

2 Introduction and Objectives of the Research

Understanding and sustaining the natural world in the 21st century depends on improving our capacity to access ecological, earth science, and human-dimension data; mining these data for new knowledge; and conveying new insights to decision-makers and the general public. Computer science and IT research can effectively address many of these issues and advance our ability to conduct ecological science. We propose a multidisciplinary research investigation to create a “Science Environment for Ecological Knowledge” (SEEK)—an IT framework and infrastructure that will be used to derive and extend ecological knowledge by facilitating the discovery, access, integration, interpretation, and analyses of distributed ecological information. SEEK will provide for the integration of local desktop data with a larger network of data and analytical tools, enabling ecologists and other researchers to tackle complex research problems that were hitherto intractable. Key computer science and IT challenges addressed in this effort are (cf. Figure 1):

- *Analysis and Modeling (AM) System*—Developing a *modeling language for ecological analysis* based on *parameter ontologies* that provide an extensible and adaptable vocabulary of ecology concepts and their relationships, and *analytical workflows*, which chain together analysis steps and bind them to suitable datasets
- *Semantic Mediation (SM) System*—Extending XML-based data integration and mediation technology by merging it with formal ontologies and other logic-based knowledge representation formalisms to facilitate *semantic mediation* over hard-to-relate schemas of data sources
- *EcoGrid (EG)* — Providing the Grid infrastructure for seamless access and manipulation of ecology data and tools by extending the expressiveness of service description languages (e.g., WSDL [1], WSFL [2]) for enhancing data and analytical service Grid capabilities

SEEK will initially target the integration and synthesis of ecological and biodiversity data. This research frontier is critical to ecological and biodiversity forecasting and to enabling managers and policymakers to anticipate environmental change and thereby deal with it in a more sustainable fashion [3]. Two research issues of critical interest to scientists and policymakers will serve to test the utility and effectiveness of SEEK: (i) detecting and understanding patterns in living resources and biodiversity and (ii) understanding the interrelationships between biodiversity and ecosystem function and how both may be affected by global change.

Multidisciplinary teams of scientists organized in collaborations—the SEEK *Working Groups*—will closely inform the design and development of the IT research areas described above, as well as directly advance our understanding of the two test-bed questions. The three Working Groups are:

- *Biological Nomenclature (BN)*—investigates solutions to mediating among multiple, often competing or inconsistent, taxonomies for naming organisms, including integration of disparate data sources (museum collections with ecological data) via enriched semantics
- *Knowledge Representation for Biocomplexity (KR)*— develops and tests ontologies for expressing concepts in ecosystems and biodiversity science that will parameterize the semantic mediation system and provide a vocabulary for the analysis and modeling system
- *Modeling and Analysis of Biodiversity and Ecological Information (MA)*—provides domain experts’ knowledge in modeling and analysis to evaluate SEEK usability for addressing biodiversity and ecological questions

2.1 The Problem: Extreme Data Heterogeneity and Complexity

Heterogeneity of biodiversity and ecological data—in syntax, schema, and semantics—prevents researchers from discovering, integrating and synthesizing the wide variety of data that has been collected and that is useful for teasing out the principles of environmental structure, function and

sustainability. These data are inherently complex, reflecting the tapestry of biotic and abiotic factors and interactions that lead to a functioning ecosystem. Semantic variability arises in ecological data not just from the use of distinct methods or research biases, but from specific research motivations and legitimate needs for specialized types of information.

2.2 Example: An Ecologist's IT Challenges (Today)

Consider a typical ecological research scenario. A scientist is interested in analyzing the *spread of invasive species* in a certain region, is already aware of major results from the literature, and has some data pertaining to the topic. First, the ecologist might try to discover additional datasets online, or from the scientific literature. For example, she searches a few “science portals” on the Web, or queries a generic search engine such as Google. This process is time-consuming and ineffective because many relevant datasets are neither referenced nor available online.

Even if potentially relevant datasets are located, the scientist encounters new problems with data access and the resolution of syntactic and semantic ambiguities in the data. For example—can the data be downloaded, or are they so large and diverse that relevant components must be extracted by queries? Are the data format and structure sufficiently well described to enable parsing and processing? Are the finer aspects of the data schema understood—e.g., do ‘blanks’ indicate missing values or ‘absence’ (structural zeroes); and does lack of an attribute indicate non-measurement or lack of presence? Are repeating values true statistical replicates?

Finally, there are even subtler semantic issues to resolve when integrating data from heterogeneous sources. Measurements among different spatial and temporal scales can sometimes be resolved by simple unit conversions (e.g., British to metric), but—depending on statistical considerations—not so simply in the case where the original *dimensions* of the measurement differ. For example, consider several data sets containing abundance estimates of organisms from different spatial locations where sampling quadrats differed in area by several orders of magnitude. Integrating these data for analyzing the presence of abundant species might be appropriate, because the result would be relatively insensitive to sampling scale. In contrast, comparing the representation of rare species across space might be inappropriate because data taken at smaller sampling scales would systematically under-represent the rare species. Results of data integration would then simply reflect sampling artifacts in the case of rare species.

To further complicate our example, biodiversity data used for tracking the spread of an invasive species would likely contain taxonomic names. Datasets collected at different times might reference competing, if not conflicting, taxonomic standards. Taxon names often change over time via “*splitting and lumping*” of lineages that cannot be reproduced by most data models used for taxonomic data storage. Thus, taxonomic names are a significant and pervasive dilemma when dealing with biodiversity data, whether the research involves biotic compositions of ecosystems, conservation of threatened species, or trends in invasive species. These and other syntactic, structural, and semantic heterogeneities present major challenges to discovery, integration, analysis, and reasoning with scientific data [4].

Assuming our scientist has finally located some data sets of interest, the analytical stage is conducted in an environment like “R” or MATLAB® or SAS®. These *analytical engines* require importing tables of data, manipulating them, then programming and executing models to interpret them. However, comparison, sharing, and extension of analyses are hindered by the wide variety of software approaches in use.

In summary, the ecological research process is severely hampered by the difficulties involved with data discovery, access, integration and analytical application. Our proposed research focuses

directly on these issues by creating a powerful and efficient environment for conducting analyses within an extensible network of data resources and software tools.

3 SEEK Project Description

SEEK is a Web-accessible IT infrastructure for deriving and extending ecological knowledge, based on semantically mediated views of data and analyses. SEEK will be designed and implemented through IT research, coupled with domain-centric Working Group activities that will test and inform the IT efforts. IT research will focus on developing a unified knowledge environment based on three components: *EcoGrid* (EG), *Semantic Mediation* (SM), and integrated *Analysis and Modeling* (AM); see Figure 1. Working group activities complement the IT research by engaging domain scientists in testing and evaluating the design and utility of the technology framework, while also using SEEK to investigate major biodiversity and ecological research questions.

3.1 Results of Prior Research

The research team includes computer scientists, biologists, and information technologists from the Partnership for Biodiversity Informatics (PBI), a consortium comprising the National Center for Ecological Analysis and Synthesis (NCEAS); the San Diego Supercomputer Center (SDSC); the University of Kansas (KU), and the University of New Mexico (UNM) and partnering institutions (Arizona State University, University of North Carolina, University of Vermont, and Napier University). Here, we briefly describe relevant projects of the PBI partners and the other participants—each of whom brings to bear some specific technical or domain expertise.

