J Intell Inf Syst
DOI 10.1007/s10844-006-0034-8

A knowledge environment for the biodiversity and ecological sciences

William K. Michener · James H. Beach · Matthew B. Jones · Bertram Ludäscher · Deana D. Pennington · Ricardo S. Pereira · Arcot Rajasekar · Mark Schildhauer

This work is based upon work supported by the National Science Foundation under Grant Nos. ITR 0225676, 0225674 and DBI-0129792, as well as DARPA (N00014-03-1-0900). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF) or DARPA.

W. K. Michener (✉) · D. D. Pennington
Biology Department, LTER Network Office, University of New Mexico, MSC03 2020, Albuquerque, NM 87131, USA
e-mail: wmichener@LTERnet.edu

D. D. Pennington
e-mail: dpennington@LTERnet.edu

J. H. Beach
Biodiversity Research Center, University of Kansas, 1345 Jayhawk Boulevard, Lawrence, KS 66045, USA
e-mail: beach@ku.edu

M. B. Jones · M. Schildhauer
National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300, Santa Barbara, CA 93101-3351, USA

M. B. Jones
e-mail: jones@nceas.ucsb.edu

M. Schildhauer
e-mail: schild@nceas.ucsb.edu

B. Ludäscher · A. Rajasekar
San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, San Diego, CA 92093-0505, USA

B. Ludäscher
e-mail: ludaesch@sdsc.edu

A. Rajasekar
e-mail: sekar@sdsc.edu

R. S. Pereira
Natural History Museum and Biodiversity Research Center, University of Kansas, 1345 Jayhawk Blvd., Lawrence, KS 66045-7561, USA
e-mail: ricardo@cria.org.br

Abstract The Science Environment for Ecological Knowledge (SEEK) is a knowledge environment that is being developed to address many of the current challenges associated with data accessibility and integration in the biodiversity and ecological sciences. The SEEK information technology infrastructure encompasses three integrated systems: (1) EcoGrid—an open architecture for data access; (2) a Semantic Mediation System based on domain-specific ontologies; and (3) an Analysis and Modeling System that supports semantically integrated analytical workflows. Multidisciplinary scientists and programmers from multiple institutions comprise the core development team. SEEK design and development are informed by three multidisciplinary teams of scientists organized in Working Groups. The Biodiversity and Ecological Analysis and Modeling Working Group informs development through evaluation of SEEK efficacy in addressing biodiversity and ecological questions. The Knowledge Representation Working Group provides knowledge representation requirements from the domain sciences and develops the corresponding knowledge representations (ontologies) to support the assembly of analytical workflows in the Analysis and Modeling System, and the intelligent data and service discovery in the EcoGrid. A Biological Classification and Nomenclature Working Group investigates solutions to mediating among multiple taxonomies for naming organisms. A multifaceted education, outreach and training program ensures that the SEEK research products, software, and information technology infrastructure optimally benefit the target communities.

Keywords Analytical workflow · Data grid · Ecoinformatics · Ecological niche model · Knowledge environment · Semantic mediation · Metadata

1 Introduction

Knowledge of the natural world is limited not just by the complexity of natural entities and processes, but also by the complexity of the data that describe them. Enhanced understanding of the natural world depends on our capacity to access and integrate data from the biological, physical, and social sciences; mine those data for new knowledge; and convey new insights to decision-makers and the general public. The integration of scientific data has proven to be particularly challenging (Boucelma et al., 2002), e.g., due to the inherent complexity and variety of scientific data, and the "hidden" implicit semantics, which often are known only to domain experts or to the scientists who created the data. Moreover, schema information and metadata, which are meant to describe the structure and content of data sets, often do not provide the comprehensive information that is required for meaningful data integration (Michener, Brunt, Helly, Kirchner, & Stafford, 1997). Typically, knowledge about naming conventions, error bounds, sampling methods, and other contextual information are either not given at all or, at best, are buried in accompanying hard copy documentation.

Over the years, information technology researchers have studied and addressed a variety of data integration problems, ranging from interoperability challenges due to heterogeneous *systems* (e.g., different hardware platforms, operating systems, and communication protocols), *syntax* (e.g., different file formats), *structure* (e.g., schema and data model mismatches), and *semantics* (e.g., different use of terminology resulting in naming conflicts) (Sheth, 1998). Since the 1980s, the database community has developed various approaches and interoperability solutions for multidatabases and federated database systems. Despite such advances, data integration issues continue to challenge biodiversity and ecological scientists.

The Science Environment for Ecological Knowledge (SEEK) represents a 5-year research investigation that addresses many of the challenges associated with facilitating knowledge discovery across disciplinary, geographic, and methodological boundaries. The project encompasses information technology (IT) research in open architectures for data access, semantic mediation using domain-specific ontologies, and semantically integrated analytical workflows. The SEEK effort focuses on developing the information technology infrastructure that is necessary for addressing data accessibility and integration challenges in the biodiversity and ecological sciences. In the remaining discussion, we present the overall SEEK design and implementation approaches.

2 Knowledge environment architecture

The overarching goal of SEEK is to 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. The resulting Internet-based infrastructure will make it possible to more easily derive ecological knowledge by flexibly composing and applying a spectrum of analysis techniques to heterogeneous data from geographically distributed sources.

The SEEK architecture includes three well integrated components: an Analysis and Modeling System that provides users with the ability to capture scientific workflows as structured objects in digital libraries for reuse and extension; a Semantic Mediation System that enables discovery and automated integration of highly relevant but heterogeneous data via formal ontologies; and the EcoGrid which encompasses the Grid infrastructure that manages the data, metadata and computational resources (Fig. 1). Together, these subsystems provide a holistic solution that enables new types of data integration, analysis, and synthesis. The Analysis and Modeling System allows scientists to compose analyses and models that utilize data products and analytical components that are stored in the EcoGrid. It utilizes the Semantic Mediation System to resolve ambiguities in meaning when chaining together analytical components and integrating diverse data for use in the analysis. The Analysis and Modeling System is designed to enable the composition of an analysis plan, which can then be dispatched to the EcoGrid for execution on one or more of the distributed computational nodes. The results of the computation will then be returned to the scientist, and optionally documented and archived as first-class products in the EcoGrid for use in additional future analyses.

