Chado for evolutionary science

Chris Mungall

HHMI (until June)

National Center for Biomedical Ontologies (after June)

Outline

- Chado key concepts
- Chado selected module tour
  - sequence: genome annotations
  - cv: ontologies and terminologies
  - phylogeny: evolutionary trees
  - phenotype: character based descriptions
    - PATO: attribute ontology

Chado: what is it?

- A Database schema for molecular biology
  - primarily model organism
- Relational schema
  - DBMS-independent
    - PostgreSQL, Sybase, Oracle, DB2
  - Has XML form
    - ChadoXML
      - comes “for free”
- Part of GMOD
  - interoperates with Apollo, GBrowse, Turnkey, ..
  - in use at various MODs and genome centers

Chado key concepts

- Integrated
  - not federated
  - foreign key relations between entities
- Modular
  - separation of concerns
  - mix-n-match
- Generic and extensible
  - uses ontologies and ‘cv’s for typing
- *Normalisation* over efficiency
- Community & open source

Chado modules

- Core
  - general (dbxrefs)
  - cv (ontologies)
  - pub (bibliographic)
  - audit
- Domains
  - sequence (genomics)
  - companalysis
  - expression
  - RAD
  - map
  - genetic
  - phenotype
  - phylogeny
  - organism
  - event

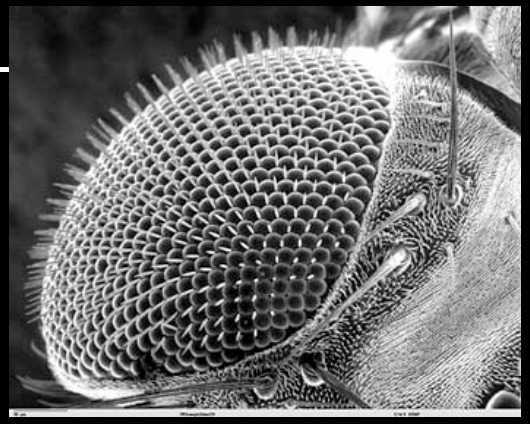
sequence
cv
phylogeny
phenotype

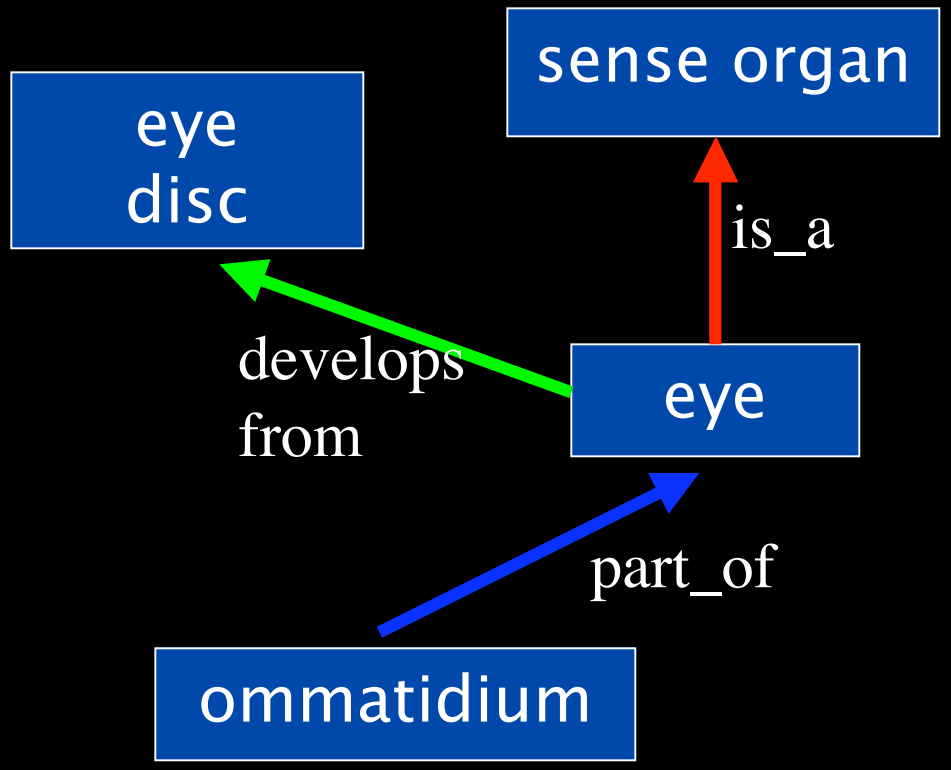
Sequence module

- some key tables:
  - feature
  - featureloc
  - feature_relationship
  - featureprop
- plays well with GFF3
- feature types come from SO (sequence ontology)

