
3. Themis-S 0.48
4. SAWSDL-MX2 0.45
5. IRS-III, SAWSDL-MX1 0.41

### Avg. Query Response Time (sec)

1. **WScolab**  **~ 0 ms**
2. SAWSDL-iMatcher .170
3. SAWSDL-MX1 .253
4. SAWSDL-MX2 .784
5. Themis-S 2.043
6. IRS-III 2.826

### Macro-averaged Recall/Precision:



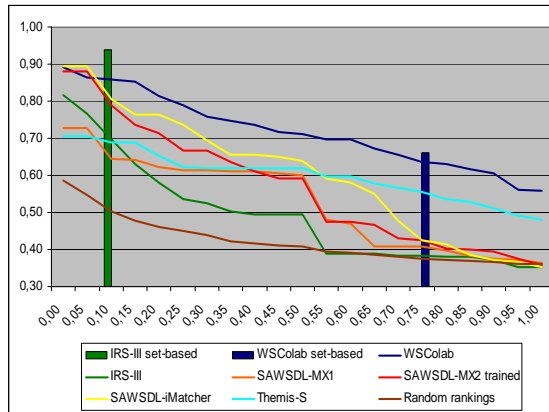
(Relaxed definition of binary relevance: JGD Binary7)  
Average Precision: Average over JGD Binary1 to 8)

## Limitation of Ranking-Based Evaluation



Set-based **binary** matchmakers (e.g. WSColab, IRS-III) **not standard comparable** with others:

- Return classical answer set with „matching“ services only:  
*No rank list of all services.*
- *Random ranking of „non-matching“ services in rank list of all services.*

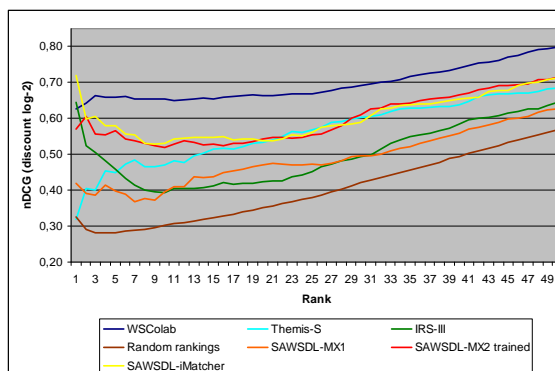


(Relaxed definition of binary relevance: JGD Binary7)

## Performance Evaluation (Graded Relevance)



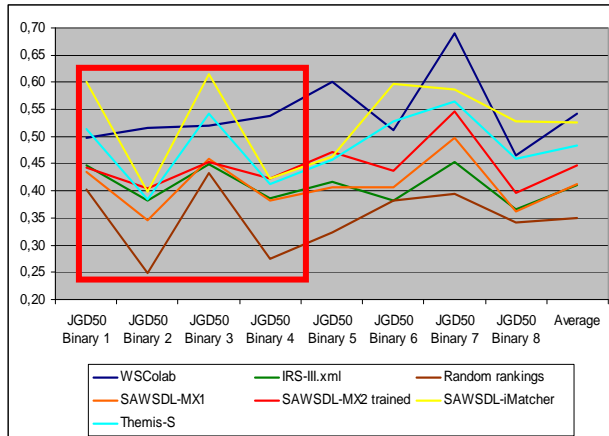
Precision	$Q_1$	$nDCG_{50}$
1. WSColab	.73	.80
2. iMatcher	.66	.71
SAWSDL-MX2	.65	.71
4. Themis-S	.66	.68
5. IRS-III	.60	.65
SAWSDL-MX1	.61	.63
7. Random	.54	.57



## Limitations of Binary Relevance



Average precision is **sensitive to different definitions of binary relevance** for JGD (different sets of relevance grades for „relevant“/„not relevant“)



- JGD Binary1: PossEqual, PossMatch, PossCompatible
- JGD Binary2: PossEqual, Partial, PossCompatible
- JGD Binary3: Approximate, PossMatch, PossCompatible
- JGD Binary4: Approximate, Partial, PossCompatible
- JGD Binary5: PossEqual, Partial, Incompatible
- JGD Binary6: Approximate, PossMatch, Incompatible
- JGD Binary7: Approximate, Partial, Incompatible
- JGD Binary8: PossEqual, PossMatch, Incompatible

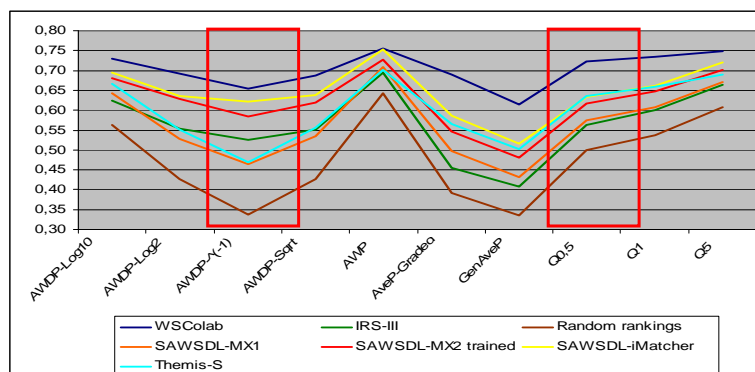
R/P can be very instable for queries with only few relevant services.