2.1 Analysis and modeling system

In current practice, ecologists frequently conceive of an analysis as a series of discrete stages that produce one or more products that they use to make judgments about a system or issue. These stages are often built in a variety of different software systems. For example, a typical scientist might use Microsoft Excel to enter data from an experimental study, export those data to a text file for statistical analysis in SAS; export the results of those analyses as text to import into Matlab to run a model, and produce some model output from Matlab. Even when particular stages of the analysis are captured in a system such as Matlab, the overall analytical process involving all of the stages is not documented formally. At best, the scientist records some notes about which versions of a model were executed using a particular version of a data set. More typically, the overall process is lost, and it is exceedingly difficult to reconstruct exactly what happened.

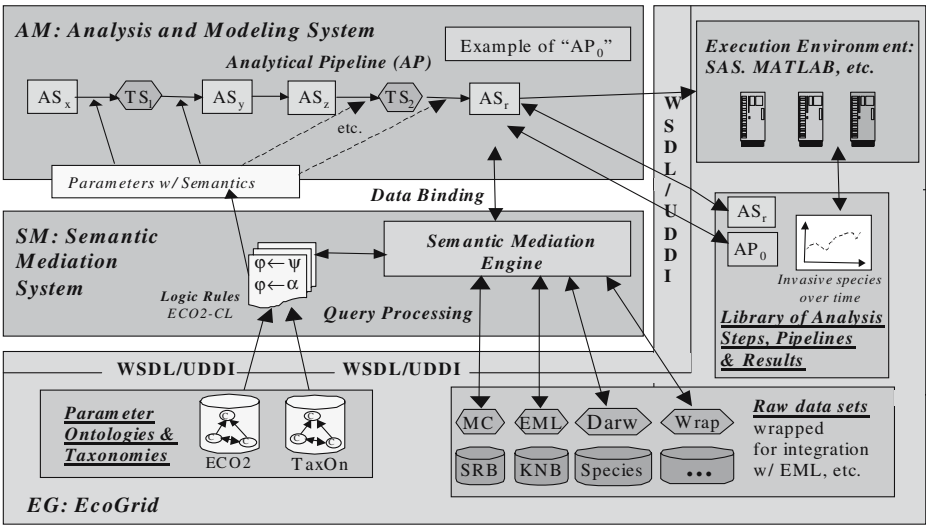

Fig. 1 SEEK architecture showing interactions among the EcoGrid, Semantic Mediation System, and Analysis and Modeling System

We have structured SEEK such that the development of analyses and linking of models is the major interface through which scientists interact with the system. The Analysis and Modeling System solves the problems of integrating scientific analyses by conceptualizing the entire analysis process as a series of data flows through discrete analytical steps. Thus, it can be referred to as an *analytical pipeline* or *scientific workflow*. Each step in such a workflow can be implemented in a different software system, and the requirements and constraints of each step are formally recorded. By recording this information about the analytical process, we enable reuse of both the individual analytical steps and the overall workflow by the same or other scientists. The information also provides an established and compact mechanism for scientists to communicate their exact procedures, which facilitates collaboration, replicability, and scientific peer review.

In SEEK, scientists can design a scientific analysis or model as an *analytical pipeline* (AP), which can be seen as a scientific workflow involving discrete analysis steps as well as data transformation steps (Fig. 1). 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 workflow 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 (cf. Bowers & Ludäscher, 2004). Each analysis step in a workflow may itself contain an analytical workflow, i.e., such networks can be nested. Analytical workflows 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.

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

(A) *Creating* Analysis Steps, e.g., by: (1) defining input and output parameters for analysis steps; (2) "*semantically typing*" these parameters by associating them with

concepts (parameter names) from a set of formalized *Parameter Ontologies*; (3) *defining pre- and post-conditions* for analysis and transformation steps, including constraining parameter values (e.g., limiting the spatial extent of a parameter); (4) implementing the analysis step in a particular analytical environment (e.g., R); and (5) *storing* the analysis and transformation steps as first-class objects in the EcoGrid

(B) *Reusing* predefined Analysis Steps (from prior analyses) from the EcoGrid

(C) *Assembling* the workflow by *linking* the outputs of one step with the inputs of a subsequent step, (a step can be an AS or a TS), provided the semantic mediator does not indicate an inconsistency in the parameter and AS semantics.

At the requirements level, it seems important to make it easy for users to define their own semantic transformations, due to the great heterogeneity and even unpredictability of such transformations, e.g., using a simple high level language based on abstract data types (ADTs). 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. See Bowers, Lin, & Ludäscher (2004) for details on this process of *semantic registration*, i.e., where data and (analytical) services are annotated with expressions from a community ontology to facilitate discovery and intelligent linking of workflow component. In this way, analyses can be treated in a manner similar to data.

During the *execution phase*, the pipeline is run in the *execution environment*, which comprises analytical engines (R, 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 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 workflow. 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); *newly derived data products* between Analytical Steps (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 workflow.

To be effective, the Analysis and Modeling System provides an intuitive user interface for composing analyses and model workflows. A variety of these systems have been prototyped by the SEEK team (Berkley, 2003; Ludäscher, Altintas, & Gupta, 2003) and by researchers in other disciplines (Khoral, 2002; Lee, 2001; London e-Science Centre, 2003; MDL Information Systems, Inc., 2003). The Analysis and Modeling System interface will produce as its output an analysis plan that can be executed, and will issue a series of instructions to the EcoGrid to execute these analytical steps on the available computational nodes of the EcoGrid. For some workflows, this may involve simply executing some analyses on the scientist's local computer; for others, it may involve dispatching analytical code to a high performance computer to achieve acceptable processing power, or dispatching it to a particular computer that already contains a large data set (e.g., a hyperspectral image) in order to avoid transferring the data over a network link with limited bandwidth. Thus, in addition to composing analyses, the Analysis and Modeling System acts as an optimizer to produce the outputs as efficiently as possible. As an implementation platform, we have adopted the Ptolemy II system from UC Berkeley (Lee, 2001) and extended it as part of the Kepler collaboration (Altintas et al., 2004).

