Title: Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity

Short title: Botanical Information and Ecology Network (BIEN)

Project Leaders

Brian J. Enquist, Associate Professor, University of Arizona, Tucson, AZ, 85721, USA; benquist@email.arizona.edu; (505) 626-3336 (main contact). Research covers the influence of plant traits, phylogeny, and diversity on larger scale ecological, ecosystem, and evolutionary patterns. Teaching interests cover general plant functional biology, macroecology, diversity, and community ecology. Outreach includes serving data from ecological forest plots via the SALVIAS network.

Richard Condit, Chief Scientist, Center for Tropical Forest Science, Global Forest Observatory Network, Smithsonian Tropical Research Institute, Unit 9100 Box 0948, DPO AA 34002; conditr@gmail.com. Research interests include population biology and community models. Teaching interests are in quantitative ecology and forest modeling, with an emphasis internationally.

Robert K. Peet, Professor, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3280, USA; uniola@email.unc.edu. Research focuses on patterns of plant diversity in North America, the vegetation of the Southeastern United States, and bioinformatics as applied to ecology and systematics. Outreach includes serving vegetation data on the VEGBANK network, developing international exchange standards for ecological and taxonomic data, and work on copyright policy in the digital age.

Mark Schildhauer, Director of Computing, National Center for Ecological Analysis and Synthesis (NCEAS), 735 State St., Suite 300, Santa Barbara, CA 93202-3351 USA; schild@nceas.ucsb.edu. Research on software tools for storing, integrating, and manipulating ecological data, through involvement with several large ecoinformatics efforts, including the Knowledge Network for Biocomplexity (KNB), and the Science Environment for Ecological Knowledge (SEEK). Outreach includes efforts to make published datasets widely available, using advanced knowledge representation techniques to facilitate data discovery and integration, and teaching ecologists the importance of informatics in their science.

Barbara M. Thiers, Director, William and Lynda Steere Herbarium at the New York Botanical Garden, 2900 Southern Blvd., Bronx, NY 10458-5126 USA; bthiers@nybg.org. Research in the application of information technology to herbarium management, and outreach efforts to increase access to specimen-based data for the scientific community.

Core Team Members

Sandy Andelman (Conservation International, Vice President TEAM initiative, vegetation plots), Brad Boyle (University of Arizona, biodiversity informatics), Jeannine Cavender-Barres (University of Minnesota, plant traits and phylogenies), Steve Dolins (Bradley University, computer science), Stephanie Hampton (Deputy Director NCEAS, outreach and training), Jesse Kennedy (Napier University, Scotland, Computer visualization, Taxonomic Concept Schema), Brian McGill (University of Arizona, ecological informatics, macroecology), Hans ter Steege (University of Utrecht, Netherlands, collections and plot networks), Jens Christian Svenning (Aarhus University, Denmark, biogeography), Nathan Swenson (Michigan State University, plant traits and macroecology), Oliver Phillips (Leeds University, UK, RAINFOR network, global change and vegetation plots), Peter Jørgensen (Missouri Botanical Garden, collections and C3Tropicos), Dave Vieglais (University of Kansas Biodiversity Research Center, Darwin Core/DiGIR), Corine Vriesendorp (Field Museum, collections and outreach); Susan Wiser (Landcare Institute, New Zealand, plant ecology, IAVS vegetation plot exchange schema).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Project Summary

To answer many of the major questions in comparative botany, ecology, and global change biology it is necessary to extrapolate across enormous geographic, temporal and taxonomic scales. Yet much ecological knowledge is still based on observations conducted within a local area or even a few hundred square meters. Understanding ecological patterns and how plants respond to global warming and human alteration of landscapes and ecosystems necessitates a holistic approach. Such an approach must be conducted at a scale that is commensurate with the breadth of the questions being asked. Further, it requires identification, retrieval, and integration of diverse data from a global confederation of collaborating scientists across a broad range of disciplines. We propose to network core databases and data networks to create a novel resource for quantitative plant biodiversity science. The grand challenge is to assemble and share the world's rapidly accumulating botanical information from plots and collections to create a distributed, web-accessible, readily analyzable data resource. With such a resource, we will answer major questions of direct relevance to plant ecology, plant evolution, plant geography, conservation, global change biology, and protection of biodiversity and ecosystem services. In particular, *how does climate influence the distribution and abundance of plant species, how does the phylogenetic diversity of plants vary across broad environmental and climatic gradients, and how are plants assembled into ecological communities?* While these and associated questions are at the core of many research endeavors in comparative botany and ecology, our past collective inability to integrate data on a large scale has significantly limited our ability to address these questions head on. This proposed Grand Challenge team will create a data resource of unprecedented size and scope together with the tools for its use, thereby empowering botanists and the general public to better address fundamental issues in plant ecology and global change biology. Although we will focus on plants of the New World, the infrastructure and protocols developed will be scalable to all geographic regions and all types of organisms. Future steps will enable cross-cutting linkages to emerging networks on plant genomics, physiology, and phylogeny, allowing us to address fundamental genetic and evolutionary questions at unprecedented spatial and temporal scales.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

I. – The grand challenge: *In a changing world, what grows where and why?*

I.a - Introduction - Understanding what controls the abundance and distribution of botanical diversity is fundamental to much research that underlies ecology, evolution, comparative plant biology, and global change biology [1-4]. For example, the geographic distributions of plant species reflect physiological tolerances [5, 6], evolutionary and climatic history, and offer insights into the traits that underlie adaptation [7-12] and the mechanisms involved in population divergence [13]. Abundance is a measure of ecological dominance and ecosystem services and often reflects fitness [3, 8, 14]. Indeed, together, information on abundance and distribution provide the ability to bridge the plant sciences by linking many central questions in plant biology [13, 14] to the great botanical diversity in nature. In addition, climate-induced shifts in the distribution and abundance of plant taxa can impact the diversity and function of local communities [15] and thereby alter ecosystem attributes [16]. In a changing world taxon abundances and geographic ranges will likely rapidly expand or contract [17-19] and some species will become extinct [20, 21] but we as a scientific community are not yet prepared to anticipate those changes [18].

Our ability to predict species' abundances and ranges, let alone how they will change, remains limited [18, 22]. In order for biologists to predict how individual taxa and entire communities will respond to a changing world requires understanding why plant taxa grow where they do and what limits their ranges. However, distributional and community shifts are broad, spanning large geographic gradients and sometimes continents. Further, range size abundance, and phylogenetic/taxonomic information have rarely been addressed in many parts of the globe, especially in the tropics due to the lack of integration of the many plot samples where abundances have been calculated. A full understanding of present and future patterns of biodiversity necessitates examination of processes and taxa across geographic and environmental gradients.

I.b - The Barriers - The lack of a global source of integrated and standardized biodiversity observation records is a fundamental impediment to advancing the plant sciences. As a result, the development of a global perspective on variation in basic floristic and ecological attributes has been limited. These problems are especially acute in the tropics where biodiversity is concentrated but poorly known [23-25].

Most datasets originate from individual researchers and span a few square kilometers [26, 27], recording varied kinds of data using idiosyncratic protocols and published (if at all) in various formats [28]. Further, even if we could integrate these original sources, we would not be able to place much confidence in the resulting list of taxa because: (*i*) there is no standardized global list to assess the validity of the names or circumscriptions of plant taxa[1]; (*ii*) there is no global standard for naming variants contained within taxa[2]; (*iii*) there are numerous technical and data quality issues with merging and serving data from disparate sources; and (*iv*) *there is no standardized process by which data on distribution and abundance of plant taxa can be combined with information from plant physiology, genomics, and phylogeny*. The last point is especially important because understanding why species are limited in where they grow requires genetic and physiological knowledge of traits that affect how an individual responds to the environment [29]. We can describe observed vegetation shifts across continents as temperatures rise, for example, but will not be able to predict future shifts until we know how individuals and taxa change across environmental gradients and how they react to changes in temperature and precipitation.

I. c - The Solution - iPlant offers a unique opportunity to create a confederated plant data network to allow scientists to address "what grows where and why". This network will also provide feedback and training to data providers as well as novel educational opportunities. The solution will require developing a

1 http://www.wakehurst.org/science/directory/projects/Target1GSPCGlobCheck.html

2 http://www.tdwg.org/about-tdwg/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

transformative cyberinfrastructure, based on proven informatics approaches coupled with cutting edge software tools to: (*i*) make easily accessible the many disparate and individually limited datasets; *(ii*) standardize these data streams, *(iii)* integrate data streams adhering to divergent taxonomic standards; *(iv)* create a constantly updated but perfectly archived data resource that biologists and the public can query seamlessly; *(v)* provide feedback to the main data providers so that they can then better serve and standardize their data; and (*vi*) integrate the plant data streams with environmental data from a range of sources (e.g., climate, land cover change, etc.). By collecting and combining vegetation censuses, botanical surveys, and specimen records from herbaria, we will achieve comprehensive, cross-taxa coverage and provide a global perspective on variation in basic floristic and ecological attributes [25, 30-32]. Ultimately, we aim to create a "Data Discovery Environment" that will serve and process basic information for academics, conservationists, and the public. We call our integrated data resource BIEN, or the Botanical Information and Ecology Network[1].

I.d - The Grand Challenge Team - We propose to address the Grand Challenge by creating an integrated global network of botany data providers and users. What the BIEN team seeks for plant science is not simply a new data network and cyberinfrastructure, but a new paradigm with respect to how data are recorded and integrated. The need for this paradigm switch is well documented [33-35], but making it happen has proven difficult. With support from the National Center for Ecological Analysis and Synthesis (NCEAS[3]), the BIEN grand challenge team held a planning meeting in December 2008. The BIEN team and additional collaborators (Table 1) broadly represents the community of plant biodiversity sciences and includes members of many major botanical institutions, data networks, and biodiversity initiatives, as well as informatics experts. Together, we identified key data sources from around the globe, the kinds of data to be extracted from each, and identified the requirements for merging disparate data into an integrated framework. In the process of initiating this botanical network we identified short- and long-term cyberinfrastructure needs, including existing technologies and protocols, programming challenges, end-user tools, and web services. Drawing from the conclusions of our initial meeting, we here detail the barriers and specific solutions needed to integrate and provide access to botanical biodiversity data.

