INTEROP: A network for enabling community-driven standards to link evolution into the global web of data (EvoIO)

Overview: Achieving Interoperability

The explosive growth of stockpiled information in the biological and earth sciences presents a wealth of opportunities for expanding bioinformatics-based analyses with respect to both the amount of data incorporated and the diversity of data types and sources to be integrated. While large-scale or integrative analyses of such data may use generic methods of machine learning, there is a theory-based comparative approach to the analysis of diverse types of biological data, in which the similarities and differences between compared things are interpreted as *evolved differences* that have arisen by a process of descent-with-modification from common ancestors. This evolutionary comparative approach, used throughout biology and paleobiology, depends fundamentally on "phylogenetic trees" representing paths of descent. While powerful tools exist for the inference of phylogenies, and while evolutionary approaches are increasingly recognized as effective, the lack of interoperability in tree-based data and services hinders large-scale and integrative analyses.

The over-arching goal of this project is to create a long-term Data Interoperability Network, EvoIO, focused on the phylogenetic trees that are used throughout biology, both as an organizing framework for knowledge (classification), and as the formal basis for rigorous methods of comparative analysis. Phylogenetic trees become useful only to the extent that they are attached to data and metadata. For instance, when trees are used in classification, taxonomic affiliations such as species names are a vital information. Other types of information include molecular or morphological data used to infer the tree, comparative data used for tree-based evolutionary inference, the sources of data or samples or procedural notes about analyses and workflows. The ability to carry out integrative analyses using such data is hindered by the lack of standards and conventions for representing and processing the data and metadata. The proposed data interoperability network will lower the barriers to data re-use and repurpose, and facilitate integrative analyses by developing and disseminating interoperability standards.

The approach advocated in this proposal is an incremental, organic, and adaptive approach based on previous work of the leadership team, most of whom were heavily involved in the work of the Evolutionary Informatics working group supported by NESCent (National Evolutionary Synthesis Center, an NSF Center). From 2006 to 2009, the EvoInfo working group oversaw development of the "EvoIO Stack", an integrated technology toolbox for addressing interoperability challenges in phylogenetics and evolutionary biology. The group used a 2009 "database interoperability hackathon" in which scientist-programmers were challenged to apply the EvoIO Stack to community data resources. Not only has the "hackathon" approach generated success stories that improve interoperability and stimulate further work, it also has empowered early-career scientists with the know-how and the connections to make ongoing impacts on interoperability. This proposal is a logical continuation and generalization of the effective network-building encouraged by hackathons, as the idea was conceived during follow-up activities from the 2009 hackathon hosted at NESCent, which was attended by the majority of PIs and Senior Personnel listed in this application.

The EvoInfo group, rather than attempting to assemble all of the stakeholders for consultation at the beginning of the process, started with a small group of stakeholders and built an initial implementation and relevant tools. Now, we are successively widening the scope to engage additional stakeholders, who will contribute valuable information for continued development and documentation of the standards. The EvoInfo group was established in 2006 by stakeholders mainly from the areas of phylogenetics, molecular evolution, and comparative genomics. By the time of the interoperability hackathon, the participants included substantial numbers of representatives from other communities, including biodiversity, paleobiology, and computer science. The EvoIO Network proposed here ultimately will include stakeholders from the various subdisciplines that use trees, including molecular evolution, phylogenetics, comparative genomics, biodiversity, phylodiversity, paleobiology, infectious diseases, and evolutionary ecology.

Results from Prior Support

NESCent Evolutionary Informatics Working Group (R. Vos and A. Stoltzfus, Co-PIs). Sept. 2006 to Apr. 2009. The National Evolutionary Synthesis Center (NESCent), an NSF Center (NSF 0423641; K. Smith), funded Vos and Stoltzfus for an Evolutionary Informatics Working Group. The working group had a flexible membership with a mailing list of 20 to 30 people, and meeting attendance of 8 to 10 non-NESCent personnel. The group helped to organize the 2006 Phyloinformatics Hackathon (NESCent's first hackathon), held 3 regular meetings in 2007 and 2008, and collaborated with NESCent staff to make its fourth meeting another hackathon (see below). This working group oversaw the development of the group of standards that will form the seed of the EvoIO Stack (NeXML, CDAO and PhyloWS), along with an informal metadata standard based on RDFa.

NESCent Phyloinformatics Hackathon (organizing team: H. Lapp, T. Vision, A. Stoltzfus, R. Vos). Dec. 2006. This hackathon supported by NESCent (NSF 0423641; K. Smith) brought together 25 participants— including programmers, documentation-writers, and end-users— to focus on improving support for phylogenetics in Bio* programming toolkits (BioPerl, BioJava, etc), and resulted in a multi-author publication [LBB+07]. Results from the hackathon include a Phylip (file format) parser in BioJava and BioRuby, an extensive set of conformance test-files for NEXUS, and new BioPerl components enabling workflows to analyze molecular evolution. NESCent's participation in the Google Summer of Code^{TM} program as a mentoring organization (organized and administered by H. Lapp) started in 2007 as a spin-off from this hackathon. The program pairs undergraduate and graduate students with senior developers from open-source software projects who act as mentors for a 12-week long remote programming internship. NESCent has meanwhile participated for three consecutive years.

NESCent Evolutionary Database Interoperability Hackathon (organizing team: H. Lapp, A. Stoltzfus, R. Vos, T. Vision, K. Schulz). Mar. 2009. This hackathon funded by NESCent (NSF 0423641; K. Smith) brought together 6 developers representing the EvoIO stack (Vos, Lapp, Pontelli, Stoltzfus, B. Chisham, R. Scherle), 17 programmers representing diverse data resources and applications, and 2 dedicated documentation-writers. Considerable effort went into planning the hackathon. After identifying 32 candidate data resources as strategic targets for interop development, the organizers requested applications and recruited the participants, most of whom had no prior association with the EvoInfo working group. At the hackathon, participants self-organized using an Open Space [OSP] approach, resulting in 5 sub-groups. Each subgroup generated a working software product to demonstrate desired interoperability improvements:

- using NeXML and CDAO, the Semantic Processing subgroup used advanced language technologies to populate a "triple store"— a logic database of subject-predicate-object triples— of phylogenetic data
- using NeXML and PhyloWS, the Phylogenetic Visualization subgroup expanded the capabilities of PhyloWidget to represent tree annotations and to retrieve and display images
- the Java API for NeXML group created a programmable Java interface to NeXML
- using NeXML and PhyloWS, the Taxonomic Intelligence group implemented a web services interface to TreeBase content
- using NeXML and PhyloWS, the Phylr group created a modular system to generate databases whose phylogenetic content can be accessed via PhyloWS

The hackathon was successful in raising the profile of NeXML, CDAO and PhyloWS, and in demonstrating that, with a modest amount of training and effort, data providers can improve interoperability, with benefits for data providers and for end users.

