UNIVERSITÄT LEIPZIG

Semantic ontology mappings

How to determine and use them

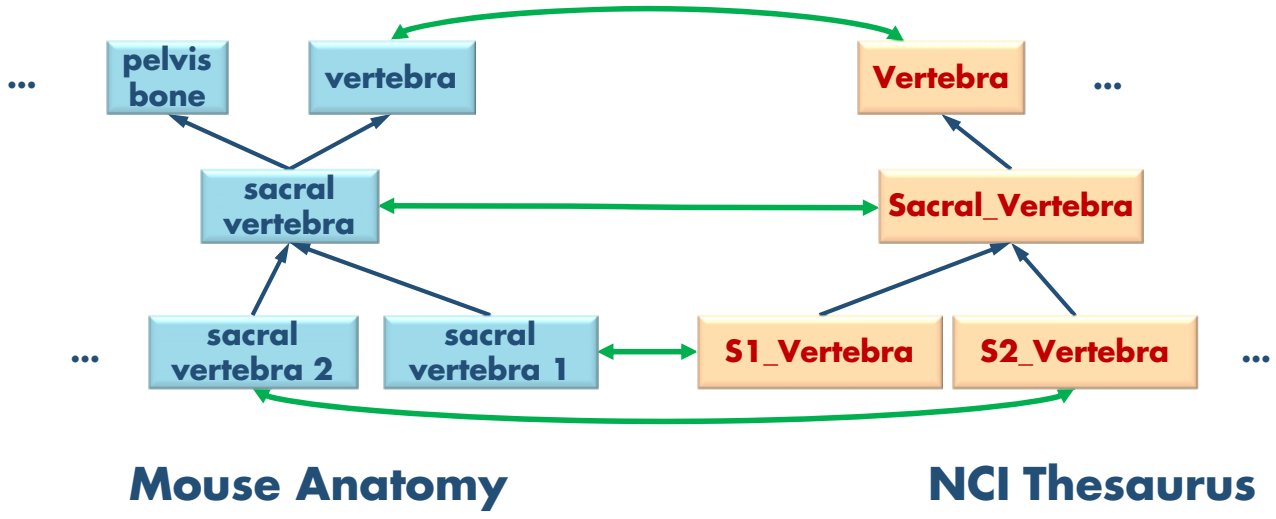
Erhard Rahm,
Patrick Arnold, Salvatore Raunich

http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

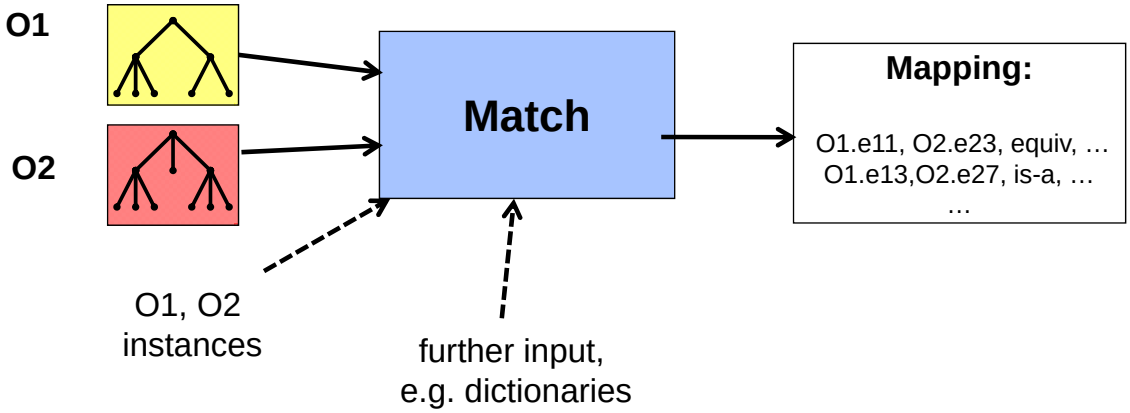
Ontology Matching

- identification of semantic correspondences between ontologies



Ontology Matching

UNIVERSITÄT LEIPZIG



- process of identifying semantic correspondences between input ontologies
  - result: ontology mapping
  - **mostly equivalence mappings**: correspondences specify equivalent ontology concepts
- Semantic Matching: support for different semantic correspondence/relationship types