The analytical perspective taken by SEEK will address issues regarding the correct integration of data. Semantic integration of 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 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.

By formally describing the data and processing semantics for an analysis, we can determine whether particular ecological data sources are appropriate for integration and use in an analysis or model. 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 well-defined *parameter ontology*. The parameter ontology creates a formal system for defining parameter semantics via (a) a controlled vocabulary of parameter names, (b) constraints (e.g., equations) relating different parameters to one another, and (c) a data type hierarchy, where base types (e.g., integer, float) come from an existing type system (e.g., XML Schema Datatypes; Biron & Malhotra, 2001). Complex derived types, however, will be specific to the ecological domain and will be drawn from a parameter ontology that will be created by a working group on Knowledge Representation.

2.2 Semantic mediation system

The Analysis and Modeling System in the SEEK architecture relies on the Semantic Mediation System. The purpose of the mediation layer is to facilitate data integration from heterogeneous data sources. In particular, it provides the Analysis and Modeling System a more abstract, conceptual (or "semantic") view of (a) the sources' data, and (b) the analytical functions and computations that work on those data (Bowers et al., 2004). The semantic mediation system performs several functions.

During the *design phase*, i.e., when analytical workflows are assembled by a domain scientist, the Semantic Mediation System provides an abstract view on the analytical steps (exported by the SEEK task repository). The WSDL description of tasks already provides XML Schema types for input and output parameters. In addition, the inputs and outputs of tasks have associated *semantic type*s (Bowers & Ludäscher, 2004). Because of these, the Semantic Mediation System often "knows" whether and how outputs of one step can be chained together with the inputs of a subsequent step. In general, the semantic type of some data is given as a set of logic formulae, acting as constraints on the possible data instances. In the simplest case this is done by associating the data with one or more *concept names* from a registered ontology, together with a role that qualifies in which way the data are an instance of the concept(s).

During the *execution phase*, i.e., when a previously designed analytical workflow is "loaded" with suitable data, the Semantic Mediation System can determine whether some registered data are compatible with the various inputs to the analytical workflow. In case the semantic type checking suggests compatible data, but the XML data types are not directly compatible, conversion routines are suggested (if available) that act as wrappers around the input data. Compilation of an analytical workflow, given a user's parameter settings and data bindings, results in a prepared or planned workflow that can be executed by the runtime system.

The SEEK semantic mediation approach builds upon a number of ideas and techniques developed in the areas of databases, knowledge representation, and logic and functional programming. Most current information integration systems build upon a wrapper-mediator framework (Wiederhold, 1992) in one way or another. In particular, the database community has developed query rewriting and evaluation techniques for *database*

mediators (Baru et al., 1999; Garcia-Molina et al., 1997; Halevy, 2001; Levy, 1999; Ludäscher, Papakonstantinou, & Velikhov, 2000). These database mediation approaches view source data as relational tables or XML documents, and several extensions have been proposed to account for the complex semantics found in scientific data sources. In particular, the importance of ontologies as a means to facilitate data integration by sharing formal representations of "glue knowledge" has been highlighted (Brodaric & Gahegan, 2002; Fonseca, Martin, & Rodreguez, 2002; Pundt & Bishir, 2002; Uschold & Gruninger, 1996), and the use of metadata (Brilhante & Robertson, 2001) and conceptual models (Gupta, Ludäscher, & Martone, 2002; Ludäscher, Gupta, & Martone, 2001, 2003; Paton et al., 2000; Peim, Franconi, Paton, & Goble, 2002) for describing and integrating scientific data are becoming more commonplace. Description logics and other logic-based approaches are also becoming more widely known as part of the Semantic Web activity, and the standards and tools produced by the latter (Manola & Miller, 2003; van Harmelen et al., 2003) can provide the basis for the formalisms used in the implementation of the SEEK semantic mediation system.

2.3 EcoGrid

The foundation of the SEEK infrastructure must be a transparent and powerful system for accessing ecologically relevant data and for executing computationally demanding analyses and simulations. These data include the heterogeneous data collected at field stations (e.g., species monitoring, hydrology, meteorology, etc.) as well as remote sensing data, data from museum collections, and much more. Models and analyses that will need to be supported include well-known biodiversity and ecosystem models such as GARP (Genetic Algorithm for Ruleset Production; Stockwell & Noble, 1992; University of Kansas Center for Research, 2002) and CENTURY (Natural Resource and Ecology Lab, 2003) as well as custom models and analyses written for a single experiment or study. The SEEK EcoGrid is being designed to provide the infrastructure for managing these data and computational resources.

The EcoGrid is intended to be a thin interface layer that allows various existing data and compute services to interoperate. For example, the Metacat system developed by the Knowledge Network for Biocomplexity (2003) is a networked data and metadata management platform with features that are similar to the Storage Resource Broker developed at the San Diego Supercomputer Center (2002). The EcoGrid will make both of these systems and others accessible through a standardized high-level programmatic API. Anybody who has developed a data management system for ecologically relevant data can become a full participant in the EcoGrid network by implementing the EcoGrid API. Thus, field stations and other sites with highly customized data management infrastructures will be able to exchange data through their common EcoGrid interfaces.

The EcoGrid will be an infrastructure that combines features of a Data Grid for ecological data management 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 features such as scalability, multiplicity of platforms (desktop to supercomputers) and storage devices, single sign-on authentication, and multi-level access control.
*Execution of analyses and models in a computational network.