Phenoscape: Linking evolution to genomics using phenotype ontologies. NSF BDI-0641025. (PIs P. Mabee, T.J. Vision, M. Westerfield, Senior Personnel H. Lapp). $1,050,945. 06/01/2007-5/31/2010. The project aims to transform free-text descriptions of morphological characters found in the systematics literature to a fully computable format with rich semantics that is based on ontologies and first-order logic (Entity-Quality (EQ) syntax), with an initial focus on Ostariophysi, a teleost group of fish that includes the zebrafish. This will allow integration of data between the vast stores of data about evolutionary phenotype diversity and the growing body of mutant phenotype data generated from genetically characterized and extensively studied model organisms, two previously disconnected domains of knowledge. Embedding semantically rich phenotype descriptions in data files of phylogenetic character state matrices is a central requirement of the project, is not met by previously existing standards such as NEXUS, and has already driven the development of the NeXML format as well as the CDAO ontology.

TOLKIN (Tree of Life Knowledge and Information Network) NSF-DEB 0827609, Collaborative Research: EuphORBia - a global inventory of the spurges (Cellinese Co-PI). $622,000. 10/01/2006-08/31/2011. (also supported by NSF-DEB 0827254, NSF-DEB 0829313, and NSF-DEB 0431258). All of the above funded projects generate rich datasets in a variety of forms. Cellinese is co-leading the development of TOLKIN [TOL], an information management and analytical web application that provides support for phylodiversity and biodiversity research projects. As a web application, collaborators in different labs and using a variety of environments can access and manipulate data in real time. Supported shared data include taxonomy, voucher specimens, morphology, DNA samples and sequences, and bibliography. Many other resources can be accessed and queried through TOLKIN, e.g., IPNI, TROPICOS, GenBank, and TreeBASE, among many others. In additions, a few tools have been integrated, e.g., BioGeomancer. A developing analytical component of TOLKIN includes workbench functionality for analysis of sequence data by automating the assemblage of FASTA files, alignments, BLASTing capability, scoring of morphological characters,and output/input of Nexus files, among others. TOLKIN has been developed in PHP and PostrgreSQL, and more recently Ruby-on-Rails.

AToL Interoperability NSF-IIS 0840702. Collaborative Research: Core Database Technologies to enable the Integration of AToL Information (Cellinese Co-PI). Cellinese is addressing the needs to integrate various phyloinformatics tools developed to facilitate AToL analyses. She has a prototype implementation of Kepler workflows evoked while embedded in TOLKIN.

Database Workshops NSF-BDI 0402795. A Workshop on Establishing a Comprehensive Database for Plant Systematics (Cellinese Co-PI). $50,000. 02/01/2004-01/31/2006. This workshop is one of several that brought together the systematics and informatics community and initiated a productive dialog, outlining broad user requirements for integrating an array of independent, distributed biological resources. Co-PI Cellinese has extensive experience in interacting with the user community and serving as a bridge between research challenges and needed infrastructure. As a result of this workshop, TOLKIN became the core of several AToL successful proposals.

CREST Center for Bioinformatics NSF-HRD 0420407. The grant funds the development of a Center for Research Excellence in bioinformatics and computational biology at NMSU (Pontelli PI; $4,500,000; 08/04-07/10). The Center is aimed at creating a core entity to coordinate research in bioinformatics across NMSU, with emphasis in areas of research like protein structure determination and discovery of gene regulatory networks. The Center is leading the creation of a graduate program in bioinformatics, and it promotes community-wide outreach efforts to develop knowledge and understanding in the areas of computational and biomedical sciences.

Background and Rationale

Phylogeny as a central and powerful principle across science

Trees play two central roles in modern biology: organizing knowledge by lines of descent, and extending knowledge through comparative analyses. Phylogeny organizes our vast, disparate, and constantly increasing data and knowledge about the past and present diversity of life on earth in a structure that has tremendous explanatory as well as predictive power. Molecular systematics has made exceptional progress in reconstructing the Tree of Life through both methodological and computational advances. This has enabled the recent and expansive growth of the second role of phylogeny, as the central tool for constructing and testing predictions and hypotheses in comparative biology. Knowledge about phylogenetic relationships informs our understanding about the patterns and processes underlying past and present biodiversity, and how diversity might change in the future in response to changes in climate, habitat, and community composition. When entities under comparison (e.g., proteins, genomes, species) are related by descent, evolutionary theory, through phylogenetic trees, provides a framework that allows for comparisons of entities while controlling for non-independence due to relatedness.

Trees are used in a wide variety of research fields and disciplines. Studies of molecular evolution have long depended on trees for a broad array of analyses such as detecting positive selection [JWA07], estimating divergence times [DB07], delimitation of species [Wie07] or estimation of diversification and extinction rates [Ric07]. Phylogenies are increasingly applied to other biological fields such as conservation biology [FO04], comparative genomics [Con07, Ell08], community ecology [WAMD02, CBKFK09] and metagenomics [WE08]. Comparative phylogenetic analyses have the potential to have enormous impact in even more diverse research areas such as paleobiology [Mac01] or epidemiology [STZ+05, SVB+09], linking evolutionary biology with the earth sciences and medical communities. Extant biodiversity is only a small fraction of the total diversity of life that has existed on earth, and understanding the patterns and constraints evident from diversity changes in response to historic events of climate change will likely prove key to forecasting future changes. Evolutionary epidemiology uses lines of descent to organize, track, and predict mutational trajectories of disease agents, incident locations, and clinical symptoms. Realizing the full potential of repurposing phylogenies across research communities and disciplines on a broad basis will require interoperability and accessibility of phylogenetic data, as well as interpretability of their meaning.

This significance of a tree depends on its data and metadata "decorations". A tree structure with no discernible content other than its structure, e.g., (w,(x,(y,z))), is of little use and cannot be re-purposed or integrated. Trees become useful when they organize data and metadata. For instance, a simple use case is attaching pictures or text to the tips of a tree, which can be an invaluable teaching tool for visualizing the shared ancestry of biodiversity. If applied on a broad basis, even such a simple annotation requires interoperability standards to automate the process of interpreting labels on the phylogeny, reconciling them with reference taxonomies, and locating corresponding image data. Trees, including their subtrees, nodes, branches, and tips, can be associated with various kinds of other data or metadata. Relevant use-cases encompass tasks such as quantitating data availability (is there sufficient density of a specific type of metadata to perform a phylogenetically-informed analysis), searching for phylogenies that relate a collection of species with metadata, visualizing large-scale data patterns across a phylogeny to enable hypothesis generation and testing, describing the methodology used to produce a phylogeny, or displaying results from an analysis in a phylogenetic context. The data and annotation to be associated may be biological (molecular sequences, specimen identifiers, morphological characteristics), but increasingly we use phylogenies with geographic (specimen locations, distribution ranges), environmental (climate data, geochemical data), ecologic (habitat and community characteristics), epidemiological (clinical symptoms, treatment regimen, patient data) or paleontological (fossil characters, locations and dates) data. An interoperability framework that is to enable truly novel research questions based on phylogeny must not only encompass the present variety of contexts and data, but also provide straightforward mechanisms by which emerging research communities can extend the standards in a bottom-up fashion.

