Data Integration and Workflow Solutions for Ecology*

William Michener[1], James Beach[2], Shawn Bowers[3], Laura Downey[1], Matthew Jones[4], Bertram Ludäscher[3], Deana Pennington[1], Arcot Rajasekar[5], Samantha Romanello[1], Mark Schildhauer[4], Dave Vieglais[2], and Jianting Zhang[1]

[1] LTER Network Office, University of New Mexico
[2] Biodiversity Research Center, University of Kansas
[3] Genome Center & Dept. of Computer Science, University of California, Davis
[4] NCEAS, University of California, Santa Barbara
[5] San Diego Supercomputer Center, University of California, San Diego

1 SEEK: Introduction and Architecture

The Science Environment for Ecological Knowledge[1] (SEEK) is designed to help ecologists overcome data integration and synthesis challenges. The SEEK environment enables ecologists to efficiently capture, organize, and search for data and analytical processes. We describe SEEK and discuss how it can benefit ecological niche modeling in which biodiversity scientists require access and integration of regional and global data as well as significant analytical resources.

SEEK is designed as a three-layer architecture. The *EcoGrid* forms the base layer and provides a uniform and simple programming interface for access to distributed resources such as data, metadata, and workflows. The KEPLER *Scientific Workflow System*[2] forms the topmost layer and provides tools that allow scientists to create and compose scientific workflows (e.g., analytical models), execute them, and archive the results. KEPLER makes extensive use of EcoGrid interfaces. For instance, through the EcoGrid, KEPLER allows scientists to search for and retrieve data and workflows stored across distributed repositories. The *Semantic Mediation System* (SMS) forms the middle layer of the architecture and mediates between heterogeneous resources in the EcoGrid and the analyses and models to be executed in KEPLER. SMS leverages ontologies to facilitate data integration and workflow composition, thereby increasing the scale and complexity of analyses that can be constructed and executed by scientists. Each of these layers is described further below.

The EcoGrid [4] layer forms the underlying cyberinfrastructure within SEEK for enabling remote data and service discovery, data sharing and access, and remote service invocation. EcoGrid services and interfaces are being built using best practices currently available in grid technology (e.g., OGSA/WSRF, SRB, and Condor). EcoGrid provides resource discovery through a registration service. Many data sets accessible through the

* We thank the other members of SEEK, including: Chad Berkley, Dan Higgins, Jessie Kennedy, Ricardo Pereira, Town Peterson, Aimee Stewart, Jing Tao, and Bing Zhu. This work is supported in part by NSF grants ITR 0225674, EF 0225665 and DBI 0129792, DARPA grant N00014-03-1-0900, and the Andrew Mellon Foundation.

[1] seek.ecoinformatics.org
[2] www.kepler-project.org

B. Ludäscher and L. Raschid (Eds.): DILS 2005, LNBI 3615, pp. 321–324, 2005.

EcoGrid have Ecological Metadata Language[3] (EML) descriptions that are also used for data discovery, access, and integration.

KEPLER is used to design and execute scientific workflows [6] (see Figure 1). KEPLER includes components (called *actors*) to access data from the EcoGrid as well as other generic scientific workflow components including R and Matlab modules for statistical analysis. From these components, customized scientific workflows can be built such as the Genetic Algorithm for Ruleset Production (GARP) discussed below. Existing components and workflows can be linked within KEPLER to form a new scientific workflow graph. The inputs and outputs of components are represented using ports, which can have structural types describing the physical representation of data (e.g., `double`) as well as semantic types describing the conceptual meaning and scientific context of data (e.g., BODYSIZE) [1, 3]. The SMS system uses structural and semantic types to help scientists construct meaningful scientific workflows.

