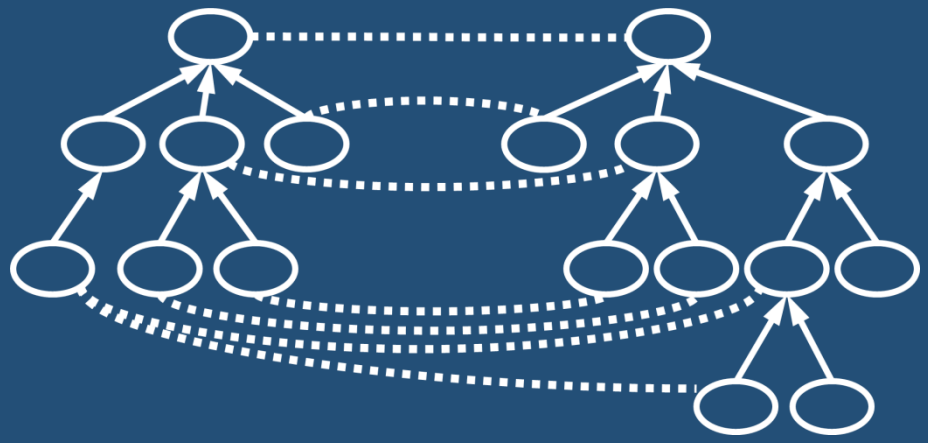


GOMMA RESULTS FOR OAEI 2012

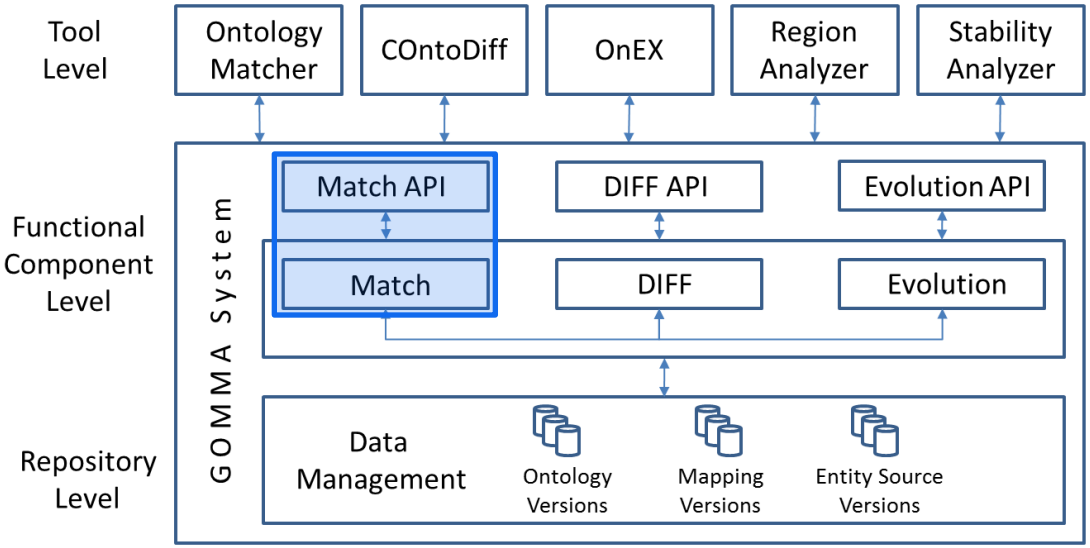
ANIKA GROSS, MICHAEL HARTUNG,
TORALF KIRSTEN, ERHARD RAHM

UNIVERSITÄT LEIPZIG

11TH NOVEMBER 2012, OM WORKSHOP, BOSTON

GOMMA

- **G**ENERIC **O**NTOLOGY **M**ATCHING AND MAPPING **MA**NAGEMENT
- Comprehensive infrastructure to manage and analyze the evolution of *life science* ontologies and mappings

