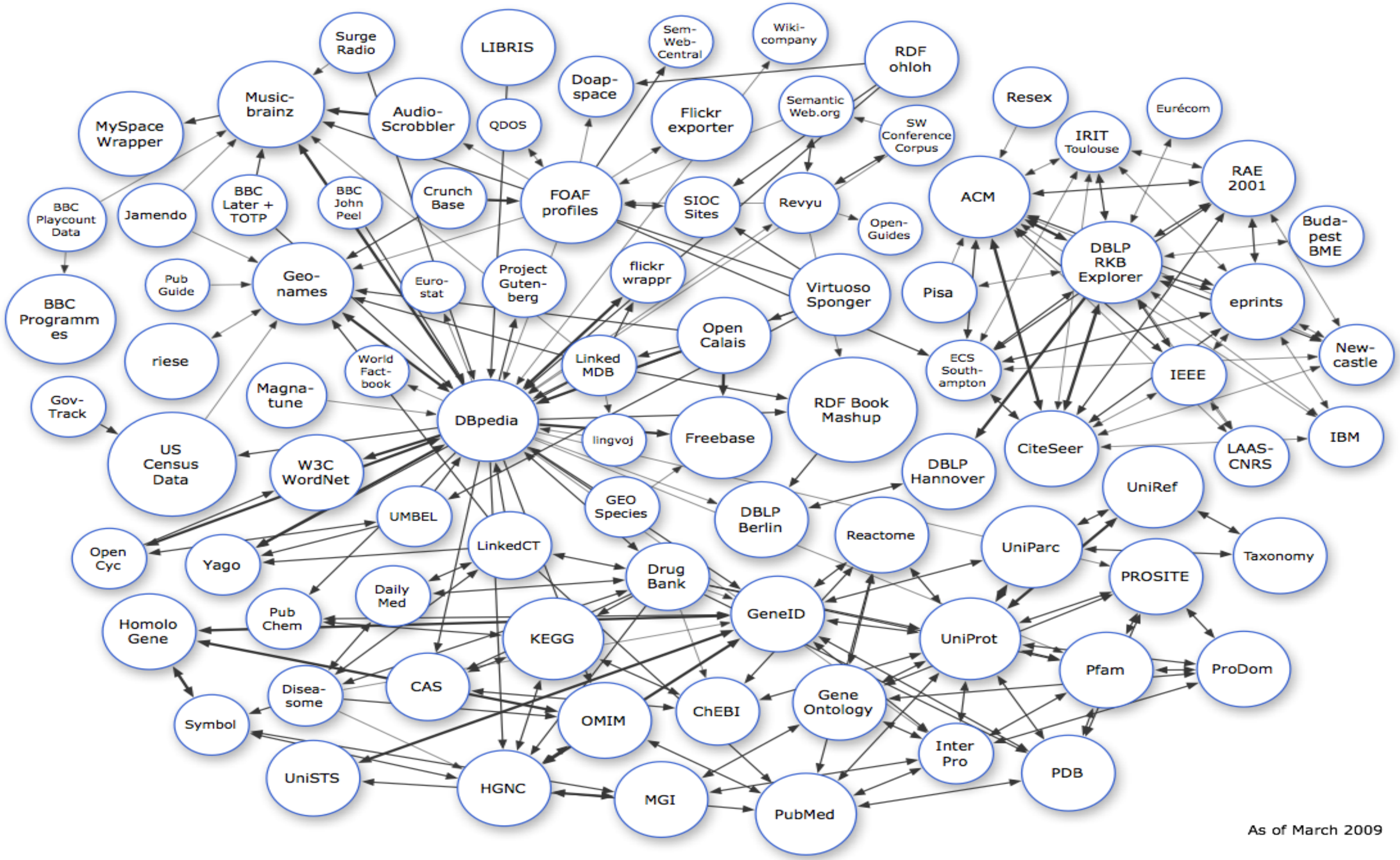
Towards data fusion in a multi-ontology environment

Andriy Nikolov
Victoria Uren
Enrico Motta

Public linked data



Issues

- Pairwise linking of datasets
  - Scale will grow
  - More effort needed to include “yet another” dataset to the cloud
- Automation would be useful

Challenges

- Instance matching
  - Aggregated attribute similarity
  - Usually configured manually for each pair of datasets and for each class
    - SILK, LinkedMDB,…
- Schema heterogeneity
  - Which datasets overlap?
  - Which attributes to compare?
- Employ automatic schema matching

Schema mismatches [11]