II. – Proposed scientific activities

Design and creation of the BIEN network will be guided by the goal of addressing the Grand Challenge question – *in a changing world, what grows where and why?* The Grand Challenge Team has identified three fundamental and tangible sub-questions that will enable us to directly address aspects of the grand challenge and demonstrate the power and utility of the cyberinfrastructure we propose to design and build. Each question addresses key problems in plant ecology, comparative plant biology, and biodiversity conservation.

Q 1: *How does climate influence the relative distribution of narrow and widespread species? Do these relationships vary in tropical and temperate environments?*

Q 2: *How are abundance and size of geographic range of taxa related? For example, do plants with small ranges tend to be rare relative to widespread species?*

Q 3: *What are the physiological, demographic, environmental, and phylogenetic correlates of rarity (small ranges, low local population size) and commonness (large range, high local population size) across environmental gradients at scales ranging from local to continental? Can these*

[1] http://www.nceas.ucsb.edu/projects/12290
[3] http://www.nceas.ucsb.edu/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

correlates be used to predict vulnerability or resistance to extinction for species and communities under differing scenarios of habitat loss and climate change?

The first two questions are fundamental to ecology [1]. The third question has strong practical implications for land management decisions related to conservation hotspots and preservation efforts [23]. All three questions lead to predictions about the vulnerability of species and communities to extinction, and all will allow us to compare temperate and tropical floras precisely [8, 36, 37]. Range size abundance, and phylogenetic/taxonomic information have rarely been addressed in many parts of the globe, especially in the tropics, and we will be able to address the questions at a global scale. Most importantly, the project will result in an integrated data resource on the distribution of plant species that will be a baseline against which to gauge responses to global warming and global change. As we discuss below, questions relating range size and abundance to physiological or genetic traits rely on other initiatives that are currently underway, some under the auspices of iPlant.

III. – Data and computational activities to address the grand challenge

III. a – The goal - Addressing the Grand Challenge will require a comprehensive, integrated and standardized data network of biodiversity observation records from across the globe. Data sources must range from the tiny datasets collected by individual scientists to data streams from large and long-lasting programs, established groups, institutions, and herbaria. We propose to develop a data integration network where plant biologists from many different disciplines, with many different research goals, upload, standardize, merge, and share data. This network will also serve as a permanent repository for legacy data. The end product will be the ability to address questions at spatial and temporal scales far exceeding the reach of any individual research program.

III. b –Biodiversity observation data - There is an enormous amount of existing data on plant distribution and abundance as well as several emerging sources of additional botanical information. Many are tiny datasets collected by a single scientist, whereas others represent much larger and long-lasting efforts. We identify two main sources of data crucial to questions about geographic range and abundance:

(1) Collection records. We estimate based on *Index Herbariorum*[4] that the world's museums hold ~300 million plant specimens, of which perhaps 15 million have been digitized and are potentially accessible for our project. Others are being steadily digitized. GBIF[5], the Global Biodiversity Information Facility, has begun the task of assembling digitized museum collection records, offering already 65 million individual species occurrence records from the Americas (including both plants and animals). There is the potential for a total of 95 million records from US herbaria alone [38]. GBIF records are available for our demonstration project. In addition, we are working directly with several US data sources, including the Missouri[6] and New York[7] Botanical Gardens, and have preliminary agreements to access other major collections in the US (Table 3).

(2) Vegetation plot records. Plant ecologists routinely delimit precise areas and assess plant species abundance, often by recording either individual trees by size or all plant taxa by percent cover. These 'plots' allow precise estimates of abundance of each species. Plot data are linked by accurate

[4] http://sciweb.nybg.org/science2/IndexHerbariorum.asp
[5] http://www.gbif.org/
[6] http://www.mobot.org/
[7] http://www.nybg.org/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

geocoordinates to specific site conditions. The grand challenge team identified vegetation plots already digitized and available to the BIEN data confederation: 1350 tropical forest plots in Central and South America and 325,000 North American vegetation plots, both forest and non-forest (Table 2). Most of these plots hold 10-100 plant species, so the total number of species occurrences is on the order of 15 million. At the first BIEN meeting we also examined the total number of digitized plots that could be integrated in a future network. We identified several hundred thousand additional North American plots potentially available, and there are over a million European vegetation plots that have been digitized in TurboVeg format alone [39] (Hennekens *pers. comm.*). Large digital plot archives such as those for New Zealand and South Africa are also potentially available. We estimate that within five years we can have networked data from in excess of 400,000 North American plots, 2000 Central and South American plots, and potentially another million plots from outside the Americas.

Based upon our first order approximations, *there is the potential for at least ~500 million taxonomic occurrence records*, where a single record is an observation of a plant characterized by a latitude and longitude coupled with a taxonomic determination of that plant. As we discuss below, accessing and integrating these two fundamental data units across multiple botanic data sources entails significant challenges in informatics. But successful integration will provide a powerful new cyberinfrastructure [38] to answer fundamental questions in ecology, evolution and global change research.

III. c Secondary Data Sources: Traits, Phylogeny, and everything -omics – Confederation of the data sources listed above will, for the first time, provide the botanical community with a baseline for the study of abundance and distribution on a global scale. It is these data that will be the core of the BIEN initiatives. However, in order to address the processes and mechanisms that influence and ultimately determine many aspects of abundance and distribution, the BIEN initiative must also merge this framework with several additional sources of data being organized by groups outside of BIEN. The BIEN data network will become even more valuable when linked to these other plant informatics efforts. Indeed, there are exciting potentials for synergistic activities with emerging groups. These groups include: (*i*) functional traits; (*ii*) genomic data; and (*iii*) phylogenies. These attributes are, in turn, best linked through botanical nomenclature (a focus of BIEN) or indirectly through phylogenetic resources such as the GC "Tree of Life" team.

Functional traits are phenotypic attributes of the organism and are defined as quantifiable morpho-physio- and phenological attributes that impact fitness indirectly via their effects on growth, reproduction and survival, the three components of individual performance [40, 41]. There are several efforts underway to actively compile and network global information on several key [42, 43] plant traits. The key trait networking efforts include GLOPNET[11], the NSF RCN funded TraitNet[12], the National Phenology Network[13], and the global TRY[14] network. Variation in functional traits often influence the performance in plants in differing environments [44]. Thus, the ability to merge plant distribution and abundance data with information on plant functional traits will allow for the mechanistic linkage between abundance and distribution with variation in phenotypes.

Ultimately, in a changing world, both genes and environments are important in determining how plants respond to the environment and ultimately where they grow. Through iPlant the BIEN team has the potential to integrate with other GC teams such as the GC team "Cyberinfrastructural Support for Genetic and Ecophysiological Studies of Plant Phenological Control in Complex and Changing Environments") utilizing new molecular- to field-level models, databases, and techniques relating to the phenology of crops and plants in natural ecosystems. These new techniques, such as gene-based, ecophysiological

[11] http://forestecology.cfans.umn.edu/glopnet.html
[12] http://www.columbia.edu/cu/traitnet/
[13] http://www.usanpn.org/
[14] http://www.try-db.org/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

approaches [12], can be used to chart how changes the abiotic environment can influence the life cycle across the geographic range of a plant species, and presumably across different species whose ranges occupy differing environments.

As we describe below, the cyberinfrastructure proposed for BIEN is key to integrating all these other GC efforts in that it provides the linkage through which all these data networks can be integrated by providing the critical semantic mediation of the many-to-many relationships between taxon names and concepts. This linkage will be central and absolutely necessary in order to merge genomic, trait, phylogenetic, distribution and abundance data.

III. d – 'Priming the Pump': Short-term needs of a global BIEN - In order to quickly address our core science questions, cyberinfrastructure needs, and obstacles to data integration, the BIEN core team is currently compiling and analyzing several data sources. Using the sources in Tables 2 and 3 we are generating the foundations of an integrated plant diversity database. During this process we have identified two critical problems that continue to limit attempts at broad-scale integration of biodiversity observation data. It is these problems that will first require the support of the iPlant collective.

- *The lack of taxonomic standardization is the most important informatics impediment in the plant sciences*. The plant sciences do not yet have the cyberinfrastructure needed for taxonomic standardization—the matching of taxon names and concepts in different data sources, and as a consequence plant scientists and the cyberinfrastructure they employ, are not prepared to provide high resolution identification of the taxa reported in literature and various botanical databases (such as occurrences, gene sequences, and traits). Tools are needed that will allow investigators to efficiently navigate a data landscape where one taxon might have many names (synonyms) and one name might refer to many taxa (taxon concepts). In addition, the system must encourage a continual update of taxonomic relationships by the taxonomic community. Members of the BIEN team have previously designed solutions to this problem using set theory mapping of the relationships among taxonomic concepts (a name as used by a specific authority) [45, 46]. Examples include design of the Taxon Concept Schema[15] recently adopted by TDWG as an international standard, and implementation of the core components in the VegBank[1] plot archive and in the SE Floristic Atlas[2]. What is missing is a cyberinfrastructure and component data that employs the critical taxon concept approach for taxon documentation and for integration of data from mixed sources that follow different taxonomic perspectives.

We propose three short-term demonstration projects aimed at solving the taxonomic barrier. First, we will create an approximation of an authoritative list of New World plants by combining a series of regional checklists, such as the USDA Plants list for North Americana north of the Mexican boundary, the Caribbean list of Acevedo (new version due in May 2009), the Missouri Botanical Garden's catalog of plants of Ecuador, Nicaragua, and Peru, Funk's catalog of plants of northeastern South America, and the Smithsonian Tropical Research Institute's Neotropical Tree Species Checklist[2]. We will attempt to place all the New World taxa reported in Tropicos either in the list or in the synonymy. Second, we will develop and deploy a taxonomic "scrubbing" tool for high-throughput detection and correction of taxonomic spelling and nomenclatural errors. This application will build on existing taxon name matching algorithms such as TaxaMatch[3] and TaxonScrubber[4]) and will use TROPICOS[5] and the IPNI Plant Names

[15] http://www.tdwg.org/standards/117/
[1] http://www.vegbank.org/
[2] http://www.herbarium.unc.edu/seflora/firstviewer.htm
[2] http://ctfs.si.edu/neotropicaltree/
[3] http://www.cmar.csiro.au/datacentre/irmng/
[4] http://www.salvias.net/pages/taxonscrubber.html
[5] http://www.tropicos.org/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Database[6] as authoritative name references. Third, we will conduct a comprehensive demonstration of the power of taxon concept mapping using the approximately 80,000 taxon concepts relationships documented for the flora of the Southeastern US in the the SE Floristic Atlas[13], expanded to allow selection of alternative taxonomic perspectives. The SE Floristic Atlas (SEFA) is the only large-scale online use of taxon concept relationships to integrate diverse occurrence data from many sources including museum collections, literature references and plot data. This regional-scale project is a working example of the sort of tools we propose to put at the service of the entire ecology and biodiversity community.

