Phyloinformatics VoCamp

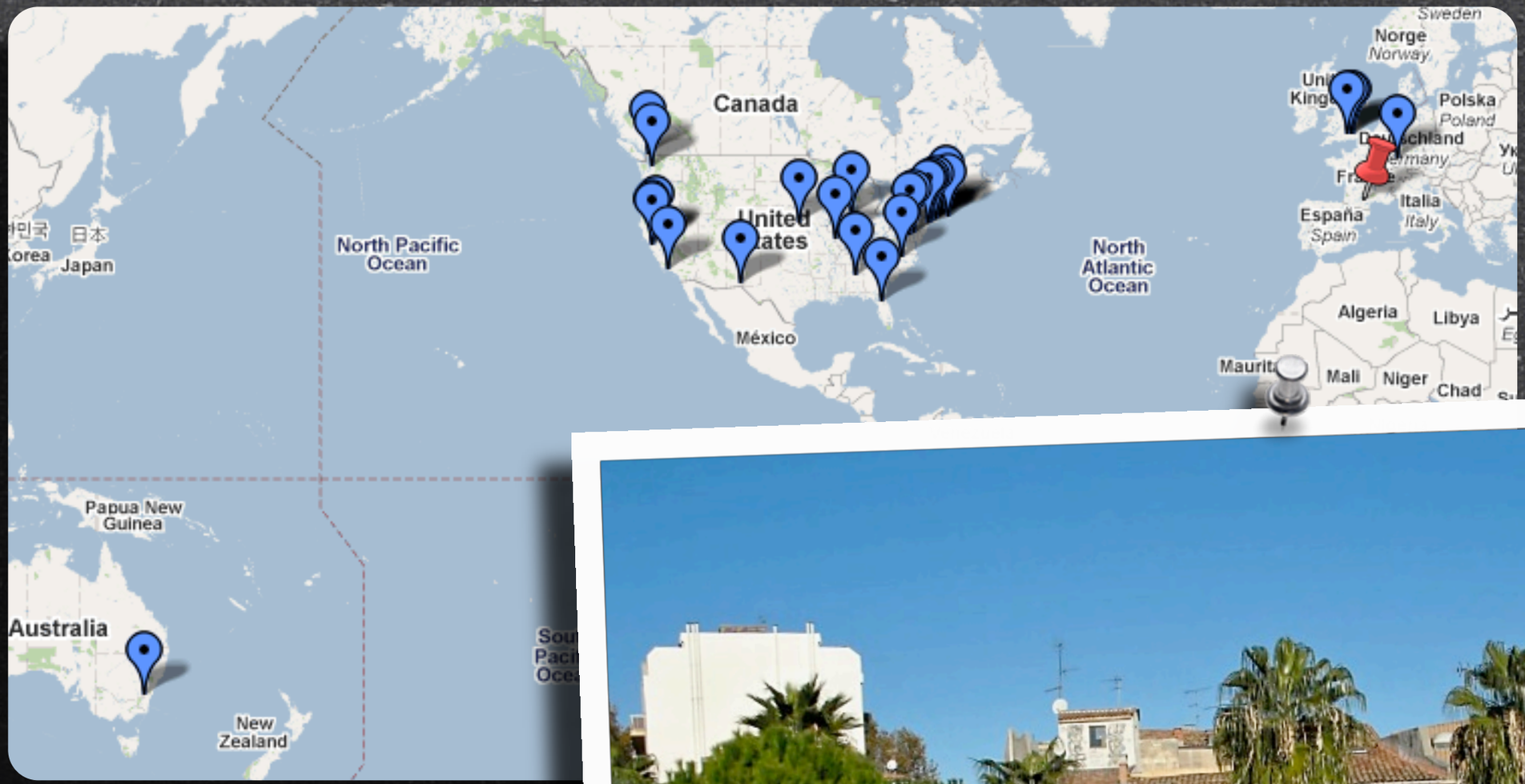
Hilmar Lapp
National Evolutionary Synthesis Center (NESCent)

