Project Description

OVERVIEW

Phylogenetic trees, which codify our knowledge of biology in the form of paths of descent, have proven extremely powerful in explaining, inferring and predicting the characteristics of past and present life on earth. Phylogenies enable us to see similarity of mortality patterns across primates [1], discover slower evolutionary rates in woody plants [2], predict species' response to climate change [3, 4] and infer the origin and spread of human pathogens [5, 6]. Owing largely to the dramatic reduction in sequencing costs and increase in computational power, phylogenies are being published at an exponentially increasing rate, with the size of data sets approaching genome-scale and covering ever higher numbers of taxa. However, most published phylogenetic results are still archived as PDF documents in journals' supplementary online materials, rather than in community data repositories in a digital form that allows effective discovery, sharing, and reuse. Scientists who use phylogenies for their research cannot easily download these trees and are therefore forced to recreate them from images, re-infer them, or replace them with much more easily accessible, but not entirely appropriate, classifications [7] [8]. Perhaps more importantly, it is difficult or nearly impossible to assess the overall state of our phylogenetic knowledge, how it has changed through time, the effect of different phylogenetic inference methods or evolutionary models, and the necessary data or re-analyses to most effectively fill the gaps in our knowledge.

Here we propose to lay the foundation for overcoming this situation through a linked set of informatics components that allow permanent archiving, efficient discovery, fast access, rich decoration, and collaborative curation and scalable synthesis of our phylogenetic knowledge. The principle challenges are three-fold: 1) Dramatically increasing the fraction of published trees that are archived in a discoverable and reusable form; 2) drastically improving the tools for integrating these trees into a continuously updated synthetic Tree of Life; and 3) engaging a diverse and growing community of users, scientists, and experts to share, curate and synthesize this data.

We propose to meet these challenges by building on the two preeminent community phyloinformatics resources. One is TreeBASE [9], the community's designated digital repository for archiving, finding, and sharing published phylogenetic trees and underlying data. The other is the Tree of Life Web Project (ToLWeb) [10], where a community of experts collaborate to present a single, curated, and richly annotated phylogeny of life. During their existence for more than a decade, both resources have accumulated considerable assets, both digital, such as user-contributed content and working source code, and social, including community recognition in the case of TreeBASE and a devoted community of expert contributors in the case of ToLWeb. The social assets are particularly valuable as it is much harder to "recreate" the engaged community than it would be to re-write software.

Despite, and in part because of the long history of TreeBASE and ToLWeb, many of their user-facing features do not meet the challenges set out above. To name a few examples, the current submission interface of TreeBASE, aimed primarily at making submitted trees useful to others, is difficult and time-consuming to operate; for both resources, finding and retrieving content is complicated and labyrinthine; ToLWeb coverage is woefully incomplete and the interface implements very few features that reward participation. In large part, this is due to the technologies and architecture paradigms prevailing (or not yet existing) at the time these resources were last developed. For example, in recent years social features that incentivize contribution have not only matured, but have also become the subject of social science research to understand why, and when such features "work" [11, 12]. Similarly, there are now freely available enterprise-level search platforms based on efficiently indexing document stores (such as Solr [13]) that offer a faster and more intuitive search experience [14] than is typically achieved with the type of relational datastores that currently underlie TreeBASE.

Perhaps more strikingly, TreeBASE and ToLWeb have never interoperated in any significant way. There are no tools to support the process of integrating newly published trees into the Tree of Life, and no cross-linking between richly-annotated large-scale synthetic trees and published phylogenetic research. By making these connections explicit, we will enable views of this data that are not currently possible, such as changing levels of support and degree of congruence for alternative topologies over time, or calculating impact metrics that quantify how a deposited research tree advances our phylogenetic knowledge. We will not only connect and integrate data, but also the people that authored the data, thus providing a social mechanism that rewards contribution of high-quality content.

Overcoming the challenges outlined above depends on how well we succeed in engaging a diverse group of people to adopt, utilize, and contribute to the resources we build. Meeting people's needs on a technical basis is obviously one prerequisite. However, data sharing and collaborative knowledge curation are social activities, and there is ample evidence that resources that support the social nature of such activities fare much better in user engagement and adoption. Neither TreeBASE nor ToLWeb have any of the social features that are now omnipresent among online websites aimed at user-contributed content, and that are key to the success of these sites. As one of the core and novel goals in this proposal we will add social capabilities to the ToLWeb and TreeBASE user interfaces. These capabilities will be guided by the considerable amount of social science research available on the factors that influence whether and how scientists participate in a collaborative cyberinfrastructure, and their effectiveness towards our objectives will be continuously evaluated as part of a social science research study.

Our development goals are designed to ultimately result in a combined community resource that 1) allows for most published phylogenetic knowledge to be archived, and to be rapidly integrated into, and remain cross-linked to, a synthetic Tree of Life; 2) allows sharing of phylogenetic knowledge in an efficiently discoverable and reusable form; and 3) uses social features and incentive mechanisms to engage people to participate and contribute, thus giving the community ownership in the content as well as in the resource itself. The goals and deliverables are as follows:

I) Cross-linking the content in TreeBASE and ToLWeb. We will compute bidirectional links between research trees in TreeBASE and their phylogenetically corresponding clade(s) in ToLWeb. Such links will be annotated with indicators of content richness (if ToLWeb is the destination) and fitness for reuse (if TreeBASE is the destination), and downloads can optionally include the content connected in this way.

II) Tools for merging research trees from TreeBASE into the Tree of Life. We will develop tools for ToLWeb clade editors that automate as many steps as possible in merging newly available research trees into ToLWeb, including notification when candidate trees become available, finding candidate trees, and "glomming" a research tree to an existing Tree of Life.

III) Usability-driven submission and ingest interface to TreeBASE. We will create a submission process to TreeBASE that consists of an API usable by 3rd party software and an interactive web interface that is focused on scalability, lowering barriers to deposition, and helping authors to resolve commonly encountered issues, all without compromising on quality and reusability of submitted data.

IV) Fast and efficient searching, discovering, and downloading of content. We will create faceted browsing and simple plus advanced search interfaces for TreeBASE and ToLWeb content. Bulk data downloads will be powered through an efficiently indexed document store regularly updated from the relational data stores native to TreeBASE and ToLWeb.

V) Social computing features and contribution incentives. We will, in a staged fashion, add a series of social computing capabilities to ToLWeb and TreeBASE to increase participation, incentivize contribution, and provide indicators of content richness and fitness for reuse.

VI) Community building and support. We will recruit staff dedicated to nurturing the community of experts, users, content authors, and external developers through outreach, coordination, help desk support, and training.