The major obstacle hindering broad availability and repurposing of trees is the lack of effective standards and a community-driven process for adopting and extending them. Existing file formats allow for representation of trees using a simple string and also the molecular or morphological character data used to infer the tree. There are no widely accepted standards for annotating tips, internal nodes or branches, and different applications have adopted unique methods for modifying the tree strings, meaning that annotations from one program may generate errors or be misinterpreted when import into another program. Other types of data/metadata, such as descriptions of evolutionary models or metadata annotations for provenance, have not seen any attempts at standardization. In order to share trees and associated metadata between data resources and analysis tools, we need a consistent syntax - with an unambiguous semantics - for describing them, that can be searched, read and understood using computational tools.

The EvoIO Stack as the foundation of community-driven interoperability

To be fully effective, a data interoperability initiative needs to encompass not only the exchange of data, but also the exchange of semantics and predictable programmatic access to interconnected data. Data exchange formats formalize the syntax in which data must be expressed to ensure that transmission from data provider to consumer incurs no loss of structure or composition (or only a documented— hence predictable— loss). However, as explained in [PCP09], syntax does not characterize the meaning of the data in an explicit way that can be extracted by computers. In recent years, the utility of formal ontologies for standardizing knowledge representation has been widely demonstrated [BR04, Skl01, SK02, SNM05], and these are therefore an integral component to achieve full interoperability of data that includes semantics.

The "EvoIO Stack" will be seeded with a triplet of emerging interoperability standards developed by an NSF-supported interoperability working group. In 2006, NESCent funded an Evolutionary Informatics working group [EVO] (PIs Stoltzfus and Vos) that gave itself the mandate to improve interoperability in evolutionary comparative analysis, so that more scientists could take advantage of the rigor and generality of the evolutionary approach. The group initially set out to support and update the aging NEXUS standard, but quickly realized that a more innovative and forward-looking approach based on modern, domain-independent information science technologies for achieving interoperability on the web would be needed to develop extensible, future-proof standards that are processable without deep domain knowledge. Based on the understanding that solving interoperability problems relies on formalizations of both the semantics (meaning) and the syntax (form) of data, metadata, and services, the group subsequently focused on a set of three "glue technologies": NeXML, a phylogenetic data exchange format based on a proven data model (namely that of NEXUS) with strict and validatable yet extensible syntax; CDAO, an ontology formalizing the semantics of evolutionary data and metadata in a machine-interpretable way; and PhyloWS, a standard programmable web-service based interface for publishing phylogenetic data to the web. Further details on each of the three standards are given below.

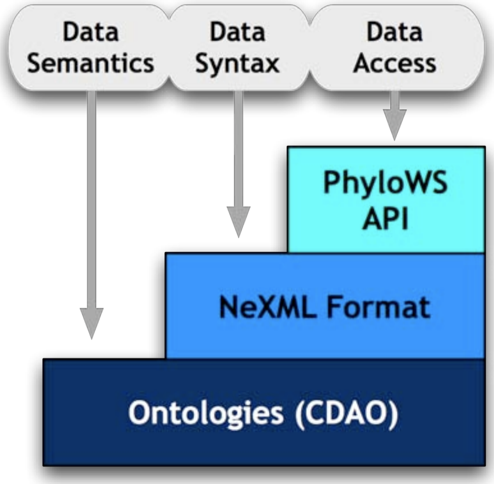

Figure 1: Relationships between the three EvoIO stack technologies

The EvoIO stack of interoperability standards for phylogenetic data will be jump-started with this triplet of technologies. This allows the proposed network to start its work on a strong foundation of intellectual as well as proof-of-concept work. The current versions of the three standards are rather preliminary and far from final, but, as shown in the 2009 Evolutionary Database Interoperability Hackathon (see Prior Support), are sufficiently mature to be applied to real-world problems so that major shortcomings or deficiencies will be revealed, and addressed, in the course of the work proposed here.

NeXML is an emerging exchange standard for phylogenetic trees, data matrices, and arbitrary metadata. Currently the most commonly used file format for trees and character data (e.g., a sequence alignment) is NEXUS [MSM97]; its usage is highest in the sub-discipline of phylogenetics, and tends to coincide with a preference for software for which NEXUS [NXS] is the preferred format (e.g., PAUP*, MrBayes). NEXUS is highly expressive to the extent that it requires a custom, context-sensitive parser (rather than standard context-free grammar-based parsers) and extensive domain knowledge for interpretation. The expressiveness has led to "flavors" of the format using different allowable approaches to serializing the same data, and the context-sensitive features of the format make syntax and consistency validation exceedingly difficult. Rutger Vos (see letter of collaboration), a co-leader of the EvoInfo working group, initiated development of NeXML [NXM], an XML schema for comparative biology that draws on the successful high-level block structure of NEXUS, but takes advantage of widespread support for XML, and harnesses the W3C-proposed RDFa standard to embed semantically rich metadata in a way that can be extracted by general purpose tools developed for the semantic web. NeXML is supported by APIs in Perl, Java, Python and JavaScript, with application support in Mesquite, BioPerl/Bio::Phylo, Phenex, DendroPy, DAMBE, TreeBASE and HIVQuery, and pledged support from the developers of PAUP* and HyPhy. While Dr. Vos remains a leader in NeXML development, a total of 14 developers (see the letter of collaboration from Vos) have worked on the NeXML schema or its implementations.

CDAO, the Comparative Data Analysis Ontology, formalizes terminology and knowledge about the biological principles inherent in phylogenetic trees, data matrices, and related information (metadata). Members of the EvoInfo working group that included Julie Thompson (U. Strasbourg, developer of clustalW [THG94] and the Multiple Alignment Ontology [MAO]) initiated the Comparative Data Analysis Ontology (CDAO) [PCP09, CDA] based on a foundation of an extensive list of use-cases [USE] for evolutionary (phylogenetic) comparative analysis reaching back to the 2006 Phyloinformatics Hackathon at NESCent (see Prior Support). The initial implementation of CDAO [PCP09] focuses on the concepts needed to represent the inference of a character history showing how a particular character (e.g., a morphological character or a position in a sequence alignment) changes during evolution, one of the core problems in comparative analysis. CDAO is represented in the Web Ontology Language (OWL), specifically the OWL-DL dialect, representing core concepts such as phylogenetic tree, Operational Taxonomic Unit, character-state data, and transition (i.e., an evolutionary change in the state of a character). An initial evaluation of the prototype has also been performed, encoding token data sets as CDAO instances and implementing simple query and reasoning tasks [PCP09].