*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.
*An EcoGrid Portal that provides a central access point for all EcoGrid data access services.

EcoGrid will tie together a variety of currently independent software systems and networks (Fig. 2). Although we will not initially include all of the services displayed in Fig. 2 due to resource and timing constraints, EcoGrid will be an open system, allowing others to develop systems according to the EcoGrid interfaces and participate in the grid.

Initial deployment of the EcoGrid is through the implementation of the EcoGrid interfaces for Metacat (Jones, Berkley, Bojilova, & Schildhauer, 2001), SRB (San Diego Supercomputer Center, 2002), DiGIR (Open Source Development Network, 2003), and Xanthoria (Center for Environmental Studies, 2002) data management systems. This provides a basic EcoGrid network that includes data from the 24 LTER sites (Long Term Ecological Research Network, 2003), over 200 field stations affiliated with the Organization of Biological Field Stations (2003), the 36 University of California Natural Reserve System (2003), the PISCO network (Partnership for Interdisciplinary Studies of Coastal Oceans, 2002), over 80 collections from museums in the Species Analyst network, and a variety of data in SRB servers at the San Diego Supercomputer Center. A wide variety of independent researchers that are not affiliated with a specific site will also be able to expose their data through the EcoGrid.

3 Implementation

3.1 Distributed development team

An extensive multidisciplinary team led by the Partnership for Biodiversity Informatics—a consortium including the University of Kansas, University of New Mexico, National Center for Ecological Analysis and Synthesis, University of California San Diego, in conjunction with other partnering institutions (e.g., Arizona State University, Napier University, University of North Carolina, University of Vermont)—is developing the core infrastructure for SEEK. Scientists, programmers, and students comprise the core development team and are distributed among the principal institutions. A project manager and a variety of communication technologies (e.g., "chat" tools, discussion groups, teleconferencing) facilitate communication among the members of the development team.

3.2 Working groups

In addition to the core development team of research scientists and programmers, multidisciplinary teams of scientists organized in collaborations—the SEEK *Working Groups*—closely inform the design and development of the IT research areas described above (Fig. 3). Collaborative working groups engage approximately 60 scientists and informatics specialists from the domains of biodiversity, ecology, earth science, and human factors to address the most critical conceptual barriers for SEEK: biological classification and nomenclature semantics, knowledge representation for the biodiversity and ecological

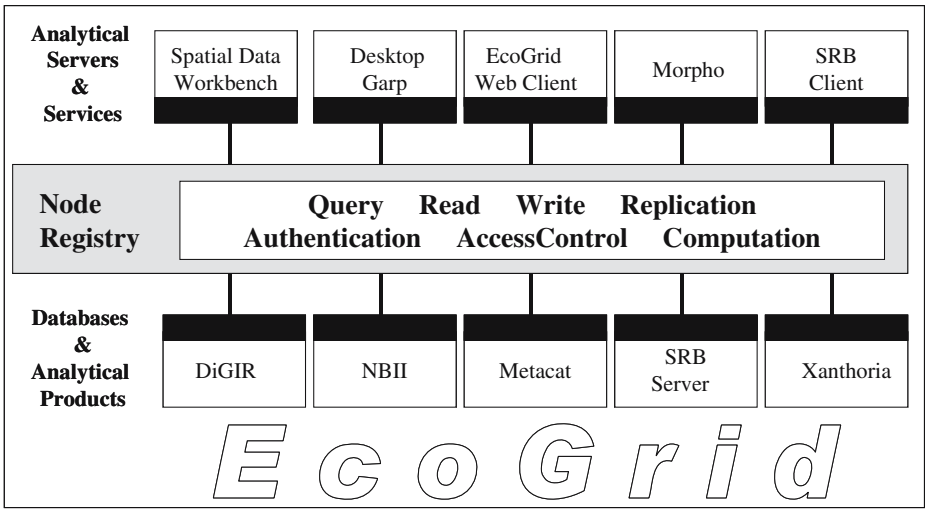

Fig. 2 A schematic of the EcoGrid interface and examples of participating nodes. Shaded areas denote SEEK development infrastructure, which interacts with nodes that will be wrapped via web services and registered for use within the EcoGrid. Core functionality of the SDSC Storage Resource Broker will be extended to manage efficient data handling and processing. Nodes function as data and analytical providers and as clients

sciences, as well as biodiversity and ecological analysis and modeling. Drawing on broad, multi-disciplinary representation, the working groups are structured so as to identify community concerns and needs, and feed that information directly into the infrastructure design process.

3.2.1 Biodiversity and ecological analysis and modeling (BEAM)

BEAM provides domain experts' knowledge in modeling and analysis to evaluate SEEK usability for addressing biodiversity and ecological questions. BEAM is initially targeting the integration and synthesis of ecological and biodiversity data. This research frontier is critical to ecological and biodiversity forecasting and necessary to enable managers and policymakers to anticipate environmental change and thereby deal with it in a more sustainable fashion. Research issues of critical interest to scientists and policymakers serve to test the utility and effectiveness of SEEK. We are first creating the information technology infrastructure that can facilitate detecting and understanding patterns in living resources and biodiversity. Future research themes will likely include understanding the interrelationships between biodiversity and ecosystem function and how they may be affected by global change, as well as other important scientific and societal issues.