Integrated with these development goals are *two research activities*: 1) A driving biological research project on synthesizing mega-phylogenies from research trees will inform the user-facing interface and tree-grafting methods development; and 2) Social science research on generalizing knowledge about the role and impact of social features in cyberinfrastructure will inform the social features most suited for our application and audience, and will continuously evaluate their effectiveness. The scope of these goals and activities is the tree of life of all eukaryotes. We exclude prokayotic research from these goals because the work required to include these domains would require significant additional resources, although we welcome the input from researchers in these areas in community engagement activities.

BACKGROUND

TreeBASE was developed for the dual purpose of serving as a digital library of phylogenetic knowledge and also a resource for phyloinformatic meta-analysis and synthesis [15, 16]. TreeBASE stores trees, matrices of molecular, discrete (morphological) and continuous character data, and important metadata for tree branches and nodes, including branch lengths, clade support, and basic specimen-related DarwinCore [17] attributes. TreeBASE is implemented in the Java programming language as a 3-tier web-application on top of a fully normalized relational database. To promote reusability of content, it reconciles user-provided taxon labels with uBio [18] and NCBI taxonomy [19] identifiers; uses persistent, globally-unique and resolvable identifiers, and delivers data in a number of different exchange and metadata standards (NeXML [20], NEXUS [21], RSS [22], RDF [23]). More recently, a handshaking protocol allows the Dryad [24] to push phylogenetic data submissions to TreeBASE, and an OAI-PMH [25] service allows Dryad to ingest TreeBASE metadata. The rate of new submissions to TreeBASE has steadily increased over time, although not nearly as fast as the growth of published phylogenies in the literature. As of July 2011, TreeBASE contained data for 2,780 publications, including 5,370 data matrices and about 8,000 trees with nearly 93,000 distinct taxon labels, of which 73,610 are mapped to taxonomic sources. There are at least 28 journals that either require or recommend that authors submit data to TreeBASE. Currently TreeBASE holds submissions for 320 different scientific journals, covering a wide spectrum of the life sciences. TreeBASE facilitates the review process by providing advanced access to reviewers, which allows assessment of not only results, but also raw data, models and parameter settings.

The **Tree of Life Web project** (ToLWeb) was the first biodiversity database to organize content in an explicit evolutionary framework. Since its creation in 1994, ToLWeb has grown to incorporate 7,957 clade pages, 5,094 species pages, and 265 education pages (called “TreeHouses”). ToLWeb has an international community of curators, with 754 registered contributors from 366 institutions and 39 countries. Some of these contributors have considerable dedication to the project: nearly 100 people have contributed to more than 10 pages each. Despite being without direct funding since 2009, the project continues to grow, with 144 new pages in 2010 and 45 new contributors, although the ToLWeb content does not cover even a fraction of the taxa available in other biodiversity databases. In 2010, the site received over 3 million unique visitors, with over 143 million views of media files (primarily images) and about 1 million views of TreeHouses. ToLWeb has been forward-thinking in its incorporation of social features, including public lists of curators and contributions, quality metrics for pages and facilities for peer-review, but these features have lacked integration with other tools and resources. ToLWeb already has great success in engaging the general public both as consumers and as content contributors. For example, students at Lead-Deadwood High School in South Dakota assembled a TreeHouse page [26] with text and images documenting indigenous plants of the Dakotas, their secondary compounds, and medicinal uses by the Sicangu Lakota people, supplemented by sound files of Lakota plant names spoken by a member of the Sicangu Sioux Tribe.

Community Engagement and Social Computing. The widespread adoption of social web features (i.e. "Web 2.0") has transformed users' expectations and behavior on the Internet. While these social aspects are most often considered in the context of online games or leisure sites like Facebook, there is significant interest in how social computing can be used to facilitate and advance science. Data and knowledge sharing in science are social activities, influenced by factors like reputation and reward [27, 28], social networks [29] and trust [30]. These social issues are even more pressing in large-scale collaborative systems.

Sharing datasets and phylogenetic trees can play an important role in bringing together different communities. These digital artifacts can be seen as boundary objects that allow multiple communities to use the same artifacts but make meaning out of them in different ways [31]. While an ecologist may seek a phylogeny as the basis for a comparative analysis, a synthetic biologist may search the database for opportunities for engineering new drugs, and high school students may draw on ToLWeb to prepare a report about evolution. To be sustainable, TreeBASE and ToLWeb should support scientific investigation across multiple communities. At the same time, it is important to support the generation of new knowledge that can happen in the space between disciplines [32]. Repositories can support these collaborations by providing opportunities to interact with artifacts in their collection to discuss and annotate the items in a way that allows community understandings to be revealed and developed.

SIGNIFICANCE

This is the first attempt to aggregate phylogenetic data across the entire tree of life, as it is published, via community collaboration. Linking TreeBASE, the only archive for core phylogenetic information, to ToLWeb will provide the conduit for new phylogenetic research to migrate unfettered into a synthetic Tree of Life. Even if in the foreseeable future only some of its branches can achieve the resolution and comprehensiveness characteristic of mega-phylogenies, they will be of immediate use in examining challenging evolutionary questions. We will demonstrate this as part of this project for fishes and asterids. For example, how have body size, growth rate, and functional traits evolved across the 450 million year evolutionary history of fishes? [33] Where are fossil taxa concentrated in the tree, and how effective are these for developing a time tree of fishes? How has flower symmetry evolved with respect to developmental constraints and adaptive value?

A key component in this proposal is the attention paid to social and user-experience aspects of cyberinfrastructure development. The work on design and evaluation of social capabilities for TreeBASE and ToLWeb provides direct benefit to these resources, but is also novel research in the social sciences. While there has been significant study of social media in general (e.g. Wikipedia, Facebook) and corporate domains, there have been few studies of social features in scientific cyberinfrastructures. The studies that have been done tend to focus on "citizen science" e.g. [34, 35], or very large scale "crowdsourcing" (e.g. [36]). Our work will explore the feasibility and impact of social media features in phylogenetics, and will provide guidance and exemplars for future cyberinfrastructure development.

DEVELOPMENT PLAN

Our development goals follow 3 overarching objectives: lower the barrier in every aspect possible (deliverables II, III and IV); create a user experience that encourages participation and rewards contribution (deliverables I-III, V); and build a diverse and healthy user community (deliverable VI). The goals combine radical redesigns to remove major barriers to using TreeBASE and ToLWeb (deliverables III-IV), while adding novel features (deliverables I, II, V).