At NCEAS (Jones, Reichman, Schildhauer), in conjunction with the Long Term Ecological Research (LTER) Network Office (Michener, Waide, Brunt) at UNM, research is progressing on the Knowledge Network for Biocomplexity [5], an integrated metadata and data network based upon an extensible domain-specific metadata content specification, Ecological Metadata Language (EML) [6, 7]. EML is based on a set of XML DTDs and Schemata. The metadata are stored as a DOM-tree in a RDBMS for scalability and performance. Metadata and data are accessed via a Java servlet called the Metacat [8, 9]. Another project is developing Morpho, a Java application for powerful interaction with the Metacat, a full-featured metadata editor and content-based data management system for scientists' desktop computers. The Monarch system links heterogeneous data sources into a variety of commercial analytical packages such as SAS and MATLAB.

At the Biodiversity Research Center at KU (Beach, Peterson, Vieglais), informatics projects are aimed at international support of biological collections data integration and analysis [10, 11]. The Species Analyst (TSA) project comprises a set of network-enabled desktop tools for querying globally distributed museum specimen databases [12]. TSA, originally based on the ISO Z39.50 information retrieval protocol standard, is being re-implemented with XML and HTTP. Sixty biological museums in several countries currently use TSA to provide access to their structured specimen data. FISHNET [13] and the "Integrated Network for Distributed Databases of Mammal Specimen Data" project are examples of taxonomic research communities employing TSA technologies.

A 1998 NSF-KDI award to KU is developing additional tools and services for biological museum data. Desktop GARP is a PC implementation of David Stockwell's (SDSC) Genetic Algorithm for Rule-set Prediction [14] for predicting species distributions based on known locations from specimen records correlated with environmental variables. Lifemapper [15] is another KDI project that has implemented GARP as a SETI@Home-like screen saver for parallel, distributed desktop computation of species prediction models. Lifemapper will serve the resulting species predictions in an open-access geospatial data archive and will soon serve images and data via SOAP and Windows Map Service.

In addition, at KU (Gauch), the OBIWAN project uses ontologies for categorizing, browsing, searching and visualizing Web information. A weighted ontology was generated for use as meta-

information by other agents [16, 17]. This reference ontology was then used to optimize multi-site browsing and provide monitoring of content changes within a web site [18]. It has become a basis for query brokering in a distributed search architecture [19, 20]. By applying the classification agent to individual users' browsing histories, weighted ontologies have been used to optimize personalized Web searching, browsing, and navigation [21, 22, 23, 24].

At SDSC (Rajasekar), research is progressing on the Storage Resource Broker or SRB [25, 26, 27, 28], a robust, database-driven distributed file system that works across platforms and can effectively provide file-level access to arbitrary, large, distributed, and heterogeneous digital objects. It operates as client-server middleware in conjunction with a metadata catalog [29] to provide content-based, rather than location or name-based, access to digital resources. It is a mature *Data-Grid* product that is actively used by several major scientific projects, including the Digital Sky Project [30] and the Protein Database [31], and is an integral part of several Digital Library Projects (e.g., CDL [32] and NSDL [33]). The SRB is also an integral part of NPACI's data access architecture [34] and houses remote-sensed data that have direct utility to ecology.

In addition, at SDSC (Ludaescher), work has been progressing on developing logic-based approaches to semantic modeling and mediation to facilitate scientific data integration [4]. For example, semantic mediation approaches based on knowledge representation techniques are enabling scientific data integration of complex neuroscience data sources [35, 36, 37, 38]. Ludaescher's work will be a linchpin to the proposed work, by extending his formal logic approach to enable advanced semantic mediation in an even more heterogeneous context than the work on neuroscience data. At UCSD (Ludaescher, Vianu), work is also progressing on XML-based mediation, optimized query evaluation, DTD inference, and semantic extensions of XML queries [39]. Also at UCSD (Goguen), the algebraic specification and transformation system OBJ has been developed and applied to data integration scenarios [40, 41, 42].

At the University of Vermont's Institute for Ecological Economics, Villa will continue developing an Integrated Modeling Architecture [43]. This activity encompasses research and software development to define a common framework for specifying and sharing analytical and modeling processes. Data, interpretive models, and processing tools (e.g., statistical software, GIS, and optimizers) are given a common semantic characterization that allows them to be assembled in arbitrary configurations and automatically checked for compatibility. An RDBMS with SOAP interface then allows libraries of "scientific modeling" objects to be stored on the network and assembled into complex and functional distributed models that can be run from a browser.

These projects all address important pieces of the major challenge—to enable vastly more efficient and powerful discovery, access, and use of scientific information. There are major benefits to confederating our research lines, because each contains critical ecological data and IT approaches. The advanced work proposed here will (a) profoundly improve the utility of all our systems and (b) advance our understanding of how these IT approaches can enhance science.

3.2 Research Activity: Analysis and Modeling System

The example in Section 2.2 illustrates that correctly integrating semantically heterogeneous data requires an understanding of how the integrated data will be used in a statistical analysis or simulation model. We recognize that two or more data sources might be *correctly* integrated for some analytical purposes but that it would be *incorrect* to integrate them for a different purpose. Thus, whether integration is appropriate depends on the specific context and semantic constraints of the analysis rather than being an inherent property of the data.

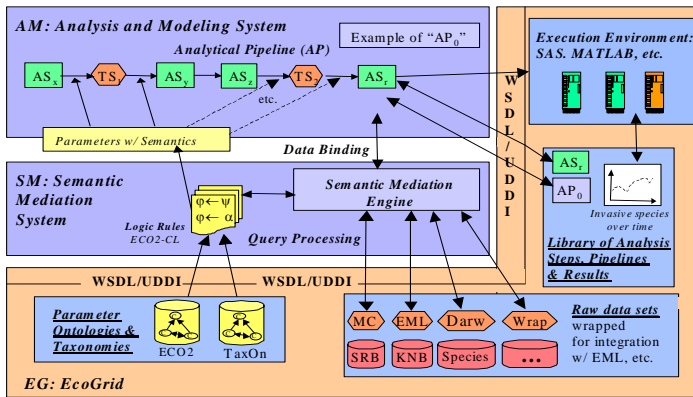


Figure 1 SEEK architecture showing interactions among the EcoGrid, Semantic Mediation System, and Analytical Pipelines. An Analytical Pipeline is a graph with nodes that represent computational components and arcs that represent input and output parameters that are explicitly tied to a parameter ontology.