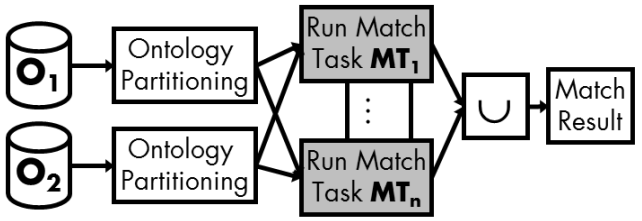


- Generic match component to semantically align ontologies

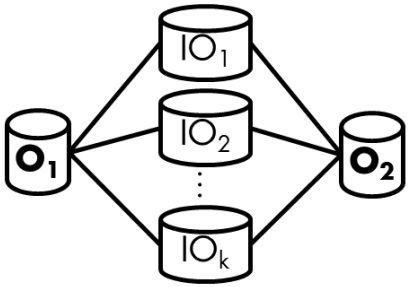
→ Could participate in 6 SEALS tracks

GOMMA's Scalable Match Techniques

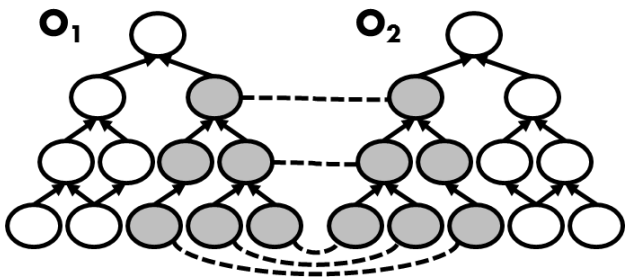
- Parallel ontology matching on multiple computing nodes and CPU cores



- Indirect computation of ontology mappings by reusing and composing previously determined ontology mappings via intermediate ontologies