PhyloWS is a web-services standard for searching, addressing, and accessing phylogenetic trees, data matrices, and their associated metadata in a predictable and programmable way. Despite the wealth of valuable and richly annotated phylogenetic data resources that are available online, the search and data access interfaces of most of these resources are incompatible, idiosyncratic, and primarily meant for human consumption, rather than being machine programmable. In recognition of this fact, H. Lapp (co-PI) and R. Vos initiated development of the PhyloWS standard at the web-services focused 2008 BioHackathon (which was sponsored by two Japanese life science centers and held in Tokyo). PhyloWS is a web-services standard that defines at a technology-agnostic, logical level use-cases, scopes, and requirements for programmatically interacting with an online phylogenetic data provider, and at a concrete, physical level at present includes a specification of a RESTful (Representational State Transfer) programming interface to trees and associated data and metadata. The API specification is currently focused on data retrieval, but specifications for data manipulations (such as removing tips, adding an annotation, changing the topology) have been started. Though relatively recent, the PhyloWS community has been growing to 10 developers on the mailing list, and PhyloWS is the subject or a key component of 2 of NESCent's Google Summer of Code projects in 2009, and 3 subgroup projects at the 2009 Database Interoperability Hackathon of 2009 (see Prior Support). Furthermore, R. Vos is implementing PhyloWS compliance in the forthcoming new version of TreeBASE.

Reference implementation

Reference implementations, especially when viewed as an essential component of standards development, help to identify problems early on, and optimize usability for the phyloinformatics community. Interoperability test-beds coupled with standard development has demonstrated benefit in other domains. As an example, geographic information managers must cope with incompatible data formats as common obstacles to interoperability and must also handle legacy data and applications. The Open Geospatial Consortium [OGC] implements interoperability experiments as part of the formalized standards development process. Geography Markup Language (GML), an OGC specification, provided a resolution to the problem of data sharing. GML is a widely adopted XML Schema that defines comprehensively data objects encountered in GIS, including geometry data, coordinate systems, measures, etc. Standard Web Service interfaces are defined based on GML, including Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), etc.

Large web systems and off-line tools are built to parse and generate data and services that make the exchange of legacy data and application convenient.

TOLKIN [TOL] is evolving as a project-based information management and analytical web application that provides informatics support for phylodiversity and biodiversity research. As a web app, collaborators in different locations and platforms can access shared data on voucher specimens, taxonomy, bibliography, morphology, DNA samples and sequences. Automated out-links connect users to relevant external data resources such as IPNI [IPN], Tropicos [TRO], GenBank [GNB], TreeBASE [TBS], Angiosperm Phylogeny Website [APW], and BioGeomancer [BGC]. The development of TOLKIN has been funded by various NSF awards to Co-PI Cellinese to support phylogenetic (AToL) and biodiversity (PBI) data. In addition, TOLKIN is expanding to serve non-plant projects, including data from Notothenoid and Labroid fish projects (NSF-ANT 0839007 & NSF-DEB 0716155). TOLKIN 2.0 will emphasize analytical “workbench” components by automating the assemblage of sequences, alignments, and output of NEXUS files, in addition to sequence BLASTing capability to support molecular systematics, and modules for defining and scoring of morphological characters for analysis and descriptive revisions, including publication of species and clade pages. Recently, prototype Kepler workflows have been integrated that run seamlessly through the TOLKIN workbench. Given the diverse data types and associated metadata, TOLKIN serves as an ideal test-bed for reference implementation of standards and specifications developed through the EvoIO community Network. In addition, current workflow development adds considerable significance for testing the applicability and benefits of the EvoIO Stack. TOLKIN incorporates open source bioinformatics solutions in its design, including BioPerl and BioRuby libraries for querying resources such as GenBank and for displaying molecular data. The data model for the molecular module is centered around the BioSQL core schema.

The EvoIO Network

Vision and Rationale for the Network

The EvoIO network will create a community-driven infrastructure of coordinated groups of stakeholders, technologists, developers, and users, with the common objective to make evolutionary data accessible, searchable, and combinable. The vision for the network draws importantly on the experience of the project leaders with "hackathons", which we conceive as a creative self-organizing group endeavor that draws on Open Space [OSP] principles and gives center stage to early-career scientist-programmers, arming them with the know-how and the opportunity to create interoperability solutions.

A *Data & Metadata Standards* working group will bring together a core group of the EvoIO standards lead developers with representatives from standards organizations, data resources, and analysis tool developers. A *Reference Implementations* working group will consist of investigators leading data resources selected as testbeds for implementing standards compliance and reviewing their utility as well as deficiencies. An *Outreach & Training* working group will organize a variety of training and workshop events to continuously increase penetration of awareness and know-how through diverse user communities. As the central mechanism to sustain the network, a series of hackathon events will continually engage participants from the different working groups, standards initiatives, evolutionary data providers, and a diverse group of stakeholder-participants who are simultaneously developers of data integration, analysis, and visualization tools.

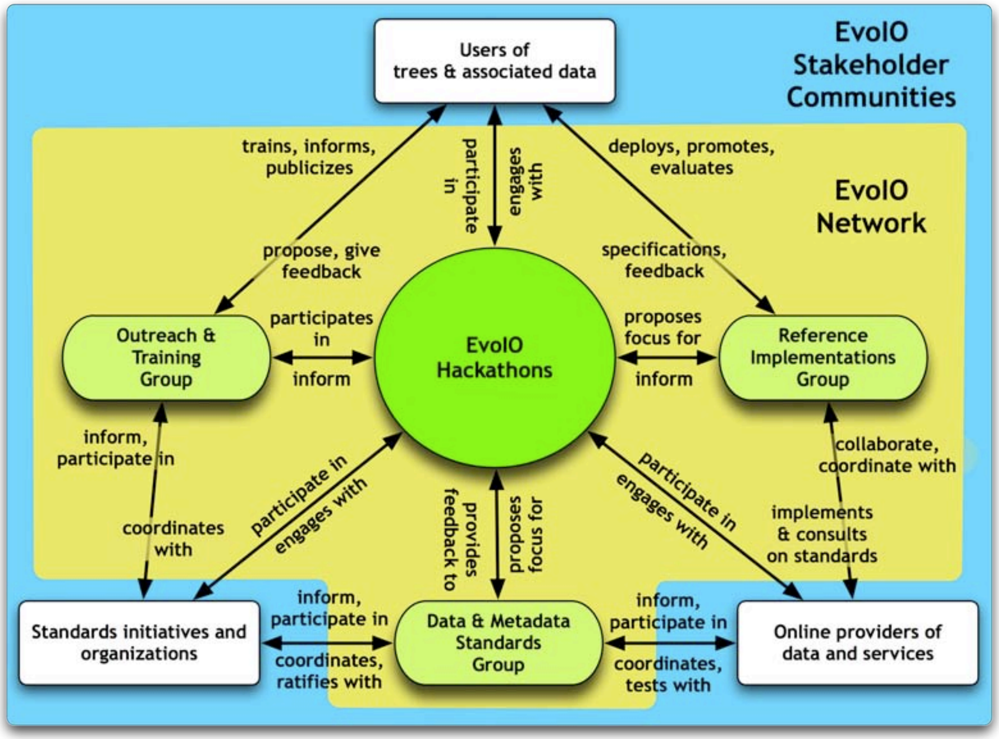
EvoIO Stakeholder Communities
EvoIO Network
Users of trees & associated data
trains, informs, publicizes
participate in
engages with
deploys, promotes, evaluates
propose, give feedback
specifications, feedback
Outreach & Training Group
participates in
inform
EvoIO Hackathons
proposes focus for
inform
Reference Implementations Group
inform, participate in
coordinates with
participate in
engages with
provides feedback to
proposes focus for
participate in
engages with
collaborate, coordinate with
implements & consults on standards
Standards initiatives and organizations
inform, participate in
coordinates, ratifies with
Data & Metadata Standards Group
inform, participate in
coordinates, tests with
Online providers of data and services