I. Cross-referencing the content in TreeBASE and ToLWeb.

The historical lack of interoperability between ToLWeb and TreeBASE makes it difficult to explore the research that underlies the synthetic Tree of Life and misses an opportunity to incentivize researchers to archive the data that represents this knowledge. Similarly, the value of TreeBASE would be greatly enhanced by the additional annotations available in ToLWeb for related clades. On ToLWeb clade pages, we will include statistics about and links to related TreeBASE trees - published studies that include taxa in this clade, authors, active publication years, phylogenetic reconstruction methods, and amount or type of data used in the reconstructions. Links will provide visual indicators that show what might be found at these links (such as number of descendent species, phylogenetic coverage, access count). We will calculate measures of topological incongruence (e.g., [37-41]) between a clade and corresponding trees in TreeBASE to identify unresolved or controversial nodes in the synthetic tree, and highlight those nodes graphically. In TreeBASE, we will link studies to ToLWeb clade pages, visual indicators to allow users to gauge the metadata richness for the clade, such as number of contributing authors, amount of text, multimedia files and diversity of annotations. Furthermore, we will highlight other TreeBASE studies that have been incorporated into that clade (see deliverable II below). All such links will also be exposed through the data access API (see deliverable IV) for use by 3rd party applications.

These links between data also connect the authors of those data within emergent social networks, which we can expose as one of the social computing features (see deliverable V) and leveraged to incentivize quality improvement. For example, if a clade editor finds that a study newly published in TreeBASE does not meet the metadata requirements for incorporation into ToLWeb, she can generate a corresponding notification, and then either the author of the study or others can collaboratively make the necessary quality improvements. To make this work, we will reconcile user identities between the two systems and migrate to a single point of authentication.

We will compute these cross-links through a two-step process of taxon name resolution (using uBio [18], NCBI [19], TNRS [42], and other reconciliation services as suitable) followed by phyloreferencing to map nodes with compatible “phylogenetic semantics”, in the sense of representing the same most recent common ancestor (MRCA). In a proof-of-concept study that uses the NCBI taxonomy for mediating between taxon names, we have developed a preliminary indexing procedure based on nested sets and transitive closure paths [43]. The procedure 1) finds the node in the NCBI tree that is the MRCA of all OTUs in a given TreeBASE tree; 2) finds all internal nodes in the ToLWeb tree that descend from this MRCA node; and 3) for each ToLWeb node found, ordered from smallest to largest clade size, cross-links it with the equivalent MRCA node in the TreeBASE tree. In a test of the procedure on 2,000 randomly chosen trees in TreeBASE, we were able to successfully cross-link over 99% of the nodes.

II. Tools for merging research trees from TreeBASE into the Tree of Life.

The NSF Assembling the Tree of Life (AToL) [44] program has radically improved our knowledge of phylogenetic relationships. Yet, finding this knowledge in synthesized or re-usable fashion remains a challenge, even for charismatic clades such as Carnivores and Marsupials with well-studied phylogenetic relationships [45-48]. While data repositories and changing community norms about archiving research data [24, 49] are improving the availability of phylogenetic data from individual studies, synthesizing these results across the Tree of Life in a form that is accessible to a broad community of researchers, educators and the general public requires significantly improving the flow of archived phylogenetic data into the Tree of Life. One of our major goals is to create tools that for the first time automate several of the tasks involved for a clade editor who wants to incorporate research trees into ToLWeb.

Specifically, we will create tools for identifying knowledge lacking in ToLWeb but for which data exist in TreeBASE, and for incorporating these data into ToLWeb in a way that maintains community confidence in the ToLWeb tree and the social incentives for authors or others to curate the data. Both components will make use of the cross-links created between the two resources (see deliverable I above), and will use these to transfer data from one repository into the other. The initial implementation will allow a user to select a specific TreeBASE study, and graft the published tree into ToLWeb using the existing curation tools. We will add an interface for choosing from studies linked to a ToLWeb node (by shared most recent common ancestor) according to metadata (such as analysis method, data type, etc) and quality criteria. One product of the research activities on mega-tree synthesis is expected to be a scientifically validated best practice protocol for combining trees, which we will apply to implementing the ToLWeb tools for incorporating research trees. We will reuse existing software for merging trees as suitable to achieve this, such as Phylomatic [50] and PhyloGrafter [51]. Later development will focus on increasing the usability of our tools, including capability for ToLWeb to handle and display alternative phylogenetic tree topologies for a clade. This will eliminate the need for phylogenetic consensus, improving the resolution of the ToLWeb tree and motivating scientific curation of alternative phylogenetic hypotheses.

III. Usability-driven submission and ingest interface to TreeBASE

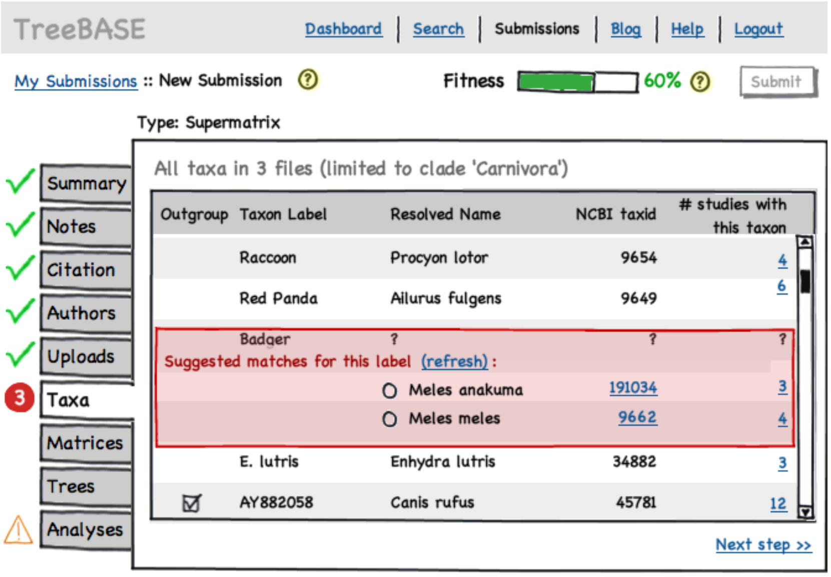
The submission interface to TreeBASE is the gateway through which published phylogenetic analysis results are archived and annotated to allow their reuse by others and their use for improving the Tree of Life. To sustain a rate of deposition that scales to the pace with which phylogenies are published, the submission process needs to meet several principle requirements: it must accept the phylogenetic results that scientists actually produce; provide a set of highly usable tools for data inspection, problem diagnosis, and quality improvement; accept metadata annotations provided by 3rd party tools; and provide a clear indication of data quality issues and how these will affect reusability of the data. The current TreeBASE deposition procedure does not meet these requirements, and places time-consuming burdens on data format preparation, which limits TreeBASE's growth capacity. This is partly because it was designed for a different objective, specifically to guarantee that once a dataset passes submission it adheres to minimal levels of consistency and completeness. While we propose a full redesign of this process and a change of its objective, our goal is to do so in a way that either maintains or improves the quality of the resulting data compared to the existing process. We will achieve this by making it considerably easier for users to meet the desired level of quality through usability-driven design and robust engineering; by providing the user with clear diagnostics of quality issues, and how they will impede reuse; and by archiving, rather than refusing, datasets with problems that the provided tools failed to correct, so that they can be corrected later, either as the tools improve, or through community effort. Level of quality (or, 'fitness for re-use') will be measured by the amount and quality of metadata: resolution of taxon labels to external databases, presence of provenance data as described in the Minimal Information for a Phylogenetic Analysis (MIAPA) [52] and other minimum reporting standards, links to voucher specimens, locality data and GenBank identifiers.

