



# Instance-based Matching of Large Ontologies

Erhard Rahm

<http://dbs.uni-leipzig.de>

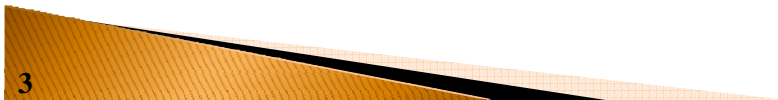
June 3, 2009

## Agenda

- ▶ Ontologies
- ▶ Ontology Matching
  - Problem
  - Match techniques and prototypes (e.g., GLUE)
- ▶ Instance-based matching in COMA++
  - Constraint- / Content-based Matching
  - Matching web directories
- ▶ Matching by Instance overlap
  - Similarity measures
  - Evaluation: Product catalogs, biomedical ontologies
- ▶ Stability of ontology mappings
- ▶ Conclusions

# Ontologies: Usage Forms

- ▶ Support a shared understanding of terms/concepts in a domain
  - Annotation of data instances by terms/concepts of an ontology
- ▶ Semantically organize information of a domain
  - Find data instances based on concepts (queries, navigation)
- ▶ Support data integration
  - e.g. by mapping data sources to shared ontology
- ▶ Sample ontologies
  - Product catalogs of companies, e.g. online shops
  - Web directories
  - Biomedical ontologies



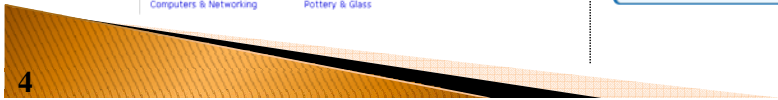
3

# Product Catalogs

- ▶ Hierarchical categorization of products
- ▶ Instances: product descriptions
- ▶ Often very large: ten thousands categories, millions of products






The screenshot displays a product catalog interface with the following elements:

- ICEcat.biz** logo with the tagline "COOL IN CATALOGUES".
- amazon.com** and **ebay** logos.
- A navigation menu with "Shop All Departments" and a list of categories: Books, Movies, Music & Games, Digital Downloads, Kindle, Computers & Office, Electronics, Home & Garden, Grocery, Health & Beauty, Toys, Kids & Baby, Apparel, Shoes & Jewelry, Sports & Outdoors, and Tools, Auto & Industrial.
- A detailed list of product categories on the left, including Antiques, Art, Baby, Books, Business & Industrial, Cameras & Photo, Cars, Boats, Vehicles & Parts, Cell Phones & PDAs, Clothing, Shoes & Accessories, Coins & Paper Money, Collectibles, Computers & Networking, Crafts, DVDs & Movies, Dolls & Bears, Electronics, Entertainment Memorabilia, Gift Certificates, Health & Beauty, Home & Garden, Jewelry & Watches, Music, Musical Instruments, Pottery & Glass, Real Estate, Specialty Services, Sporting Goods, Sports Mem. Cards & Fan Shop, Stamps, Tickets, Toys & Hobbies, Travel, Video Games, and Everthing Else.
- Four highlighted product categories on the right:
  - kitchen & houseware**: refrigerators, (freestanding) cookers, vacuum cleaners, washing machines.
  - office equipment, supplies & accessories**: paper cutters, laminators, paper shredders, paper perforators, binding machines.
  - personal care**: men's shavers, hairdryers, electric toothbrushes, solaria.
  - clothing**: women's clothing, men's clothing.



4

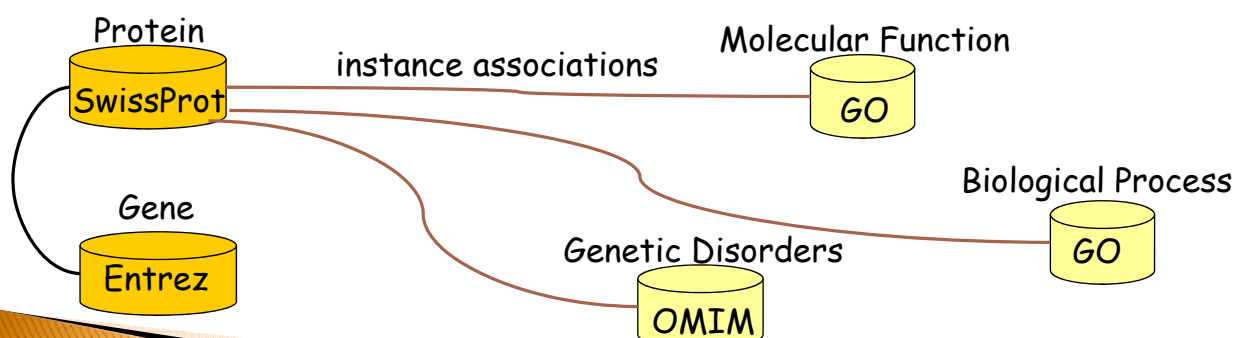
# Web Directories

- ▶ Categorization of websites
- ▶ Instances: website descriptions (URL, name, content description)
- ▶ Manual vs. automated category assignment of instances
- ▶ General lists or specialized (per region, topic, etc.), e.g.
  - Yahoo! Directory 
  - Dmoz – Open Directory Project (ODP) 
  - Google Directory – based on Dmoz 
  - Business.com 
  - Vfunk: Global Dance Music Directory 

5

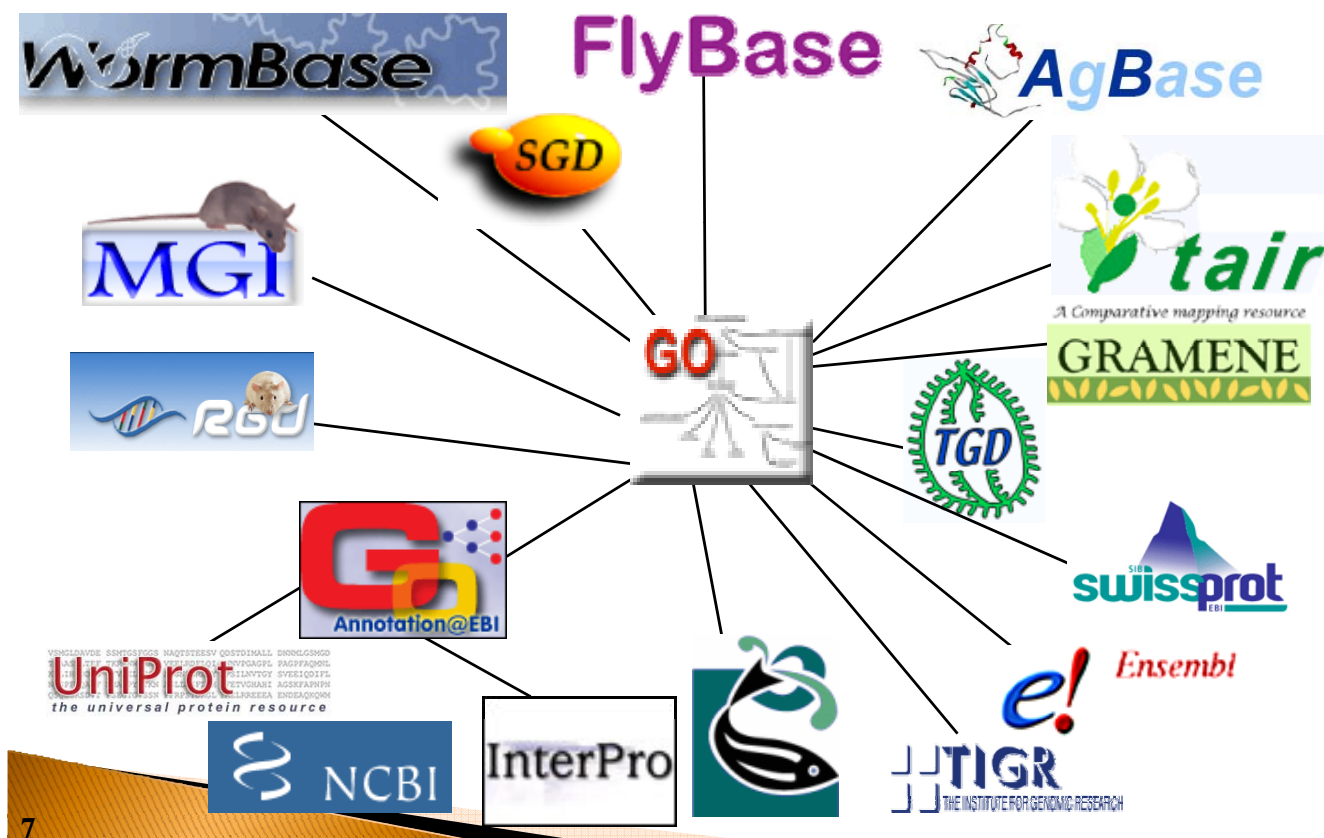
# Life Science Ontologies

- ▶ Many ontologies for different disciplines, e.g.
  - Molecular Biology, Anatomy, Health etc.
- ▶ Largest ontologies (> 10,000 concepts), e.g., Gene Ontology (GO), NCI Thesaurus
- ▶ Ontologies used to annotate genes and proteins
  - ✓ Support for “functional” data analysis
- ▶ Instances: annotated objects; separate from ontology



6

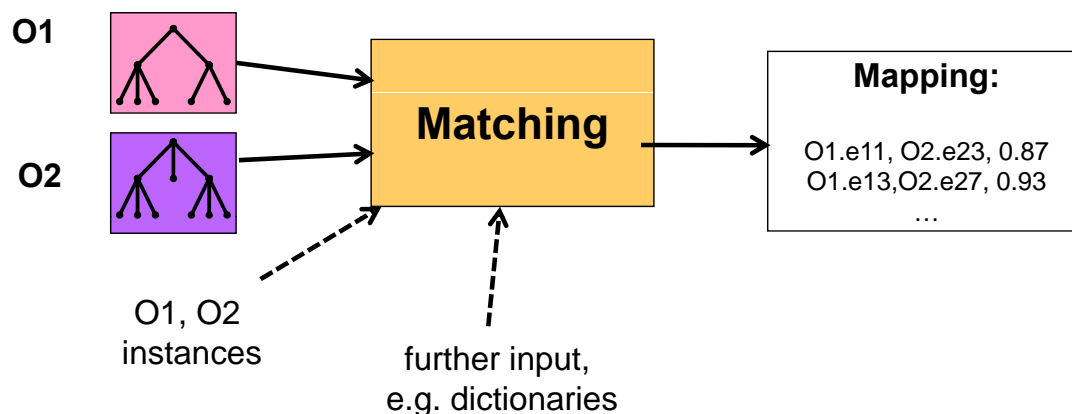
# Example: Widespread Usage of GO



## Assumed Ontology Model

- ▶ Focus on practically used ontologies
- ▶ Ontology O consists of a set of *concepts/categories* interconnected by *relationships* (e.g. of type „is-a“ or „part-of“). O is represented by a **DAG** and has a designated root concept.
  - Concepts have *attributes*, e.g. Id, Name, Description
  - Concepts may have associated *instances*
- ▶ Ontologies may be versioned
- ▶ Instances
  - May be managed together with ontology or independently
  - May be associated to several concepts
  - May have heterogeneous schemas, even per concept

