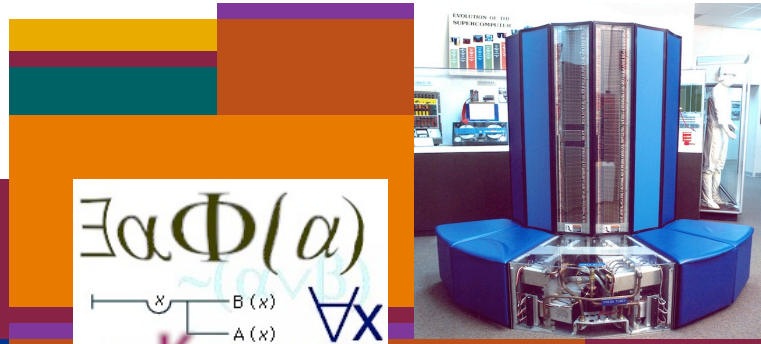
Phylogenetics, Informatics, and Answer Set Programming

Enrico Pontelli

Department of Computer Science

New Mexico State University



What are we going to talk about?

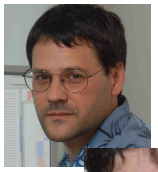
- The EvoIO Initiative
  - Interoperability in Phylogenetic Inference
  - The EvoIO Stack
- ASP in the EvoIO Stack
- Future Directions of EvoIO

The History of EvoIO

- National Evolutionary Synthesis Center

- EvoInfo Working Group
  - Fall 2006
  - Promote use of *Phyloinformatics* solutions
    - Address issues of interoperability
    - Promote the development of standards and community cohesion
    - Support software development efforts

EvoIO

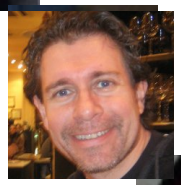


Arlin Stoltzfus
National Institute of Standards and Technology,
Gaithersburg, MD

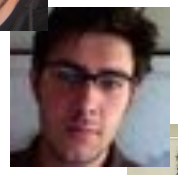
www.evoio.org

Karen Cranston
Duke University & OpenTree
Durham, NC

Hilmar Lapp
Duke University
Durham, NC



Enrico Pontelli
New Mexico State University,
Las Cruces, NM

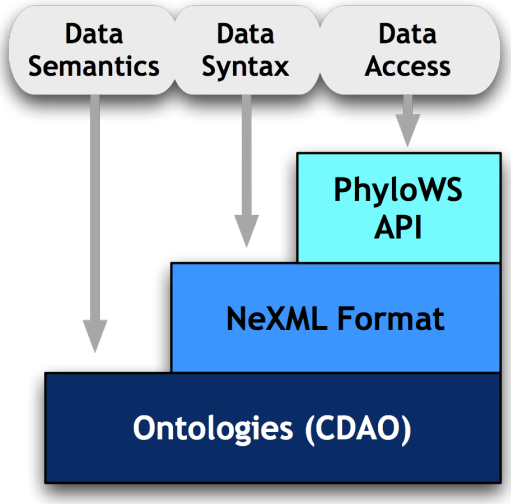


Rutger Vos
Naturalis
Amsterdam

Brian O'Meara
University of Tennessee
Knoxville, TN

EvoIO

- EvoIO and its sub-groups
  - Promote the EvoIO Stack



  - Promote community efforts to address phyloinformatics interoperability
    - **HIP: H**ackatons for **I**nteroperability in **P**hylogenetics

PHYLOINFORMATICS AND INTEROPERABILITY

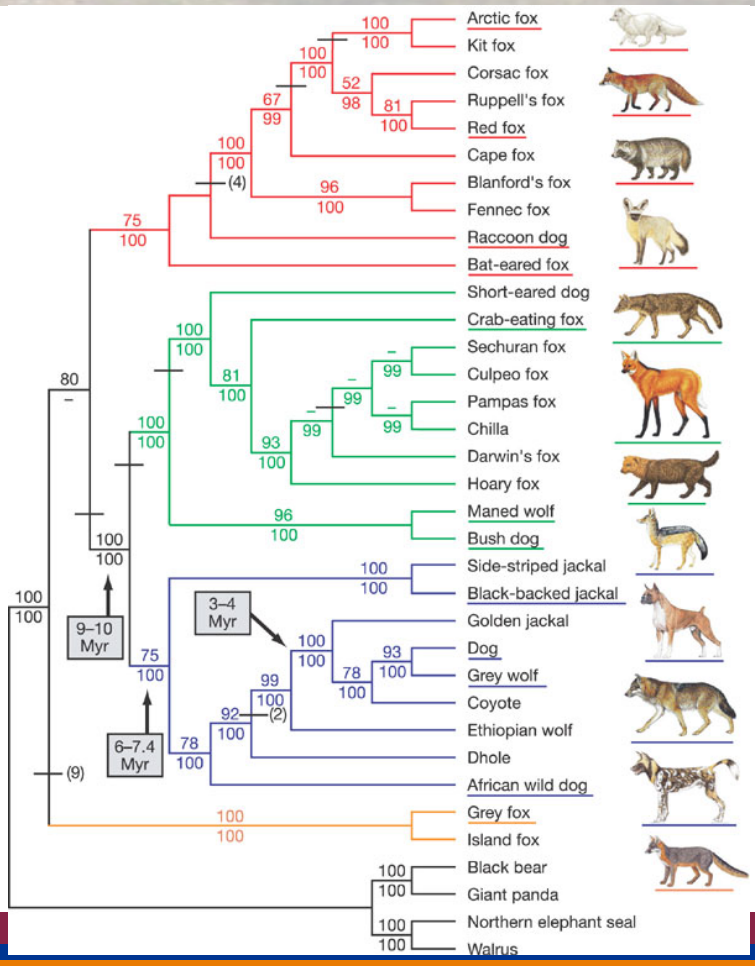
What is Phyloinformatics?

Phylogenetics:

"The systematic study of organism relationships based on evolutionary similarities and differences."

Informatics:

"The sciences concerned with gathering, manipulating, storing, retrieving, and classifying recorded information."



Why should you care?

Firstly,

"Nothing in evolution makes sense except in the light of phylogeny"
(Society of Systematic Biologists)

"Nothing in biology makes sense except in the light of evolution"
(Theodosius Dobzhanksy)

Surely, "gathering, manipulating, storing, retrieving and classifying" such information is worthwhile?

But if that doesn't convince you…

As a consumer of phylogenetic data

The “New Biology” is coming:

> “***Major advances will take place via integration and synthesis, rather than decomposition and reduction***”
> (Committee on a New Biology for the 21st Century, 2009)

Integrative aspects present at the level of software – need to compose analysis steps into workflows

Presumably, this will involve retrieving and classifying.

As a producer of phylogenetic data

- Many journals require proper storage of data described in a manuscript.

- Funding agencies require dissemination and sharing of research results.

The Past (Present?)

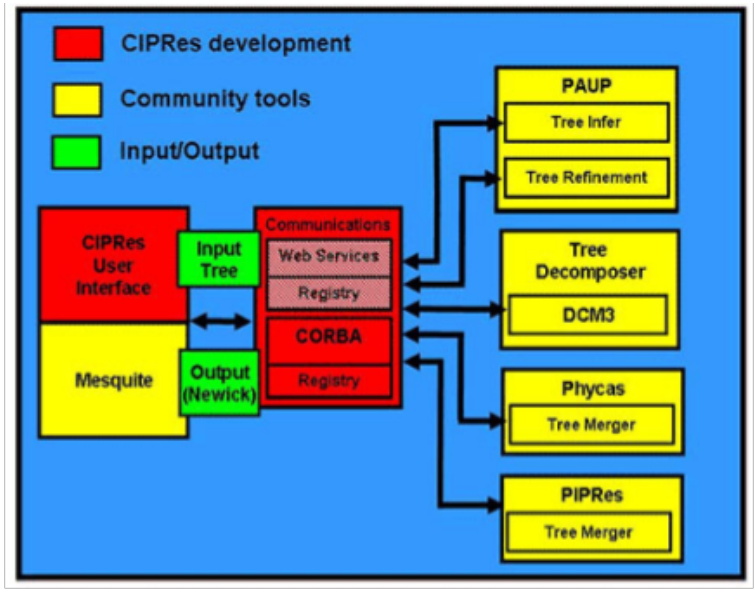
- Everything closed:
  - Idiosyncratic, private data
  - “pay-walls”
  - Closed source software
  - Lack of accessible publishing media
  - Data “Mine”ing
  - **Non-existent interoperability**

An Interoperability Disaster Story

The Great Baltimore Fire of 1904

The Past

- Cyberinfrastructure for Phylogenetic Research (CIPRES)
  - enable large-scale phylogenetic reconstructions on a scale that supports analyses of huge data sets containing hundreds of thousands of bio molecular sequences.



CIPRES

NM STATE

The Past

- NEXUS
  - Valuable effort but…
    1. Inconsistent specification and implementation (e.g., PAUP-Nexus, Mesquite-Nexus, etc.)
    2. Not Extensible
    3. A number of missing features (inconsistently implemented as private blocks)
  - The TreeBase experience

*__Maddison, Swofford and Maddison__, 1997. NEXUS: An Extensible File Format for Systematic Information. *Syst. Biol.* **46**(4):590-621

Challenges

- Powerful tools for evolutionary biology are underutilized and difficult to apply
- Tools are mainly used in **an expert-supervised approach**, which is time-consuming, difficult to document, error-prone, and not scalable
- It is necessary to better document and automatize **the whole pipeline** used for evolutionary analysis

DNA extraction	Sequencing and Base-calling	Ortholog searches	Multiple Alignment	Alignment refinement	Phylogenetic reconstruction	Statistical analysis
Extraction kits Conditions	PCR conditions Sequencer PHRED	BLAST BBH COGnitor PSI-BLAST Phylogeny	Clustal T-Coffe MAFFT MultAlign	Manual Leon REFINER HMM	Parsimony Max Likelihood PAUP Phylip	Bootstrap Jacknife Bayesian MCMC

The Present/Future

Science is opening up:

- Open data
- Open access publishing
- Open source software

Publishing is now accessible to everyone, online

Growing number of repositories

TreeBASE

MorphoBank
Morphological Homology on the Web

DRYAD

TIME TREE
THE TIMESCALE OF LIFE

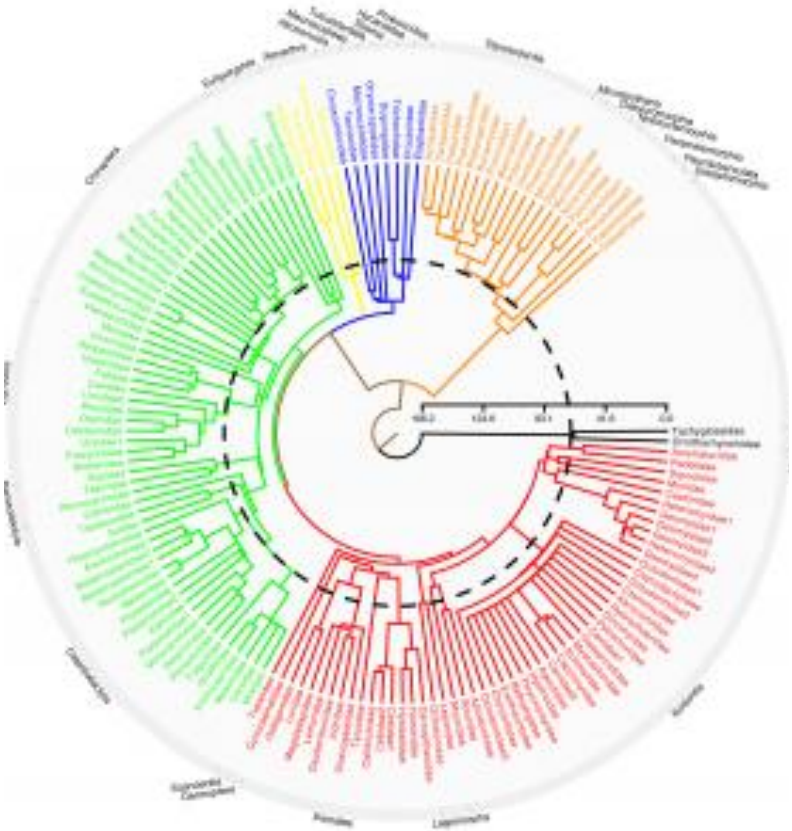
Treefam

Our current nightmare

Hunts Needle in a Haystack

HOW LONG does it take to find a needle in a haystack? Jim Moran, Washington, D. C., publicity man, recently dropped a needle into a convenient pile of hay, hopped in after it, and began an intensive search for (a) some publicity and (b) the needle. Having found the former, Moran abandoned the needle hunt.

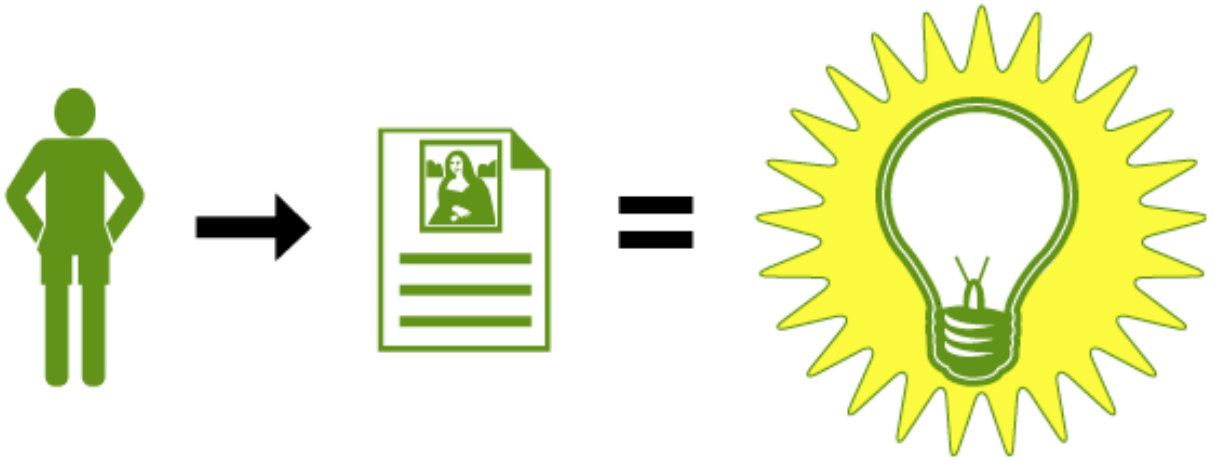
This is too hard



- **O. R. P. Bininda-Emonds, M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman** and **A. Purvis**, 2007. The delayed rise of present-day mammals. *Nature* **446**: 507-512.

Let's delegate that

The current web makes sense to us

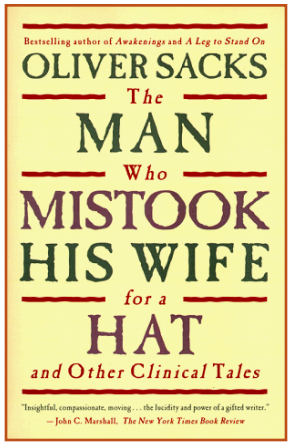


But not to a machine

Mozilla

Fichier Edition Affichage Aller à Marque-pages Outils Fenêtre Aide

http://www.reazyread.com/sacks.html



The Man Who Mistook His Wife for a Hat : And Other Clinical Tales *by* Oliver Sacks

In his most extraordinary book, "one of the great clinical writers of the 20th century" (The New York Times) recounts the case histories of patients lost in the bizarre, apparently inescapable world of neurological disorders. Oliver Sacks's The Man Who Mistook His Wife for a Hat tells the stories of individuals affli
lost their memorie
recognize people a
shout involuntary
retarded yet are gi

If inconceivably strange, these brilliant tal
are studies of life struggling against increc
impaired, to imagine with our hearts what
medicine's ultimate responsibility: "the su

Our rating : ★★★★☆

Find other books in : ☐ Neurology

Search books by terms :

Google - Mozilla

Fichier Edition Affichage Aller à Marque-pages Outils Fenêtre Aide

http://www.google.com/

Google™

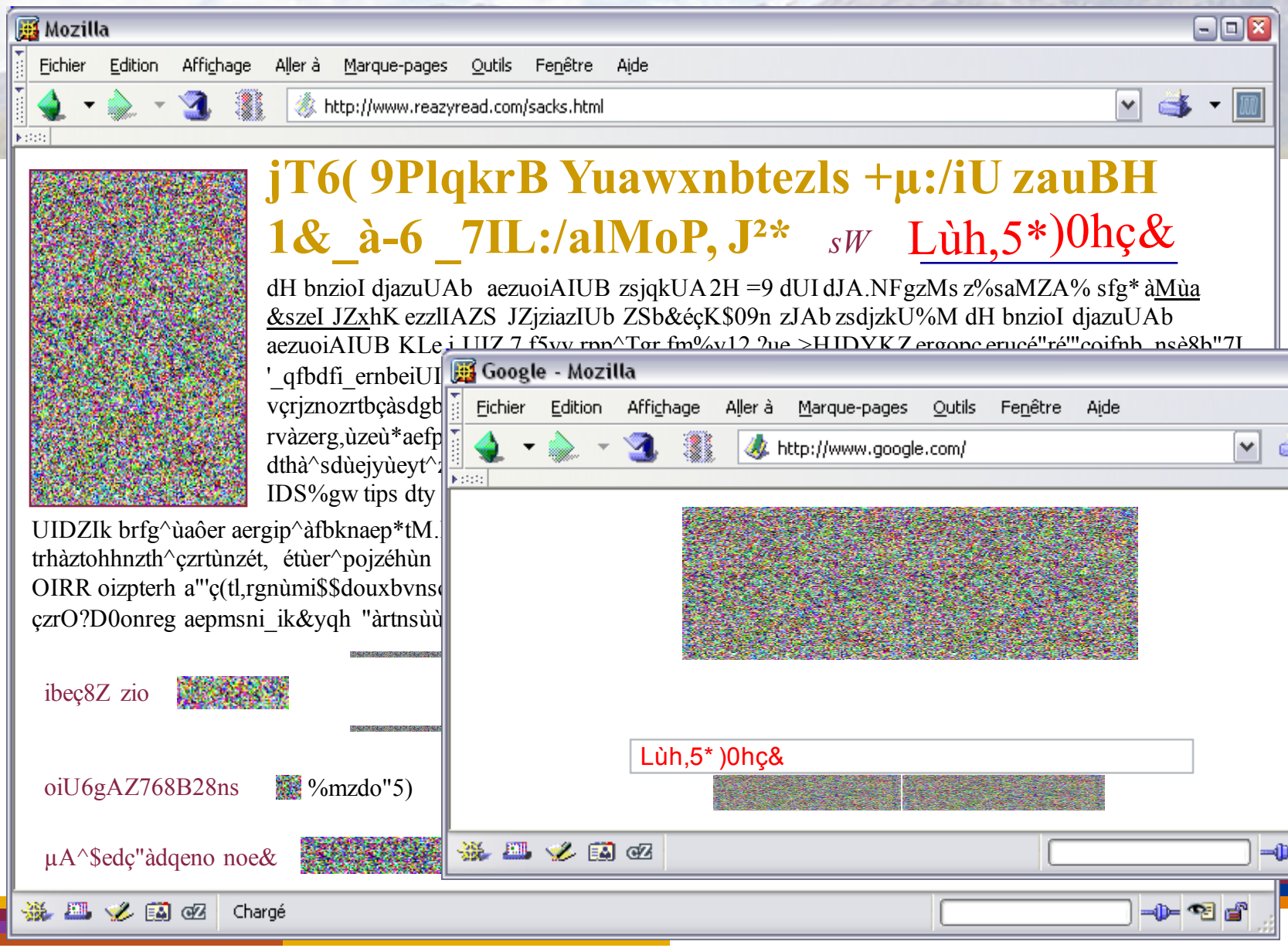
Web Images Groups News Froogle Local more »

Oliver Sacks

Google Search I'm Feeling Lucky

Advanced Search
Preferences
Language Tools

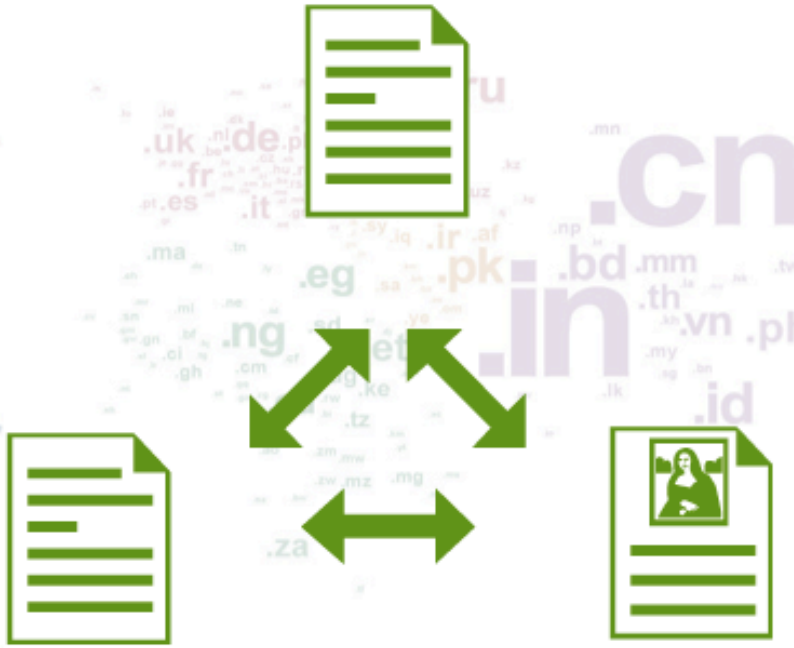
Chargé


Mozilla
Fichier Edition Affichage Aller à Marque-pages Outils Fenêtre Aide
http://www.reazyread.com/sacks.html
jT6(9PlqkrB Yuawxnbtezls +µ:/iU zauBH
1&_à-6 _7IL:/alMoP, J²* sW Lùh,5*)0hç&
dH bnzioI djazuUAb aezuoiAIUB zsjqkUA2H =9 dUI dJA.NFgzMs z%saMZA% sfg* àMùa
&szeI JZxhK ezzlIAZS JZjziazIUb ZSb&éçK$09n zJAb zsdjzkU%M dH bnzioI djazuUAb
aezuoiAIUB KLe
'_qfbdfi_ernbeiUI
vçrjznozrtbçàsdgb
rvàzerg,ùzeù*aefp
dthà^sdùejyùeyt^
IDS%gw tips dty
UIDZIk brfg^ùaôer aergip^àfbknaep*tM.
trhàztohhnzth^çzrtùnzét, étùer^pojzéhùn
OIRR oizpterh a"'ç(tl,rgnùmi$$douxbvns
çzrO?D0onreg aepmsni_ik&yqh "àrtnsùù
ibeç8Z zio
oiU6gAZ768B28ns %mzdo"5)
µA^$edç"àdqeno noe&
Google - Mozilla
Fichier Edition Affichage Aller à Marque-pages Outils Fenêtre Aide
http://www.google.com/
Lùh,5*)0hç&
Chargé

What was informatics again?