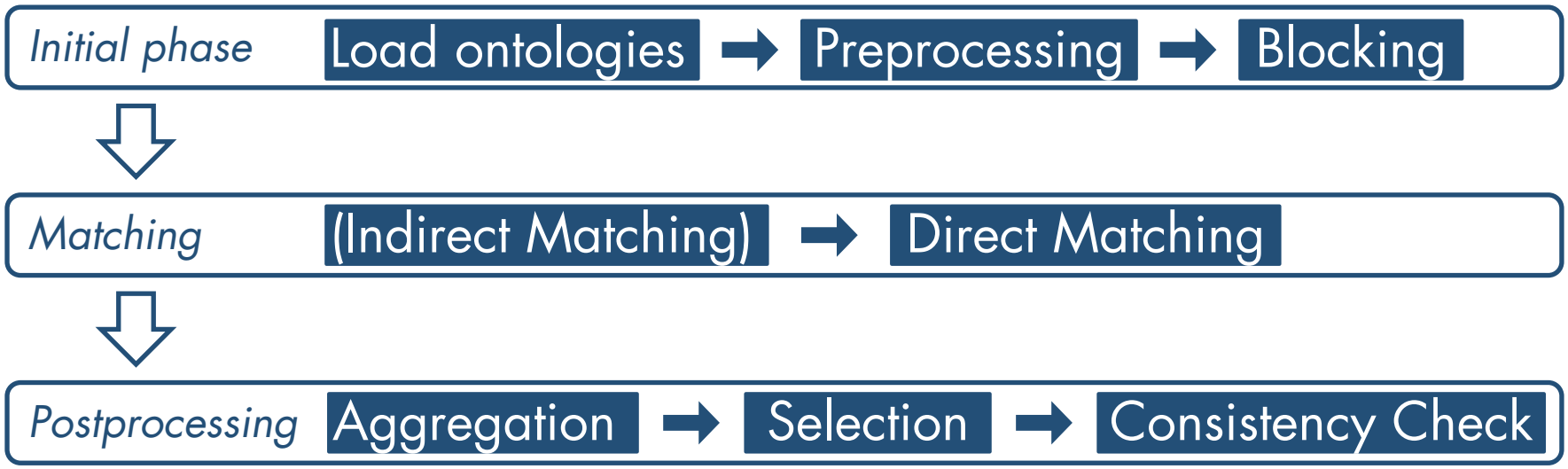


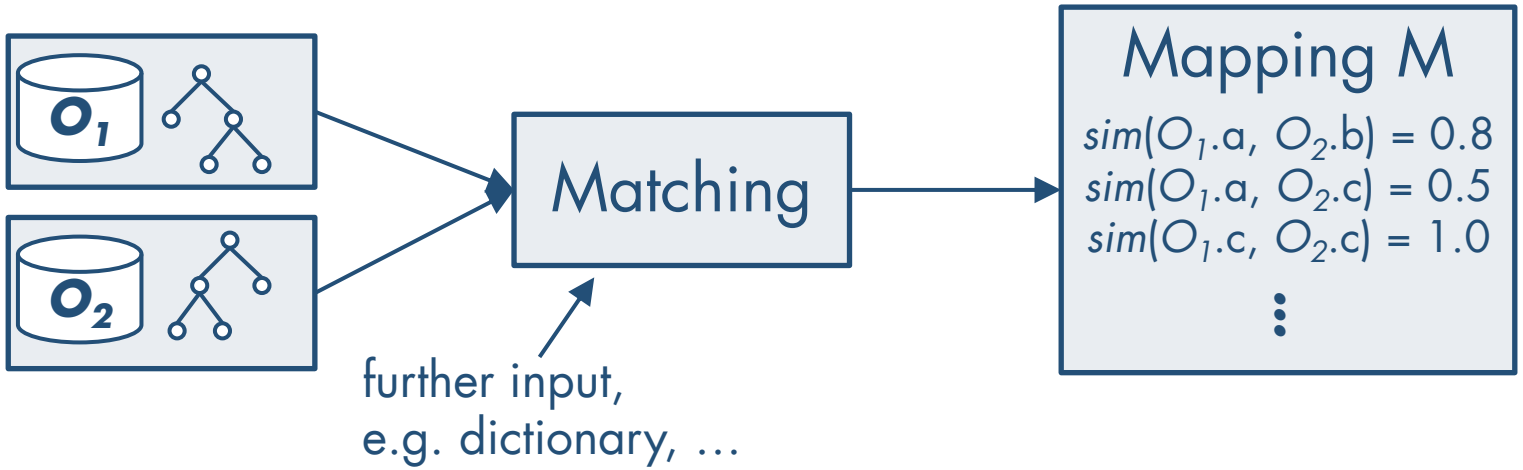
- Reduction of search space (blocking) by restricting matching to overlapping ontology parts



→ Efficient and effective matching of very large ontologies

GOMMA Matching Workflow





Initial Phase

- Parse and load ontologies
- Assign all relevant information to concepts
- Text attributes for string-based comparison (name, synonym, instance, …)
- Preprocessing (Normalization, Translation …)

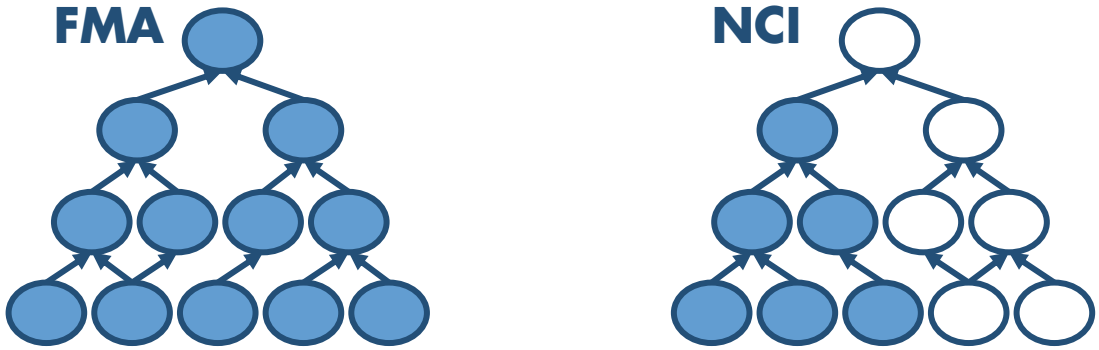
Translation for non-English ontologies:

- Use translation API (`http://mymemory.translated.net/`) to iteratively established a dictionary for non-English terms
- Add translated terms as synonyms

```
ID:      http://iasted_fr#c-1203110-3646755
Name:    pause café
Synonym: coffee break
Synonym: break
```

BLOCKING

- Aim: Reduce number of comparisons for large ontologies
- Useful for "asymmetric" match problems:
  match a specific ontology to a broader ontology (from which only a part is relevant)



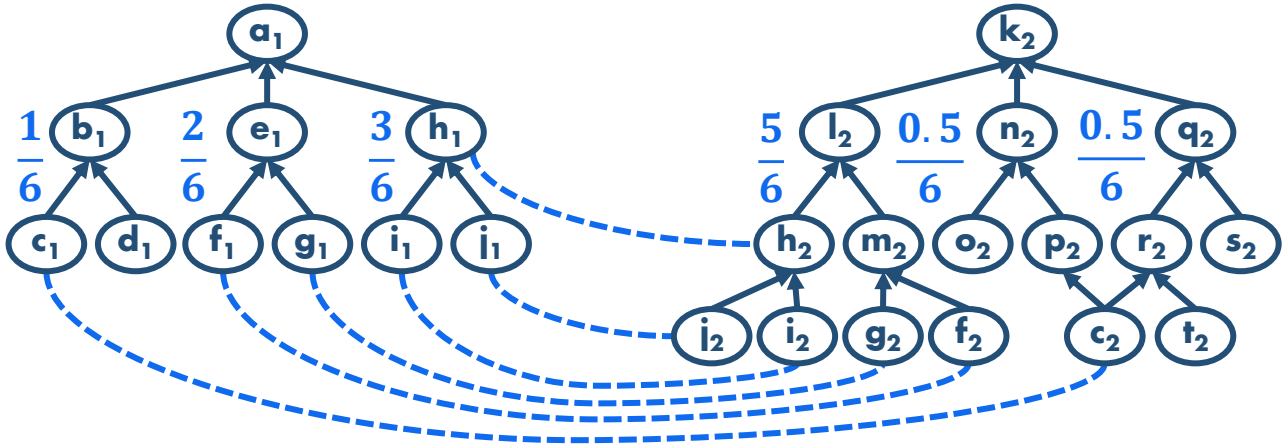
- Automatically identify the relevant part of the broader ontology
- Match only this part with the more specific ontology