A biological ontology is:

- A precise representation of some aspect of biological reality
  - what *kinds* of things exist?





sequence
cv
phylogeny
phenotype

cv module

- Ontologies and controlled vocabularies ('cv's) are ubiquitious in Chado
- key tables
  - `cvterm` (a term, or class in an ontology)
  - `cvterm_relationship`
  - `cv`
- can represent any ontology in OBO
- compatible with RDF/RDFS

sequence
cv
phylogeny
phenotype

phylogeny module

- key tables:
  - phylotree
  - phylonode
    - has one parent; branchlength
  - phylonodeprop
    - extensible tag-value pairs; eg statistics
- Also:
  - phylonode_feature
    - can make trees from any feature type
  - phylonode_organism

phylogeny module status

sequence
cv
phylogeny
phenotype

- Status: new
  - needs more community input!
- SQL Query support
  - nested set representation
  - SQL functions
- Currently no links to phenotype module

sequence
cv
phylogeny
phenotype

phenotype module

- Originally intended for model organism mutant phenotypes
  - adaptable?
- Based on 2003 EAV model
  - requires extensions?
    - e.g. Diederich 1997
- Currently mixed in with genetics module
  - in production use at FlyBase

sequence
cv
phylogeny
phenotype

key tables

- EAV model
- phenotype
  - entity_id (a cvterm from e.g. anatomy)
  - attribute_id (cvterm from PATO)
  - value_id (cvterm from PATO)
  - value (free text alternative to above)
  - stage_id (cvterm from e.g. dev stage)
- phenotype_comparison

sequence
cv
phylogeny
phenotype

PATO: Attributes

- Attributes (qualities)
  - physical attribute
    - weight
    - color
    - morphology
      - shape
      - size
  - temporal (process) attribute
    - duration
    - frequency

sequence
cv
phylogeny
phenotype

PATO: Attributes

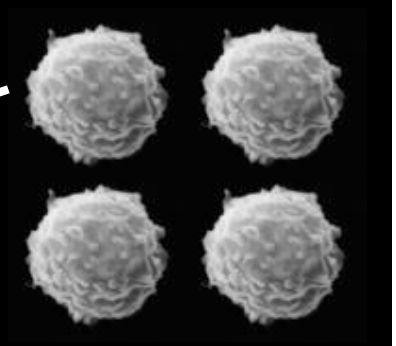
- Attributes (qualities): values (states)
  - physical attribute
    - weight: heavy, light
    - color: red, magenta
    - morphology
      - shape: branched, cleft, coiled
      - size: large, small, hypertrophied
  - temporal (process) attribute
    - duration: long, short
    - frequency: frequent, infrequent

sequence
cv
phylogeny
phenotype

Combining entity terms and PATO terms

- Entity-attribute structured annotations
  - Entity term; *PATO term*
    - tail fin ZDB:020702-16; *ventralized* PATO:0000636
    - kidney ZDB:011143-431; *hypertrophied* PATO:0000584
    - midface ZDB:020702-16; *hypoplastic* PATO:0000645

- Pre-composed phenotype terms
  - Mammalian Phenotype Ontology
    - "increased activated B-cell number" MPO:0000319
    - "pink fur hue" MPO:0000374



sequence
cv
phylogeny
phenotype

Extensions to simple EAV?