Figure 2: Proposed structure of the EvoIO network. The hackathons are the central organizing event that bring together the working groups with the broader community of data providers, users and standards organizations

Such a structured, coordinated, yet fully community-driven grass-roots network is both novel for the domain of evolutionary data, and a fundamental enabling infrastructure for numerous and innovative uses of evolutionary data in research, education, and integrative large-scale information management. At present, the communities and groups of people and resources who would form the stakeholders of the proposed network are uncoordinated, disjoint in membership, unaligned in their goals, and seldom interact or collaborate. In short, the EvoIO Network will generate the cohesion currently lacking among people and groups, and will foster meaningful and effective practices for online data exchange, dissemination, and interlinking. To ensure a broad representation of stakeholder communities, we have initially identified the following areas (with illustrative examples of active projects) as targets for interoperability of evolutionary data:

- molecular evolution & phylogenetics (e.g., TreeBase, TreeFam)
- comparative genomics (e.g., Ensembl, Gramene)
- evolutionary ecology, biodiversity, ToL (e.g., EoL, TOLKIN, ToLWeb, APweb)
- paleobiology (e.g., PaleoDB, PaleoPortal, Timetree)
- epidemiology (e.g., LANL HIV databases, GISAID)

An expanded version of this list (developed for this project) will be used as the basis for recruiting participants to a virtual stakeholder group (an email list), from which the working group members are expected to emerge as energetic and collaborative participants. We will remain responsive to the appearance of new data resources, inviting key personnel to participate in email lists and hackathon events. We will locate new resources through application notes in relevant journals (such as Systematic Biology, Bioinformatics and BMC Bioinformatics), the yearly Nucleic Acids Research database server and web server issues, The TDWG Biodiversity Information Networks Database [BIN] list, new Encyclopedia of Life content providers and NSF Assembling the Tree of Life (AToL) awards.

The following table lists projects already committed to participating in the network, either through direct affiliations with the PIs or through letters of collaboration from key personnel.

Project	Resource description	Collaboration type
Encyclopedia of Life [EOL]	Aggregator of biodiversity data including descriptive text, images, video, biogeography, classification, specimens; will be major user of standards	Biodiversity Synthesis Center at FMNH will host and fund meetings; Co-PI Cranston is a postdoc at FMNH; letter of collaboration from BioSynC director Mark Westneat.
iPlant Collaborative [IPC]	Cyberinfrastructure for plant biology; supporting “tree of life for green plants” grand challenge; will be major user of standards	Send personnel to meetings and training workshops; fund programmer time; letter of collaboration from Project Director Steve Goff; Co-PI McKay is leader of iPlant tree of life engagement team
Biodiversity Information Standards (TDWG) [TDW]	International organization facilitating standards for collaboration across biological databases	Co-PIs Cellinese and Lapp are co-conveners of the phylogenetics standards interest group
CREST Center for Bioinformatics and Computational Biology [CRS]	Cross-disciplinary center (Computer Science, Biology, Chemistry, Biochemistry, Plant and Environmental Sciences, and Mathematical Sciences) for training, research and education bioinformatics	Will host and fund meeting; Co-PI Pontelli is Director of CREST
PhyLoTA browser [PLTA]	Sequence clusters, alignments and trees based on data from GenBank	Co-PI Cranston is a developer
TOLKIN [TOL]	Shared data on voucher specimens, taxonomy, bibliography, morphology, DNA samples and sequences linked to external resources	Will serve as a reference implementation; Co-PI Cellinese is co-leading project development

With increasing penetration of outreach practices into more and more diverse scientific communities, the activities of the proposed network will, on a long-term basis, result in evolutionary data from a broad array of data providers ranging from large database resources to individual research labs being accessible as well as addressable online, with consistent and computable semantics. Because the latter will include resolvable links to other data types connected to various elements of a phylogenetic tree, this will fully integrate data about the history of life into the emerging interconnected web of data from various domains of science.

Network Organization

The core structure of the EvoIO network consists of a Leadership Team, three working groups, and a central integrating hackathon series that convenes once per year. The work of all groups and participants will be coordinated by a Leadership Team, which consults with an Advisory Board, as described in the Evaluation and Management plans.

The working groups will consist of 8-10 people comprised of a core of network leaders and additional representatives recruited from a variety of stakeholder groups and organizations on a rotating basis. Most of the interactions of working group members will be electronic. For face-to-face meetings, participants will be sponsored by the network for travel expenses, complemented where possible by externally sponsored stakeholder participants (see letters of collaboration). The main charge of the working groups is to continually review and chart the technology, implementation, and outreach vision for the EvoIO set of standards, ensure that the network is responsive to emerging or changing technologies as well as to the needs of its stakeholder communities, and make necessary decisions to align network activities with the long-term objectives of evolutionary data interoperability and interconnectedness. Each working group will have an email list that reaches a larger stakeholder group. Our experience with virtual groups is that a working group of a half-dozen members can be supplemented with a several-times-larger email list. In the ongoing electronic discussions that accompany development and planning, some individuals emerge as energetic contributers while others lurk in the background.

The **Data and Metadata Standards Working Group** will be co-led by Stoltzfus and Lapp, and will include Pontelli and others to be chosen as described further below. This group is responsible for development of the standards and will provide an interface between developers and data providers. This includes identification of stakeholders in key subject areas. The Data and Metadata group will hold teleconferences on a quarterly basis and will hold face-to-face meetings in the first and second years. A key task for the first year is to engage a large group of potential stakeholders (as described above in Vision and Rationale).

