assembling a draft overall tree of life from phylogenetic trees and taxonomic databases

Jonathan A Rees
US National Evolutionary Synthesis Center
Duke University
rees@nescent.org
TDWG, 31 October 2013
software team:

Jim Allman
Joseph Brown
Karen Cranston
Cody Hinchliff
Mark Holder
Jonathan Leto
Emily McTavish
Peter Midford
Rick Ree
Stephen Smith

funding:

US NSF
what is open tree of life?

1. collect
2. normalize
3. synthesize
4. expose
1. **collect** phylogenetic trees for best possible coverage of entire tree of life

2. **normalize** tips so that they match between source trees

<table>
<thead>
<tr>
<th>label</th>
<th>normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Hemsleya amabilis HS454</em></td>
<td>524163 <em>Hemsleya amabilis</em></td>
</tr>
<tr>
<td><em>Theria</em></td>
<td>4267989 <em>Theria</em> in <em>Arthropoda</em></td>
</tr>
<tr>
<td><em>Nicotiana suaveolans var excelsior</em></td>
<td>232354 <em>Nicotiana rotundifolia</em></td>
</tr>
<tr>
<td><em>Selysia prunifera</em></td>
<td>949305 <em>Cayaponia prunifera</em></td>
</tr>
</tbody>
</table>
3. **synthesize** a single ‘big tree’ algorithmically from the source trees

4. **expose** source trees and ‘big tree’ in various ways
exposing provenance

- links to studies
- links to data deposits (e.g. treebase)
- links to taxonomic database records
- methods documentation
- versioning
reference taxonomy

- used for normalization, internal node labeling, gap-filling
- need NCBI taxonomy
- supplement with GBIF
- patch system
- future: other sources
‘open’

trees are not creative expression

... ergo no © protection

... ergo © licensing is meaningless

... CC0 is nice (and required by Dryad), but no CC0 for legacy data or NCBI
lessons

• NeXML and badgerfish are good
• machine-processable tip identity would be awfully nice
• we were surprised by tree rooting problem
• provenance is an uphill battle
• to be seen: github for data curation?