*"The sciences concerned with gathering, manipulating, storing, **retrieving**, and **classifying** recorded information."*

Instead of linked documents



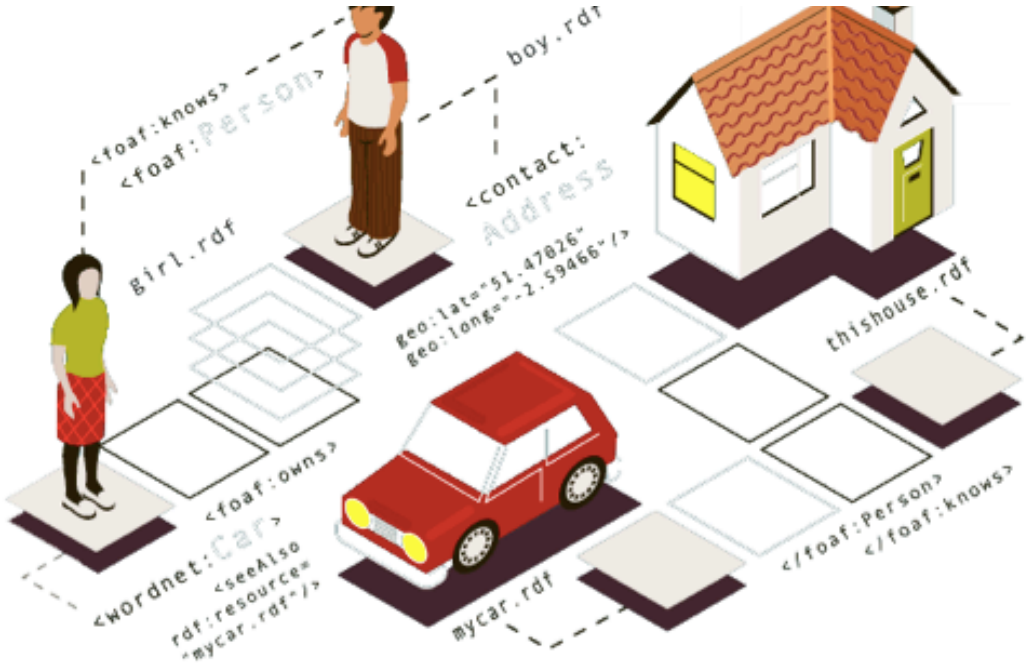
something is needed

some knowledge

Names/**concepts**

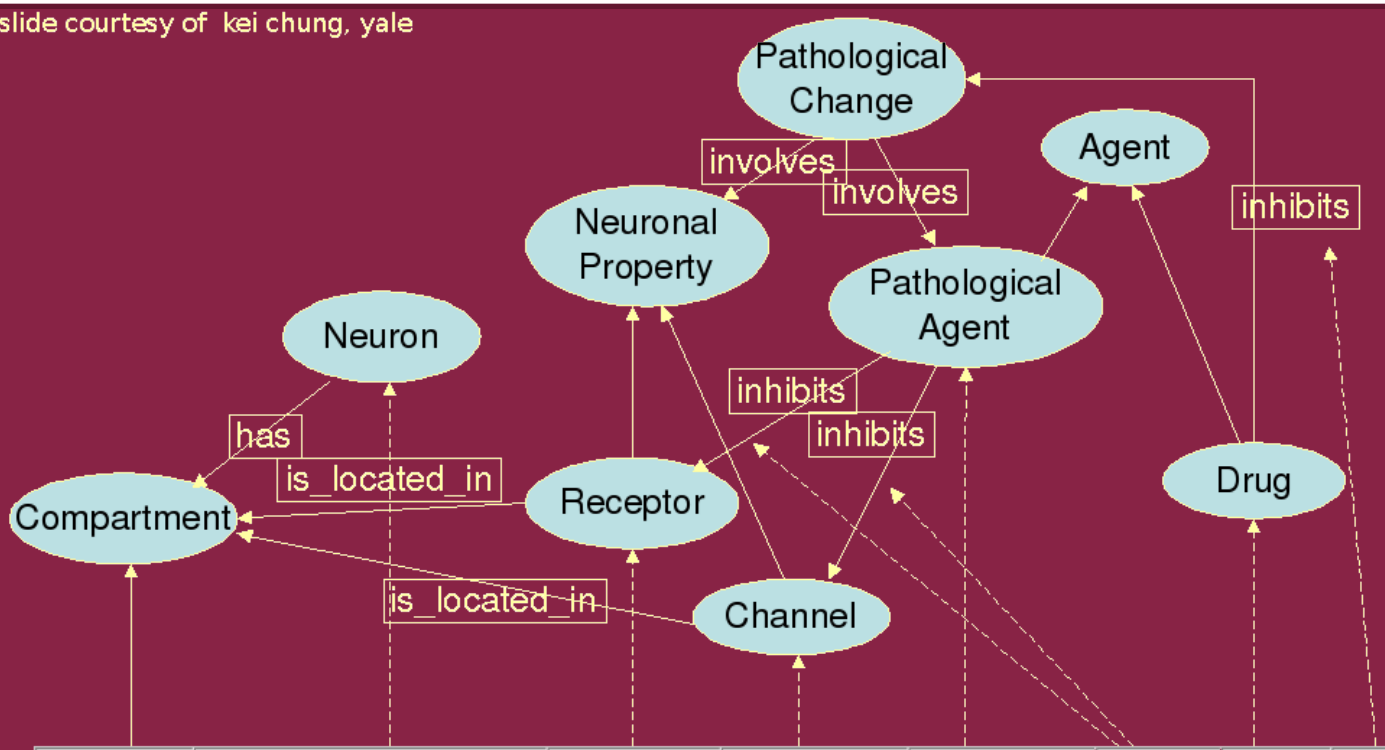
"Now! *That* should clear up a few things around here!"

A web of linked **concepts**



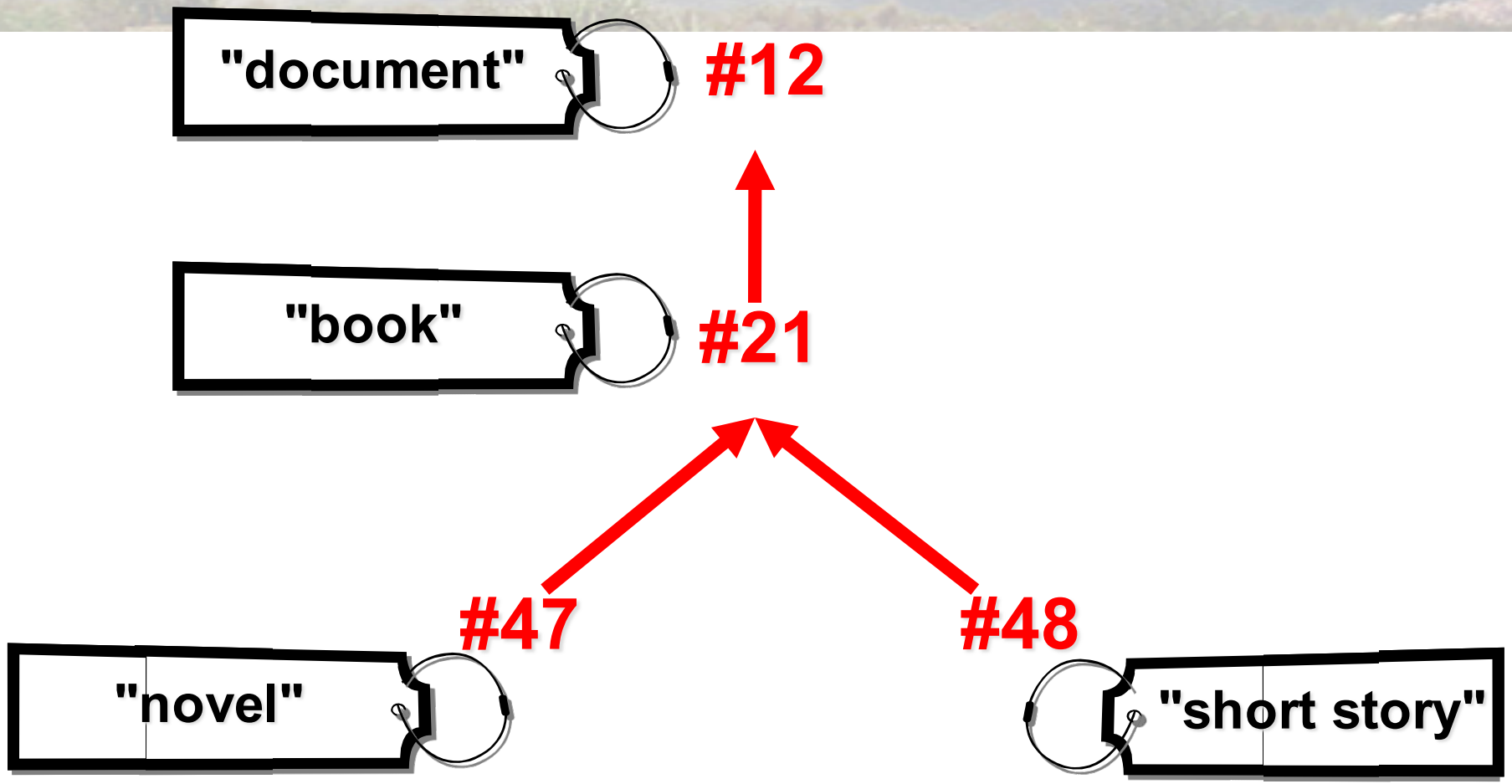
And concepts can span repositories

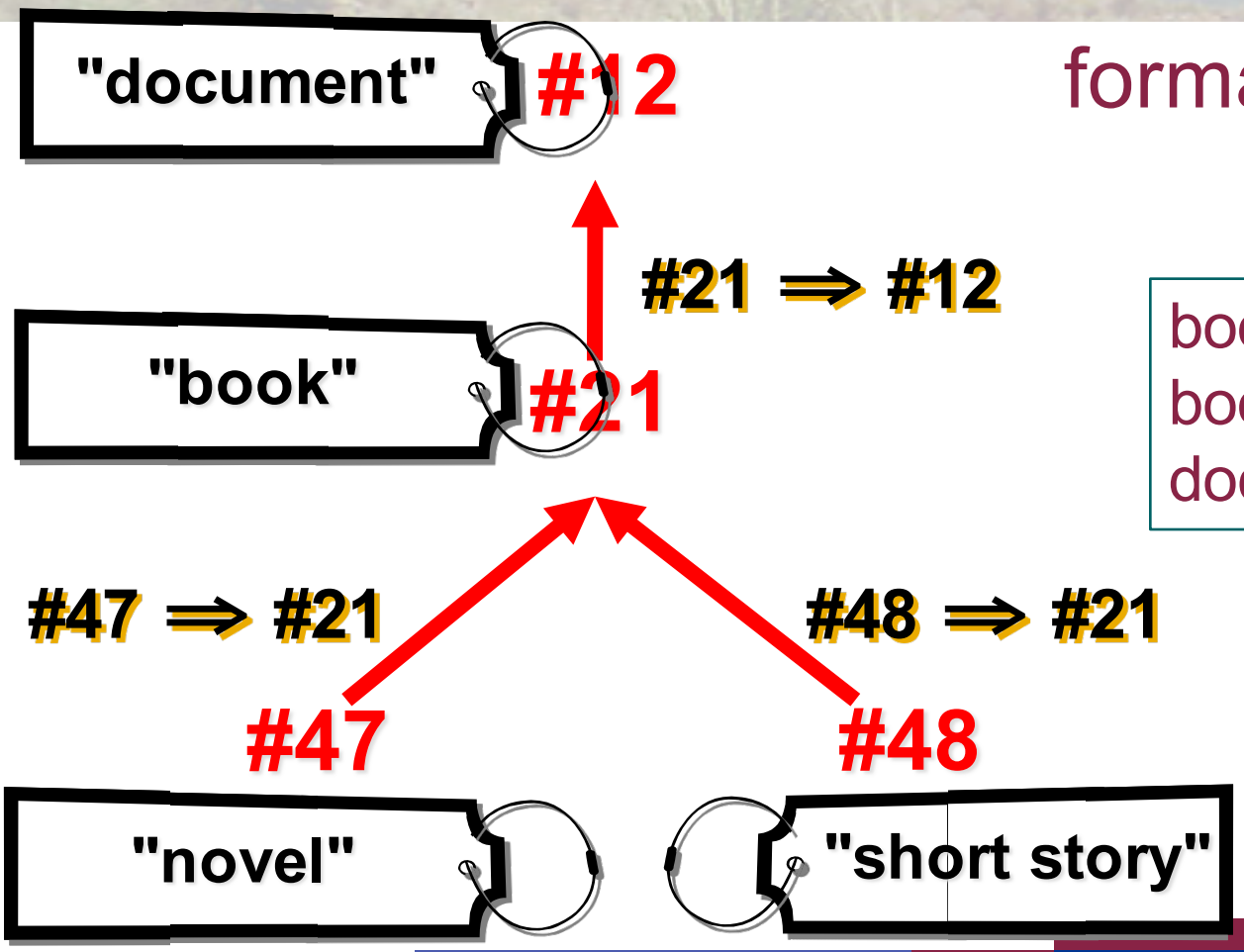
slide courtesy of kei chung, yale



Compartment	Cell: NeuronDB	Receptor	Channel	Pathological Agent (PA)	PA Action	Drug	Drug Action	Stage	Note	Detail
Soma	CA1 pyramidal neuron		I A	beta Amyloid	Inhibits			Early	View	66240
	Olfactory bulb mitral cell	GabaA						Early	View	66750
Dendrite	CA1 pyramidal neuron		I A	beta Amyloid	Inhibits			Early	View	66240
	Olfactory bulb mitral cell	GabaA						Early	View	66750
Unspecified	Oocyte		I L high threshold	beta Amyloid	Inhibits			Early	View	66252
								Early	View	66753
	CA1 pyramidal neuron			beta Amyloid	Inhibits			Early	View	66758
	CA1 pyramidal neuron	NMDA	I Calcium	beta Amyloid	Inhibits		Inhibits		View	66250

kind of





formalized ontological

knowledge

book(X) :- novel(X)
book(X) :- shortstory(X)
document(X) :- book(X)

Ontology

is not a synonym of

Taxonomy

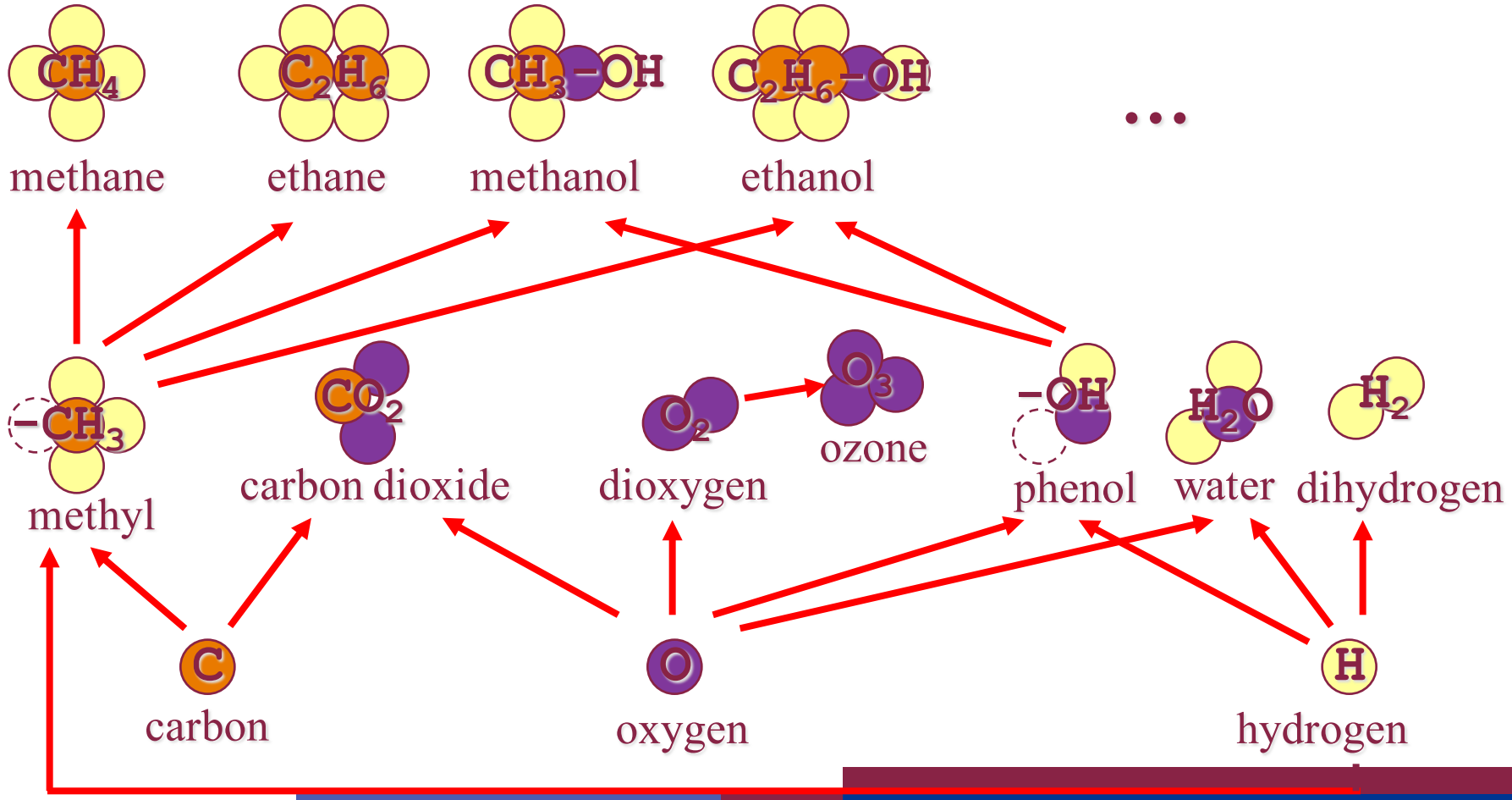
Taxonomical

knowledge is a kind of

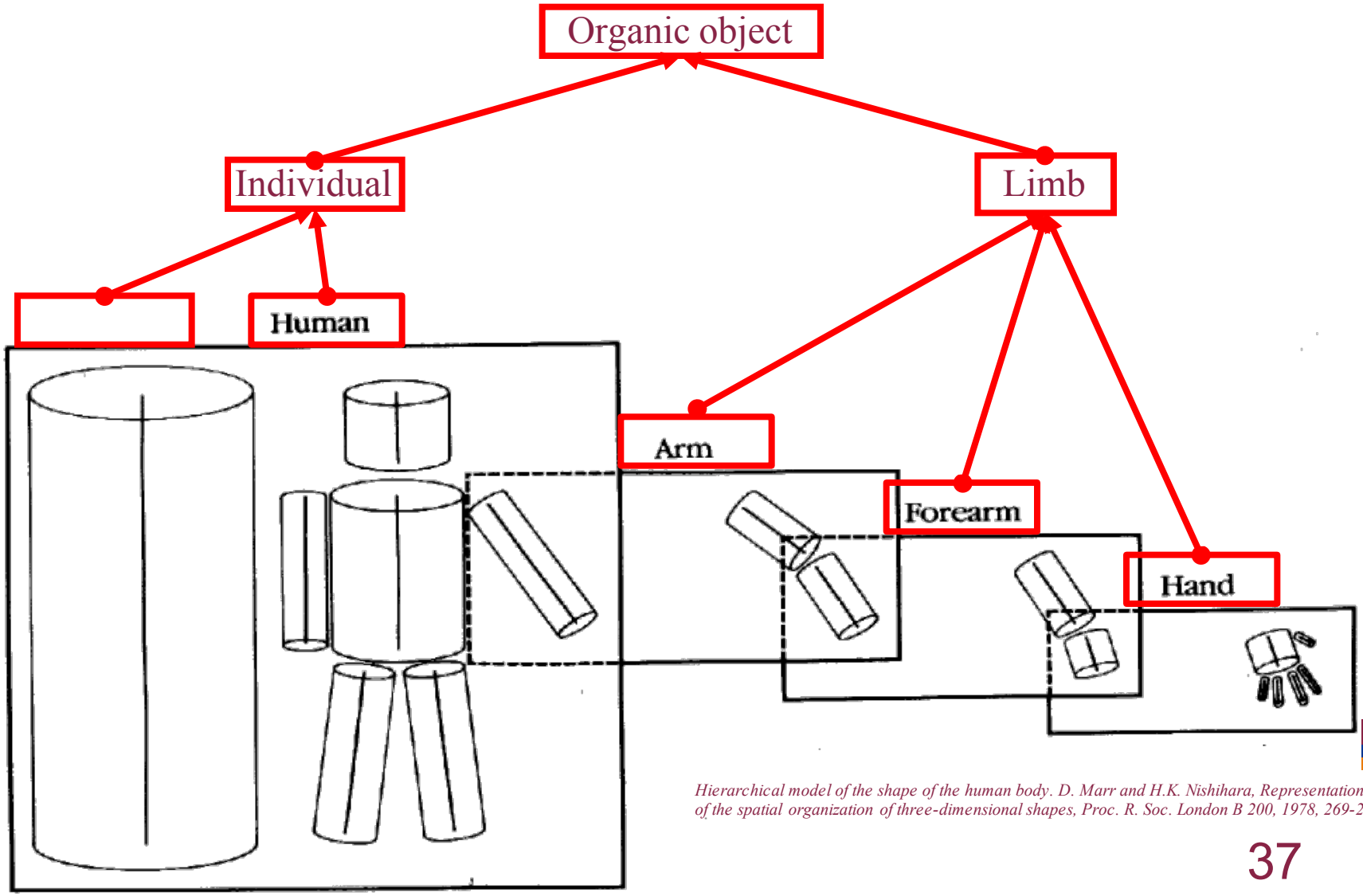
ontological

knowledge among others

part of


CH4
C2H6
CH3–OH
C2H6–OH
...
methane
ethane
methanol
ethanol
–CH3
CO2
O2
O3
–OH
H2O
H2
methyl
carbon dioxide
dioxygen
ozone
phenol
water
dihydrogen
C
O
H
carbon
oxygen
hydrogen

combine different kinds of ontological knowledge



Hierarchical model of the shape of the human body. D. Marr and H.K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, Proc. R. Soc. London B 200, 1978, 269-294.