Why Matching is important

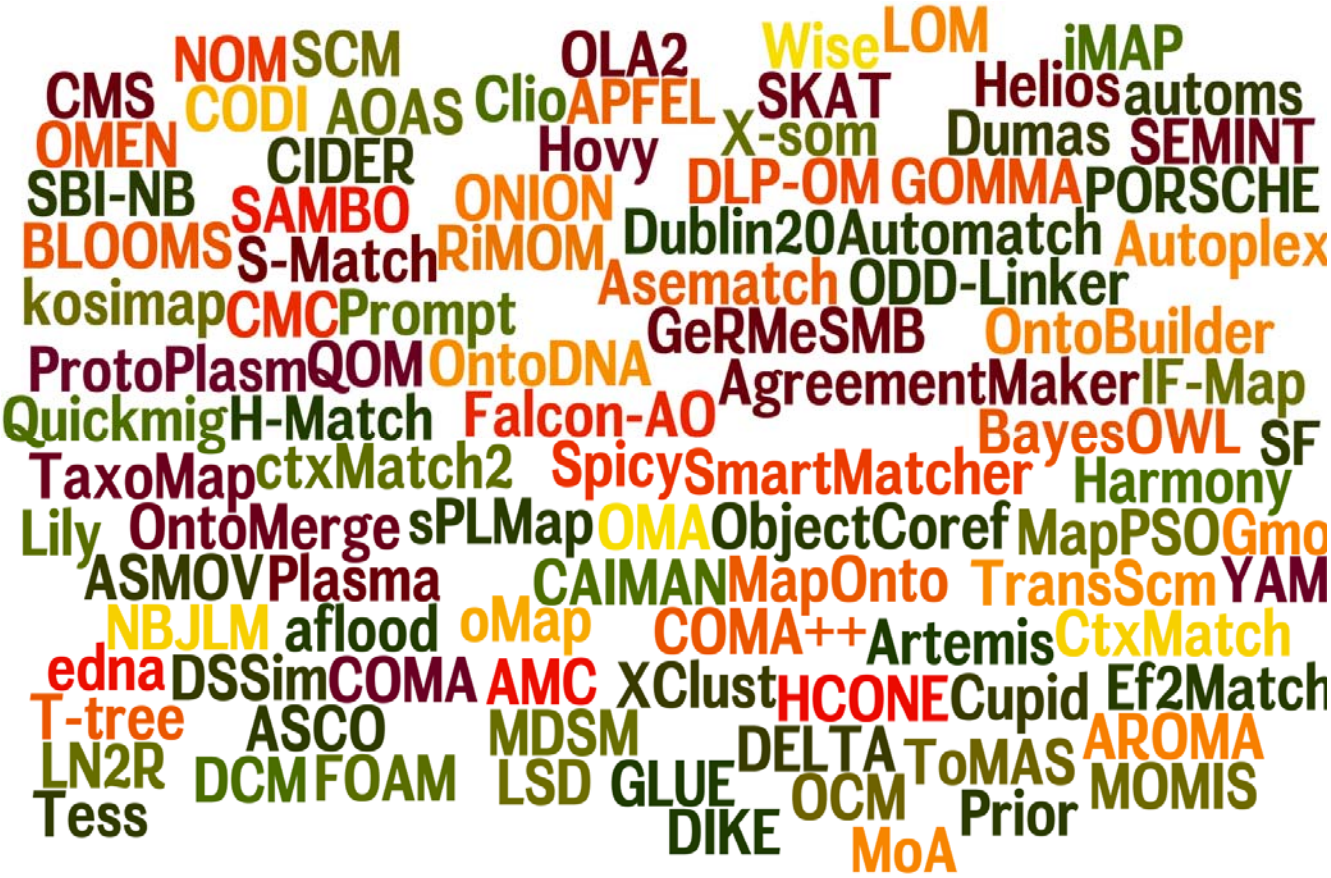
UNIVERSITÄT LEIPZIG

- semantic data linking and integration
- interrelate concepts to find/query additional annotations
- integrate / unify / merge overlapping/related ontologies or schemas
- support for ontology evolution

Match techniques

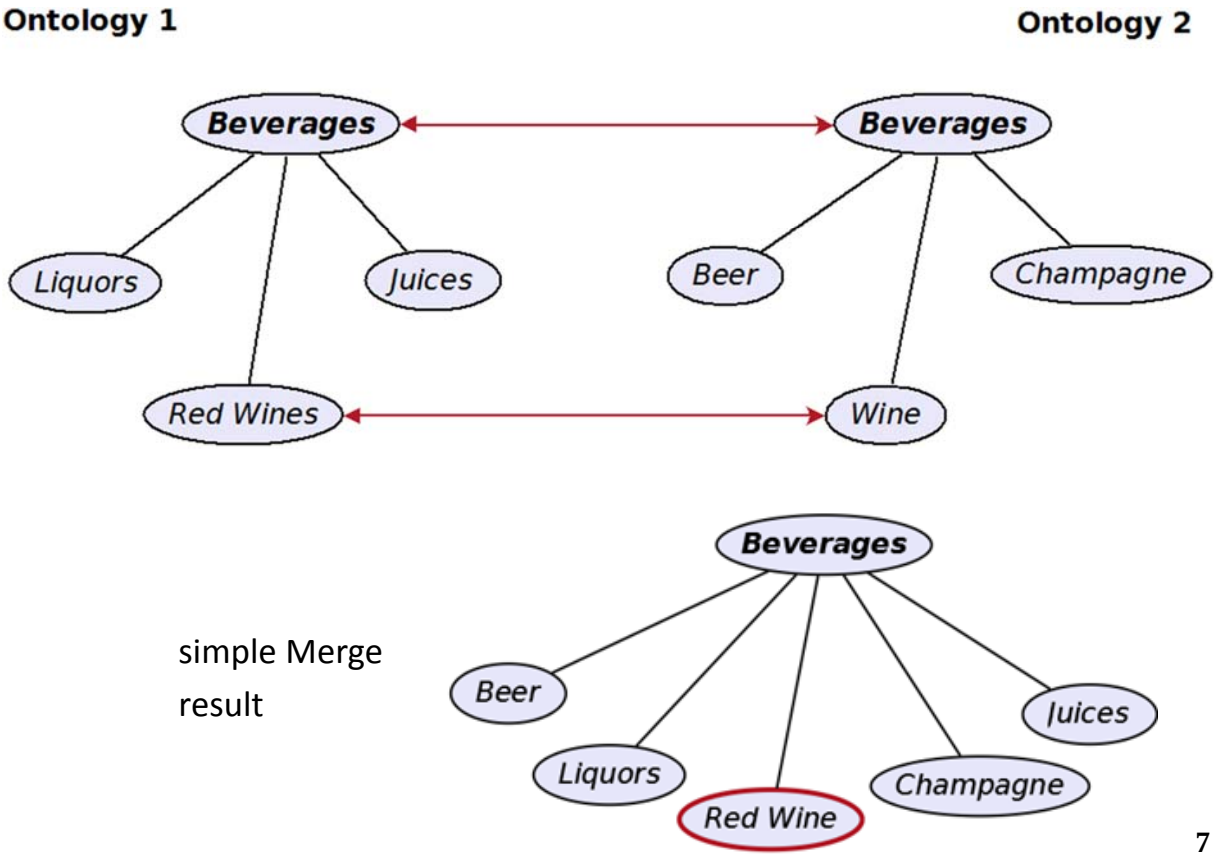
- Linguistic matchers
  - similarity of concept/element names
  - use of dictionaries/thesauri, e.g. WordNet / UMLS
- Structure-based matchers
  - consider similarity of ancestors/descendants
- Instance-based matchers
  - consider similarity of instances/annotations
- need to combine several match techniques („matchers“)