We are presently testing the efficacy of the analytical workflow approach by assessing species distribution predictions with ecological niche models to predict biodiversity phenomena in response to environmental change. Ecological niche modeling is used to convert known locations of species' occurrences into spatially continuous predictions across a landscape (Peterson et al., 2001). An artificial intelligence method that has shown robust performance with improved accuracy in niche modeling is the Genetic Algorithm for Ruleset Production (GARP; Peterson & Vieglais, 2001). GARP is used to predict species

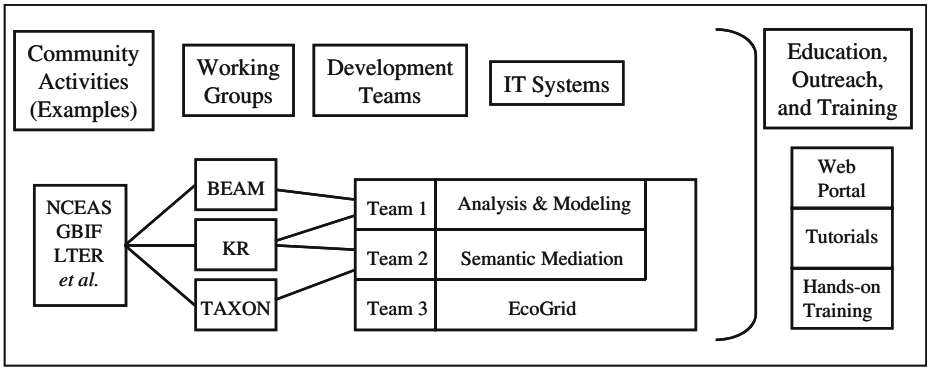

Fig. 3 Organization of development teams and working groups. Each of the three IT systems, EcoGrid, Semantic Mediation System, and Analysis and Modeling System, consists of a multidisciplinary development team of computer scientists and ecologists. Development teams interact extensively with three working groups, whose function is to address the most critical conceptual barriers for SEEK: biological classification and nomenclature semantics (TAXON), knowledge representation for the biodiversity and ecological sciences (KR), and biodiversity and ecological analysis and modeling (BEAM). The working groups are comprised of domain scientists who are engaged in the broader community (*NCEAS* National Center for Ecological Analysis and Synthesis, *GBIF* Global Biodiversity Information Facility, *LTER* Long Term Ecological Research Network). Many participants in both the development teams and working groups are cross-trained in both ecology and computer science, or have worked extensively on ecological IT problems in the past

distributions from environmental characteristics of known locations (Stockwell & Noble, 1992; Stockwell & Peters, 1999). The genetic algorithm is trained on environmental characteristics of a subset of known occurrence points. Rule fitness is tested by predictive accuracy on resampled occurrence points. The result is a set of 5 to 50 different rules in the form of "IF...THEN" statements that together define the dimensions of the species' ecological niche.

An implementation of this analytical workflow within the SEEK Analysis Workflow System is illustrated in Fig. 4. It is clear that this type of research question requires assembling a framework of modeling and integration that interconnects very diverse data streams-ecosystem function, species' geographic occurrences, and remotely sensed information (Fig. 5).

3.2.2 Knowledge representation (KR)

KR consists of ecologists, and informatics and knowledge-modeling specialists, working together to develop formal approaches that will assist in the parameterization of the Analysis and Modeling and Semantic Mediation Systems. The KR group will develop a generalized constraint language for expressing the pre- and post-conditions necessary for the Semantic Mediation System to resolve the compatibility of analysis steps and suitability of data for analytical workflows in the Analysis and Modeling System. Significant domain expertise will be necessary for the next task, which is to identify and explicitly specify key ecological concepts and their inter-relationships using techniques endorsed by the knowledge researchers. The resulting formal parameter ontologies will assist in the assembly of analytical workflows, as the function of the individual analytical steps will be described by familiar, domain-specific concepts drawn from ontologies. Initial research is focusing on developing ontologies for biodiversity and ecosystem function, and approaches

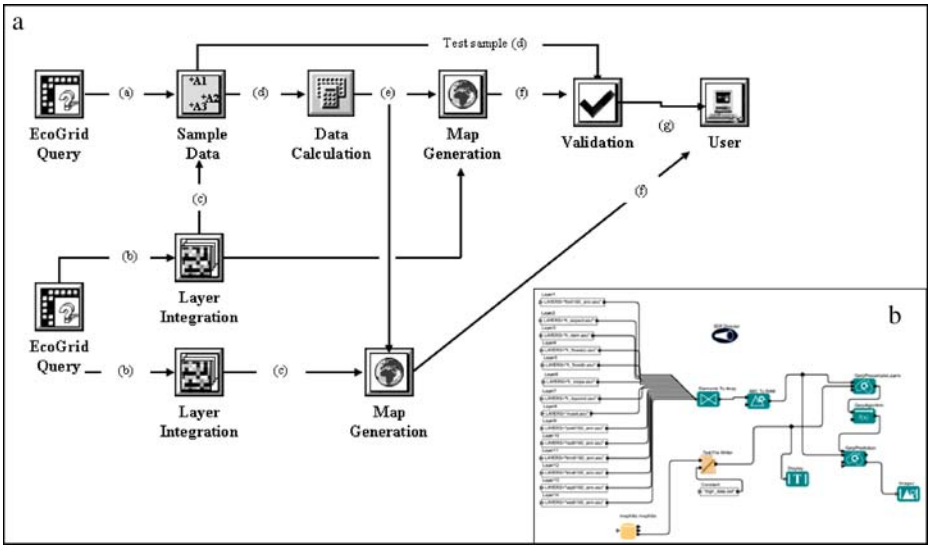

Fig. 4 Modeling the distribution of a species using appropriate ecological niche modeling algorithms (e.g., GARP—the Genetic Algorithm for Ruleset Production). **A** Abstract conceptualization of the workflow. **a** The EcoGrid is queried for data specifying the presence or absence of a particular species in a given area. These data are semantically integrated. **b** Multiple environmental layers relevant to the species' distribution are selected with a second EcoGrid query. **c** Environmental layers, representing current conditions and potential conditions after climate change, are spatially integrated. **d** Samples are selected from the presence/absence data, and the corresponding values from the environmental layers are retrieved. The sample is divided into a training set and a testing set. **e** The GARP algorithm is run on the training set. The GARP ruleset is then applied to the entire area, creating predictive maps under current conditions (native range) and after climate change (changed range). **f** Predictive maps are sent to the scientist's workstation for further analysis. **g** A comparison is made between the ground truth occurrence data that were set aside as test data, and the corresponding location on the predictive map of current conditions. Error measures providing an indication of model quality are sent to the scientist's workstation. **B** Implementation in Kepler, an analytical workflow environment

to ecological niche modeling, but the technology framework and languages chosen for this effort will be general and extensible to encompass additional topics, potentially far broader than ecology.