- Scope
  - dbpedia:Company vs sweto:Company ∩ sweto:Bank
- Granularity
  - foaf:Person vs dbPedia:Politician
- Modelling style
  - “red” vs #FF0000
- Terminological
  - Company vs Corporation

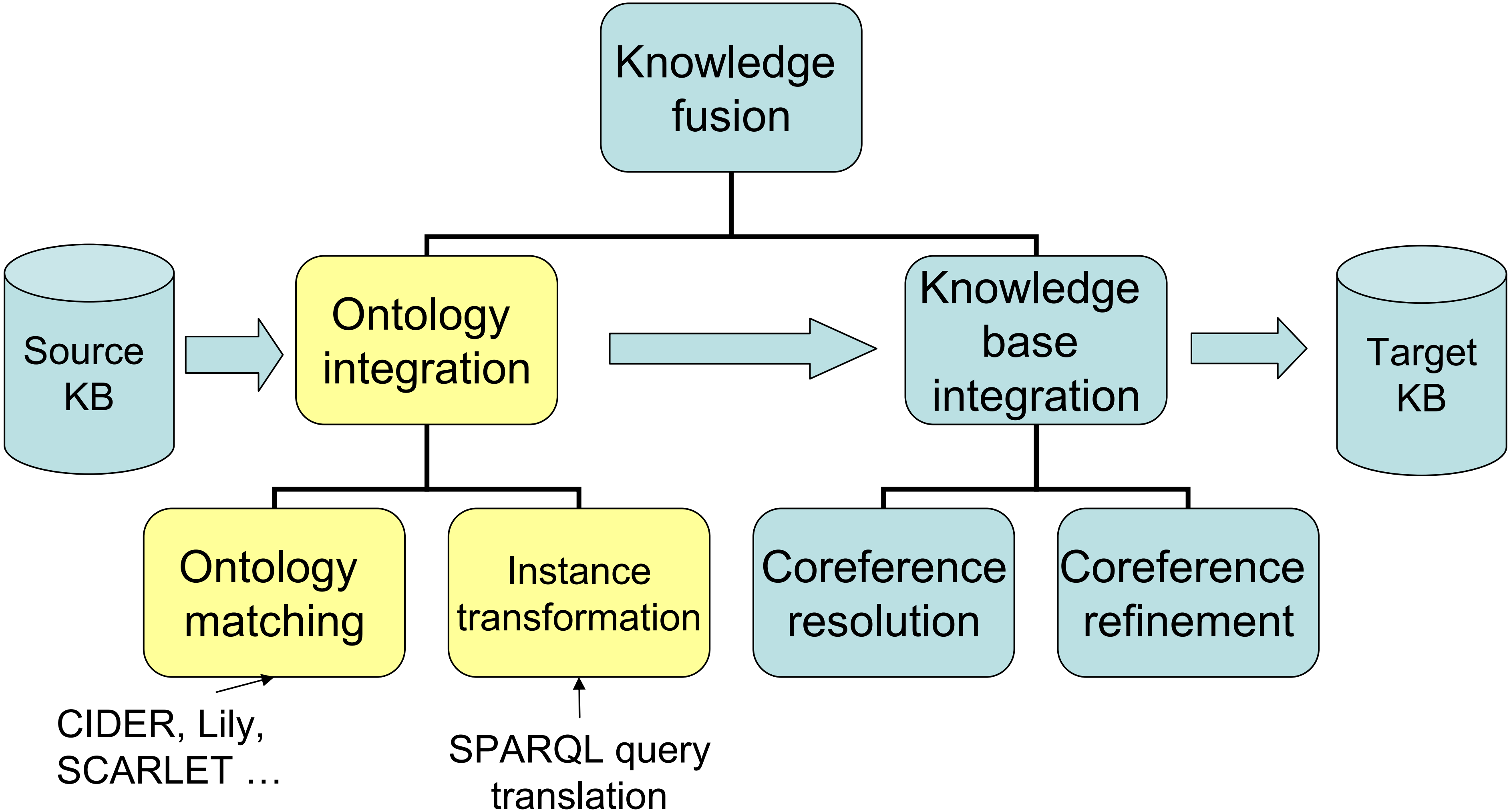
Schema matching

- Many existing tools (OAEI)
  - Lily
  - Falcon-AO
  - CIDER,
  - …
- Features
  - Produce DL relations between concepts and attributes ($\equiv$, $\sqsubseteq$)
  - Focus on terminological mismatches