Match Prototypes



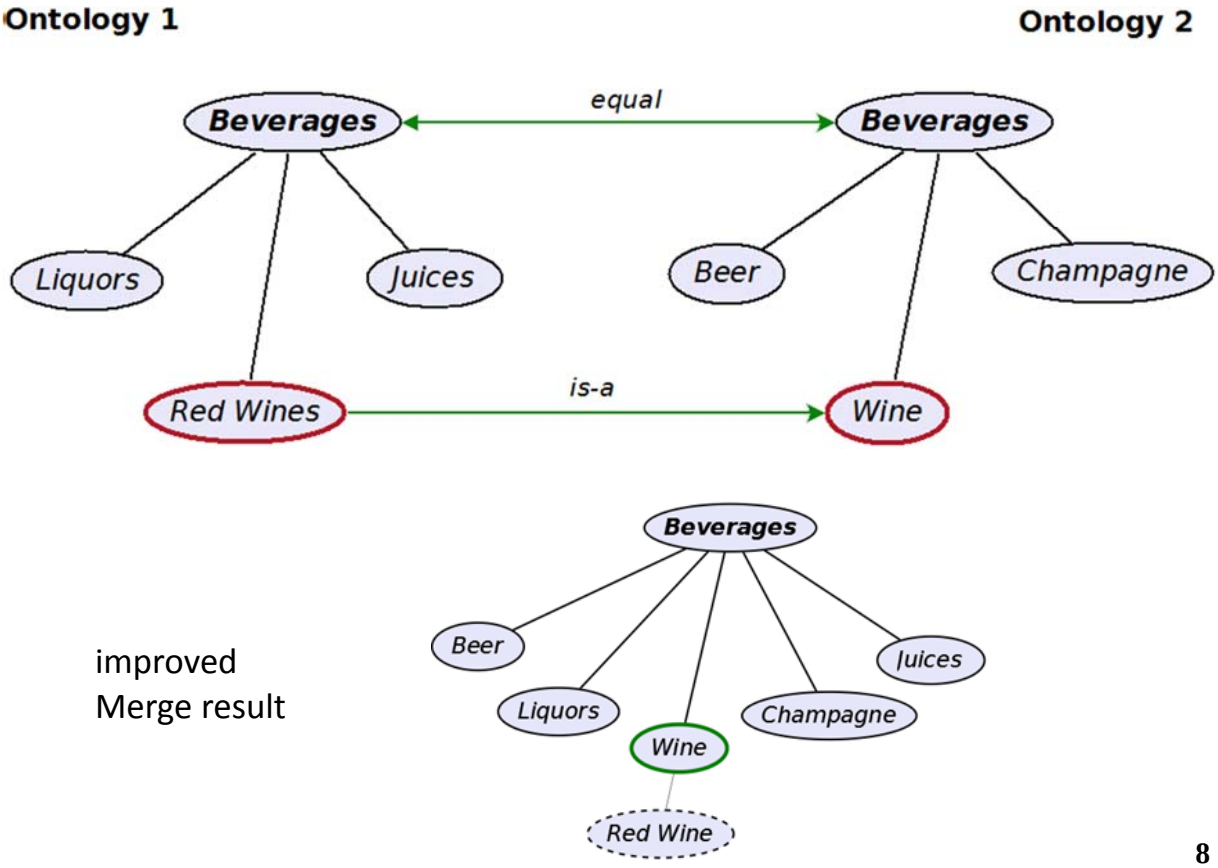
Simple Match + Merge

UNIVERSITÄT LEIPZIG



Semantic Match + Merge

UNIVERSITÄT LEIPZIG



Agenda

- Ontology matching
- Semantic match approach STROMA
  - Enrichment approach and component strategies
  - Evaluation
- Broadening background knowledge
  - Extracting semantic relations from Wikipedia
  - Semantic mapping repository SemRep
- Ontology merging
  - Symmetric vs. asymmetric merge
  - ATOM approach
- Conclusions and outlook

STROMA*

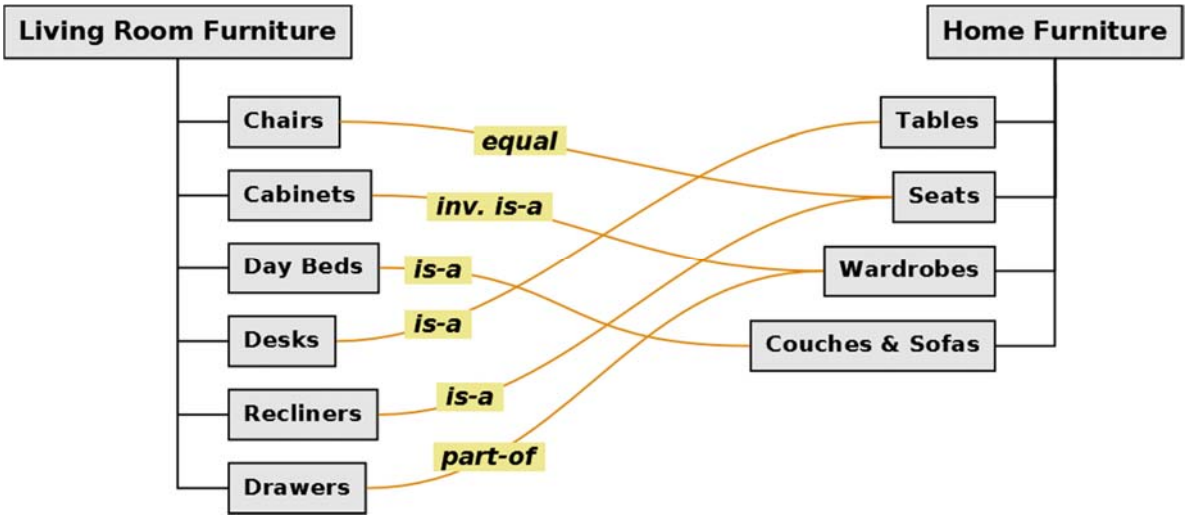
- STROMA: SemanTic Refinement (Enrichment) of Ontology MAppings
- STROMA follows a 2-step approach to semantic matching
  - step 1: Standard Matching (Coma, AgreementMaker, …)
  - step 2: Semantic enrichment of equivalence mapping
- Characteristics
  - different match tools can be used for step 1
  - step 2 is relatively cheap since only relatively few correspondences need to be evaluated
  - strong dependency on step 1 (tool, mapping)

* Arnold, P., Rahm, E.: *Enriching Ontology Mappings with Semantic Relations.* Data and Knowledge Engineering. 2014

Supported Relation Types

Type	Linguistic Terminology	Example
Equal	Synonym – Synonym	Car – Automobile
Is-A	Hyponym – Hypernym	Car – Vehicle
Inv. Is-A	Hypernym – Hyponym	Vehicle – Car
Part-Of	Meronym – Holonym	Engine – Car
Has-A	Holonym – Meronym	Car – Engine
Related	Cohyponym – Cohyponym	Car – Bike