The **Reference Implementations Working Group** will be led by Cellinese, and will include Lapp, Cranston, and post-docs at UMBI and UF (to be named). The Reference Implementations Working group will hold virtual meetings (by teleconference) on an as-needed basis and a face-to-face meeting in year three. This group is responsible for implementations of the standards. These can be direct application to data sources (for example, the TOLKIN reference implementation project) or the development of APIs (for example, Java libraries for reading and writing NeXML files). Subscribers to the implementations email list will grow in number with each hackathon.

The **Outreach Working Group** will be co-led by Cranston and McKay. It will hold monthly teleconferences leading up to face-to-face meeting in the first year. Thereafter, teleconferences as needed to prepare for outreach events. The Outreach Working Group will focus on developing and disseminating documentation and training resources for the EvoIO set of standards. These tasks fall into two basic categories: development of online resources and organization of training events.

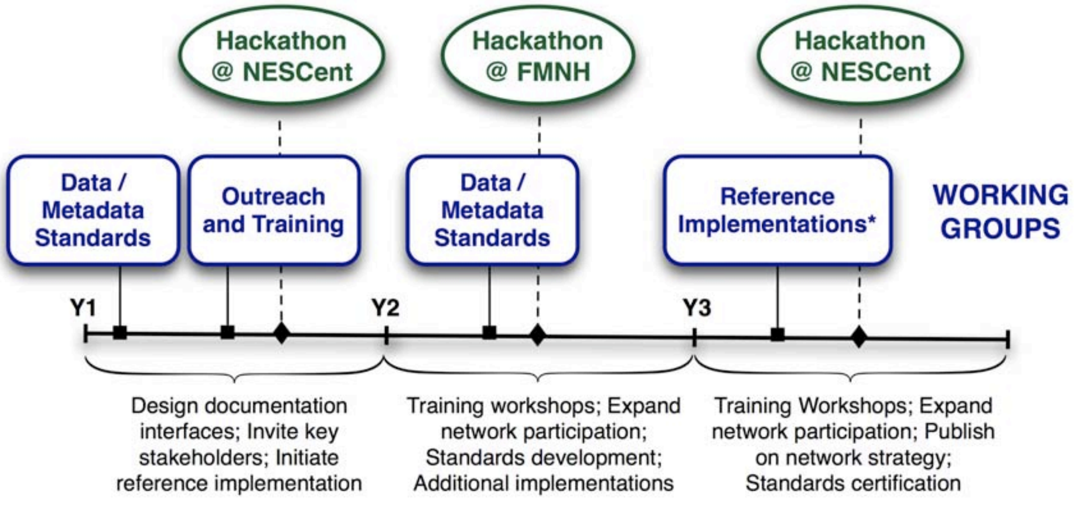
Hackathon @ NESCent
Hackathon @ FMNH
Hackathon @ NESCent
Data / Metadata Standards
Outreach and Training
Data / Metadata Standards
Reference Implementations*
WORKING GROUPS
Y1
Y2
Y3
Design documentation interfaces; Invite key stakeholders; Initiate reference implementation
Training workshops; Expand network participation; Standards development; Additional implementations
Training Workshops; Expand network participation; Publish on network strategy; Standards certification

Figure 3: The timeline of proposed network activities over the duration of the project, with examples of key tasks in each of three years. The Reference Implementations meeting in year 3 will be a larger hackathon-sized meeting held jointly with the TreeVisualization group at FMNH.

The **Hackathons** are the events that integrate members of the working groups with members of the community of standards developers, data providers and users. Each event will consist of 17-20 people, the majority of whom will be sponsored by the network for travel expenses, with some additional externally sponsored participants (see letters of collaboration). Organization of hackathon events will be done by personnel from all three working groups, including selection of attendees from the list of potential stakeholders. The organizing team for each hackathon will hold virtual meetings (teleconferences) and will document their work using a wiki (as we have done in the past for the 2006 and 2009 NESCent hackathons: see Prior Support). We have not assigned themes to these events in advance. Instead, we will allow the current state of network to inform both the list of participants and overall organizational goals. Then, participants at the event determine the agenda for the event that best utilize their collective expertise and meet their needs. The hackathon approach, integrated with Open Space [OSP] technology, has been shown to be an effective mechanism to inspire creative collaborations and solutions.

Network Activities

Technical support for standards development

The continued development of standards will be accomplished through an ongoing interaction between developers and data providers mediated by network personnel and using the hackathon model. The technical staff (the gradate student and two half-time postdocs) funded by this proposal will provide test data and examples of resources to the working groups and hackathons for the development of standards. This staff will also write APIs to support usage of the standards across programming platforms. The Data and Metadata Standards working group meetings, both face to face and online, will provide a forum for developers to interact across the three EvoIO component projects. Co-PIs Lapp and Cellinese, as co-conveners of the Biodiversity Information Standards (TDWG) Phylogenetics Interest Group, will promote formal adoption of the EvoIO stack ontologies, standards, specifications, and reference implementations.

Hackathon planning, hosting, and follow-up

The hackathon event series is the central integrating activity of the network that brings the leading technology and standards developers together with representatives from the working groups, stakeholder groups, standards initiatives, evolutionary data providers, and the diverse group of users who are simultaneously developers of data integration, analysis, and visualization tools. The hackathon format has a unique strength not only in the intense face-to-face interactions that it facilitates, but also through the collaborative software coding work, which gives every participant a direct sense of ownership in the products. They are hence the primary mechanism for engendering, sustaining, and broadening participation and know-how in applying as well as extending the EvoIO standards. As products, the hackathons will generate an on-going stream of prototypes showcasing the added value of interoperable data to research and dissemination, targets for outreach and training, opportunities for reference implementations, and assessments of utility and gaps in the EvoIO set of standards based on trying to apply them to concrete real-life problems.

TOLKIN and other reference implementations

Reference Implementations and interoperability test-beds will be built concurrently with standards and specifications developed for the EvoIO Stack. TOLKIN will serve as either a reference implementation or a production implementation based on the reference for certain specifications from the EvoIO Stack. The implementation of PhyloWS will prove to be a logical extension as part of a reference implementation. All reference implementation code libraries based on the Stack will be made publicly available through the open source code repositories. TOLKIN developers currently devote enormous efforts to address file formats and implementation issues. The adoption of NeXML as a standard data exchange format will allow TOLKIN to provide a user-friendly workflow runtime system and a library of commonly used workflows in phylogenetic systematics, molecular evolution, and population genetic studies. Formalized digital workflows are a convenient way to express the complex processing pipelines encountered in different research fields, which might consist of intermixed database query, statistical inference, algorithmic analysis, rendering, etc. The objects and tools involved can be locally installed or accessed remotely through a standard interface. Computational models provided by existing workflow system (e.g., Kepler) eliminate the need for researchers to cope with the heterogeneity of concurrency, data consumption ratio, etc.

