Mapping Composition for Matching Large Life Science Ontologies

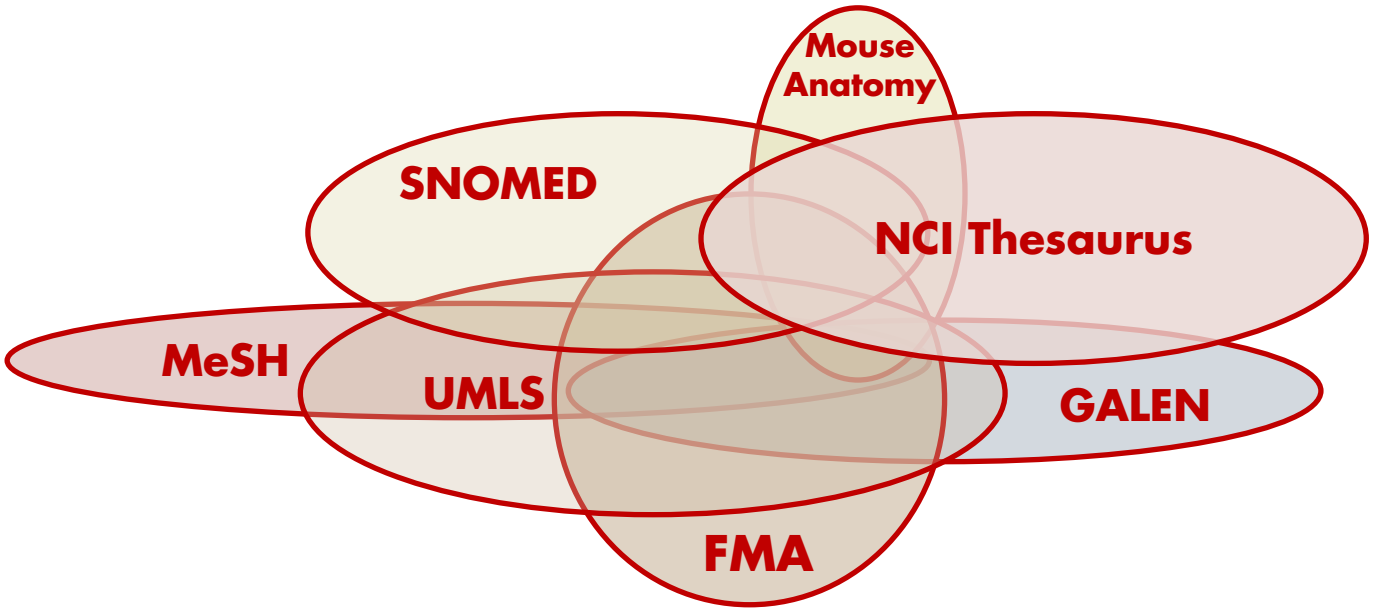
Anika Gross, Michael Hartung,
Toralf Kirsten, Erhard Rahm

UNIVERSITÄT LEIPZIG

29th July 2011, ICBO, Buffalo

Ontologies

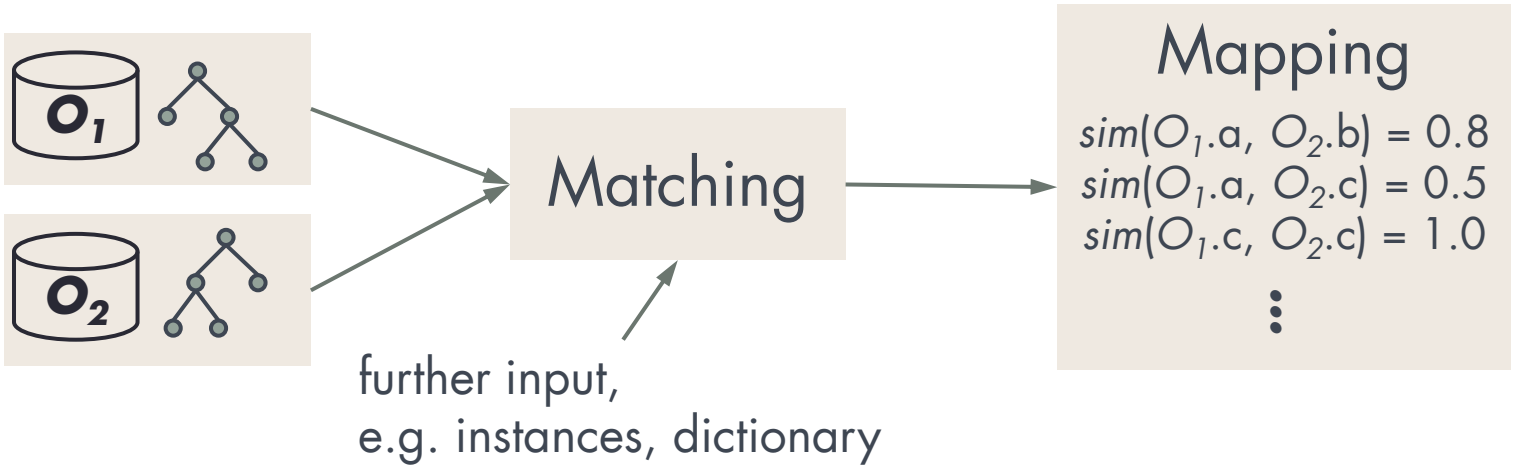
- Multiple interrelated ontologies in a domain (e.g. anatomy)