STROMA Match example



Semantic Match Tools

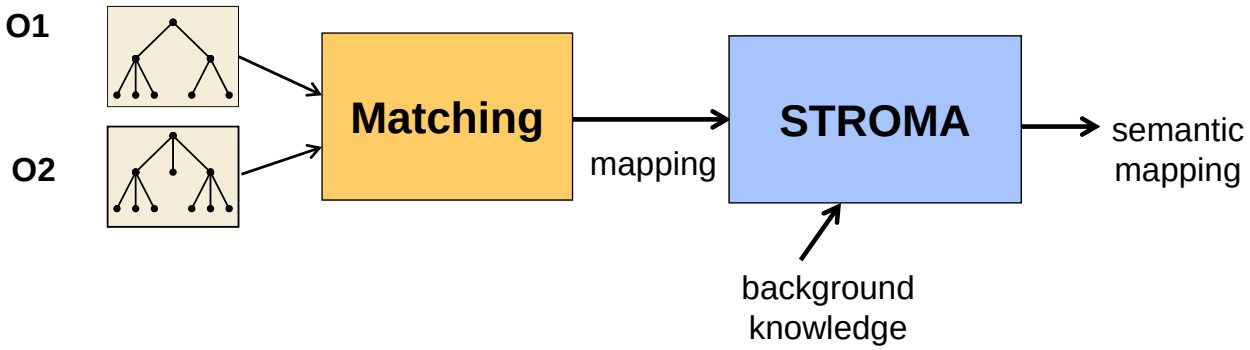
UNIVERSITÄT LEIPZIG

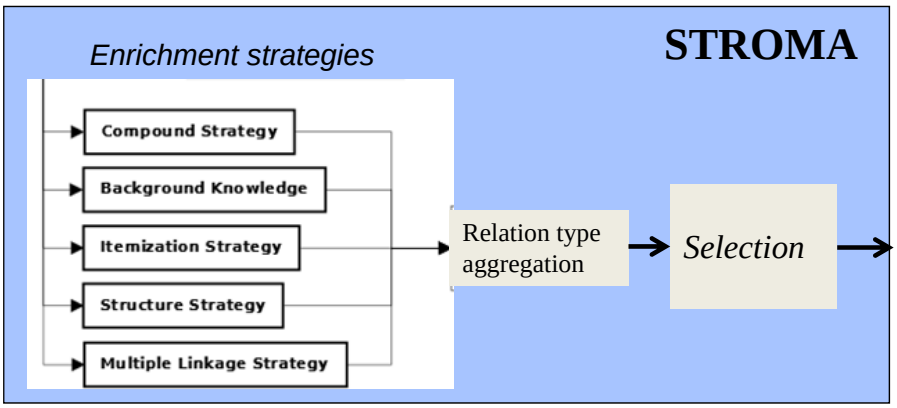
	S-Match	TaxoMap	Aroma	ASMOV	STROMA
Architecture	1-step	1-step	1-step	1-step	2-step
Background knowledge sources	WordNet	WordNet		WordNet	WordNet, UMLS, OpenThesaurus, Wikipedia
Main techniques	linguistic	linguistic	probabi-listic, instance-based	linguistic, structural, instance-based	linguistic, structural

	equal	is-a / inv. is-a	part-of / has-a	related
S-Match	X	X		X
TaxoMap	X	X		X
ASMOV	X	X		
AROMA	X	X		
STROMA	X	X	X	X

STROMA architecture

UNIVERSITÄT LEIPZIG





Compound Strategy

- discovers compound words in a correspondence
  - *red wine – wine*
  - *bus-driver - driver*
  - *blackboard - board*
- compound relations frequently express *is-a* relations

- Modifier term should have minimal length (e.g., 3)
  - inconsistency – consistency
  - retail - tail

Apply Background Knowledge

- critical to find semantic relations for concept names without any linguistic similarity
  - car *is-a* vehicle
  - engine *part-of* car
- support for diverse BK resources
  - initially: WordNet, UMLS, OpenTheasurus
  - new **SemRep** repository includes automatically derived relations from Wikipedia
- fast lookup needed

Apply Background Knowledge (2)

- **Gradual modifier removal** technique
  - compound words are often missing in BK resources, e.g. *US vice president* not in WordNet
  - incrementaly remove modifier terms from the left until term can be found
  - Allows to find correspondences such as

    US vice president is-a person

Itemization Strategy

- concept names with itemizations need special treatment
  - beer & wine - wine
  - not *is-a* but *inverse is-a*
- Approach
  - determine component terms (items) in itemizations I1, I2

    { books, e-books, movies, films, cds } - { novels, compact discs }
  - remove hyponyms and synonyms per itemization

    { books, movies, cds } - { novels, compact discs }
  - remove snonyms between concepts

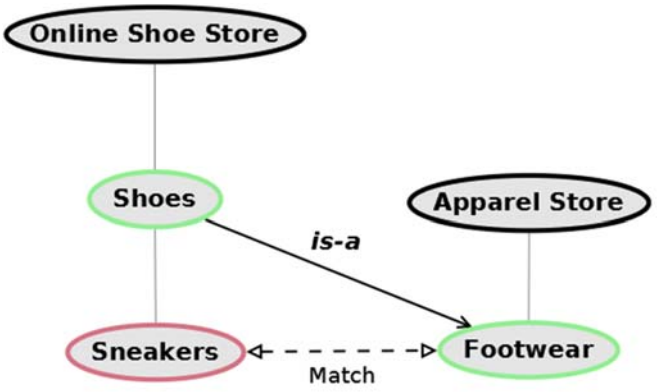
    { books, movies} - { novels}
  - remove items in I2/I1 that are a hyponym of a term in I1/I2

    { books, movies} - {}
  - determine relation type based on cardinality of I1 and I2:

    equal (both empty), *is-a* (only I1 empty), *inv is-a* (only I2 empty), *undecided* (both not empty)
- uses compound and BK strategies for finding synonyms/homonyms