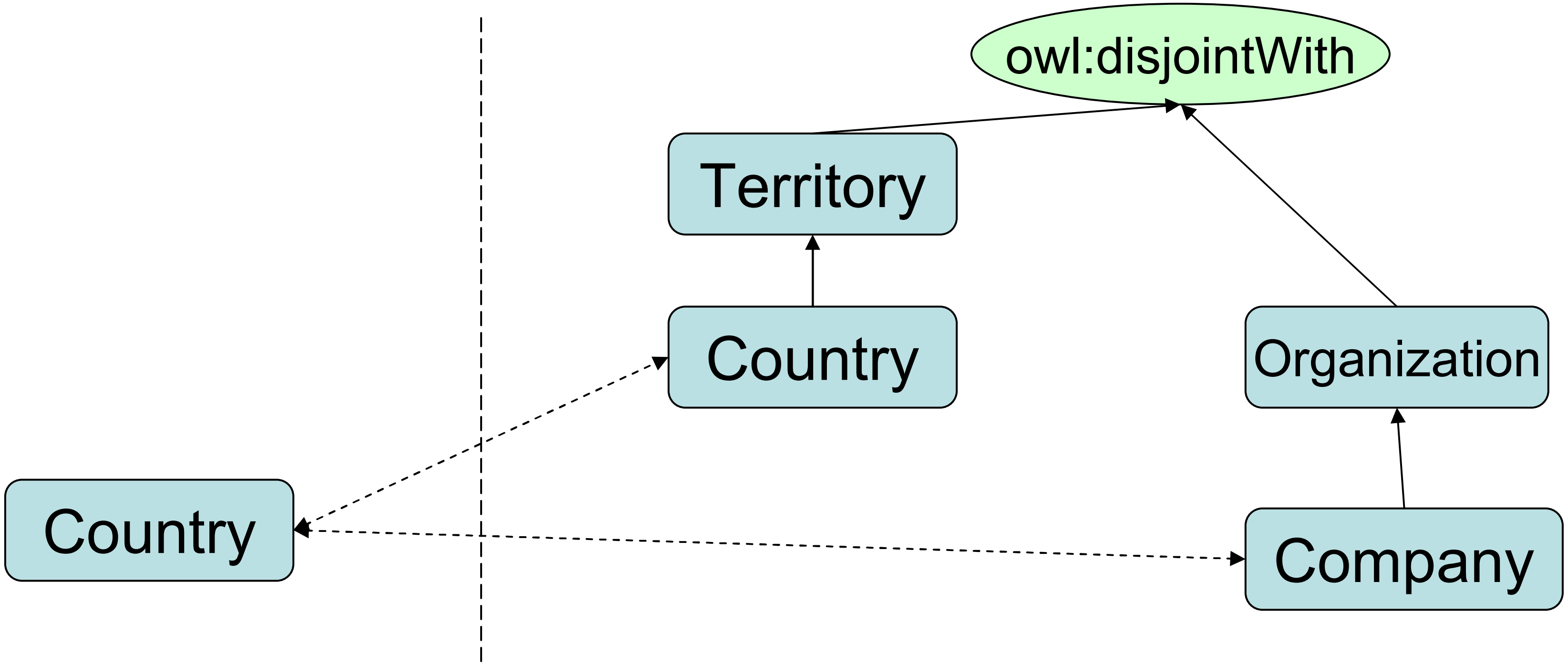
KnoFuss

- Designed for the corporate knowledge management scenario
- Single common schema
- Workflow
  - Coreference resolution
    - Attribute-based similarity
  - Coreference refinement
    - Analysis of links, constraints and provenance
- Extendable library of methods

Task decomposition



Filtering



- Produce candidate mappings
- Remove conflicting mappings based on the similarity score

Query translation

```
SELECT ?uri WHERE {
  ?uri rdf:type sweto:Computer_Science_Researcher }

SELECT ?uri WHERE {
  { ?uri rdf:type tap:ComputerScientist }
  UNION
  { ?uri rdf:type tap:MedicalScientist }
  UNION
  { ?uri rdf:type tap:CMUPerson } }
```

Setup

- Datasets
  - TAP
  - SWETO
  - DBPedia
- Ontology matching
  - CIDER (Gracia & Mena, 2008)
  - Lily (Wang & Xu, 2008)
- Instance coreference resolution
  - String similarity (Jaro-Winkler, L2 Jaro-Winkler)

Tests (F1-measure)

Datasets	manual	CIDER	Lily
TAP/SWETO	0.77	0.76	0.42
TAP/DBPedia	0.88	0.66	0.44
SWETO/DBPedia	0.89	0.81	0.70

- Instance coreference resolution
  - String similarity (Jaro-Winkler, L2 Jaro-Winkler)

Conclusions

- Schema-level recall is important (even at the expense of precision)
  - CIDER outperformed Lily
  - Finding overlapping classes
- Restrictions are very useful
  - Disjointness, cardinality
  - Public reference ontology may help?
- Provenance of linksets is crucial
  - Extending coreference bundles?