The SMS layer provides ontology-based services to KEPLER including support for data integration, workflow composition, and concept-based searching. The SEEK Knowledge Representation Team (KR) includes ecologists and knowledge engineers who jointly develop and maintain formal ontologies to be used by the SMS. These ontologies cover a number of different areas including measurement, time and space, basic ecological concepts, biodiversity, and unit systems. Also as part of KR/SMS, the SEEK Taxonomic Object Service [5] is being developed to help resolve progressive changes of taxonomic names to sets of taxonomic concepts, providing well defined, authoritative, and (ideally) unambiguous information about the identification of organisms.

2 Use Case: Ecological Niche Modeling

A new and promising paradigm in ecology is the use of ecological niche modeling (ENM) to extrapolate implications of global climate change for biological diversity [7]. Figure 1 shows the KEPLER implementation of an ENM workflow that assesses the implications of climate change for mammals of the Western Hemisphere. Such broad-scale comparative analyses of effects of different climate-change modeling scenarios are difficult to implement due to their computational complexity, which includes data discovery ($>$ 3,000 mammal species), data integration (20 climate scenarios), and analytical complexity ($>$ 180,000 model runs).

ENM incorporates both spatially explicit point data indicating where a species has been found as well as spatially explicit environmental data such as descriptions of climate, hydrology, and soils. Within KEPLER, scientists can use EcoGrid-based search interfaces to discover occurrence data (e.g., within the DiGIR network) as well as environmental data (e.g., located within Metacat or SRB collections)[4]. These searches leverage metadata about objects to locate relevant items of interest and present them to the user. For ENM, partitioning the relevant taxonomic data into species groupings may be difficult as a result of changes in taxonomic names. The Taxonomic Object Service can be used to help resolve these taxonomic clustering issues.

[3] knb.ecoinformatics.org/software/eml/

[4] digir.sourceforge.net/, knb.ecoinformatics.org/software/metacat/, www.sdsc.edu/srb/

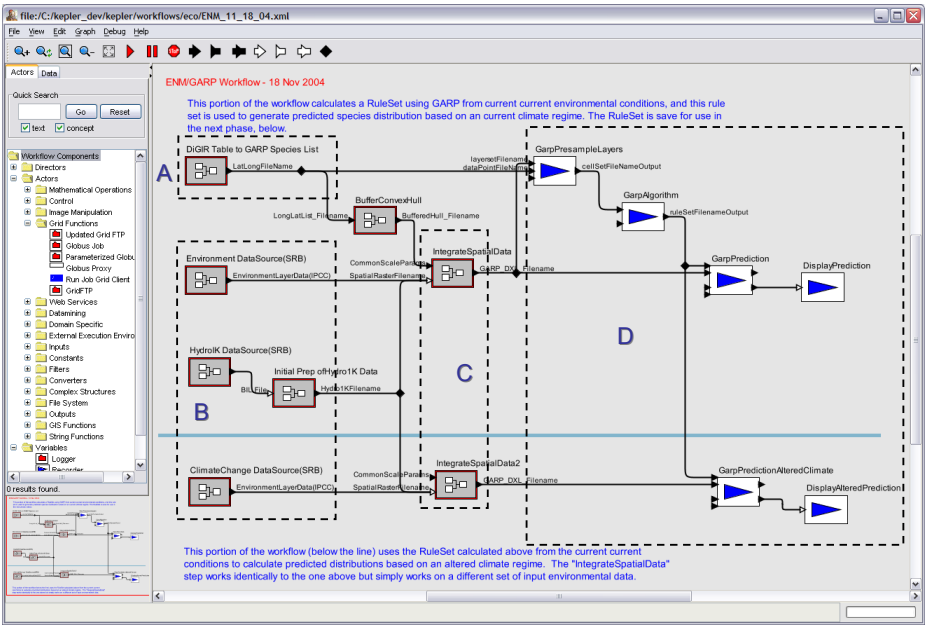

Fig. 1. The ENM workflow in KEPLER with components for: (A) accessing and pre-processing DiGIR species occurrence data; (B) accessing and pre-processing SRB environmental data; (C) integrating occurrence data and environmental layers; and (D) GARP modeling steps

Once relevant data sets are discovered they can be directly imported into KEPLER. Data access components are provided by KEPLER that can use the detailed descriptions of the physical data structure of EML to automate the process of importing data. Thus, using EML-described data sets in a workflow simply involves dragging their associated icons onto the workflow canvas (Figure 1, A and B). KEPLER parses the metadata and exposes output ports that represent each attribute within the data. It also provides a Query-By-Example extension for user-friendly SQL query construction.