- Identify overlapping information between ontologies
- Create mappings

Ontology Matching

- Manual creation of mappings between very large ontologies is too labor-intensive
- Semi-automatic generation of semantic correspondences (linguistic, structural, instance-based ontology matching)

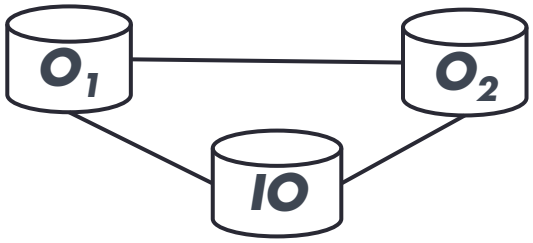


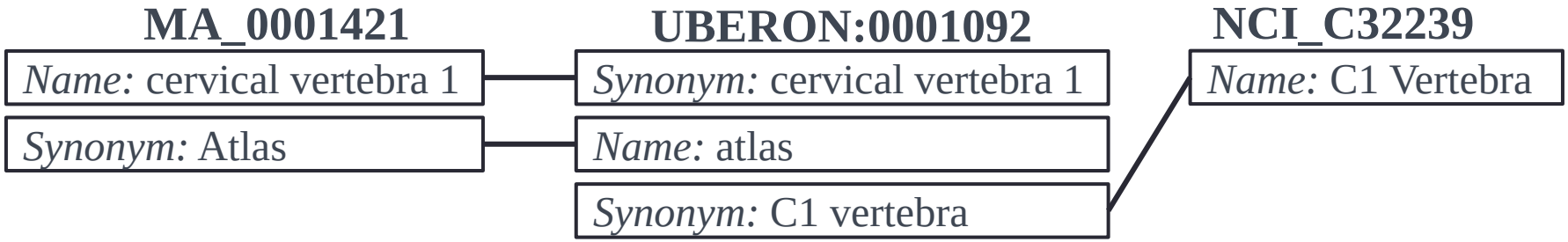
→ Interrelation of ontologies

→ Integration of heterogeneous information sources

COMPOSING

- Indirect composition-based matching
- Via intermediate ontology (*IO*): important hub ontology, synonym dictionary, …





→Find new correspondences via composition

→Reuse existing mappings to

→Increase match quality

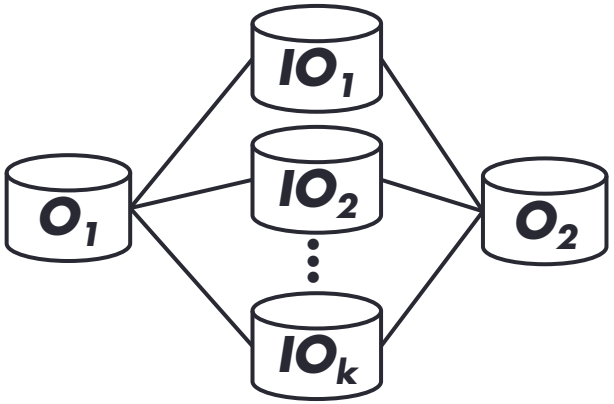
→Save computation time

CONTRIBUTIONS

- Composition-based ontology matching approach, reuse of previously determined mappings →*composeMatch*
- Optional match step to improve composition-based match quality → *extendMatch*
- Use of ontology and mapping operators: compose, match and extract
- Evaluation: indirect matching of *MA* and *NCIT* using large intermediate ontologies (*UMLS*, *FMA*, *Uberon*, *RadLex*)

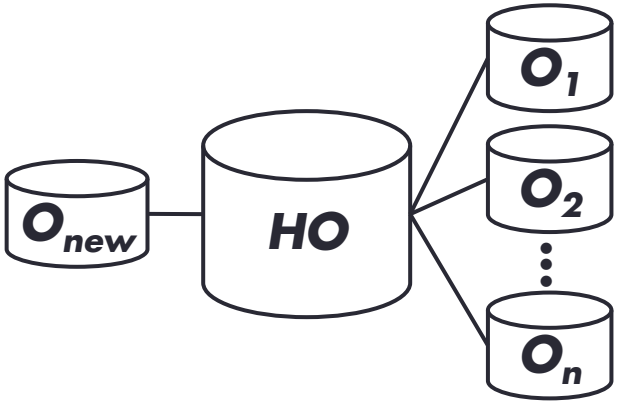
Indirect Matching

- Use mappings to intermediate ontologies IO_1, …, IO_k to indirectly match O_1 and O_2
- Reduce matching effort by reusing mappings to IO
  → very fast composition