Concepts are linked

Linked by statements called “triples”

A triple is a statement

subject predicate object

Any part of a triple may have to be uniquely identifiable. For this we use URLs.

An applied example

Triple 1

Subject: **<http://example.org/data/tree1>**
Predicate: **<http://example.org/terms/hasLikelihood>**
Object: **2342.323**
i.e. -lnL(tree1) = 2342.323

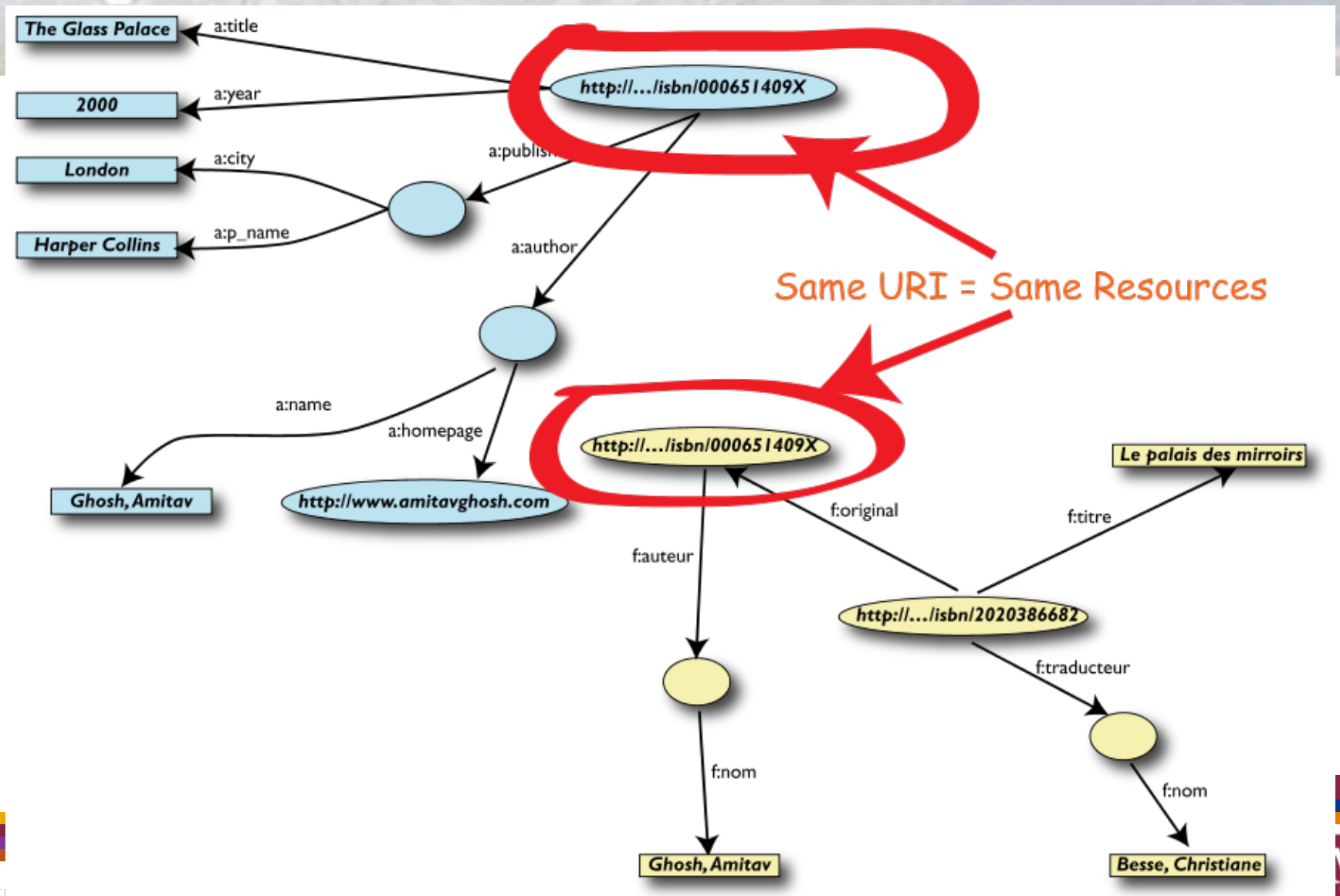
Triple 2

Subject: **<http://example.org/data/tree2>**
Predicate: **<http://example.org/terms/hasLikelihood>**
Object: **2341.184**
i.e. -lnL(tree2) = 2341.184

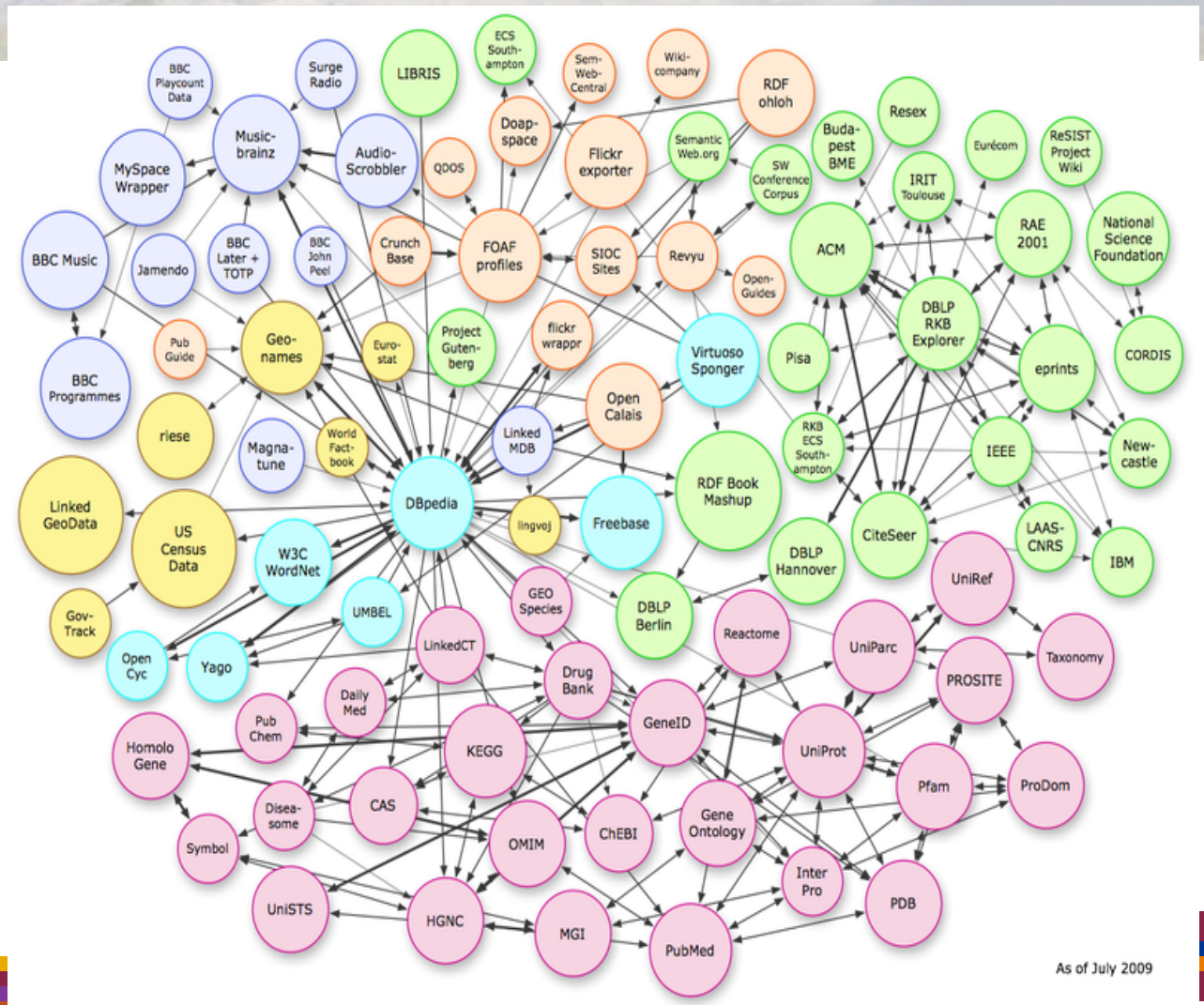
What's the better tree?

- The ontology defines what a likelihood is and how to compare negative log likelihoods.
- Hence, automated reasoning can conclude that tree2 is the better tree.

Compose networks of concepts



Concepts connected by statements



KLAP Laboratory

NM STATE

Ontology

a logical theory which gives an explicit, partial account of a conceptualization *i.e.* an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality ; the aim of ontologies is to define which primitives, provided with their associated semantics, are necessary for knowledge representation in a given context.

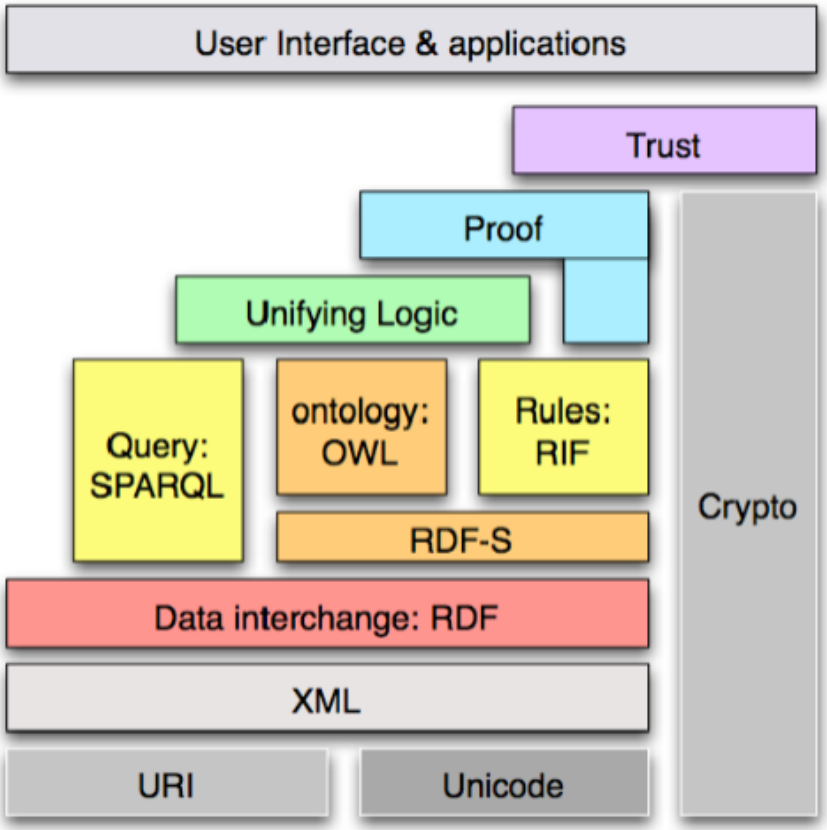
[Gruber, 1993] [Guarino & Giaretta, 1995] [Bachimont, 2000]

Semantic web

- "Semantic Web technologies"
  - A family of technology standards that 'play nice together', including:
    - Flexible data model (RDF)
    - Expressive ontology language (OWL)
    - Distributed query language (SPARQL)
  - Drive Web sites, enterprise applications

The technologies enable us to build applications and solutions that were not possible, practical, or feasible traditionally.

The (In)Famous Layer Stack



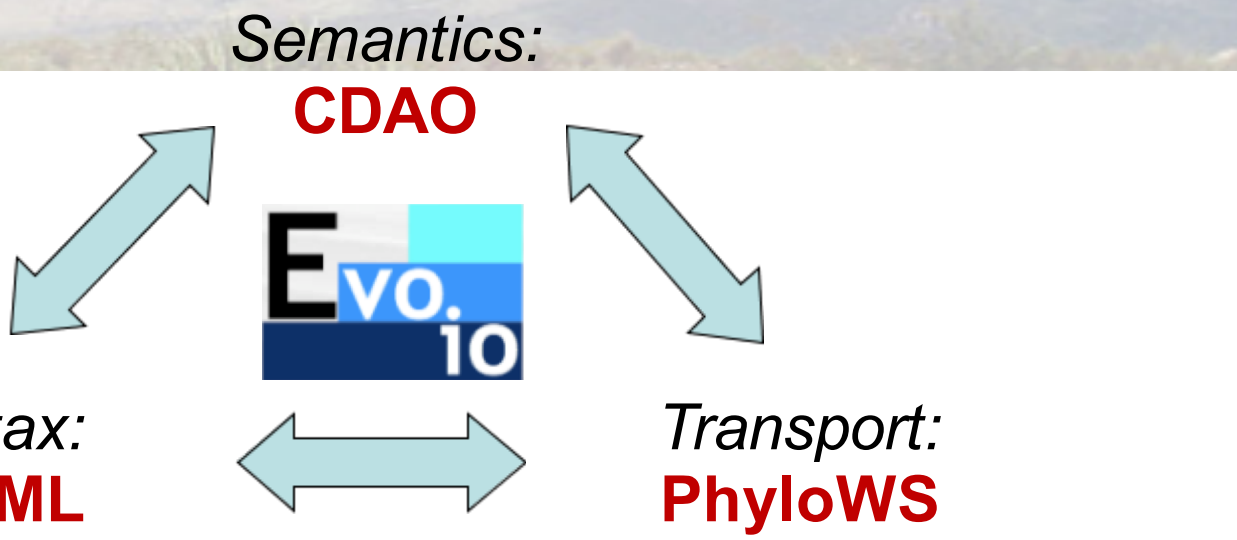
www.evoio.org

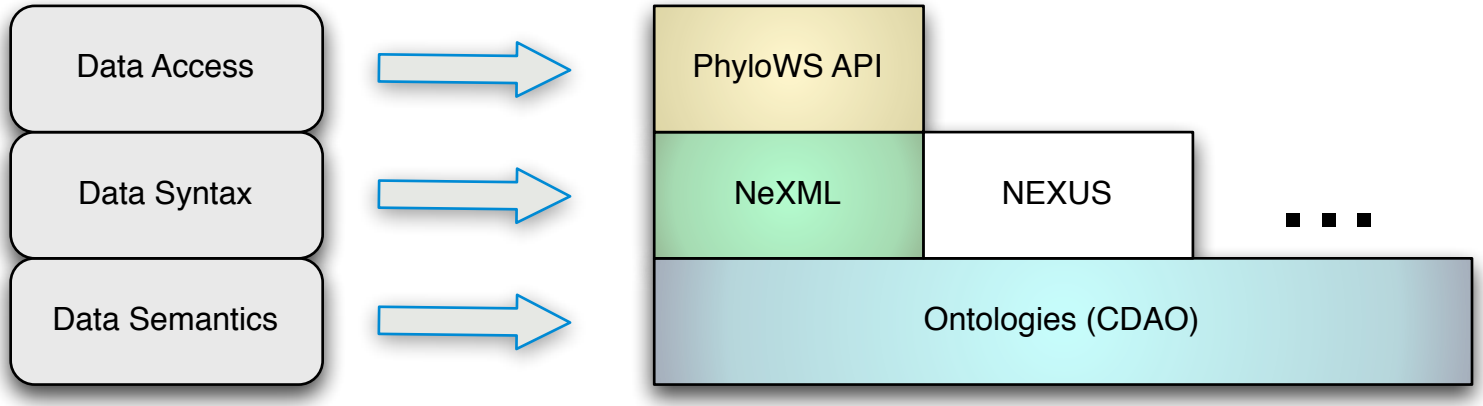
EVOIO STACK

KLAP Laboratory

NM STATE

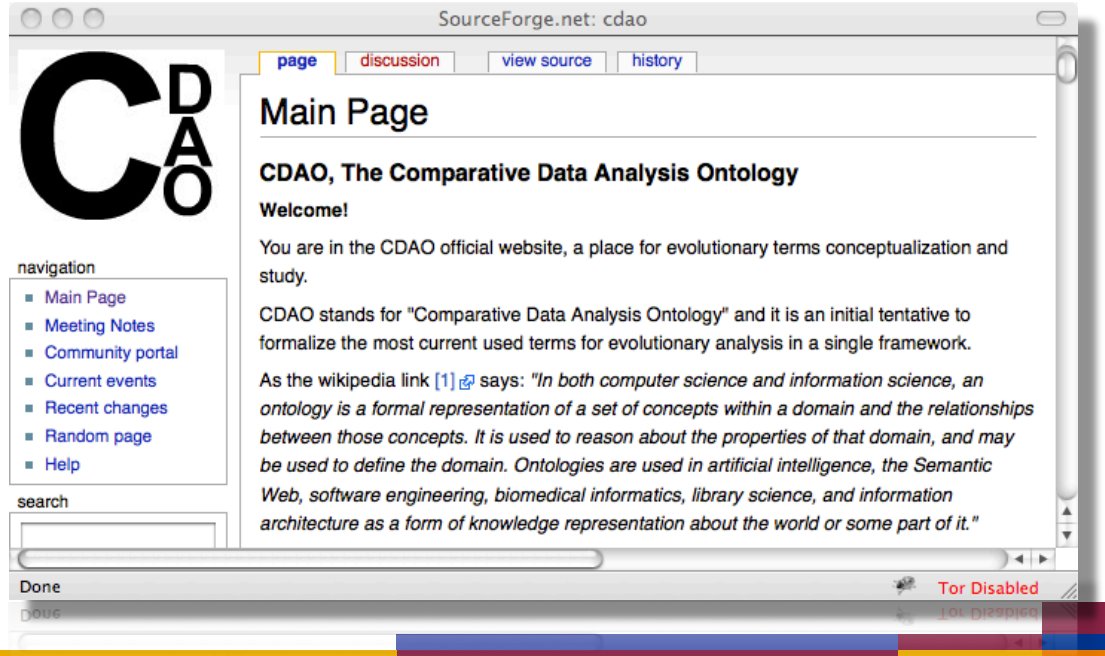
EVOIO STACK


Semantics:
CDAO
Evo.io
Syntax:
NeXML
Transport:
PhyloWS

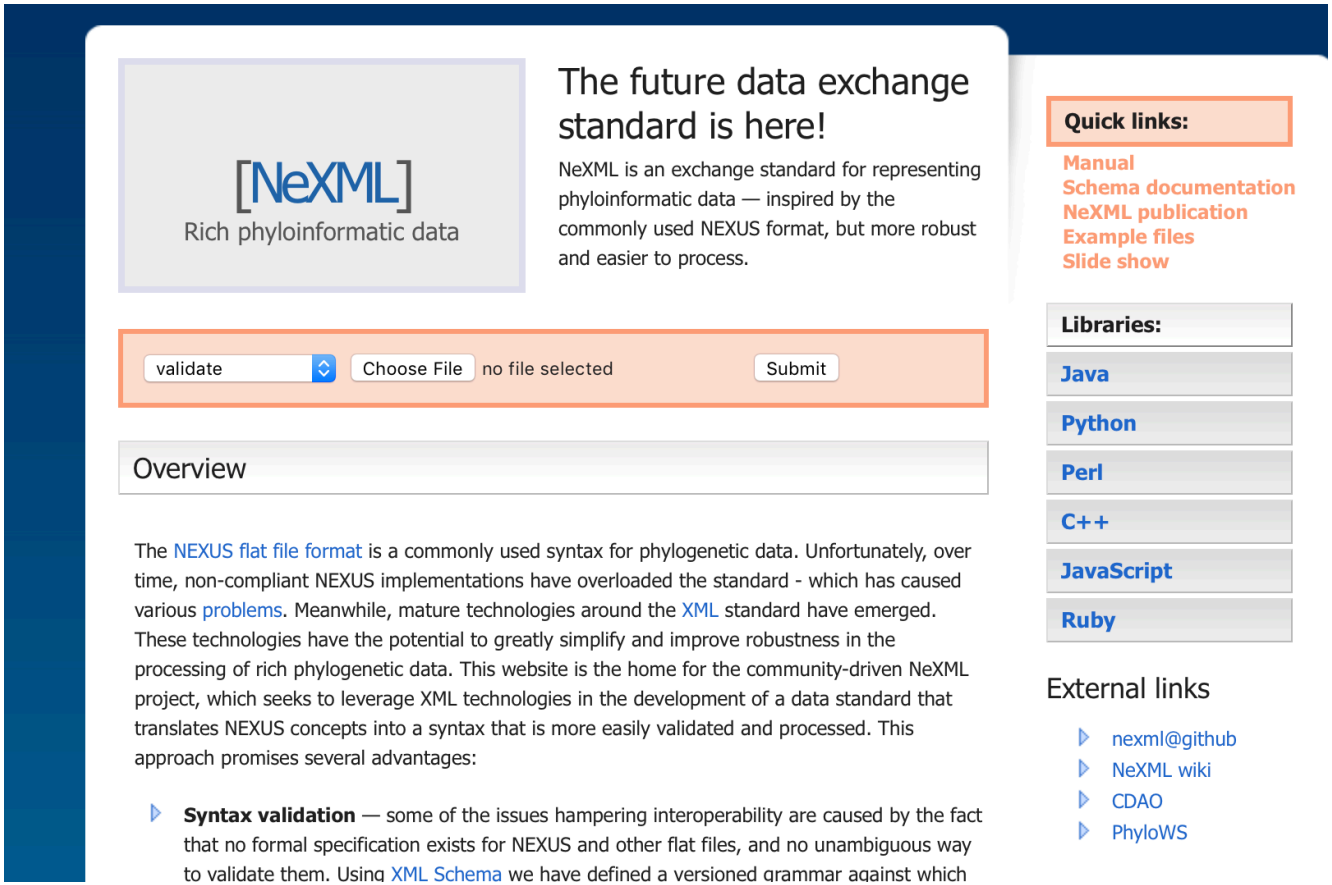

Data Access
Data Syntax
Data Semantics
PhyloWS API
NeXML
NEXUS
. . .
Ontologies (CDAO)

Concepts are defined in ontologies

"An ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to describe the domain."