To meet the principle requirements outlined above, we will make changes to the TreeBASE backend and database, create a programmable interface (API) for data ingest into TreeBASE, and implement a new web-based submission tool that uses the ingest API. TreeBASE backend and database changes are necessary to allow TreeBASE to accept the data that scientists produce. Specifically, the data model must be changed to accommodate new analysis methods, such as species trees inferred from sub-sampled gene trees in BEAST [53]), the increasing use of genome-scale data, and studies with a large number of trees (e.g. from a Bayesian MCMC run). In addition, the metadata model needs to be converted to a weakly typed model so that it can accept any annotation. The ingest API will be developed from the existing Dryad-TreeBASE handshaking protocol implementation [54], which is based on BagIt, a hierarchical file packaging format standard [55].

The design of the new web-based submission tool will include a high degree of interactivity, correction of common format errors, diagnostic feedback and progress indicators (see Figure 1). While necessary, these will not be sufficient; how a human interacts with an interface is also strongly influenced by factors such as contextual arrangement of information and navigational flow. This tool will therefore be developed in a highly iterative manner, with usability tests conducted early on with PIs and trainees in the mega-phylogeny research, and later with external scientists in conjunction with the training workshops.

IV. Fast and efficient searching, discovering, and downloading of content

The value of a community resource to users is not only determined by the content itself, but also by how easily content can be discovered, found, and downloaded, both by humans and by machines. The use-case for discovery is fundamentally different from search. In the latter case a user knows the criteria that desired content must match, or even knows that matching content exists. But oftentimes users are not familiar with what content to expect from a resource, or how the content is made searchable, and therefore how to find what might be interesting to them. Neither TreeBASE nor ToLWeb currently provide an interface to browse their content for the purposes of discovering "interesting" things, and the search interface to TreeBASE is slow and assumes that a user knows well how to match their search criteria to the metadata attributes. Downloading data from TreeBASE can be extremely slow or impossible for large datasets, and there is no mechanism to download the tree and annotations for clade pages in ToLWeb.



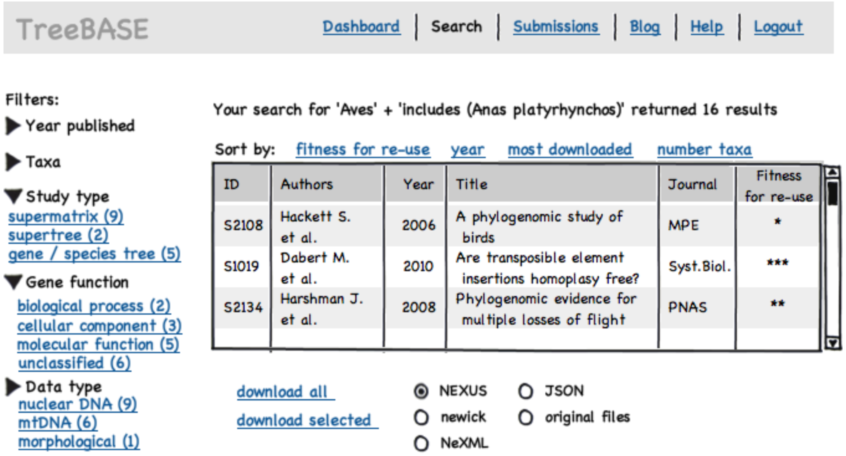
Type: Supermatrix

All taxa in 3 files (limited to clade 'Carnivora')

Outgroup	Taxon Label	Resolved Name	NCBI taxid	# studies with this taxon
	Raccoon	Procyon lotor	9654	4
	Red Panda	Ailurus fulgens	9649	6
	Badger	?	?	?
	Suggested matches for this label (refresh):			
		○ Meles anakuma	191034	3
		○ Meles meles	9662	4
	E. lutris	Enhydra lutris	34882	3
☑	AY882058	Canis rufus	45781	12

Figure 1: TreeBASE submission mockup, showing problem resolution, progress indicators and Fitness for Re-use.

To this end, we will create interfaces designed for usability that will transparently access the content of both TreeBASE and ToLWeb through a single homogeneous interface. Simple search will mirror the prototypical simplicity of the Google [56] search, and accommodate the frequent case when a user does not know which metadata attribute to choose, or where the search term is so specific (such as an identifier) that providing metadata is unlikely to improve specificity. Advanced search will allow users to specify metadata attributes for search terms. For content discovery and narrowing of search results, we will create a faceted browsing interface that allows for exploration based on drilling down along a set of orthogonal metadata (see Figure 2). Faceted browsing has become very popular in applications from internet commerce to library catalogues, and have shown to provide a highly intuitive user experience. Approaches to optimizing their design, layout, and implementation have been well studied [14, 57-60]. These search interfaces will be powered by a document store indexed by Solr [13], a fast



Your search for 'Aves' + 'includes (Anas platyrhynchos)' returned 16 results

Sort by: fitness for re-use year most downloaded number taxa

ID	Authors	Year	Title	Journal	Fitness for re-use
S2108	Hackett S. et al.	2006	A phylogenomic study of birds	MPE	*
S1019	Dabert M. et al.	2010	Are transposible element insertions homoplasy free?	Syst.Biol.	***
S2134	Harshman J. et al.	2008	Phylogenomic evidence for multiple losses of flight	PNAS	**

● NEXUS ○ JSON
○ newick ○ original files
○ NeXML

Figure 2: TreeBASE search mockup, with sorting, filtering and annotations that indicate quality.

open-source enterprise search platform with efficient support for full-text indexing and faceted search. The document store will be regularly updated from the relational data stores native to TreeBASE and ToLWeb. Users will be able to download single TreeBASE studies or all search results in bulk, or the phylogeny and annotations for ToLWeb clade pages. To implement fast download, these functions will also harness the Solr-indexed document store rather than assembling the data first from the relational data stores.