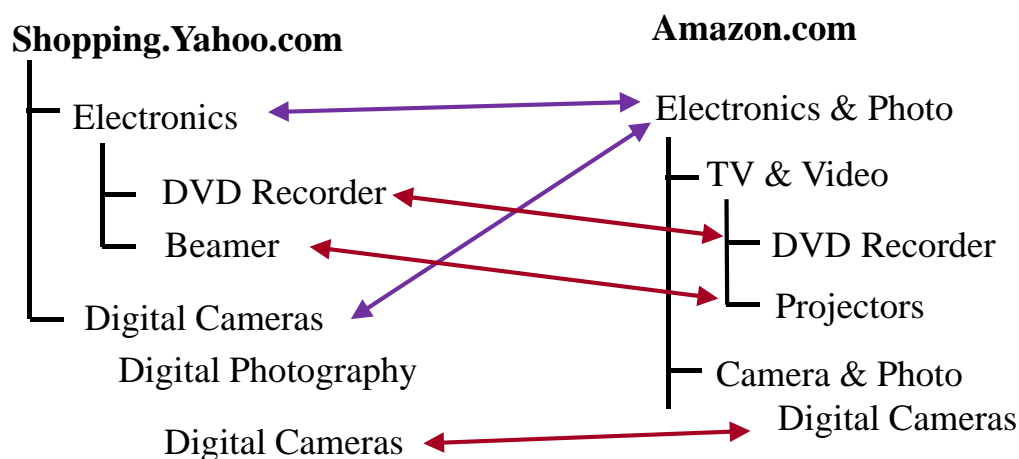
# Ontology Matching / Alignment



- ▶ Process of identifying semantic **correspondences** between 2 ontologies
  - Result: **ontology mapping**
  - Mostly equivalence mappings: correspondences specify equivalent ontology concepts
- ▶ Variation of schema matching problem

9

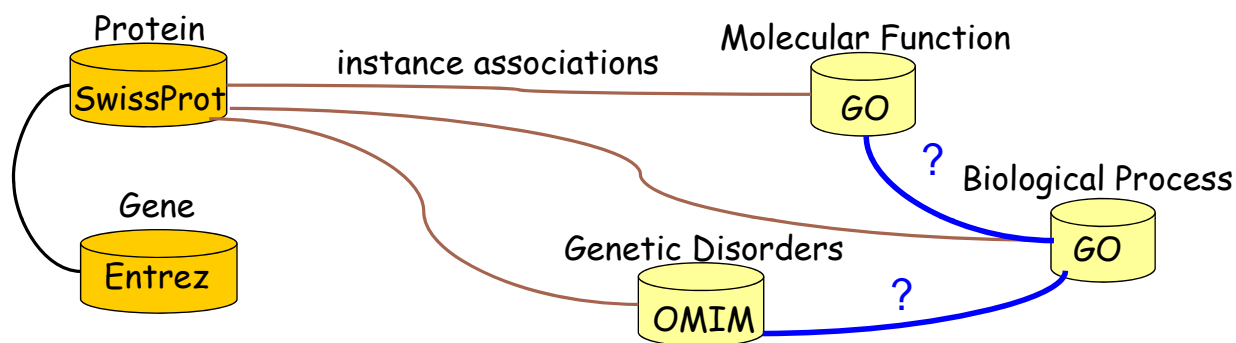
# Matching of Product Catalogs



- ▶ Ontology mappings useful for
  - Improving query results, e.g. to find specific products
  - Advanced (cross-site) product recommendations
  - Automatic categorization of products in different catalogs
  - Merging catalogs

10

# Matching Life Science Ontologies



- ▶ Ontology mappings useful for
  - ▶ Improved analysis
    - ▶ Answering questions such as “Which Molecular Functions are involved in which Biological Processes?”
  - ▶ Validation (curation) and recommendation of instance associations
  - ▶ Ontology merge or curation, e.g. to reduce overlap between ontologies

11

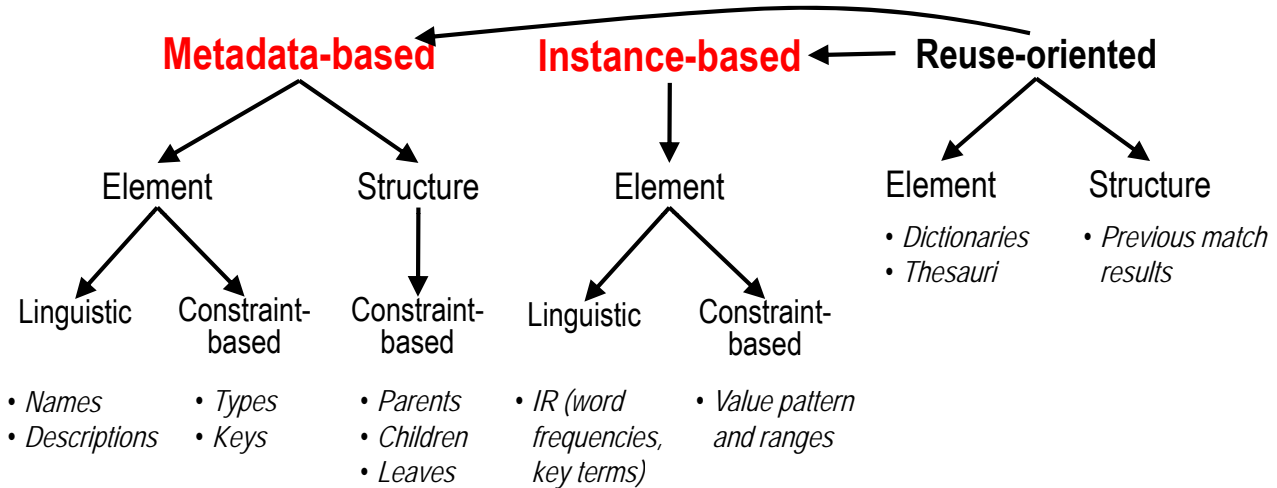
## Ontology Matching is challenging

- ▶ High degree of semantic heterogeneity in independently developed ontologies
- ▶ **Syntactic differences**
  - Different models and languages
- ▶ **Structural differences**
  - Different is-a and part-of hierarchies
  - Overlapping categories
- ▶ **Semantic differences**
  - Naming ambiguities and conflicts
- ▶ **Modeling errors / inconsistencies**
- ▶ **Instance / content differences**
  - ▶ Different scope
  - ▶ Heterogeneous instance representations
- ▶ Fully automatic, generic solutions ?

12



# Automatic Match Techniques\*

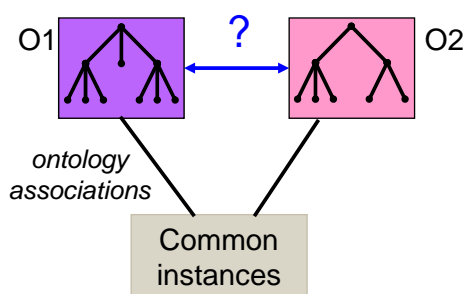


- ▶ **Matcher combinations**
  - ▶ Hybrid matchers
  - ▶ Composite matchers

\* Rahm, E., P.A. Bernstein: *A Survey of Approaches to Automatic Schema Matching*. VLDB Journal 10(4), 2001

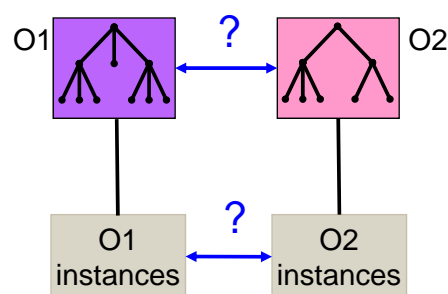
## Instance-based matching

- ▶ semantics of a category may be better expressed by the instances associated to category than by metadata (e.g. concept name, description)
  - Categories with most similar instances should match
- ▶ Main problem: Availability of (shared/similar) instances for most/all concepts
- ▶ **Common cases:**



a) **Common instances (separate from ontologies)**

Example: Documents/Objects annotated by O1, O2 terms / concepts



b) **Ontology-specific instances**

b1) with shared instances  
b2) without shared (but similar) instances

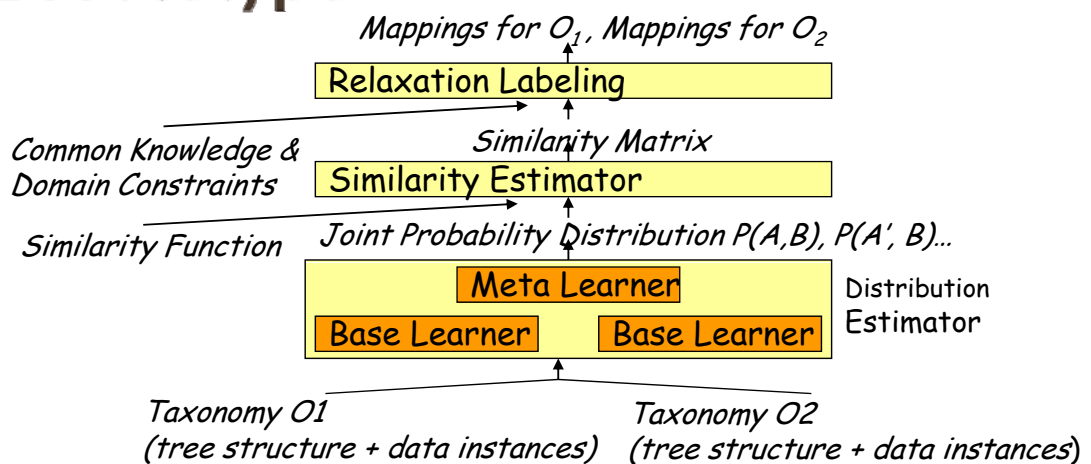
# Match Prototypes

- ▶ Many prototypes for schema or ontology matching \*
- ▶ Instance-based schema matching (XML, relational)
  - SEMINT
  - LSD
  - Clio
  - iMap
  - Dumas
- ▶ Instance-based ontology matching (OWL)
  - GLUE, U of Washington
  - COMA++, U Leipzig (supports schema + ont. matching)
  - FOAM / QOM, U Karlsruhe
  - Sambo, Linköping U, Sweden
  - Falcon-AO, South East U, China
  - RiMOM, Tsinghua U, China

