Semantics of and for the diversity of life: Opportunities and perils of trying to reason on the frontier

Hilmar Lapp
National Evolutionary Synthesis Center (NESCent)
CSHALS, Boston, February 28, 2014
The diversity of life is stunning.
Extinct biodiversity is even greater than extant
• About 1.2-2.1 billion natural history specimens
• > 7,000 natural history collections registered

Comparative features documented in meticulous detail, in free text.
Finding similar information in free-text is difficult

<table>
<thead>
<tr>
<th>Description</th>
<th>Author(s)</th>
<th>OMIM query</th>
<th>#records</th>
</tr>
</thead>
<tbody>
<tr>
<td>“lacrymal bone...flat”</td>
<td>Mayden 1989</td>
<td>“large bone”</td>
<td>1083</td>
</tr>
<tr>
<td>“lacrim...small, flat”</td>
<td>Grande and Poyato-Ariza 1999</td>
<td>“enlarged bone”</td>
<td>224</td>
</tr>
<tr>
<td>“lacri...triangular”</td>
<td>Royero 1999</td>
<td>“big bones”</td>
<td>21</td>
</tr>
<tr>
<td>“first infraorbital (lachrimal) shape...flattened”</td>
<td>Kailola 2004</td>
<td>“huge bones”</td>
<td>4</td>
</tr>
<tr>
<td>“fourth infraorbital...anterior and posterior margins...in parallel”</td>
<td>Zanata and Vari 2005</td>
<td>“massive bones”</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“hyperplastic bones”</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“hyperplastic bone”</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“bone hyperplasia”</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“increased bone growth”</td>
<td>879</td>
</tr>
</tbody>
</table>
Comparative features underlie knowledge about evolutionary history.

Sereno (1999)

Vs.

Dos Reis et al (2014) Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol Lett 10
Even though, much of biodiversity is not well described.

Content-rich = richness score > 0.4
Only families with > 10 species.
Dark areas in the Tree of Life

EOL/BHL Research Sprint
Jessica Oswald, Karen Cranston, Gordon Burleigh, and Cyndy Parr
Our knowledge is conflicting
Making biodiversity knowledge part of Big & Linked Data faces major challenges

- The Linnean system for taxonomy was not designed for Linked Data.
  - Names as identifiers, and track provenance.
  - No canonical comprehensive taxonomy.
- Specimens referenced by a combination of metadata (with unreliable provision & uniqueness).
  - No common resolver to identifiers or URLs.
- Most knowledge and data is in free text using the expressivity of natural language.
Also huge opportunities for advancing science

- Organizing and linking data to biodiversity
- Data mining morphology, traits, habitats, ecological interactions, and other descriptive data
Using reasoning to make linking data by taxon less perilous
The perils of linking data by taxon

Tetrapoda

Taxonomy ID: 32523
Genbank common name: tetrapods
Inherited blast name: vertebrates
Rank: no rank
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)

Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha
How to render the definition of taxon names computable?

1. Tetrapoda (crown group)
2. Tetrapoda (with digits)
3. Tetrapoda (stem group)
Naming everything in the Tree of Life is impractical.
Phyloreferencing: A universal computable coordinate system for life

What if this were not just a signpost but a coordinate computable against any tree?
Phyloreferencing: A universal computable coordinate system for life

Requirements:

• Any node, branch, subtree is referencable
• References are unambiguous
• References are portable
• References are computable
• Adapts easily to new and changing knowledge
Phylogenetic clade definitions

Node-based

Branch-based

Apomorphy-based

Form the basis of Phylogenetic Nomenclature

We already know how to render the semantics of names computable

- Conjunctive OWL class expressions
- Necessary and sufficient conditions for class membership

**Class:**
lateral_ventricle_choroid_plexus_stroma

**EquivalentTo:**
uberon:choroid_plexus_stroma and bfo:part_of some uberon:telencephalic_ventricle
Ontologies for evolutionary relationships exist