- A second major infrastructural problem arises from interaction and networking among data sources, which leads to serious challenges with regards to data quality and data provenance. For example, experts may detect and correct errors in the raw data for their own use, but this secondary improvement often does not flow back to the original data source. As a result efforts are wasted, original data sources remain uncorrected, and it is often challenging to determine what constitutes the best, or even, unique, set of information. A cyberinfrastructure is needed that allows seamless feedback between data providers and data users in a process of data annotation and correction. This feedback would include revision of data, real-time addition of new data, perfect archiving so that data available at a given date can be easily viewed for reanalysis, and feedback to the data sources with respect to suggested revisions.

We will process approximately 1,350 tropical forest plots from Central and South America, 400,000 North American vegetation plots, and perhaps 400,0000 museum collections to validate names and geocoordinates. We will implement a workflow where all suggested changes are piped back to the source for validation and we will subsequently track the level of response achieved to guide us in design of more efficient and user-friendly tools.

In order to demonstrate the feasibility of an enormous, cross-continent, taxon-occurrence data network, the BIEN team will immediately use the data networks created through these projects to begin addressing our guiding scientific questions. A table will be created for all plot data that contains geocoordinates, survey date, and abundance for each species; a parallel table giving geocoordinates, date, and species name will be assembled from the specimen data. The end result will be, for the first time, the creation of data resources containing standardized and error-checked geographic occurrence and abundance records of several tens of thousands of plant species in the Americas.

III.e –Long-term feasibility of a global BIEN: Overview – With iPlant, we aspire to develop a global data integration network where plant biologists from many different disciplines can upload, standardize, merge, and share data. This network will also serve as a permanent repository for legacy data and provide a benchmark against which to quantify the impact of climate change. The end product of such a network will be the ability to address questions at spatial and temporal scales far exceeding the reach of any individual research program.

Data integration at this scale will be an immense challenge, one that we believe will require an innovative hybrid solution with features of both a data warehouse and a data network. Proximately, as a data warehouse, the solution must allow for import and standardization of diverse data according to a common schema and vocabulary. Ultimately, as a distributed data network, the solution must empower the community to contribute, manage and update their own data. Engaging the community will be essential for long-term sustainability in the face of frequent updates.

[6] http://www.ipni.org/ipni/plantnamesearchpage.do
[13] http://www.herbarium.unc.edu/seflora/firstviewer.htm

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Figure 1 provides an overview of the central activities of the BIEN network. Producing an integrated data network will require several main steps (labeled in the figure) including:

1) ***Coordination of core botanical data streams****:* Provide tools for the creation, coordination and uploading (push) or harvesting (pull) of disparate data sources into the confederated iPlant Cyberinfrastructure, using common data communication protocols and exchange schema. The two main data streams are ecological plots/surveys and specimen records. In addition, we intend to also interact with outside groups representing the trait, phlogenetic and genomic communities.
2) ***Data integration and quality control:*** Here we will enable integration of disparate data according to a consistent model, while coupling concept-based taxonomic authority information with the raw data. This system identifies and iteratively refines taxonomic ambiguities and errors, while providing comprehensive feedback to original data provider.
3) ***Global scale, extensible, confederated database:*** Here we will create a web-based framework with data communication protocols and exchange schema that are compatible with a) broader ecological and environmental networks (e.g. NSF OCI/DataNet efforts and NSF OCI/Interop efforts in ecological and earth sciences, NEON), and b) other data frameworks (phylogenetic, traits, genomics)
4) ***Web-accessible end-user resource:*** Next, we propose to create a flexible and logical interface for powerful data querying, discovery and analysis, with download formats readily usable by all field of plant biology research
5) ***A Data-Discovery Environment (DDE)*****:** As discussed below, the main user interface of the coordinated BIEN network will be the "Data Discovery Environment" or DDE. The DDE is the end result of data integration, standardization, and confederation, accessed through a flexible querying interface that allows for deep exploration and analysis of the confederated database.
6) ***Iterative feedback to the community and data providers*****:** The BIEN project will not be static but instead will be a dynamic network. The process of #1-5 will be the creation of cyberinfrastructure (tools for data standardization, scrubbing, and exploration) that will allow for iterative feedback to the original data providers who can then modify their original data sources. Thus, over time, botanical diversity data are increasingly corrected and improved.

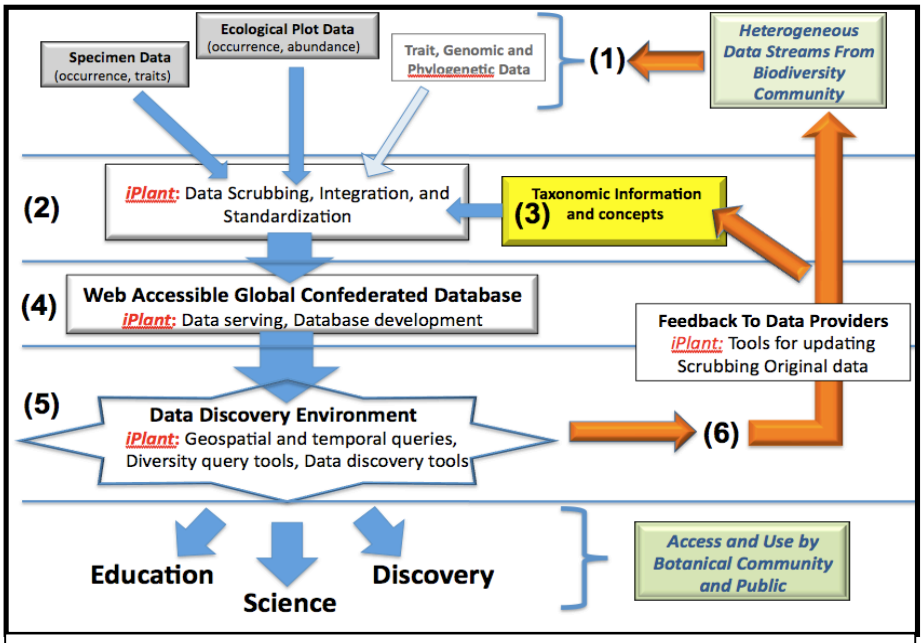

Figure 1. Overview of the proposed logical flow and important cyberinformatics steps in the creation of a Botanical Information and Ecology Network (BIEN). As described in the text, this global network is an iterative dynamic network that provides feedback to data providers as well as, over time, and increasingly improved source of standardized biodiversity data.

III. e - Long-term feasibility of a global BIEN: Data Discovery Environment - The main user interface of the coordinated BIEN network will be the "Data Discovery Environment" or DDE. The DDE will be the access point to the global confederated database for both academics and the general public. The DDE

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

will allow for: (*i*) multiple opportunities for outreach and education (see section V. below); (*ii*) allow the user to view the data as originally collected or under alternative taxonomies and phylogenies; (*iii*) to visualize the distribution and density of data points; (*iv*) to create species-lists with linked species attributes; and (*v*) to pool different data sources (biological observations, traits, physiology, climate) at differing temporal and spatial scales. Appendix B provides a more detailed outline of the steps involved in the data integration and discovery process. Below we detail the key steps in the creation of BIEN.

IV. Proposed long-term cyberinfrastructure: Tools and Web Services

Addressing the Grand Challenge question will require not only the compilation of data but also the *maintenance of a comprehensive, integrated and standardized global data network*. Creation and maintenance of this resource will require a cyberinfrastructure composed of numerous tools and services. We define cyberinfrastructure as highly extensible, broadly compatible, highly useful information resources that are compatible with but extend the existing technology solutions on which the community currently relies. Figure 1 provides an overview of five central sets of activities the BIEN team will need to support by identifying, modifying or creating tools and services.

Modern botanical science is being dramatically altered by access to expanding quantities of data. Our grand challenge represents but one of many such topics on organizing and serving of such large data quantities. Of special interest here is that the core components of each of the six central sets of activities identified in Fig. 1 will be of critical value to scientists addressing other questions and will be of much broader value beyond our proposed project. Importantly, the cyberinfrastructure we are proposing can be generalized to observations of all types or organisms at all spatial scales.

Proposed cyberinfrastructure to create tools to empower and motivate sharing of biodiversity data

- **Taxon concept mapping tools**
- **Taxon concept resolution services**
- **Tools for taxonomists mapping concepts**
- **Tools for mass import**
- **Tools for aggregators mapping concepts**
- **Tools for the community to contribute**
- **Tools for data integration**
- **Tools for taxon mapping and prediction**

IV.a. The creation, coordination, and digestion of multiple core botanical data streams, each conforming to a specified exchange schema - To assure efficient and accurate access to biodiversity data, those data must be provided via community-sanctioned protocols that are supported by a suite of tools for efficient data export, discover, revision, and import. The protocols and tools are needed in part to empower and motivate the community to share biodiversity, and in part to provide assured direct and efficient access to large quantities of quality-controlled, standardized data. Moreover, providers of data, while often willing to share data, generally do not have the resources to develop idiosyncratic exports for individual users, but need to employ a single data export mechanism. The plant biodiversity and ecological communities have made major progress in establishing necessary protocols, and several implementations of these have significantly improved access to specimen data in particular over the past decade (e.g. Darwin Core). Yet much work remains to be done as not all types of observation data are yet supported, current implementations are often spotty in their data holdings, and errors and redundancies in the data need to be resolved.

- Creation of standardized biodiversity data-exchange schemas - Accurate and efficient data migration and ingestion requires broad international acceptance and compliance with established data exchange standards. Only with acceptance and widespread application of data exchange standards will we be able to absorb data from a broad array of sources. We propose working with existing confederation schemas, or where necessary, developing novel schemas, that preserve the richness of information contributed by heterogeneous data providers, and implementing these in production-ready systems available to researchers around the world. This effort will resolve the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

complex challenge of integrating vegetation, collection and observation data collected over vast spatiotemporal scales, using numerous collection methodologies, and currently stored in disparate data systems.