15

\* Euzenat/Shvaiko: Ontology matching. Springer 2007

## GLUE Prototype\*



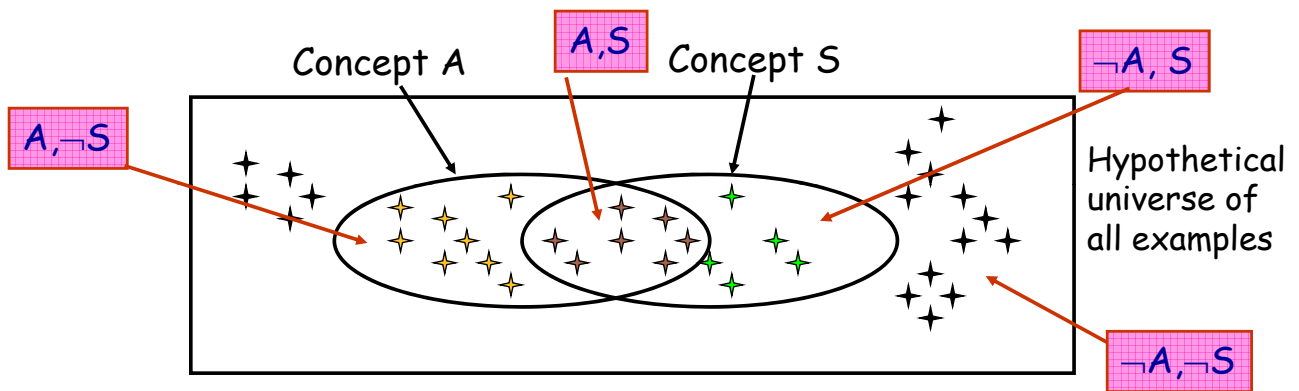
- ▶ Use of machine learning to find ontology mappings
- ▶ Base learners use concept names + data instances (description)
- ▶ Similarity measures computed from “joint probability distribution” of concepts
- ▶ Evaluation on comparatively small ontologies: 3 match tasks, per ontology: 34–331 concepts, 6–30 non-leaf concepts, 1500–14000 instances, 34–236 correspondences

16

\* Doan, AH; et al: *Learning to Match Ontologies on the Semantic Web*. VLDB Journal, 12(4):303-319, 2003



# GLUE: Concept Similarity



$$\text{Sim}(\text{Concept A, Concept S}) = \frac{P(A \cap S)}{P(A \cup S)} = \frac{P(A, S)}{P(A, \neg S) + P(A, S) + P(\neg A, S)}$$

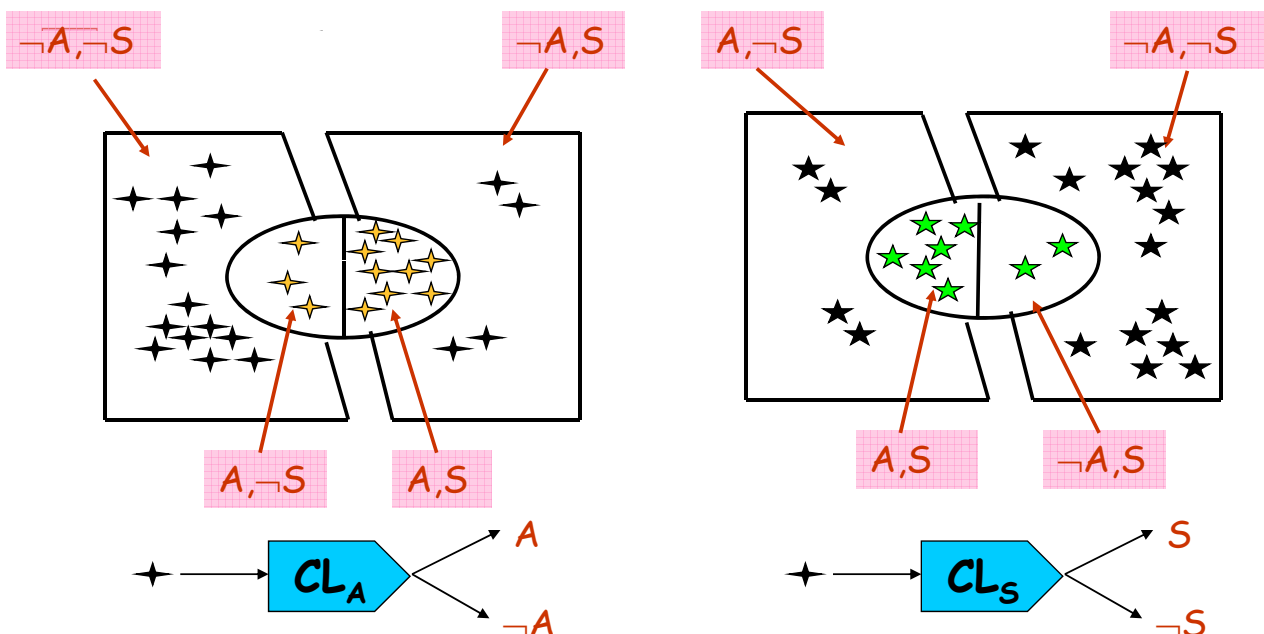
[Jaccard]

Joint Probability Distribution:  $P(A, S), P(\neg A, S), P(A, \neg S), P(\neg A, \neg S)$

different similarity measures usable based on JPD

# GLUE: Machine Learning for Computing Similarities

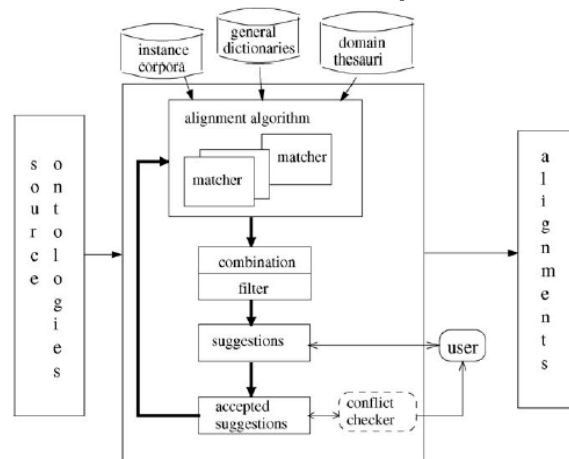
Mutual use of trained classifiers to determine instance-concept associations (requires no shared but only similar instances)



JPD estimated by counting the sizes of the partitions

# SAMBO Prototype\*

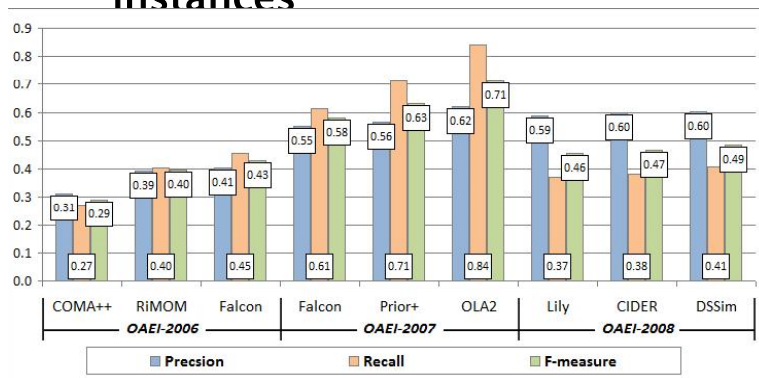
- ▶ System for aligning and merging biomedical ontologies
- ▶ Framework to find similar concepts in overlapping OWL ontologies for alignment and merge tasks
  - ▶ Combined use of different matchers and auxiliary information
    - ▶ Linguistic, structure-based, constraint-based
    - ▶ Instance-based matching
      - Based on texts (e.g., papers)
      - Two concepts are similar if a document describes both concepts
  - ▶ description logic reasoner checks results for ontology consistency and cycles



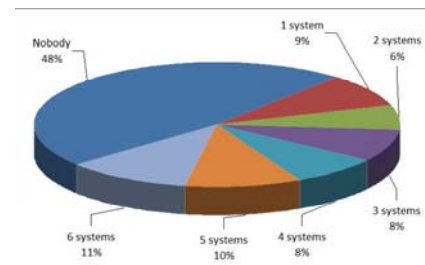
\* Lambrix, P; Tan, H.: SAMBO – A system for aligning and merging biomedical ontologies. Journal of Web Semantics, 4(3):196-206, 2006

# OAEI\*: Directory Results

- ▶ Dataset
  - extracted from Google, Yahoo and Looksmart web directory
  - More than 4,500 simple node matching tasks, no instances



Comparison of matching quality results (top-3 systems of each year)



In 2008 the systems together did not manage to discover 48% of the total number of positive correspondences

- OAEI (Ontology Alignment Evaluation Initiative) Alignment Contest, <http://oaei.ontologymatching.org>

