The challenge of expressive reasoning over large knowledgebases

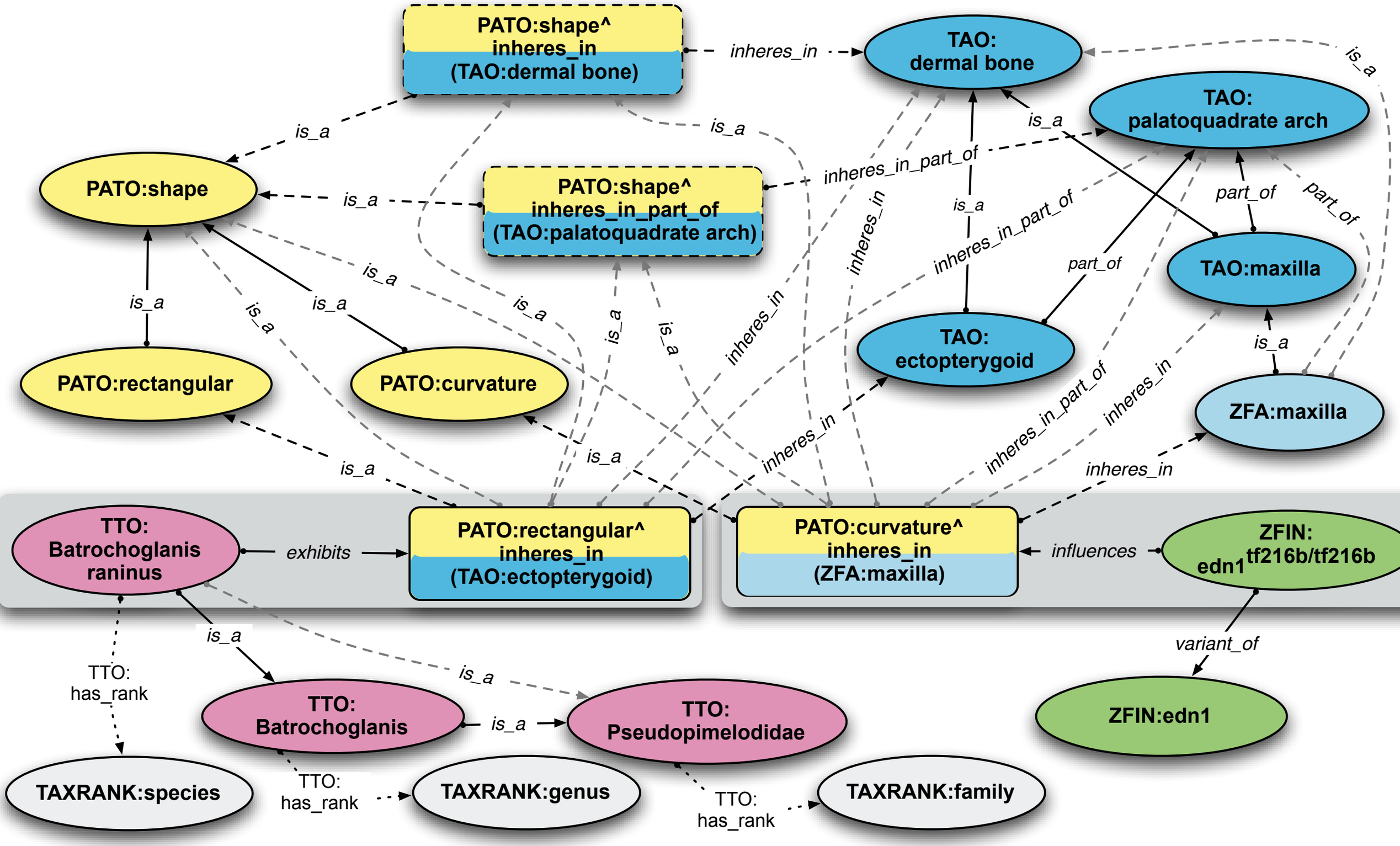
Hilmar Lapp (NESCent)
Jim Balhoff (NESCent)

EQ phenotypes

- Entity-Quality is a template for creating logical expressions with explicit semantics
- Curator selects terms from ontologies to fill in E and Q slots
- E = ‘tooth’; Q = ‘rounded’
- Semantics formalized in OWL (Web Ontology Language) class expression:
  - '*has_part* some (**tooth** and *bearer_of* some **rounded**)'
  - To create a phenotype annotation, an organism can be asserted to be a member of this class

Reasoning allows...

