Using a git-based datastore for community curated phylogenies

June 16, 2014

Abstract

Motivation: The Open Tree of Life project is collating and synthesizing hundreds of phylogenies into a unified tree of life. The base data for for this project are community contributed phylogenies. Via the OpenTree curator tool, these phylogenies are updated to reflect consistent tip names across studies, and accurate rooting. **Results:** Using a git-based datastore to hold these phylogenies automatically version controls data as updates are made. In addition, hosting this datastore on GitHub provides a straightforward and familiar method for researchers to access these data. **Availability:** Both the code and the datastore are available on GitHub.

Outline:

Introduction:

Set up the problem to be solved:

1. OpenTree concept

The Open Tree of life project is a collaborative efforet to synthesise phylogenetic knowledge across the entire tree of life ([**?**]). This ambitious undertaking uses supertree methods to combine large numbers of individual phylogenies with a taxonomic backbone into a synthetic tree comprising all named species. In addition, this data set is itself useful to other researchers in biology. Most published phylogenies are available only as images in the archives of the journals in which they were publishe([**?**]) This format makes it nearly impossible for researchers to re-use this data. A secondary goal of the OpenTree of life project is to make the datastore of published phylogenies which makes up the synthetic tree available and accessable to researchers in a usable format.

While several projects have been developed to store phylogenetic information, most notably Treebase [**?**] and Phylografter [**?**].

builds upon a datastore of individually published phylogenies with their associated metadata. 3. Community curation

4. Choice of Git

The Open Tree of Life project is creating a synthetic phylogeney, incorporating published phylogenies from across the tree of life. We are using a git based datastore, mirror to github, to simulataneously track study curation and disseminate the phylogenies.

Features:

1. Transparent Fine grained versioning

+1 Straightforward and automatic with git.
-1 Files are not line based, so diffs can get weird.
+1 branching
-1 complexity of file org makes auto merges hard and maual merges confusing.

2. Familiarity

3. Shareable

+1 Whole datastore is currently cloneable on github!
+1 API's can request indivual studies fomr github rather than relying on the OpenTree server.

4. Comparison to alternatives:

Not structured (SQL)
annotations etc can be easily added
Searching/indexing implemented using OpenTree Index (OTI)
Mongo, couch, ???

Implementation

1. Stucture:

Git repo on server pushed to mirror, mirror pushes to git hub.
- Github web hook nudges OTI to update on push * Editing (curation) occures via the Opentree Curator webapp.
Curation of a branch creates work in progress (WIP) branch of repo.
- When curator saves edits they are automatically merged to master if they are FF
- Otherwise (i.e. another WIP has been mergered into master since split?) curator must view and accept merged changes
(this relies on reasonable diffs being returned to curator)
Library to interact with Data store as well as other OpenTree componenets descibed in Peyotl

2. Future:

Appropriately formatted files could be edited and merged into Phylesystem via pull request.

Is this model applicable for other datastores?

Conclusions