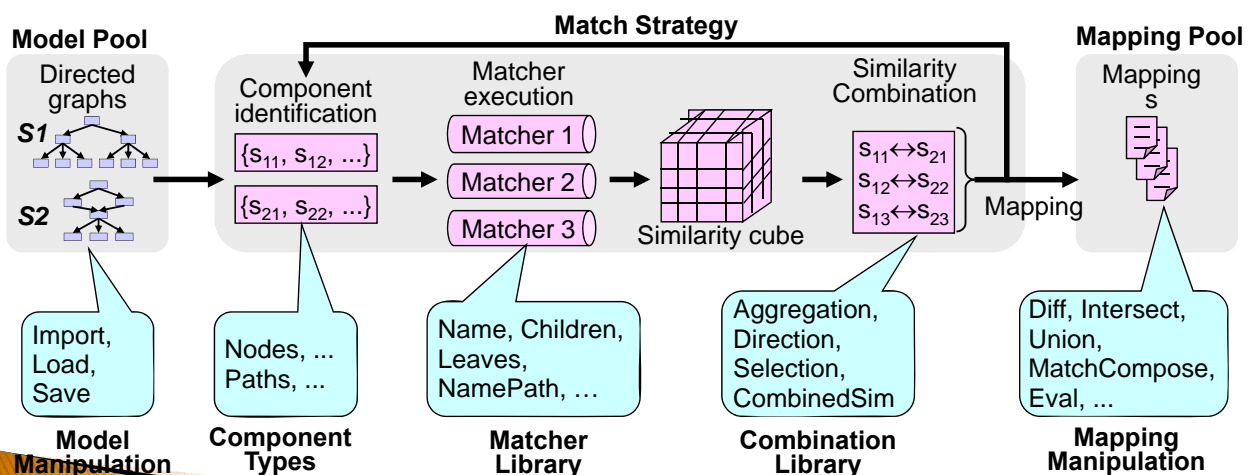
# Agenda

- ▶ Ontologies
- ▶ Ontology Matching
  - Problem
  - Match techniques and prototypes (e.g., GLUE)
- ▶ Instance-based matching in COMA++
  - Constraint- / Content-based Matching
  - Matching web directories
- ▶ Matching by Instance overlap
  - Similarity measures
  - Evaluation: Product catalogs, biomedical ontologies
- ▶ Stability of ontology mappings
- ▶ Conclusions

21



- ▶ Extends previous COMA prototype (VLDB2002)
- ▶ Matching of XML & rel. Schemas and OWL ontologies
- ▶ Several match strategies: Parallel (composite) and sequential matching; Fragment-based matching for large schemas; Reuse of previous match results



22

# COMA++ GUI Overview

Repository (persistent) & Workspace (in-memory)

Current Mapping

Domains

Schemas/  
Ontologies

Mappings

Schema/  
mapping info

The screenshot shows the COMA++ GUI with the following components:

- Repository/Workspace:** A sidebar on the left containing a tree view of Domains (e.g., OAEI Ontology Alignment Contest), Schemas (e.g., Apertum), and Mappings (e.g., Excel\_Apertum, Excel\_Noris). A table at the bottom provides details for the selected mapping.
- Mapping View:** The main area displays a mapping between 'Excel (XDR)' (Source Schema) and 'Noris (XDR)' (Target Schema). It shows hierarchical structures with nodes like PurchaseOrder, Contact, Address, and DeliverTo, connected by lines representing correspondences.
- Match Mapping View:** A top bar with a color scale from 0.0 to 1.0 and a 'Match Mapping View' button.
- Status Bar:** A bottom bar with a search field and the text 'Select a node to display its correspondences'.

# Matcher & Match Strategies

Configuration of matcher

The 'Existing Matchers' dialog box displays a table of matchers with the following categories:

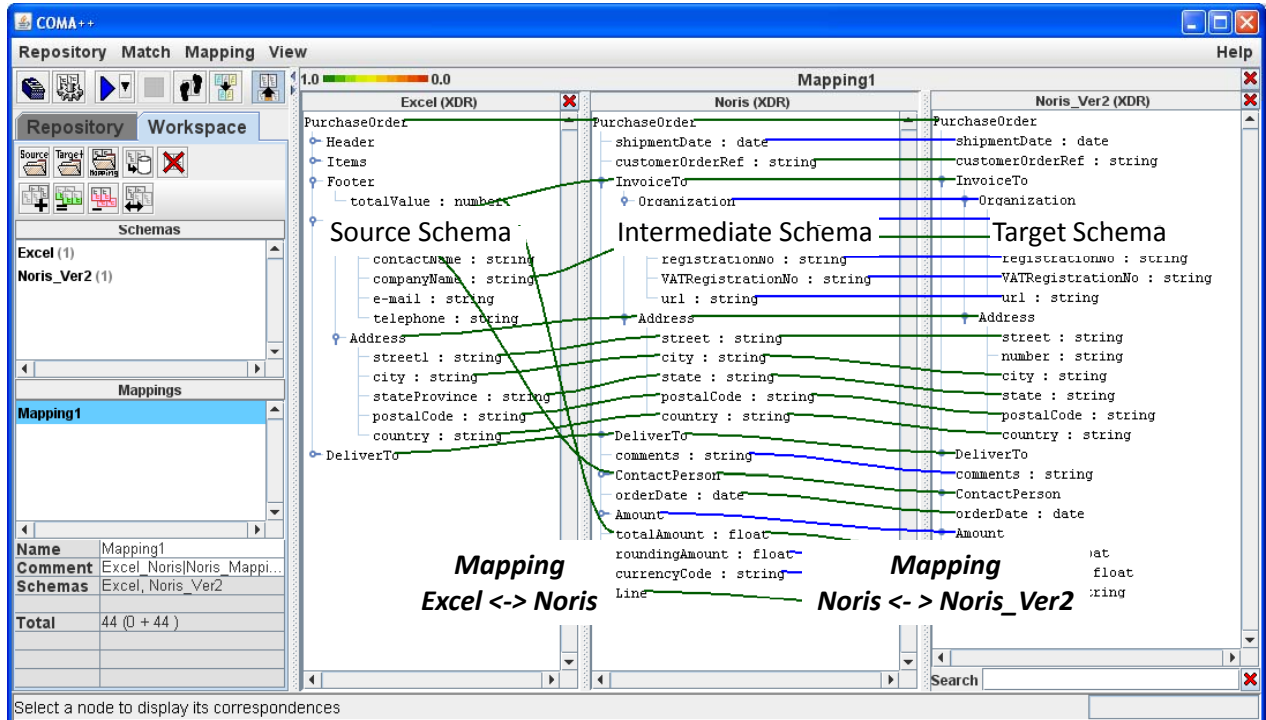
- Metadata-based:** CHILDREN, COMA, COMA\_OPT, COMMENT, CONTEXTS, DATATYPE, INSTANCES, LEAVES, NAME, NAMESTAT, NAMESTYPE, NODES, PARENTS, PATH, SIBLINGS, STATISTICS, STATTYPEINST.
- Reuse-based:** REUSE.
- Instance-based:** INST\_ALL\_CONTENT, INST\_CONSTRAINT, INST\_DIRECT\_CONTENT.
- User-programmed:** Stem dg MaxDelta001, Stem dg MaxN0.

Configuration of match strategies

The 'Configure Strategy (Advanced)' dialog box shows the following configuration options:

- Change to Basic:** A button to switch to a simpler configuration mode.
- Context:** Selected as '(COMA default strategy)'. Context Matcher is set to 'COMA'. FilteredContext is checked with Node Matcher set to 'NODES'.
- Nodes:** Node Matcher is set to 'NAMETYPE'.
- Reuse:** Option to 'use existing mapping paths'.
- Fragment:** Fragment Identification is set to 'CURSCHEMA'. Match Strategy is set to 'FilteredContext'.

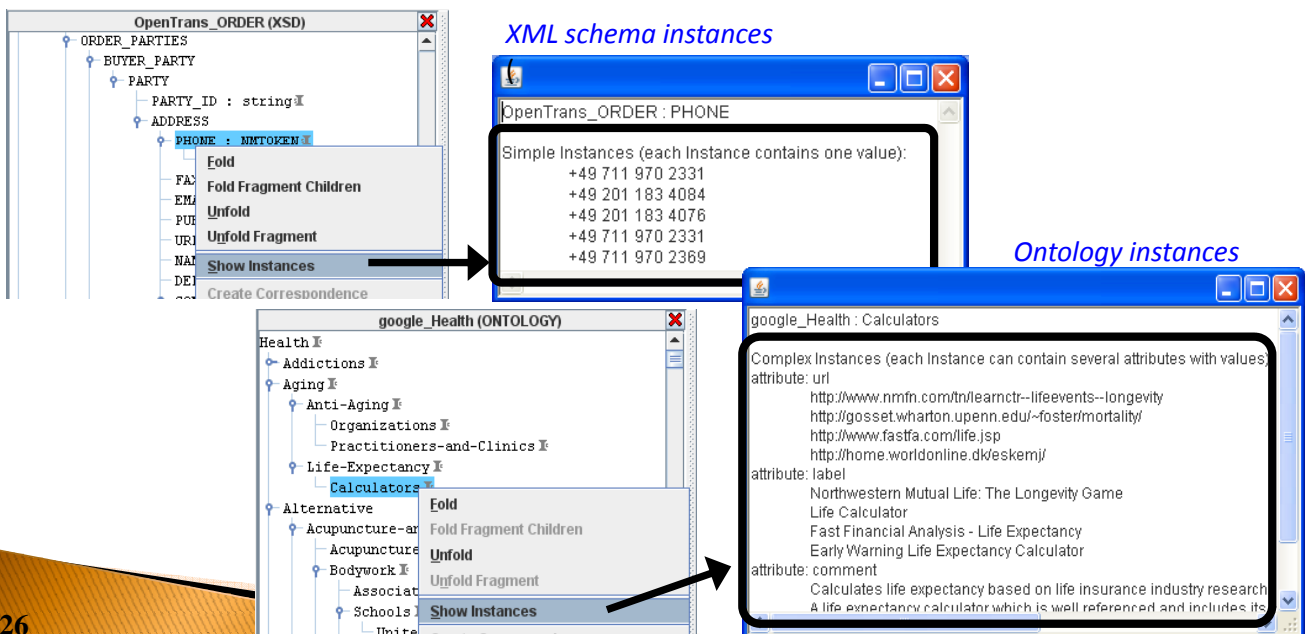
# Reuse of Mappings



25

## Instance-based Matching in COMA++

- ▶ Instance matchers introduced in 2006
  - ▶ Constraint-based matching
  - ▶ Content-based matching: 2 variations
- ▶ Coma++ maintains *instance value set* per element