- New phenotype XML schema being developed for:
  - Relational attributes
  - Separation of measurements from attribute states
  - Composing entity terms
  - Value/state sets
  - Relative states
  - Variation in space and time
  - Is the 'A' superfluous?
  - Alternative to absence/number
- See *Diederich 1997*

sequence
cv
phylogeny
phenotype

OBD and NCBO

- National Center for Biomedical Ontology
  - driving biological project:
    - genotype-phenotype-disease
- Tool development
  - phenotype annotation (EAV)
- OBD
  - Data counterpart to OBO
  - Generic metamodel (RDFS/OWL)
  - Chado relational views

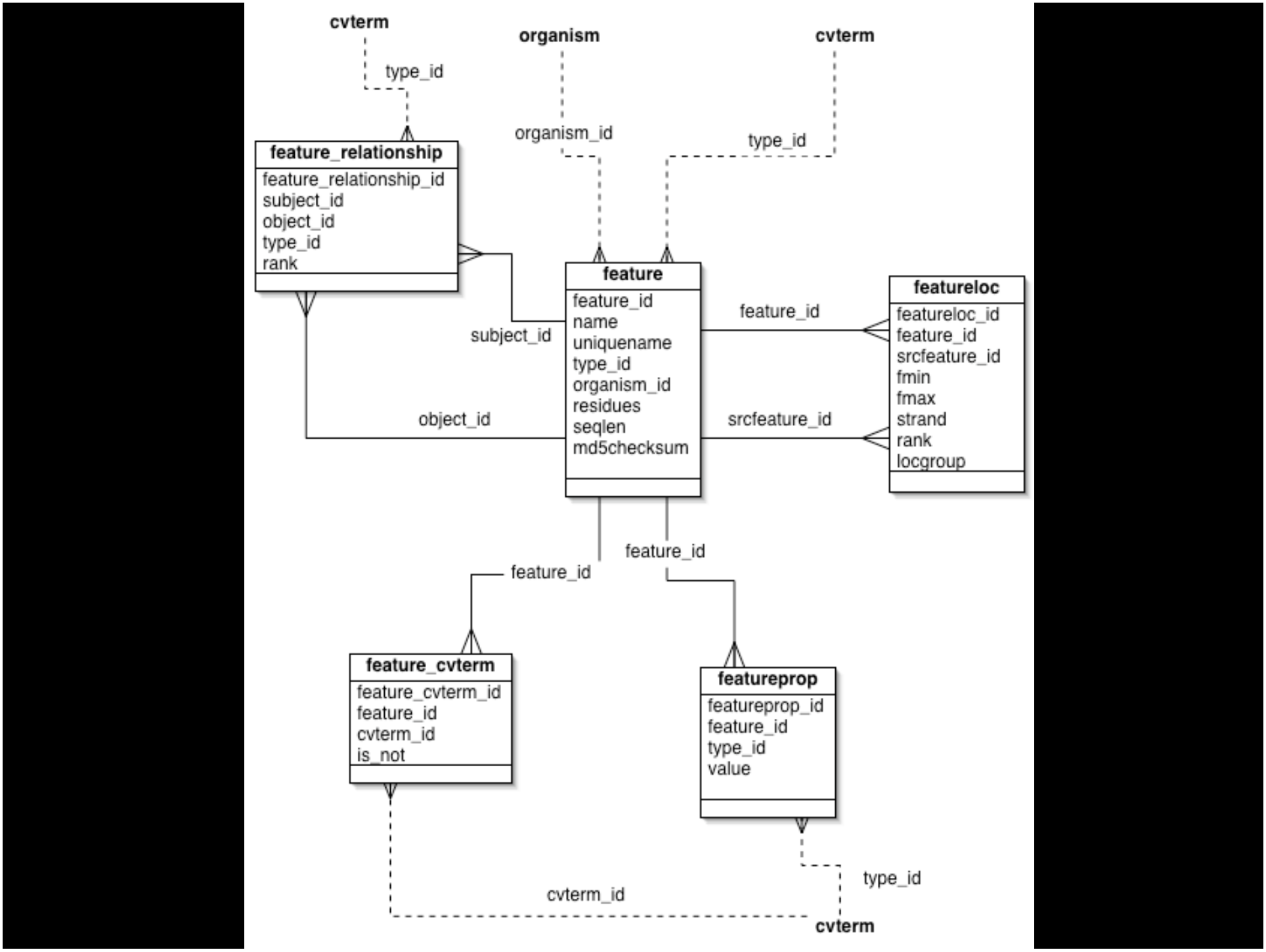
Summary

- cv module is stable
- genomics part of chado is stable
    - weeellllll, the meta-schema is still evolving…
- phylogeny and phenotype untested
  - more input required
  - complexity of phenotypes
- PATO may need to evolve a lot
- a lot of software still to be written