Expressing concepts in data syntax



[NeXML]
Rich phyloinformatic data

The future data exchange standard is here!

NeXML is an exchange standard for representing phyloinformatic data — inspired by the commonly used NEXUS format, but more robust and easier to process.

validate | Choose File | no file selected | Submit

Quick links:

Manual
Schema documentation
NeXML publication
Example files
Slide show

Libraries:

Java
Python
Perl
C++
JavaScript
Ruby

External links

- nexml@github
- NeXML wiki
- CDAO
- PhyloWS

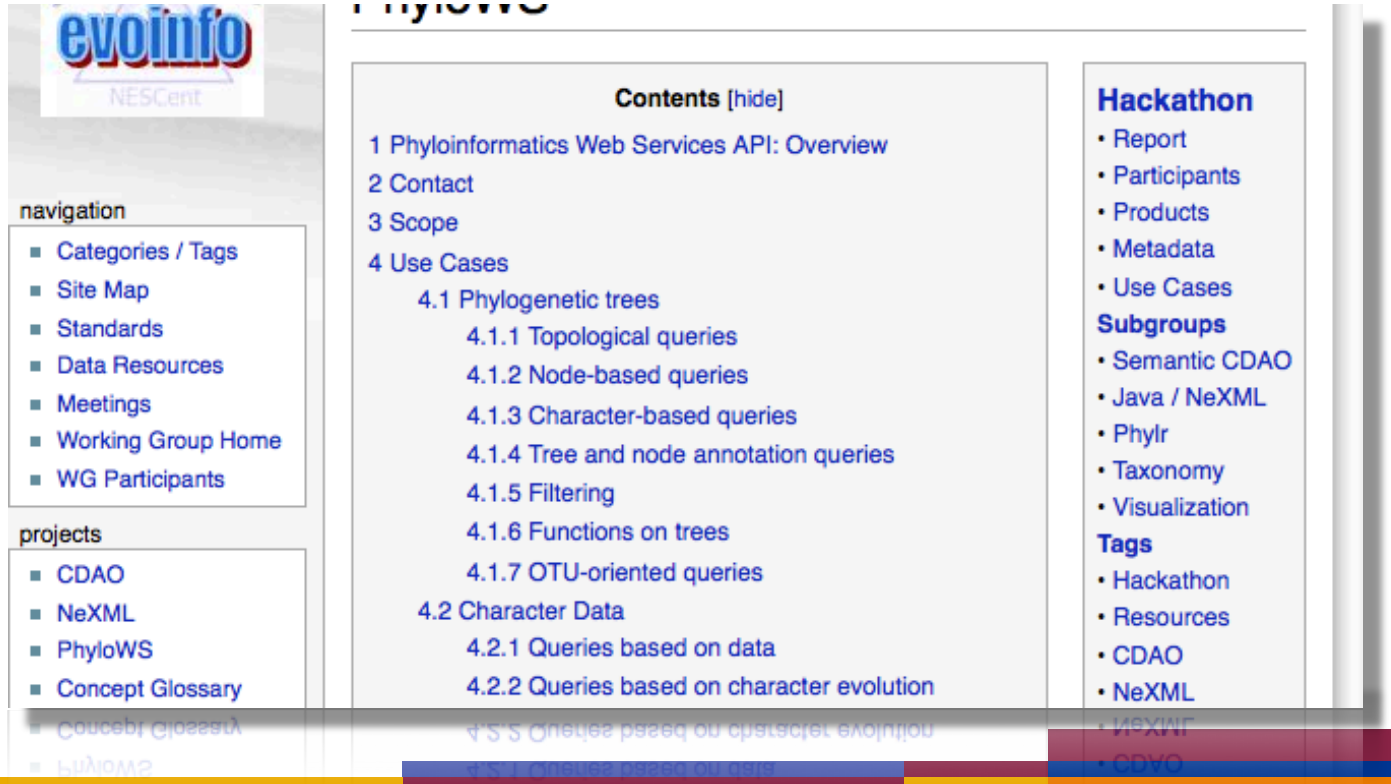
Overview

The NEXUS flat file format is a commonly used syntax for phylogenetic data. Unfortunately, over time, non-compliant NEXUS implementations have overloaded the standard - which has caused various problems. Meanwhile, mature technologies around the XML standard have emerged. These technologies have the potential to greatly simplify and improve robustness in the processing of rich phylogenetic data. This website is the home for the community-driven NeXML project, which seeks to leverage XML technologies in the development of a data standard that translates NEXUS concepts into a syntax that is more easily validated and processed. This approach promises several advantages:

- **Syntax validation** — some of the issues hampering interoperability are caused by the fact that no formal specification exists for NEXUS and other flat files, and no unambiguous way to validate them. Using XML Schema we have defined a versioned grammar against which

URLs for phylogenetics

PhyloWS doesn't just provide an anchor to identify phylogenetic data, it also enables searching and retrieval.



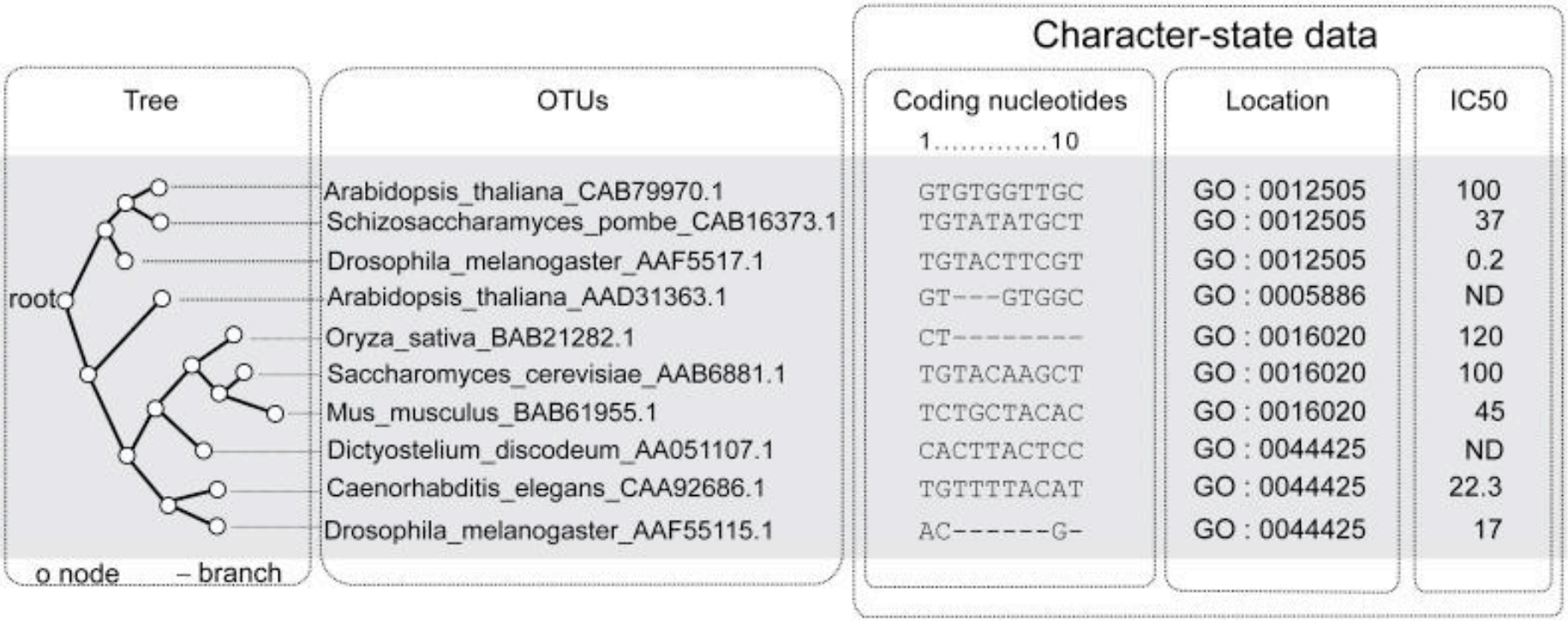
www.evolutionaryontology.org

THE COMPARATIVE DATA ANALYSIS ONTOLOGY (CDAO)

KLAP Laboratory

NM STATE

Character-state data model



CDAO development

Specification:
Study use-cases to clarify scope

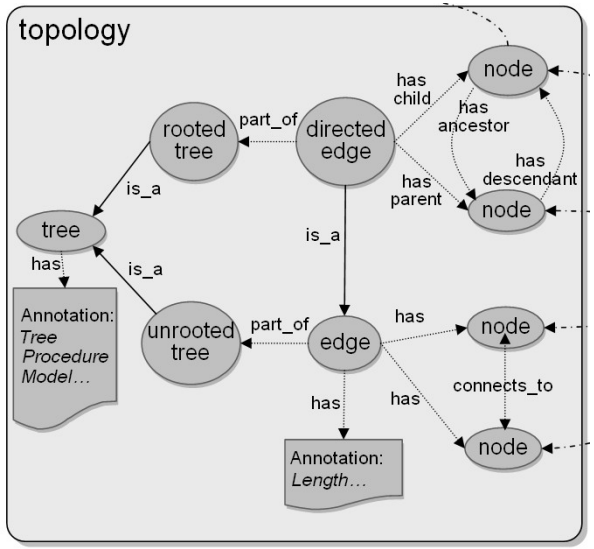
Choice of representation:
Choose language and development tools

Conceptualization:
- Identify terms from use cases, artefacts
- Build concept glossary
- Classify key concepts and relations

Implementation:
Formalize the concepts and relations using the chosen language and tools

Evaluation:
Test the ontology for its ability to represent data called for in the use cases, and to support reasoning

Ontology refinement

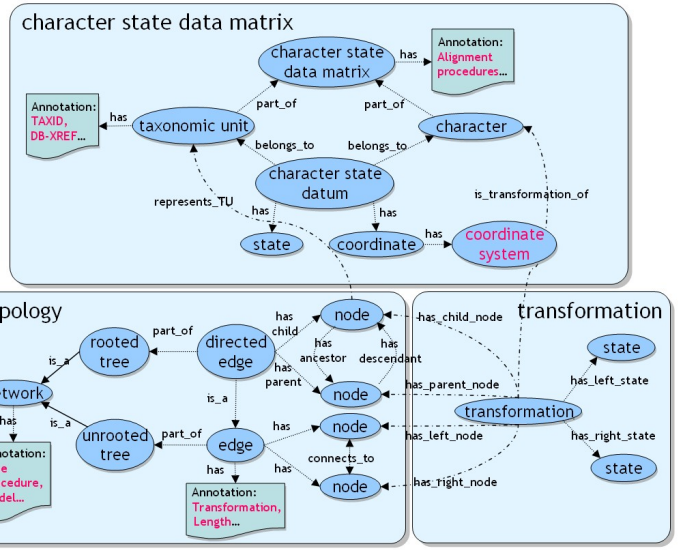


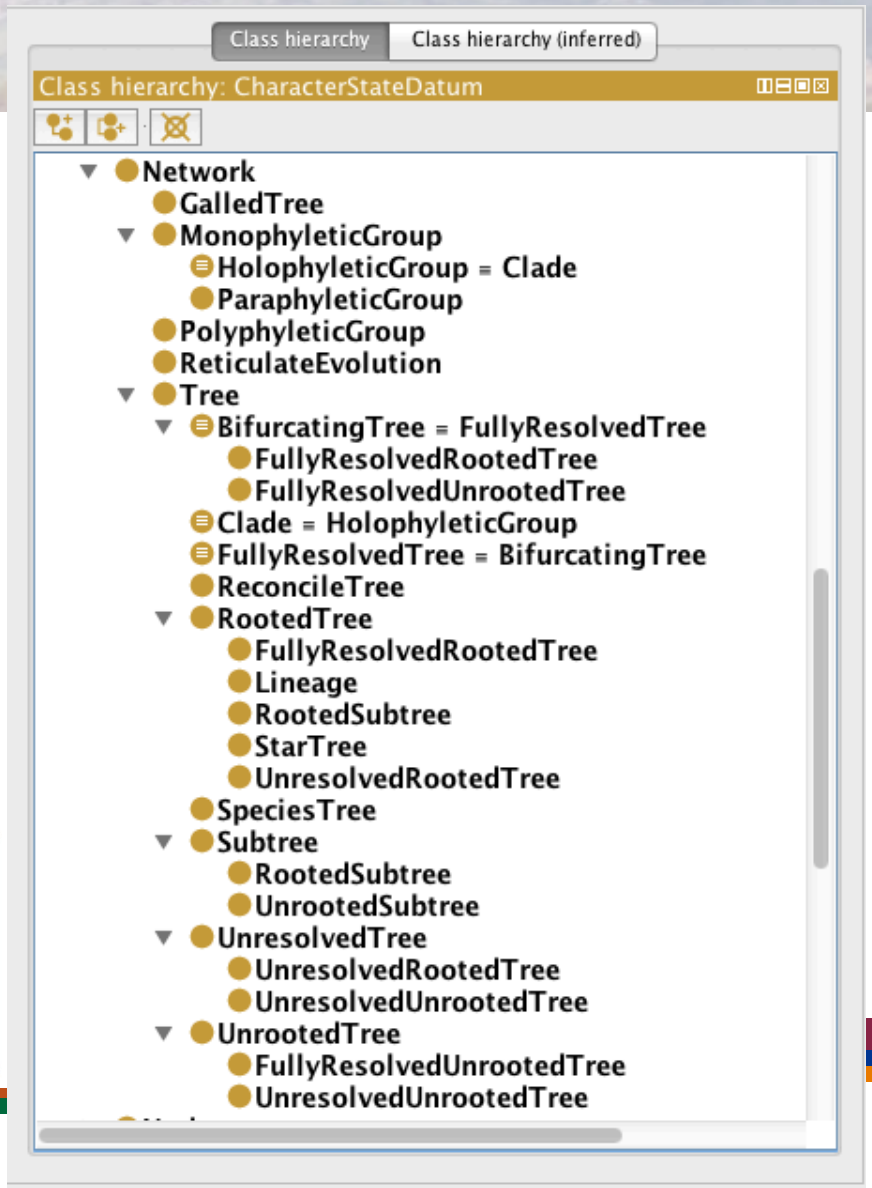
Some CDAO concepts and relations in regard to trees

KLAP Laboratory

CDAO Organization

- Topologies
  - Nodes and Edges
  - Networks
  - Rooted and Unrooted Trees
- Characters and Character State
- Data Matrices
  - Continuous
  - Categorical (DNA, RNA, Amino Acids)
  - Compound
- TUs
- Annotations
  - Lengths
  - Methods





www.nexml.org

NEXML

KLAP Laboratory

NM STATE

Motivations and design principles

- NeXML is an encoding format that should meet the following requirements
  - Easy to produce by tools
  - Easy to parse and manipulate
  - Easy to transmit and store
  - Easy to extend
- **XML** formats provide most of these features
  - Extensible markup language to represent and serialize structured documents
- Design
  - General “block” structure similar to NEXUS
  - Trees/Networks described following GraphML
  - Each entity can have a list of properties attached (e.g., metadata)

Current Support

- **OTUs**
- **characters**: dna, rna, nucleotide, protein, categorical, continuous, restriction
- **trees**: graphml trees and networks, various edge formats and rootings

Character Classes

	Sequence	Cells
DNA	`DnaSeqs`	`DnaCells`
RNA	`RnaSeqs`	`RnaCells`
Protein	`ProteinSeqs`	`ProteinCells`
Standard	`StandardSeqs`	`StandardCells`
Continuous	`ContinuousSeqs`	`ContinuousCells`
Restriction	`RestrictionSeqs`	`RestrictionCells`

Tree Classes

	Float	Int
Network	`FloatNetwork`	`IntNetwork`
Tree	`FloatTree`	`IntTree`

neXML

```
<!-
    nested inside /nexml/characters element
-->
    <format>
        <states id="states1">
            <state id="s1" symbol="1"/>
            <state id="s2" symbol="2"/>
            <state id="s3" symbol="3"/>
            <state id="s4" symbol="4"/>
            <polymorphic_state_set id="s5" symbol="5">
                <mapping state="s1"/>
                <mapping state="s2"/>
            </polymorphic_state_set>
            <uncertain_state_set id="s6" symbol="6">
                <mapping state="s3"/>
                <mapping state="s4"/>
            </uncertain_state_set>
        </states>
        <char states="states1" id="c1"/>
        <char states="states1" id="c2"/>
    </format>
    <!-- row elements follow -->
```

- ✓ formally defined syntax
- ✓ OTS tools to validate
- ✓ extensible
- ✓ versioned

evoinfo.nescent.org/**PhyloWS**

PHYLOWS

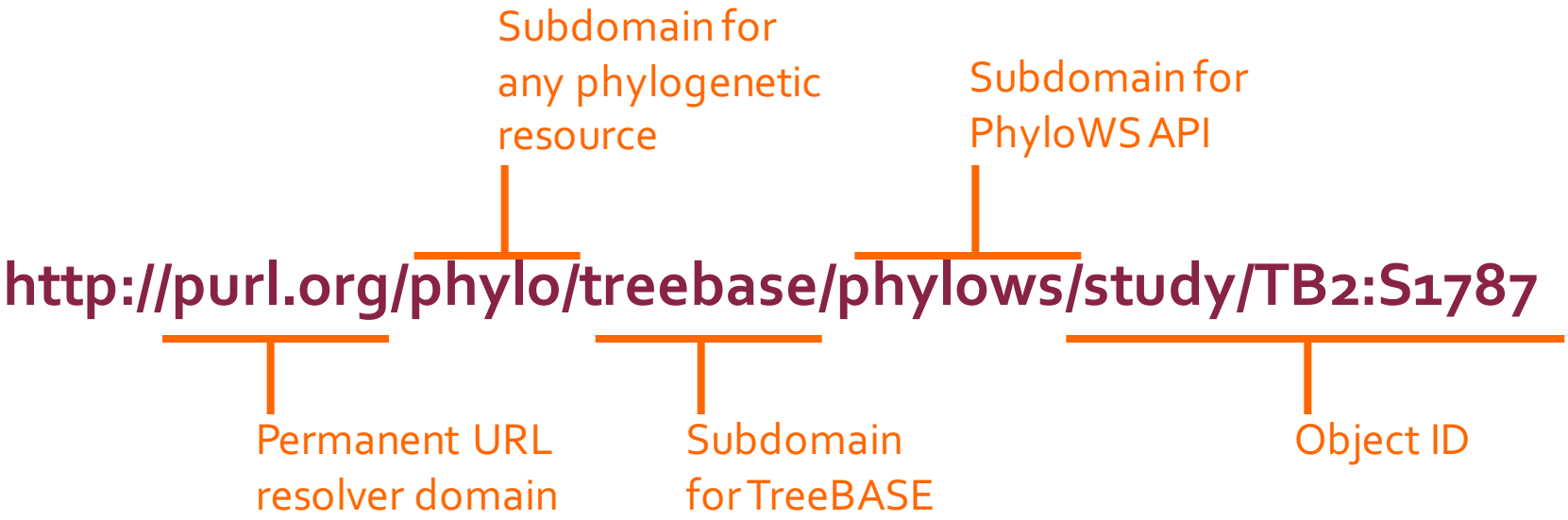
KLAP Laboratory

NM STATE

Phylows

- Web Service Interface – i.e., a web-based API to phylogenetic repositories
- Supports a variety of queries, e.g.,
  - Topological queries (e.g., common ancestors, node information, character-based queries)
  - Filtering queries (e.g., filter trees based on metrics, characteristics, nodes)
  - Functions on trees (e.g., rerooting, comparison, aggregation)
  - Character queries (e.g., compare OTUs, retrieve matrices)