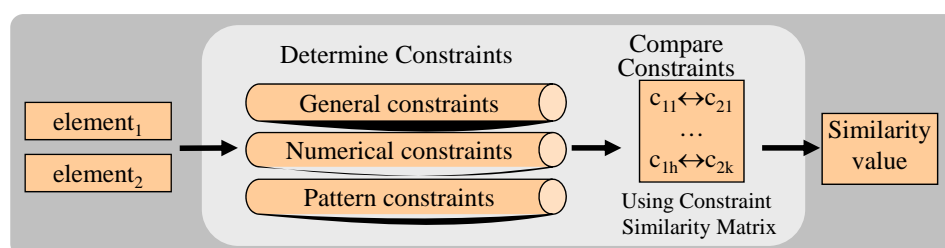


26

# Constraint-based Matching

- ▶ Instance constraints are assigned to schema elements
  - **General constraints:** always applicable  
*Example:* average length and used characters (letters, numeral, special char.)
  - **Numerical constraints:** for numerical instance values  
*Example:* positive or negative, integer or float
  - **Pattern constraints:**  
*Example:* Email and URL
- ▶ Use of constraint similarity matrix to determine element similarity (like data type matching)
- ▶ Simple and efficient approach
  - Effectiveness depends on availability of constrained value ranges / pattern
- ▶ Approach does not require shared instances

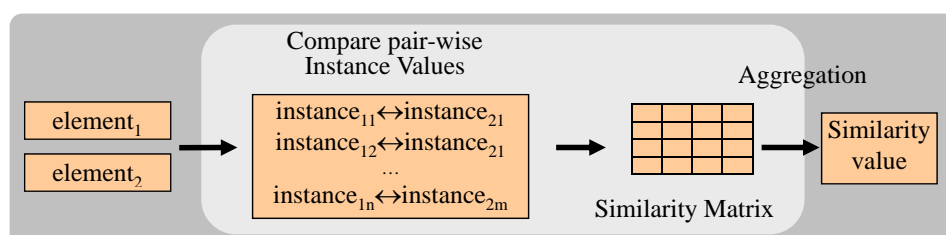
“My@email.com” vs.  
“Your@email.org”



27

# Content-based Matching

- ▶ 2 variations
  - *Value Matching:* pairwise similarity comparison of instance values
  - *Document (value set) matching:* combine all instances into a virtual document and compare documents
  - Both approaches do not require shared instances
- ▶ Value matching
  - Use any similarity measure for pairwise value comparison
  - Aggregate individual similarity values (similarity matrix) into a combined concept similarity (e.g., based on Dice)

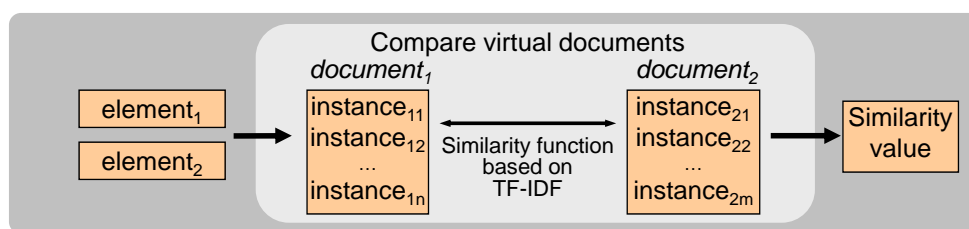


28



# Content-based Matching 2

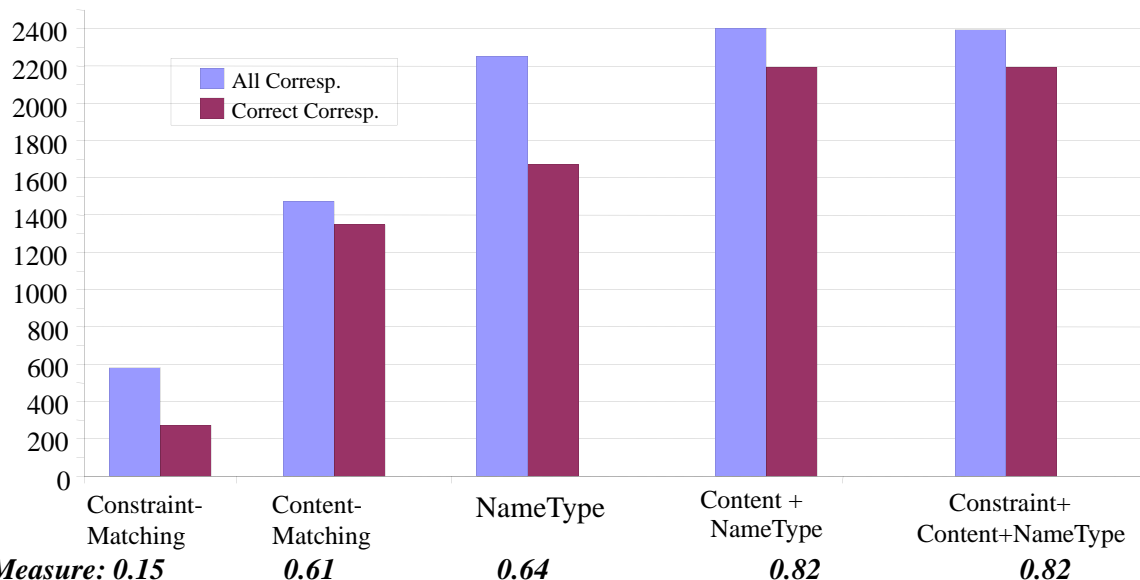
- ▶ Document matching
  - 1 instance document per category or selected string category attribute (e.g. description)
  - Document comparison based on TF-IDF to focus on most significant terms
- ▶ Two options to deal with multiple string attributes
  - All values for these attributes are handled as one virtual document
  - Independent matching per attribute and aggregation of the similarity values



29

## Evaluation for OAEI benchmark test

- ▶ 39 of 51 test cases based on instances
- ▶ 2966 correspondences in reference alignment

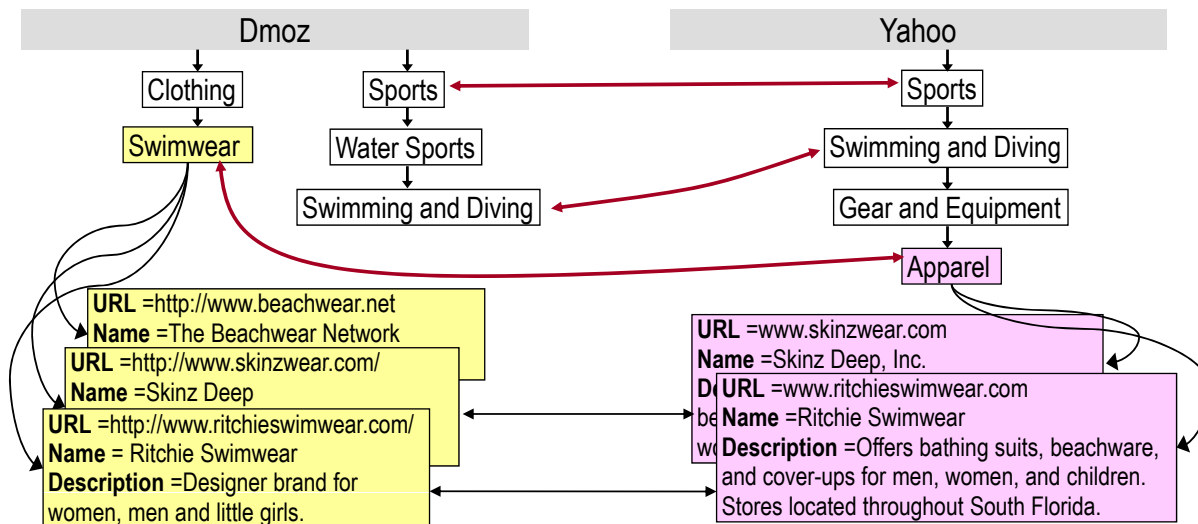


30

# Use case: Web Directory Matching \*

- ▶ Instance-based matching between 4 web directories, limited to online shops

	Dmoz	Google	Web	Yahoo
#Categories	746	728	418	3,234
#Direct instances	15,304	15,082	13,673	34,949



\* Massmann, S., Rahm, E.: *Evaluating . Evaluating Instance-based Matching of Web Directories*. Proc. WebDB 2008

## Web Directory Matching

- ▶ Instances are shop websites
- ▶ Instance-based matching on 3 attributes: shop URL, name, description
  - Use of directly and indirectly associated instances
- ▶ *URL matcher* based on value matching
  - After URL preprocessing, equal URLs are needed (same shops in different directories) to find matching categories



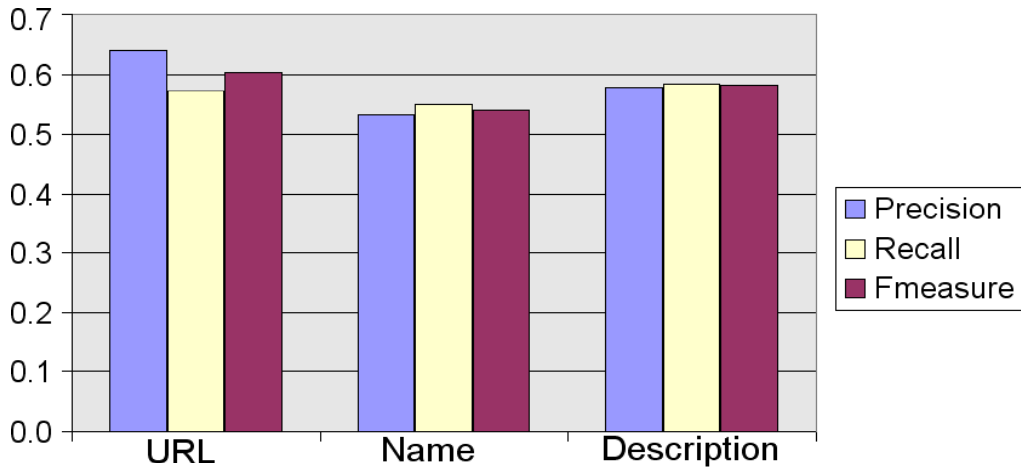
- ▶ *Name matcher* based on value matching
- ▶ *Description matcher* based on document matching
- ▶ Name / description matching do not need shared instances

# Results of instance-based Matchers

- ▶ Six match tasks → six reference mappings (manually created)