→ IO should have a significant overlap with O_1 and O_2

→ IO_1, …, IO_k may complement each other



→ Centralized hub HO

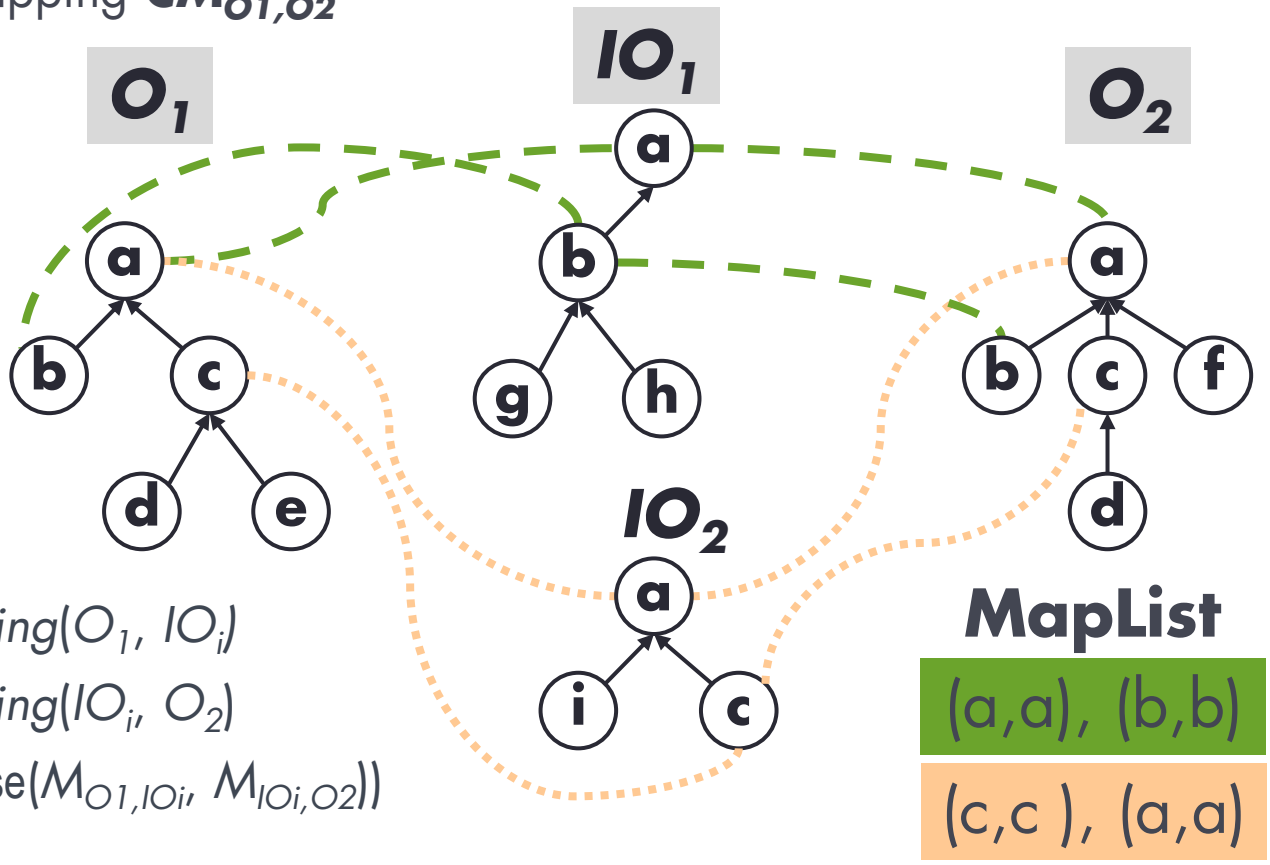
→ many mappings to other ontologies

→ O_{new} aligned with any O_i via HO

(1) ComposeMatch

Input: Two ontologies O_1 and O_2, list of intermediate ontologies $IO_1 \ldots IO_k$, occurrence count ***occ***

Output: Composed mapping $CM_{O1,O2}$



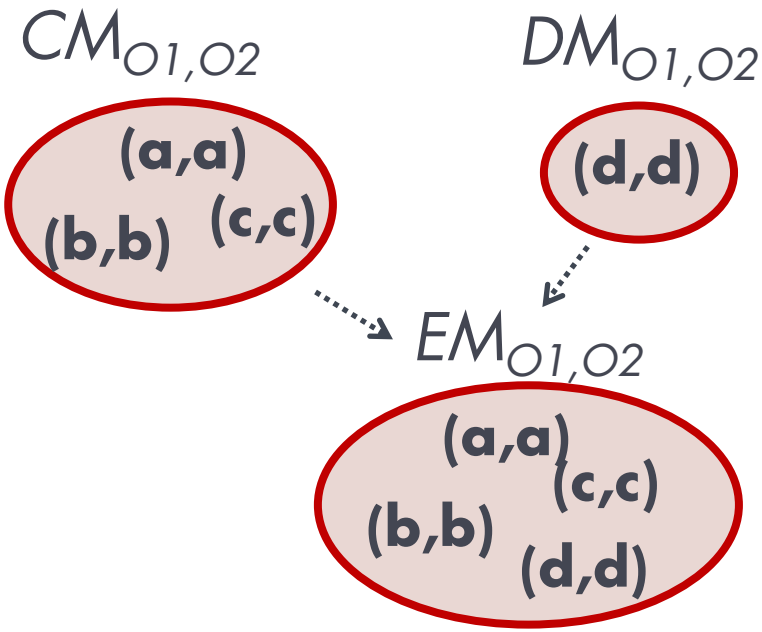
MapList ← empty
for each $IO_i \in IO$ **do**
  $M_{O1,IOi} \leftarrow getMapping(O_1, IO_i)$
  $M_{IOi,O2} \leftarrow getMapping(IO_i, O_2)$
  MapList.add(compose($M_{O1,IOi}$, $M_{IOi,O2}$))
end for
return merge(*MapList, occ*)