- Tools to export, search and insert schema data - To efficiently employ biodiversity data streams conforming to a common schema, tools will be needed to generate those data streams, view and edit them, and absorb them into the greater cyber infrastructure In some cases standard schema are already supported, such as in the transmission of collection records conforming to DarwinCore. However, to achieve the critical buy-in from the community we will need to provide interface tools for the other data systems, such as the Specify[1]and EMu[2] systems for collection management and the TEAM[3] (Conservation International), VegBank, and TurboVeg systems for vegetation plot data. In addition protocols and scripts should be provided for incorporation in new data systems to ease communication.

A single, widely used data exchange standard—Darwin Core[4]— exists for biological collections data, although there are several common variants that are not completely compatible. Nevertheless, the widespread use of this standard via the DiGIR[5] and TAPIR[6] data exchange protocols will simplify extraction of data from specimen databases. Darwin Core, however, only provides a description of a subset of the ecologically relevant information that might be available for a specimen record, and falls far short of the content needed for a broader range of taxon occurrence records such as those from vegetation plots where, for example, ecologically important measurements of association and relative abundance can vastly enhance our ability to understand the mechanisms influencing distribution and co-existence of plant species. Overall, a much more complex exchange schema is needed to bring together the numerous formats in which vegetation data are captured, as well as to harmonize vegetation, taxon occurrence and specimen data. Fortunately, development of a confederated vegetation data exchange schema, provisionally titled VegX, is nearing completion and will soon be available as an official TWDG[7] and IAVS standard[8]. This schema will be deployed as the common template for the import of specimen and vegetation data into the core BIEN database (see *Data import*, below).

Exchange standards are required for several types of data. Fortunately, BIEN team members play leading roles in the on-going development of many of these standards including those for collection data, vegetation plot data, general taxon observation data, and taxon concept data, BIEN members closely involved in standards development would work closely with the iPlant technologists to assure that emerging international standards for these data types are appropriately incorporated into the iPlant products.

IV. b. Quality control and standardization of the content of the data streams - Access to large quantities of biodiversity observation data does not assure the data consistency, accuracy and integration needed to address the grand challenge question. Quality control of the data stream must be assured through standard services and workflows. We here identify several key components for which tools and workflows will be required.

- Tools for taxonomic scrubbing - Names of organisms are infamously prone to spelling errors and orthographic variants. The only solution is to match names against standardized lists. The two

[1] http://www.specifysoftware.org/Specify/specify/
[2] http://www.kesoftware.com/content/view/512/356/lang,en/
[3] http://www.teamnetwork.org/en/
[4] http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome
[5] http://digir.sourceforge.net/
[6] http://wiki.tdwg.org/TAPIR
[7] http://www.tdwg.org/
[8] http://www.bio.unc.edu/Faculty/Peet/vegdata/standards.htm

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

primary lists available are W3Tropicos[9] and IPNI[10], both of which have administrators who have expressed interest in cooperating with us, and work on a centralized nomenclature is underway at GBIF/EoL. Applications are needed to perform exact and fuzzy matching of names, correct spelling errors, map synonymies and quantify taxonomic uncertainty. Existing applications such as TaxaMatch[11] and our own Taxon Scrubber[12] point the way to more sophisticated solutions (see 15, 18]). In addition. Consistency will be greatly enhanced by adoption of a set of universal identifiers (GUIDS) for taxon names as advocated by TDWG[13] and already implemented by IPNI[14] in the form of LSIDs[16]. We will consistently apply taxon name GUIDS as available and work to unify such systems across name providers.

- Georeferencing Tools - Although latitude and longitude for the collection site are today standard standard components of a plant observation and collection records, most older specimens lack these data. At The New York Botanical Garden, only an estimated 30% of specimens include geocoordinates. In order for herbarium specimens to be used in vegetation analyses, they must be georeferenced. Adding coordinates to specimens retroactively is time-consuming, sometimes requiring as much time as digitization of all other specimen data. A cyberinfrastructure tool is needed to help supply these missing data, and to check existing geocoordinates against locality descriptions. Although several georeferencing applications are already available (e.g., Biogeomancer[17], SpeciesLink[18]), none are capable of the high-throughput georeferencing required by BIEN. One of our primary goals is to engage the community and build upon existing expertise wherever possible with the goal of, improving and assisting in the deployment of these applications as web services.

- Tool for the Detection of duplicates. Standard plant specimen techniques include collection in multiple sets and distributed to multiple herbaria. Though this practice is beneficial for users of individual herbaria, it can be an insidious source of error and pseudo-replication in analyses based on georeferenced specimen data. A cyberinfrastructure tool is needed to identify duplicates and remove all but one instance from a given analysis. Some work toward this end is already under way in the form of the Filtered-Push network[19] (see also http://mantis.cs.umb.edu/wiki/index.php/Main_Page)

- Tools to Compile Names of Collectors, Determiners, and Taxonomists. Authoritative lists of persons involved in collection and determination (plus linked GUIDS) are necessary to maximize data quality and consistency within the biodiversity community. This information will serve many purposes, the most important of which is inferring identification quality—an issue distinct from taxonomy. We will build upon existing resources within the community by accessing existing authoritative data sources such as the Harvard Names Database[20], TROPICOS[21], and the New York Botanical Garden's internal database of 150,000 plant collectors and taxonomists

IV. c. Taxonomic integration - Taxonomy is the common language for describing biodiversity, but as with any language, taxonomic names change over time as discoveries lead to new interpretations of how

[9] http://www.tropicos.org/
[10] http://www.ipni.org/
[11] http://www.cmar.csiro.au/datacentre/irmng/
[12] http://www.salvias.net/pages/taxonscrubber.html
[13] http://www.tdwg.org/
[14] http://www.ipni.org/
[16] http://en.wikipedia.org/wiki/LSID
[17] http://www.biogeomancer.org/
[18] http://splink.cria.org.br/tools?criaLANG=en
[19] http://www.tdwg.org/proceedings/rt/metadata/351/0
[20] http://asaweb.huh.harvard.edu:8080/databases/botanist_index.html
[21] http://www.tropicos.org/PersonSearch.aspx

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

biological entities should be classified. Millions of observations and collections of plant specimens exist that are "identified" through these taxonomic names, but since the meanings of these names can change, with the consequence that the interpretation and definitive understanding of "what occurred where" is compromised, yet there remains no clear-cut mechanism for updating determinations through time. Clarifying the meanings and relationships of names applied to organisms by different researchers at different times is a fundamental challenge when integrating biological data.

Ambiguity can arise due to spelling errors, variant spellings, nomenclatural synonymy, but also due to taxonomic revisions where splitting and lumping change the circumscription of specimens associated with names [19], *making taxonomic standardization not only a major challenge for BIEN, but also a major impediment to merging data within any area of the biological sciences where the taxon is the linking variable*. Methods and services for taxonomic standardization must be provided that successful navigate a data landscape where one taxon might have many names and one name might refer to many taxa. Moreover, solutions must support a world where different institutions have different preferred taxonomies, and the accepted solutions are constantly changing as new information becomes available, yet must be perfectly achieved to capture the content at any time in the past.

Resolution of the many-to-many relationship between names and taxon concepts is to be found in application of taxon concept relationships as first described by Berendsohn 1997 [47] and subsequently articulated by BIEN personnel in the recently adopted TDWG taxon concept schema and in related publications (e.g. [34, 35]), Variants of this schema have already been embedded in the VegBank plot data archive and in the SE Flora NatureServe Biotics and the Euro+Med flora biodiversity systems. Planning documents for a future release of USDA PLANTS also include taxon concepts. However, the major impediment to adoption of the approach is the need to create, update and resolve the relationships among taxon concepts for which specific tools and services are needed.

- Taxon concept mapping tools. Relationships among taxon concepts are routinely asserted by taxonomists when studying specific groups but are not captured in the data in any standardized format. In addition, aggregators of biodiversity information also make these decisions on a regular basis. We need to provide software tools to facilitate the capture and documentation of this process so that the information can be used for future data integration. As part of the SEEK project and to support the SE Atlas project, we developed a prototype taxon mapping tool, ConceptMapper[1]. ConceptMapper is a desktop tool to assist taxonomists to relate taxonomic concepts from one classification to another and to manage taxonomic concept metadata that precisely define taxonomic concepts. Concept data stored in the system can be retrieved, visualized and changed through the ConceptMapper user interface. Main functions include importing, exporting, querying and viewing concept data, adding and editing relationships, concepts, references and specimens.

- Taxon concept service. A service containing a central cache of taxon concepts and their relationships is needed so that users and data systems that wish to integrate data can find related concepts and make accurate matches. Preliminary design specifications and a prototype were developed as part of the recent SEEK project[2].

- Taxon concept matching tools. Users wishing to integrate datasets need to discover taxon concept relationships and be guided through the required integration decisions. In particular, we need a tool that increases our ability to import sets of concepts and find matches; suggesting when

[1] http://cvs.ecoinformatics.org/cvs/cvsweb.cgi/~checkout~/seek/projects/taxon/conceptmapper/
[2] http://seek.ecoinformatics.org/Wiki.jsp?page=SEEKTaxonCommunity

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

the ecological and biodiversity information can be merged and how to merge ambiguous matches. This is a complex task commonly required in meta-analysis.

IV.d. Data warehouse and network - Mechanisms and infrastructure must be provided to allow efficient access to and analysis of the available data and incorporation of other data types linked by commonality of taxon or location or both**.** Data integration at this scale will be an immense challenge, one that we believe will require an innovative hybrid solution with features of both a data warehouse and a data network. As a data warehouse, the solution must allow for import and standardization of diverse data according to a common schema and vocabulary and support efficient querying and data exploration. As a network, the solution must empower the community to contribute, manage and update their own data. Engaging the community will be essential for long-term sustainability in the face of frequent updates. We anticipate that the greater BIEN data network will be distributed across many institutions, and that this is unavoidable because of issues of ownership, confidentiality and funding. However, we anticipate drawing on the many sources to create a centralized data warehouse, optimized for user access and efficiency. This application will interact with all of the preceding applications and will manage mapping of source data to a common exchange schema, transfer to temporary staging tables, correction and standardization within the staging tables of taxonomy, locality information and geocoordinates, and final import and normalization to the BIEN core database. A user interface will allow for supervised execution of these processes, and will enable users to generate error reports that can be used to correct source data, thus providing feedback to the community and improving data quality during future imports.