TOLKIN resources can be remotely accessed using a REST-based interface. Features of the application are accessed by users providing resource identifiers and requested action types in the HTTP request. GET, POST, PUT and DELETE requests are supported for each resource. PhyloWS would provide a logical and useful extension of the services already provided. TOLKIN already follows a set of standards for the list, retrieve, create, update, and delete resource functions. Having a standards-oriented API for accessing other site functions would make TOLKIN more open and accessible to users and other developers in the scientific community.

Co-PI Cellinese will manage and coordinate project collaborators and other interested parties to ensure that user requirements are effectively translated into standards and specifications, and can be implemented as useful community software tools. The requested FLMNH Postdoc (mentored by Cellinese) will take a primary role in building sample implementations representing a suite of interoperability test-beds that will be built in Perl and/or as web applications with Ruby-on-Rails. As part of this effort, the Postdoc will also implement a set of test suites that document successful code implementations. Sample implementations will include conversion services from Nexus to NeXML and implementations of the PhyloWS API client. In addition, we will explore the possibility of exporting TOLKIN projects into CDAO ontologies, which will foster data permanency and repurposing.

Documentation and online resources

Adoption of new technologies is highly dependent on the availability of documentation that is clearly written and easily accessed and on technical support from developers or expert users. EvoIO.org will be the point of entry for all documentation and support with links to existing resources for NeXML [NXM] and CDAO [CDA]. The main EvoIO.org site will be developed using wiki technology, similar to perl.org or evoinfo.nescent.org. The use of a wiki as a communication forum at past hackathons has proved to be an excellent method for capturing discussions, ideas, proposals and documentation from the participants in an open format that can be accessed by the wider community. There are currently developer mailing lists for NeXML and PhyloWS. We will archive existing content in a single location using Nabble or a similar service, and also add lists aimed at supporting users rather than developers. We will provide several introductory tutorials, including example projects such as the TOLKIN or other reference implementations. We also plan to document the process of community building and standard development itself [LBB+07] to answer questions about why the hackathon model works, and how to run a successful hackathon. Co-PIs McKay and Cranston will be primarily responsible for organization and initial content through the Outreach working group. We plan to spend year 1 setting up the infrastructure for these resources as well as writing basic introductory content. In years 2 and 3, we will extend the content based on reference implementations, hackathons and results from the other working groups. We will engage the user and development community to contribute content through the use of wikis and Google Groups / mailing lists.

Training events

We will hold training events beginning in year two to allow for initial development of online documentation and reference implementations in year one. Training events will focus on end-users of evolutionary data, e.g., on locating, querying and using evolutionary data using the EvoIO standards. In order to maximize outreach and participation, we will co-localize these training events with existing meetings of evolutionary biologists. We plan to hold at least one workshop at each of Evolution (joint meeting of the Society for the Study of Evolution, the American Society of Naturalists and the Society of Systematic Biologists), SMBE (Society of Molecular Biology) and Botany & Mycology (joint meeting of several plant societies). These events will be lead by co-PIs McKay and Cranston and co-PI Lapp and Vos (letter of collaboration). We will also incorporate training on usage of the EvoIO standards into the annual NESCent Computational Phyloinformatics course. Co-PI Lapp and Vos have been faculty at this course since its inception in 2006. Teaching materials at the training events will include online resources at EvoIO.org as well as the TOLKIN reference implementation. Later in the project, this will expand to include resources and tools developed at the hackathons.

Evaluation plan

The project will implement a thorough evaluation plan, which will address both the establishment of the interoperation network and the quality of the technical products developed and disseminated through the proposed activities. The evaluation plan will be designed and implemented by the project director, with the assistance of an external evaluator (Dr. Cummings) and with consultation from an Advisory Board. The Advisory Board will consist of 4 individuals not affiliated with EvoIO projects, but chosen to represent larger initiatives such as W3C and OBO, and preferably with an international scope. The UMBI budget provides funds for face-to-face Advisory Board meetings in years 1 and 3: in order to take advantage of synergies and to expose the board to Network activities, these meetings will be co-localized with a working group meeting or hackathon. The Advisory Board will have joint teleconferences with the Leadership Team on a quarterly basis in year 1, and twice-yearly thereafter.

The main goals of the projects to be measured by the evaluation process are: **(1)** Establishment of an effective network that promotes wide adoption of the EvoIO standards; **(2)** Development and dissemination of EvoIO standards, including documentation, reference implementations; **(3)** Creation of an outreach and training infrastructure to grow and expand the collaborative network.

There are four phases of the evaluation plan: (a) *Design Evaluation* will be conducted immediately after the approval, to complete the details of the interop network infrastructure design; (b) *Implementation Evaluation* will proceed concurrently with the development of the EvoIO network. It will ensure that the construction of the infrastructure is performed per the design guidelines, and that the research and outreach processes are properly activated; (c) *Process Evaluation* will measure the effectiveness of the implemented mechanisms and progress towards the goals throughout the project. This concurrent evaluation will be critical in providing feedback to the investigators, allowing them to take corrective actions [Lan95, Nie93]; (d) *Exit Evaluation* will measure the final outcomes of the development and outreach infrastructure, using quantitative (e.g., publications, citations, new projects, number of participants) and qualitative (e.g., surveys, feedback reports) metrics. The exit evaluation will also validate the potential for long-term impact, e.g., adoption of EvoIO standards in established and new tools, conversion of data repositories to EvoIO standards, use of EvoIO standards in submissions to journals.

The evaluation process will be multi-platform. The project director will oversee the evaluation of the overall network, while the co-PIs will evaluate the progress of their respective working groups. For the evaluation, we will collaborate with Professor Jonathon Cummings (Associate Professor of Management) from the Fuqua School of Business at Duke University. Professor Cummings has conducted several evaluations of NSF programs, including projects in Knowledge and Distributed Intelligence (KDI; [CK05]) and Information Technology Research (ITR; [CK07]). His current work focuses on Virtual Organizations, and specifically how to increase the likelihood of success for research collaborators across institutions [CFF+08]. He has also developed software to facilitate the collection, analysis, and visualization of evaluation data [NCP]. With NESCent, Professor Cummings has evaluated informatics hackathons. His formative evaluation [Scr67] includes observations, interviews, and surveys to provide feedback on the effectiveness of NESCent. A key advantage of formative evaluation is that experimental methods can be used to assess the impact that interventions have on the programs. For example, he could compare productivity and community-building in physical versus “virtual” hackathons. The external evaluator will be responsible to implement the evaluation metrics for the specific activities, filtering the collected data and providing a periodic assessment of accomplishments. The Advisory Board provide comprehensive external evaluations of the network’s accomplishments.