Enabling similarly effective reuse of the data by 3rd party applications requires fast and programmable (API) access to the phylogenetic trees and attached data in machine-readable standard exchange formats (NEXUS [21], NeXML [20], JSON [61]). TreeBASE already has a data access API following PhyloWS [62], an emerging standard for programmatic access phylogenetic data providers. However, the performance of the current implementation can be poor because it queries the relational datastore. We will change this to instead use the Solr-indexed document store, and we will extend the API to expose ToLWeb content as well. To further improve utility for content aggregation applications, which often make heavy use of client-side JavaScript, we will add JSON support. Two applications will validate the effectiveness of the API. Phylocom [63] is a popular comparative phylogenetic analysis package for plant ecologists, and includes the Phylomatic application, which takes an input a list of species, compares the list to a reference tree, and returns a tree that is pruned to those species. The current default reference tree for Phylomatic is based on the Angiosperm Phylogeny Group (APG) III phylogeny [64]. Cam Webb, the author of these tools, has agreed (see letter) to use the API to make the ToLWeb tree available for users beyond the plant community. The second is the navigation of biodiversity data in an evolutionary framework. The Encyclopedia of Life (EOL) [65], a widely recognized biodiversity content aggregator with the aim to provide a web page for every species, and VertNet [66], a distributed database network for natural history collections, have agreed (see letters) to provide the updated ToLWeb tree as a means for users to browse their content phylogenetically.

V. Social computing features and contribution incentives.

The challenges in cyberinfrastructure are as much social as technological, and therefore issues such as incentives, credit attribution, and interdisciplinarity are crucial. We propose to: 1) characterize current knowledge sharing practices in phylogenetics, 2) develop and evaluate social features in ToLWeb and TreeBASE, and 3) leverage activity traces for exploration and navigation.

V.1 Knowledge Sharing Practices in Phylogenetics. Attitudes toward data sharing and reuse vary among disciplines depending on both the context of production [67] and use [27]. Even in fields like genomics where data sharing is the norm, data sharing can be problematic [68]. So that our development work is based in stakeholder needs, during year 1 we will undertake a study of knowledge sharing practices and attitudes in the phylogenetic community. This study will have three components: an online survey; qualitative interviews and observations; and analysis of activity within TreeBASE and ToLWeb. The *survey* will be uncover current knowledge sharing attitudes and social concerns across a wide variety of expected stakeholders of TreeBASE and ToLWeb. It will extend findings from a recent general data sharing survey [69] by providing field-specific insights in order to inform the development of TreeBASE and ToLWeb. The survey will be supplemented with *qualitative interviews and observations* to reveal scientific practices. Interviewing scientists and conducting observations will provide detailed and useful information about how data flows from the field through the lab and into publications and repositories. We expect to conduct 40-60 hours of observations and interviews in two to three laboratories, and also interview additional individuals about specific areas of concern identified in the survey. Lastly, we will explore anonymous log files, TreeBASE helpdesk requests, and public content in TreeBASE and ToLWeb. This will provide an understanding of how the site is currently used, especially most common communities, current frustrations with the interface, and temporal, geographic, and individual patterns of participation. This will also provide an evaluation baseline as we move forward with the project.

V.2 Developing Social Features in ToLWeb and TreeBASE. This is a key goal of this proposal. While we hope that including social features will make these resources more enjoyable to use, our intent is to leverage social interaction to meet the challenges of large-scale, distributed, data-intensive science. We will focus our development on three main concerns: 1) providing a comprehensive collection of reusable data; 2) ensuring high quality data and metadata; and 3) maintaining and sustaining a curated community-owned resource. In order to build a comprehensive database, we will develop social incentive systems to motivate scientists to submit data and metadata. Status indicators and social comparisons can be very powerful motivators for contributions in online communities [70]. Other incentives will rely on leveraging existing reward systems in science. Highlighting individual contributions makes it possible for scientists to better document their own participation for tenure and promotion reviews.

In order to encourage high quality content, we will develop annotation and discussion features around the data. While we will maintain the integrity of original submissions, we will also enable collaborative development and discussion of content [71]. These features allow value to be added to existing data records, including allowing others to fill in missing or community-specific metadata such as a teacher annotating a tree with links to online lesson plans; and documenting conversations as contextual information for later users. Revealing this activity around trees can also provide an important motivator for scientists to contribute their data to the system. To support community ownership and maintenance of the resources, we will develop functionality to create groups and join communities of interest around particular clades or methods. These groups can engage in discussion, “adopt” particular trees for curation, and be notified about changes or open issues with community-specific resources.

V.3 Leveraging Activity Traces. The digital traces left by users (e.g. logs of tree views, downloads, comments, etc.) reveal social patterns, and we will develop functionality that uses these to support navigation, exploration, and serendipitous discovery. We hypothesize that these data can be used to generate scientifically meaningful relationships and reveal trends in scientific interests. Knowing, for example, that two trees are frequently viewed together may reveal a non-obvious similarity or contrast.

V.4 Evaluation. We will employ mixed methods including surveys, interviews, observations, and log file analysis to evaluate the success of the social features. New methods like "digital trace ethnography" can help to understand how practices change in response to technological features [72, 73]. We also intend to conduct surveys with users to investigate satisfaction with the system and the impact of new features. The hackathons and training sessions will provide feedback through interviews and usability tests.

VI. Community building and support.

The long-term sustainability of TreeBASE and ToLWeb are highly dependent on active user, contributor and developer communities. To actively nurture and build these communities, we will establish a dedicated staff to provide community coordination and user support functions, including triaging of trouble reports; documentation for users and developers; training; content curation; advocacy and outreach at conferences. These functions can be remarkably effective at creating and growing a healthy community, as has for example been demonstrated by the success of the Generic Model Organism Database (GMOD) [74] HelpDesk [75], an NIH-funded position with similar responsibilities and managed at NESCent. Since the GMOD HelpDesk was created in 2007, a variety of metrics, including mailing list traffic and software downloads, show a steadliy growing and healthy community across GMOD’s array of sub-projects. Here, the role will be split between Katja Schulz, former managing editor for ToLWeb and current Species Pages Coordinator for the Encyclopedia of Life, and William Piel, one of the original TreeBASE authors and current monitor of the TreeBASE helpdesk.

We will also hold engagement activities to provide outreach to the community and obtain feedback for the project. The ***Stakeholders meeting*** in year 1 will broaden awareness of the project and ensure that our planned implementations meet current and forthcoming needs. Attendees will be recruited through a mixture of direct invitation and an open call for participation, and will include current developers, users and contributors as well as new potential collaborators, such as comparative biologists and those working in clades with a comparably low number of contributors (e.g. environmental genomics projects such as iSEEM [76]). Two ***Training and usability workshops*** will train producers (e.g., systematists) and users of trees (e.g., comparative biologists) on the new submission, search and browse interfaces. The first will be held in Year 2 co-localized with the Evolution 2013 meetings. The second, held in Year 3 at the Society for Integrative Biology (SICB) meeting, will also include training on the ToLWeb social curation features. At both conferences, we will arrange short usability testing sessions with select participants to quantitatively assess the new interfaces.