IV.e. Data discovery and analysis environment **-** The data discovery environment will consist of a rich set of tools for user-driven data exploration, amalgamation and extraction. These tools will assist in the querying and compilation of candidate data sets and subsequent analysis. A common language of geospatial and phylogenetic queries will enable the user to build linkages between normally disparate datasets such as ecological inventories, specimen occurrences, trait databases and environmental observations.

- Tools for the implementation of workflows - Many types of data manipulation and analysis will be complex, yet will need to be repeated by numerous users. Examples include efficient phylogenetically structured queries, queries linking either children or parents, queries oriented around taxon concept resolution, flexible geospatial queries and data mapping, and integration of genomic, physiological or trait data. Often these may need to be established in advance to be run from a simple web interface. . We anticipate building on the Kepler[1] workflow environment and system developed for the Ecological and Geoscience communities. Such workflow infrastructure tools will allow automated analyses capability to be readily available to domain-level scientists and even the general public through web portals.
- Tools for data query and discovery – Two types of tools are necessary for performing data discovery: (1) A data environment allowing the extraction of any subset of data, and (2) a suite of analytical tools. Obviously useful subsets of the data would be single species or locations, but more complex queries, such as phylogenetic or trait-based, would also be useful. Analytical tools include mapping of many features, range prediction (thus rely on links to climatic data), and comparisons accounting to alternative taxonomic concepts.

[1] http://kepler-project.org/Wiki.jsp?page=KeplerProject

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

- Tools for taxon/clade mapping, exploration, and analytical range modeling - Features of the data discovery environment must also include: flexible geospatial querying and mapping of data, efficient, phylogenetically-structured queries, and the ability of the user to retrieve all records for child taxa linked to a particular parent taxon/clade that occurs within a user specified geographic region. The BIEN team is interested in also working with iPlant to also provide simultaneous access to public domain climate, geophysical and satellite data. In addition, we will incorporate new research [48-51] into how to best 'predict' the geographic range of a given taxon/calde. Range predictions based on widely-used ecological niche modeling techniques (see [48-51] for a review) will be constructed by combining taxon observations from plots and specimens with public domain climate, geophysical and satellite data. These predictions will be cached and available for searching in the same manner as actual taxon observations, enabling users to predict species occurrences and richness in unexplored regions. The current University of Arizona Biodiversity Informatics Initiative (BDII[2]) provides an example of the potential power of such applications.

Proposed Cyberinfrastructure for Data Discovery Environment

- **Flexible geospatial querying and mapping of data.**
- **Efficient, phylogenetically-structured queries, enabling the user to retrieve all records for child taxa linked to a particular parent taxon.**
- **Support of alternative taxonomies and phylogenetic perspectives**
- **Integration with phylogenetic, genomic, physiological and trait datasets**
- **Simultaneous access to public domain climate, geophysical and satellite data**
- **Analytical tools for range modeling**

IV. f. Update and checking tools - - Data are constantly changing. Any comprehensive and robust biodiversity cyberinfrastructure must efficiently handle the dynamic nature of data where new data and corrections to previously available data are seamlessly incorporated and integrated on a continuous basis. This will require either that sources push corrections to the system are queried for updates on a routine basis. In addition, if data are to be cited and available for reanalysis, it will be necessary to have perfect versioning so that the data source can be viewed as it was at any time in the past, a capability already incorporated into some data sources, such as VegBank. Finally, as errors are identified and as value is added to records, this information should be returned to the source. Indeed, many of the tools we will be developing will provide useful feedback to collections as they continue to digitize their specimens.

V. – Education, Outreach, and Training (EOT)

Our grand challenge question, "in a changing world, what grows where and why", provides a unique opportunity to build a core program in botanical education, outreach and training (EOT). The BIEN project's focus on biodiversity, taxonomy, and ecology will make a critical and inherently appealing aspect of plant biology available to a much larger audience. BIEN offers exceptional opportunities for generally raising awareness of the importance of plants and plant diversity for students as well as the lay public. The main goals for education, outreach, and training are:

1. To develop innovative cyberinfrastructure that will support advanced plant biology research, but also enable students ranging from K-12 through college as well as citizen scientists, to discover and explore botanical diversity throughout the world at differing scales (Amazon forest, Manaus Brazil, Pima county, Sabino Canyon etc.) using integrated taxononomies (Maple trees, *Acer* sp., Rosids etc.) and community-sanctioned information about functional forms and traits (grasses, orchids, trees,etc.; specific leaf area, seed size, flower color, plant size (dbh), etc.) of plants. Bringing together information on plant distributions, abundances, and taxonomic/phylogenetic relationships will

[2] http://loco.biosci.arizona.edu/bdii/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

provide a compelling and accessible entrée for students of all ages, into many other areas of plant biology.
2. To increase the participation of underrepresented groups within botany and plant ecology by partnering with ongoing efforts at: (i) NCEAS (ii) the Ecological Society of America Strategies for Ecology, Education, Diversity, and Sustainability (SEEDS[1]) program; and (iii) the New York Botanical Garden's undergraduate internship program, that specifically recruits from underrepresented groups in the Bronx and New York City in general
3. To promote the active participation of graduate students and postdoctoral researchers in education and outreach activities.
4. To educate the public, and K-12 students, on the effects of global climate change on plant distribution, diversity, and abundances.
5. To build upon existing outreach programs at NCEAS in order to educate botanists and ecologists in advanced knowledge representation techniques to facilitate data discovery and integration, and teaching the importance of informatics in their science.
6. To inform conservation planners, policy makers, agriculturalists, plant breeders, and the interested public via tools that quantitatively predict how changes in climate and land use may influence plant diversity and plant distributions at differing spatial and temporal scales. This information would be relevant to real- time decision and policy making.

We will develop these modules in conjunction with the iPlant Education and Outreach Team (see Appendix) by hosting EOT training workshops as well as developing targeted cyberinfrastructure that integrates with our Data Discovery Environment.

V. a. EOT Training Workshops - A highlight of the BIEN project will be the extent to which the tools are highly usable and understandable, by hiding the technical details of the advanced confederation of distributed data-- giving the appearance of a unified repository for accessing plant specimen and occurrence data, and grouping this information according to the latest and historical taxonomies of interest. Rapid adoption by the scientific community, however, will be greatly facilitated via training workshops, directed at "teaching the teachers" about these new tools. We believe that efforts directed at training the future generations of instructors provide maximum leverage of investments in technology transfer.

The National Center for Ecological Analysis and Synthesis, NCEAS, one of the major partners on this effort, is already well suited for hosting training workshops of this sort, with excellent on-site computing capabilities and support. We propose that two training workshops per year be held in the outer years of this project, once the new tools and approaches are deployable. This is to avoid the prospect of scientists investing their time in learning tools which are not yet stable nor highly usable. Our 'core team' is already well poised to interact with iPlant on the development of our EOT plan. Stephanie Hampton, Deputy Director of NCEAS, is already involved with education, outreach, and training in projects related to cyberinfrastructure development for the ecological sciences, and will provide strong connections of the BIEN EOT with these other efforts. Sandy Andelman, director of the TEAM initiative for Conservation International, has a long history of outreach to the conservation planning community and development of cyberinfrastructure for the ecological and conservation sciences. Barbara Theirs, director of the New York Botanical Garden, is actively involved in public botanical outreach to under represented groups and inner-city children. Other project personnel (Peet, Schildhauer) have had major involvement with EOT efforts on other large cyberinfrastructure projects for the ecological and plant sciences (NSF funded KDI and ITR programs[1]), and are familiar with the special challenges of technology training for this domain. The BIEN personnel look forward to working with the iPlant technologists, and in particular, iPlant

[1] http://www.esa.org/seeds/
[1] : http://knb.ecoinformatics.org; http://seek.ecoinformatics.org

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

experts in technology training, to develop and host workshops that catalyze awareness, proficiency, and adoption of these new tools by the plant science community.

V. b. Cyberinfrastructure tools for EOT - As discussed above, the EOT component of BIEN will develop, in partnership with iPlant, additional CI tools within the Data Discovery Environment (DDE). These CI tools will be important for community and public outreach, training faculty, training young researchers/future instructors, building community participation/discourse, and helping develop assessments for technology transfer. These tools (summarized in the CI for EOT Box) will enable a broad range of botanical discovery including: (*i*) the ability of diverse users (from K-12 to academics) to quickly identify plants online, as well as to uniquely explore the botanical diversity of different geographic regions and places ; and (*ii*) the enable the management of specimen and ecological observation data (in both herbaria as well as for ecologists in the field). The EOT CI should greatly assist with the development of model curricula, academic research beyond the immediate scope of the goals of BIEN, and citizen science. In short, the EOT CI tools will raise awareness of plant biodiversity at every educational level.

CyberInfrastructure for EOT

Data Discovery for Plant Diversity and Climate Change

- Citizen, student, and academic discovery and visualization tools based on level of interest and expertise.
- Geographic, climate, and future climate overlay tools for the generation of species/clade range and distribution maps, climate maps, and diversity maps.

Identification

- On-line taxon identification tools for academics, students, and citizen scientists.
- Generation of plant species lists, guides, images based on geographic queries and exploration.

Specimen/Observation management –

- Batch taxonomy updates.
- Specimen management - Identifying duplicate specimens.

We envision that the proposed EOT CI will enable researchers, students, and citizen scientists to access and discover patterns and reveal insights from the confluence of specimen data, species traits, species range maps, linkages between climatic attributes and plant distributions, as well as with well-defined taxonomy/phylogenetic relationships. Such CI tools will be a tremendous improvement to the current system of access to specimens and information of plant biodiversity. As a practical outreach example, the proposed CI tools will enable unique outreach to smaller-scale herbaria, such as those in the developing world. These herbaria are crippled by poorly-curated collections, limited resources to identify specimens, and almost no access to floristic treatments. Moreover, maintaining up-to-date taxonomy is nearly as slow as it was a century ago; taxonomic changes still need to be evaluated name-by-name, and updated by hand, specimen-by specimen. This creates a tremendous obstacle for specimen identification and training, as well as hampering local scientists' abilities to conduct ecological and botanical research. The proposed EOT CI would greatly facilitate the improvement, management, and networking of biodiversity data in the developing world.