Structure Strategy

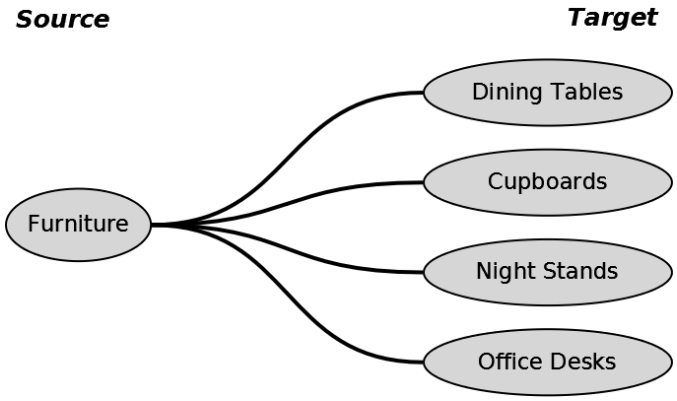
- **issue:** A relation between two matching concepts X, Y cannot be derived
  - consider father concepts X‘ or Y‘
  - check relations X‘-Y resp. X - Y'



Shoes is-a Footwear → *Sneakers is-a Footwear*

Multiple Linkage Strategy

- Heuristic
  - If node X has multiple relations to nodes Y1, ..., Yn, X is probably more general than Y1, ..., Yn.
  - Inv. is-a (or has-a) can be assumed



Aggregating relation type decisions

- each input correspondence is handled by each of the implemented strategies
- combine individual results based on confidence weights
- example: *high-school* ⇔ *school*

Strategy	Conf.	Equal	Is-a	Inv. Is-a	Part-of	Has-a	Related
Compound	*1*	0	1	0	0	0	0
Background Kn.	*2*	0	2	0	0	0	0
Structure	*0.5*	0.5	0	0	0	0	0
Itemization	*1*	0	0	0	0	0	0
Multiple Link.	*0.5*	0	0	0.5	0	0	0
Result		0.5	**3**	0.5	0	0	0

STROMA evaluation

6 benchmarks with perfect mappings

	Benchmark	#Corr.	Equal	Is-a	Part-of	related
1	Web directories	340	278	52	5	5
2	Diseases	395	354	40	1	0
3	Text Mining taxon.	762	70	692	0	0
4a	Furniture: Amazon-Ebay	136	15	111	10	0
4b	Furniture: Wikipedia-Ebay	87	3	83	0	1
4c	Furniture: Amazon-Wikipedia	138	16	115	7	0

STROMA evaluation (2)

- Perfect input mapping
  - F-Measure 87-96% for tasks 1-3
  - lower quality for furniture tasks (39-67% F-measure overall, 49-77% für non-equal types)
- COMA input mapping
  - missing ontologies for task 3
  - missing/wrong correspondences in input mapping reduces overall F-measure to 60-67% for tasks 1-2 and 21-63% for furniture tasks

Comparative tool evaluation

Overall F-Measure results (Furniture benchmarks)

- S-Match (2011 version), TaxoMap V3.5 (different thresholds)
- COMA + STROMA using WordNet
- Wikipedia tests challenging due to very specific concepts , e.g. *bean bag (is-a chair)* or *cassone (is-a chest/coffer)*

	Benchmark	S-Match	TaxoMap	(COMA +) STROMA
4a	Amazon-Ebay	0.37	0.31	0.63
4b	Wikipedia-Ebay	0.04	0	0.34
4c	Amazon-Wikipedia	0.15	0	0.21

Agenda

- Ontology matching
- Semantic match approach STROMA
  - Enrichment approach and component strategies
  - Evaluation
- Broadening background knowledge
  - Extracting semantic relations from Wikipedia
  - Semantic mapping repository SemRep
- Ontology merging
  - Symmetric vs. asymmetric merge
  - ATOM approach
- Conclusions and outlook

Background Knowledge Acquisition

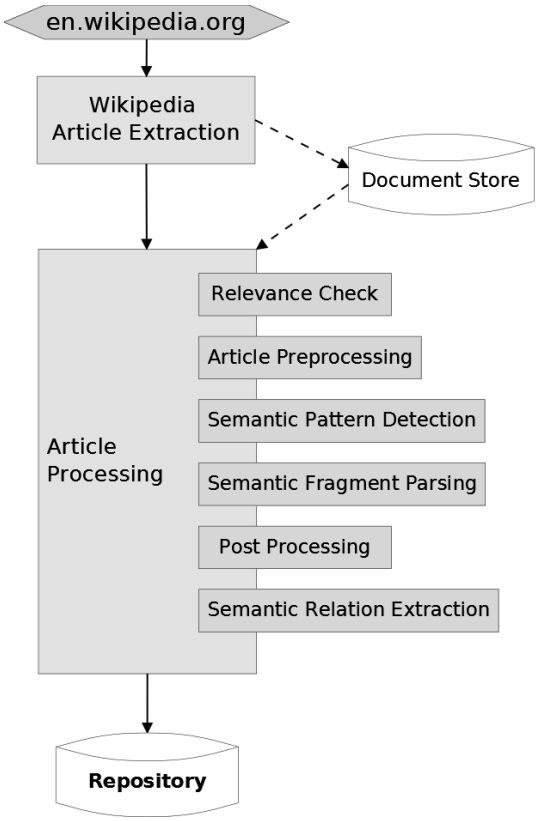
- Background knowledge valuable
- Limited coverage of existing BK resources
  - WordNet
  - domain-specific resources like UMLS
- Goals
  - gather additional background knowledge (semantic concept relations) from Wikipedia
  - build up comprehensive repository (SemRep) with combined knowledge from different sources

Extracting semantic relations from Wikipedia*

- Wikipedia
  - comprehensive, curated and current definitions of almost any noun of the English language
  - definition sentence usually contains an is-a relation
    - sometimes also has-a or part-of relations
- Approach
  - extract and process definition sentence per noun / concept name
  - determine relations based on typical patterns using finite-state machines

*Arnold, P.; Rahm, E.: *Extracting Semantic Concept Relations from Wikipedia*. Proc. 4th Int. Conf. Web Intelligence, Mining and Semantics (WIMS), 2014