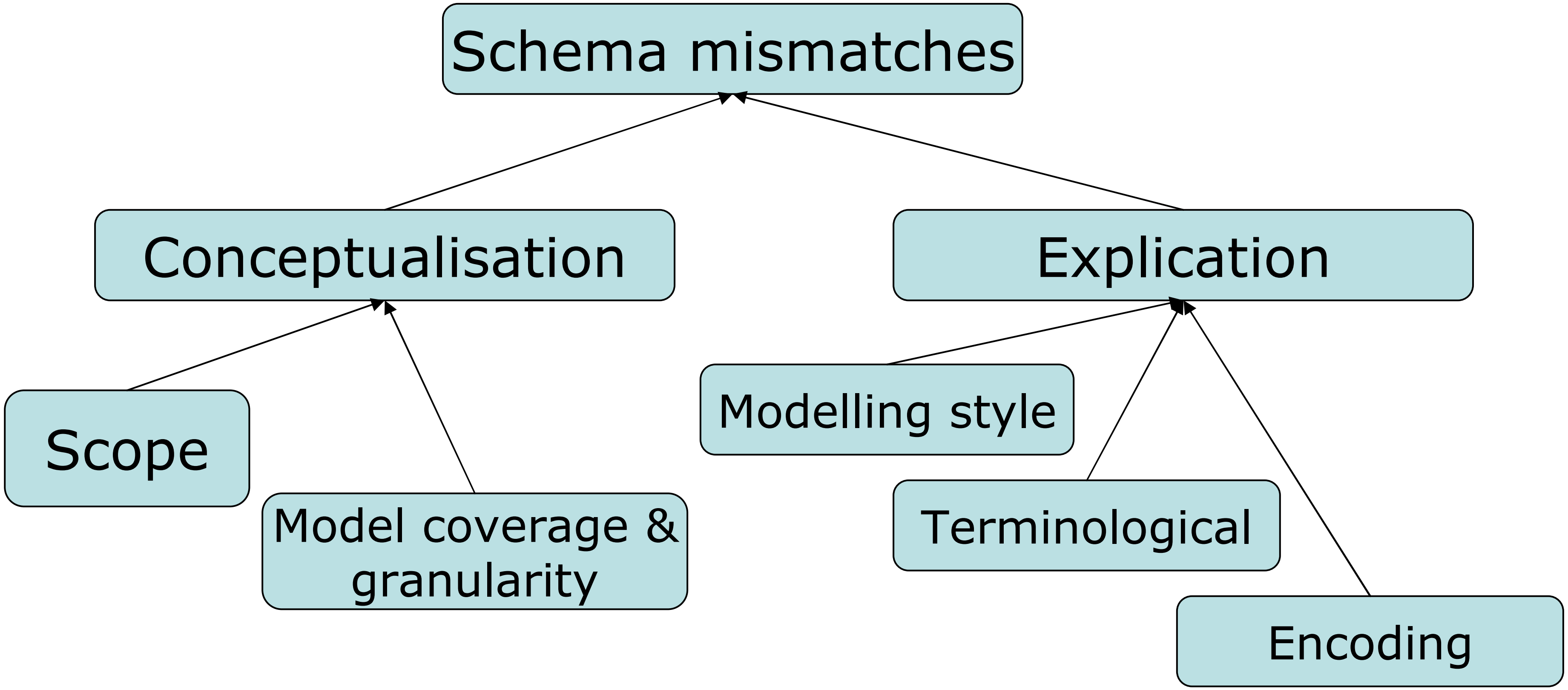
Questions?

Thanks for your attention

Tests

- CIDER
  - All schema mappings above the threshold are accepted
- Lily
  - One-to-one schema mappings
  - “Competitive” schema mappings are removed
  - (+) Higher schema alignment precision
  - (-) Negative impact at the data level

Schema mismatches [11]



Future work

- Original version
  - Sequential workflow
  - Schema integration -> data integration
  - Omitted schema mappings – lower data-level recall
- To do:
  - Iterative workflow (as in (Udrea et al., 2007))
  - Discovery of omitted schema mappings based on instance-level matches