## Performance Evaluation (Graded Relevance)



Performance over different graded relevance measures is relatively stable (*Stable*: No change of matchmaker ranking for different measures)

nDCG variants (AWDP-\*) *sometimes* rank *differently* than e.g. Q measures with integrated AP



## Track 3: Some Lessons Learned



1. *Bottleneck of describing semantic services*
  - >> Annotation of JGD200 overcharged participants → fall back to JGD50 but this is clearly too small of a collection
  - >> More active participation in test collection building required (e.g. joint project funding, online portal(s), special TREC-like developer conference, ..), de-facto standards OWLS-TC and SAWSDL-TC to start with.
2. *Non-logic-based selection tools performed as good as logic-based ones*
  - >> Where are the logical or hybrid IOPE matchmakers that can do better?
  - >> What were the most hard implementation problems to cope with?
3. *Evaluation for graded relevance much more stable than for binary relevance*
  - >> Further test collection building should include graded relevance sets

## Outlook on 4th S3 Contest in 2010



- **New** semantic service matchmaker entries already confirmed ... ☺
  - **MOD** (A-STAR, Singapore)
  - **iSeM** (DFKI, D) - hybrid, adaptive; Combined/IOPE
  - **SAWSDL.SAG** (TU Darmstadt/Software AG, D)
- **Location** of final presentation/discussion of results **TBD**
- **Improved test collections** OWLS-TC4, SAWSDL-TC2. What about WSML-TC?
- Continuation of **cross-evaluation** track **TBD**

**NEW: SAWSDL-TC 2.0**  
with *new* geoservices domain (JGD50-SAWSDL),  
*more* SAWSDL services and queries, and *additional* graded relevance sets  
@semwebcentral.org, December 2009



## References/Contacts



### Track 1:

- ALIVE - jim.andreou@gmail.com
- JIAC-OWLSM - nils.masuch@dai-labor.de
- Opossum - erantoch@gmail.com  
E. Toch, A. Gal, I. Reinhartz-Berger, D. Dori: A Semantic Approach to Approximate Service Retrieval  
ACM Transactions on Internet Technology (TOIT), 8(1), 2007.
- OWLS-MX3 - klusch@dfki.de  
Klusch, M.; Kapahnke, P. (2009): OWLS-MX3: An Adaptive Hybrid Semantic Service Matchmaker for OWL-S.  
CEUR Proceedings of 3rd International Workshop on Semantic Matchmaking and Resource Retrieval (SMR2) at  
ISWC, Washington, USA
- OWLS-MX2 - klusch@dfki.de  
Klusch, M.; Fries, B.; Sycara, K. (2009): OWLS-MX: A Hybrid Semantic Web  
Service Matchmaker for OWL-S Services. Web Semantics, 7(2), Elsevier
- OWLS-iMatcher - dengping@ifi.uzh.ch  
Christoph Kiefer, Abraham Bernstein. The Creation and Evaluation of iSPARQL Strategies for Matchmaking.  
Proceedings of the 5th European Semantic Web Conference (ESWC). Tenerife, Spain, June 1-5, 2008.
- SPARQLent - marco.sbodio@gmail.com

Source: Klusch

33

## References/Contacts



### Track 2:

- URBE - plebani@elet.polimi.it
- SAWSDL-MX - klusch@dfki.de  
Klusch, M.; Kapahnke, P. (2008): Semantic Web Service Selection with SAWSDL-MX. CEUR Proceedings of 2nd  
International Workshop on Semantic Matchmaking and Resource Retrieval (SMR2) at ISWC, Karlsruhe, Germany
- SAWSDL-MX2 - klusch@dfki.de  
Klusch, M.; Kapahnke, P.; Zinnikus, I. (2009): SAWSDL-MX2: A Machine-Learning Approach for Integrating Semantic Web  
Service Matchmaking Variants. Proceedings of IEEE 7th International Conference on Web Services (ICWS), LA, USA
- COM4SWS - schulte@kom.tu-darmstadt.de
- SAWSDL-iMatcher - dengping@ifi.uzh.ch

### Track 3:

- IRS-III - L.S.Cabral@open.ac.uk  
Dietze, S., Benn, N., Domingue, J., Conconi, A., and Cattaneo, F. (2009) Two-Fold Semantic Web Service Matchmaking -  
Applying Ontology Mapping for Service Discovery, 4th Asian Semantic Web Conference, Shanghai, China
- Themis-S - oliver.mueller@ercis.uni-muenster.de
- WSColab - mgawinecki@gmail.com; <http://mars.ing.unimo.it/wscolab/new.php>

Source: Klusch

34



... Thanks for your attention !

Any **QUESTIONS?**



... Next year with *your brand new ultra mega beat'em all matchmaker !?* 😊

Source: Klusch

35