Extraction workflow

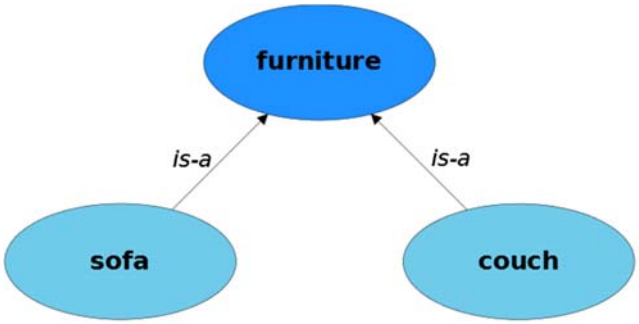


Use of semantic patterns

- semantic relations can be identified by typical patterns in definition sentences
- Is-a patterns
  - ... is a ...
  - ... is a specific form of ...
  - ... is typically a kind of ...
- Has-a patterns
  - ... consisting of ...
  - ... having ...
- find the concepts that are connected by the patterns

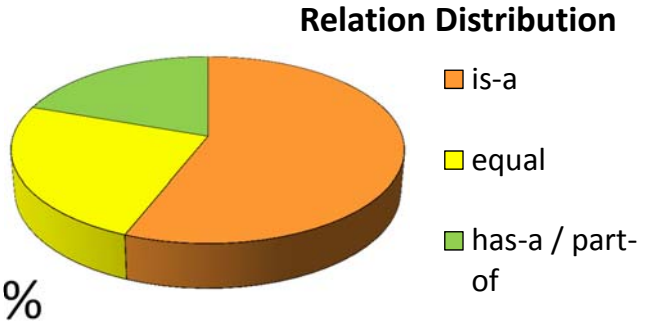
Example

A **couch** or **sofa** is a piece of furniture for seating two or more people [...]



Extraction results

- Wikipedia relation extraction yields about...
  - 5.3 million concepts
  - 11.1 million relations



- precision is about 80 %
  - ca. 20 % false relations
  - need for post-processing and quality improvement

Semantic Repository SemRep

- combines semantic relations from different resources to support matching etc.

Resource	Lang.	Creation	#Concepts	#Relations	File size
WordNet	EN	Manually	116,326	1,694,505	45 MB
Wikipedia	EN	Automatically	5,300,428	11,071,622	374 MB
UMLS	EN	Manually	109,599	281,972	19 MB
OpenThesaurus	DE	Manually	58,473	914,864	25 MB

- key features
  - comprehensive, multi-lingual, extensible
  - fast lookup
  - can indirectly derive relations (complex paths)

Querying SemRep

- queries to request semantic relation between concepts A and B
- if there is no direct relation, SemRep evaluates paths between A and B up to length k (2-4)
  - for each path, the relation type and a confidence value is calculated
  - path type of the highest scored path is returned (or a ranked list of several candidates)

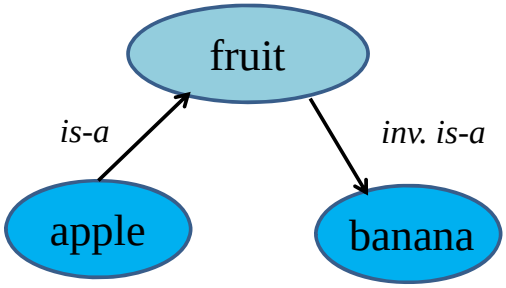
Determining relation type for paths

- easy for homogeneous paths with same relation type per step
- otherwise: use aggregation matrix
  - *convertible is-a car has-a engine*
  - *combustion engine is-a engine part-of car inv. is-a convertible*

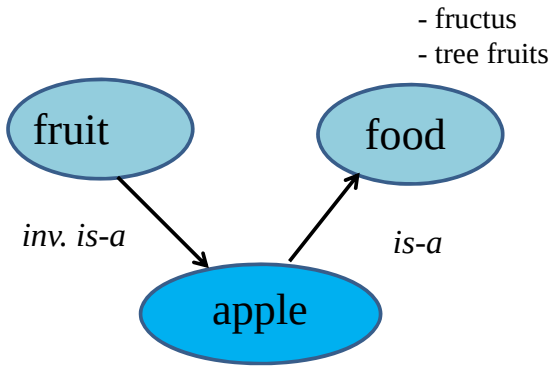
c1 – c2 \ c2 – c3	equal	is-a	inv. is-a	part-of	has-a
equal	*equal*	*is-a*	*inv. is-a*	*part-of*	*has-a*
is-a	*is-a*	*is-a*	*related*	*part-of*	*has-a*
inv. is-a	*inv. is-a*	–	*inv. is-a*	*part-of*	*has-a*
part-of	*part-of*	*part-of*	*part-of*	*part-of*	*related*
has-a	*has-a*	*has-a*	*has-a*	–	*has-a*

Relation type for paths (2)

- Problem cases with inverse relations



result type: *related*



different result types possible (is-a, inv is-a, equal): *undecided*

SemRep implementation

- implemented in Java, using a hash map for concepts and relations
  - relational and graph DBMS too slow
- path search is implemented as a bi-directional breadth first search
- 5-10 ms execution time for maximal path length 2

SemRep evaluation

- STROMA results using WordNet vs. STROMA using SemRep

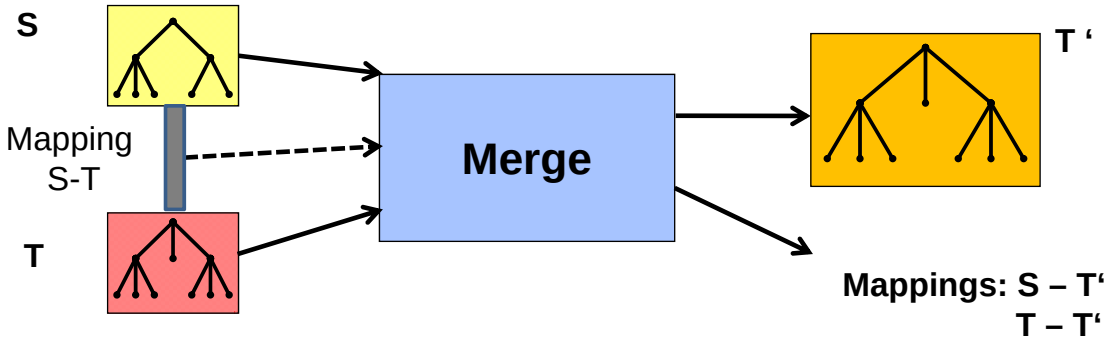
Benchmark	#Corresp.	F-Measure (STROMA + WordNet)	F-Measure (STROMA + SemRep)
Furniture	136	66.9 %	77.1 %
Groceries	169	38.4 %	49.1 %
Clothing	144	44.1 %	68.5 %