TDWG Conference 2009, Montpellier

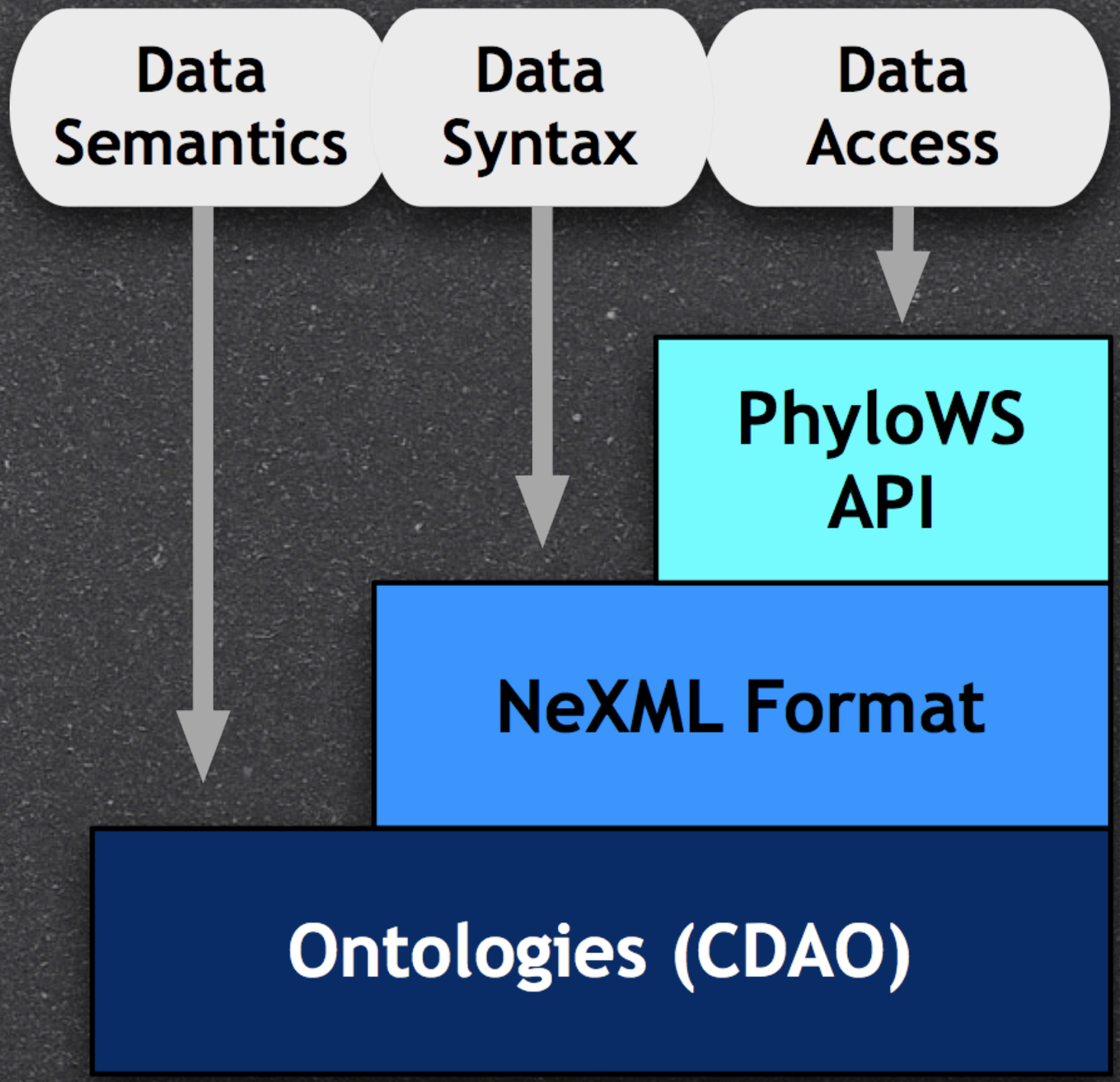
http://evoio.org/wiki/VoCamp1

Acknowledgments

- Organizing Committee:
  - Arlin Stoltzfus (Chair)
  - Nico Cellinese
  - Karen Cranston
  - Hilmar Lapp
  - Sheldon McKay
  - Enrico Pontelli
- 30 Participants from 7 countries, 28 institutions, and >40 projects
- Sponsors:
  - NESCent
  - TDWG
  - LIRMM
- The VoCamp is a Phylogenetics Standards Interest Group activity.


Canada
United States
México
North Pacific Ocean
North Atlantic Ocean
Japan
Papua New Guinea
Australia
New Zealand
Norge
Norway
Polska
Poland
España
Spain
Italia
Italy
Algeria
Libya
Mali
Niger
Chad

Evolutionary Informatics Working Group (2006-2009)