	Dmoz ↔ Google	Dmoz ↔ Web	Dmoz ↔ Yahoo	Google ↔ Web	Google ↔ Yahoo	Web ↔ Yahoo
# Corresp	729	218	436	211	416	235

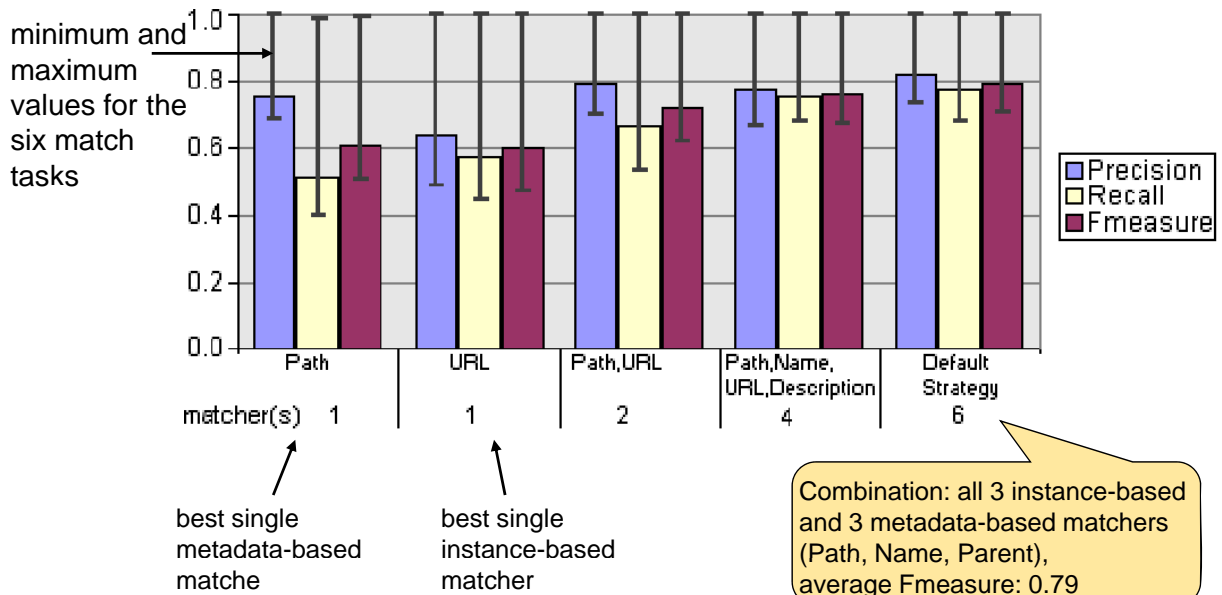
$\Sigma$  2245



33

## Instance- + metadata-based matching

- ▶ Combination of three instance-based matchers (URL, name, description) and six metadata-based matchers



34

# Agenda

- ▶ Ontologies
- ▶ Ontology Matching
  - Problem
  - Match techniques and prototypes (e.g., GLUE)
- ▶ Instance-based matching in COMA++
  - Constraint- / Content-based Matching
  - Matching web directories
- ▶ Matching by Instance overlap
  - Similarity measures
  - Evaluation: Product catalogs, biomedical ontologies
- ▶ Stability of ontology mappings
- ▶ Conclusions

35

## Matching by Instance Overlap \*

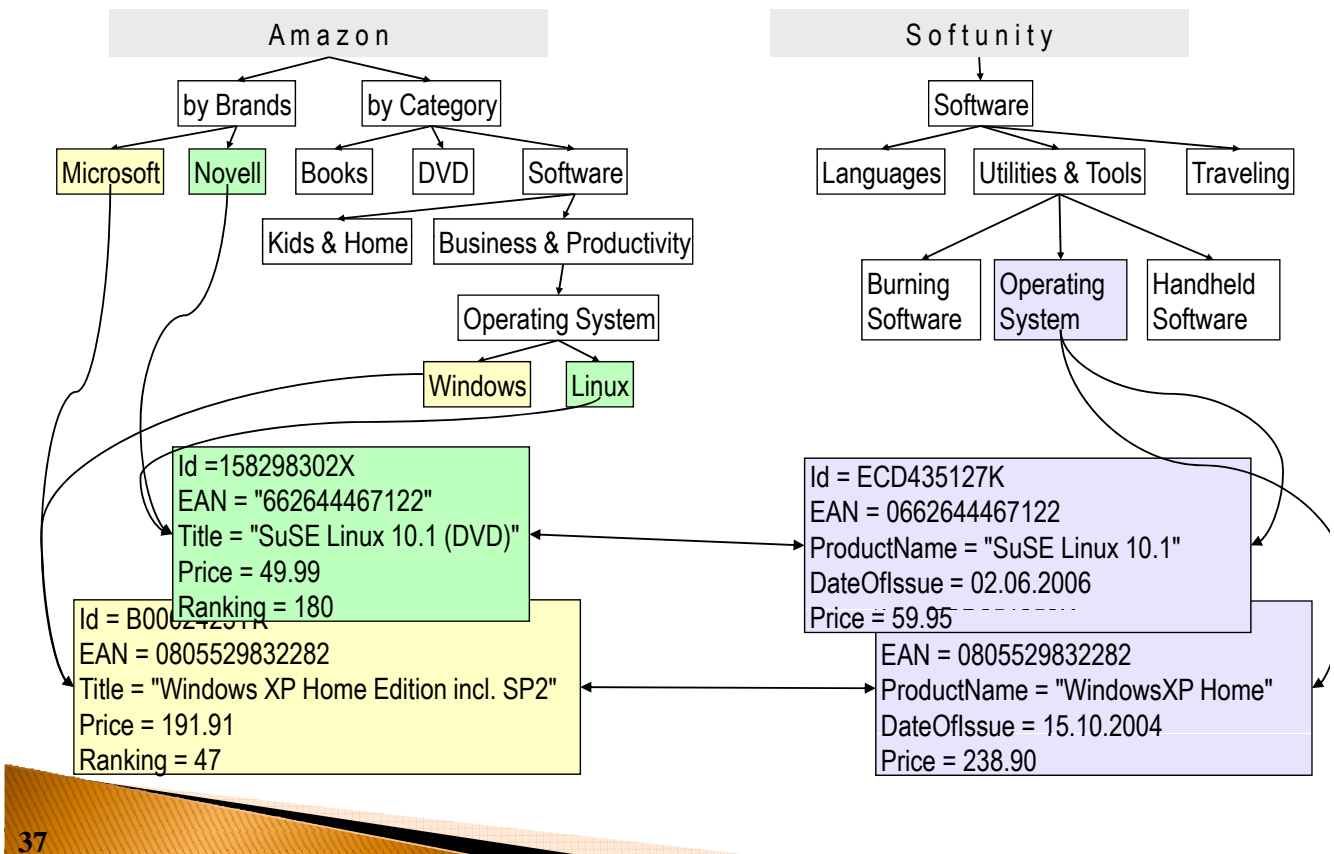
- ▶ **Use of instance overlap** for ontology matching: two concepts are related / similar if they share a significant number of associated objects
- ▶ Different measures to determine the instance-based similarity
  - Base-K; Dice, Min, Jaccard ...
- ▶ Extensions:
  - Consideration of indirect instance associations
  - Combination with other match approaches
  - Consideration of similar (but non-identical) objects

\* Thor, A; Kirsten, T; Rahm, E.: *Instance-based matching of hierarchical ontologies*. Proc. BTW, 2007

Kirsten, T, Thor, A; Rahm, E.: *Instance-based matching of large life science ontologies*. Proc. DILS, 2007

36

# Example: Product Catalogs



## Similarity Measures

- ▶ Baseline similarity  $Sim_{BaseK}$

$$Sim_{BaseK}(c_1, c_2) = \begin{cases} 1, & \text{if } N_{c_1c_2} \geq K \\ 0, & \text{if } N_{c_1c_2} < K \end{cases}$$

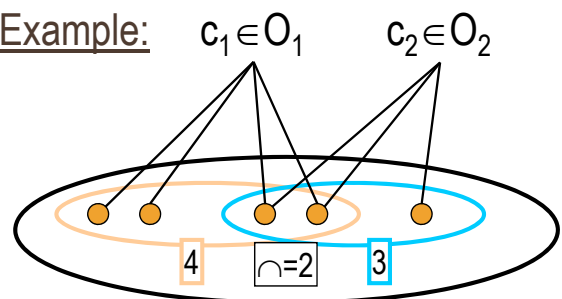
- Dice similarity  $Sim_{Dice}$

$$Sim_{Dice}(c_1, c_2) = \frac{2 \cdot N_{c_1c_2}}{N_{c_1} + N_{c_2}}$$

- Minimum similarity  $Sim_{Min}$

$$Sim_{Min}(c_1, c_2) = \frac{N_{c_1c_2}}{\min(N_{c_1}, N_{c_2})}$$

Example:



$$Sim_{Base1} = Sim_{Base2} = 1, Sim_{Base3} = 0$$

$$Sim_{Dice} = 2 \cdot 2 / (4 + 3) = 0.57$$

$$Sim_{Min} = 2/3 = 0.67$$

$$0 \leq Sim_{Dice} \leq Sim_{Min} \leq Sim_{Base1} \leq 1$$

# Approximate Evaluation Measures

- ▶ Computation of precision & recall needs a perfect mapping (reference alignment)
  - Laborious for large ontologies
  - Might not be well-defined
- ▶ Syntactic measures to “approximate” recall / precision
- ▶ **Match coverage**: fraction of matched categories

$$MatchCoverage_{o_1} = \frac{|C_{o_1-Match}|}{|C_{o_1}|} \in [0...1] \quad InstMatchCoverage_{Combined} = \frac{|C_{o_1-Match}| + |C_{o_2-Match}|}{|C_{o_1-Inst}| + |C_{o_2-Inst}|}$$