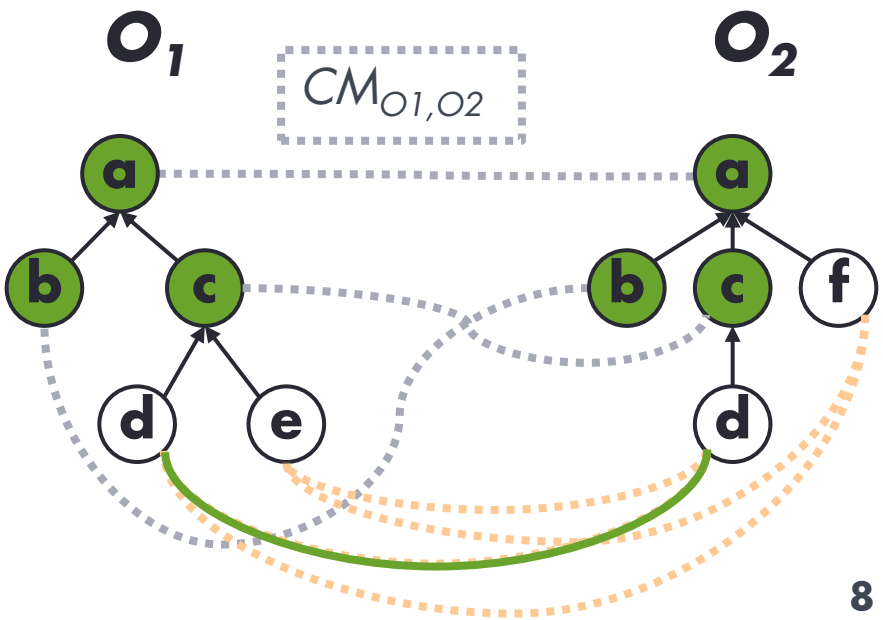
occ = 1: $CM_{O1,O2}$ = {(a,a),(b,b),(c,c)}
occ = 2: $CM_{O1,O2}$ = {(a,a)}

(2) ExtendMatch

Input: Two ontologies O_1 and O_2, composed mapping $CM_{O1,O2}$
Output: Extended Mapping $EM_{O1,O2}$

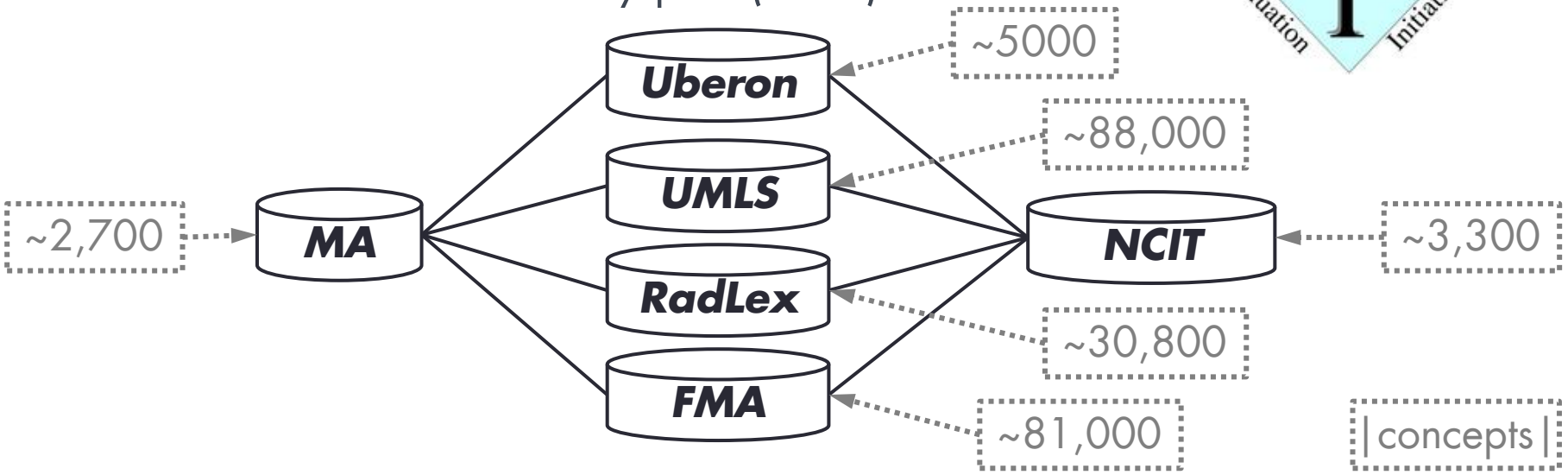
$\Delta O_1 \leftarrow$ extract(O_1, $CM_{O1,O2}$)
$\Delta O_2 \leftarrow$ extract(O_2, inverse($CM_{O1,O2}$))
$DM_{\Delta O1 \Delta O2} \leftarrow$ match(ΔO_1, ΔO_2) *//Direct Mapping*
$EM_{O1,O2} \leftarrow$ merge({$CM_{O1,O2}$, $DM_{\Delta O1 \Delta O2}$}, 1)
return $EM_{O1,O2}$