PhyloWS URL API



HISTORY AND APPLICATIONS

Process

Evolutionary Informatics Working Group

WG1 → NESCent

WG2 → NESCent

WG3 → NESCent

WG4 → NESCent

2007

Planning meeting (Philly)

Phylohackathon (Durham)

NESCent

neXML

2008

CDAO (Durham)

PhyloWS (Tokyo)

NESCent

Google Summer-of-Code

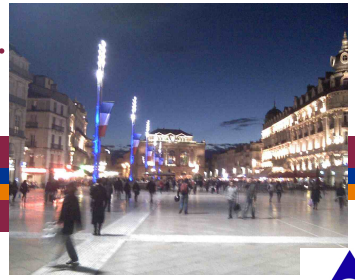
2009

DB interop hackathon (Durham)

NESCent

Google Summer-of-Code

2010



Phyloinformatics VoCamp

Google Summer-of-Code

2011

Hackatons, Interop., Phylogenies

NESCent

Evo.io

Biodiversity Information Standards TDWG

KLAP Laboratory

Early adopters

- **Nexml parsers and writers**:
  - **mesquite** (java NeXML class libraries)
  - **Bio::Phylo** (BioPerl compatible)
  - **pyNexml** (python)
  - **DAMBE**
  - **NCL** (C++)
  - **JavaScript**

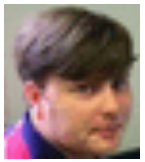
Full Adopters:

TreeBase

PhenoScape

Implementations

Nexml IO implementations
- Bio::Phylo, BioPerl, HIVQuery (Perl)
- DAMBE
- Mesquite, TreeBase (Java) *
- **Phenex (Java via XMLbeans)**
- DendroPy (Python)
- NCL (C++)
- BioRuby (Ruby)



Jim Balhoff
National Evolutionary Synthesis Center,
Durham, NC

PhyloWS
- Phylr
- **TreeBase2**
- Bio::Phylo (ToLWeb and TimeTree via screen-scraping)
- PhenoScape

Rutger Vos
Naturalis,
Netherlands

Vivek Gopalan
Bioinformatics and Computational
Biosciences Branch (BCBB), NIAI

CDAO
- **Nexplorer3**
- CDAO-store
- TreeBase2
- PhenoScape

Enrico Pontelli
New Mexico State University,
Las Cruces, NM

www.cs.nmsu.edu/~cdaostore

CDAOSTORE = ASP+CDAO+PHYLOWS

KLAP Laboratory

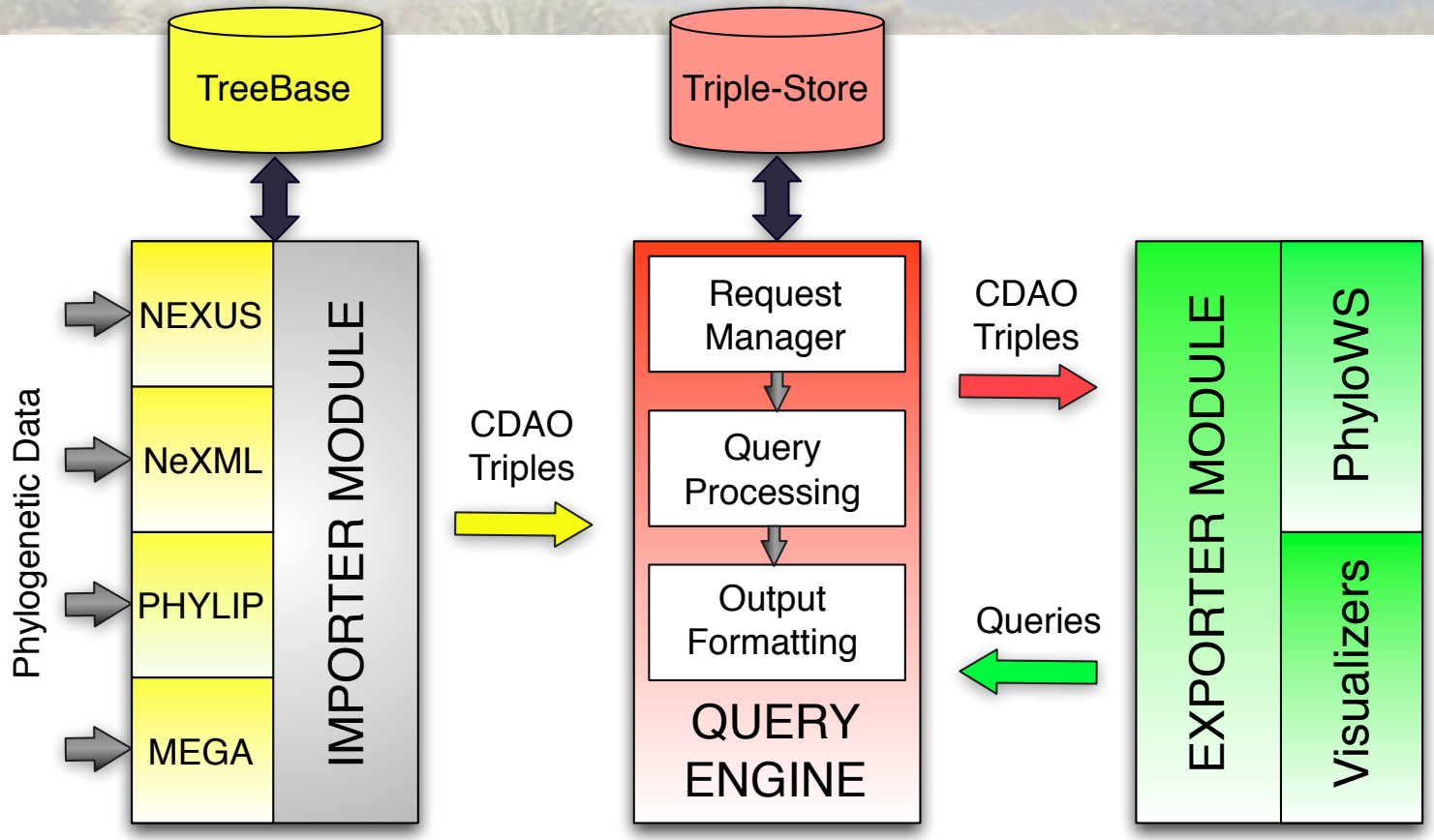
NM STATE

CDAOStore

- **Public repository of phylogenies annotated with CDAO**
  includes a complete dump of TreeBase
- **Ability to import most commonly used data formats**
  NEXUS, NeXML, MEGA, PHYLIP
- **Ability to provide RDF, NeXML, and graphical output**
- SPARQL and domain-specific querying capabilities

"CDAO-Store: Ontology-driven Data Integration for Phylogenetic Analysis" **B. Chisham, B. Wright, T. Le, T. Son, E. Pontelli. BMC Bioinformatics, 12:98, 2011**

CDAOStore


TreeBase
Triple-Store
Phylogenetic Data
NEXUS
NeXML
PHYLIP
MEGA
IMPORTER MODULE
CDAO Triples
Request Manager
Query Processing
Output Formatting
QUERY ENGINE
CDAO Triples
Queries
EXPORTER MODULE
PhyloWS
Visualizers

Why LP? Why ASP?

- Why LP?
  - Ease of modeling phylogenies and character matrices
  - Ease of encoding tree operations
  - Ability to modularly add knowledge and constraints (e.g., topological constraints)
- Why ASP?
  - Scalable engines
  - Link with RDF been explored
  - Elegant
- Some issues, e.g.,
  - Real numbers (Lua extensions)
  - Scale

Implementation Considerations

- RDF triples from CDAO are imported as ASP facts

```
tree(t_id).                          % t_id is a tree
tree_label(t_id,lab).                % t_id has label "lab"
tree_is_defined_by(t_id,s_id).       % t_id is studied in s_id
tree_ntax(t_id,n_Taxa).              % t_id has n_Taxa taxa
edge(t_id,n1,n2).                    % t_id contains an edge from n1 to n2
edge_length(t_id,n1,n2,l).           % l is the length of the edge (n1,n2) in t_id
represents_TU(t_id,n1,tu_id).        % node n1 of t-id represents tu_id
taxon_id(tu_id,taxon_id).            % tu_id represents taxon_id
matrix_type(m_id,m_type).            % matrix m_id is of the type "m_type"
belongs_to_TU(m_id,cell,tu_id).      % cell in matrix m_id belongs to tu_id
```

PhyloWS

- Web Service API specification for Phylogenetic analysis services and repositories
- Designed from community studies
- Scope
  - Creation of phylogenies
  - Maintenance of repositories
  - Application of Phylogenies

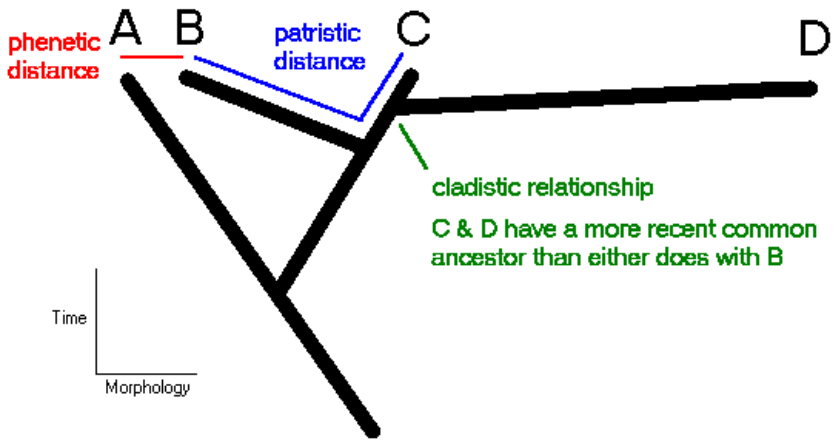
Vaguely Specified
Application Dependent

  - **Querying and Retrieval**

Our Focus

Classes of Queries: Node Queries

- Node-Oriented Queries
  - Retrieval of nodes of phylogenies
  - Various types of search criteria
- Determine Nearest-Common Ancestor of two or more leaves
- Patristic distance among two nodes
- Determine nodes that have distance ≥ c from the root
- Determine lineage of a taxa



Classes of Queries: Node Queries

- Sample Encodings:
  - Most recent common ancestor:

```
mrca(N, S) :- tree(T), node(T,N), set_of_taxa(S),
              common_ancestor(T,N,S),
              not non_min(T,N,S).
non_min(T,N,S) :- common_ancestor(T,N1,S),
                  ancestor(T,N,N1), N1 != N.
```

  - Patristic distance:

```
patristic_distance(T,N1,N2,D):- mrca(M, s), D=L1+L2,
              distance_to_ancestor(T,M,N1,L1),
              distance_to_ancestor(T,M,N2,L2).
distance_to_ancestor(T,N1,N2,L):- parent(T,N1,N2),
              edge_length(T,N1,N2,L).
distance_to_ancestor(T,N1,N2,D):- parent(T,Nb,N2),
              edge_length(T,Nb,N2,L),
              distance_to_ancestor(T,N1,Nb,L2), D=L+L2.
```

Classes of Queries: Clade Queries

- Focus on computing Clades of a phylogeny satisfying given properties
- Determine minimum spanning clade of a set of taxa
  - Popular query – e.g., Phylomatic
- Determine clades whose taxa have a certain character

Classes of Queries: Clade Queries

- Minimum spanning clade:

```
minimum_clade(S,N):- tree(T), node(T,N), mrca(N, S).
minimum_clade(S,D):- tree(T), node(T,N), mrca(N, S),
                     ancestor(T,N,D).
label(D,TU_Label):- tree(T), minimum_clade(_,D),
                     represents_TU(T,D,TU),
                     tu_label(T,TU,TU_Label).
```

- Find clade whose taxa have a given character

```
clade(s).
{in_clade(s,N) : leaf(t,N)}.
member(N,s):- in_clade(s,N).
  :- minimum_clade(s, N), not in_clade(s, N).
  :- in_clade(s,N), represents_TU(T,N,TU),
     belongs_to_TU(M,Cell,TU), not belongs_to_Character(M,Cell,c).
```

Classes of Queries: Tree Queries

- Queries focused on selecting phylogenies meeting certain criteria or determining properties of phylogenies
- Some Examples
  1. Extract phylogenies satisfying certain topological constraints, e.g.,
     - Bound on overall width
     - Bound on number of taxa

```
matching_ntax(T, T1, Cnt):-
        tree_ntax(T1,N), tree_ntax(T,Cnt),
        Cnt <= N+c, Cnt>=N-c.
```

Classes of Queries: Tree Queries

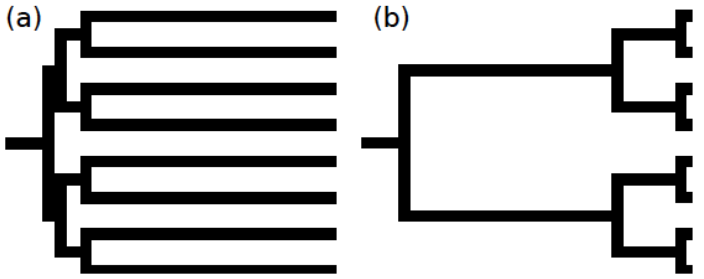
2. Extract phylogenies containing a given set of taxa

```
connect_tu(TU,S,T): tree(T), otu_set(S),
        member(TU,S),
        represents_TU(T,Node,TU).
tree_otus(T,S):- tree(T), otu_set(S),
        not not_representative(T,S).
not_representative(T,S) :- member(TU,S),
                           not connect_tu(TU,S,T).
```

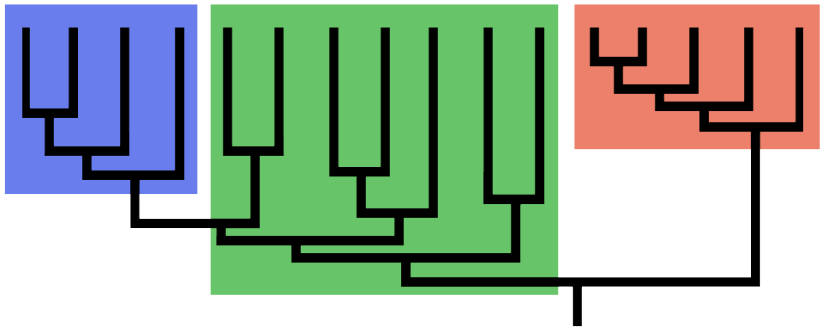
2. Determine structural metrics of phylogenies (e.g., imbalance factors)
   - Colless Imbalance coefficient
   - I2 Imbalance coefficient
   - Pybus Gamma statistics

Classes of Queries: Tree Queries

4. Measure and search phylogenies based on stemminess



5. Determine Monophyletic Groups



Classes of Queries: Data Queries

- Determine metadata of phylogenies or data underlying the construction of the phylogeny
- Metadata queries, e.g.,
  - Types of characters used
  - Authors, publications, study id
- Character Data queries, e.g.,
  - Determine matrices describing a taxon
  - Determine all characters describing a certain taxon
  - Project a matrix over a set of taxa or characters

Classes of Queries: Data Queries

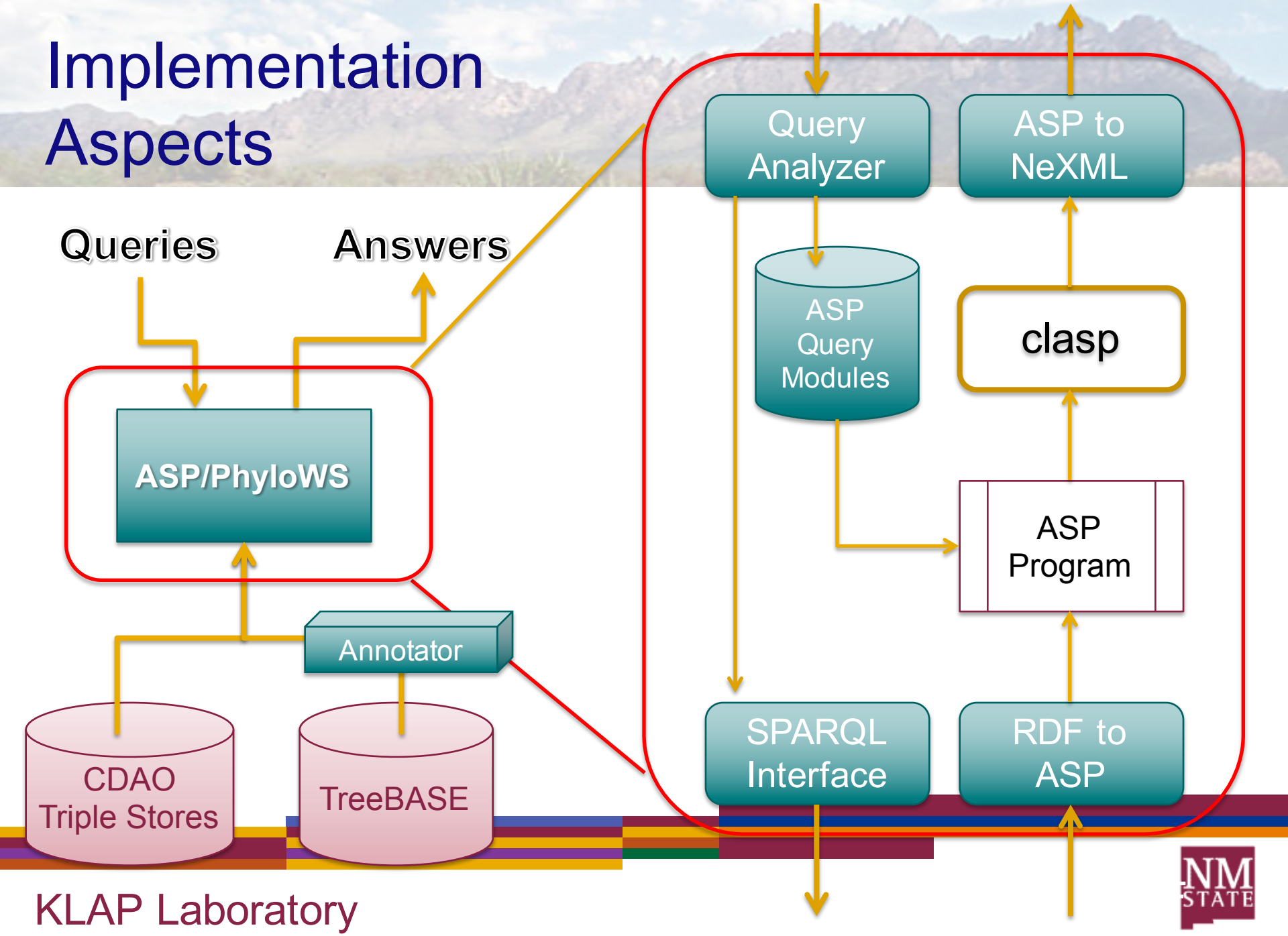
– identify the character that appears in all matrices containing data for a given set of OTUs.

```
matching_matrices(M,S):- otu_set(S), has_TU(M,_),
                                  not not_complete(S,M).
not_complete(S,M) :- member(E,S), not has_TU(M,E).
has_character(M,C):-
                belongs_to_character(M,_,C).
character_in_all_matrices(C):-
                matching_matrices(M,S),
                has_character(M,C),
                not incomplete_mat_char(M,C).
incomplete_mat_char(M,C) :- matching_matrices(M1,S),
                not has_character(M1,C).
```

Transformation Queries

- Phylogenies can be manipulated to meet the needs of downstream applications
- Comparison of trees
  - E.g., Robinson-Foulds distance between trees
- Transforming phylogenies
  - Remove a set of taxa from a phylogeny
  - Rerooting a phylogeny
  - Modifying branch lengths (e.g., ultrametricizing)
- Combining phylogenies
  - Computing consensus tree
  - Computing supertree

Implementation Aspects



Some Implementation Considerations

- ASP modules for each query
- Need extensions
  - Support for Real Numbers: Lua functions (clasp-supported extensions)
  - Issues of scale – recursive ASP
- Most encodings are very compact and elegant