SMS provides a generic set of ontology-based languages and tools for storing and exploiting semantic annotations [1, 3], 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 [2], component and data discovery via concept-based searching, as well as basic services used in workflow modeling such as ensuring that workflows are "semantically" type-safe (based on annotations). In the ENM case, each of the data types must undergo a series of transformations for integration, including re-projection to a common geographic coordinate system, re-scaling to a common resolution, and re-orientation to center the imagery on the same point on the globe. The locations of occurrence points are used to sample the environmental data and create vectors containing many bands of information associated with each occurrence point (Figure 1, C).

The ENM workflow analyzes native distributions of species using a genetic algorithm (GARP; Figure 1, D) written in C++ [8]. The GARP algorithm generates a rule-

based model from input data. The workflow is run a number of times to generate a set of distinct models. The models are then used to construct a probabilistic prediction of the full distribution range under current climate conditions, and potential distributions under various climate change scenarios. The workflow consists of more than fifty components in approximately ten nested workflows including the GARP algorithm, grid access and query components, GIS components in GRASS and GDAL, statistical components developed in R, and image processing and viewing components developed in ImageJ. This workflow is reusable by multiple biodiversity scientists in many different applications in its current form, and can readily be modified for additional applications.

Finally, data derived during the execution of the ENM workflow can be saved to the EcoGrid. KEPLER workflows can be configured to allow any output from a component to be written to the EcoGrid with appropriate metadata, completing the "analysis cycle" by allowing future work to seamlessly use the results of an existing workflow.

3 Conclusion and Future Project Directions

SEEK encompasses many cyberinfrastructure tools needed to integrate complex ecological data and enable rapid development and re-use of complex scientific analyses. Nevertheless, many challenges remain. Future work includes: (1) exploration of new ways to leverage and extend the Taxonomic Object Service; (2) use of the Geographical Markup Language to achieve greater interoperability of spatial/GIS data; (3) additional support for semantic annotations in workflow design and execution [3]; (4) native support for scheduling of compute-intensive, distributed scientific workflows; (5) additional geospatial semantics for ontology-based ENM workflow compositions; and (6) usability engineering to improve SEEK tools.

References

1. C. Berkley, S. Bowers, M. Jones, B. Ludäscher, M. Schildhauer, and J. Tao. Incorporating semantics in scientific workflow authoring. In *SSDBM*, 2005.
2. S. Bowers and B. Ludäscher. An ontology-driven framework for data transformation in scientific workflows. In *Intl. Workshop on Data Integration in the Life Sciences*, 2004.
3. S. Bowers and B. Ludäscher. Actor-oriented design of scientific workflows. In *Intl. Conf. on Conceptual Modeling (ER)*, 2005.
4. M. Jones. SEEK EcoGrid: Integrating data and computation resources for ecology. *DataBits: An electronic newsletter for Information Managers*, Spring 2003.
5. J. Kennedy, T. Paterson, and R. Kukla. Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. In *Intl. Workshop on Data Integration in the Life Sciences*, 2005.
6. B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practic & Experience, Special Issue on Scientific Workflows*, 2005.
7. A. Peterson, M. Ortega-Huerta, J. Bartley, V. Sanchez-Cordero, J. Soberon, R. Buddemeier, and D. Stockwell. Future projections for mexican faunas under global climate change scenarios. *Nature*, 416, April 2002.
8. D. Stockwell and D. Peters. The GARP modelling system: Problems and solutions to automated spatial prediction. *Intl. Journal of Geographical Information Science*, 13, 1999.