The pPOD Core Data Model


Special thanks to Brent Mishler, David Maddison, Jeff Oliver, Rutger Vos, Francois Lutzoni, Martin Ramirez, Jonathan Coddington, Wayne Maddison, Fan Ge, Ashley Green, Jin Ruan, Martin Wu, John Lundberg, John Sullivan
The Core Data Model (CDM) under development in the pPOD project will serve the following purposes:

• It will allow experimentation with the **modeling of provenance** in phylogenetic pipelines.

• It will serve as a **schema for a persistence tool**, to work (1) in standalone mode, (2) with our lab notebook suite and (3) integrated with Mesquite as a module.

3. It will serve as a **target for schema mappings** used to connect other AToL databases, resources like TreeBASE, etc.
The Role of Provenance

Backwards provenance “query”
Starting from a research “product”, eg. a tree, a supertree, a matrix, track backwards through stored objects to all the raw input information that led to this product.

Forwards provenance “query”
Starting from a raw input, eg., a specimen, an image, a sequence, track forwards through stored objects to all research products that this input contributed to.

In both cases, navigate biological assumptions in both directions, eg., homology assumptions.
Persistence Tool

query

CDM (an OO schema)

store commands

provenance query

schema mappings

AToL AAA

TreeBASE

Mesquite module

RDBMS

Kepler-based workflow tool

(phylogenetic query language)
AToL Data that needs to be modeled in CDM
(not an exhaustive list)

Analyzed data:  
- trees,
- matrices, cells, (row) segments,
- operational taxonomic units (OTUs),
- taxa,
- standard characters and their states,
- genes, gene fragments

Raw data:  
- standard views, images,
- sequences, chromatograms,
CDM: Phylogeny Inference Data

Analyzed data: trees, matrices, operational taxonomic units (OTUs), standard taxa

Diagram:
- Tree
  - provenance
- Matrix
  - StdMatrix
  - SeqMatrix
- StdTaxon
  - authority
- OTU
  - taxon
  - isA
  - Set
  - List
Modeling Provenance (1)

Must be modeled and stored explicitly!

But it can be provided by automatic workflow tools
“Kinds” of Provenance

- Relationship between stored objects
  - Eg., tree T123 was obtained from matrix M456 by Joe Bio on 01/31/2001 using PAUP with parameters... (SEE PREVIOUS SLIDE)
  - Tracking through copy or cut/paste operations, possibly across repositories

- Trace of data moving through a workflow
  - Sequence of timestamps, tool invocations (parameters), authors

Trace of data through a logically expressed view/query
  - Can be computed automatically as the view/query output is computed
CDM: Morphological Data

Analyzed data: standard matrices, cells, standard characters and their states,

Raw data: standard views, images, specimens, collections
Modeling Provenance (2)
Example of Phylogenetic Query

Find all standard matrices with some character C whose label contains the substring "elytra" and some OTU whose state for character C contains the substring "transverse"; return all such matrices, together with their characters, OTUs and states satisfying the conditions.
Semi-formalized (OQL) query example

SELECT  \( M \), label of \( C \), label of \( X \), label of state encoded in cell \( E \)
FROM  \( M \) over all standard matrices,
\( C \) over all characters of \( M \),
\( X \) over all OTUs of \( M \),
\( E \) is the cell corresponding to \( C \) and \( X \) in \( M \)
WHERE  the label of \( C \) is like "*elytra*"
AND  the label of the state encoded in cell \( E \) is like "*transverse*"
Molecular Data

**Analyzed data:** sequence matrices, (row) segments, genes, gene fragments

**Raw data:** sequences, chromatograms, specimens, samples, collections
molecular matrix

OTU1

OTU2

row segment (from some sequence) from some contig

from different specimens

gene frag 1
gene frag 2