V. c EOT Assessment - The BIEN PI's collectively have an outstanding record of developing and deploying advanced technology tools, as well as being highly productive and creative in their individual plant biology research specialties. Our EOT team and core team members well aware of the growing number of modalities for engaging with scientists in EOT, including the use of wiki-style sites for accreting and evolving understanding, videoteleconferencing for facilitating remote interaction, Flash and other methods for presenting scientific visualizations and bite-size snippets of technology transfer via the Web, etc. What is not clear is which of these approaches are most effective for communicating the advances on a project such as this. Thus, the BIEN personnel look forward to working with the iPlant technologists, and in particular, iPlant experts in technology and education training, not only to develop and host workshops to catalyze awareness, proficiency, and adoption of these new tools by the plant

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

science community, but also to investigate which advanced technology delivery methods are most effective for EOT on a cyberinfrastructure effort such as iPlant.

VII. Project Management:

Much of the CI development will be done with assistance from Working Groups containing expertise in plant species, representative user groups, education and outreach audiences, evaluation teams, etc. We propose five Core Activities Working Groups to define and motivate the primary science and cyberinfrastructural goals of this project:

A. **Science Working Group:** will oversee and set the agenda for BIEN so that the development of cyberinfrastructure is guided by Science needs. This group will also initiate and conduct, with collaboration of with BIEN team members and collaborators, the research underlying our core research questions.
B. **Specimens, Plots, and Occurrences**: to bring together the significant data resource providers on this proposal, identify potential future providers, and clarify domain-related challenges to integrating these disparate data sources
C. **Confederation Data Model and Exchange Schema**: working in close conjunction with Group A, to develop a formal model that can integrate the relevant data resources, and provide a well-specified set of protocols to allow for future community participation
D. **Taxonomy**: to focus on specifying and resolving the taxonomic names issues with a robust exchange schema
E. **Georeferencing**: to focus on developing a standards-compliant approach to resolving and harmonizing any of the plant data having a georeferenced context.
F. **Data System Features and usability**: work closely with Working Group A to help define the end-user requirements for this framework, to assure their relevance in meeting the most critical needs of the targeted research audience especially in the Data Discovery Environment.

We also propose two Synergistic Activities Working Groups, to acknowledge the importance of linkages with these other areas of concern, and to potentially feed into the core cyberinfrastructure development activity:

A. **Phylogenetic Issues**: with much closely aligned work focused on phylogenetic analyses of plant communities, this working group will help identify and potentially coordinate participation of technology efforts in phylogenetics with BIEN.
B. **Trait/Genomic Integration**: many efforts are underway to more effectively enabling querying on functional traits of vegetation rather than taxonomic names, and this working group will identify and potentially coordinate the participation of those efforts with BIEN.
C. **Education Training and Outreach (EOT)**: As described above, this group will work with iPlant in the design of training workshop and digital training tools. These tools will be important for (*i*) mapping global distribution of plant distributions, assessing the linkages between climate change and plant diversity, and specimen management tools for herbaria and their associates; (*ii*) for conservation biology professionals; (*iii*) for instructors/educators to use the tool. In addition, this group will help design geographic discovery tools for predicting species in a polygon that can be used locally anywhere in the Americas, and then a usability assessment in several settings, Linkages to information to aid in identification

The associated domain scientists as well as key community personnel associated with each working group are detailed in the Appendix C. Once work groups are formed, a joint workshop will be held with other GC Teams with common CI interests and iPlant staff. We envision that iPlant could sponsor at least 1-2 meetings per year for each of the Working Groups., Several of these Working Groups might meet in

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

close succession or simultaneously, to enable program-wide coordination of their efforts. (see suggested relative time line on Appendix D).

VI. b. Project progress monitoring and evaluation plan - The BIEN core team will establish an external evaluation team comprised of plant biology community. This evaluation team will be responsible for overseeing that the central goals as finally agreed upon between BIEN and iPlant are effectively and efficiently implemented, but also that the needs of the representative botanical communities are best met. Further, external evaluators will be responsible for assessing Cyberinfrastructure as well as educational and outreach.

VIII – Impact of successful infrastructural development on the broader field

Taxonomy underlies all biology and will be central to the development of synthetic biodiversity science at iPlant. We propose to address the taxonomy challenge head-on and provide tools that assure the necessary matching of names and concepts. Solving the 'taxonomy problem' is the key enabler and would be widely useful to botanists and zoologists alike. Indeed, we see taxonomy tools as central to integration within all of iPlant. We will work closely with other groups within iPlant and elsewhere (TRAITNET, TRY, TEAM, Tree of Life, Bar Code Consortium) to assure that taxonomic tools we create are widely used. We emphasize that it is the *integration* of taxonomic information with the significant information resources being brought together under this proposal that will significantly improve access and use of biodiversity data within the plant sciences. Indeed we see taxonomy tools as central to integration within all of iPlant, and a key requirement for creating the massive, global-scale confederation of plant biodiversity data that we propose here. The PI's and associated collaborators represent the range of institutions, authority, and competencies to provide the iPlant engineers with top-notch expertise in plant ecology and biodiversity cyberinfrastructural needs, along with close awareness of the relevant existing technological implementations underway within these areas. This combination of expertise is what we believe is necessary to create a relevant and transformative information resource for plant biology.

The BIEN network will provide biologists, ecologists, and conservation biologists easy access to widely dispersed data that everyone knows about but few can ever work with. The BIEN tools will be valuable to virtually every branch of ecology, even if simply to provide background information on a species of interest. Finally, by working with iPlant, we hope to remain near the forefront of bioinformatics, developing software that helps to assure the integrity, accuracy, and timeliness of widely used data. At least some of these tools may be relevant in other aspects of informatics.

Appendix A - Tables

Table 1. BIEN core team and collaborators outside the team of PIs. Core Team members responsibilities involve serving as a central domain scientist in one or more of the core working groups as well as overseeing the development of cyberinfrastructure. Collaborators may serve in BIEN working groups and/or advice the development of tools. All have pledged data or collaboration in the development of software tools and to represent the botanical communities that they represent.

Name	Institution	Core Team or Collaborator	Attended BIEN Workshop	Contribution
Sandy Andelman	Conservation International	Core	x	Data, tools, outreach, science
Jeannine Cavender-Bares	University of Minnesota	Core	x	Data, science

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Name	Institution	Core Team or Collaborator	Attended BIEN Workshop	Contribution
Steven Dolins	Bradley University	Core	x	IT expert
James Edwards	Encyclopedia of Life	Collaborator		Tools
Stephanie Hampton	NCEAS[2]	Core		Outreach, Education
John Janovecs	Botanical Research Institute of Texas, ATRIUM	Collaborator		Data, tools
Peter Jørgensen	Missouri Botanical Garden	Core	x	Data & taxonomy
Jessie Kennedy	Napier University (UK)	Core		Taxonomy, tools
Nick King	Global Biodiversity Information Facility (Denmark)	Collaborator		Data, tools
James Macklin	Harvard University Herbarium	Collaborator		Taxonomy, tools
Patrick Miles	U.S. Forest Service	Collaborator		Data
Brian McGill	University of Arizona	Core	x	informatics, science
Oliver Phillips	Leeds University (UK)	Core	x	Data, science
Tony Rees	CSIRO[1] (Australia)	Collaborator		Tools
Hans ter Steege	National Herbarium (Netherlands)	Core	x	Data, science
Corine Vriesendorp	Field Museum	Core	x	Outreach, Education
Nathan Swenson	Michigan State University	Core	x	Data, science
David Vieglais	University of Kansas	Core		Tools
Susan Wiser	Landcare Research (New Zealand)	Core	x	Data standards
Kerry Woods	Bennington College	Collaborator	x	Data
Josh Madin	ARC-NZ Research Network for Vegetation Function[3], & Computational Ecology Group, Macquarie University (Australia)	Collaborator		IT, Tools

2 National Center for Ecological Analysis and Synthesis, US
1 Commonwealth Scientific and Research Organization, Australia
3 http://www.vegfunction.net/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Table 2. Plots and other vegetation censuses currently available for the BIEN short-term data integration network. We also list external collaborators (*) who have been contacted and are willing to share data but at this point are satisfied with remaining external to the BIEN group.

Organization	Location	Contact	Number of units	Unit size (ha)
CTFS[7]	S. America	R. Condit	53	1
RAINFOR[8]	S. America	O. Phillips	252	1
ATDN[9]	S. America	H. ter Steege	251	1
SALVIAS[10]	S. America	B. Enquist	233	0.1
TEAM[11]	S. America	S. Andelman	90	1
Missouri Botanical Garden	Bolivia	P. Jørgensen	578	0.1
BRIT/Atrium[12]	Peru	J. Janovecs	81	0.1
US Forest Service, FIA[13]	USA	P. Miles	300,000	0.01
US National Park Service	USA	C. Lea*	5,000	0.1
US Forest Service, NRS[14]	Michigan	K. Woods	445	0.01
VegBank	USA	R. Peet	21,000	0.1
West Virginia Heritage	WV	J. Vanderhorst*	2,900	0.1
Virginia Heritage Program	VA	K. Paterson*	3,900	0.1
US Forest Service, Landfire	USA	D. Long*	365,896	various
Carolina Vegetation Survey	S.E. USA	R. Peet	8200	0.1

7 Center for Tropical Forest Science
8 Amazon Forest Inventory Network
9 Amazon Tree Diversity Network
10 Synthesis and Analysis of Location Vegetation Inventories
11 Tropical Ecology Assessment and Monitoring Network
12 Botanical Research Institute of Texas
13 Forest Inventory and Analysis
14 Northern Research Station

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Table 3. Herbarium, museum and other occurrence records presently available for the BIEN data integration program. Most are original sources of data, but GBIF is a secondary collection of the other sources, plus many others. We also list external collaborators (*) who have been contacted and are willing to share data but at this point are satisfied with remaining external to the BIEN group.

Organization	**Contact**	**Number of records**
Missouri Botanical Garden	Peter Jørgensen	3,000,000
New York Botanical Garden	Barbara Thiers	900,000
Smithsonian Institution	Warren Wagner*	1,000,000
University of Arizona	Brad Boyle	173,000
Harbard University Herbaria	James Macklin	200,000
Field Museum	Robin Foster*	90,000
University of Aarhus	Jens Christian Svenning	100,000
Utrecht University	Hans ter Steege	114,000
University of North Carolina	Robert Peet	110,000
GBIF[1]	Nick King	65,626,000 (animals & plants)

[1] Global Biodiversity Information Facility

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Appendix B. The following outlines the work-flow and principal applications of the proposed BIEN Data Integration and Discovery Network.