Some Implementation Considerations

- Given a phylogeny T, transform branch lengths using ultrametricization

```
#begin_lua
   function greater(a,b)
      if tonumber(a) > tonumber(b)
         then return 1 else return 0


     end
  end
#end_lua.
```

```
tree_length(T,L) :- leaf(T,N2),
    node_length(T,N2,L),
    not dominated(T,L).
dominated(T,L) :- leaf(T,N),
     node_length(T,Nb,L1),
     1==@greater(L1,L).
edge_length(t2,N1,N2,L) :-
     edge_length(t,N1,N2,L),
     not leaf(t,N2).
edge_length(t2,N1,N2,L) :-
     edge_length(t,N1,N2,L1), leaf(t,N2),
     tree_length(t,L2),
     node_length(t,N1,L3),
     L:=@subtract(L2,L3).
```

Status of the Project

- Framework implemented as a self-contained applications
  - SPARQL interface to CDAO triple stores
  - NeXML exporter
  - Relatively simple query analyzer
- Queries
  - 4 types of Node queries
  - 2 types of Clade queries
  - 14 types of Tree queries
  - 4 types of Data queries
  - 8 types of Phylogeny Function queries

Status of the Project

- Effective queries computation over the CDAOStore triple store
  - Complete import of all phylogenetic studies from TreeBASE
  - ~3,000 studies
  - 5,794 character data matrices
  - 8,621 phylogenies
  - Over 470GB of RDF triples
- Majority of queries require from a few seconds to a couple of minutes

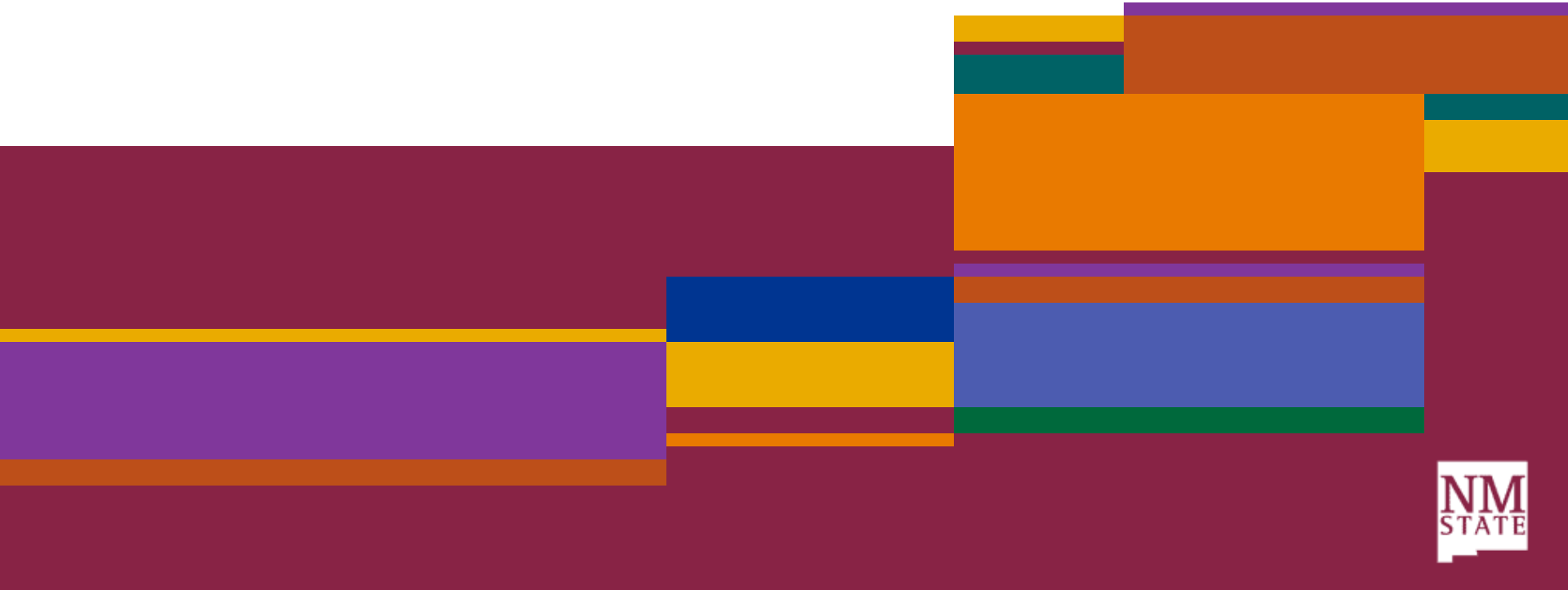
Preliminary Results

Query	Data Size	Time	Query	Data Size	Time
N1	644KB	2.3	N2	698KB	2.84
N3	685KB	1.06			
C1	1.5MB	2.94	C2	31.9MB	13.42
T1	698KB	1.21	T2	698KB	1.00
T3	698KB	0.81	T4	1.2MB	0.45
T5	65.6KB	0.02	T6	31.9MB	1913.8
T7	644KB	0.83	T8	6.2KB	0.82
D1	44MB	15.27	D2	34MB	10.31
D3	34.9MB	14.02	D4	19.1MB	5.94

Preliminary Results

- Some concerns
  - Some queries require a creative implementation
    - Supertree Computation: infeasible in clasp for larger trees; recursive computation
    - Lua calls are not always handled efficiently by clasp
  - Some queries require an intelligent query analysis
    - Clasp unable to handle fact bases larger than ~500MB
  - Had to restrict some experiments to ~83% of the studies in TreeBASE

Current and Future Directions

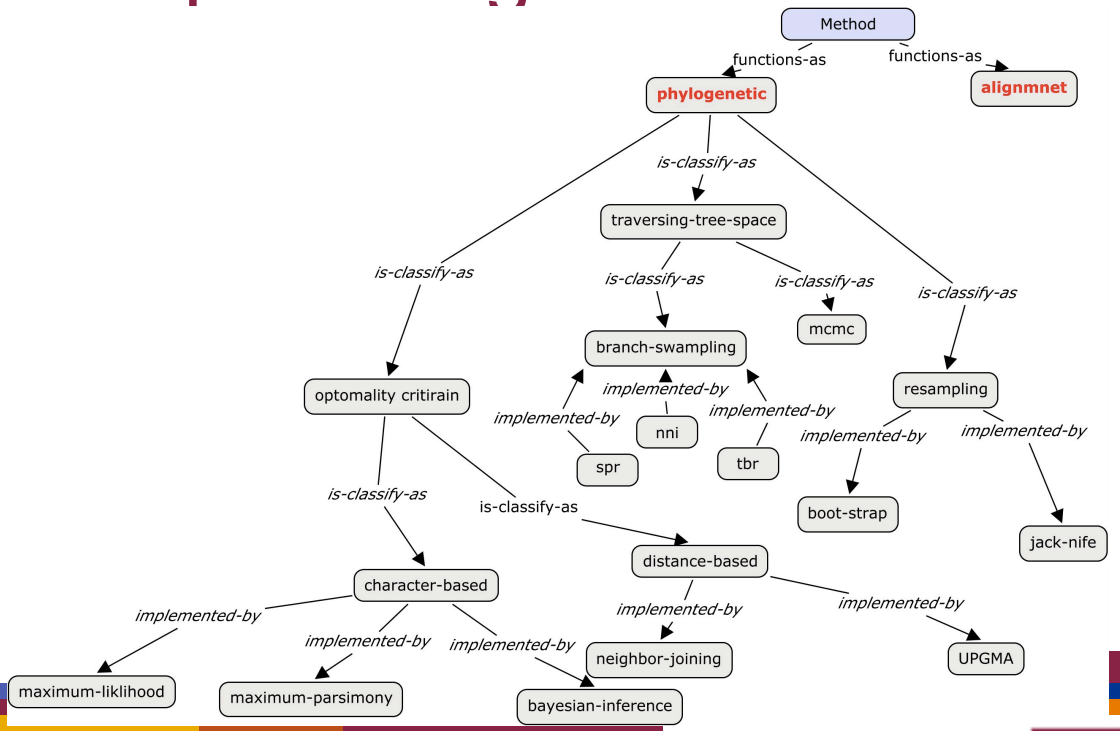


From CDAO to MIAPA

- Minimal Reporting Standards
- Community agreed-upon regularized set of the available metadata ('data about the data') pertaining to an experiment, making explicit both the biological and methodological contexts
- MIAME checklist required by journals and funding agencies for micro-array experiments
- MIBBI: project (https://biosharing.org/) maintains a web-based, freely accessible resource for checklist projects
- **Minimal Information About a Phylogenetic Analysis (MIAPA)**:
  - Seminal development path 2006 (Leebens-Mack et al.)
  - TDWG
  - Identify reporting standards to enable data reuse and experiment replication
  - 2013 Hackathon annotation exercise

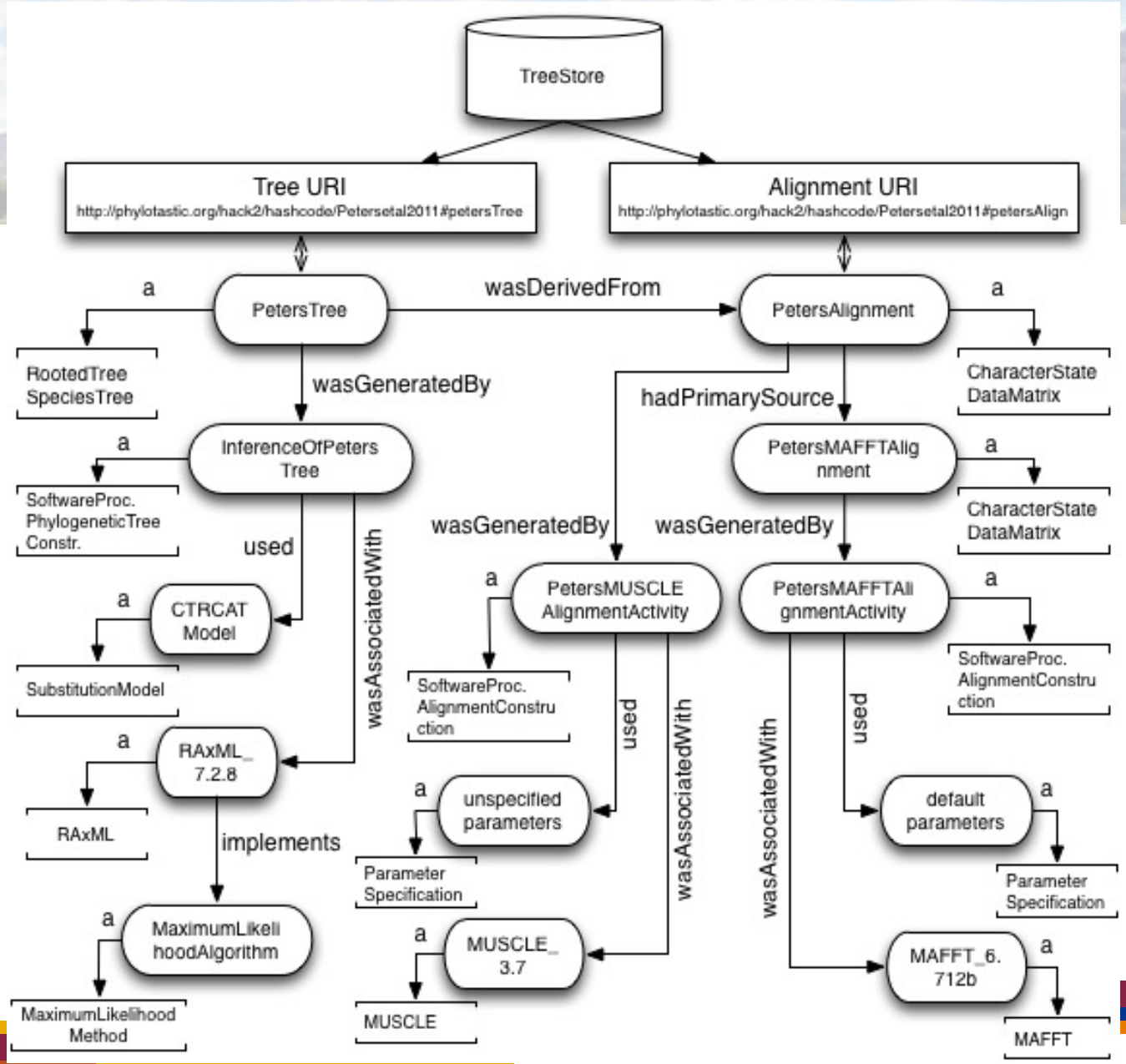
From CDAO to MIAPA

- CDAO as a data description standard
- Formalizing description of generation processes



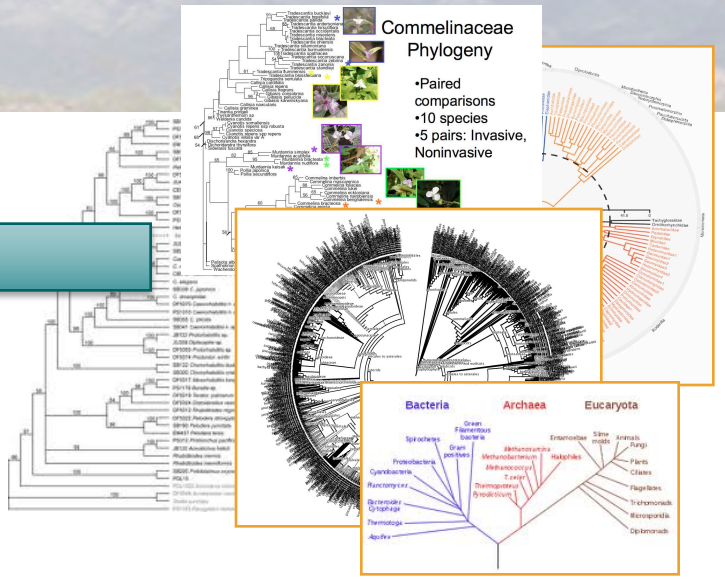
MIAPA Ontology

- OWL
- Imports several ontologies
  - CDAO
  - BIBO
  - PROV
  - IAO
- https://github.com/miapa/miapa/blob/master/ontology/miapa.owl



KLAP Laboratory

Phylotastic: the problem



Researchers, educators, & resource-providers who could use a good species tree

Tree-of-Life knowledge

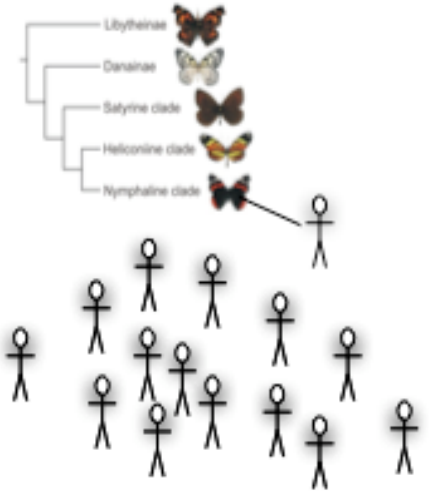
- >10^4 published trees
- mostly locked in pics (< 5% archived)
- → Hard to discover & access
- Inconsistent encoding
- Incomplete annotation
- Name-matching problematic
- → Hard to use

KLAP Laboratory

NM STATE

SHARING & RE-USE OF TREES

Producers

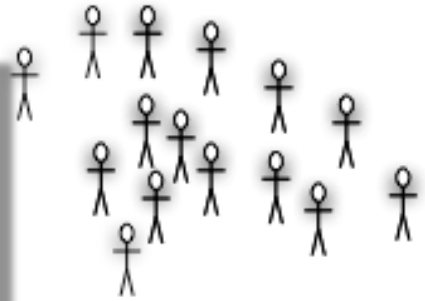


tree image from nymphalidae.net

?

- ✓ analyzed data in public archives
- ✓ studied data sharing policies
- ✓ examined phylogen* articles
- ✓ examined Evolution and Am J. Bot
- ✓ interviewed users
- ✓ studied examples in more depth

Consumers (re-users)



EXAMPLES OF RE-USE OF TREES

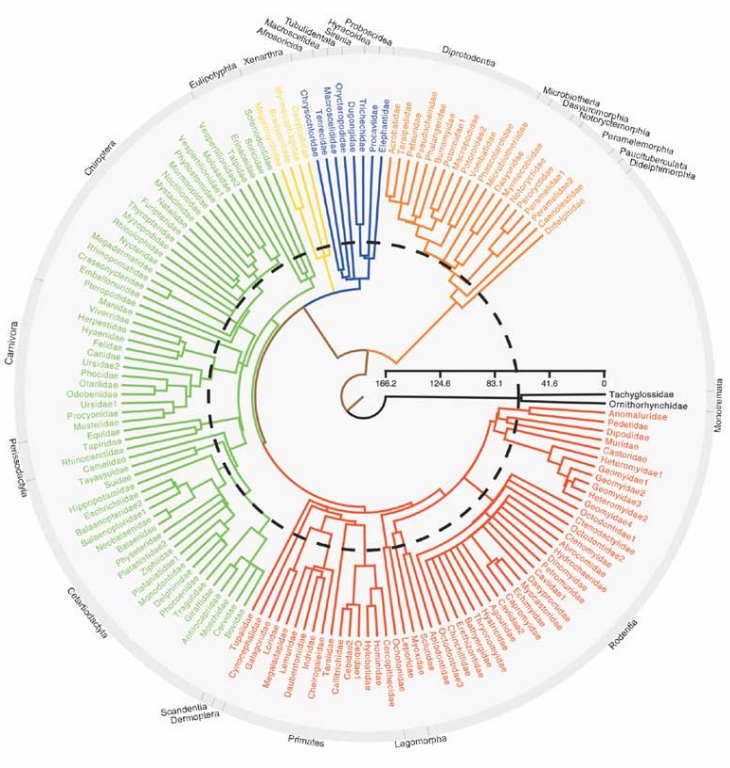
Input tree	Research problem	Reference
APG tree	niche-diversity correlations	Burns & Strauss, 2011
APG tree	spatial distribution of wood traits	Zhang, et al., 2011
APG tree	spatial patterns of diversity	Morlon, et al., 2011
APG tree; Davies, et al tree	leaf veins & functions	Walls, 2011
Bininda-Emonds mammal tree	allometry of milk properties	Riek, 2011
APG tree	patch diversity	Duarte, 2010

Some big trees*

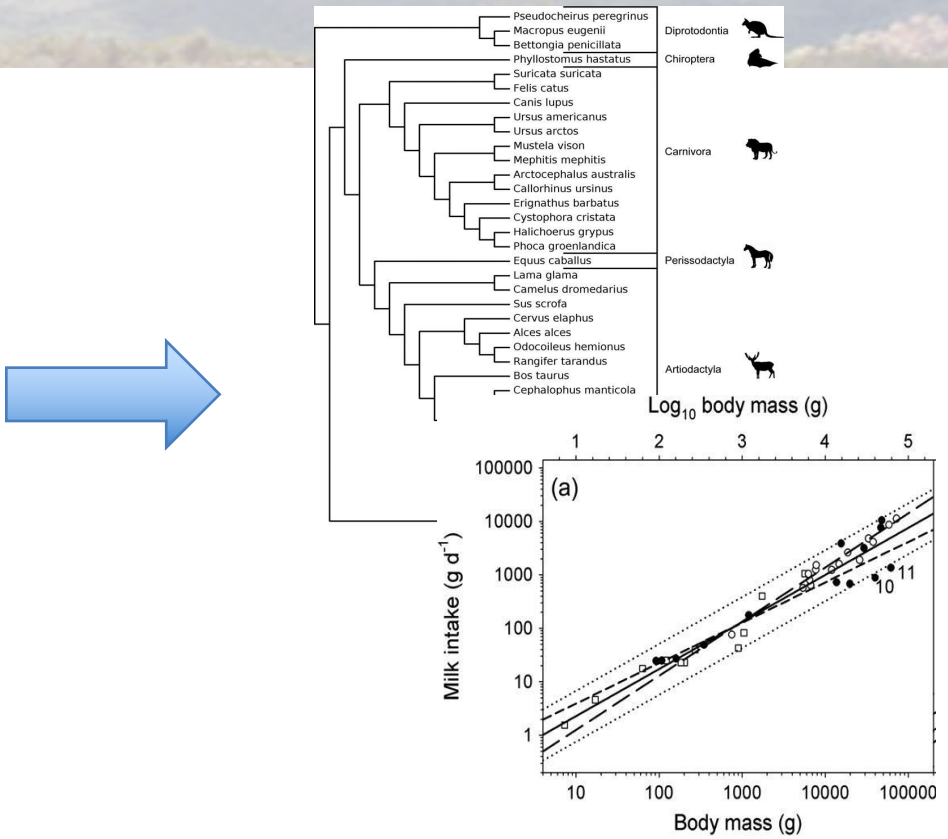
- 4,500 mammals (Bininda-Emonds et al. 2007)
- 55,473 angiosperms (Smith et al. 2011)
- 1,827 angiosperm taxa in APG tree (Phylomatic)
- 800 fish families (Westneat & Lundberg unpub.)
- 16,000 taxa in ToLWeb (Maddison, et al)
- 73,060 eukaryotes (Goloboff et al. 2009)
- 400,000 prokaryotic 16S rDNAs (McDonald, et al 2012)