SEEK will address these issues by taking an analytical perspective on data integration. By formally describing the data and processing semantics for an analysis, we can determine whether particular ecological data sources are appropriate for integration and in an analysis or model. In SEEK (Figure 1), scientists can design a scientific analysis or model as an *analytical pipeline* (AP), which can be seen as an “analysis workflow” involving discrete analysis steps as well as data transformation steps. Analytical steps (AS) implement *mathematical models, simulations, and analyses* that are useful in ecology; transformation steps (TS) implement *conversions, querying, and restructuring* of data. Each step in the pipeline is linked in the graph such that the semantic requirements of the inputs (*preconditions*) of one step are consistent with the semantic characteristics of the output (*post conditions*) from the previous step. Each analysis step in a pipeline may itself contain a nested analytical pipeline, i.e., such networks can be nested. Analytical pipelines can be modeled as labeled directed graphs in which analysis steps are represented as nodes, parameters as edges, and parameter semantics and pre- and post-conditions as edge labels and node labels, respectively.

Requirements of each analysis step and transformation step are defined by *semantically* typing the input parameters and output parameters of the step, using terms from a *parameter ontology* (Section 3.6). The parameter ontology creates a formal system for defining parameter semantics via (i) a controlled vocabulary of parameter names, (ii) constraints (e.g., equations) relating different parameters to one another, and (iii) a data type hierarchy, where base types (e.g., integer, float) come from an existing type system (e.g., XML Schema Datatypes [40, 44]). Complex derived types, however, will be specific to the ecological domain. For example, an analysis step that calculates the *Shannon Diversity Index H'* (Figure 2) has one output parameter that represents the diversity index (floating point number), and two input parameters representing a *Species Count* (integer i) and *Proportional Abundance* of a species (array of floating point numbers p). These concepts from the ecological sciences domain will be drawn from the parameter ontology that will be created by the Working Group on Knowledge Representation (Figure 3, Section 3.6).

During the **design phase**, a scientist can develop an analytical pipeline (or reassemble pieces from a library of existing analysis and transformation steps) by:

- (A) *creating* Analysis Steps, e.g., by: (i) defining input and output parameters for analysis steps; (ii) “*semantically typing*” these parameters by associating them with *concepts* (parameter names) from a set of formalized *parameter ontologies*; (iii) *defining pre- and post-conditions* for analysis and transformation steps, including constraining parameter values (e.g., limiting the spatial extent of a parameter); (iv) implementing the analysis step in a particular analytical

environment (e.g., SAS); and (v) *storing* the analysis and transformation steps as first-class objects in the EcoGrid

(B) *reusing* predefined Analysis Steps (from previous analyses) from the EcoGrid

(C) *assembling* the Pipeline by *linking* the outputs of one step with the inputs of a subsequent step, (a step can be an analysis step or a transformation step), provided the semantic mediator does not indicate an inconsistency in the parameter and analysis step semantics (Section 3.3).

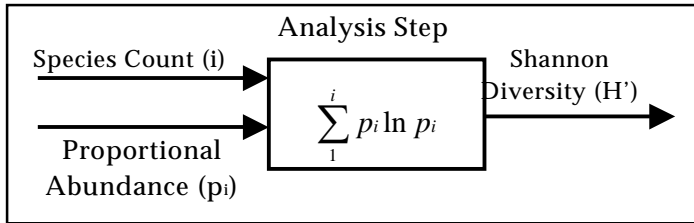


Figure 2: Analysis steps specify computational details in a language supported by the execution system. Arcs include parameter specifications and refer to a parameter ontology. Pipelines are created by linking input and output parameters.

By formalizing the Analysis Steps (AS) in a declarative language, they become first-class objects to be stored along with relevant data in the EcoGrid. In this way, analyses (virtual data) can be used similarly to “normal” data. At the requirements level, it seems important to make it easy for users to define their own semantic transformations because of the great heterogeneity and even unpredictability of such transformations, e.g., using a simple high level language based on abstract data types (ADTs). Some successful experiments have been done using OBJ [40,42].

During the **execution phase**, the pipeline is run in the *execution environment*, which comprises analytical engines (SAS, Matlab) and other environments, as well as data access, querying and transformation steps coordinated by the query processing component. The results, e.g., a time series plot of “*incidence and spread of invasive species*,” are returned to the scientist. These, and intermediate results, are stored in the EcoGrid (Section 3.4) as first-class data products. To this end, the semantic mediation and execution environments keep track of all information that is necessary to *reproduce* and *interpret* as well as *refine* and *rerun* the analytical pipeline. Among other things, this information includes: *references* to the selected input data sets (“raw” or “derived”; in the latter case, also data *provenance* and *processing* history [45]); *newly derived data products* between ASs (having a unique identifier for reuse); the *transformation logic* (a.k.a. view definition); and specifications for the Analysis Steps, Transformation Steps, and the Analytical Pipeline.

3.2.1 IT Research Challenges

Semantic Analysis and Parameter Models: When modeling parameter semantics via formal ontologies, their interrelationships via constraint formulas, and the pre- and post-conditions of Analysis Steps, one can choose from many formal languages. The challenge is in (i) choosing languages that balance the trade-off between expressive power and tractability, and (ii) devising automated deduction procedures and reasoning algorithms that can be implemented effectively and, ideally, efficiently. We will investigate deductive languages and techniques for analytical pipeline construction, e.g., [46, 47], as well as emerging business-oriented work on XML Pipeline Definition Language [48] and Web Services Flow Language [49]. In particular, we will examine whether and how workflow languages such as WSFL and XML Pipeline Definition Language can accommodate the description and composition of scientific calculations. We also plan to study OBJ extensions [40] and active logic rules [50,51] for the development of pipeline definition languages.

Several systems exist for visual, component-based analysis, e.g., Khoros and Ptolemy. The Khoros system allows researchers to work with various types of data and analyses and perform sophisticated analyses on the data [52]. The user is limited, however, to those analyses that exist or can be implemented in a language like C. The Ptolemy system provides a visual programming environment for modeling embedded systems in which the basic model components are implemented in Java [53, 54] and described in XML [55]. Neither environment provides the user with information about the semantics of the analytical components beyond simple data typing, nor can they reason about semantic constraints associated with analysis steps. We will investigate analytical workflow approaches that build upon systems such as Khoros and Ptolemy but that include rich capabilities for semantic processing of the analytical models.

Similarly, for checking the consistency of formal ontologies (in particular, parameter ontologies), corresponding languages and reasoning procedures have to be devised. Decidable fragments of first-order logic, called *description logics* [56], e.g., have been used to model ontologies in the neuroscience domain [36, 57] and allow reasoning about consistency and structure of ontologies.

3.2.2 Deliverables

- Formal languages for describing ecology-related *parameters*, *concepts*, and their *relationships*: ECO² (Ecology Ontology; Note: ECO² *content* will be developed by Working Group 2)
- formal languages to describe analysis steps, i.e., languages for expressing *constraints* over ECO² and the analytical pipeline: ECO²CL (ECO² *Constraint Language*); investigation of the use of description logics (DLs), and DL reasoners [58]
- a resolution mechanism for ECO²CL, i.e., algorithms that (i) check consistency of an analytical pipeline, and (ii) retrieve marked-up data sets that are consistent with the pipeline
- GUI for creating analysis and transformation steps, assembling pipelines, storing various pipeline components, and monitoring pipeline execution. This extends Morpho and utilizes visual model assembly tools from the Ptolomy "Virgil" interface [54] and formal approaches [59].
- an Extensible Execution system for executing analysis and transformation steps on various systems such as SAS, Matlab, R, ArcView, and custom simulation systems. This will extend our current Monarch work to a broader array of environments.