(I). DATA INTEGRATION

1. **Import.** Primary data are mapped and imported to VegX schema-compliant staging tables. Only minimal standardizations needed to match source data to schema are performed at this stage. Plots and other taxon observations will generally require a separate mapping script for each dataset; most specimen dataset can be imported from existing Darwin Core extracts.
2. **Primary standardization.** Standardization is performed independently for each data source. Basic error-checking performed at this stage. No versioning. Some user intervention required via import/error-checking interface. Error reports are generated for correction of source data. Taxonomic and geographic errors must be corrected prior to normalization.
    a. **Taxonomy**. Taxa are checked to ensure they match to nomenclaturally-valid names stored in core database. Only spelling errors and nomenclaturally-invalid names are corrected.
    b. **Geography**. Coordinates checked for numeric errors (out of bounds, etc.). Coordinates checked against locality fields to detect mis-matches. Political divisions are checked against to ensure they match to standard values in core database.
    c. **Missing data**. User is prompted to provide missing data, especially metadata.
3. **Normalization.** Data from staging table are merged field by field to normalized core database.
4. **Secondary standardization & indexing.** These are performed once data is in core database. User intervention required. Changes are versioned from this point forward.
    a. **Taxonomy.** Taxonomic concepts are specified and taxonomic synonymys are adjusted at the discretion of the data owner, according to authority lists linked to core database.
    b. **Georeferencing.** Georeferencing tools may be used to add coordinates to non-georeference specimens at this stage
    c. **Duplicate detection.** Duplicate specimen records are detected and removed.
    d. **Indexing of collectors and determiners.**
5. **Data management.** A rich user interface will allow users to perform ongoing management and update of data directly within core database. Existing data sets may be imported in steps 1-4 above, or entered directly via data management interface.
    a. Correction of existing data.
    b. **Direct entry of new data** within controlled environment of data model, including both addition of new observations to existing data, or entry of entire new data sets.
    c. **Tracking of determinations from voucher specimens.** Updated identifications may be applied to ecological observations by monitoring determination status of voucher specimens deposited as herbarium specimens.
    d. **Adjustment of taxonomy.** Changes in taxonomic status can be tracked and propagated across data sets.
    e. **Upload of additional media.** Additional media linked to observations (images, recordings, environmental measurements) may be uploaded on an ongoing basis.
    f. **Data access.** User sets access level and field embargos (if any) for data Access levels are: 1-hidden; 2-metada visible, data by request only; 3-full data freely available. These permissions may be set as defaults, or assigned by the data owner to particular users for particular datasets. Record and field-specific embargoes may also be applied if necessary (e.g., locality fields hidden to protect threatened and endangered species). See The

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

SALVIAS Project[15] and VegBank[16] for working examples of user-managed data access and field embargoes.

(II). DATA DISCOVERY

1. **Data selection**. User may query by any attribute in database, including spatial joins and map browsing.
2. **Data sharing.** User may request access to full data for any restricted datasets or embargoed fields (e.g., endangered species). Interface facilitates direct communication between data owners and requesting parties. Data owners can track data access logs.
3. **Additional taxonomic standardization.** Alternative synonymies and taxonomic concepts may be applied to aggregated data.
4. **Additional data sources.** Data external to BIEN can be linked to primary data and accessed via taxonomy or geography (spatial joins).
    - Phylogeny. Taxa can be mapped to alternative phylogenies, and searched via hierarchically-structured queries capable of retrieving all ancestors (parents) or all descendents (children) of a given taxon.
    - Traits and physiological data
    - Molecular sequence data
    - Climate or other environmental data
    - Satellite imagery
5. **Analysis tools**. Examples include:
    - Mapping
    - Range modeling, under current, past and projected climate scenarios
    - Statistics of diversity and abundance
    - Ordination
    - Phylogenetically structured analyses (e.g., phylogenetic diversity)
6. **Download.** Full data available for download in a variety of formats.
7. **Versioning.** Users may explore alternative versions of data. Content of downloaded datasets are timestamped and archived, and can be retrieved and examine at any time.

[15] www.salvias.net
[16] www.vegbank.org

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Appendix C. Working groups and proposed membership for major components of the BIEN Data Integration and Discovery Network. Each working group consists of both IT experts with a record of relevant application development and domain scientists familiar with the principal cyberinfrastructural challenges. BIEN project leaders and core team members indicated by asterisk.

Role	Name	Institutional affiliation	Relevant applications or activities
1. Specimens, Plots and Observations Working Group			
Domain/IT	Bob Peet*	University of North Carolina	VegBank[17]; SE Floristic Atlas[18]
Domain	Peter Jorgensen*	Missouri Botanical Garden	TROPICOS[19]; The Madidi Project[20]
Domain	Rick Condit	CTFS, Smithsonian	
Domain/IT	Brad Boyle*	University of Arizona	SALVIAS[21], BDII[22]
Domain/IT	Brian Enquist	University of Arizona	SALVIAS
Domain/IT	Bob Magill	Missouri Botanical Garden	TROPICOS
IT	Chris Freeland	Missouri Botanical Garden	TROPICOS
IT	Jesse Kennedy*	Napier University	TDWG[23]
2. Data Model and Exchange Schemas Working Group			
Domain/IT	Bob Peet*	University of North Carolina	VegBank; SE Floristic Atlas
Domain/IT	Brad Boyle*	University of Arizona	SALVIAS, BDII
Domaiin	Susan Wiser*	Landcare Research, New Zealand	NVS[24]
IT	David Hearn	University of Arizona	BDII
IT	Mark Schildhauer*	NCEAS	TDWG-OSR[38]
IT	Josh Madin	Macquarie University	ARC-NZ Research Network for Vegetation Function[39], and Computational Ecology Group
Domain/IT	Nick Spencer	Landcare Research, New Zealand	Vegetation Observations Exchange Schema (VegX)[25], TDWG

17 http://www.vegbank.org
18 http://www.herbarium.unc.edu/seflora/firstviewer.htm
19 http://www.tropicos.org/
20 http://www.mobot.org/MOBOT/Research/madidi/
21 http://www.salvias.net
22 http://loco.biosci.arizona.edu/bdii/
23 http://www.tdwg.org/
24 http://nvs.landcareresearch.co.nz/
39 http://www.vegfunction.net/
25 http://wiki.tdwg.org/twiki/bin/view/Vegetation/WebHome

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Domain/IT	Miguel Cáceres	Universitat de Barcelona, Spain	VegX[25], VegAna[30]
Domain/IT	Martin Kleinkamp	Bundesamt für Naturschutz (BfN), Germany	VegX[25], VegetWeb
IT	Jesse Kennedy*	Napier University	TDWG
IT	Dave Vieglais*	University of Kansas Biodiversity Research Center	Darwin Core[26], DiGIR[27], Specify[28], Plantcollections.org[29]

3. Taxonomy Working Group

Domain	Bob Peet*	University of North Carolina	VegBank
Domain	Jerry Cooper	Landcare Research (New Zealand)	TWDG-Darwin Core[26], TCS[30]
Domain	Peter Jorgensen*	Missouri Botanical Garden	TROPICOS; taxonomic specialist
Domain	Barbara Thiers*	New York Botanical Garden	Taxonomic specialist
IT	Jesse Kennedy*	Napier University	TDWG
IT	Dave Vieglais*	University of Kansas Biodiversity Research Center	Darwin Core, DiGIR, Plantcollections.org
Domain/IT	Bob Magill	Missouri Botanical Garden	TROPICOS, taxonomic specialist
Domain/IT	Brad Boyle*	University of Arizona	SALVIAS, TaxonScrubber[31]
Domain/IT	Tony Rees	CSIRO	TaxaMatch[32]

4. Data System Features and User Interface Working Group

Domain/IT	Bob Peet*	University of North Carolina	VegBank
Domain	Peter Jorgensen*	Missouri Botanical Garden	TROPICOS; The Madidi Project
Domain/IT	John Janovec	BRIT	Atrium[41]
IT	Matthias Tobler	BRIT	Atrium[41]

30 http://biodiver.bio.ub.es/vegana/
26 http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome
27 http://digir.sourceforge.net/
28 http://www.specifysoftware.org/Specify
29 http://plantcollections.pathf.com/
30 http://www.tdwg.org/standards/117/
31 http://www.salvias.net/pages/taxonscrubber.html
32 http://www.cmar.csiro.au/datacentre/irmng/
41 http://atrium.andesamazon.org/
41 http://atrium.andesamazon.org/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

IT	Dave Vieglais*	University of Kansas Biodiversity Research Center	Specify
IT	Mark Schildhauer*	NCEAS	SEEK[33], EarthGRID

5. Phylogenetic Issues Working Group

Domain/IT	Michael Sanderson	University of Arizona	TOL[34], Phylota Browser[35], BDII
Domain/IT	Cam Webb	The Arnold Arboretum of Harvard University	Angiosperm Phylogeny Website[36], TOL
Domain	Peter Stevens	Missouri Botanical Garden	Angiosperm Phylogeny Website
Domain	Brad Boyle	University of Arizona	BDII
Domain/IT	Reed Beeman	University of Florida	TOL

6. Trait Integration Working Group

Domain	Brian Enquist*	University of Arizona	SALVIAS, TraitNet
Domain	Jeannine Cavendar-Bares*	University of Minnesota	Traitnet[37]
Domain	Nate Swenson	Michigan State University	SALVIAS, TraitNet
IT	Mark Schildhauer*	NCEAS	TraitNet, TDWG-OSR[38]
IT	Josh Madin	Macquarie University	ARC-NZ Research Network for Vegetation Function, and Computational Ecology Group

7. Education and outreach working group

IT	Matthias Tobler	BRIT	ATRIUM
Domain	Corrine Vriesendorp*	Field Musuem	Field Tropical Plant Guides[39]
Domain	Sandy Andelman*	TEAM, Conservation International	Conservation outreach
Domain	Mark Schildhauer*	NCEAS	Informatics, computational, and IT tools
Domain	Stephanie Hampton*	NCEAS	Education outreach, NCEAS and Ecological Society of America
Domain	Robert Peet*	University of North Carolina	Informatics tools
Domain	Barbara Theirs	New York Botanical Garden	Public outreach

[33] http://seek.ecoinformatics.org/
[34] http://www.tolweb.org/tree/
[35] http://loco.biosci.arizona.edu/pb/
[36] http://www.mobot.org/mobot/research/apweb/welcome.html
[37] http://www.columbia.edu/cu/traitnet/
[38] http://www.tdwg.org/activities/osr/
[39] http://fm2.fieldmuseum.org/plantguides/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

8. Science working group			
Domain	Brian Enquist	University of Arizona	Scaling approaches, macroecology and physiological/functional ecology
Domain	Rick Condit	Center for Tropical Forest Science	Macroecology and tropical ecology
Domain	Brian McGill	University of Arizona	Macroecology, ecoinformatics and biogeography
Domain	Nate Swenson	Harvard University Herbaria	Plant physiological/functional ecology, phylogenetic approaches in botany.
Domain	Jeannine Cavendar-Bares*	University of Minnesota	Plant evolutionary biology, phylogenetic ecology, and trait based ecology
Domain	Sandy Andelman	TEAM, Conservation International	Conservation Biology and population biology.