Thanks

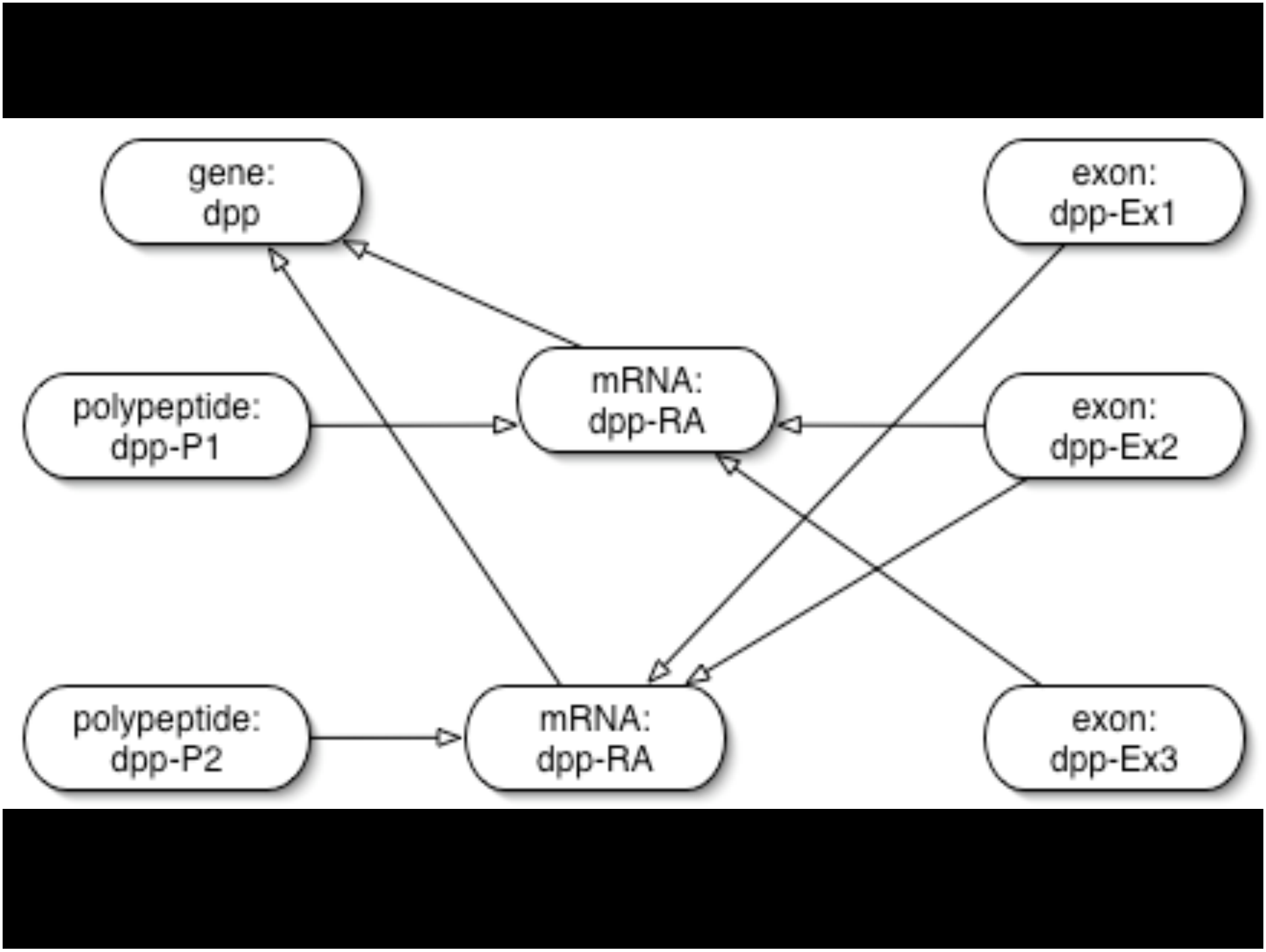
- Chado
  - Dave Emmert
  - Stan Letovsky
  - Shengqiang Shu
  - Pinglei Zhou
  - Aubrey de Grey
  - Scott Cain
  - Lincoln Stein
  - Mark Gibson
  - Peili Zhang
  - Colin Wiel
  - Richard Bruskiewitz
  - Allen Day
  - Bill Gelbart
  - Gerry Rubin
  - Suzanna Lewis
- PATO
  - Georgios Gkoutos
  - Monte Westerfield
  - Fabian Neuhaus
  - John Day-Richter
  - Rachel Drysdale
  - Michael Ashburner