Evaluation Setup

- Match problem
  - Adult Mouse Anatomy (MA)
  - NCI Thesaurus Anatomy part (NCIT)



- Gold standard ~1500 correspondences
- Precompute mappings using a match strategy

Preprocessing
Normalization

→

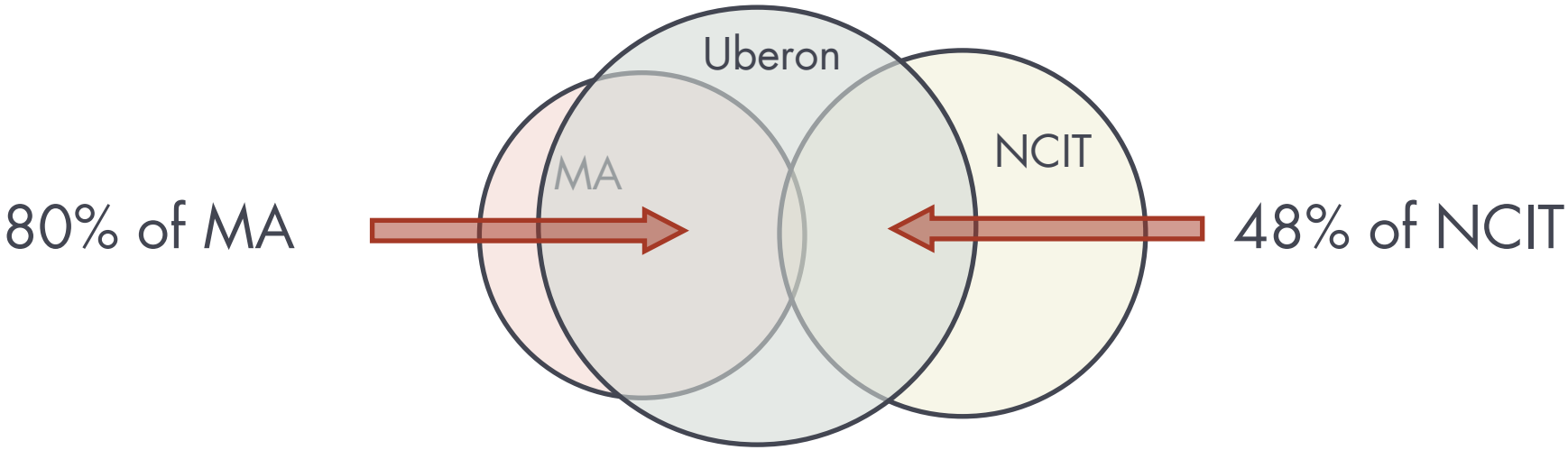
Linguistic Matcher
(Name, synonyms, Trigram t = 0.8)

→

Selection & Postprocessing

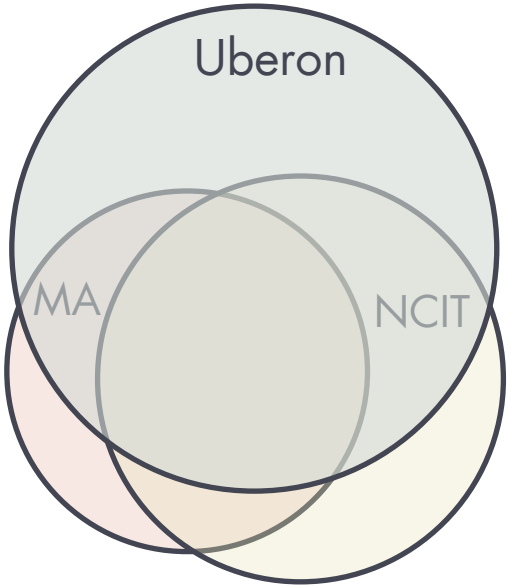
Mapping Statistics

Is there a good coverage of MA and NCIT by intermediate ontologies?



Mapping Statistics

Is there a good coverage of MA and NCIT by intermediate ontologies?