Appendix D

Proposed initial timeline for the BIEN project.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Relative Timeline											
Preparation Stage											
Create five core working groups	■										
Create synergistic Activities group	■										
Create External Evaluation Team	■										
Build BIEN Network											
Identify data sources	■	■	■								
Compile data	■	■	■								
Develop tools for data standardization				■	■	■	■				
Develop tools for feedback to data providers				■	■	■	■				
Create Data Discovery Environment								■	■	■	■
Education, Outreach and training											
Develop tools for data providers									■	■	
Develop tools for conservation biology									■	■	
Provide links to identification tools											■
Training sessions for users											■
Meetings											
iPlant Challenge/BIEN Team	■		■		■		■		■		■
Working groups (core and synergistic)	■		■		■		■		■		■
External evaluation	■				■				■		■

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

Literature Cited

1. Brown, J.H., *On the Relationship Between Abundance and Distribution of Species.* American Naturalist, 1984. **124**(#2): p. 255-279.
2. Lomolino, M.V., B.R. Riddle, and J.H. Brown, *Biogeography*. 2006: Sinauer.
3. McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D. , Anderson, M.J., Benecha, H.K., Dornela, M., Enquist, B.J. , Green, J.L., He, F., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling , A., Soykan, C.U., Ugland, K.I. and E. P. White, *Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework.* Ecology Letters, 2007: p. 995-1015.
4. Woodward, F.I., *Climate and plant distribution*. 1987, Cambridge: Cambridge University Press.
5. Nobel, P.S. and E.G. Bobich, *Plant frequency, stem and root characteristics, and CO_2 uptake for Opuntia acanthocarpa: elevational correlates in the northwestern Sonoran Desert.* Oecologia, 2002. **130**: p. 165-172.
6. Osmond, C.B., et al., *Stress physiology and the distribution of plants.* Bioscience, 1987. **37**: p. 38-48.
7. Feild, T.S., T. Brodribb, and M. Holbrook, *Hardly a relict: Freezing and the evolution of vesselless wood in winteraceae.* Evolution, 2002. **56**(3): p. 464-478.
8. Swenson, N.G. and B.J. Enquist, *Ecological and evolutionary determinants of a key plant functional trait: Wood density and its community-wide variation across latitude and elevation.* American Journal of Botany, 2007. **94**(3): p. 451-459.
9. Sebastiaan, V., et al., *Life-history traits are correlated with geographical distribution patterns of western European forest herb species.* Journal of Biogeography. **34**: p. 1723-1735.
10. Condit, R., S.P. Hubbell, and R.B. Foster, *Changes in tree species abundance in a Neotropical forest: Impact of climate change.* Journal of Tropical Ecology ;, 1996. **12**(pt.2)): p. 231-256.
11. Eckhart, V.M., M.A. Geber, and C.M. McGuire, *Experimental studies of adaptation in Clarkia zantiana I. Sources of variation across a subspecies border.* . Evolution, 2003. **58**: p. 59-70.
12. Wilczek, A.M., et al., *Effects of Genetic Perturbation on Seasonal Life History Plasticity.* Science, 2009. DOI: 10.1126/science.1165826.
13. Pauwelsa, N.M., et al., *When population genetics serves genomics: putting adaptation back in a spatial and historical context.* Current Opinion in Plant Biology, 2008(2): p. 129-134.
14. Leimu, R. and M. Fischer, *A meta-analysis of local adaptation in plants.* PLoS ONE, 2008. **3**: p. e4010.
15. Hooper, D.U., et al., *Effects of biodiversity on ecosystem functioning: A consensus of current knowledge.* Ecological Monographs, 2005. **75**(1): p. 3-35.
16. Bergengren, J.C., et al., *Modeling global climate-vegetation interactions in a doubled CO2 world.* Climatic Change, 2001. **50**(1-2): p. 31-75.
17. Williams, J.W., et al., *Rapid vegetation responses to past climate change.* . Geology 2002(30): p. 971-974. .
18. Williams, J.W. and S.T. Jackson, *Novel Climates, No-Analog Plant Communities, and Ecological Surprises: Past and Future.* Frontiers in Ecology and Evolution 2007(5): p. 475-482. .
19. Shafer, S.L., P.J. Bartlein, and R.S. Thompson, *Potential changes in the distributions of western North America tree and shrub taxa under future climate scenarios.* Ecosystems ;, 2001. **4**(3): p. 200-215.
20. Stephen P. Hubbell, S.P., et al., *How many tree species and how many of them are there in the Amazon will go extinct?* PNAS, 2008. **105**: p. 11498-11504.
21. Colwell, R.K., et al., *Lowland Biotic Attrition in the Wet Tropics Global Warming, Elevational Range Shifts, and Lowland Biotic Attrition in the Wet Tropics.* Science, 2008. **322**: p. 258-261.
22. Loiselle, B.A., et al., *Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes?* 2008. **35**: p. 105-116.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

23. Myers, N., et al., *Biodiversity hotspots for conservation priorities.* . Nature, 2000(403): p. 853-858.
24. Pitman, N.C.A., et al., *Dominance and distribution of tree species in upper Amazonian terra firme forests.* Ecology, 2001. **82**: p. 2101-2117.
25. Weiser, M.D., et al., *Latitudinal patterns of range size and species richness of New World woody plants.* Global Ecology and Biogeography, 2007. **16**(5): p. 679-688.
26. Williams, P.H., K.J. Gaston, and C.J. Humphries, *Mapping biodiversity value worldwide: Combining higher-taxon richness from different groups.* Proceedings of the Royal Society of London Series B-Biological Sciences, 1997. **264**(1378): p. 141-148.
27. NRC, *NEON: Addressing the nation's environmental challenges.* . 2003, Washington, DC: National Research Council.
28. Bortolus, A., *Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology.* Ambio, 2008. **37**: p. 114-118.
29. Reich, P.B., et al., *The evolution of plant functional variation: Traits, spectra, and strategies.* International Journal of Plant Sciences, 2003. **164**(3): p. S143-S164.
30. Gentry, A.H., *Changes in plant community diveristy and floristic composition on environmental and geographic gradients.* Annals of the Missouri Botanical Garden., 1988. **75**: p. 1-34.
31. Gentry, A.H., *Tropical forest biodiversity: distributional patterns and their conservational significance.* Oikos, 1992. **63**: p. 19-28.
32. ter Steege, H., et al., *Continental-scale patterns of canopy tree composition and function across Amazonia.* Nature, 2006. **443**: p. 444-447.
33. Berendsohn, W.G., *A taxonomic information model for botanical databases the IOPI Model.* . Taxon, 1997. **46**: p. 283-309.
34. Kennedy, J., R. Kukla, and T. Paterson, *Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration*, in *Data Integration in the Life Sciences: Proceedings of the Second International Workshop,*, B.L.R. Ludäscher, Editor. 2005, DILS 2005, LNBI: San Diego, CA.
35. Franz, N.M., R.K. Peet, and A.S. Weakley, *On the use of taxonomic concepts in support of biodiversity research and taxonomy.* , in *The New Taxonomy. Systematics Association Special Volume (74) Symposium Proceedings,*, Q.D. Wheeler, Editor. 2008, Taylor & Francis: Boca Raton, FL, . p. 63-86.
36. Enquist, B.J. and K.J. Niklas, *Invariant scaling relations across tree-dominated communities.* Nature, 2001. **410**: p. 655-660.
37. Swenson, N.G., et al., *The influence of spatial and size scale on phylogenetic relatedness in tropical forest communities.* Ecology, 2007. **88**(7): p. 1770-1780.
38. Rabeler, R. and R.K. Macklin, *Herbarium networks in the United States: Towards creating a toolkit to advance specimen data capture.* Collection Forum, 2007. **21**: p. 223-231.
39. Stephan, M.H. and J.H.J. Schaminée, *TURBOVEG, a comprehensive data base management system for vegetation data* Journal of the Vegetation Science, 2001. **12**: p. 589-591.
40. Violle, C., et al., *Let the concept of trait be functional!* Oikos, 2007. **116**: p. 882-892.
41. McGill, B., et al., *Rebuilding community ecology from functional traits.* Trends in Ecology & Evolution, 2006. **21**: p. 178-185.
42. Westoby, M., et al., *Plant ecological strategies: Some leading dimensions of variation between species.* 2002.
43. Wright, I.J., et al., *The worldwide leaf economics spectrum.* Nature, 2004. **428**(6985): p. 821-827.
44. Poorter, L., et al., *Are functional traits good predictors of demographic rates? Evidence from five neotropical forests.* Ecology, 2008. **89**(7): p. 1908-1920.
45. Franz, N., R.K. Peet, and A.S. Weakley, *On the use of taxonomic concepts in support of biodiversity research and taxonomy.* 2008.
46. Franz, N.M. and R.K. Peer, *Toward a language for mapping relaionships among taxonomic concepts.* Systematics and Biodiversity, 2008. doi:10.1017/S147720000800282X

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ:

47. Berendsohn, W.G., *A taxonomic information model for botanical databases: the IOPI Model.* . Taxon, 1997. **46**: p. 283-309.
48. Graham, C.H., et al., *New developments in museum-based informatics and application in biodiversity analysis.* Trends in Ecology and Evolution 2004. **19**: p. 497-503.
49. Elith, J., et al., *Novel methods improve prediction of species' distributions from occurrence data.* Ecography, 2006. **29**: p. 129-151.
50. Randin, C.F.e.a., *Are niche-based species distribution models transferable in space?* . Journal of Biogeography, 2006(33): p. 1689-1703.
51. Philips, S.J., *Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al.* . Ecography, 2008. **31**: p. 272-278.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2615v2 | CC BY 4.0 Open Access | rec: 6 Dec 2016, publ: