Creating and Providing Data Management Services for the Biological and Ecological Sciences: Science Environment for Ecological Knowledge

Samantha Romanello[1], James Beach[2], Shawn Bowers[4], Matthew Jones[3], Bertram Ludäscher [4], William Michener[1], Deana Pennington[1], Arcot Rajasekar[5], Mark Schildhauer[3]
[1]Univ. New Mexico; [2]Univ. Kansas; [3]UC-Santa Barbara; [4]UC-Davis; [5]UC-San Diego
sroman@LTERnet.edu, beach@ku.edu, sbowers@ucdavis.edu, jones@nceas.ucsb.edu, ludaesch@ucdavis.edu, wmichene@lternet.edu, dpennington@lternet.edu, sekar@sdsc.edu, schild@nceas.ucsb.edu

Abstract

The Science Environment for Ecological Knowledge (SEEK) [1] is an information technology project designed to address the many challenges associated with data accessibility and integration of large-scale biocomplexity data in the ecological sciences. The SEEK project is creating cyberinfrastructure encompassing three integrated systems: EcoGrid, a Semantic Mediation System (SMS) and an Analysis and Modeling System (AMS). SEEK enables ecologists to efficiently capture, organize and search for data and analytical processes (i.e., scientific workflows) from their desktop in a user friendly interface -- ultimately providing access to global data and analytical resources typically out of reach for many ecologists. The prototype application is ecological niche modeling.

1. Introduction

The spread of the West Nile Virus, the emergence of invasive species and the effects of climate change on biodiversity and the environment are challenging ecological issues that rely heavily on acquisition of data from diverse sources and intensive computational effort. Ecological issues like these and others highlight the critical need of scientists, researchers and policy makers' to have rapid access to available data. The objective of the SEEK project is to increase the speed and efficiency of data acquisition, integration, analysis and synthesis in the biological and ecological sciences. SEEK scientists and developers are building a three-tiered information technology infrastructure composed of the EcoGrid, the Semantic Mediation System and the Analysis and Modeling System (Figure 1). The EcoGrid is an open architecture for data access across organizational and institutional boundaries. The Semantic Mediation System (SMS) is a "smart" data discovery and integration system based on domain-specific ontologies. The Analysis and Modeling System (AMS) implemented thru the Kepler workflow system supports semantically integrated analytical workflows. With the development of this infrastructure SEEK is poised, not only to provide global access to ecological data and information but also to facilitate ecological and biodiversity forecasting.

SEEK enhances the national and global capacity for observing, studying, and understanding biological and environmental complexity in several ways. First, through the development of intelligent analytical tools and an infrastructure capable of semantically integrating diverse, distributed data sources, it removes key barriers to knowledge discovery. Second, SEEK enables scientists to exercise powerful new methods for capturing, reproducing, and extending the analysis process. Third, by expanding access to distributed and heterogeneous ecological data, information, and knowledge, SEEK creates new opportunities for scientists, resource managers, policy makers and the public to make informed decisions about the environment. Finally, it provides an infrastructure for educating and training the next generation of ecologists in the information technology skills that are critical for scientific breakthroughs in the future. This paper begins with a brief description of the SEEK project, describes the three-tiered information technology infrastructure of the SEEK project and the prototype application, and concludes with a report on significant findings to date.

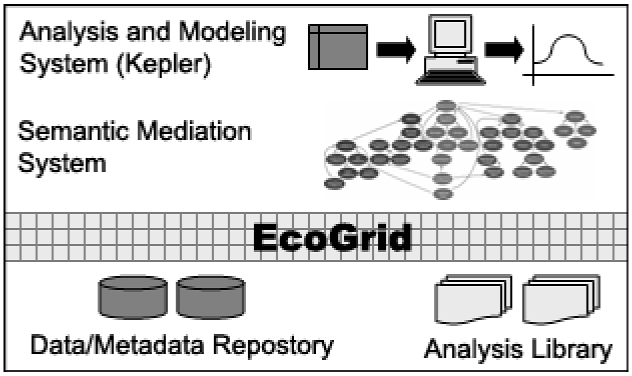
Analysis and Modeling System (Kepler)
Semantic Mediation System
EcoGrid
Data/Metadata Repostory
Analysis Library

Figure 1. SEEK architecture

2. SEEK Community

SEEK is a multi-disciplinary, multi-institutional and multi-national effort designed to create cyberinfrastructure for ecological, environmental, and biodiversity research and to educate the ecological community about ecoinformatics. SEEK infrastructure development is supported by software engineers and computer scientists dispersed across the eight institutions involved in the project. The design and development of the SEEK cyberinfrastructure is informed by three multidisciplinary teams of scientists organized in Working Groups. The Biodiversity and Ecological Analysis and Modeling Working Group (BEAM) informs development through evaluation of SEEK efficacy in addressing biodiversity and ecological questions. A Knowledge Representation Working Group (KR) develops formal ontologies that enable the assembly of analytical workflows in the Analysis and Modeling System and access to source data in EcoGrid. A Biological Classification and Nomenclature Working Group (Taxon) investigates solutions to mediating among multiple taxonomies for naming organisms. Additionally, a multifaceted Education, Outreach and Training (EOT) program ensures that the SEEK research products, software, and information technology infrastructure optimally benefit the target communities via the project website (http://seek.ecoinformatics.org) [2, 3].

3. The EcoGrid

The EcoGrid [4] is a collection of distributed ecological, biodiversity and environmental data and analytic resources (data, metadata, analytic workflows and processors) that are often located at different sites and in different organizations. The EcoGrid uses the Open Grid Services Architecture (OGSA, http://www.globus.org/ogsa/) framework to provide a set of standardized interfaces for accessing data resources through a service-oriented framework. The current prototype implementation of the EcoGrid uses the OGSA 'Factory" service to enable scalable deployment of the access services, but in other aspects is more similar to a traditional web service implementation. As the Web Services Resource Framework (WSRF) matures we will investigate migrating our services to the WSRF specification. EcoGrid combines features of a Data Grid for ecological data management and a Compute Grid for analysis and modeling services. EcoGrid forms the underlying framework for data and service discovery, data sharing and access and analytical service sharing and invocation.

Biodiversity and ecological data include, but are not limited to the heterogeneous data collected at field stations, as well as remote sensing data and data from museum collections. Computational models and analyses include well-known biodiversity and ecosystem models such as GARP (Genetic Algorithm for Ruleset Production; Stockwell and Noble, 1992 [5]; University of Kansas Center for Research, 2002 [6]) and CENTURY (Natural Resource and Ecology Lab, 2003 [7]) 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 diverse data and computational resources.

4. Analysis and Modeling System

The Analysis and Modeling System in SEEK is a multiplatform, open-source, visual programming tool (i.e., the Kepler [8, 9, 10] workflow system) that allows users to create executable analytical pipelines and workflows based on research models. Kepler, based on Ptolemy II [11], is a collaborative project involving contributing members from SEEK, the GEOsciences Network (GEON) http://www.geongrid.org/, the Scientific Data Management Center (SDM) part of the Scientific http://sdm.lbl.gov/sdmcenter/, the Ptolemy Project http://ptolemy.eecs.berkeley.edu/ptolemyII/, the ROADNet (Real-time Observatories, Applications, and Data Management Network) Project http://roadnet.ucsd.edu/ and the EOL (Encyclopedia of Life) http://eol.sdsc.edu/. Scientific workflows are a formalization of the scientific research process (Figure 2). That is, typically a scientist will generate a research question, collect data, analyze the data using several models, programs, software and hardware, and physically coordinate the data transformation, exporting and importing.

Kepler allows scientists to design, execute, monitor, re-run and communicate analytic procedures with minimal effort. Therefore, scientific workflows in Kepler include the analysis steps as well as the data acquisition, integration, transformation, synthesis and archival information. Scientists can create and save workflows on the EcoGrid within Kepler. These workflows can then be searched and downloaded by other researchers for replicating or expanding upon the analysis.

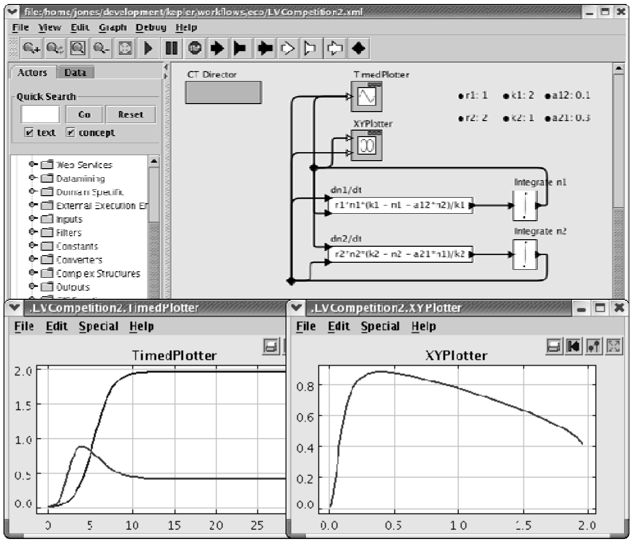

Figure 2. The Kepler workflow system showing the Lotka-Volterra predator-prey model.