Evolutionary Database Interoperability Hackathon

NESCent, March 2009

PhyloWidget

iPlant
Empowering a new plant biology

eol
Encyclopedia of Life

MESQUITE
W.P. Maddison
D.R. Maddison

PHENOSCAPE

GMOD

MorphoBank
Homology of phenotypes over the web

Biodiversity Collections Index

Morphbank

PANDIT

The Paleobiology Database

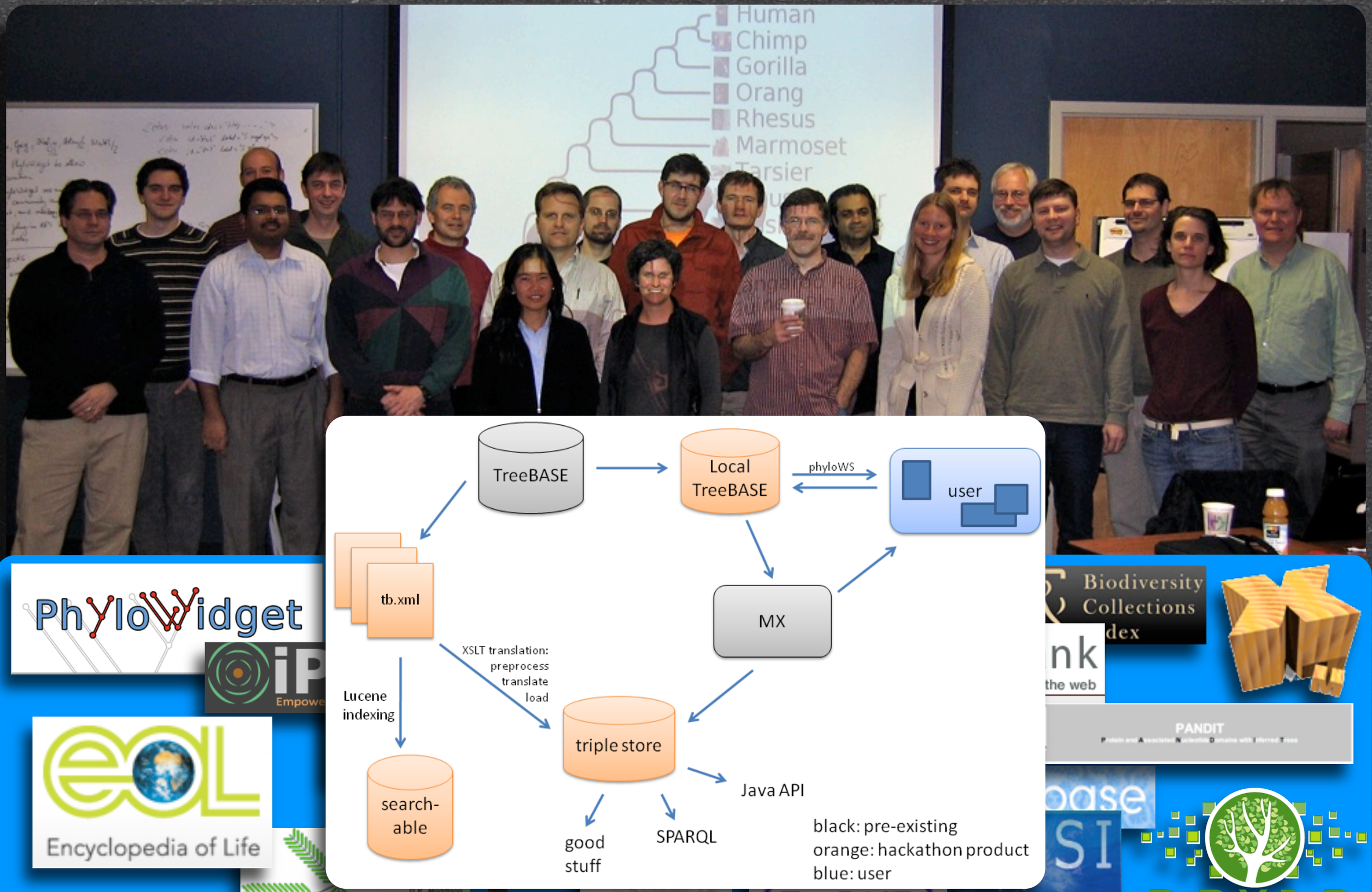
PESI
EU-NOMEN

DRYAD

TreeBASE

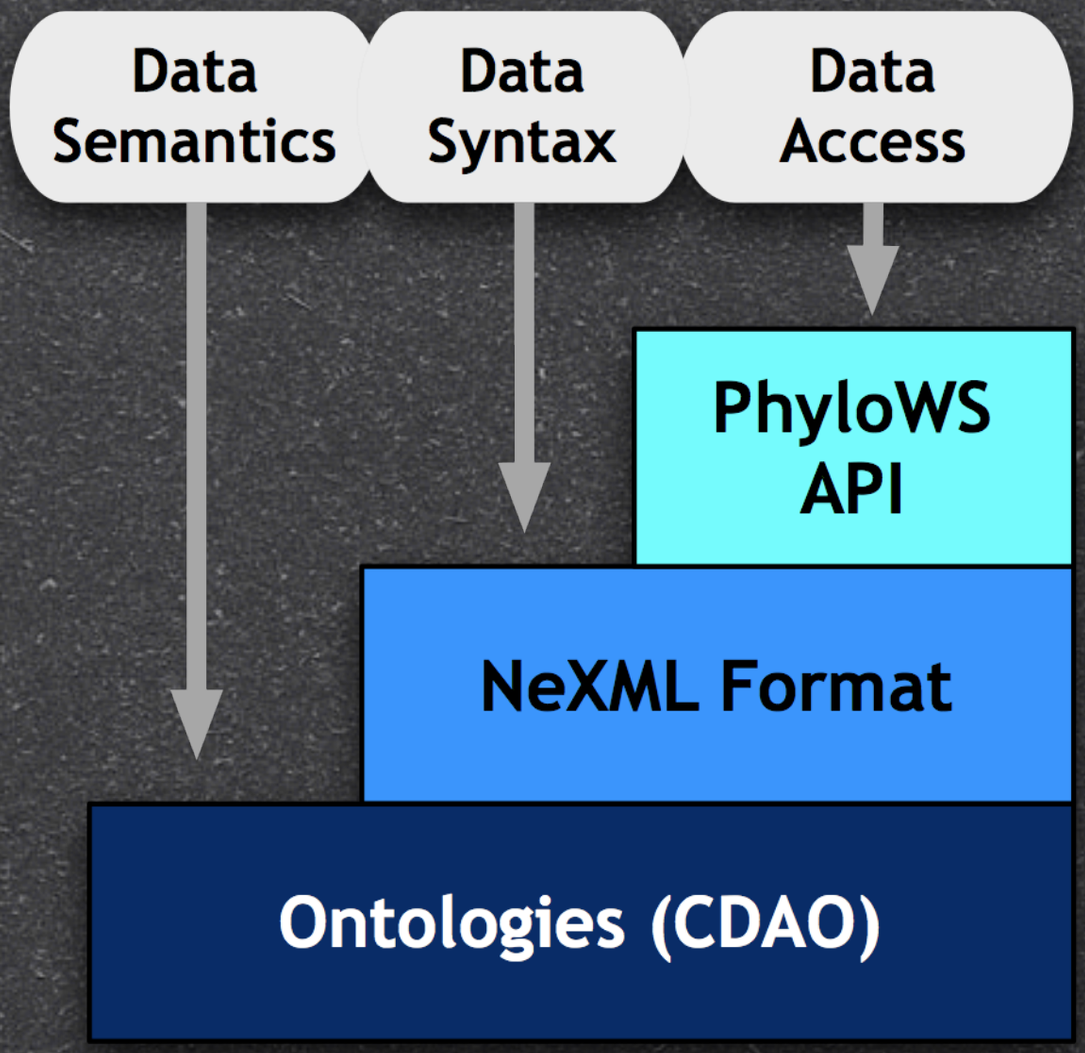
PB

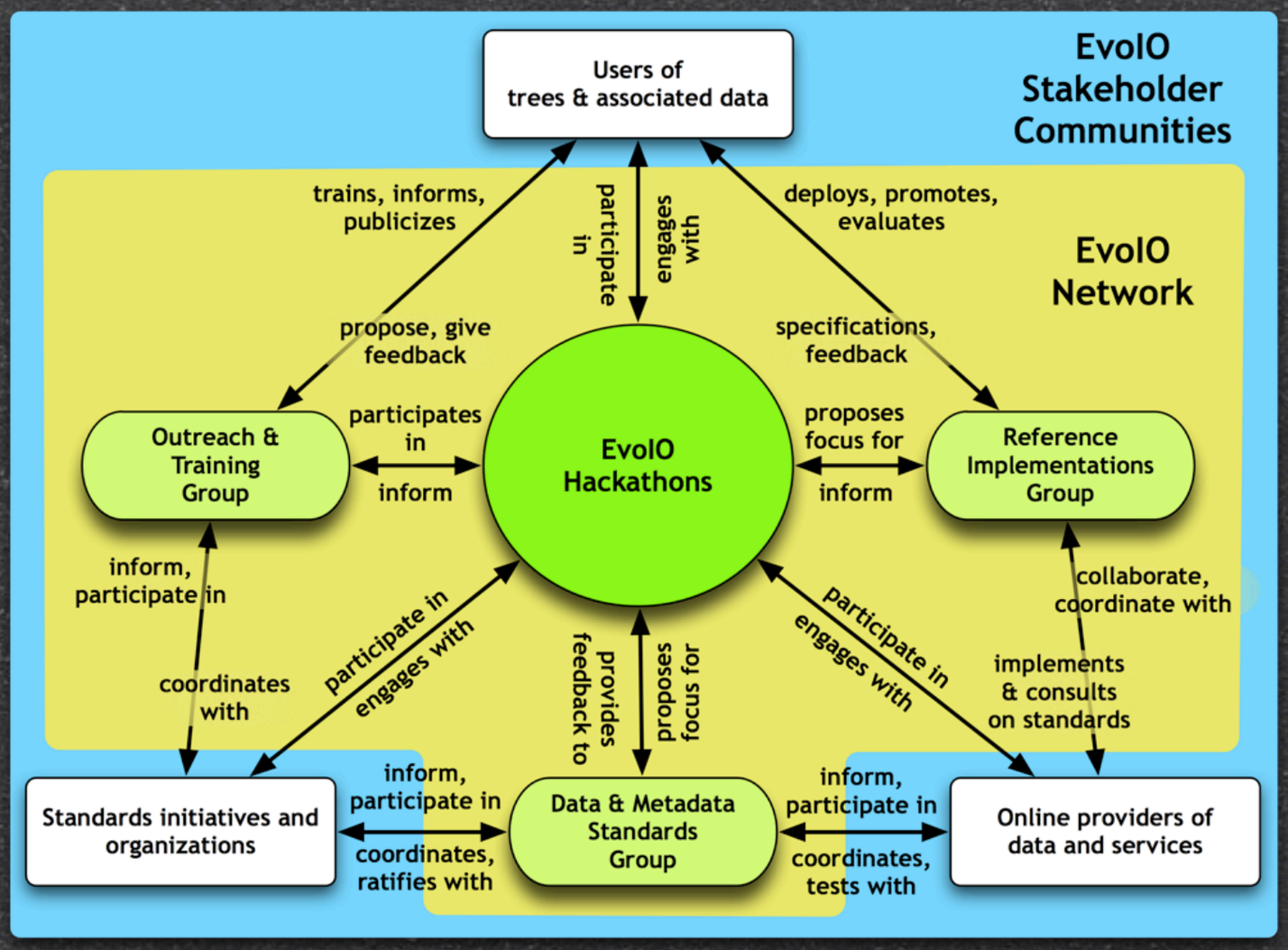
pPOD


Human
Chimp
Gorilla
Orang
Rhesus
Marmoset
Tarsier
TreeBASE
Local TreeBASE
phyloWS
user
tb.xml
MX
XSLT translation:
preprocess
translate
load
Lucene
indexing
triple store
search-
able
Java API
good
stuff
SPARQL
black: pre-existing
orange: hackathon product
blue: user

NSF INTEROP Proposal*:

A network for enabling community-driven standards to link evolution into the global web of data (EvoIO)