High overlap of covered MA and NCIT concepts
→ promising for composition-based match results

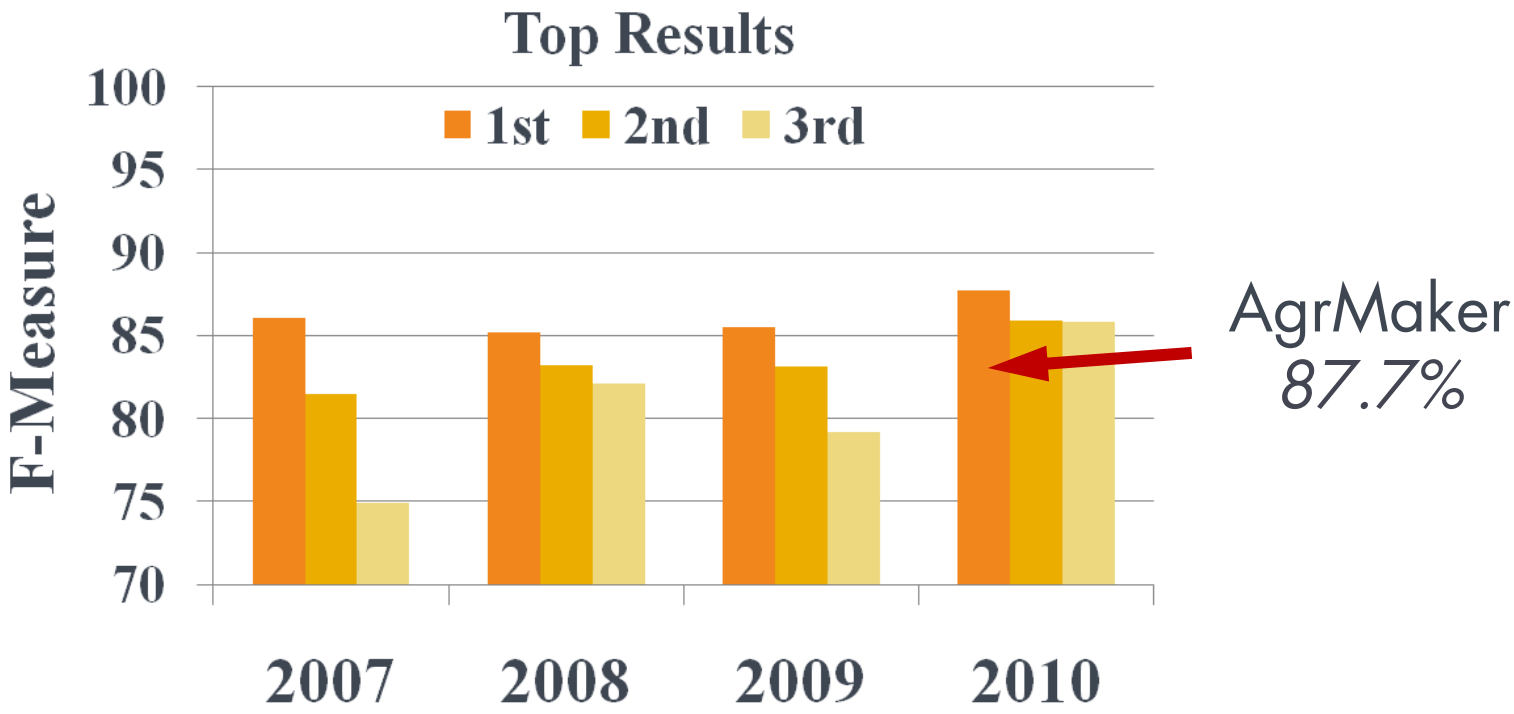
Mapping Statistics

Is there a good coverage of MA and NCIT by intermediate ontologies?

Mapping	Coverage Domain	Coverage Range	Mapping size
MA-Uberon	**80%**	45%	2300
Uberon-NCIT	33%	**48%**	1703
MA-UMLS	**69%**	3%	2975
UMLS-NCIT	5%	**87%**	4214
MA-RadLex	**39%**	3%	1082
RadLex-NCIT	4%	**40%**	1347
MA-FMA	**57%**	2%	1601
FMA-NCIT	3%	**67%**	2337

High overlap of covered MA and NCIT concepts
→ promising for composition-based match results