→ Can dramatically improve efficiency in applicable cases

→ Improve match quality due to fewer false positive correspondences

Blocking

1. Identify $M_{initial}$ (efficient match method)
2. Identify a set of subgraph roots below the top root and propagate correspondence counts from the leaf level upwards to the roots
3. Compute correspondence fractions
4. Select most valuable root(s), concepts in subgraphs used for matching; No root exceeds the threshold → blocking not applied

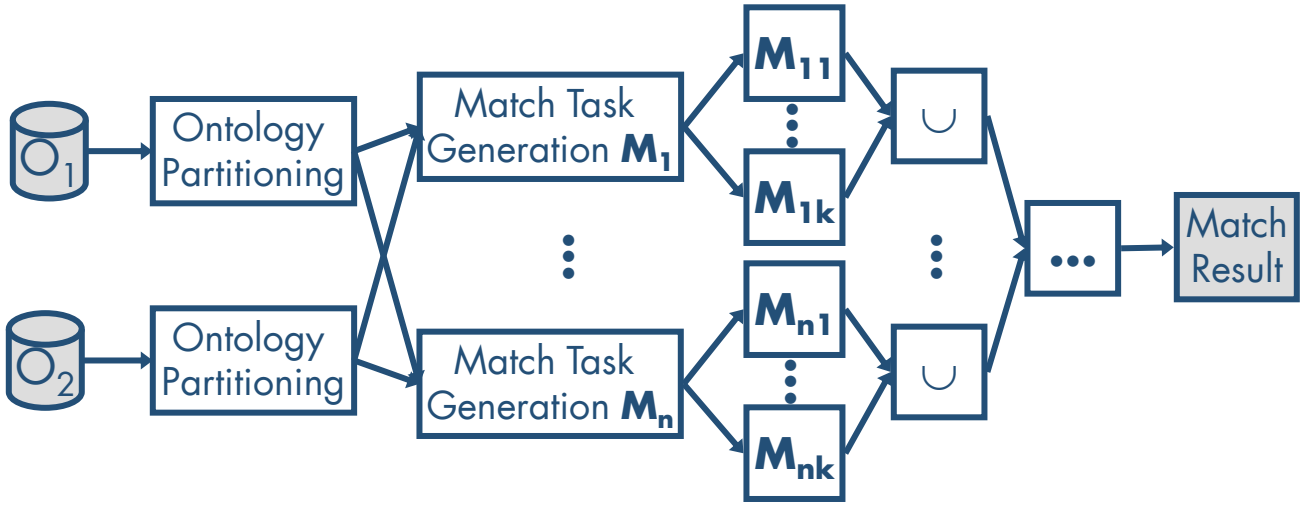


$$corrFrac(root) = \frac{|M_{initial}(subgraph(root))|}{|M_{initial}|}$$

Direct Matching

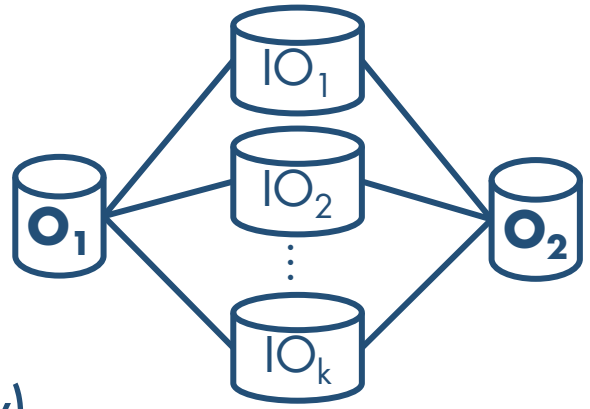
- Use of internal ontology knowledge like concept associated information
- *NameSynonym* matcher: determine the maximal string similarity for names and multi-valued synonyms per concept pair
- Optionally (if available):
  apply a *Comment* matcher and *Instance* matcher

- Intra-matcher parallelization: (for OAEI only threading)