- Inference of relationships not directly asserted
  - e.g. via transitivity (tooth *part_of* jaw, jaw *part_of* head -> tooth *part_of* head)
- Semantic queries can return all data meeting logical criteria
  - e.g. '*has_part* some (fin and *bearer_of* some blue)' matches a fish with a blue dorsal fin
  - or '*has_part* some ((*part_of* some head) and (*bearer_of* some blue))'
- OWL 2 provides a reasoning framework with good support for web standards and useful expressivity


PATO:shape^
inheres_in
(TAO:dermal bone)
inheres_in
TAO:
dermal bone
is_a
is_a
TAO:
palatoquadrate arch
PATO:shape
is_a
PATO:shape^
inheres_in_part_of
(TAO:palatoquadrate arch)
inheres_in_part_of
is_a
inheres_in_part_of
part_of
part_of
is_a
is_a
is_a
is_a
is_a
is_a
is_a
is_a
inheres_in
inheres_in
part_of
TAO:maxilla
PATO:rectangular
PATO:curvature
TAO:
ectopterygoid
is_a
ZFA:maxilla
is_a
is_a
inheres_in
inheres_in
inheres_in_part_of
inheres_in
inheres_in
TTO:
Batrochoglanis
raninus
exhibits
PATO:rectangular^
inheres_in
(TAO:ectopterygoid)
PATO:curvature^
inheres_in
(ZFA:maxilla)
influences
ZFIN:
edn1tf216b/tf216b
is_a
variant_of
TTO:
has_rank
TTO:
Batrochoglanis
is_a
TTO:
Pseudopimelodidae
ZFIN:edn1
TAXRANK:species
TTO:
has_rank
TAXRANK:genus
TTO:
has_rank
TAXRANK:family

Challenges

- Available reasoners are insufficient
  - Either lack the expressivity we need, or they do not scale to the data sets we have.
- OWL-DL reasoners
  - complete, but don't scale to large data sets
  - although significant advances from hyper-tableaux-based reasoners
- Rule-based reasoners
  - amenable to iterative inference of deductive closure
  - lack full expressivity
- RDF triple store reasoners
  - Highly scalable (billions of triples)
  - Often limited to subclass reasoning, transitivity
- Common benchmarks exist but are biased
  - Contain data sets with either large ontologies (T-box) or large instance data sets (A-box), but not ones that combine both (which is what we have).

Our data: Phenoscape Knowledgebase

- 3 million OWL axioms (including ontologies and evolutionary data)
  - About 100,000 ontology terms from over a dozen ontologies
  - About 500,000 instance data assertions
- Fastest DL reasoner (FaCT++) completes in-memory classification after 12.5 hours
  - Attempt to generate inferred axioms aborted after 2 weeks runtime
  - Makes use of about 10 GB memory (out of 96 GB), 1 cpu
  - For comparison, a dataset with similar data of about 35k axioms takes less than a minute on a laptop

Approaches

- "Live" query-time reasoning
- Advance materialization of inferences

Live reasoning

- OWL 2 DL reasoners
  - Allow full expressivity, complete inference/query answering
  - Typically work on dataset in memory
  - Extremely slow to classify data and answer queries for anything beyond a "small" dataset
  - Available reasoners are single-cpu only
  - Most seem optimized for ontology classification; much slower when instance data is included
  - Minimum 12 hours from application startup before ready to answer first query using fastest DL reasoner (FaCT++) on all Phenoscape evolutionary data

Live reasoning

- Performance can be improved by restricting expressivity in specific ways
  - OWL 2 provides 3 "profiles" of restricted expressivity (EL, QL, RL)
  - Implementations of various restricted profiles are available for query-time reasoning within some RDF triple stores
  - Unfortunately these expressivity restrictions are too limiting for our data

“Problem” expressions

- Restricted reasoning profiles tend to exclude one or more of these from use
- Phenotypes involving absence (universal quantification and class negation)
  - *has_part* only (not scale)
- Phenotypes involving counts (cardinality restrictions)
  - *has_part* some (head and *has_granular_part* exactly 2 carina)
  - *has_granular_part* min 20 vertebra
- Transitive properties, inverse properties, property chains

Advance materialization

- Reasoner performance is less of an issue
- Changes to any data may involve re-running entire reasoning process
- OWL 2 DL reasoners
  - Initial classification step slow
  - Attempt to generate inferred axioms on full dataset has not been successful
  - Default generated axioms seem to not be that useful for queries
- Rule-based reasoners can generate inferences more quickly, but do not support full expressivity or may be incomplete
  - But it may be practical to materialize all the *required* inferences - if the required expressivity is supported

Possible strategies

- Precompute inferred axioms using an OWL DL reasoner for smaller datasets (one study at a time), then merge
  - Must still reason over all ontologies with each dataset
  - Some inferences will be missed that would make use of facts across datasets
  - Not yet fully attempted - FaCT++ crashes when generating inferred axioms
- Precompute specific inferred axioms with target queries in mind, using rule-based reasoner
  - But only anticipated queries can be completely answered
  - Current strategy; limited expressivity