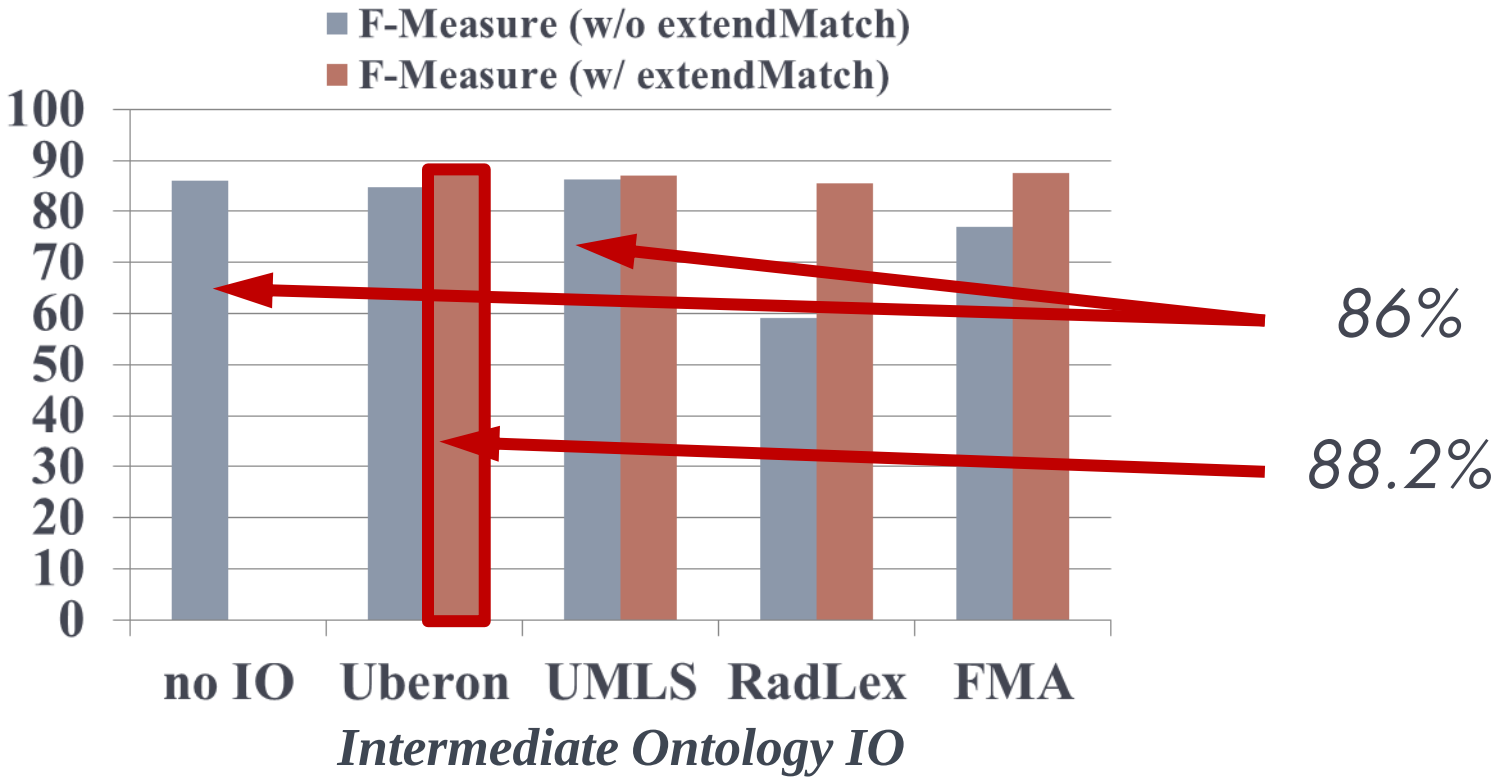
OAEI Anatomy Track



http://oaei.ontologymatching.org/[year]/anatomy

EVALUATION

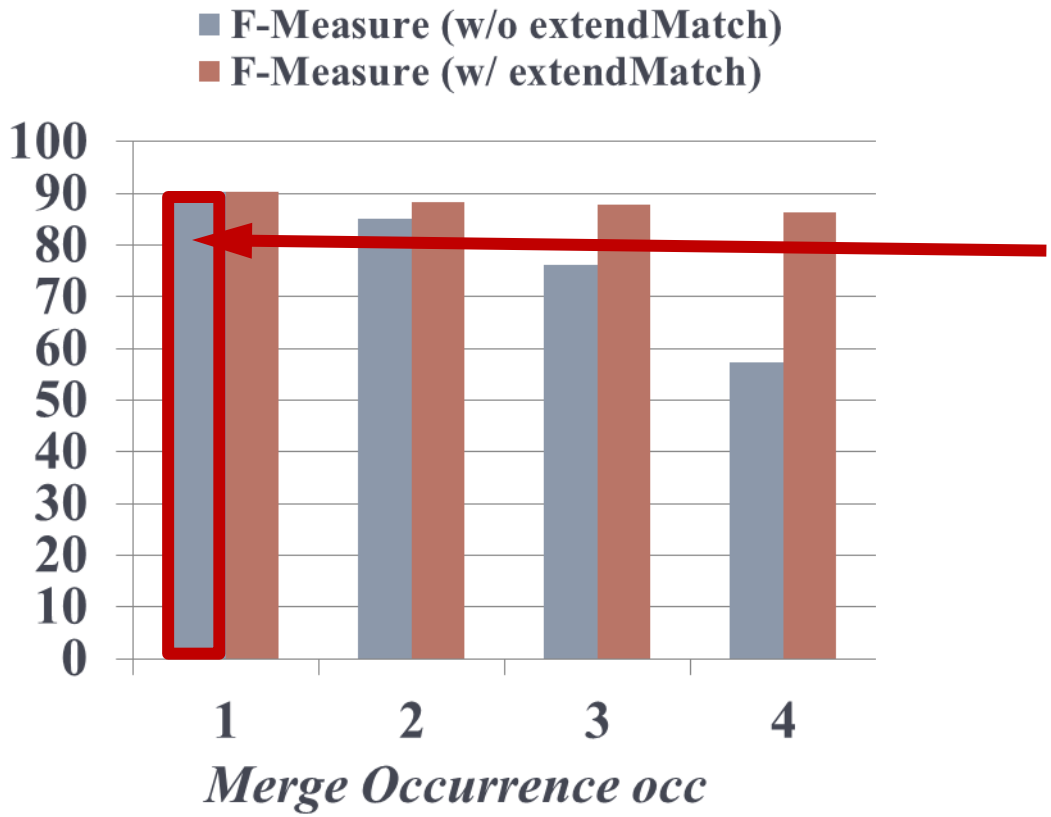
- Direct match result compared to *composeMatch* via each hub
- Additional matching of unmatched parts (*extendMatch*)