We will hold also two ***hackathons**,* intense face-to-face hands-on collaborative coding events [77] that bring together developers with diverse backgrounds and affiliations to work face to face on shared objectives. The first is planned for year 2 to engage developers of online resources that incorporate phylogenetic data, such as Phylomatic [63] or TimeTree [78], or comparative analysis tools such as the phylogenetic analysis packages in the R statistical analysis system (see [79]) in using the new search and download APIs of TreeBASE and ToLWeb. A second hackathon in year 3 follows maturation of the data submission API for TreeBASE, and will engage developers of phylogenetic data management, including Mesquite [80], and inference software in developing means to format and submit data directly from the software that researchers use to produce phylogenetic data. For both events, a core group of attendees will be directly identified by involvement with key software projects, but the majority of attendees will be recruited through open calls for participation, which has proven an effective mix at previous NESCent hackathons towards achieving both productivity and outreach.

Driving research: Synthesizing mega-phylogenies from research trees

As an integral part of our proposal, a set of domain-driven research activities on synthesizing mega-phylogenies from smaller research trees is fully aligned with our development plan. The research addresses key questions in the evolution of diverse phylogenetic groups of high interest, with the following goals: (a) directly inform, validate, and drive the development outcomes on a continuous basis, specifically the new data deposition, curation, search, synthesis and collaboration features; (b) demonstrate that these features advance our ability to build a synthetic Tree of Life through data reuse; (d) show how these new features inform future research by identifying knowledge gaps; and (e) engage large existing research communities in the proposed platform. To accomplish these, we have selected two very large branches of the Tree of Life: all fishes at the family level (~32,000 species) and all asterid flowering plants (Asteridae, ~65,000 species). We will engage the existing community of fish phylogenetics researchers of DeepFin (a highly successful NSF RCN team led by G. Orti), plant researchers from the Angiosperm AToL group, and the 2008 BioSynC “Mega-phylogeny Assembly by Literature-mining and Grafting” [81] workshop to participate in testing the proposed user interfaces, social networking tools, and interoperability links between TreeBASE and ToLWeb.

As part of the mega-phylogeny synthesis research, we will seed TreeBASE with up to 500 published, highly-annotated tree topologies for fishes and asterids, including, wherever possible, the datasets upon which they are based. Literature searches indicate approximately 2700 phylogenetic publications on fishes from the past 30 years, with about 1000 high-value studies (in terms of coverage and data richness). For asterids, there are currently 353 studies in TreeBASE, summing to more than 8,500 distinct species. We will submit at least 200 additional studies, including 124 that have already been assembled as part of a preliminary effort at Yale to build a campanulid tree with 4995 species. This phase will inform development of the TreeBASE submission interface (deliverable III). For mega-tree assembly we will use prune-and-graft operations, which have been shown as a rapid and effective mechanism for creating a synthetic view of the Tree of Life [82-85]. Our preliminary efforts have produced a grafted mega-tree of all fishes at the family level [84], containing over 850 fish families. This backbone structure will be migrated to ToLWeb, as will our preliminary assembly of an asterid tree including 4995 species in the campanulid branch and 3894 species in the lamiids. In the course of this work, we have assembled provisional protocols for mega-tree construction, which will be further refined and inform the development of the tools for merging TreeBASE trees into ToLWeb (deliverable II). After the initial deposition, our mega-trees will be constantly updated as new studies are added to TreeBASE, which will provide rigorous testing of new features for cross-linking between ToLWeb and TreeBASE (deliverable I), for incorporating new research trees into ToLWeb (deliverable II), and for identifying topological incongruence with the synthetic tree (deliverable I). We will also improve the quality of the TreeBASE submissions through additional metadata and annotations. For example, we will link fish phylogenies to phenotypic and genotypic data via taxonomic names and gene ontology terms in FishBase [86] and Phenoscape.

Finally, we will demonstrate the power of a synthesized mega-phylogeny, and thus of well-populated branches of the ToLWeb Tree of Life, in examining challenging evolutionary questions across our target clades. We will address questions pertaining to evolutionary patterns in functional trait, character, and flower evolution (see e.g., [33]), and also questions that begin to assess the phylogenetic knowledge itself, such as how effective the distribution of fossil taxa with fishes is for developing a time tree; and how research effort is distributed across the trees, and whether individual researchers or large collaboratives have the largest impact on our phylogenetic knowledge.

MANAGEMENT PLAN

Specific Aim	Responsibility				Timeline of Activity											
	Lapp, Cranston	Donoghue, Piel	Westneat	Bietz	Year 1				Year 2				Year 3			
I) Linking TreeBASE and ToLWeb content	■	▨				▨	▨	■	■	■	■	▨	▨	■	■	▨
II) Tools for merging trees into ToLWeb	■	▨	▨				▨	▨	■	■	■	■	■	■	▨	▨
III) Usability-driven submission interface	■	▨	▨	▨	▨	■	■	■	■	■	▨	▨	▨	■	■	▨
IV) Faceted browsing, fast download	■	▨			■	■	■	■	▨	▨	▨	■	■	▨		
V) Social computing, contribution incentives	■			■	▨	▨	▨	■	■	■	■	■	■	■	■	■
VI) Community building and support	■	▨	▨		■	▨	▨	■	■	■	■	■	■	■	■	■
Driviing research on synthesis of mega-trees		■	■		■	■	■	■	■	■	■	■	■	■	■	■
Migrate hosting, DataONE member node	▨	■									▨	▨	■	■	■	■
Training Workshops	■	▨		▨					▨	■	▨	▨	■	▨		
Hackathons	■	▨								▨	■	▨	▨	■	▨	▨

Dark shade: primary responsibility or activity; Light shading: secondary responsibility or activity.

A. Responsibilities and timelines. The team assembled here combines deep expertise in several areas that are all critical to the outcomes of the project. Responsibilities are distributed accordingly. Lapp has extensive experience in evolutionary bioinformatics, promoting open and collaborative software development practices, and managing globally distributed software projects. He will oversee and coordinate all software development, project communication, reports, and outreach. Project staff located at NESCent, under the direction of Lapp, will have primary responsibility for the development of all user-facing software components. Lapp will also be responsible for directing the Community Coordinator and ToLWeb Helpdesk (see deliverable VI for responsibilities). Cranston, a trained evolutionary scientist, will coordinate the community and stakeholder engagement activities, and will be specifically responsible for translating stakeholder-driven use-cases into software requirements and priorities. Piel is one of the co-founders of TreeBASE and has been overseeing the resource as well as served as its Managing Editor for more than a decade. He will continue this role, and also direct the programming staff located at Yale, which are responsible for all backend, long-term hosting, and data permanence work (see Sustainability Plan) on TreeBASE and ToLWeb. Donoghue and Westneat will jointly lead the guiding research project on synthesizing mega-phylogenies from smaller research trees, and will be responsible for informing software and method development on an on-going basis from the results. Bietz has extensive experience studying the use and development of cyberinfrastructure for scientific collaboration, and will use the findings from that to inform the social computing and incentivizing contribution goals of the project. He is responsible for identifying, designing, and evaluating the social capabilities to be developed for TreeBASE and ToLWeb, and he will also oversee and evaluate the usability testing activities.