- ▶ **Match ratio**: #correspondences per matched concept

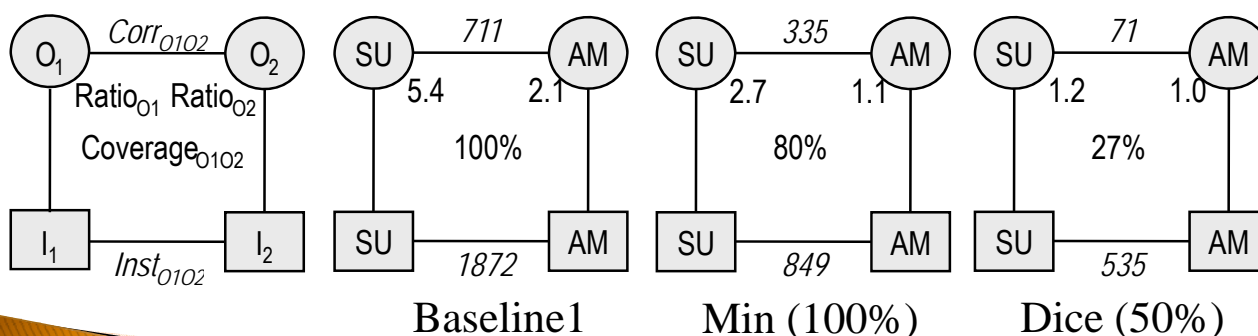
$$MatchRatio_{o_1} = \frac{|Corr_{o_1-o_2}|}{|C_{o_1-Match}|} \geq 1 \quad CombinedMatchRatio = \frac{2 \cdot |Corr_{o_1-o_2}|}{|C_{o_1-Match}| + |C_{o_2-Match}|} \geq 1$$

- ▶ Goal: high Match Coverage with low Match Ratio

## Results for Product Catalog Matching

- ▶ Amazon (AM) vs. Softunity (SU)
- ▶ Baseline1: max. Match Coverage, high Match Ratios
- ▶ Sim<sub>Min</sub>: good Match Coverage, moderate Match Ratios
- ▶ Sim<sub>Dice</sub>: low Match Coverage, low Match Ratios

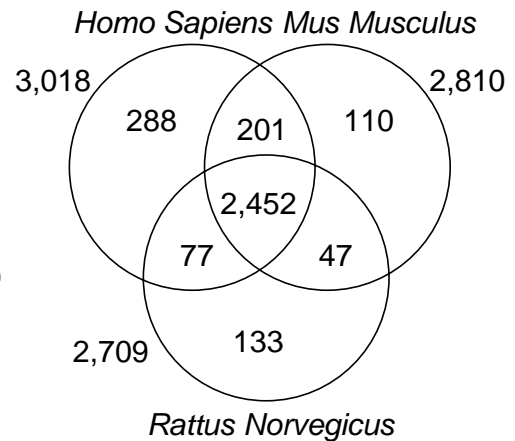
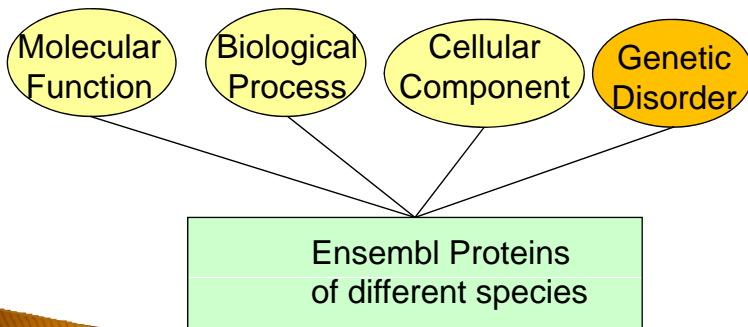
	SU	AM
# concepts (product categories)	470	1,856
# concepts having instances	170	1,723
# instances (products)	2,576	18,024
# direct associations	2,576	25,448
# associations / # instances	1	≈ 1.4
# Instances / #concepts	≈15	≈15





# Life Science Match Scenario

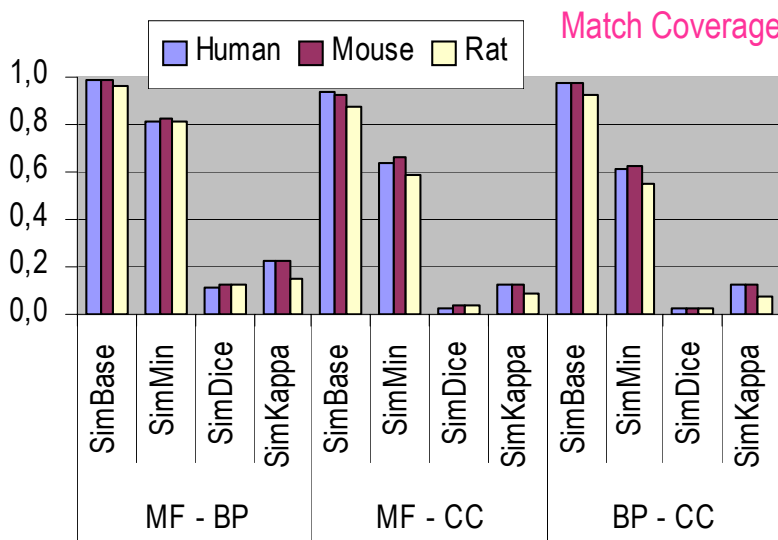
- ▶ Ontologies
  - 3 subontologies of GeneOntology
  - Genetic disorders of OMIM
- ▶ Instances: Ensembl proteins of 3 species, i.e. homo sapiens, mouse, rat
- ▶ Only subset of concepts has associated instances



Number of associated Biological Processes (total # processes: 12,555)

## Match Results for GO tasks

- ▶  $Sim_{Base}$ : high Match Coverage (99%) w.r.t. concepts having instances, very high Match Ratios
- ▶  $Sim_{Dice}$ : low Coverage (< 20%) and low Match Ratios
- ▶  $Sim_{Min}$ : good Coverage (60%–80%) with moderate Match Ratios



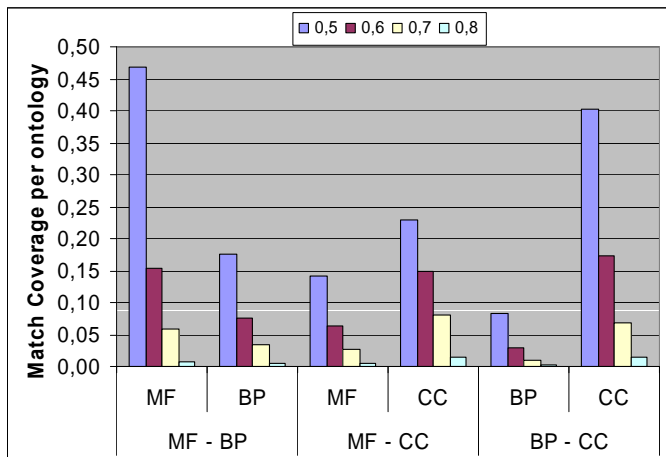
Match Ratios per ontology

	MF - BP	
	MF	BP
Base	20.4	17.0
Min	4.4	4.0
Dice	1.3	1.2

(Match Ratios for Homo Sapiens, MF-BP task)

# Metadata-based GO Matching

- ▶ Simple matcher on concept names
- ▶ Relatively low Match Coverage (however w.r.t. all concepts including instance-free concepts)
  - ▶ No correspondences for similarity  $\geq 0.9$
  - ▶ Low similarity thresholds (e.g.  $< 0.6$ ) too imprecise



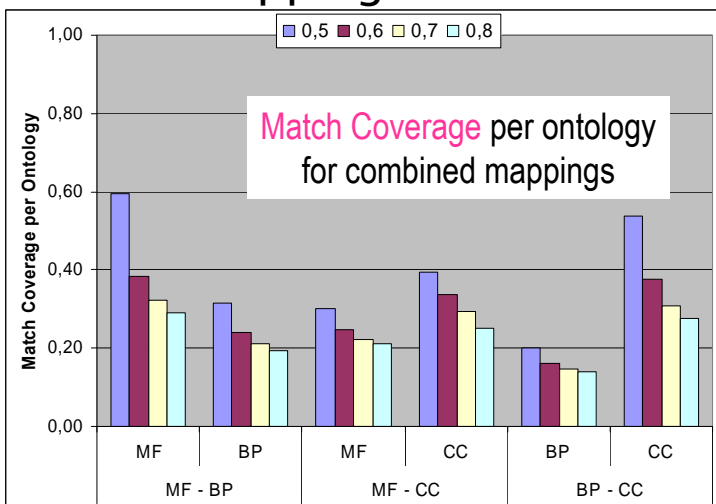
Match Coverage per ontology

	MF - BP	
	MF	BP
0.5	4.4	6.9
0.6	2.4	2.9
0.7	1.4	1.4
0.8	1.1	1.1

Match Ratios per ontology

## Results for Match Combinations

- ▶ Combinations between instance- ( $Sim_{Min}$ ) and metadata-based match approach
  - ▶ Union: Increased Match Coverage and Match Ratios
  - ▶ Intersection: Low Match Coverage ( $< 1\%$ )
- ▶ Low overlap between instance- and metadata-based mappings



Match Ratios per ontology  
(Name threshold 0.7)

	MF - BP	
	MF	BP
$\cup$	4.1	3.7
$\cap$	1.0	1.0

( $Sim_{Min} = 1.0$ , Homo Sapiens)

# Impact of Annotation Provenance

- ▶ Automatically vs. manually assigned annotations
- ▶ Example: Annotations in Ensembl (July 2008) – 46,704 proteins

	MF		BP	
Automatically assigned	82466	82%	57824	72%
Manually assigned	17729	18%	22951	28%
Sum	100195		80775	

- ▶ Ontology mappings for Base3,Min
  - Restriction to manual annotations returns small mappings of likely improved quality

		Corr <sub>BP_MF</sub>	C <sub>BP</sub>	C <sub>MF</sub>
all	Base3	21386	1939	1393
	Min ∩ Base3	3275	1107	1107
man	Base3	3835	899	533
	Min ∩ Base3	758	435	285

		MC <sub>BP</sub>	MC <sub>MF</sub>	MR <sub>BP</sub>	MR <sub>MF</sub>
all	Base3	0,13	0,17	11,0	15,4
	Min ∩ Base3	0,08	0,13	3,0	3,0
man	Base3	0,06	0,06	4,3	7,2
	Min ∩ Base3	0,03	0,03	1,7	2,7

45

## Agenda

- ▶ Ontologies
  - ▶ Ontology Matching
    - Problem
    - Match techniques and prototypes (e.g., GLUE)
  - ▶ Instance-based matching in COMA++
    - Constraint- / Content-based Matching
    - Matching web directories
  - ▶ Matching by Instance overlap
    - Similarity measures
    - Evaluation: Product catalogs, biomedical ontologies
- ▶ Stability of ontology mappings
  - ▶ Conclusions