Indirect Matching

- Composition-based matching
- Aim: Reuse existing high quality mappings to efficiently match two so far unmatched ontologies
- Composing mappings via one or more intermediate "hub" ontologies (*IO*)



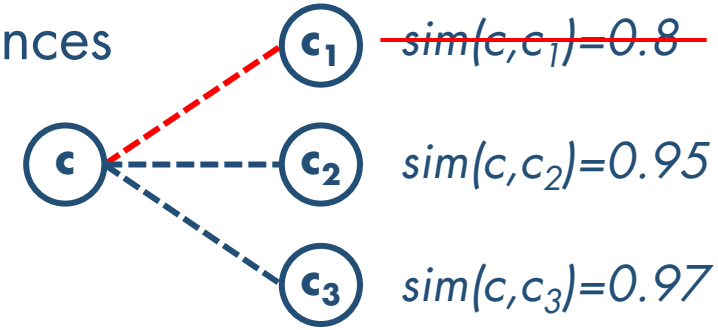
- For OAEI: Precompute several mappings (using the direct match strategy) from source and target to different *IO*s and compose these mappings

- Result mapping might still be incomplete:

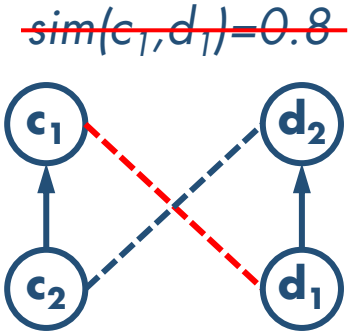
→ *Extend* result mapping: Identify unmatched source and target concepts and match them directly

POSTPROCESSING

- Combination of directly and indirectly determined mappings (union mappings, take average of similarity values)
- Select most likely correspondences
  - Similarity threshold
  - *MaxDelta* selection



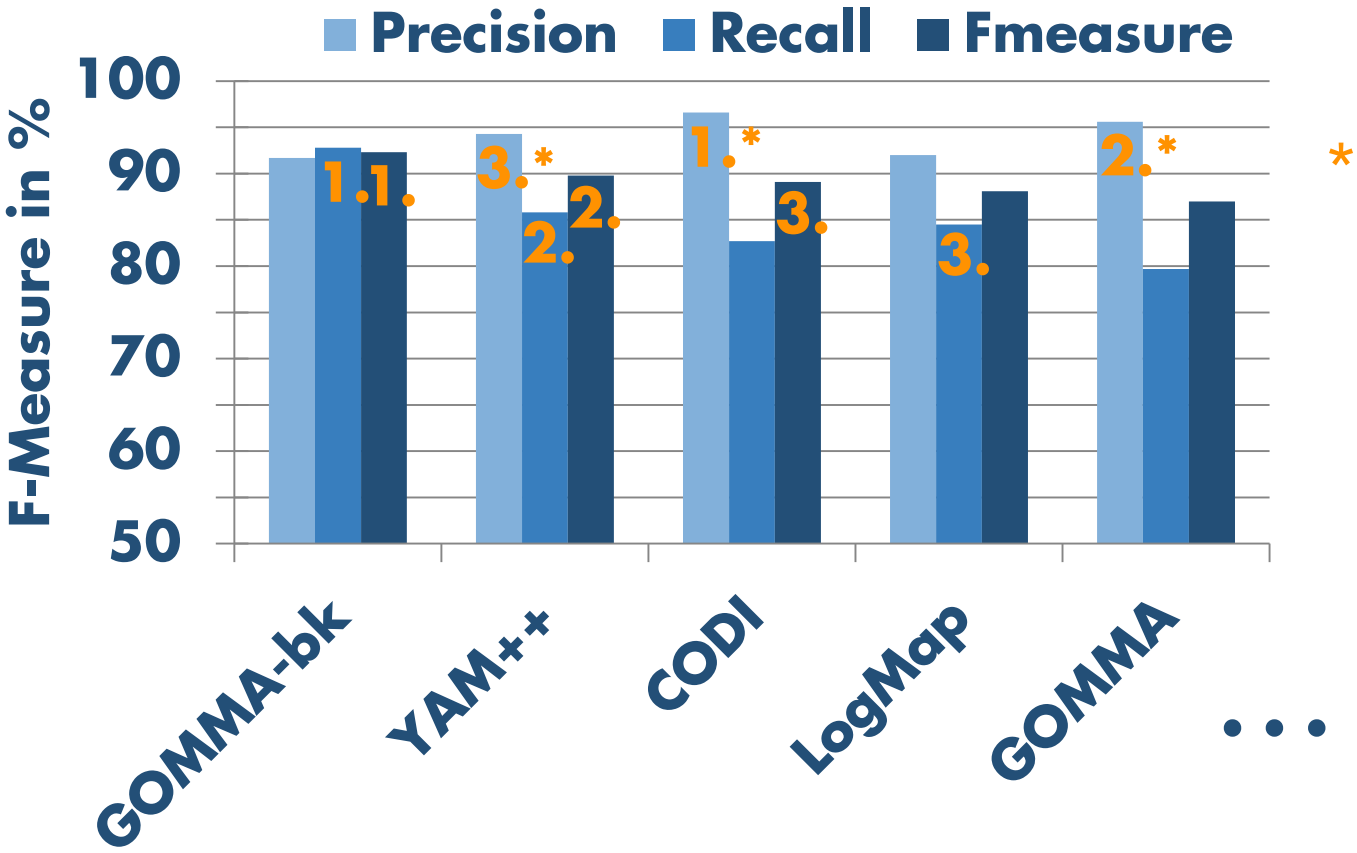
- Consistency checking
  - Remove CrissCross
  - Datatype Consistency
  - ParentChild Extension
  - Property Extension