* Proper phylogenies as well as phylogeny-based taxonomic hierarchies

USE CASE: EVOLUTION OF LACTATION



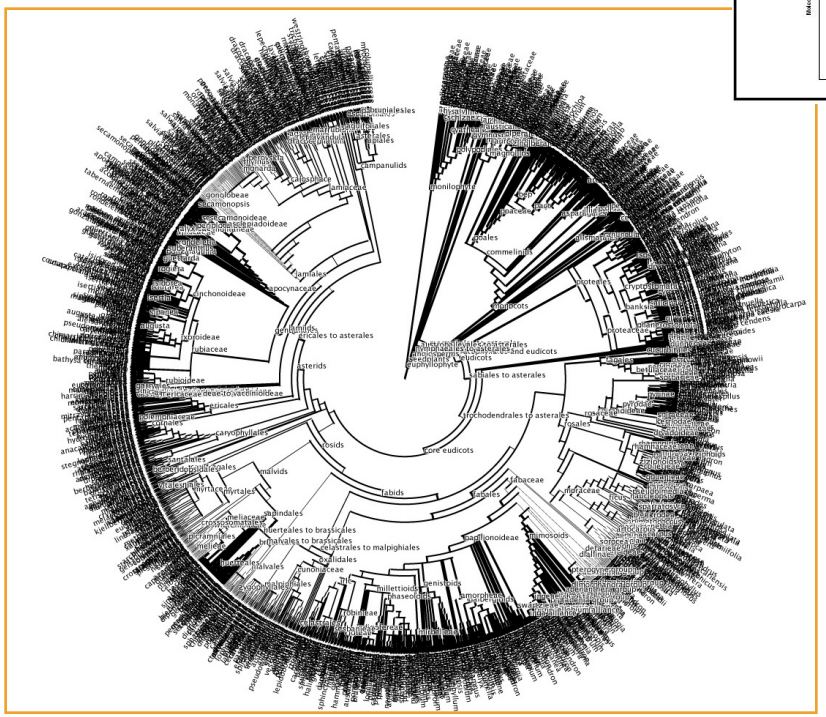
Low-resolution view of mammal supertree
Bininda-Emonds et a. 2007



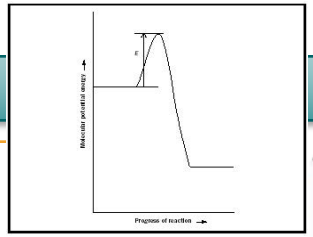
Pruned Tree and Allometric scaling using
Phylogenetic Independent Contrasts (Riek, 2011)

Overcoming barriers

Actual state of ToL



APG framework tree



What users need

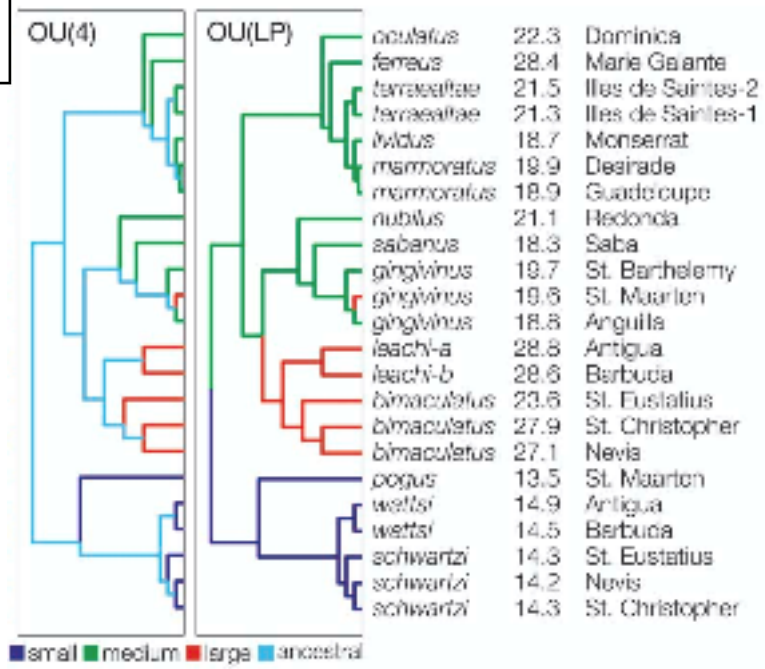
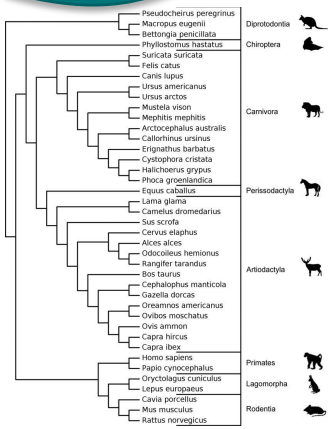


Fig. 5 from Butler & King, 2004

Inconsistent Names as a barrier to integration

Riek, 2011 (Mammalian Biology 76(1):3-11)



40 species

auto-extract species names from text

<1 minute

Phylotastic

36 species + 2 extras*

Copy & paste species named in Table 1

5 minutes

Phylotastic

33 species

Manually key in species list from tree image

15 minutes

Phylotastic

36 species

Manually reconcile names with names in source tree

Hours or days

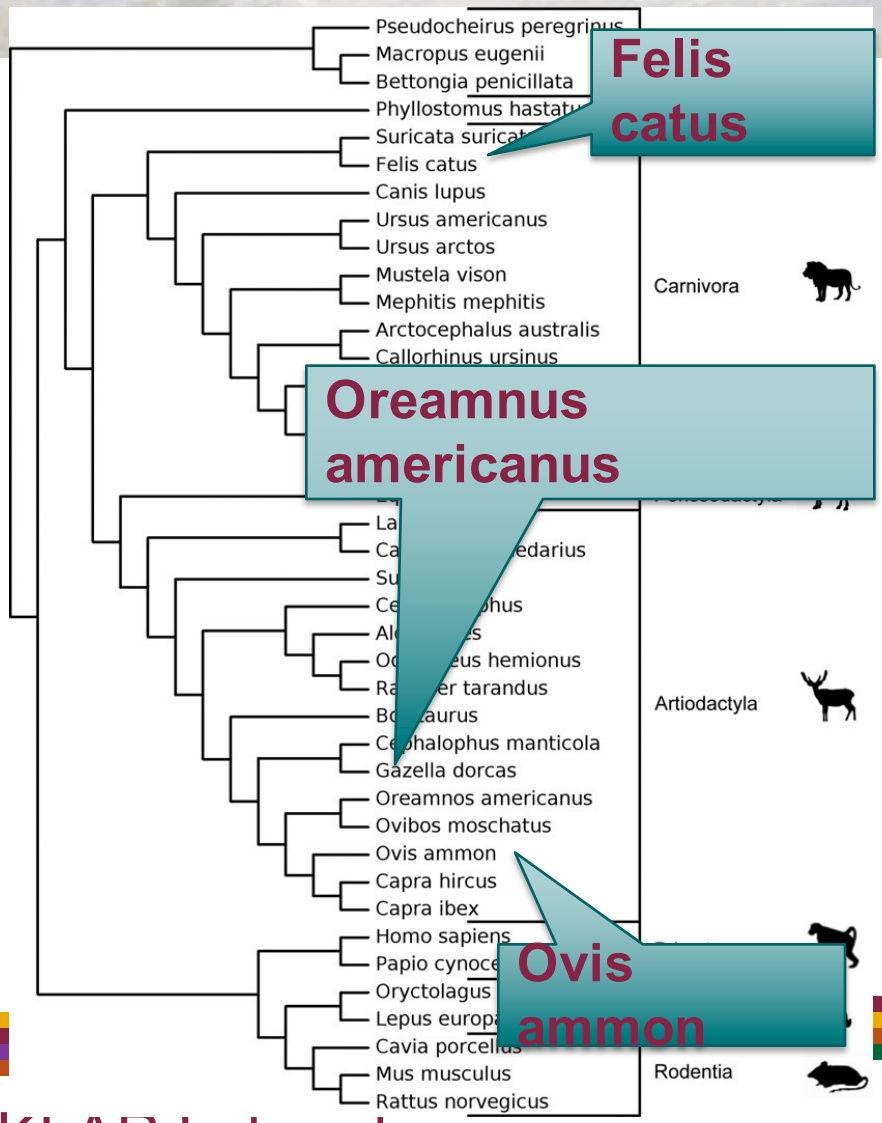
Phylotastic

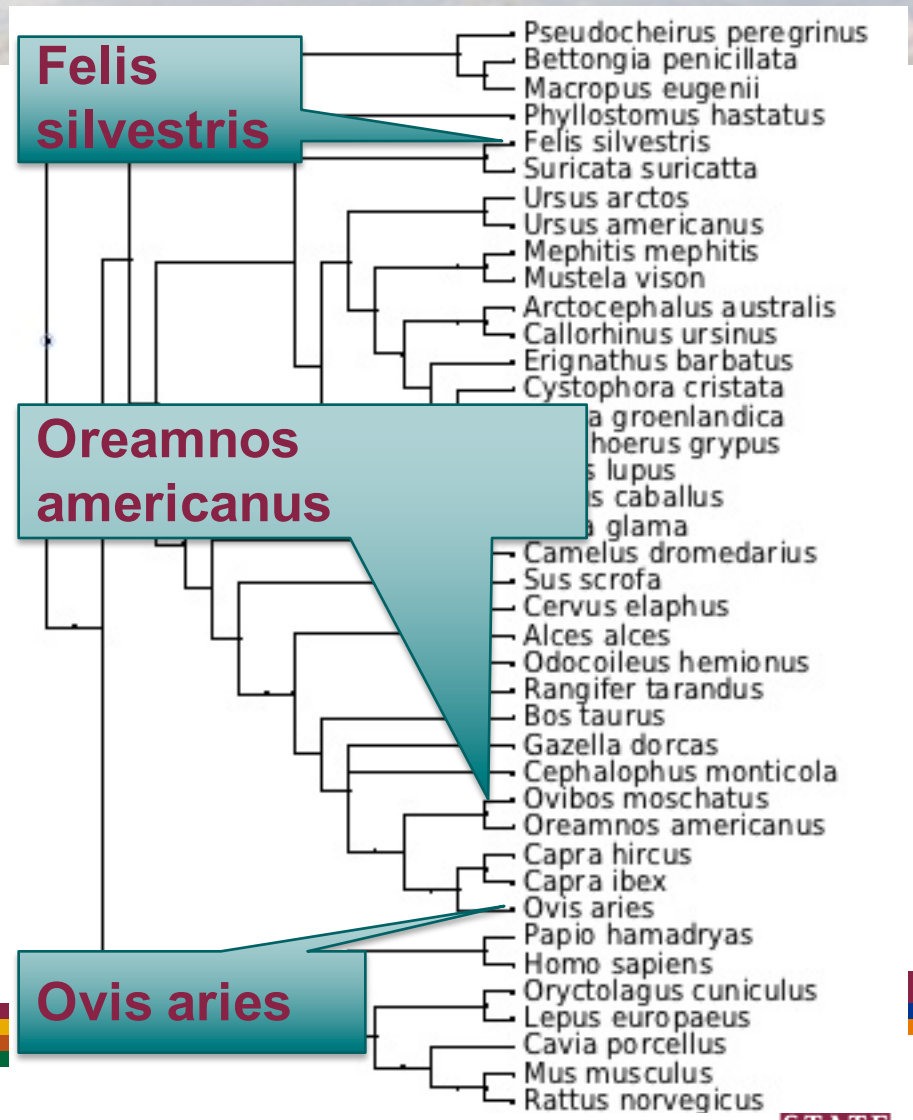
40 species

KLAP Laboratory

* named in text but not used in phylogenetic analysis

~ 5% to 15% non-standard names





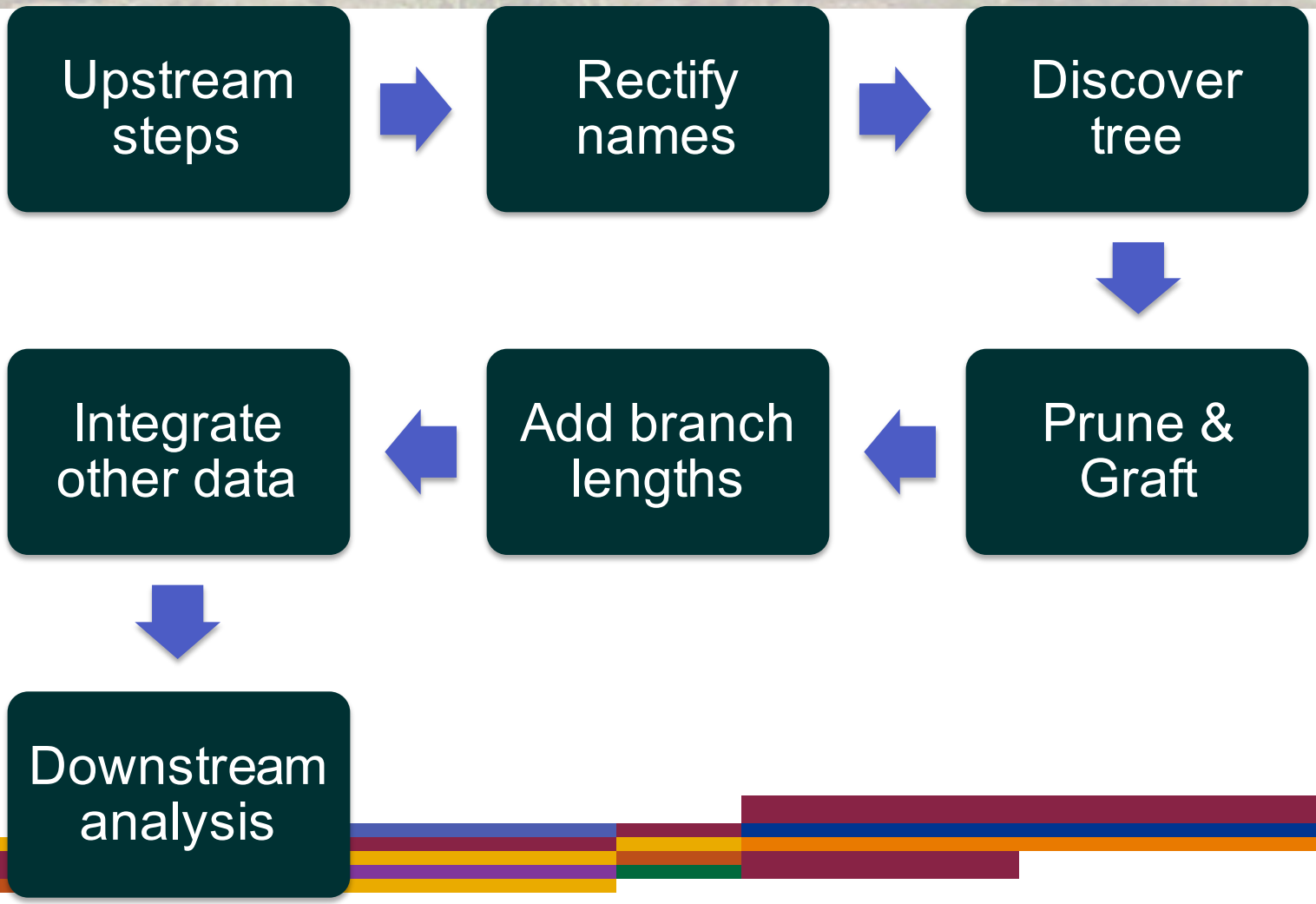
KLAP Laboratory

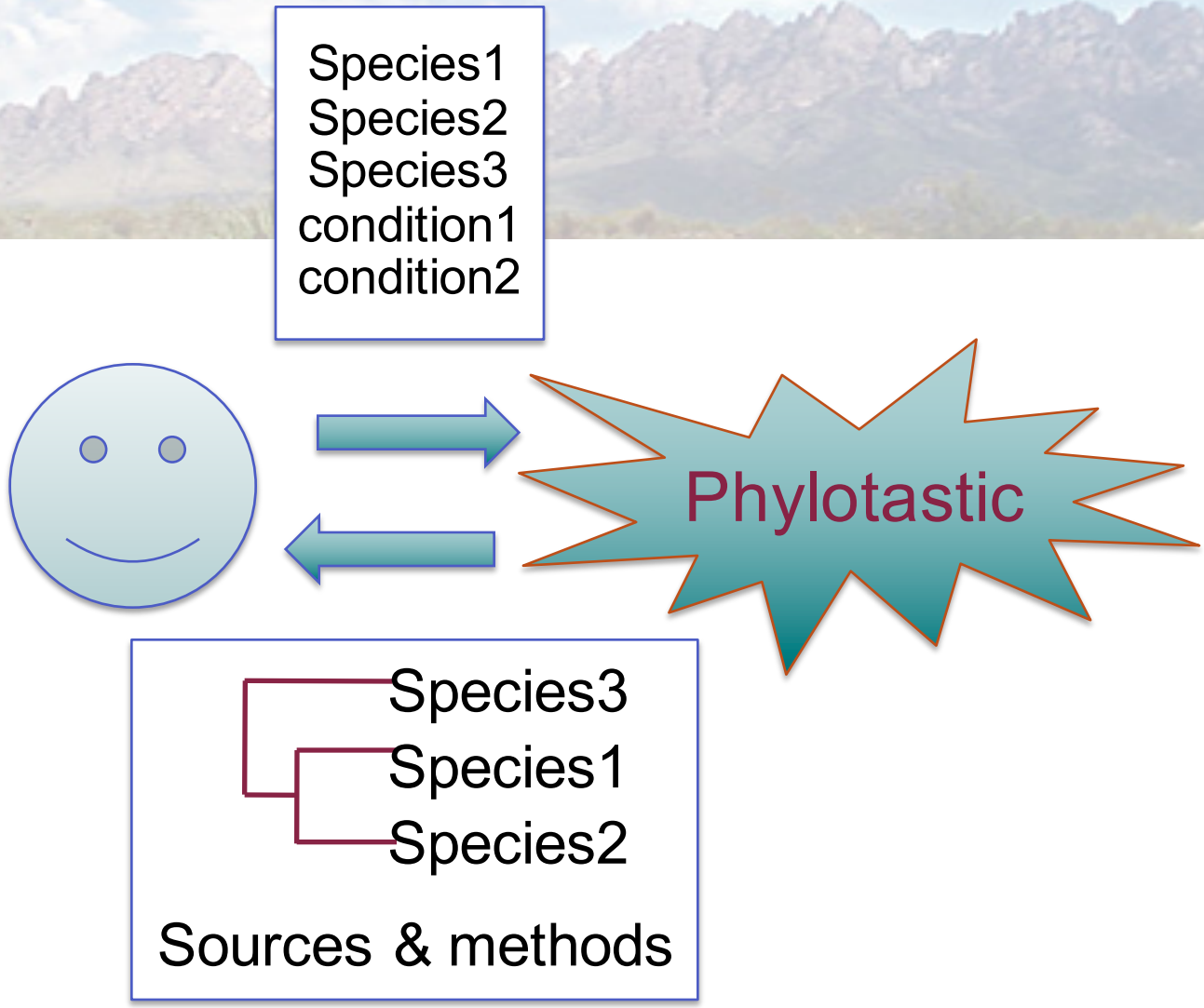
Figure from Riek, 2011

Phylotastic tree after name correct

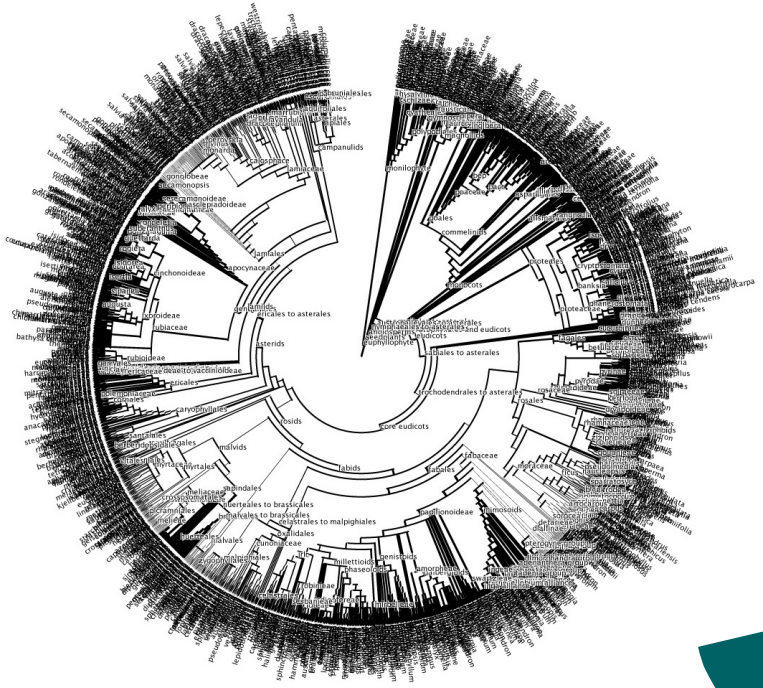
STATE

Typical user's workflow

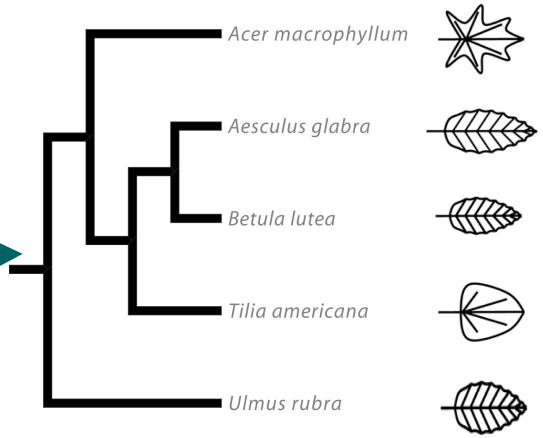


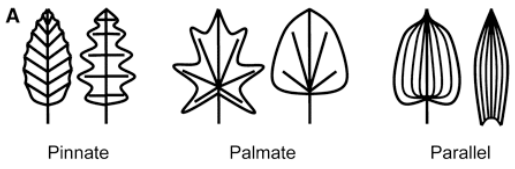

Species1
Species2
Species3
condition1
condition2
Phylotastic
Species3
Species1
Species2
Sources & methods