3.2.3 Biological classification and nomenclature (TAXON)

TAXON is focused on the challenge of mapping related taxonomic concepts to each other across multiple classifications, as well as the problem of disambiguating the many-to-many relationship that exists between scientific names and formal taxonomic concepts. The codified rules of biological nomenclature, for purposes of protecting the historical precedence and professional credit of authors who name new species, have unintentionally created monumental complexity for unambiguously mapping scientific names to species concepts. In spite of the difficulty, the precise mapping between them is critically important for scientists however, e.g., for purposes of search and retrieval when names are used as an index to associated data, such as occurrence and distribution data from field surveys or DNA sequence information from molecular analysis. Without some certainty measure of which version of a taxonomic concept was intended when a particular name was applied to

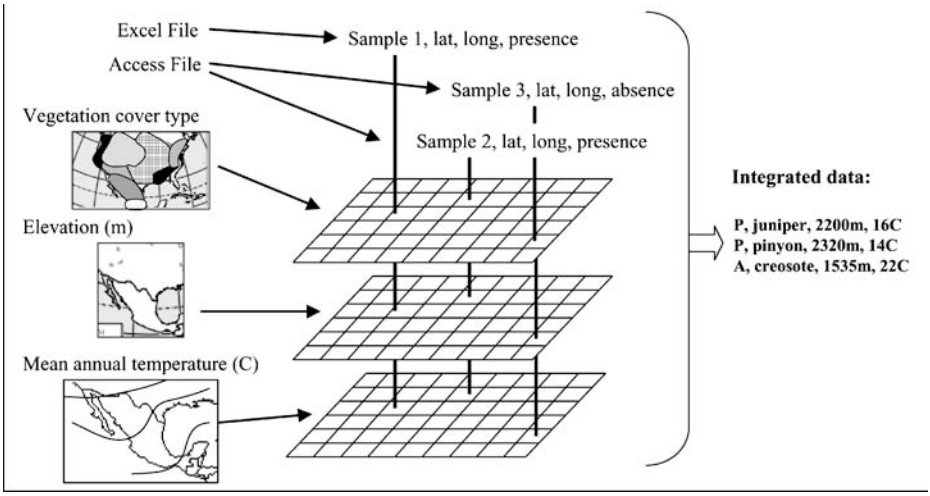

Fig. 5 Integration of heterogeneous data formats. Semantically-integrated species occurrence data are combined with spatially-integrated environmental data, to produce sample data consisting of species' occurrence (*P* present, *A* absent), vegetation type, elevation (m), and mean annual temperature (C)

a specimen or a DNA sample—it would is be impossible to know which species biological researchers were actually studying.

The Classification and Taxon working group is investigating information retrieval, natural language processing, and semantic mediation techniques to disambiguate the concepts behind the repeated application of a scientific name in different publications, data sets, or other contexts, and to determine the information relationships among multiple related taxonomic concepts. Where we do not have enough information to precisely map concepts, we will likely use probabilistic approaches for handling irresolvable ambiguity.

By developing a model, indexing and query techniques, and software tools to approach this problem, we hope to clarify the historical vagaries associated with the application of Latin names to concepts, and provide web-based mechanisms to utilize those distinctions for more precise queries and retrieval of ecological, museum and molecular data sets within the services of the Ecogrid and the overall SEEK project architecture. The work will build on considerable prior work in this area by working group members Beach, Pramanik, and Beaman (1993), Chaffee and Gauch (2000), Gauch (2002), Graham, Watson, and Kennedy (2002), Peet (2002), Pretschner and Gauch (1999), Pullan, Watson, Kennedy, Raguenaud, and Hyam (2000), Raguenaud and Kennedy (2002), Raguenaud, Kennedy, and Barclay (2000), Raguenaud, Graham, and Kennedy (2001), and Raguenaud et al., (2002).

3.3 Education, outreach and training

SEEK employs 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 is ensured through the direct inclusion of IT and domain scientists from the international scientific community in Working Groups. Working Groups include participants from the U.S. and International Long-Term Ecological Research Networks, the

Integrated Taxonomic Information System, the Organization of Biological Field Stations, Global Biodiversity Information Facility, the National Biological Information Infrastructure, BIOSIS and other relevant organizations.

A WWW portal, http://seek.ecoinformatics.org, houses and points to Internet-accessible resources (software, archives, research products and technical information) that are easily discovered, and freely accessible to the scientific community. SEEK is firmly committed to supporting the Open Source Initiative (2003).

Informatics training is coordinated and provided through tutorials at the San Diego Supercomputer Center and an intensive two-week course in informatics (supported through a National Science Foundation Research Coordination Network project) for staff and students associated with biological field stations and marine laboratories that is offered at the University of New Mexico. SEEK supports instructors and provide training materials and content for these courses.

A key element of our community outreach is an innovative annual symposium and training program that focuses on information technology transfer to young investigators and students, particularly those from underrepresented groups. Young faculty members and post-doctoral associates participate in a weeklong symposium in which the participants gain hands-on experience with the latest information technology, including products resulting from SEEK. Participants are 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.

4 Conclusion

It is anticipated that when complete, SEEK will enhance the national and global capacity for observing, studying, and understanding biological and environmental complexity in several ways. Through the development of intelligent analytical tools and an infrastructure capable of semantically integrating diverse, distributed data sources, it will remove key barriers to knowledge discovery. SEEK will enable scientists to exercise powerful new methods for capturing, reproducing, and extending the analysis process. By expanding access to distributed and heterogeneous ecological data, information, and knowledge, SEEK will create new opportunities for scientists, resource managers, policy makers and the public to make informed decisions about the environment. Finally, it will provide an infrastructure for educating and training the next generation of ecologists in the information technology skills that will be critical for scientific breakthroughs in the future.