Evaluation Results

- GOMMA participated in
  - Anatomy
  - Large Biomedical Ontologies
  - Library
  - Conference
  - Multifarm
  - Benchmarks

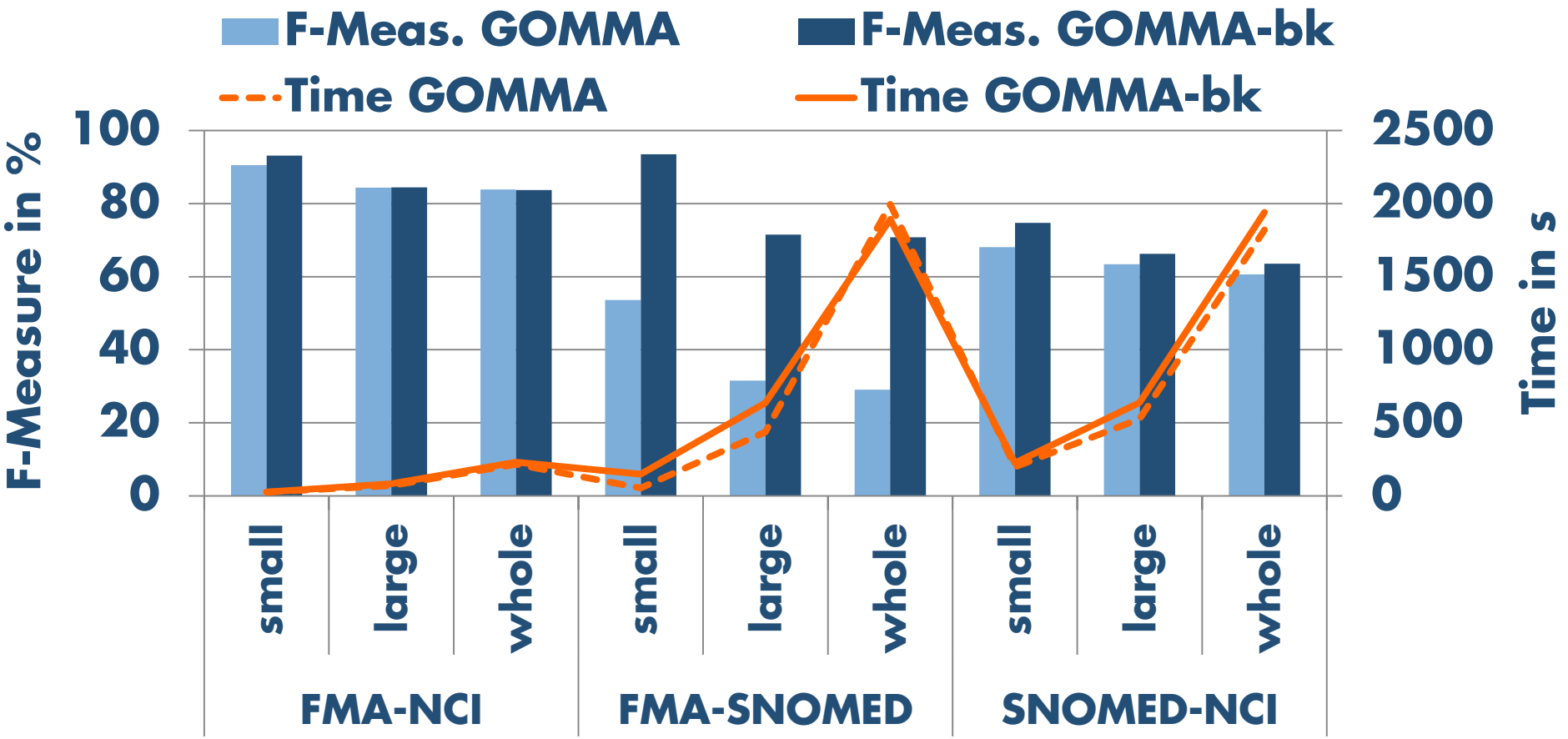
ANATOMY



* In Top5

- Most systems favor precision over recall
- Highest recall: GOMMA-bk
  - Composition and reuse of mappings to UMLS, Uberon and FMA
- Best F-Measure = 92.3
- GOMMA Runtime: 15-17 seconds