supplemental slides follow…


cvterm
type_id
organism
organism_id
cvterm
type_id
feature_relationship
feature_relationship_id
subject_id
object_id
type_id
rank
feature
feature_id
name
uniquename
type_id
organism_id
residues
seqlen
md5checksum
featureloc
featureloc_id
feature_id
srcfeature_id
fmin
fmax
strand
rank
locgroup
subject_id
feature_id
object_id
srcfeature_id
feature_id
feature_id
feature_cvterm
feature_cvterm_id
feature_id
cvterm_id
is_not
featureprop
featureprop_id
feature_id
type_id
value
type_id
cvterm_id
cvterm

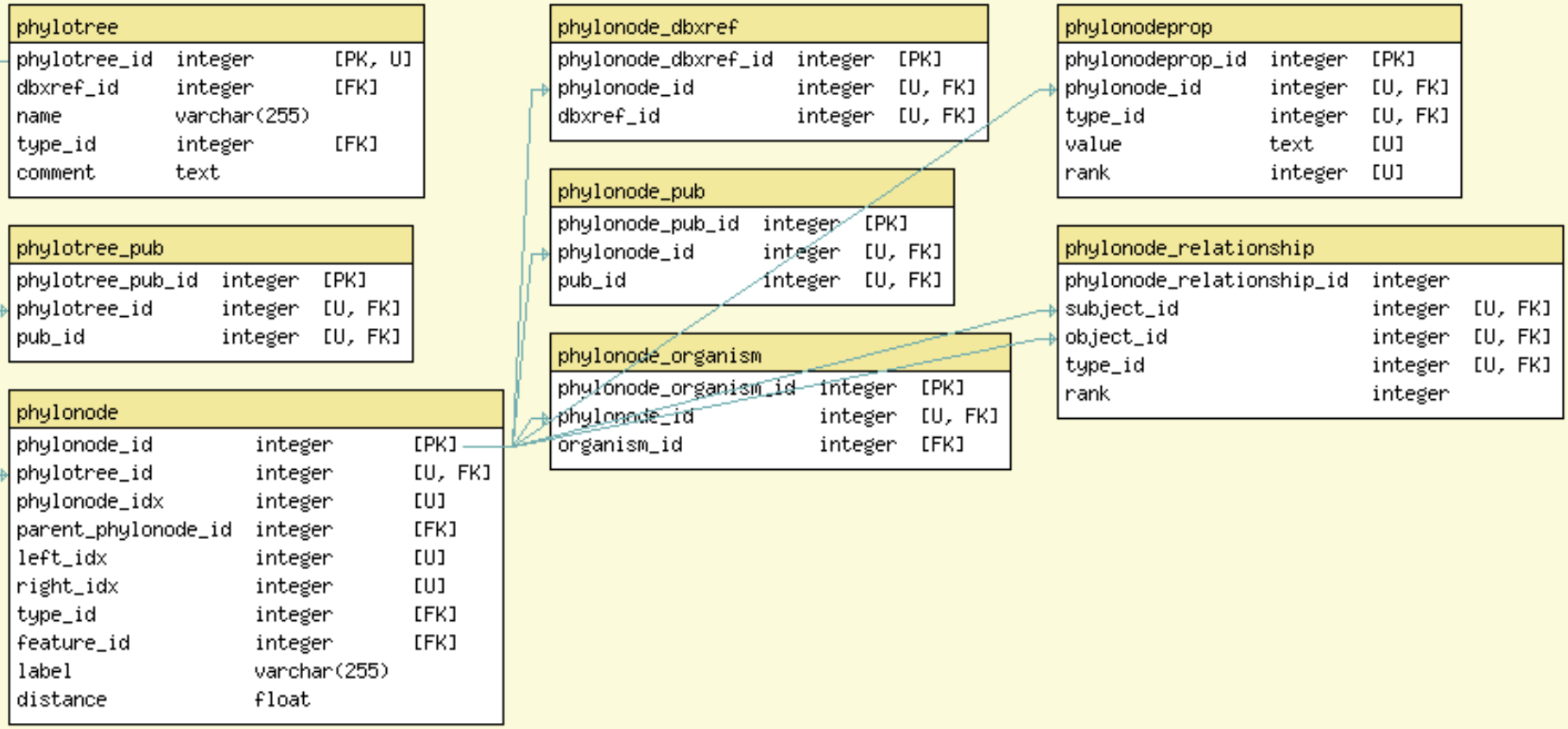
the feature_relationship table

- Feature graphs
- Relationships between pairs of features
  - this exon *part_of* that mRNA
  - this polypeptide *derives_from* that mRNA
- Feature graphs constrained by SO
  - (all) exon *part_of* (some) mRNA
  - (all) polypeptide *derives_from* (some) mRNA


gene:
dpp
exon:
dpp-Ex1
polypeptide:
dpp-P1
mRNA:
dpp-RA
exon:
dpp-Ex2
polypeptide:
dpp-P2
mRNA:
dpp-RA
exon:
dpp-Ex3

organism module

- ultra-simple
- key (only!) entity
  - `organism` (actually: taxon)
- NCBI taxon compatible
  - `unique(genus, species)`
- Each feature *must* be linked to an organism
- Strains?
- But what about evolution?…