Network Management Plan

The management plan is based on the structure of the Network, takes into account the skills and experience of the leadership team, and addresses communication, accountability, supervision, and adapting to changed conditions. The multiple-PI structure for this project is dictated by the scope of the proposed research and the need for complementary expertise in evolution, bioinformatics, and computer science. The PI and 5 Co-PIs are located at 6 geographically dispersed work sites (UMBI, CSHL, FMNH, NESCent, NMSU, UF); project staff include 1 graduate student, and 2 part-time post-docs at different sites. Members of the leadership team are adept at using remote collaborative technologies including email, text chat, video chat, teleconference and wiki. The more senior members of the leadership team (Pontelli, Stoltzfus, Lapp) have considerable experience in project leadership or in managing group collaborations. This group of 6 has had many teleconferences and has brainstormed and solved conflicts using principles of consensus decision-making.

The *Leadership Team* consisting of the PI and Co-PIs will act as an executive committee that manages the working groups and all aspects of the project. As chair of the Leadership Team, Stoltzfus (UMBI) will ensure that the Leadership Team stays focused on the goals of the project, and will provide an interface between the lead institution and NSF with regard to project accountability and reporting. The Leadership Team, in turn, will ensure that all groups and participants continue to document their efforts, to communicate with others, and to stay focused on the goals of the project. The Leadership Team will communicate by an email list, and will conduct its meetings by teleconference, at least monthly during the first year of the project, and at least quarterly during the second and third years. Each monthly meeting will result in a progress report posted on the project web site at evoIO.org. The Leadership Team will consult with the Advisory Board on any major decisions, will report yearly to the board, and will make the consideration of any feedback from the board a regular agenda item at its meetings. Unforeseen collaborations with the Network may arise, and for this reason, the Leadership Team may be expanded (by consensus of existing members) to add members representing projects that have made a clear commitment to the goals of the Network and to its collaborative philosophy.

Each working group is led or co-led by members of the Leadership Team. This facilitates communication and coordination between the working groups and the Leadership Team. Although none of the working groups has a face-to-face meeting every year, the members will have numerous opportunities to interact directly at Network-sponsored hackathons and workshops. In the case of conflicts between individuals of two or more teams, the PIs will address the issue as a team and will decide, by consensus, what steps will be taken to address the conflict. Oversight of expenditures at the work sites is not expected to be problematic, given that there is very little left to the discretion of the PIs.

Broader Impacts

The research areas affected by this proposal — all those areas in which phylogenetic trees are used — are diverse and currently are not unified by professional organizations, software platforms, or standards. By bringing together scientists from various disciplines, we will develop awareness of the need for standards, cohesion around preferred approaches to interoperability, and ultimately a broad consensus on specific standards. A successful approach for community building will also be an important resource. Our approach builds on the momentum of work done under prior NSF funding via NESCent, including our innovative hackathon model. This model empowers early-career scientists with the know-how and the connections to build interoperable solutions. Through this mechanism, user requirements will be effectively translated into standards and specifications, and can be implemented as useful community software tools. While a core group of hackathon participants will be selected by the organizers to take advantage of targets of opportunity, the remainder will be chosen in response to a broad solicitation throughout the entire biodiversity, systematics, genomics, and phylogenetics communities. Hackathons will take place in eastern, western, and central locations to maximize diversity in impact. The development of outreach activities, particularly those that are online or take place at conferences, will make the products of this consortium widely accessible to all.

Co-PIs McKay and Lapp have strong ties to the Generic Model Organism Database (GMOD) project, which has been very effective in promoting genomic data and software interoperability through collaborative development and active outreach through workshops, courses and online documentation [GMO]. A measure of the success of this approach is the widespread adoption of GMOD software components, such as the Generic Genome Browser, by almost all major model organism databases and more than a hundred other online databases. In addition to the successful hackathon model, the network will also employ outreach activities modeled after GMOD to promote awareness and adoption of the resources generated by this consortium. The development of the EvoIO network will serve as a model of effective collaboration and community building for the phylogenetics and related communities in much the same way as GMOD has done for the genomics communities.

Education, Outreach and Training

The development of informatics infrastructure to support the proposed standards will provide excellent opportunities for wide involvement of graduate and undergraduate students. In particular, we envision selecting some of the development tasks as topics for the Senior Project class in the Computer Science program at New Mexico State University. This is a required class for all undergraduate students, where teams of students put in practice their software engineering and programming skills in a simulated "real world" task. NMSU is a federally recognized Hispanic-serving institution. The Department of Computer Science and the CREST Center are dedicating extensive resources to developing educational and research programs that builds on the diversity of the local student population. Special efforts will be made to involve graduate students from traditionally under-represented backgrounds in this proposed project. CREST is also offering summer programs for high school students, teachers and community college students in bioinformatics. The issues of interoperability being addressed in this proposal are an ideal subject for a portion of such a bioinformatics course.

Outreach to younger students and the public will come from collaboration with FMNH / Encyclopedia of Life. The standards developed in this protocol will allow providers of evolutionary data to become data partners with the EOL, increasing exposure of EOL users to the tree of life. The collaboration agreement includes a joint meeting between the EvoIO network and the FMNH Tree Visualization working group. This group has a very strong outreach component focused on creation of phylogeny visualization tools for students and the general public. The ability to annotate phylogenies with highly visual metadata such as images or geographic locations will be critical for the success of such endeavors to forward tree-thinking. Our hackathon model engages young researchers in real-world biodiversity informatics problems while also introducing them to a network of more experienced researchers. In our experience, these are very productive events and often produce results that can either be published or followed up on as ongoing collaborations. Participants, including graduate students, are encouraged to present the results of their work at meetings, and projects started at the hackathons can also be expanded into full NSF proposals involving both early-career and experienced personnel. This proposal itself is an example of this type of mentoring, as three of our PIs are early-career scientists and two are women in computational biology.

As they mature, adoption of the standards and associated tools by users that stand to benefit from them is an ultimate goal of this network. We will emphasize an additional outreach component directed at invested members of the phyloinformatics community and also non-specialist users who are interested in having access the the software, etc, made possible by the implementation of the NeXML and PhyloWS standards. This outreach will be in the form of workshops at held at major conferences, such as SMBE (Society for Molecular Biology and Evolution) and the "Evolution" meeting (joint annual meeting of the Society for the Study of Evolution, the Society of Systematic Biologists, and the American Society of Naturalists). These workshops will increase accessibility to the potential user community, including underrepresented groups, because they will be held at conferences where many researchers, who may not otherwise be aware of the network, are present. The workshops will have 3-4 instructors, who will present high-level overviews of the standards, as well as engaging the attendees with hands-on instruction for implementation and use or EvoIO resources. The iPlant collaborative, who has committed to developing infrastructure components in concert with the standards developed by this network, has also committed to sponsoring the iPlant staff to attend and serve as instructors at these workshops.