5. Semantic Mediation System

The Analysis and Modeling System (i.e., Kepler) in the SEEK architecture leverages the Semantic Mediation System (SMS) [12, 13]. The goal of the SMS layer is to support scientists' workflow modeling and design processes. In particular, SMS exploits domain ontologies to facilitate (1) "smart discovery" of data sets and components (individual actors and complete workflows), (2) "smart binding" of data sets to components, and (3) "smart linking" of components to each other as part of the overall design process.

The Semantic Mediation System provides a generic set of ontology-based languages and tools for storing and exploiting "superimposed" semantic annotations [15], which explicitly link existing data sets and workflow components to ontologies. Through semantic annotations, the mediation layer provides knowledge-based data integration and workflow composition services [13,14], as well as basic services used in workflow modeling, such as ensuring that workflows are "semantically" type-safe (based on annotations) and component and data discovery via concept-based searching.

Ontology development is a major part of SEEK. We have developed initial ontologies for ecological data and workflows, focusing on measurements, basic ecological concepts, symbiosis, and biodiversity. We are also building tools to support the editing and curation of ontologies, with the goal of making these tools accessible and easy to use for domain scientists.

6. Prototype application

A new and promising paradigm in biodiversity informatics is the use of ecological niche modeling to extrapolate and anticipate implications of global climate change for biological diversity [17, 18, 19]. Future scenarios based on general circulation models (GCMs) present diverse visions of global climate futures. The implications of these different futures for biodiversity are only now being explored. While data suggest that climates are changing, the implications of these changes remain unclear and little explored. At this time there are no hemisphere-wide evaluations or broad comparative analyses of implications of different GCM modeling scenarios, due to the prohibitive time costs for large-scale analyses. With the use of distributed resources and the building of analytic workflows for automated processing of climate change and biodiversity analyses, the first application for the SEEK project is a large scale ecological niche modeling assessment of mammals of Western Hemisphere to look at the implications of climate change on current and projected habitat range. This application models distributions of all mammal species in the Western Hemisphere and generates projections of distribution change under multiple Intergovernmental Panel on Climate Change scenarios (http://www.ipcc.ch). This project includes the analysis of integrated field data for over 3000 mammal species, under 20+ climate scenarios, using 2-3 dispersal scenarios (180,000+ model runs).

7. Significant results

To date, a variety of SEEK tools have been created including a protoype of the Ecogrid has been created (Table 1).

1. Currently the Ecogrid provides: access to different data catalogs (Metacat, SRB/MCAT, DiGIR); a search, read and write interface; and uploading of metadata and data.
2. The first alpha-quality users' release of Kepler was in May 2004. Kepler has an Ecological Metadata Language -aware data plug-in, an EcoGrid plug-in, web service actors and a web service harvester.
3. Toolkit for reasoning and data conversion, and an access API for ontologies called “Sparrow” was developed. A user-friendly editor for OWL ontologies (grOWL) is currently in its second alpha release.

4. The Biological Classification and Nomenclature (Taxon) working group has created a “Taxonomic Object Service” (TOS) that provides information about the relationships among taxa via SOAP and web interfaces.

Table 1. SEEK Tools

Application	Description	url
Kepler	Is a flexible workflow system designed to process and ingest heterogeneous ecological data from ecologists and other domain scientists.	http://kepler.ecoinformatics.org
Sparrow	Aims at combining algorithms and techniques from logic-based knowledge representation and databases into a single, open-source toolkit.	http://seek.ecoinformatics.org/
GrOWL	Is a visualization and editing tool for Ontology Web Language (OWL) and Description Logics (DL) ontologies based on a semantic network knowledge representation paradigm	http://ecoinformatics.uvm.edu/dmaps/growl
EcoGrid	Is a thin layer to allow various data and computer services already in existence to interoperate base on GRID technology.	http://seek.ecoinformatics.org/

Additional information about these and other SEEK tools can be found at http://seek.ecoinformatics.org.