Comparative Data Analysis Ontology:

- OWL ontology
- Scope: phylogenetic data and trees
- Rich set of axioms
The Tree of Life as an ontology

Class: Tetrapoda_Total
EquivalentTo:

cda:has_Descendant value taxon:Amniota
and phyloref:excludes_lineage value taxon:Dipnoi
The Tree of Life as an ontology

Class: Tetrapoda_Crown
EquivalentTo:
  cdao:has_Descendant value taxon:Amniota
  and phyloref:excludes_lineage value taxon:Crassigyrinus
The Tree of Life as an ontology

**Class:** Tetrapoda_Digit_hand  
**EquivalentTo:**
- cdao:has_Descendant value taxon:Amniota  
- and phyloref:has_Progenitor some (bfo:has_part some uberon:manaul_digit)
Phyloreferences can be named or anonymous

- Phyloreference expressions can be anonymous, or named and registered
- Semantics is exactly the same
- Naming has benefits: promotes reuse, consistency, usability and accessibility by non-experts
- Ontology of computable clade names

Class: AGF4-SHRO-3560
EquivalentTo:
  cdao:has_Descendant value taxon:Amniota
  and phyloref:has_Progenitor some (bfo:has_part some uberon:'manual digit')

VS.

Class: Tetrapoda_Digit_hand
Annotations:
  rdfs:label “Tetrapoda with limbs with digits”
  dc:description “the first sarcopterygian to have possessed digits homologous with Amniotes”
EquivalentTo:
  cdao:has_Descendant value taxon:Amniota
  and phyloref:has_Progenitor some (bfo:has_part some uberon:'manual digit')
Common coordinate systems also require standards

- Common formally defined language (OWL)
- Common ontologies to encode knowledge and queries (CDAO, PhyloRef)
- Common identifiers for phyloreference specifiers
  - Phenotype and anatomy ontologies
  - Canonical taxonomy for OTU names
  - Canonical GUIDs for specimens
Using ontologies & reasoning to mine our knowledge of trait diversity
Challenge: What do we know about the evolution of morphological biodiversity?


http://www-news.uchicago.edu/releases/06/060405.tiktaalik.shtml
Vast stores of data, but all free text
Manual effort of experts is not scalable

Fig. 7, Sereno (2009)

Fig. 6, Sereno (2009)
Challenge: What can we hypothesize about genes involved in the evolution of biodiversity?
Model organism & health literature contains descriptions of gene and disease phenotypes

Fig. 2 Ossifications in the young and homozygous edn1 mutant. (A) The top, of negative images with Blackman in larvae at bones of the pharyngeal arches (Maboe, 1996) by their lateral plates. B) Ossifications are labeled on the right. (B) Many of the anterior structures (in the first two arches) are missing in the edn1 mutant. Ceratobranchial 5 and the ceratohyal are present, shortened and somewhat malformed. In the mandibular arch dermal bones (max/den) are present but severely malformed, an example of the 'wicket' phenotype discussed in the text (see also Fig. 3). In the hyoid arch the opercle is present and its joint region (upper part of the bone) is markedly expanded, a mild example of the 'opercle-gain' phenotype described in the text and other figures. Scale bar: 100 μm.
Translational bioinformatics

Model organism
->
Human

Fig. 1, Washington et al (2009)

Fig. 3, Washington et al (2009)
Translational biodiversity informatics

Model organism genes -> Evolutionary diversity by semantic similarity of phenotypes

Fig. 1, Washington et al (2009)
Focus: vertebrate fin/limb transition
# Phenotype annotation of taxon