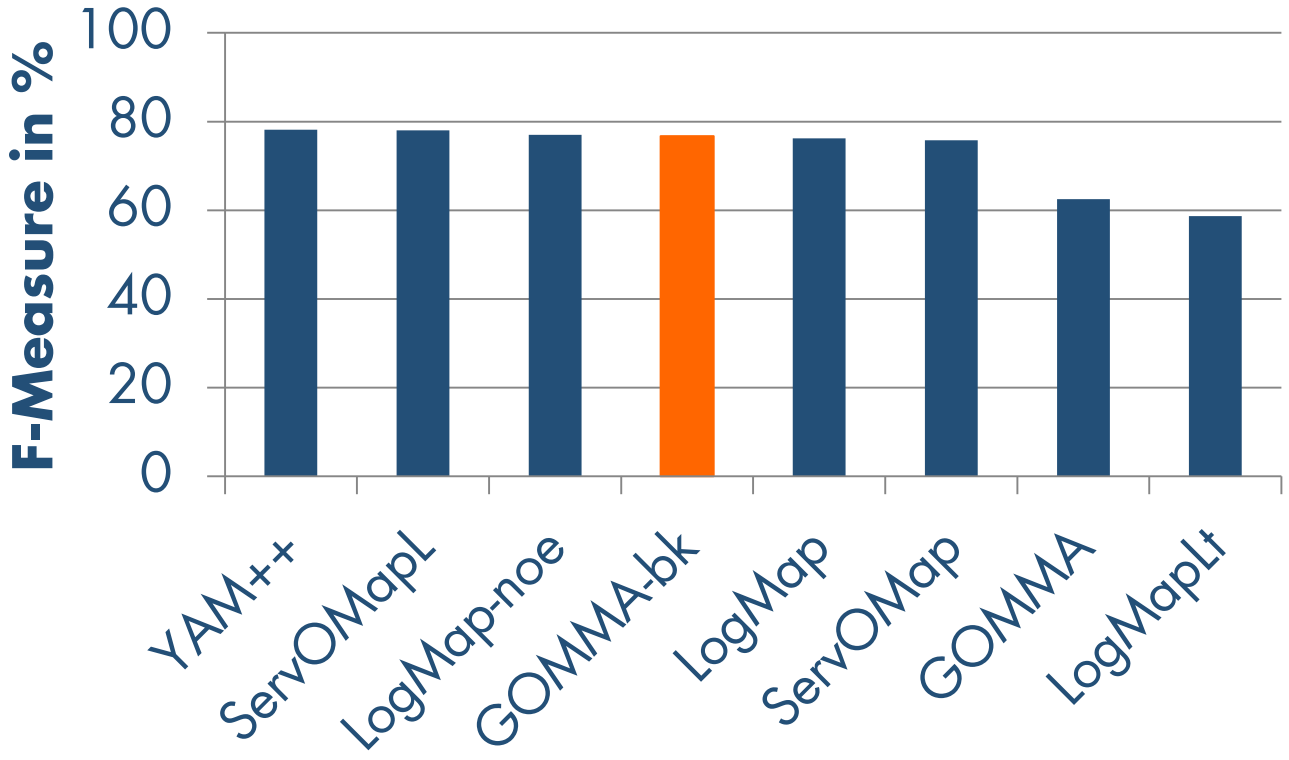
Large Biomedical Ontologies



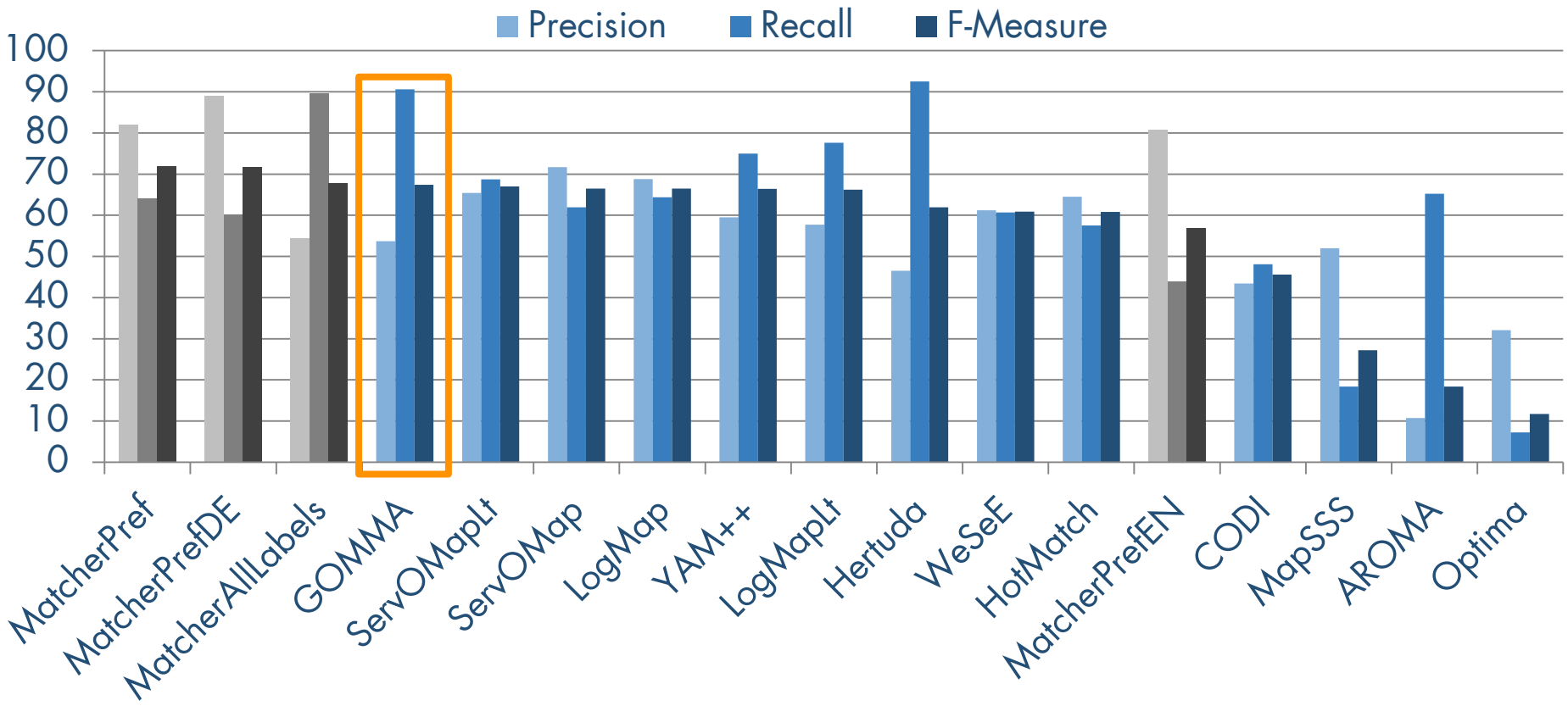
- SNOMED-related tasks more difficult
- GOMMA-bk: F-Measure ↑, best F-Measure for small tasks (up to 94%)
- Blocking and parallel matching useful to achieve good runtimes: 97 min for all 9 tasks

Large Biomedical Ontologies

- 15 out of 23 participating systems/configurations solved at least one subtask
- 8 systems could complete all 9 *largeBio* tasks:



LIBRARY



- No system is better than three basic string-based strategies provided by the organizers
- GOMMA: (marginal) best F-measure for participating systems (67.4%)
- Especially high recall → similar to basic strategy *MatcherAllLabels*

≈ GOMMA takes maximum similarity for name and synonyms

Summary

- GOMMA achieves very good quality with good runtimes
- Best system for Anatomy & Library
- Among top systems for LargeBio,Conference,Multifarm&Benchmark

GOMMA's strength

- Scalable matching due to blocking, parallel matching and mapping composition
- Improvement of match quality by using domain knowledge
  - Mapping composition via domain-specific hub ontologies
  - Application of multi-language translation services for improved synonyms

Future Improvements

- Additional consistency checks
- Improved blocking techniques → reduction of search space

GOMMA RESULTS FOR OAEI 2012

Funding: German Research Foundation Grant RA497/18-1
"EVOLUTION OF ONTOLOGIES AND MAPPINGS"