We welcome contributions to the SEEK project as well as feedback on the technical approaches that are being taken. SEEK updates, as well as information about subscribing to mailing lists, are available through http://seek.ecoinformatics.org.

References

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludäscher, B., & Mock, S. (2004). Kepler: An extensible system for design and execution of scientific workflows. I. In *16th International Conference on Scientific and Statistical Database Management (SSDBM'04), Santorini Island, Greece* (pp. 21–23).

Baru, C., Gupta, A., Ludäscher, B., Marciano, R., Papakonstantinou, Y., Velikhov, P., et al. (1999). XML-based information mediation with MIX. In *ACM Intl. Conference on Management of Data (SIGMOD), Philadelphia, PA* (pp. 597–599).

Beach, J. H., Pramanik, S., & Beaman, J. H. (1993). Hierarchic taxonomic databases. In R. Fortuner (Ed.), *Advances in computer methods for systematic biology* (pp. 241–256). Baltimore, MD: The Johns Hopkins University Press.

Berkley, C. (2003). *Monarch: Metadata-driven analytical processing. LTER DataBits*. Retrieved Spring 2003, from http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/03spring/.

Biron, P., & Malhotra, A. (2001). *XML schema part 2: Datatypes. W3C Recommendation*. Retrieved May 02, 2001, from http://www.w3.org/TR/xmlschema-2/.

Boucelma, O., Castano, S., Goble, C. A., Josifovski, V., Lacroix, Z., & Ludäscher, B. (2002). Scientific data integration—Report on the EDBT'02 panel. *SIGMOD Record, 31*(4), 107–112.

Bowers, S., Lin, K., & Ludäscher, B. (2004). On integrating scientific resources through semantic registration. In *16th International Conference on Scientific and Statistical Database Management (SSDBM'04), Santorini Island, Greece* (pp. 21–23).

Bowers, S., & Ludäscher, B. (2004). An ontology-driven framework for data transformation in scientific workflows. In *Intl. Workshop on Data Integration in the Life Sciences (DILS'04), Leipzig, Germany*.

Brilhante, V., & Robertson, D. S. (2001). Metadata-supported automated ecological modelling. In C. Rautenstrauch & S. Patig (Eds.), *Environmental information systems in industry and public administration*. Hershey, PA: Idea Group Publishing (http://www.dai.ed.ac.uk/groups/ssp/psfiles/virginia/EnvIS_chap.aw.ps).

Brodaric, B., & Gahegan, M. (2002). Distinguishing instances and evidence of geographical concepts for geospatial database design. In M. J. Egenhofer & D. M. Mark (Eds.), *Geographic Information Science, 2nd Intl. Conference (GIScience), number 2478 in LNCS*. Boulder, CO: Springer (September). Retrieved from http://link.springer-ny.com/link/service/series/0558/tocs/t2478.htm.

Center for Environmental Studies. (2002). *Xanthoria: A distributed query system for XML encoded data. Arizona State University*. Retrieved from http://ces.asu.edu/bdi/Subjects/Xanthoria/.

Chaffee, J., & Gauch, S. (2000). Personal ontologies for web navigation. In *Ninth International Conference on Information and Knowledge Management (CIKM 2000)* (pp. 227–234).

Fonseca, F., Martin, J., & Rodreguez, M. A. (2002). From geo- to eco-ontologies. In M. J. Egenhofer & D. M. Mark (Eds.), *Geographic Information Science, 2nd Intl. Conference (GIScience), number 2478 in LNCS*. Boulder, CO: Springer (http://link.springer-ny.com/link/service/series/0558/tocs/t2478.htm).

Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., et al. (1997). The TSIMMIS approach to mediation: Data models and languages. *Journal of Intelligent Information Systems, 8*(2), 117–132.

Gauch, S. (2002). *Biodiversity Information Organization using Taxonomy (BIOT)*. In *Proc. of the National Conference on Digital Government Research, Los Angeles, CA* (pp. 169–174).

Graham, M., Watson, M., & Kennedy, J. (2002). Novel visualisation techniques for working with multiple, overlapping classification hierarchies. *Taxon, 51*(2), 351–358.

Gupta, A., Ludäscher, B., & Martone, M. E. (2002). *Registering scientific information sources for semantic mediation*. In *21st Intl. Conference on Conceptual Modeling (ER), LNCS 2503, Tampere, Finland* (pp. 182–198). Berlin Heidelberg New York: Springer.

Halevy, A. (2001). Answering queries using views: A survey. *VLDB Journal, 10*(4), 270–294 (http://link.springer.de/link/service/journals/00778/bibs/1010004/10100270.htm).

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *IEEE Internet Computing, 5*(5), 59–68.

Khoral. (2002). *Khoros Pro 2001 integrated development environment*. Retrieved from http://www.khoral.com/.

Knowledge Network for Biocomplexity. (2003). Retrieved from http://knb.ecoinformatics.org/.

Lee, E. A. (2001). *Overview of the Ptolemy project*. Technical memorandum UCB/ERL M01/11. University of California, Berkeley, CA.

Levy, A. Y. (1999). Logic-based techniques in data integration. In J. Minker (Ed.), *Workshop on logic-based artificial intelligence*. Washington, DC, College Park, MD.

London e-Science Centre (2003). *Discovery net. Imperial College of London*. Retrieved from http://www.lesc.ic.ac.uk/projects/dnet.html.

Long Term Ecological Research Network (2003). *The U.S. long term ecological research network*. Retrieved from http://www.lternet.edu/.

Ludäscher, B., Altintas, I., & Gupta, A. (2003). *Compiling abstract scientific workflows into web service workflows*. In *15th Intl. Conference on Scientific and Statistical Database Management (SSDBM), Boston, MA*.