# Evolution of Life Science Ontologies

- ▶ Continuous evolution of ontologies (many versions)
- ▶ Evolution analysis of 16 life science ontologies:
  - Average of 60% growth in last four years
  - Deletes and changes also common

Ontology	size	C  <sub>start</sub>	C  <sub>last</sub>	grow <sub> C , start, last</sub>
NCI Thesaurus		35,814	63,924	1.78
GeneOntology		17,368	25,995	1.50
-- Biological Process	large	8,625	15,001	1.74
-- Molecular Function		7,336	8,818	1.20
-- Cellular Components		1,407	2,176	1.55
ChemicalEntities		10,236	18,007	1.76



[www.izbi.de/onex](http://www.izbi.de/onex)

Ontology	Full period (May. 04 - Feb. 08)							Last year (Feb. 07 - Feb. 08)		
	Add	Del	Obs	adr	add-frac	del-frac	obs-frac	Add	Del	Obs
NCI Thesaurus	627	2	12	42.4	1.3%	0.0%	0.0%	416	0	5
GeneOntology	200	12	4	12.2	0.9%	0.1%	0.0%	222	20	5
-- Biological Process	146	7	2	16.2	1.2%	0.1%	0.0%	133	10	2
-- Molecular Function	36	3	2	6.8	0.4%	0.0%	0.0%	69	7	3
-- Cellular Components	18	2	0	8.9	1.0%	0.1%	0.0%	19	3	0
ChemicalEntities	256	62	0	4.1	1.8%	0.5%	0.0%	384	67	0

#monthly changes:

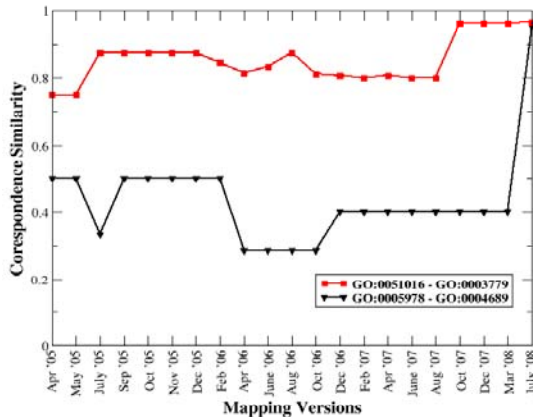
Hartung, M; Kirsten, T; Rahm, E.: *Analyzing the Evolution of Life Science Ontologies and Mappings*. Proc. 5<sup>th</sup> Intl. Workshop on Data Integration in the Life Sciences (DILS), 2008

## Stability of Ontology Mappings

- ▶ High change rates in
  - Ontologies
  - Instances
  - Annotations (instance–concept associations)
- ▶ Ontology mappings (between versions of two ontologies) also change frequently, especially for instance–based match approaches
  - correspondences may disappear in newer mapping versions
- ▶ Consideration of instance overlap or metadata–bases similarity may not be sufficient for determining „good“ ontology mappings

# Stability of Ontology Mappings

- ▶ Standard match approaches only consider information about current ontology versions and ignore evolution history



Is the **black** correspondence as good as the **red** one?

- Possible instabilities of match correspondences due to evolution of ontologies and/or related data source

- Idea: Consider the **evolution** of a match correspondence to assess its **stability/quality** in the current version

\* Thor, A; Hartung, M; Gross, A; Kirsten, T; Rahm, E.: *An Evolution-based Approach for Assessing Ontology Mappings - A Case Study in the Life Sciences*. Proc. BTW, 2009

# Stability Measures

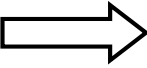
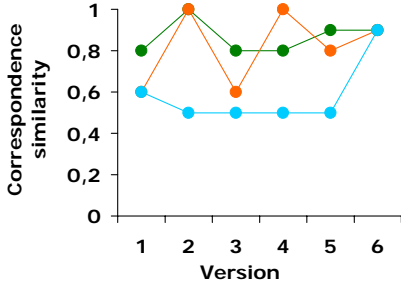
- ▶ Average Stability

$$stabAvg_{n,k}(a,b,m) = 1 - \frac{1}{k} \cdot \sum_{i=n-k}^{n-1} |sim_{i+1}(a,b,m) - sim_i(a,b,m)| \in [0,1]$$

- ▶ Weighted Maximum Stability

- Proximity of similarities in the last versions compared to the current version

$$stabWM_{n,k}(a,b,m) = 1 - \max_{i=1...k} \left[ \frac{|sim_n(a,b,m) - sim_{n-i}(a,b,m)|}{i} \right] \in [0,1]$$



	stabAvg <sub>6,5</sub>	stabWM <sub>6,5</sub>
(a <sub>1</sub> , b <sub>1</sub> )	0.9	0.95
(a <sub>2</sub> , b <sub>2</sub> )	0.7	0.9
(a <sub>3</sub> , b <sub>3</sub> )	0.9	0.6

# Stability Evaluation

- ▶ Setting
  - ▶ Mapping GO *Biological Processes* to *Molecular Functions*
  - ▶ Instance based matching (using Ensembl source)
  - ▶ Result: 2,497 correspondences ( $\text{Base3} \cap \text{Min} \geq 0.8$ ) which existed in the last 5 versions
- ▶ Selection of correspondences based on similarity and stability

accepted

candidates

questionable

$\text{sim}_{26} > 0.9$	$\text{stabWM} > 0.95$	$0.95 \geq \text{stabWM} \geq 0.85$	$0.85 > \text{stabWM}$	$\Sigma$
$\text{stabAvg} > 0.95$	424 55%	37 55%	11 25%	596
$0.95 \geq \text{stabAvg} \geq 0.85$	863 96%	15%	235 125%	1734
$0.85 > \text{stabAvg}$	17 5%	13 16%	30% 85%	167
$\Sigma$	1449	536	512	2497

51

## Conclusions

- ▶ Instance-based match approaches
  - ▶ Important since instances reflect well semantics of categories
  - ▶ Availability of usable instances may be restricted to subset of concepts (consideration of indirectly associated instances helpful)
  - ▶ Need to be combined with metadata-based techniques
- ▶ Correct ontology mappings NOT limited to 1:1 correspondences
- ▶ High change rates for ontologies/instances may result in unstable ontology mappings
- ▶ Matching based on shared instances
  - ▶ Different similarity measures to consider instance overlap
  - ▶ Especially applicable in bioinformatics (frequent annotations)
- ▶ Instance-based matching in COMA++
  - ▶ 3 basic instance matchers (constraint-based, content-based) not requiring shared instances
  - ▶ Flexible combination with many metadata-based approaches and different match strategies

52



# Some Areas for Further Work

- ▶ Evaluation and validation of large ontology mappings
- ▶ Combined study of ontology matching and instance (entity) matching
  - Correspondences based on instance similarity not equality
  - Entity matching utilizing category similarity
  - Automatic instance categorization
- ▶ Scalable instance match approaches based on machine learning
- ▶ Ontology Evolution
- ▶ Ontology Merging

# Online Bibliography

<http://se-pubs.dbs.uni-leipzig.de>

**Schema Evolution** publication categorizer

Keyword search:  Search More options

Guided search: Click a term to initiate a search.

Schema Evolution: Schema Evolution (218), Sch. Matching/Mapping (173), Model Management (67), Information integration (21)

Paper type: no paper type (224), Theory (100), Prototype (31), Survey / Rihl (31)

Search: Ontology Alignment

Results

Title/Author	Year	Citation:	Result
<i>Doan, AnHai; Madhavan, Jayant; Domingos, Pedro; Halevy, Alon</i> Learning to Map between Ontologies on the Semantic Web	2002	666	
<i>Noy, N.F.; Musen, M.A.</i> The PROMPT suite: interactive tools for ontology merging and mapping	2003	330	

Chiang Chou Churchill Cimpian Claypool Cleve CURINO Dadam Davidson Del Fabro Deutsch Dittrich Doan Domingos Dominguez Dong Easterbrook Edberg Ehrig Engmann Euzenat Fagin Fensel Ferrandini Fletcher Gal Garcia-Molina Giunchiglia Grandi Gray Guerrini Haas Haase Hainaut Halevy Hartung Henrard Hernandez Hick Ho Ichise Jamil Jarke Jonker Jouault Jouault Kedad Keller Kensche Kim Kirsten Kleir Klettke Koeller Kolaitis Kong Kosky Kramer Kuno Lakshmanan Lambrix Lammel Laurent Lautemann Lechtenboerger Lechtenborger Lee Lee Lenzerini Leonard Leonardi Lerner Li Lieberherr Lin Lloret Lopes Loscic Madhavan Maedche Massmann McBrien Mecca Melnik Mens Mesiti Meyer Miller Mocan Modici Moon Morishima Mork Morzy Motik Motro Motz Musen Mylopoulos Nash Naumann Navathe Nejadi Nica Noy

# References

- ▶ Aumüller, D., Do, H., Massmann, S., Rahm, E.: *Schema and ontology matching with COMA++*. Proc. ACM SIGMOD, 2005
- ▶ Hartung M, Kirsten T., Rahm E.: *Analyzing the evolution of life science ontologies and Mappings*. Proc. DILS 2008. Springer LNCS 5109
- ▶ Kirsten T., Thor A., Rahm E.: *Instance-based matching of large life science ontologies*. Proc. DILS 2007. Springer LNCS 4544
- ▶ Massmann, S.; Rahm, E.: *Evaluating Instance-based matching of web directories*. Proc. 11th Int. Workshop on the Web and Databases (WebDB), 2008
- ▶ Rahm, E., Bernstein, P.: *A survey of approaches to automatic schema matching*. The VLDB Journal, 10(4): 334–350, 2001.
- ▶ Thor A., Hartung M., Groß A., Kirsten T., Rahm E.: *An evolution-based approach for assessing ontology mappings – A case study in the life sciences*. Proc. 13<sup>th</sup> German Database Conf. (BTW), 2009
- ▶ Thor, A., Kirsten, T., Rahm, E.: *Instance-based matching of hierarchical ontologies*. Proc. 12<sup>th</sup> German Database Conf. (BTW), 2007