## Vertebrate Taxonomy Ontology
- **Terms**: 106,927
- **Synonyms**: 104,506

## Uberon Anatomy Ontology
- **Terms**: 11,300
- **Skeletal**: 3,327

## Quality Ontology (PATO)
- **Terms**: 2,050

## Table

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Entity</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cheirolepis</strong></td>
<td>pectoral fin radial</td>
<td>present</td>
</tr>
<tr>
<td></td>
<td>cleithrum</td>
<td>pointed</td>
</tr>
<tr>
<td><strong>Amphibamus</strong></td>
<td>humerus</td>
<td>present</td>
</tr>
<tr>
<td></td>
<td>cleithrum</td>
<td>slender</td>
</tr>
</tbody>
</table>
Computable via shared ontologies, rich semantics, OWL reasoning
Ontology-annotated descriptions integrate across studies & fields

Comparative studies + Model organism datasets = Phenoscape Knowledgebase
Phenoscape KB content

- 16,000 character states from >120 comparative morphological datasets, linked to 4,000 vertebrate taxa.
- Imported genetic phenotype and expression data from ZFIN, Xenbase, MGI, and Human Phenotype project.
- Shared semantics: Uberon (anatomy), PATO (phenotypic qualities), Entity-Quality (EQ) OWL axioms (phenotype observations)
- Plus a dozen other ontologies …
KB enables candidate gene hypothesis generation

Mutation of eda gene in Danio:

Ictalurus punctatus:

Harris et al., 2007

Copyright © Jean Ricardo Simões Vitule, All Rights Reserved
Eda expression is lacking in the epidermis
KB enables candidate gene hypothesis generation

Caudal fin, absent

taxa: Mola, Gymnotiformes

genes: smc3, tll1, yap1

Did Mola lose its caudal fin because of changes in regulation of yap1?
KB enables candidate gene hypothesis generation

**Basihyal, absent**

**taxa:** Siluriformes

**genes:** (11)

*Is the basihyal absent in catfish because of changes in brpf1?*

Photo: Richard Edmunds

Ictalurus punctatus

Laue et al., 2008
Work on vertebrate fin/limb transition candidate genes is ongoing.
Using reasoning to synthesize data implied by what we know
What do we know about the evolution of morphological biodiversity?

Do we really not know more?

Presence / absence of digits

Author-asserted:
74 present
25 absent
11% of taxa
Character synthesis by inference

• Most phenotypic descriptions of some feature of a structure implies its presence or absence:
  
  • “Humerus slender and elongate: with length more than three times the diameter of its distal end” → humerus must be present

• Partonomy axioms in the ontology allow inferring presence or absence:
  
  • ‘all humerus part_of some forelimb’ → forelimb must be present if humerus is; humerus must be absent if forelimb is

• Developmental origin axioms allow inferring presence or absence:
  
  • ‘all limb develops_from some limb bud’ → limb bud must be present if limb is; limb must be absent if limb bud is
A reasoner can fill in lots of gaps

Presence / absence of digits

Asserted & inferred:
645 inferred present
74 asserted present
25 asserted absent
82% of taxa
A reasoner can fill in lots of gaps asserted presence/absence with inference

Mesquite “birds-eye view”
Even knowing only presence/absence of traits can be powerful.
Synthesis highlights conflict and gaps

- Conflicting interpretations in studies
  - supinator process of humerus: both absent & present in *Strepsodus* (Zhu et al. 1999 vs. Ruta 2011)

- Gaps in knowledge
  - acetabulum present or absent?

- Same term, different meaning?
  - *Acanthostega*— “radials, jointed” (Swartz 2012)
  - but doesn’t have radials...

- Uneven taxon sampling
How to reason and query at the scale of biodiversity?
Rich axioms allow rich inferences, but make scaling a challenge

- Anatomy ontologies and EQ annotation employ rich OWL semantics → requires DL reasoner

- Classifying and querying over large dataset (~25 million RDF triples) does not scale well

- Presently, the only feasible OWL reasoner is ELK
  - constrained to OWL EL profile → limits kinds of expressions that can be used
  - best performance over class axioms only → data must be modeled so as to avoid need for classifying instances
kb-owl-tools

**OWL conversion**
- Includes translation of EQ to OWL expressions

**Axiom generation**
- "Absence" classes for OWL EL negation classification workaround
- SPARQL facilitation (e.g. materialized existential hierarchies such as part_of)