3.3 Research Activity: Semantic Mediation System

We have described a system that represents analytical procedures modularly using semantically typed parameters. The *semantic mediation system* forms a middleware component between the analytical pipeline and the data and metadata sources available in EcoGrid. It must perform query processing and reasoning tasks over formal ontologies, the constraints of the analytical pipeline, and the schemata and constraints of the given data sources. Hence, in addition to constraint checking at design time, the system must blend resolution steps and query rewriting with traditional query processing steps. The logic-based system must:

- 1) determine candidate analysis steps for inclusion in a pipeline by presenting those steps that satisfy the semantic requirements of the previous step,
- 2) derive transformation steps needed to convert between two analysis steps (unit conversion, coded-value conversion, schema conversion), and
- 3) construct an integrated data view that satisfies the semantic requirements of the input parameters for an analytical pipeline and provides access to a materialized view.

During the **design phase**, scientists build an analysis to address particular questions. We envision them starting from the desired product (e.g., a plot of invasive species spread versus

time) and working backwards to determine how to create an analysis that can be linked to available data. For any given analysis step, the semantic mediation engine tries to determine the suite of prior steps that have semantically compatible outputs. For example, the analysis step in Figure 2 requires proportional abundance as an input. The semantic mediation engine would query the EcoGrid for existing steps that produce proportional abundance as their output, or produce a semantically compatible output that could be transformed into proportional abundance. To this end, the mediation engine employs automated deduction techniques over the given constraints, i.e., expressed as parameter constraints, source constraints, and pipeline constraints in ECO²CL. Reasoning may also proceed in a forward manner, starting from available data sources and having the user select from the constraint-satisfying analysis steps in the analysis repository.

The EcoGrid (Section 3.4) will contain a large number of ecological data sets, each of which is described through its “semantic metadata” wrapper in terms of the parameters that it contains, including “semantic types” for parameters. To facilitate the abovementioned form of semantic mediation, semantic *source registration procedures* must be developed that allow the data providers to mark up their raw data according to ECO² concepts and parameter semantics. These semantic *data annotations* are drawn from the same ontologies used for modeling analysis steps.

During the **data discovery and binding phase**, i.e., after the pipeline has been defined, the semantic mediation system determines, from the set of “semantically wrapped” data sources, those sources that satisfy the constraints imposed by the parameter specifications within the pipeline. The data sets can directly match the required parameters or indirectly match them via a transformation step constructed by the mediation engine. In an interactive step, the user selects those datasets that she wants to actually bind to analysis steps as input to the pipeline. Because both data sources and analysis steps are described using parameter ontologies, the semantic mediation system can treat them similarly when trying to construct the analytical pipeline.

3.3.1 IT Research Challenges

Semantic Data Integration: The goal of data integration and mediation systems [60, 61, 62, 63] is to provide uniform access to many heterogeneous data sources including databases, web source, digital libraries, and flat files. The two basic approaches for defining mediator systems are *global-as-view (GAV)*, in which the integrated view is defined over the source schema, and *local-as-view (LAV)*, in which the sources are defined in terms of the integrated schema. Both approaches have their advantages and shortcomings, e.g., GAV requires redesigning the global view every time a new source is added or a source schema changes, whereas with LAV, the global view is automatically computed by “inverting” the local views. A serious problem with LAV is that for complex mediated views and sources with limited information, local views often cannot be inverted. Also, research is still in its infancy for answering queries in either the GAV or LAV setting as soon as one leaves limited RDBMS languages and conjunctive queries (e.g., select, join) and moves to the very active area of mediators with XML-based query languages [61, 64, 65, 66].

Although it is generally agreed that semistructured data models such as XML are very flexible and most promising for general mediator systems, efficient XML query processing and rewriting are still major research areas with numerous partial solutions and unsolved problems. A promising approach for semantic mediation is to employ deductive object-oriented languages in order to express the rich semantics necessary in SEEK scenarios. Indeed, *semantic mediation* has been identified as a promising approach for *scientific data integration* [4], and first results in the domain of bioinformatics and neurosciences are very encouraging [35, 36, 57]. SEEK data integration problems are “complex, multiple worlds problems” [LGM01] because of disjoint schemas, hidden expert knowledge, and assumptions that need to be made explicit in order to make data sources “joinable”. These investigations will be critical for the success of SEEK.

3.3.2 Deliverables

- *formal languages* that can express ECO²CL constraints and deductive integrated view definitions over complex object-oriented models
- *query processing and reasoning procedures* that combine view definitions, user queries, and the given constraints to produce effective and efficient query plans
- *source registration and mediation methodology*, i.e., mechanisms for registering the semantics of both data sources and “analytical sources” (i.e., analysis steps) with the mediator and APIs for querying these registered sources at runtime
- *a semantic mediation system* that implements the abovementioned query processing and reasoning procedures for both data and analytical sources. This will extend our previous work on logic engines used in the KIND prototype [36, 57].

3.4 Research Activity: EcoGrid

The *Grid* (a.k.a. *Compute Grid*) is a system that links multiple computational resources such as computers, sensors, data, and people [67]. A *Data Grid* [68, 69] denotes a network of storage resources, from archival systems to caches and to databases that are linked across a distributed network with an emphasis on high performance and throughput. *Service Grids* are extensions to data and compute grids that provide common public interfaces for discovering and invoking services on behalf of remote users and applications.

We propose to research and build *EcoGrid*, an infrastructure that combines features of a *Data Grid* for ecological data and a *Compute Grid* for analysis and modeling services. EcoGrid will form the underlying framework for data and service discovery, data sharing and access, and analytical service sharing and invocation. Specifically, the EcoGrid will provide

- Seamless access to data and metadata stored at distributed EcoGrid nodes, including scalability, multiplicity of platforms (desktop to super computers) and storage devices, authentication through single sign-on authentication and multi-level access control,
- Execution of analytical pipelines (Section 3.2) via web services (WSFL, and WSDL, XPipe),
- EcoGrid node registry for data and compute nodes based on UDDI,
- Rapid incorporation of new data sources as well as decades of legacy ecological data,
- Extensible, ecologically relevant metadata based on the Ecological Metadata Language,
- Replication of data to provide fault tolerance, disaster recovery and load balancing.

This will result in an integration of data-, compute-, and service-grids for ecology.

3.4.1 IT Research Challenges

Scientific Extensions to Web Services: WSDL is a language for describing web service interfaces. It details how one invokes a service and how the response is returned [1, 70]. The UDDI environment provides a scalable mechanism for registering the location of data and analytical services [71]. WSDL as it stands now is geared toward business applications and its extensions to scientific domains need to be examined. WSDL is based on XML, the expressive power of which has been studied extensively [72, 73, 74, 75]. However, the expressive power of WSDL and suitable extensions for scientific domains has not been studied.