- Uberon & UMLS → best evaluated intermediate ontologies

Evaluation

- Combination of the four composed mappings
- Correspondences have to occur in at least 1,…,4 mappings



union(occ=1)

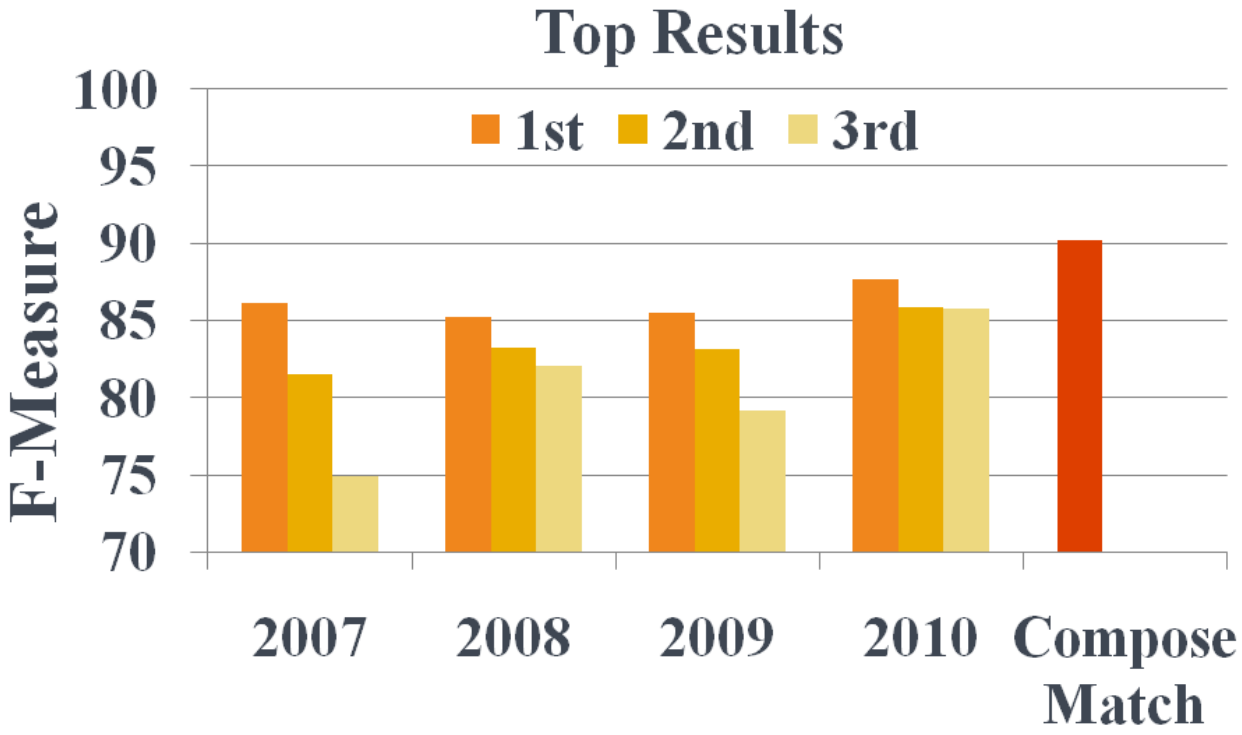
F-Measure	**90.2**
Precision	92.7
Recall	87.8

Higher occurrences
→ Recall ↓

extendMatch
→ Recall ↑

Evaluation

- Combination of the four composed mappings
- Correspondences have to occur in at least 1,…,4 mappings



Conclusions

- Composition-based approach for indirect matching of large life science ontologies (*composeMatch*, *extendMatch*)
- Reuse mappings for improved match efficiency and quality (>90%)
- Evaluated several intermediate ontologies
  → Uberon and UMLS: very effective, suited as hub ontologies in the anatomy domain

Future Work

- Investigate composition-based ontology matching for further domains

- Study the impact of additional mappings
  - Determined by structural matching
  - Existing mappings from BioPortal

Mapping Composition for Matching Large Life Science Ontologies

Funding

German Research Foundation DFG Deutsche Forschungsgemeinschaft

Grant RA497/18-1

"Evolution of Ontologies and Mappings"