(*) submitted July 2009, currently under review

Motivation

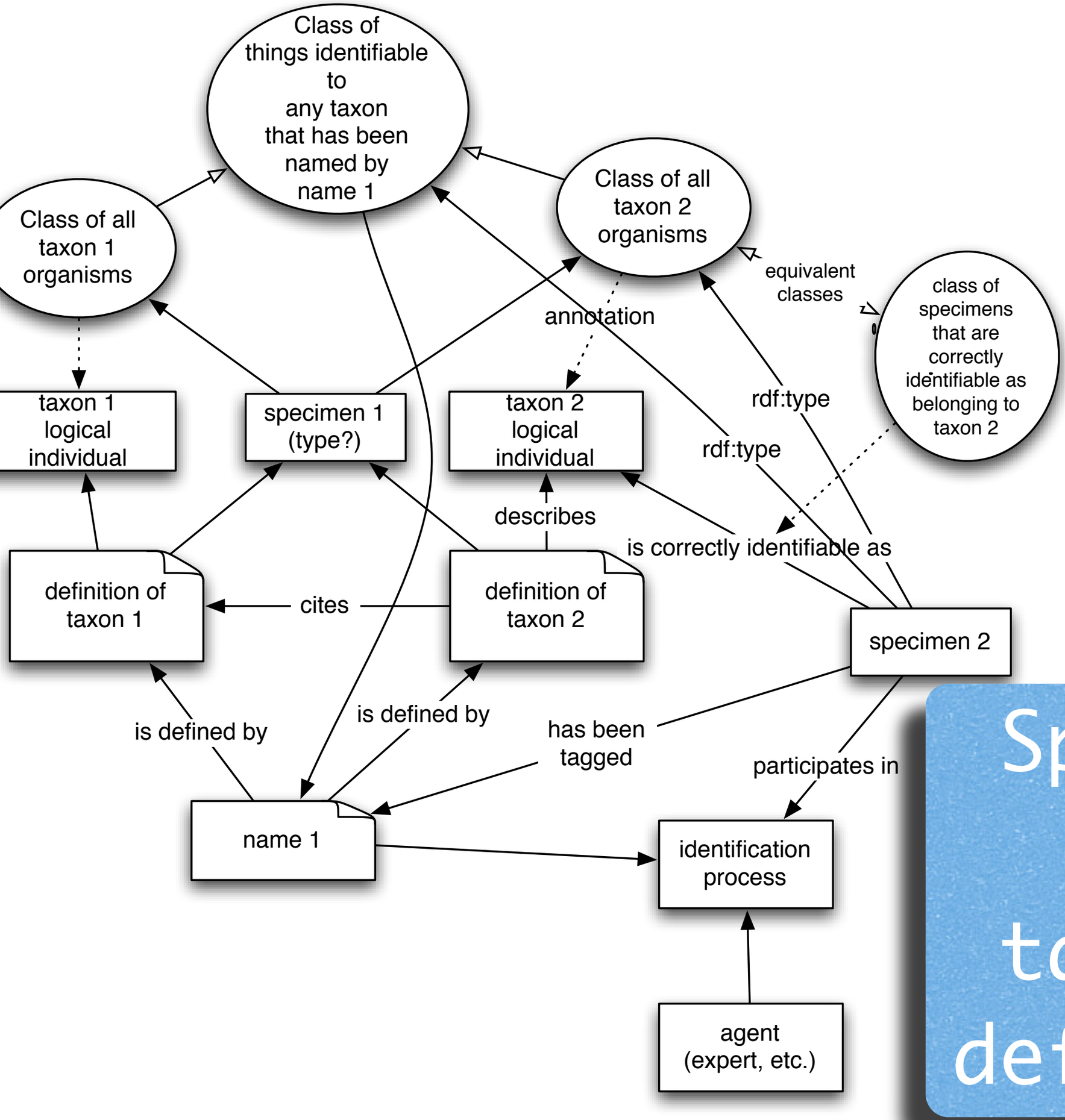
- Integrating data with the evolution of historical and extant biodiversity is a grand challenge.
- Data interoperability rests on formalized, shared vocabularies.
- Such ontologies have emerged as a key gap, and developing them is a community-based project.
- Hackathons have been highly successful at collaborative development of open & standards-supporting software.

Phyloinformatics VoCamp: The event

- Hands-on, face-to-face collaboration
- Participants: 1/2 invited, 1/2 through open CfP
- Interdisciplinary composition fosters cross-pollination, learning
- 4 days total, 2 segments
- Self-organized into 6 subgroups

Publishing Taxonomies Subgroup

- Goal: Determine RDF modeling rules that best support reuse and inference over biological classifications published as linked-data RDF.
- Outcomes:
  - Illuminated modeling taxa as individuals versus as classes.
  - Explored the problems with expressing and inferring monophyly in OWL-DL.
  - Conceptual model of inferring the species of specimens based on taxonomic definitions.



SpeciesID using taxonomic definitions

Taxonomic Reasoning Subgroup

- Goal: Use-case questions requiring inference over phylogenies and related data, and how to answer those with data, ontologies and reasoners.
- Outcomes:
  - Core question: Location -> species present there -> traits of those.
    Most recent ancestor -> all descendants -> traits of those.
  - Requires geo-referenced species observations, trait ontology with subsumption hierarchy, phylogeny.
  - Many lessons. E.g., real data for this are hard to come by. Iterative development rested on interdisciplinary group.

Integrating Ontologies Subgroup

- Goal: Determine best practices for the building, maintenance and integration of ontologies shared across domains.
- Outcomes:
  - Use-case query: In which environments do we find social tuco-tucos? (Integrates DarwinCore records with phylogenetic character state data.)
  - SPARQL query representing the above.
  - Recommendations for ontology development and annotation based on encountered obstacles.

Triplestore Subgroup

- Goal: Geospatial reasoning over GBIF occurrence data integrated with the IUCN Redlist, using RDF and Franz' AllegroGraph.
- Outcomes:
  - Worked out many bug and data loading issues with Franz.
  - Identified RDF parsing and spec issues of the RDF data sources.

Phyloreferencing Subgroup

- Goal: Define syntax, semantics, and query expressions for nodes (and their subtrees) defined by phylogenetic ancestry.
- Outcomes:
  - 3 principal use-case queries
  - Phyloreference expressions for each
  - Vocabulary for query and specifier semantics started (needs definitions)
  - PhyloWS query specification (CQL-based) started

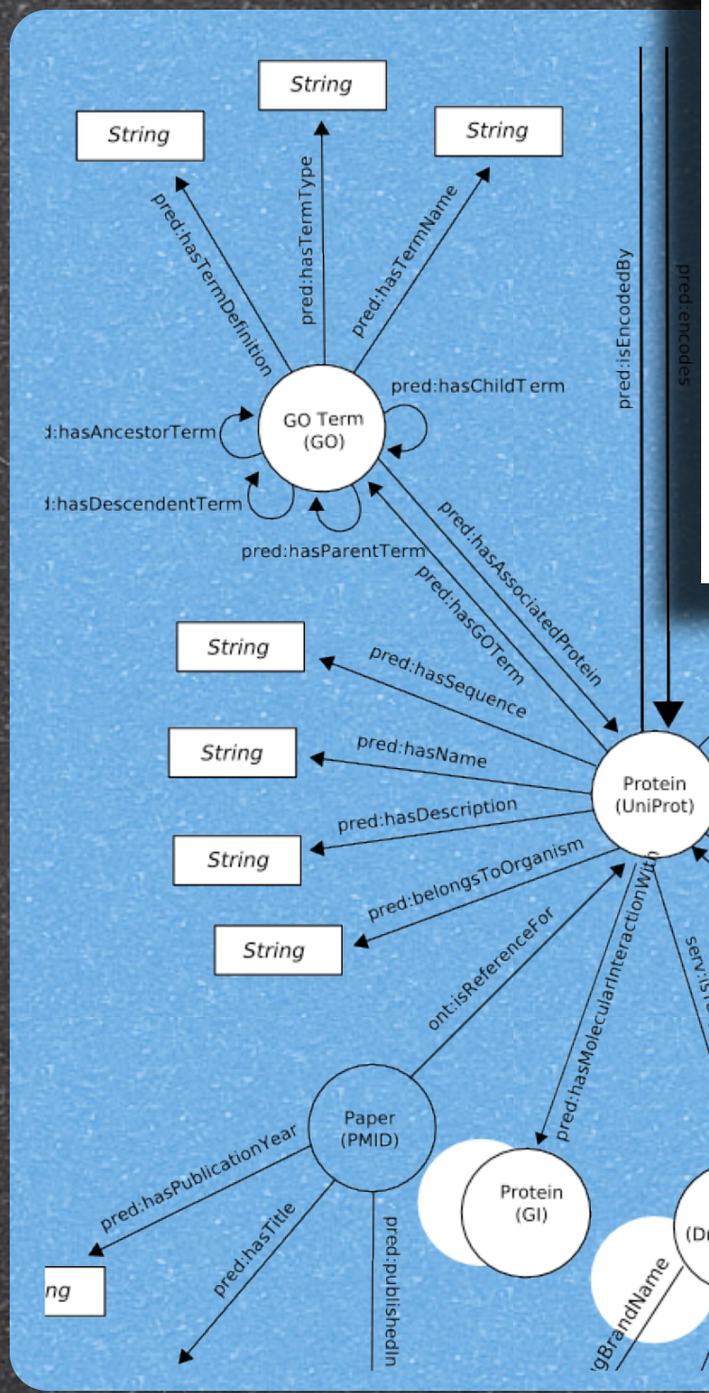
Phyloreferencing: Example

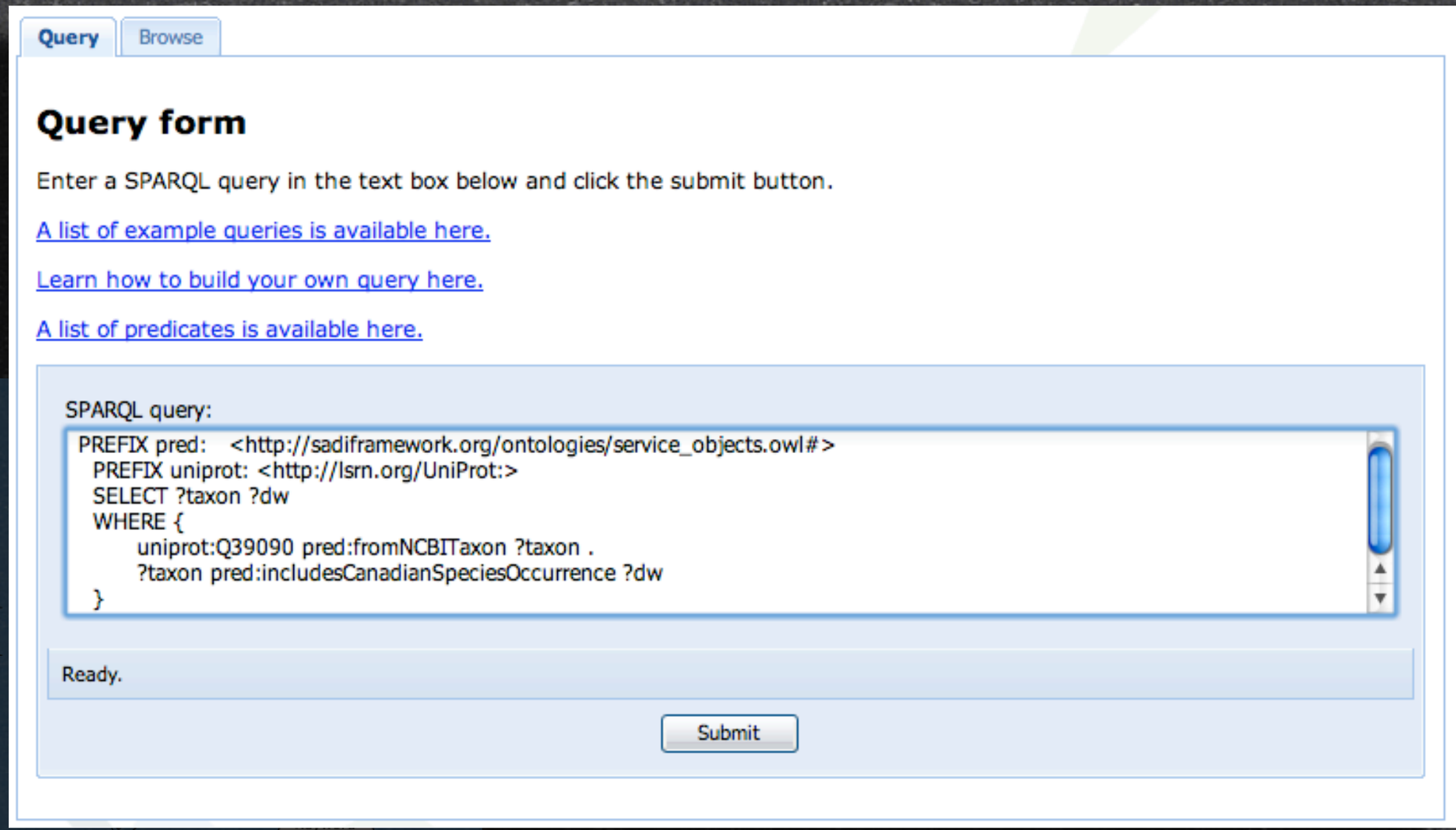
- Give me a subtree for my group of interest. E.g., Magnoliaceae from a tree of all plants.
- Formally: the clade originating with the most recent common ancestor of S_1,..., and S_n, with $n \geq 2$.
- S_n are node specifiers: e.g., taxon name, taxonID, specimen, sequence accession
- Phyloreference: **<S_1&...&S_n**
- Specifier and query resolution semantics are defined by vocabulary terms.

SADI Subgroup

- Goal: Use the SADI* web service framework to link species occurrence data from GBIF to molecular data in UniProt.
- Outcomes:
  - Used taxon names from UniProt to integrate GBIF occurrence records.
  - Created web services that enable locality queries on UniProt through SPARQL

(*) See M. Wilkinson's Wild Ideas presentation this afternoon.





SADI example

Lessons learnt

- Bootcamps help tremendously, but only with the right level of detail. An expert in the group can be as or more effective.
- Good use-cases take much more time than you think.
- Real data is much less available than you think. Use made-up data for learning.

Non-Tangible Outcomes

- “I learned a lot.”
- “We have been talking about possible collaborations and grant proposals.”
- Shared goals and cohesion across communities: e.g., DwC, CDAO, TDWG, GCP
- Tried to integrate data beyond own domains. E.g., UniProt and GBIF

VoCamp and TDWG

- More than 10 participants who would not have come otherwise experienced TDWG.
- Future such events need better coordination with program committee.
- Can we leverage TDWG to organize a 2010 follow-up event?
- How do we sustain future such events at TDWG? Can TDWG help with securing funding?

URLs

- Phyloinformatics VoCamp: http://evoio.org/wiki/VoCamp1
- Evolutionary Informatics Working Group: http://evoinfo.nescent.org
- NESCent-sponsored hackathons: http://hackathon.nescent.org
- EvoIO Interoperability network (NSF proposal under review): http://evoio.org