For example, the WSDL/UDDI combination alone is not sufficient to describe how web services might interact with one another in the construction of a scientific analysis system. In ecological analyses and models there can be strong effects of platform-specific characteristics such as precision of floating-point calculations or microprocessor architecture. WSDL does not provide for describing these types of constraints as part of the service description. Also, the taxonomies used in current implementations of UDDI do not provide sufficient categorization of services and data

for full exploitation by a complex system such as SEEK. The problems that we face in making WSDL useful for scientific applications are

- An alternative, extensible taxonomy must be incorporated into the UDDI to allow for the richer types of service categorization required for scientific computation in SEEK,
- It may be necessary to provide an additional metadata layer based on the MetaCat and MCAT architectures for additional functionality beyond that intended by the UDDI registry.

The development or adoption of a language that provides expressive power beyond WSDL to describe scientifically relevant issues, e.g., precision, triggers, and rules are required in the construction of an analytical pipeline. Transactional properties of WSDL, e.g., associativity, commutativity, and transitivity will be studied, enabling us to enhance parallelizability and composability of WSDL programs. Because WSDL will be used for service composition and possible parallel execution on distributed systems, these questions will be important not only for the SEEK project but also for other WSDL-based architectures [76, 77, 78].

3.4.2 Deliverables

- EcoGrid—a distributed, open system for (i) accessing ecological data sources, (ii) subscribing new and legacy data resources, and (iii) managing and accessing analytical services
- Deployment of EcoGrid nodes across >200 Metacat, SRB, and Species Analyst sites

3.5 Working Group 1: Biological Classification and Nomenclature Semantics

Systematics focuses on the phylogenetic reconstruction, circumscription, classification and naming of species and higher taxa. The classification of a species often changes over time as new data and analyses allow more precise determination of a taxon's affinities to other organisms. Coupled with classification are codified protocols for scientific naming. The various taxonomic disciplines have collectively developed many formal naming systems for organisms that are typically based on references to archived "type" specimens. Most scientific nomenclature rules mandate that a species be renamed if the taxonomic definition of the organism is narrowed so that it no longer subsumes the same "type" specimen, or if the definition is enlarged and subsequently includes other earlier named "types" that have priority. As a consequence of the "type" naming method, the same scientific name can be applied over time to multiple, distinct taxonomic concepts, and conversely the re-classification or placement of a taxonomic entity can result in the same taxon having multiple names at different times [79, 80, 81, 82, 83]. The result is 75 years of many-to-many mappings between species concepts and scientific names.

The semantic ambiguity resulting from nomenclatural systems that trump name publication priority and type specimen relationships over inviolate, unique taxonomic concept identifiers, presents a frustrating and significant challenge to researchers investigating biodiversity phenomena when they attempt to utilize species data sets from multiple sources or those assembled through time periods of greater than a few years. For meaningful, precise, ad-hoc integration and analysis of any species data sets in ecology and systematics, we must map and mediate among the critical semantic distinctions hidden beneath biological nomenclature.

The use of a scientific name in a field survey, on specimen labels or in a publication constitutes an *assertion* of a taxonomic concept. Semantic mediation of data labeled to species will require an Internet service with explicit mappings among the assertions. The scale of the mapping enterprise argues for a distributed, self-assembling database architecture; for higher plants there are perhaps 300,000 species, over a million validly published scientific names, and many millions of assertions.

We intend to guide the working group's activities by initially leveraging previous modeling and implementations for assertions and taxon concepts, particularly in collaboration with the Database and Object Systems Group, Napier University, Edinburgh (subcontract PI, J. Kennedy),

creators of the Prometheus classification data management system [81]. In addition to the researchers from the four collaborating institutions, the WG will include domain scientists, computer scientists and informatics specialists from BIOSIS, Inc. (see letter), the Taxonomic Databases Working Group, the FGDC (R. Peet, subcontract PI), the Integrated Taxonomic Information System (see letter) and from other international initiatives. We will offer our collaboration to the Global Biodiversity Information Facility's Electronic Catalog of Names Working Group, which has a similar vision and objectives.

3.5.1 IT Research Challenges

- Development of a comprehensive conceptual model that can represent all relevant aspects of biological classification and nomenclature semantics, specifically models of *multiple interpretations* depending on explicit representations of context information, e.g., temporal, hierarchical and circumscription dimensions.
- Development of logic representations that allow *reasoning* about the consistency and consequences of multiple, possibly competing interpretations. For example, using formalizations in *modal* and *many-valued logic* [84, 85], an automated deduction system may be devised that allows one to systematically compute all consequences of different taxonomic interpretations and feed those into the semantic mediation system, which in turn would show the different data and analysis views arising from the different nomenclature interpretations.
- Deducing *concepts* rather than species name strings from distributed taxonomic data sources

3.5.2 Deliverables

- **Conceptual schema and data model** for concept-based nomenclature data leveraging previous research by collaborators and colleagues.
- **Data entry software** that allows scientists to add new or published assertions as needed and to map institutional and personal perspectives on the relationships among assertions.
- **Desktop visualization tools for data discovery and management of multiple classifications** will be based on previous work by the Napier University Prometheus Project and others.
- **Database implementation** for an operational, Web-accessible prototype database with representative data from several different taxonomic groups (e.g. higher plants, fishes) aimed ultimately at a global, distributed and federated system of taxonomic concept servers.
- **Internet service for automated name/concept resolution**, accessible via EcoGrid, for several groups using information from synonymies currently available in public databases.
- **Usability analysis of the functional requirements** by working group members would evaluate all applications and tools developed for nomenclature resolution.

Significance. An Internet taxonomic concept (assertion) resolution service employing a semantic mediation engine would exploit the SEEK architecture to enable precise species concept based data discovery and integration. This specific concept identity resolution problem is representative of a large class of problems (e.g. with classifications for biotic communities, soils, rocks, places) where there exist many-to-many relationships between concepts and names. The solutions we develop should have fundamental utility far beyond biological nomenclature and biodiversity.

3.6 Working Group 2: Knowledge Representation for Ecology

Although the EcoGrid will provide for discovery of and access to a wide variety of ecological data and analytical services, greater advances could be made if the data were available in a common semantic framework. The objective for this working group is to create a series of ecologically relevant ontologies that can be used by the semantic mediation system (section 3.3) to reason about data and appropriate analysis steps. These tasks involve domain scientists with

expertise in ecology, biodiversity, and environmental sciences to drive the conceptual development of the ontologies. We will also involve a core group of computer scientists who will choose the framework and languages for representing the ontologies. Because there are many possible representations of complex scientific domains (multiple worlds problem), part of this research will focus on how to mediate among conflicting or ambiguous ontologies.

A primary product of this working group will be a parameter ontology (ECO²) for biodiversity and productivity science that can be used by the analytical pipeline (Figure 3). Each parameter will be formally defined in terms of its semantic relationship to other terms, starting from a basic ontology of scientific units (e.g., XML Schema Datatypes [44] or the Ontology for Engineering Mathematics [86]). Parameters will be related to each other in terms of "part-of" relationships that show their derivation and in terms of properties (slots) that describe how they are calculated from their components. In addition, parameters will be grouped into classes that show relationships in terms of what the parameters measure (e.g., "species count" and "species density" might both be subtypes of "abundance").