One of the overarching principles behind our development approach is to provide value to users early and often throughout the project. The removals of major barriers to using the resources are released first (deliverables III and IV: first releases in Q3 of year 1 and Q2 of year 2, respectively). Otherwise, major release milestones of all user-facing features (deliverables I,II,V) are driven by and coincide with the training workshops (Q2 and Q1 of years 2 and 3, respectively), so that users external to the project can be engaged in their further refinement right away. Major milestones for data access and ingest APIs are similarly driven by the hackathons in year 2 and 3. All deliverables will receive renewed development effort following outreach events to respond to user feedback and the results of usability tests.

B. Advisory Board. The Phyloinformatics Research Foundation (PRF) [87] was formed in 2010 with the purpose of providing governance, direction, and stewardship for both ToLWeb and TreeBASE under a single umbrella. Several of the PIs and senior personnel also serve on the PRF Board, which ensures direct communication between the PRF and the development team on project progress and long-term vision. In conjunction with the PRF, a complementary Advisory Board of four members will meet annually to advise on project direction, its utility to stakeholder communities not represented among the PRF, and on compatibility with emerging technological trends. Committed members (see letters) include Jamie Taylor (Google; co-founder of Freebase [88]) and Sam Donovan (U. Pittsburgh, expert on teaching tree-thinking). Additional members will be recruited when the project is funded; candidates include a publisher of biodiversity or evolutionary titles, and an expert in usability or information visualization.

C. Project Coordination. Central coordination of activities (see A. above for responsibilities) and frequent communication are key to mitigating the risks from multiple institutions collaborating with distributed responsibilities. The project will start with an annually held All-hands face-to-face meeting of all project personnel to foster team cohesion, to review goals and past progress, and to establish a shared understanding of work plans, milestones, and priorities for the following 12 months. The Advisory Board meetings, which includes all project PIs, and the PRF Board meetings, which includes several project personnel, will be held back-to-back with the project's All-hands meetings to minimize travel. Outside of the annual meetings, the senior personnel of the project will have biweekly teleconferences to discuss progress, obstacles, and short-term priorities. The personnel responsible for software development and community coordination, located or managed at NESCent and Yale, will continuously discuss all design, implementation, and trouble report questions on public email groups. They will also use social network-enabled tools, such as micro-blogging sites (e.g., Twitter [89]) to reach out frequently to the wider evolutionary and biodiversity informatics community, allowing the coordination with related efforts to emerge where useful. The team will support the incorporation of contributed code through continuous integration and unit testing. Where appropriate, we will adopt agile methods for rapid design and iteration of minor features, while keeping APIs and complex workflows stable for existing users. We will use the GitHub wiki and issue tracker features for collaborative documentation of information for developers (software architecture and deployment, APIs) and users (manuals, help pages) as well as tracking bug reports and feature requests.

The senior personnel of this project have been involved in collaborative research and development before, but the team as such has not worked together before. To mitigate inherent risks, a team led by Dr. Gary Olson (University of California, Irvine) will evaluate our collaboration using their Collaboration Success Wizard (CSW) [90] (see letter). The CSW is a web-based survey based on over twenty years experience studying scientific collaborations. It identifies potential vulnerabilities and suggests strategies to improve our collaborative process. We will complete the CSW at the beginning of the project, and again near the midpoint to gauge improvement from implementing the suggested practices.

D. Collaboration and coordination with other projects. Several of the PIs are involved, often in leading roles, in a number of cyberinfrastructure and standards initiatives for evolutionary science, and will provide direct connections to and coordination. **NeXML** [20], the Comparative Data Analysis Ontology (**CDAO**) [91], and **PhyloWS** [62] are emerging standards for the exchange, semantics, and programmatic access to evolutionary data. The Minimum Information for a Phylogenetic Analysis (**MIAPA**) [52] standard tries to identify the metadata attributes required to make a tree or analysis reusable. To inform these standards from biodiversity science requirements, the **Phylogenetics Standards Interest Group** [92] of **TDWG** reaches out to biodiversity information scientists. **Phenoscape** [93] transforms natural language phenotype descriptions from the literature into ontological expressions with fully computable semantics [94-96]. The Scientific Observations Network (**SONet**) [97] brings together a wide range of efforts to standardize the semantics of scientific observation data, including morphological character observations. The Data Observation Network for Earth (**DataONE**) [98] provides sustainable preservation of and universal access to data from a virtual network of repositories in the earth sciences, ecology, and evolution. **Dryad** [24] is a digital repository for data underlying scientific publications, and is a DataONE member node. Generic Model Organism Database (**GMOD**) [74] is a consortium of intercompatible software components for managing genome-scale biological data, including phylogenies. The **iPlant Collaborative** [99] develops scalable cyberinfrastructure for solving grand challenge problems in plant science, such as a mega-phylogeny of all plants. The Encyclopedia of Life (**EOL**) [65] aggregates biodiversity information across all of life from a huge number of partner projects (including ToLWeb). Large-scale biological systematics research networks, including the Cypriniform and Angiosperm **AToL** projects, and **DeepFin** [100], assemble data and develop methods for phylogenetic inference.

E. Dissemination plan. All software will be developed in the form of distributed collaborative open-source projects, with source code available immediately under an OSI license (specifically, a modified BSD license [101]). TreeBASE has been publicly hosted on SourceForge since 2010 [102]. Both projects will be moved to Github for source code hosting to better accommodate the social aspects of collaborative software development and micro-contributions. We will create public email groups for developers and users of ToLWeb (they exist already for TreeBASE). For the dissemination of project progress and content updates, we will convert the existing ToLWeb news page [103] to a blog, dubbed PhyloCommons, and regularly post updates. As a result of the development goals, the content of TreeBASE and ToLWeb permitted for reuse (TreeBASE studies marked published by depositors, Creative Commons-licensed content in ToLWeb) will be available for bulk download in standard formats (NEXUS, NeXML, and Newick).