  - substantial improvement of mapping quality
  - using only BK strategy achieves F-Measure of 12-24% (WordNet) vs 43-53% (SemRep)

Agenda

- Ontology matching
- Semantic match approach STROMA
  - Enrichment approach and component strategies
  - Evaluation
- Broadening background knowledge
  - Extracting semantic relations from Wikipedia
  - Semantic mapping repository SemRep
- Ontology merging
  - Symmetric vs. asymmetric merge
  - ATOM approach
- Conclusions and outlook

Ontology Merging

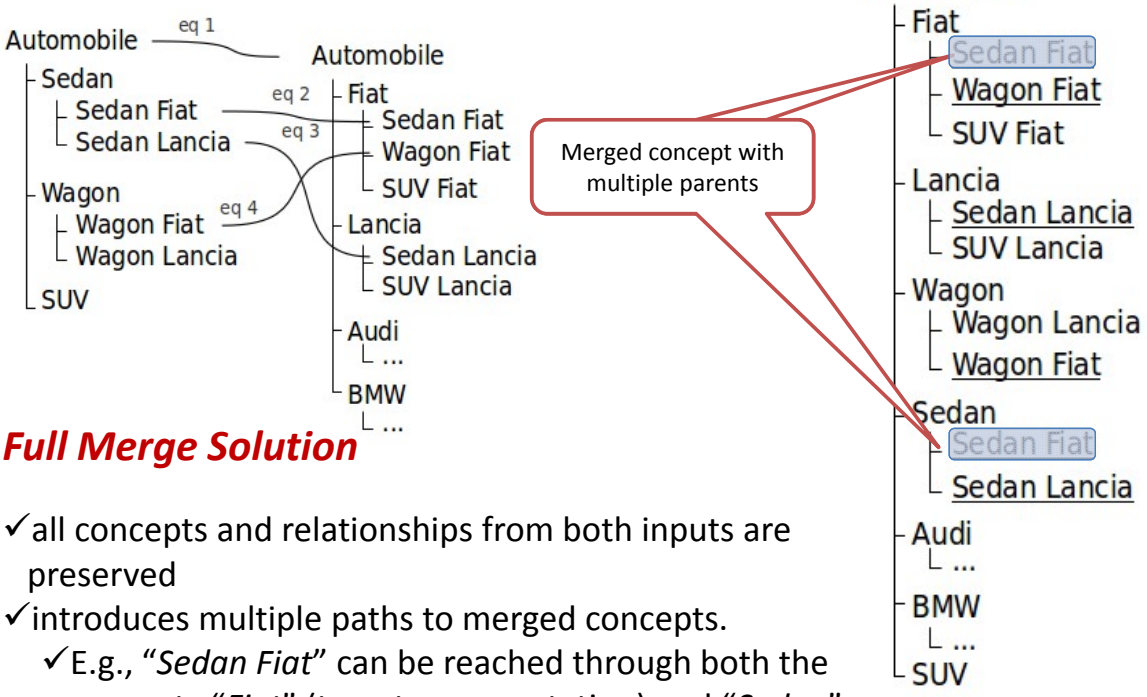


- Process of merging input ontologies into integrated ontology
  - symmetric merge or
  - asymmetric, target-driven merge
- optional use of match mapping between input ontologies

Previous work

- huge amount of work on schema integration
  - mostly addressed both matching and merging
  - complex solutions with high degree of manual interaction
- recent schema merging approaches based on predetermined match mapping
  - [Pottinger and Bernstein 2003], [Pottinger and Bernstein 2008]
  - [Chiticariu, Kolaitis, Popa 2008], [Radvan, Popa , Stanoi, Younis 2009]
  - ...
- relatively few approaches for ontology merging
  - PROMPT (1999-2000), Chimaera (2000), FCA-Merge (2001), ...
  - combined approaches for match and merge
  - high degree of user intervention needed
  - symmetric merge (full preservation of both input ontologies)

Merge example



Full Merge Solution

- ✓all concepts and relationships from both inputs are preserved
- ✓introduces multiple paths to merged concepts.
  - ✓E.g., "*Sedan Fiat*" can be reached through both the concepts "*Fiat*" (target representation) and "*Sedan*"
- ✓reduced understandability by mixing different categorizations

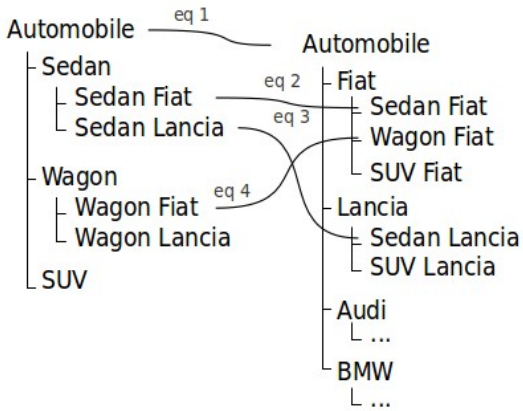
Symmetric Merge

- combines equivalent concepts and maintains all remaining input concepts and relationships of both input ontologies (**Full Merge**)
- maintaining different organizations of the same information can reduce understandability and introduce **semantic overlap**
  - e.g. multiple paths to the same information / multiple inheritance
- **reduced stability** for a preferred input ontology such as mediator ontology
  - e.g. product catalog of a price comparison portal

ATOM approach*

- **A**utomatic **T**arget-Driven **O**ntology **M**erging
- asymmetric, target-driven merge approach
- aims at reduced semantic overlap in merge result
  - preserves target ontology but drops source concepts and relationships that would introduce redundancy in the merge result
- utilization of input match mapping
  - base version: equivalence correspondences
  - improved version: is-a / inverse-is-a correspondences
- automatic generation of default solution(s)
  - result may interactively be adapted by users if needed

* Raunich, S., Rahm, E.: *Target-driven Merging of Taxonomies with ATOM.* Information Systems, 2014



ATOM Solution

Automobile*
- Fiat
  - Sedan Fiat*
  - Wagon Fiat*
  - SUV Fiat
- Lancia
  - Sedan Lancia*
  - SUV Lancia
- Wagon
  - Wagon Lancia
- BMW
  - ...
- Audi
  - ...
- SUV