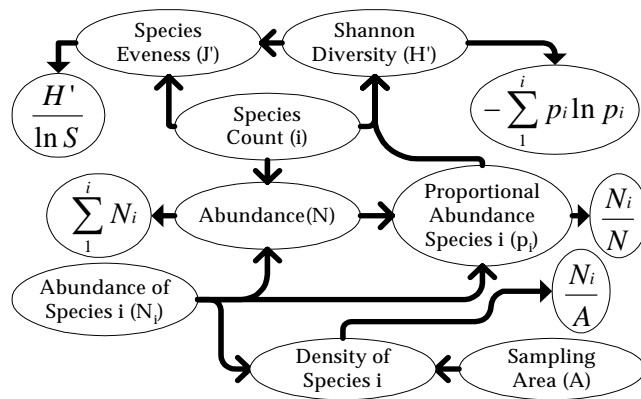


Figure 3: Simplified parameter ontology showing relationships among measured parameters in a fraction of the biodiversity domain. Arrows indicate part-of (composition) relationships, property relationships, and their equations. Actual ontologies will be developed using formal languages (e.g., DAML+OIL,KIF).

The working group will also develop a constraint language (ECO²CL) for expressing pre- and post-conditions for the parameters in analysis steps. This is critical for describing the semantic requirements of analysis steps. This working group will include domain expertise for specifying the types of constraints needed by the language as well as for developing functioning analysis steps that use the constraint language and the parameter ontology.

Finally, because developing ontologies for the ecological domain is a massive task, the working group will utilize automated feature extraction techniques to assist in this effort. The project will build on earlier work in corpus linguistics that identified semantically and syntactically related words from unstructured corpora [87]. Our research will extend these techniques for extracting specific features, rather than just related terms, based on statistical and rule-based patterns. We feel that this is an excellent application of feature extraction techniques because our domain is more limited in scope than the natural language or image processing domains to which these techniques are typically applied.

3.6.1 IT Research Challenges

Current research in ontologies often focuses on the mapping of content to correct locations in ontologies, based primarily on text categorization algorithms. We will explore automatic feature extraction techniques and categorize items based on a combination of textual and *other* knowledge features. This working group will concentrate on identifying the relevant features underlying relevant categories and on developing the feature extraction algorithms.

This group will also focus on the important problem of reconciling multiple, inconsistent ontologies. Ontologies are not static, but rather grow and evolve as new information is acquired.

This leads to a proliferation of ontology versions as well as the emergence of multiple ontologies. Thus, content classified under one ontology will need to be merged with content classified under another. Current ontology-mapping techniques apply statistical or knowledge-based techniques to effect these transformations and tend to ignore the structure of the ontology itself. This project will conduct novel research exploring the use of graph matching algorithms that exploit the link structure among similar concepts across ontologies, for facilitating the mapping process.

3.6.2 Deliverables

- ECO² and ECO²CL vocabularies (concepts and relationships) and grammars
- feature extraction techniques for generating parameter ontologies from scientific literature
- suite of ecologically relevant analysis steps specified in terms of ECO² and ECO²CL

3.7 Working Group 3: Biodiversity and Ecological Modeling and Analysis

This working group will serve as a test-bed to evaluate the effectiveness of SEEK. Challenges faced by the Working Group in addressing ecological case studies will reflect broad community needs for discovery of and access to relevant data sources (EcoGrid), for new virtual (mediated) views of input data sets (Semantic Mediation), and for automated resolution of input and output requirements for analysis algorithms (Analysis and Modeling System).

Three related scientific challenges provide the basis for the test cases. First, biodiversity patterns (e.g., species richness) have properties that can be described as regional self-organization [88, 89, 90, 91] and that appear to result from the scaling behavior of networks and the biophysical laws that govern the partitioning of materials in ecosystems [92]. These patterns have yet to be fully understood at broad spatial scales. Second, causal links appear to exist between biodiversity patterns and ecosystem function [93], which leads to predictable relationships between ecosystem properties such as productivity and features of biodiversity such as species richness [94]. Such relationships have only begun to be examined across ecosystems. Third, effects of global climate change on biodiversity have been modeled in two distinct manners, one focusing on ecosystem function [95, 96] and another focusing on species autecology [97, 98]; the relationship between these model types has yet to be investigated. Progress in addressing these three challenges has been slow because of the difficulties in acquiring relevant data and in scaling analyses to accommodate heterogeneous data streams and cover broad geographic ranges.

We will address these questions by assessing species distribution predictions with geospatial niche models to predict biodiversity phenomena in response to environmental change [97]. Building on previous studies of scale-dependence in the relationship between biodiversity and ecosystem function [99, 100, 101] as well as investigations on the relationship between biodiversity and primary productivity initiated as part of the KNB project, we will examine interrelationships between biodiversity and ecosystem function and how both may be affected by global change

This working group has a broad and complex challenge, namely that of assembling a framework of modeling and integration that interconnects the diverse data streams that are required for these types of analyses. These data streams represent fields as diverse as ecosystem function, species' geographic occurrences, and remotely sensed information. The strength of the overall project—that of integrating such diverse data streams for the first time—is demonstrable via the new classes of knowledge that emerge from their integration. Hence, this working group will focus on bridging the gaps among the diverse fields that produce and analyze these data streams.

3.7.1 Applications of IT Advances

Access to data sources via EcoGrid will provide species occurrence data derived from ecological, museum, and observational studies, including the NSF's network of Long-term Ecological

Research (LTER) sites, as well as modeled distribution data, precise data on ecosystem function, and remotely sensed information to permit extrapolation across scales.

Integration of data sources via the Semantic Mediation System will provide new opportunities for scaling analyses across time, space, and taxa, and transforming data for particular analyses.

The Analysis and Modeling System will allow formal descriptions of analyses and models so that investigators can discover, reuse, extend, and communicate about precise analytical approaches.

3.7.2 Deliverables

- Test cases and specification requirements for parameter ontologies and SEEK analytical pipelines, considering biodiversity patterns, interrelationships between biodiversity patterns and ecosystem function, and models for global climate change
- Usability evaluation of SEEK environment for specific ecological questions and user interface design issues will be addressed [102, 103]
- Evaluation of the extent to which new classes of knowledge are enabled by integration of highly heterogeneous data streams in SEEK.

4 Community Outreach

The Partnership for Biodiversity Informatics and our SEEK collaboration comprise a diverse group of investigators from multiple institutions (including three EPSCoR states—Kansas, New Mexico, and Vermont) and are firmly committed to community-building efforts and promoting a diverse workforce (e.g., UNM is a Hispanic-serving institution; both it and KU have long traditions in educating Native Americans).

We will employ a multi-faceted approach to insure that the research products, software, and information technology infrastructure resulting from SEEK optimally benefit science, education, and the public. Outreach includes community involvement, a WWW portal, informatics training, and an innovative annual IT transfer symposium. Broad community participation in SEEK will be ensured through the direct inclusion of IT and domain scientists from the international scientific community in Working Groups and on the Science & Technology Advisory Board (see Project Management). Working Groups will include participants from the U.S. and International Long-Term Ecological Research Networks, the California Institute for Telecommunications and Information Technology, the Integrated Taxonomic Information System, the Organization of Biological Field Stations, Scripps Institution of Oceanography, the National Biological Information Infrastructure, BIOSIS and other organizations. [See accompanying letters of participation.] In addition, one of our PIs, Dave Vieglais, serves on the Global Biodiversity Information Facility (GBIF) and the Executive Director of GBIF, Jim Edwards, has agreed to serve on our S&T Advisory Board (see accompanying letter). We will target Working Group composition to include at least 25% minorities and underrepresented groups.