**Identifier cleanup**
- URIs for standard properties across ontologies are a mess

**Materialize inferred subclass axioms**
- ELK reasoner using extracted tbox axioms only (not feasible with individuals included)

**Assertion of absence hierarchy**
- Based on inverse of hierarchy of negated classes computed by ELK

**MOD curators**
- MODs phenotype & gene expression annotations gene identifiers

**Phenoscape curators**
- NeXML matrices

**Ontology curators**
- ontologies

**Bigdata triplestore**
- RDF (all)
- OWL
- Owllet SPARQL endpoint
- SPARQL queries
- applications

**Phenoscape curators**
- SPARQL queries

**MOD curators**
- MODs phenotype & gene expression annotations gene identifiers

**Ontology curators**
- ontologies

**Bigdata triplestore**
- RDF (all)
- OWL
- Owllet SPARQL endpoint
- SPARQL queries
- applications
Want to allow arbitrary selection of structures of interest, using rich semantics:

(part_of some (limb/fin or girdle skeleton))

or (connected_to some girdle skeleton)

RDF triplestores provide very limited reasoning expressivity, and scale poorly with large ontologies.

However, ELK can answer class expression queries within seconds.
What if instead of this (*):

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ao: <http://purl.obolibrary.org/obo/my-anatomy-ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT DISTINCT ?gene
WHERE
{
?structure rdf:type ?structure_class .
# Triple pattern selecting structure:
?structure_class rdfs:subClassOf "ao:muscle" .
?structure_class rdfs:subClassOf ?restriction
?restriction owl:someValuesFrom "ao:head" .
}
```

we could do this:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ao: <http://purl.obolibrary.org/obo/my-anatomy-ontology/>
PREFIX ow: <http://purl.org/phenoscape/owlet/syntax#>
SELECT DISTINCT ?gene
WHERE
{
?structure rdf:type ?structure_class .
# Triple pattern containing an OWL expression:
?structure_class rdfs:subClassOf "ao:muscle and (ao:part_of some ao:head)"^^ow:omn .
}
```
owlet: SPARQL query expansion with in-memory OWL reasoner

- **owlet** interprets OWL class expressions embedded within SPARQL queries
- Uses any OWL API-based reasoner to preprocess query.
  - We use ELK that holds terminology in memory.
- Replaces OWL expression with FILTER statement listing matching terms
- [http://github.com/phenoscape/owlet](http://github.com/phenoscape/owlet)
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ao: <http://purl.obolibrary.org/obo/my-anatomy-ontology/>
PREFIX ow: <http://purl.org/phenoscape/owlet/syntax#>

SELECT DISTINCT ?gene
WHERE
{
?structure rdf:type ?structure_class .

# Triple pattern containing an OWL expression:
?structure_class rdfs:subClassOf "ao:muscle and (ao:part_of some ao:head)"^^ow:omn .
}

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ao: <http://purl.obolibrary.org/obo/my-anatomy-ontology/>
PREFIX ow: <http://purl.org/phenoscape/owlet/syntax#>

SELECT DISTINCT ?gene
WHERE
{
?structure rdf:type ?structure_class .

# Filter constraining ?structure_class to the terms returned by the OWL query:
FILTER(?structure_class IN (ao:adductor_mandibulae, ao:constrictor_dorsalis, ...))
}
How to annotate descriptive biodiversity data at scale?
1. Students: 
gather publications 
(scan hard copies, 
produce OCR PDFs)

2. Students: 
Manual entry of free 
text character 
descriptions, matrix, 
taxon list, specimens 
and museum numbers 
using Phenex