Ludäscher, B., Gupta, A., & Martone, M. E. (2001). *Model-based mediation with domain maps*. In *17th Intl. Conf. on Data Engineering (ICDE), Heidelberg, Germany*.

Ludäscher, B., Gupta, A., & Martone, M. E. (2003). A model-based mediator system for scientific data management. In T. Critchlow & Z. Lacroix (Eds.), *Bioinformatics: Managing scientific data*. San Mateo, CA: Morgan Kaufmann.

Ludäscher, B., Papakonstantinou, Y., & Velikhov, P. (2000). Navigation-driven evaluation of virtual mediated views. In *Intl. Conference on Extending Database Technology (EDBT), LNCS 1777, Konstanz, Germany* (pp. 150–165). Berlin Heidelberg New York: Springer.

Manola, F., & Miller, E. (2003). *RDF primer, W3C working draft*. Retrieved from http://www.w3.org/TR/rdf-primer/.

MDL Information Systems, Inc. (2003). *Pipeline pilot*. Retrieved from http://www.mdli.com/products/pipelinepilot.html.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications, 7*(1), 330–342.

Natural Resource and Ecology Lab. (2003). *CENTURY soil organic matter model v. 5. Colorado State University*. Retrieved from http://www.nrel.colostate.edu/projects/century5/.

Open Source Development Network. (2003). *Distributed Generic Info Retrieval (DiGIR)*. Retrieved from http://sourceforge.net/projects/digir/.

Open Source Initiative. (2003). Retrieved from http://www.opensource.org/.

Organization of Biological Field Stations. (2003). Retrieved from http://www.obfs.org/.

Partnership for Interdisciplinary Studies of Coastal Oceans. (2002). Retrieved from http://www.piscoweb.org/.

Paton, N. W., Khan, S. A., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., et al. (2000). Conceptual modelling of genomic information. *Bioinformatics, 16*(6), 548–557.

Peet, R. K. (2002). *The VegBank taxonomic datamodel. NBII All-Node Meeting. Davis, CA*. Retrieved from http://www.bio.unc.edu/faculty/peet/pubs/NBII_Taxa.ppt.

Peim, M., Franconi, E., Paton, N. W., & Goble, C. A. (2002). Query processing with description logic ontologies over object-wrapped databases. In *14th Intl. Conference on Scientific and Statistical Database Management (SSDBM), Edinburgh, Scotland*. Retrieved from http://www.computer.org/proceedings/ssdbm/1632/16320027abs.htm.

Peterson, A. T., Sanchez-Cordero, V., Soberon, J., Bartley, J., Buddemeier, R. H., & Navarro-Siguenza, A. G. (2001). Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling, 144*, 21–30.

Peterson, A. T., & Vieglais, D. A. (2001). Predicting species invasions using ecological niche modeling: New approaches from bioinformatics attack a pressing problem. *BioScience, 51*(5), 363–371.

Pretschner, A., & Gauch, S. (1999). Ontology-based personalized search. In *Proceedings of the Eleventh IEEE International Conference on Tools with Artificial Intelligence (ICTAI '99), Chicago IL* (pp. 391–398).

Pullan, M. R., Watson, M., Kennedy, J., Raguenaud, C., & Hyam, R. (2000). The Prometheus taxonomic model: A practical approach to representing multiple classifications. *Taxon, 49*, 55–75.

Pundt, H., & Bishir, Y. (2002). Domain ontologies for data sharing—An example from environmental monitoring using field GIS. *Computers & Geosciences, 28*(1), 95–102 (http://dx.doi.org/10.1016/S0098-3004(01)00018-8).

Raguenaud, D., Graham, M., & Kennedy, J. (2001). *Two approaches to representing multiple overlapping classifications: A comparison, 13th international conference on scientific and statistical database management -SSDBM 2001* (pp. 239–244). Fairfax, Virginia: George Mason University.

Raguenaud, C., & Kennedy, J. (2002). Multiple overlapping classifications: Issues and solutions. In J. Kennedy (Ed.), *14th International conference on scientific and statistical database management—SSDBM 2002, Edinburgh, Scotland* (pp. 77–86).

Raguenaud, D., Kennedy, J., & Barclay, P. (2000). The Prometheus database for taxonomy. In O. Gunther & J. Lenz-H. (Eds.), *12th International Conference on Scientific and Statistical Database Management, SSDBM 2000, Berlin, Germany* (pp. 250–252).

Raguenaud, C., Pullan, M. R., Watson, M., Kennedy, J., Newman, M., & Barclay, P. (2002). Implementation of the Prometheus taxonomic model: A comparison of database systems. *Taxon, 51*(1), 131–142.

San Diego Supercomputer Center. (2002). *Storage resource broker*. Retrieved from http://www.npaci.edu/dice/srb/.

Sheth, A. (1998). Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In M. Goodchild, M. Egenhofer, R. Fegeas, & C. Kottman (Eds.), *Interoperating geographic information systems* (pp. 5–30). Kluwer (http://lsdis.cs.uga.edu/lib/1998.html.

Stockwell, D. R. B., & Noble, I. R. (1992). Induction of sets of rules from animal distribution data: A robust and informative method of data analysis. *Math and Computers in Simulation, 33*, 385–390.

Stockwell, D. R. B., & Peters, D. (1999). The GARP modeling system: Problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science, 13*, 143–158.

University of California Natural Reserve System. (2003). Retrieved from http://nrs.ucop.edu/default.htm.
University of Kansas Center for Research. (2002). *Desktop GARP (Genetic Algorithm for Ruleset Production)*. Retrieved from http://www.lifemapper.org/desktopgarp/.
Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods, and applications. *Knowledge Engineering Review, 11*(2), 93–155 (http://citeseer.nj.nec.com/uschold96ontologie.html).
van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2003). *OWL web ontology language reference W3C working draft*. Retrieved March 31, 2003, from http://www.w3.org/TR/owl-ref/.
Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Computer, 25*(3), 38–49.