A WWW portal, evolving from www.ecoinformatics.org, will house and/or point to Internet-accessible resources (software, archives, research products and technical information) that are easily discovered, and freely accessible to the scientific community. We are firmly committed and will adhere to Open Source Foundation guidelines. As part of our commitment to building the informatics capacity in the biodiversity and ecological sciences, 3.25 FTEs (two Post Doctoral Associates, one Education Coordinator and half of one of the PI's time (0.25 FTE, Michener)) will be devoted to preparing web-based informatics training materials that will be available through the SEEK portal. Also, Paul Tooby, Senior Science Writer at SDSC, will assist in producing overview descriptions of research and Working Group activities that will be web-accessible.

Informatics training will be coordinated and provided through twice-yearly tutorials at SDSC (40-100 trainees annually) and an intensive two-week course in informatics (funded by an NSF

RCN grant) for staff and students associated with biological field stations and marine laboratories that will be offered at UNM (20 trainees annually). SEEK will support instructors and provide training materials and content for these courses, as well as a new distributed graduate seminar series that will be offered at SEEK institutions and made available over the WWW.

A key element of our community outreach will be an innovative annual symposium and training program that focuses on information technology transfer to young investigators and students, particularly those from underrepresented groups. We will recruit 30-40 young faculty members and post-doctoral associates for a week-long symposium in which the participants will gain hands-on experience with the latest information technology, including products resulting from SEEK. Participants will be provided with web-based materials that they can use in developing courses at their home institutions. Our objective in this regard is to “train the teachers” thereby extending our outreach to the broadest possible community. We will recruit participants through ads in *Science* and newsletters. AAAS will help recruit minority and women scholars.

5 Significance

The foremost intellectual merit of the SEEK project is its professional level of IT engineering for the biodiversity and ecological science communities, with its strong emphasis on standards-based, open architectures. The SEEK consortium and its partner institutions are a multidisciplinary and multi-sector intellectual team uniting the biotic, environmental and information sciences in knowledge networking research and infrastructure development.

The SEEK project will deploy IT research to enable integration, analysis, and synthesis of earth and human systems data on unprecedented scales. Products of this collaboration directly address a grand challenge for the 21st century identified by the NSF, USGS, and NASA [104], to understand biological and environmental systems in all their complexity in order to use them in a sustainable fashion [105]. This understanding is critical to science and society—for managing natural resources, for sustaining human health, for maintaining economic stability, and for improving the quality of human life. The need for this knowledge is urgent as the daily conversion of natural systems to human-managed systems accelerates the decline of biological and ecological diversity.

In the information economy, access to information for knowledge creation and decision-making is as valuable as the information itself. This project will enable bringing the intellectual content of biodiversity and ecological information into currency for science and society and the use of that information across research, education, commerce and government. In recognition of the overall value of this project for science and society, the partnering institutions are committing \$ 2.5 M in cash cost-share, in addition to extensive in-kind contributions.

Examples of significant project outcomes include: (i) revolutionizing discovery, access to and integration of ecological, earth, and human dimension data and information through the SEEK infrastructure; (ii) developing intelligent analytical tools and infrastructure to support the needs of scientists, decision-makers, and the general public; (iii) education and training of the next generation of ecologists in information technology skills; and (iv) improving the opportunities for scientists, resource managers, policy makers, and the public to make scientifically-informed decisions about the environment by expanding access to ecological data, information, and knowledge.

6 Project Organization and Management Plan

SEEK will be managed through the Partnership for Biodiversity Informatics (PBI)—a consortium comprised of the Biodiversity Research Center at the University of Kansas, the Long-Term Ecological Research (LTER) Network Office at the University of New Mexico, the National Center for Ecological Analysis and Synthesis (NCEAS) at the University of California Santa Barbara, and the San Diego Supercomputer Center (SDSC). The PBI mission is to promote the integration and synthesis of ecological and biodiversity information through information technology in order to enhance the national and global capacity for observing, studying and understanding environmental complexity. PBI has successfully collaborated on IT research, training, and community outreach efforts, and has the experience and infrastructure needed to manage large projects such as SEEK.

6.1 Overall management and organization

PBI and all participants will prepare a detailed project execution plan that will specify project milestones, budget estimates, and project and task organization after the project is initiated. Weekly Executive Council videoconferences will be scheduled to review progress, assess project milestones, coordinate new activities, and resolve problems. Meetings of the Executive Council and all Working Group and Research Activity leaders will take place quarterly to review progress, assess project milestones, coordinate new activities, and resolve problems. The Project Manager will maintain daily contact and oversight of developers, as well as coordination with Executive Council members to facilitate optimal productivity of project participants.

Figure 4 illustrates the project management structure that builds upon approaches that have worked effectively in our prior large NSF collaborative projects. An Executive Council (Beach, Ludaescher, Michener, Reichman) comprised of one representative from each of the primary PBI institutions plus the Project Manager (Jones, who will serve as ex officio member) will coordinate overall project operations, as well as manage day-to-day operations at their individual sites. The Executive Council will define project tasks and deliverables, and establish project milestones and schedules. Members of the Executive Council are highly qualified for their roles and have had considerable experience working with the other team members on this project and on related activities.

Dr. Michener is the Associate Director of the LTER Network Office and Principal Investigator for the Resource Discovery Initiative for Field Stations (an NSF-funded Research Coordination Network involving numerous US and international institutions). Formerly, he was Director of the Ecology and Biocomplexity Programs at NSF from 1999 through 2000, where he managed a combined annual budget of approximately \$60M.