3. Character 
annotation by experts: 
Entry of phenotypes 
and homology 
assertions using 
Phenex

4. Build pipeline for 
the Phenoscape KB

139 Publications
8,073 characters
18,811 states
for 4397 taxa
~6.7 FTE years
CharaParser: Using NLP to scale annotation

Cui H (2012) CharaParser for fine-grained semantic annotation of organism morphological descriptions. JASIST 63: 738-754
Most of the effort may actually be data digitization, ontology building, etc.
Building semantically rich community ontologies is hard

Uberon: cross-species anatomy ontology

222 terms

790 terms
Taxonomy ontology needs to be synthesized

Vertebrate Taxonomy Ontology (VTO)

- Gnathostomata
  - Chondrichthyes
  - Teleostomi
    - Acanthodii†
    - Euteleostomi
      - Actinopterygii
        - Latimeria
      - Dipnotetrapodomorpha
        - Dipterus
      - Rhipidipodia
      - Tetrapodomorpha
        - Eotetrapodiformes
          - Elpistostegalia
            - Elpistostege†
            - Panderichthys†
          - Tetrapoda
            - Acanthostega†
            - Amniota
              - Sauropsida
                - Anapsida
                - Testudines
              - Diapsida
                - Sauria
                - Synapsida
                  - Mammalia
                  - Amphibia
                    - Anura
                    - Caudata

- VTO: 107,004
- NCBI: 65,747
- TTO: 38,640
- AWeb: 7,854
- PaleoDB: 35,937

Midford et al., in press JBMS
How to evaluate semantic similarity metrics meaningfully?
Many different metrics

What if there is no “gold standard”?

CRAFT: the Colorado Richly Annotated Full Text Corpus

Quick Facts

- 67 full text articles
- >560,000 Tokens
- >21,000 Sentences
- ~100,000 concept annotations to 7 different biomedical ontologies/terminologies
  - Chemical Entities of Biological Interest
  - Cell Ontology
  - Entrez Gene
  - Gene Ontology (biological process, cellular component, and molecular function)
  - NCBI Taxonomy
  - Protein Ontology
  - Sequence Ontology
- Penn Treebank markup for each sentence
- Multiple output formats available
- Integrated with UIMA

The Colorado Richly Annotated Full Text Corpus (CRAFT) is a manually annotated corpus consisting of 67 full-text biomedical journal articles. Each article is a member of the PubMed Central Open Access Subset.
Evaluate by tendency to maximizing inter-curcurator similarity

- Can also be used to assess payoff from annotation granularity
Summary: Big obstacles, but also big opportunities for semantics-driven discovery

- Vast stores of published knowledge and data waiting to be exploited
- Countless possibilities for knowledge discovery by connecting data
- Scaling out expressive reasoning is hard
- Workforce training is a serious issue
Phenoscape project team

- National Evolutionary Synthesis Center (NESCent)
  - Todd Vision (also University of North Carolina at Chapel Hill)
  - Hilmar Lapp
  - Jim Balhoff
  - Prashanti Manda
- University of South Dakota
  - Paula Mabee
  - Wasila Dahdul
  - Alex Dececchi
- University of Chicago
  - Paul Sereno
  - Nizar Ibrahim
- Mouse Genome Informatics
  - Judith Blake
  - Terry Hayamizu
- University of Oregon (Zebrafish Information Network)
  - Monte Westerfield
  - Yvonne Bradford
  - Ceri Van Slyke
- Cincinnati Children's Hospital (Xenbase)
  - Aaron Zorn
  - Christina James-Zorn
  - Virgilio Ponferrada
- California Academy of Sciences
  - David Blackburn
- University of Arizona
  - Hong Cui
- Oregon Health & Science University
  - Melissa Haendel
- Lawrence Berkeley National Labs
  - Chris Mungall
Acknowledgements

- Phenoscape personnel, PIs, curators & workshop participants
- National Evolutionary Synthesis Center (NESCent)
- Nico Cellinese (U. Florida): Phyloreferencing
- NSF (DBI-1062404, DBI-1062542)
- Karen Cranston (Open Tree of Life)
- Cynthia Parr (EOL)
- William Ulate (BHL)