8. Acknowledgements

This work is supported in part by NSF grants ITR 0225674 (SEEK), 0225673 (GEON), EF 0225665 and DBI 0129792; DARPA N00014-03-1-0900 and DOE SciDAC DE-FC02-01ER25486 (SDM).

9. References

[1]SEEK: Science Environment for Ecological Knowledge, http://seek.ecoinformatics.org

[2]W. K. Michener, J. H. Beach, M. B. Jones, B. Ludaescher, D. D. Pennington, R. S. Pereira, A. Rajasekar, and M. Schildhauer, "A Knowledge Environment for the Biodiversity and Ecological Sciences", *Journal of Intelligent Information Systems*, 2004.

[3]W K. Michener, "Building SEEK: the Science Environment for Ecological Knowledge ", *DataBits: An electronic newsletter for Information Managers*, Spring, 2003.

[4]M. B. Jones, "SEEK EcoGrid: Integrating Data and Computational Resources for Ecology", *DataBits: An electronic newsletter for Information Managers*, Spring 2003.

[5]D.R.B.Stockwell, and I.R. Noble. “Induction of Sets of Rules From Animal Distribution Data: A Robust And Informative Method of Data Analysis”. *Mathematics and Computers in Simulation*, 32, 1992, pp. 249–254.

[6]http://www.lifemapper.org/desktopgarp/#acknowledge

[7] http://www.nrel.colostate.edu/projects/century/

[8]Kepler: An Extensible System for Scientific Workflows, http://kepler.ecoinformatics.org

[9]I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock, Kepler: Towards a Grid-Enabled System for Scientific Workflows, *Workflow in Grid Systems, GGF10,* Berlin, 2004.

[10] I. E. Altintas, E. Jaeger, K. Lin, B. Ludaescher, and A. Memon. A Web Service Composition and Deployment Framework for Scientific Workflows. In *2nd Intl. Conference on Web Services (ICWS),* San Diego, California, July 2004.

[11] E. A. Lee, "Overview of the Ptolemy Project," *Technical Memorandum UCB/ERL M03/25*, University of California, Berkeley, CA, July 2, 2003.

[12] C. Berkley, S. Bowers, M. Jones, B. Ludaescher, M. Schildhauer, J. Tao. Incorporating semantics in scientific workflow authoring. In *Proceedings of the 17th International Conference on Scientific and Statistical Database Managemen*t (SSDBM'05)

[13]B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, Y. Zhao, “Scientific Workflow Management and the Kepler System,” *Concurrency and Computation: Practice & Experience,* Special Issue on Scientific Workflows, to appear, 2005

[14] S. Bowers and B. Ludaescher, “An Ontology-Driven Framework for Data Transformation in Scientific Workflows,” In *Proceedings of the International Workshop on Data Integration in the Life Sciences (DILS'04),* volume 2994.A, Springer, LNCS, Leipzig, Germany, March 25-26, 2004

[15] S. Bower, D. Thau, R. Williams and B. Ludaescher, “Data Procurement for Enabling Scientific Workflows: On Exploring Inter-Ant Parasitism” In *Proceedings of the 2nd International Workshop on Semantic Web and Database*s (SWDB'04), Toronto, Canada, 29-30 August, 2004.

[16] S. Bowers, K Lin, and B. Ludaescher, “On Integrating Scientific Resources through Semantic Registration,” In *Proceedings of the 16th International Conference on Scientific and Statistical Database Managemen*t (SSDBM'04), 21-23 June 2004, Santorini Island, Greece.

[17]A.T. Peterson, and D. A. Vieglais. “Predicting Species Invasions Using Ecological Niche Modeling,” *BioScience*, 51, 2001, pp. 363-371.

[18]Peterson, A.T., H. Tian, E. Martínez-Meyer, B. Huntley, J. Soberón, and V. Sánchez-Cordero. Modeling distributional shifts of individual species and biomes. In T. E. Lovejoy and L. Hannah (eds.), *Biodiversity and Climate Change*, Yale University Press, New Haven, Conn, 2005f.

[19]A. T. Peterson, M. A. Ortega-Huerta, J. Bartley, V. Sanchez-Cordero, J. Soberon, R. H. Buddemeier, and D. R. B. Stockwell, “Future Projections for Mexican Faunas Under Global Climate Change Scenarios,” *Nature*, 416, 2002b, pp. 626-629.