BROADER IMPACTS

For **biological research**, the proposed platform will greatly enhance the digitization of phylogenetic data by providing a means to easily disseminate research data in a way that provides measurable scientific rewards. The lack of existing resources for comprehensive phylogenetic data cannot be overstated. The **connections to biodiversity resources** will provide the long-awaited comparative framework for scientific users to browse content and can help non-scientist users to understand the shared evolutionary history of species in these databases. **Social cyberinfrastrucuture development** is not only an important outreach component of the project, but is also designed to advance the understanding of how informatics resources allow people to collaborate more effectively. We will **engage developer communities** through hackathons that explore novel uses and implementations of the APIs. New student developers will be introduced to these phyloinformatics resources through NESCent's participation in the Google Summer of Code program [104]. Our **cross-disciplinary training programs** will involve postdocs and graduate students, as well as a significant number of undergraduate students, focusing on historically underrepresented groups. The **educational impact** of ToLWeb will be enhanced through more complete coverage, richer content and simpler interfaces. User surveys indicate significant use of ToLWeb in the home schooling community as well as traditional K-12 classrooms, making this a key resource for teaching evolution and tree-thinking.

SUSTAINABILITY PLAN

Identifying long-term strategies for TreeBASE and ToLWeb to support future innovation as well as sustain a baseline of software maintenance is one of the main tasks of the recently formed PRF, the umbrella foundation to ToLWeb and TreeBASE (see Advisory Board). As annual meetings of the PRF are already supported through 2014 (by NESCent), we focus our sustainability plan on two main objectives. First, we aim to nurture a growing community of participating users, developers, and contributors who share ownership in the resources and their content. The social capabilities that we will create for these resources (deliverable V) and the community building, training, and engagement activities (deliverable VI) are designed to accomplish this, as is the move to of source code hosting to Github [105].

Second, we undertake steps to ensure the continued availability of TreeBASE and ToLWeb after the end of this project and to permanently preserve their data holdings. The Yale Peabody Museum has made a commitment (see letter) to serve as the institutional home for both TreeBASE (currently hosted at NESCent) and ToLWeb (currently hosted at U. Arizona) for 5 years after the end of funding from this grant, by providing the required physical and human resources. Both resources will be migrated to their new host environment at the Peabody Museum starting in Year 2. To ensure permanent preservation of their data, including protection from corruption, we will register TreeBASE and ToLWeb as “member nodes” with the Data Observation Network for Earth (DataONE) [98], a federated network of data repositories in the earth sciences, ecology, and evolution. The DataONE architecture consists of a small number of “coordinating nodes” (CNs) that transparently federate search, discovery, and data access across the network, and a growing number of data repositories that function as member nodes (MNs). To register TreeBASE and ToLWeb as MNs, we will implement the required DataONE MN AP [106]. DataONE would welcome TreeBASE and ToLWeb as new DataONE MNs, and supports candidate MNs in joining the network through training workshops and a helpdesk (see letter from W. Michener). As a significant byproduct, becoming a DataONE MN would also make TreeBASE records discoverable across DataONE’s network, which includes a far larger set of biological domains than the scientists who are typically aware of TreeBASE (or ToLWeb).

RESULTS FROM PRIOR NSF SUPPORT

Hilmar Lapp is co-PI on NSF DBI-0753138, “INTEROP: Creation of an international virtual data center for the biodiversity, ecological and environmental sciences”, University of New Mexico, $50,803 to Duke University, 7/15/08 – 6/30/12. He co-leads the design and implementation of the project’s remote summer internship mentoring program for students, which was mirrored after the Google Summer of Code™ program, and is a member of DataONE’s Data Integration and Semantics Working Group. 3 annual remote internship summer programs so far; mentored or co-mentored 2 remote student interns, one on semantic phyloinformatics web-services, and one on exposing data repository holdngs to the Linked Open Data cloud. H. Lapp is co-PI on NSF DBI-1062404, “ABI Development: Ontology-enabled reasoning across phenotypes from evolution and model organisms”, University of South Dakota, $479,175 to Duke University, 7/1/11 – 6/30/15. The project builds on NSF DBI-0641025 “Linking evolution to genomics using phenotype ontologies”, 6/07 – 6/11, on which Lapp was Senior Personnel overseeing the software development and implementation of a knowledgebase for semantic integration of phenotype observations. Produced and contributed to several knowledge organizations systems, including the Teleost Anatomy Ontology [107]; Phenex software for data curation [108]; and the Phenoscape Knowledgebase [109]. Three publications so far, plus one in revision and two in preparation. **Michael Donoghue** is PI on DEB-0431258, “AToL Collaborative Research: Resolving the Trunk of the Angiosperm Tree and Twelve of its Thorniest Branches.” 8 institutions, $1,015,267 to Yale. 9/1/04-

8/31/10. His lab resolved the Campanulidae using ~25,000 bp of DNA sequence from 122 taxa. The Yale team developed TOLKIN (www.tolkin.org), featuring online access to taxonomic, geographic, specimen, and sequence data; an exhibit, “Travels in the Great Tree of Life,”, developed at the Yale Peabody Museum; three postdocs and one graduate student were funded; 21 publications. M. Donoghue is PI on IOS-0842800 “Collaborative Research: The evolution of leaf form in Viburnum (Adoxaceae).” 3 institutions, $210,952 to Yale. 6/1/09-5/31/12. His lab is expanding molecular phylogenetic analyses of Viburnum to provide a framework for analyzing the evolution of leaf form and physiological ecology. Sampling has increased from 40 to 115 species, and from 5 to 10 DNA markers. One postdoc, one graduate, and one undergraduate have been funded. Two oral presentations and one poster at meetings; 5 publications to date. **William Piel** received a subcontract on DBI-0743720: “A digital repository for preservation and sharing of data underlying published works in evolutionary biology.” 5 institutions, $280,000 to Yale. 6/1/08-8/31/11. Building a handshaking protocol between Dryad and TreeBASE, exposing TreeBASE metadata using OAI-PMH, and improving metadata in TreeBASE. The first two activities are complete. Piel is Co-PI on IOS-0818731: “The evolution of serial homology: a case study with nymphalid butterfly eyespots.” PI: A. Monteiro. $540,000. 7/1/08-6/30/12. Building / populating core IT infrastructure, including a web application and database that tracks butterfly taxa (6,900 species), specimens (9,120 records), images (14,853 photos), and scoring of wing pattern elements (~4,000 so far). **Mark Westneat** is PI on DEB-0844745, Phylogenetic Relationships and Evolution of Skull Mechanisms in Perciform Coral Reef Fishes. 3/1/2009- 2/28/20012; $412,000. Major progress is being made in our understanding of phylogenetic relationships among coral reef fishes and higher level percomorph fishes based on molecular analysis, including a large megatree of all fish families. 18 publications, 2 Ph.D. theses, and 3 current Ph.D. candidates; fostered collaboration with 6 colleagues, 4 postdocs, 3 graduate and 3 undergraduates. Developed several open source software applications for biomechanics and phylogenetics and coral reef games in virtual worlds.