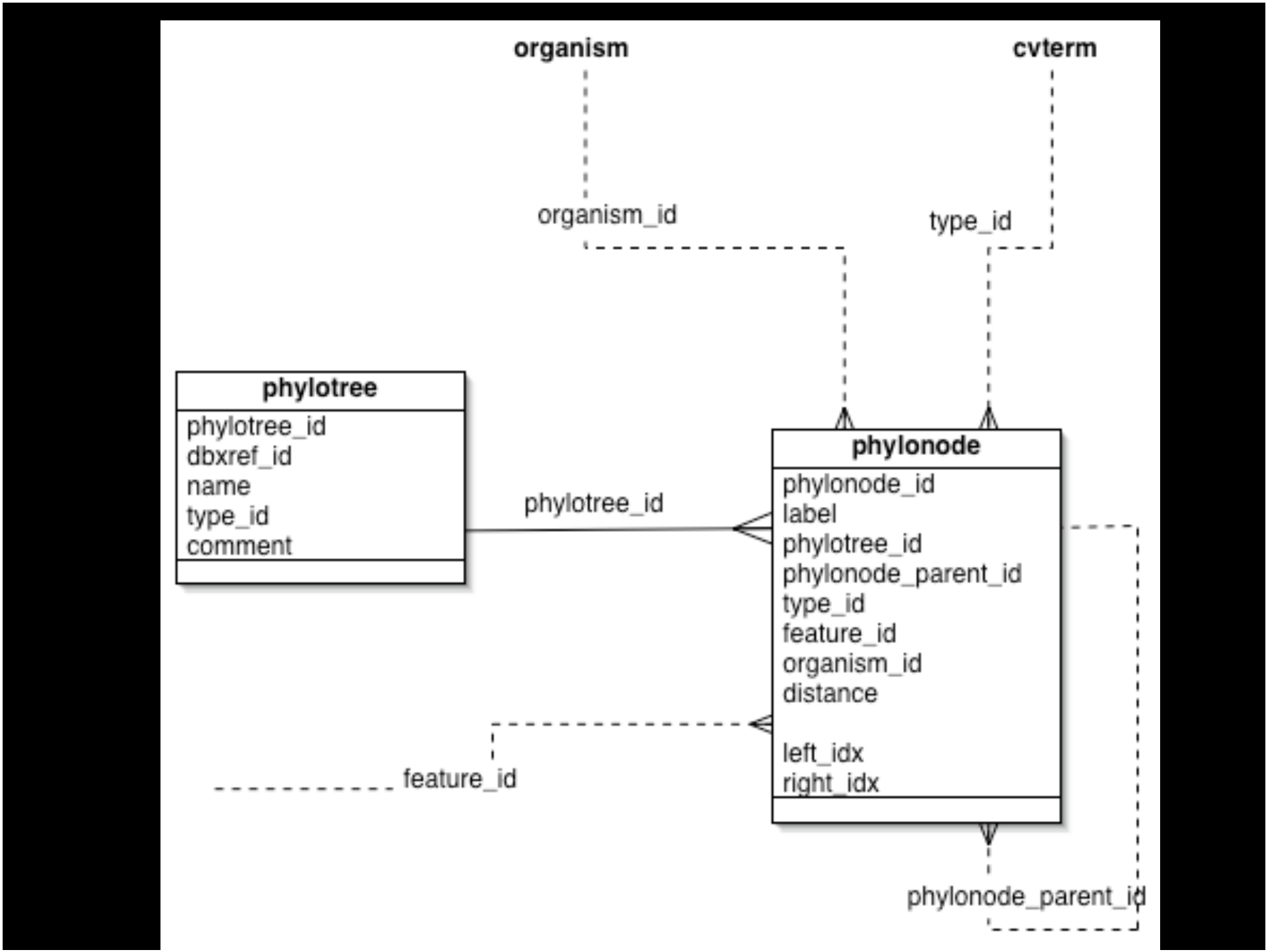
Legend

[FK] Foreign Key

[U] Unique constraint

[PK] Primary key

Created by SQL::Translator 0.07


organism
cvterm
organism_id
type_id
phylotree
phylotree_id
dbxref_id
name
type_id
comment
phylotree_id
phylonode
phylonode_id
label
phylotree_id
phylonode_parent_id
type_id
feature_id
organism_id
distance
left_idx
right_idx
feature_id
phylonode_parent_id

phylogeny module

- Status: mature, not yet in production use
- Original design:
  - Richard Bruskiewitz (IRRI)
  - Adapter from Aaron Mackey's design
- Key concept: trees
  - nodes
  - nodes have a single parent (except root)
    - branch length
  - node belongs to one tree
  - nodes can be linked to other chado entities
  - nodes can have data attached (e.g. statistics)

phylogeny: SQL functions

- Views and SQL functions can greatly enhance queryability
- Implemented:
  - `phylonode_height(phylonode_id)`
  - `phylonode_depth(phylonode_id)`
- TODO
  - `is_monophyletic(phylonode_id[ ], outgroup_id)`
  - etc

phylogeny module

- Can make trees of anything (within reason)
  - taxonomies
    - table: phylonode_organism
  - features of any type in SO
    - table: phylonode_feature
      - polypeptides (protein)
      - introns
      - …
  - links to alignments, etc

phylogeny module: open questions

- controversial?
  - *optional* phylonode_relationship (DAGs!)
- Query efficiency vs redundancy
  - nested set representation
- Attaching characters (phenotypes)
  - nothing in place at present

Linking phylogeny to phenotype

- phylonode already linked to feature
  - (features have sequences)
- how do we link phylonode to phenotype
  - nature of linkage
    - essentialist
    - statistical
- what have we missed?
  - e.g. environment

phenotype module: status

- Is basic EAV model sufficient for
  - model organism mutant phenotypes
  - systematics
- Extension required?
  - Diederich 1997
- Managing multiple versions
  - compatibility layers

Software integration

- Molecular phylogeny
  - easy?
    - eg via Bio::Tree in bioperl
- Character-based phylogeny
  - NEXUS etc
  - easy/hard???