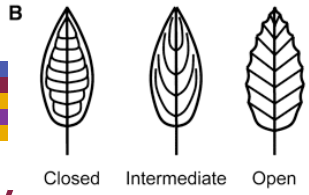
Tree pruning to accommodate comparative analysis of trait data (using PhyloMatic)



Acer macrophyllum
Betula lutea
Aesculus glabra
Tilia americana
Ulmus rubra







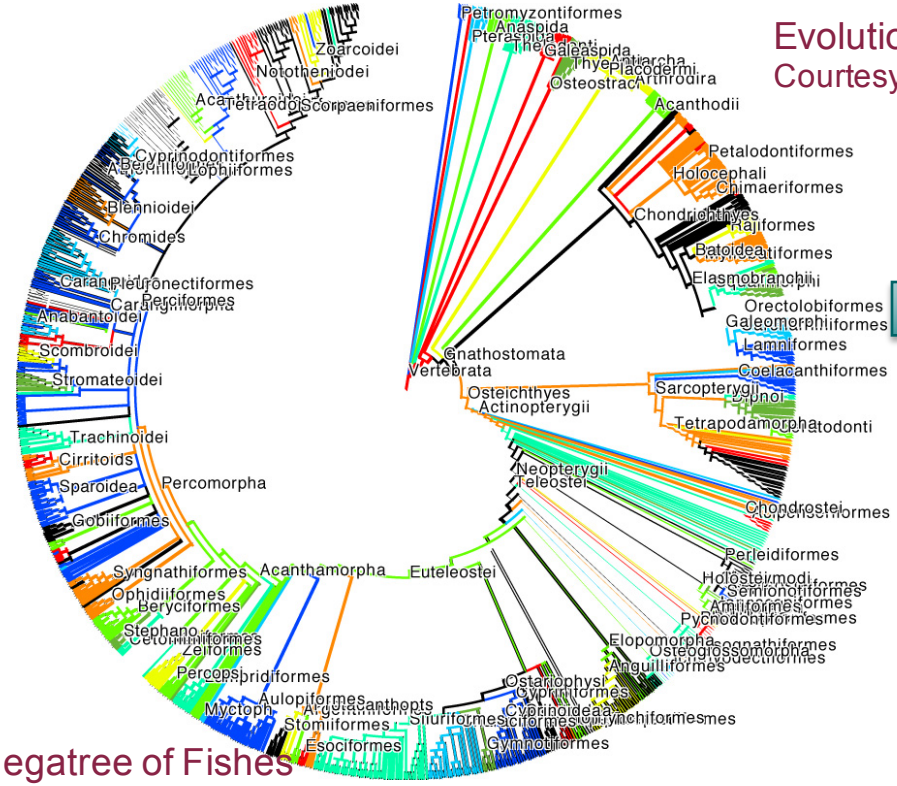
Leaf patterns image from Walls RL: Angiosperm leaf vein patterns are linked to leaf functions in a global-scale data set. American Journal of Botany 2011, 98(2):244-253.

KLAP Laboratory

NM STATE

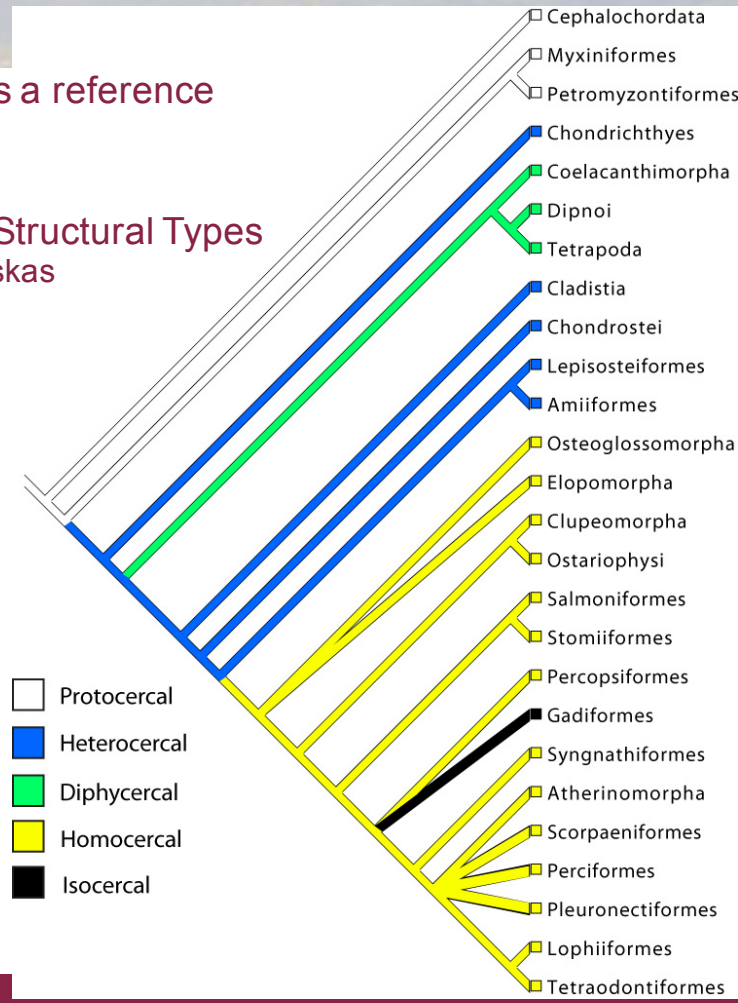
Teaching morphological evolution with reference to the community-consensus megatree of fishes

Righthand tree assembled by hand in Mesquite using the left as a reference because no easy pruning and grafting automation exists.



Megatree of Fishes
Courtesy of Brian Sidlauskas

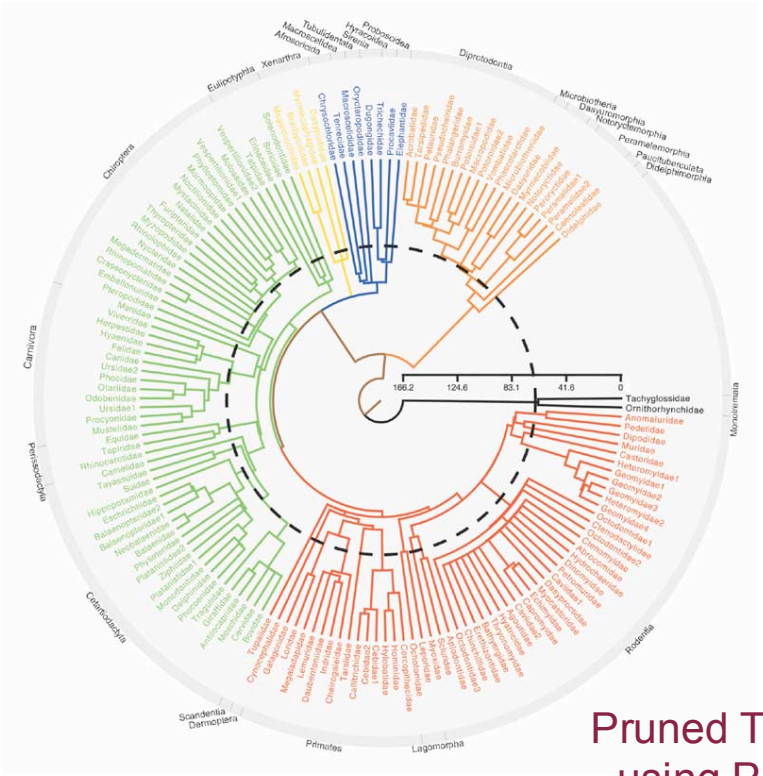
Evolution of Fish Tail Structural Types
Courtesy of Mark Westneat

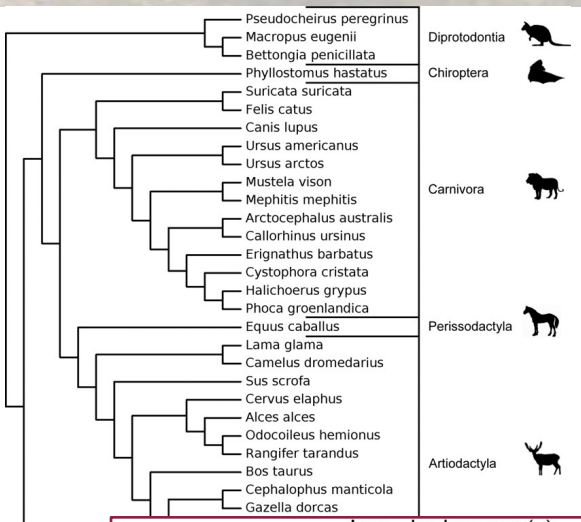


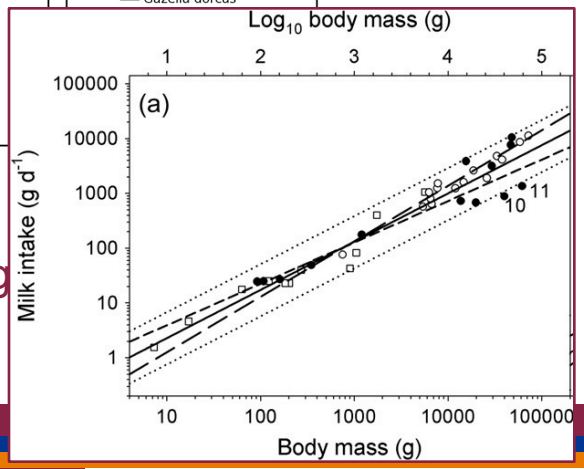
KLAP Laboratory

NM STATE

Allometric analysis of the evolution of lactation using the supertree of mammals







Pruned Tree and Allometric scaling using Phylogenetic Independent Contrasts (Riek, 2011)

Supertree of Mammals
Bininda-Emonds et a. 2007

KLAP Laboratory

HIP Hackathons

HIP Leadership Team

Participants

National Evolutionary Synthesis Center

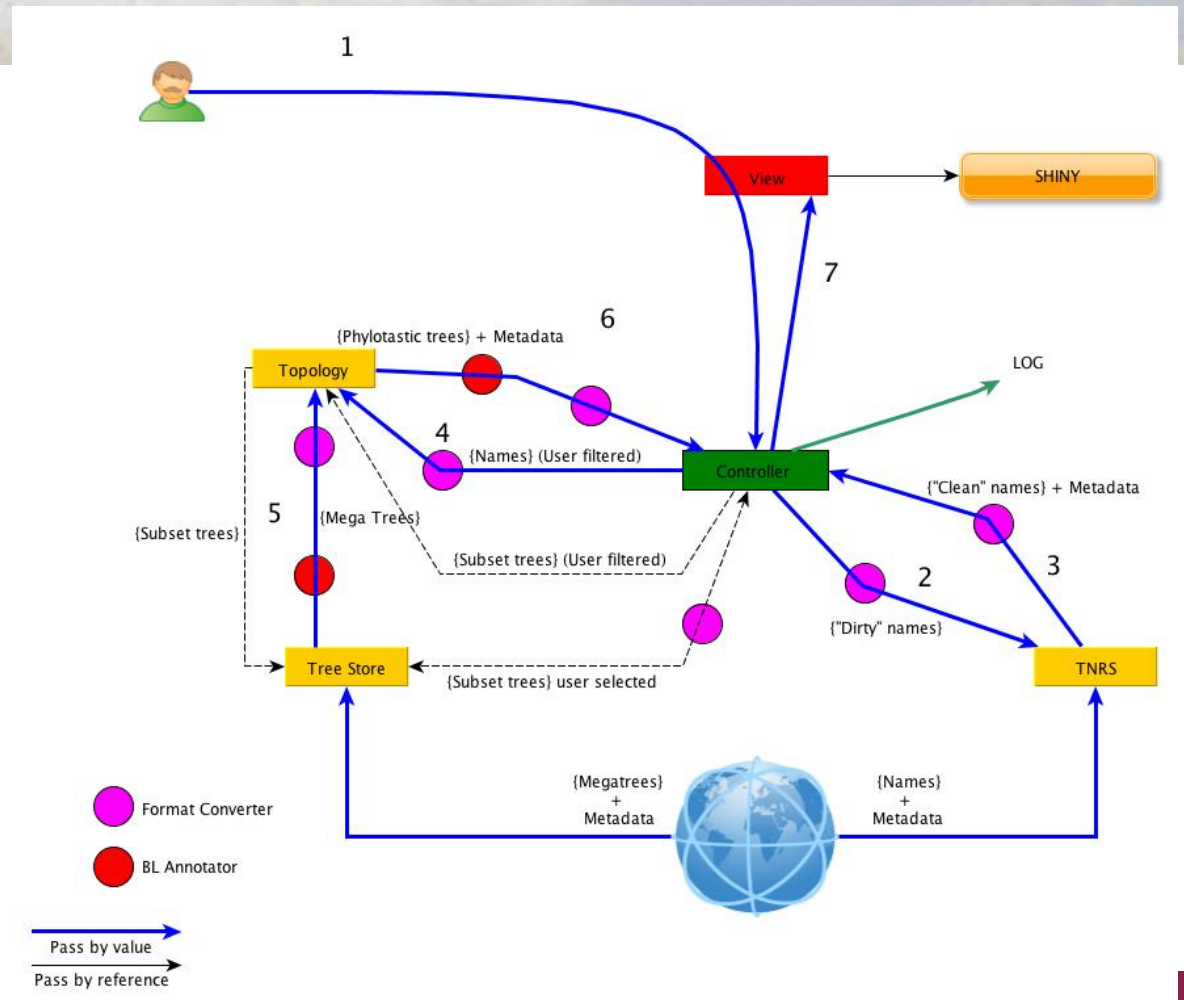
NESCent

iPlant Collaborative™

The EOL Biodiversity Synthesis Group

NM STATE

Phylotastic

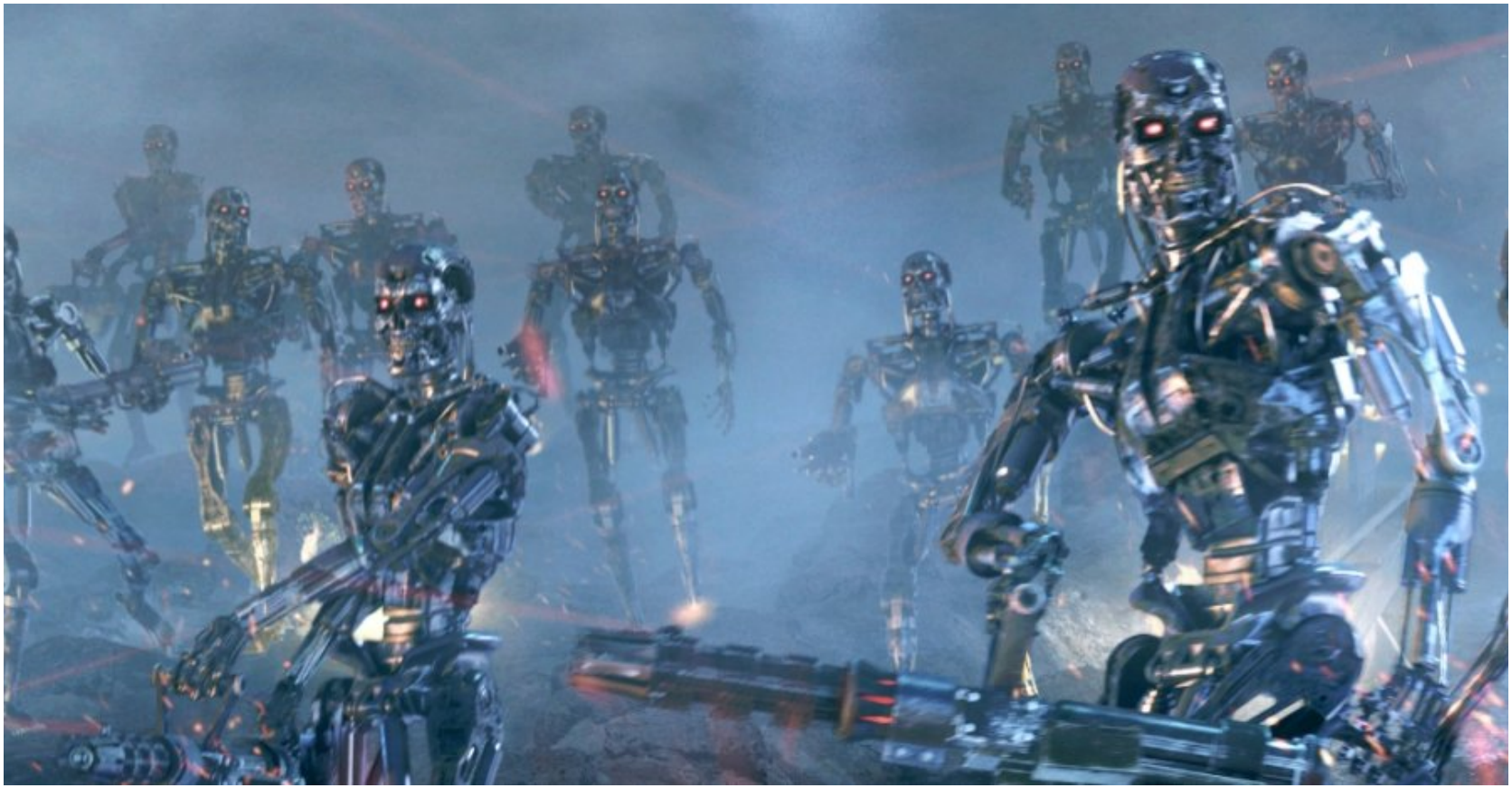


Many online data repositories

Challenges

- Fragile: web services are often unstable
- Data gets bigger and bigger
- Many concepts not yet in ontologies
- Many data still “locked in” in publications

The Future



standards are voluntary

City (greatest to least populated)	Hose connection		Pumper Connection	
	Diameter (inch) NS = 2 ½	Thread (threads/inch) NST= 7 ½	Diameter (inch) NS = 4 ½	Thread (threads/inch) NST = 4
New York City, NY	2 ½	8	NS	NST
Los Angeles, CA	NS	NST	4	4
Chicago, IL	-	-	4 ½	6
Houston, TX	NS	NST	NS	NST
Philadelphia, PA	NS	NST	NS	NST
Phoenix, AZ	2 ½	6	4	6
San Diego, CA	NS	NST	4	4
Dallas, TX	NS	NST	4	4
San Antonio, TX	NS	NST	4	4
Detroit, MI	NS	NST	2 ½	6

Seck & Evans, 2004. NISTIR 7158

The Oakland Firestorm of 1991

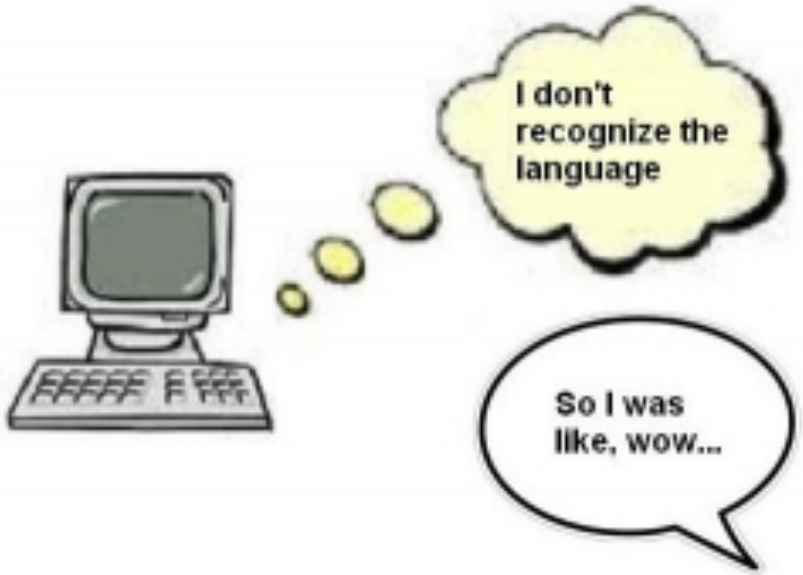
Image: wikipedia

Oakland conforms to standard

KLAP Laboratory

Interpreting locked in knowledge

- Text and images meant for humans are being processed by machines. Examples:
  - Taxon name mining (BHL)
  - Tree figure processing
  - Automated annotation



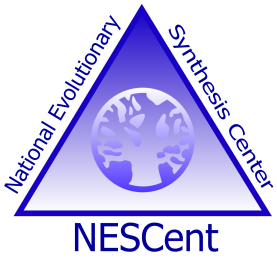
Summary

- Phyloinformatics is moving from **closed** to **open** to **linked** data
- Concepts and syntax are increasingly **formalized** and **machine readable**
- **Automated** queries across **integrated** resources will enable **synthetic research**
- Still lots to do to **deploy these technologies** and **unlock legacy data**

Acknowledgements

Thank you for your attention!
Also, many thanks to:

The EvoInfo group
The EvoIO group
J. Leebens-Mack
Hilmar Lapp
Arlin Stoltzfus
Rutger Vos
Brian O'Meara



Thank You! Questions?