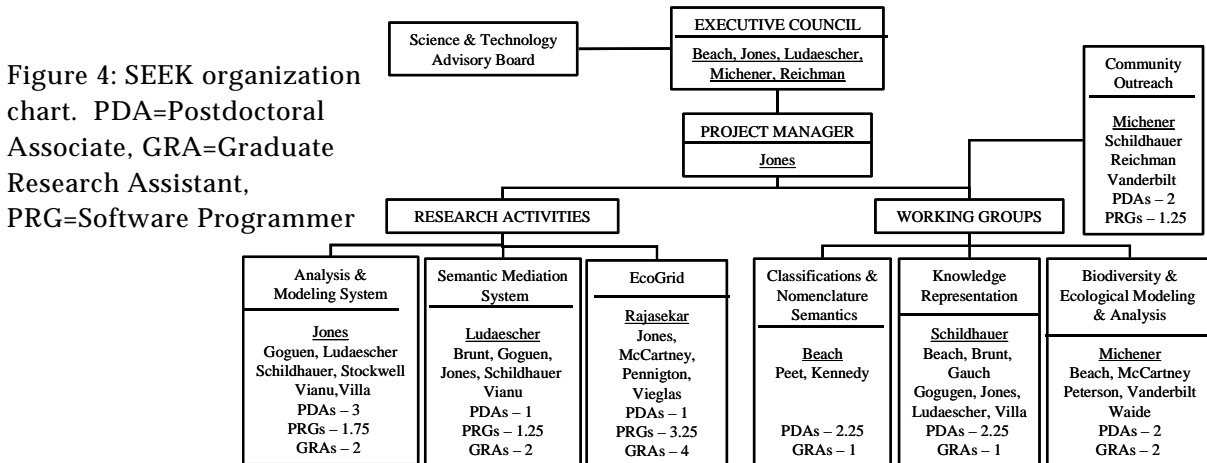
Dr. Beach is Assistant Director for Informatics at its Biodiversity Research Center and Principal Investigator of a \$2M NSF KDI Award for Knowledge Networking of Biodiversity Information. He is also co-PI of the Specify Software Project, an NSF-funded biodiversity software infrastructure project for museum biological collections. He currently serves as chair of the Board of Trustees of BIOSIS, a \$20M, not-for profit, biological database publisher and worked for two years at NSF as Director of the Biological Databases and Informatics Program

Dr. Ludaescher is leading the Knowledge-Based Integration Lab at the SDSC and is co-investigator of the BIRN Coordinating Center at UCSD, which coordinates the data integration activities of the \$20M collaborative NIH Biomedical Informatics Research Network [BIRN]. He is also co-PI in the \$12M collaborative DOE Scientific Data Management Center project [SDM] of the SciDAC (Scientific Discovery through Advanced Computing) program, where he leads the knowledge-based data federation activity of SDSC.

Dr. Reichman is Director of the NSF-funded NCEAS. The Center has hosted more than 2,500 different scientists who participate as sabbatical visitors, postdoctoral associates, and members of working groups. Previously, Dr. Reichman was the Director of the Konza Prairie Biological Station and Assistant Director for Research of the National Biological Survey.

The Project Manager (Jones) is Co-PI on the NSF-funded Knowledge Network for Biocomplexity project and has extensive experience managing a distributed, multi-institutional team of software developers. Jones will oversee and coordinate research and development efforts among all institutions. All participants will use “chat” and videoconference tools and will be expected to communicate regularly with their colleagues and the Project Manager.

Programming and research tasks will be distributed among the partnering institutions where facilitation by co-PIs can be most effective. Although each co-PI has responsibility for particular research tasks, daily oversight and coordination remain under the Project Manager. An Education Coordinator will focus on outreach activities, and two part-time administrative assistants will handle accounting, communications, meeting planning, and administrative functions.



Each Working Group leader will organize their Working Groups to best advantage, although the Executive Council will play a large role in facilitating ethnic, gender, and geographic participation. Working Group Leaders must participate in all WG meetings or arrange for another member of the Executive Council to cover this responsibility. Typically, at least two members of the Executive Council will be involved in any given Working Group meeting, but the management line runs from one designated Co-PI to each Working Group to avoid the problem of divided (and therefore unclear) responsibilities.

6.2 Schedule

All years—Project Management: Revise project plan, recruitment, create project infrastructure, Annual meeting for all project participants; **ALL:** usability testing; **EOT:** Annual Informatics Training Symposium, SEEK portal development, develop Informatics Training materials.

Year 1—EG: Prototype and web client at KU, UCSB, UNM, SDSC; **SM:** Develop languages for ontology and constraints; **AM:** Develop analysis step and pipeline languages; **KR:** Survey biodiversity parameters; **BN:** Concept-based taxonomy model; **MA:** Use cases and requirements.

Year 2—EG: Develop Web Services, Desktop client, Add LTER nodes; **SM:** Query processing & source registration research; **AM:** Prototype execution system; **KR:** Manually generate biodiversity ontologies; **BN:** Database implementation; **MA:** Create models (analysis steps and pipelines).

Year 3—EG: Web Services for computation, add OBFS nodes; **SM:** Mediation system for analysis step selection & data binding; **AM:** Execution system web service, software for analysis step & pipeline creation; **KR:** Ontology feature extraction; **BN:** Concept-based taxa entry software.

Year 4—EG: Service Description Language research, complete node deployment; **SM:** Query efficiency tuning; **AM:** pipeline validation using SM; **KR:** Build broad ontologies using feature extraction; **BN:** Visualization of multiple classifications; **MA:** Overall system evaluation.

Year 5—EG: System Integration, Performance, International nodes added; **AM:** Revised client; **KR:** Use & evaluate feature extraction for concept-based taxonomies; **BN:** Web service for name/concept resolution; **MA:** Publish results of research.

6.3 Milestones, performance metrics, reports, and reviews

Project management software will be used to specify project research and education milestones, as well as display, track, and coordinate key scheduling and resource information. The PBI Executive Council will review financial reports for the participating institutions on a quarterly basis; expenditures will be examined in relation to meeting project milestones and any reserves will be reallocated as deemed necessary to promote timely completion of project tasks.

Project performance metrics for project objectives (including project reviews, software development milestones, and publications), will be developed by the Executive Council and Working Group leaders in consultation with the Science & Technology Advisory Board. Costs and timelines will be maintained for each research activity and working group. Quarterly reviews and reports will be prepared by the Executive Council highlighting accomplishments, progress, problems, and revisions in budgeting and project scheduling. Reviews will be shared with all senior personnel as well as NSF and the Science & Technology Advisory Board.

Annual reviews will be performed by the Science & Technology Advisory Board, i.e., six to eight individuals chosen in consultation with NSF who will review the science, technology, cost, and schedule. The Board will consist of individuals with ties to the ecology, biodiversity, and information technology communities to insure the broadest possible representation. Initially, the following individuals/organizations have agreed to serve on the Advisory Board: Jim Edwards, Director – Global Biodiversity Information Facility (Copenhagen); Linda Sacks, Vice President for Marketing and Development – BIOSIS Corporation; Peter Arzberger, representative from the California Institute for Telecommunications and Information Technology; and a representative from the National Biological Information Infrastructure (Gladys Cotter or her designated appointee). Three to four remaining S & T Advisory Board members will be chosen in consultation with NSF. Advisory Board reports will be disseminated among all senior personnel and NSF.

6.4 International participation

Dr. Jessie Kennedy (Napier University, Scotland) will participate in the definition of the taxonomic model, prototype software tools to aid the migration of existing data sources to the new taxonomic database model, and investigate a distributed database architecture and visualization tools to support the semantic mediation of plant taxonomic classifications. Drs. Baillargeon and Munro, from the Canadian R&D team for the Integrated Taxonomic Information System (ITIS) for the United States, Canada, and Mexico, will develop appropriate interfaces between the international, multilingual version of ITIS and the software tools for data entry, multiple classification visualization, and Internet name services for automated name/concept mapping that will emerge from SEEK. The Executive Director of the Global Biodiversity Information Facility, Jim Edwards, will serve on our S&T Advisory Board; Dr. Vieglais, PI from University of Kansas, serves on GBIF, and we anticipate participation of GBIF staff members in working groups.