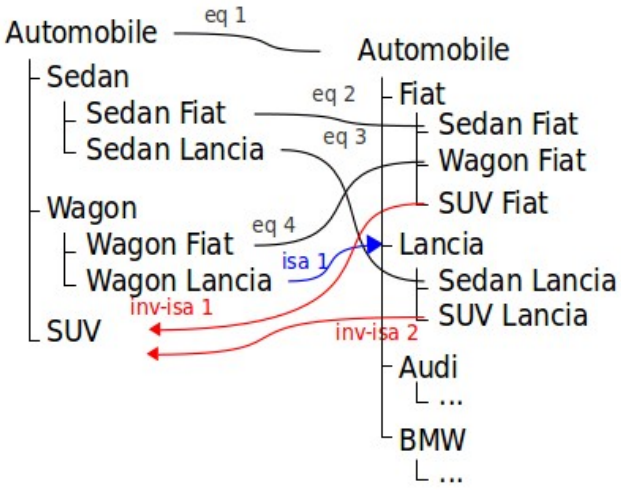
Full Merge Solution

Automobile
- Fiat
  - Sedan Fiat
  - Wagon Fiat
  - SUV Fiat
- Lancia
  - Sedan Lancia
  - SUV Lancia
- Wagon
  - Wagon Lancia
  - Wagon Fiat
- Sedan
  - Sedan Fiat
  - Sedan Lancia
- Audi
  - ...
- BMW
  - ...
- SUV

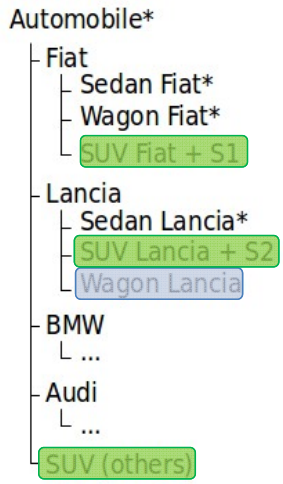
ATOM Solution

- preserves the target ontology
- more compact than the full merge solution
- no multiple inheritance

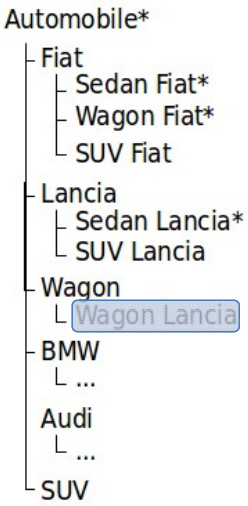
- semantic overlap is only partially reduced
  - some concepts could be better placed (e.g. Wagon Lancia)
  - overlap between general SUV concept and SUV Fiat and SUV Lancia
- more semantic input mapping allows further improvement



ATOM with Extended Mapping



ATOM with only Equivalence Mapping



ATOM Solution with Extended Input Mapping

- ***is-a*** and ***inverse-is-a*** relationships in addition to equivalence correspondences

- concept *Wagon Lancia* is now well placed
- no more overlap between general SUV concept and the more specific concepts SUV Fiat and SUV Lancia

Main Properties of ATOM approach

- P1 – Target Preservation
  - all target concepts and relationships remain in the result
- P2 – Correspondence Preservation
- P3 – Instance Preservation
  - for both input ontologies
  - mappings S-T‘ and T-T‘ specify where instances should migrate
- P4 – Controlled Semantic Overlap
  - for each target (T) concept t in the merge result the number of root paths to t in the merge result T’ is not higher than in T
  - limits multiple paths to leaf nodes / nodes with associated instances

Experimental Results

Merge Example		Anatomy		eBay Catalog	
		Mouse	NCI	v94	v93
Input size	Concepts	2,700	3,300	21,000	23,500
	Leaf paths	2,300	**2,600**	18,400	19,700
# correspondences		~1,000		~20,200	
FULL solution	Concepts	5,100		23,400	
	Leaf paths	**12,900**		21,600	
ATOM solution	Concepts	5,000		23,300	
	Leaf paths	**6,900**		20,400	
Execution time		1 second		7 seconds	

Agenda

- Ontology matching
- Semantic match approach STROMA
  - Enrichment approach and component strategies
  - Evaluation
- Broadening background knowledge
  - Extracting semantic relations from Wikipedia
  - Semantic mapping repository SemRep
- Ontology merging
  - Symmetric vs. asymmetric merge
  - ATOM approach
- Conclusions and outlook

Conclusions and outlook

- STROMA: Determine relation type of correspondences
  - generic strategies: Compounds, itemizations, …
  - background knowledge decisive for semantic matching
- Extension of existing background knowledge
  - automatic extraction of semantic relations from Wikipedia
  - integrating extracted relations and existing resources in SemRep
  - use of SemRep can significantly improve match quality
- Ontology merging
  - can be improved by providing semantic ontology mappings
  - target-driven approach like ATOM supports largely automatic approach and compact merge results

Some open challenges

- improve quality of extracted semantic relations
- better repository support for homonyms (e.g. table)
- domain-specific solutions for semantic matching (e.g., life sciences)
- evaluation of merge quality (many valid solutions)
- holistic merging of (>>2) taxonomies/ontologies

References

- Arnold, P.; Rahm, E.: Semantic Enrichment of Ontology Mappings: A Linguistic-based Approach. Proc. ADBIS Conf., 2013
- Arnold, P.; Rahm, E.: Enriching Ontology Mappings with Semantic Relations. Data and Knowledge Engineering. 2014
- Arnold, P.; Rahm, E.: Extracting Semantic Concept Relations from Wikipedia. Proc. 4th Int. Conf. Web Intelligence, Mining and Semantics (WIMS), 2014
- Arnold, P.; Rahm, E.: SemRep – A Repository for Semantic Mapping. Univ. of Leipzig., submitted for publiction. 2014
- Bellahsene, Z.; Bonifati, A.; Rahm, E. (eds.): Schema Matching and Mapping. Springer-Verlag, 2011
- Rahm, E.: Towards large-scale schema and ontology matching. In: Schema Matching and Mapping, Springer-Verlag, 2011
- Raunich, S., Rahm, E.: ATOM: Automatic Target-driven Ontology Merging, Proc. ICDE 2011
- Raunich, S., Rahm, E.: Target-driven Merging of Taxonomies with ATOM. Information Systems, 2014