SEEK: Data Integration and Workflow Solutions for Ecology

Shawn Bowers (UC Davis Genome Center) and the SEEK Team

Science Environment for Ecological Knowledge

NSF-funded, multidisciplinary research project to facilitate ...

Access to ecological, environmental, and biodiversity data
- Enable data sharing & reuse
- Enhance data discovery at global scales

Scalable analysis and synthesis
- Taxonomic, Spatial, Temporal, Conceptual integration of data, addressing data heterogeneity issues
- Enable communication and collaboration for analysis
- Enable reuse of analytical components
Main Components:

- **Kepler**
  - Problem-solving environment for scientific data analysis and visualization ➔ “scientific workflows”

- **EcoGrid**
  - Distributed data network for environmental, ecological, and systematics data
  - Making diverse environmental data systems interoperate

- **Semantic Mediation System**
  - Ontology-based data and component discovery and integration

- Knowledge Representation WG
- Taxon WG
- BEAM WG
- Education, Outreach, Training
The Kepler Scientific Workflow System

Based on Ptolemy II
- Developed by electrical engineering community as a visual dataflow programming application
- Hierarchical workflows (actors) and explicit computation models, e.g., process networks, continuous time, discrete event, etc.

Kepler adds
- Compute & Data Intensive
- Web-service support
- Generic & domain-specific actor libraries (R, data transport, niche-modeling, phylogenetics, etc.)
- EML-based metadata tools
- Distributed Execution
- EcoGrid access/query
- Ontology and integration support

Example: Ecological niche modeling workflow to assess implications of global climate change for western hemisphere mammals
The Kepler Scientific Workflow System

- Access and pre-process species occurrence data (DiGIR)
The Kepler Scientific Workflow System

This portion of the workflow calculates a RuleSet using GARP from current current environmental conditions, and this rule set is used to generate predicted species distribution based on an actual climate regime. The RuleSet is saved for use in the next phase, below.

This portion of the workflow (below the line) uses the RuleSet calculated above from the current current conditions to calculate predicted distributions based on an altered climate regime. The "IntegrateSpatialData" step works identically to the one above but simply works on a different set of input environmental data.

- Access and pre-process environmental data (SRB)
Integrate occurrence data and environmental layers
The Kepler Scientific Workflow System

- GARP modeling steps and visualization
In general, SEEK scientific workflows are complex, involving multiple, distributed, and heterogeneous data sets and analytical steps …
Seek EcoGrid

Allow diverse environmental data systems to interoperate
- Hide complexity of underlying systems using lightweight interfaces
- Integrate diverse data networks from ecology, biodiversity, and environmental sciences

Data systems
- Systems contribute by implementing EcoGrid Grid-service interfaces
- Prototypes exist for: Metacat, SRB, DiGIR, and Xanthoria

Supports multiple XML-based metadata standards
- EML & Darwin Core as foci
EcoGrid Queries in Kepler

A simple example of using EML data. First, a search is done in the Data pane to locate an EML-described data set, which is dragged onto the workflow canvas. The EML data source is added to the workflow, and then it contacts the EcoGrid server to download the data and configure the ports. After being configured, it displays the ports from the EML data source, which are then mapped into an XY scatterplot.
EcoGrid Sources in Kepler
A simple example of using EML data. First, a search is done in the Data pane to locate an EML-described data set, which is dragged onto the workflow canvas. The EML data source is added to the workflow, and then it contacts the EcoGrid server to download the data and configure the ports. After being configured, it displays the ports from the EML data source, which are then mapped into an XY scatter plot.
A simple example of using EML data. First, a search is done in the Data pane to locate an EML-described dataset, which is dragged onto the workflow canvas. The EML data source is added to the workflow, and it contacts the EcoGrid server to download the data and configure the ports. After being configured, it displays the ports from the EML data source, which are then mapped into an XY scatterplot.
SEEK Knowledge Representation (KR)

Developing **OWL-DL ontologies** for ecology

Measurements, Biodiversity concepts, Units, Spatial/Temporal, Location, Food Webs, Symbiosis, Models, Statistics, …

Kepler supports **semantic types** – annotations relative to the ontologies – allowing workflow designers to:

- semantically describe components
- check that a given workflow is well-typed
- perform ontology-based browsing and search
Typing Workflow Components

Semantic Type Editor is used to assign one or more semantic types to the component or to the component’s input and output ports. In the simplest case, a semantic type is a class taken from an OWL-DL ontology. Multiple types define a conjoined concept expression.

A simple ontology browser is provided in Kepler to navigate a classified OWL-DL ontology. Classes can be searched for and selected as a semantic type.
Kepler can **statically perform semantic and structural type checking** of connections. A type checker allows the user to see potentially mismatched port connections as well as known type conflicts before workflow execution.

The user can **navigate the unsafe and potentially unsafe** channels using the Kepler Type Checker dialog. When a channel is selected: (a) it is highlighted on the canvas, (b) the structural type and status is shown (here, the channel is structurally well typed), and (c) the semantic type and status is shown (here, the connection produce a semantic type error).
Kepler Actor-Library

- Ontology-based actor organization / browsing
- Customizable libraries based on ontologies
- Text search with concept-based expansion

Users can discover ImageJ using various search terms. Here, ImageJ shows up in multiple tree locations based on its given annotations. The library search permits text-based matching against the component’s metadata (its given name and certain properties), expanded with concept matches.
Kepler provides a more advanced ontology-based search mechanism. Users can start the **Semantic Search** dialog, where components can be search for based on their semantic types.

The Semantic Search dialog allows a user to search components by any combination of actor, input, and output semantic types.
Concluding Remarks

Some ongoing and future integration work within SEEK:

– Adding additional adaptor (“shim”) components:
  • Wrapping external components into Kepler
  • Developing new actors for specific integration tasks
  • Applying these to ENM and Biodiversity workflows

– Data procurement and pre-processing
  • (Semi-) automating the merging and integration of data sets via semantic types

– Implementing advanced EcoGrid-based metadata query UI

– Employing semantic types and schema matching approaches for component composition
Acknowledgements

This material is based upon work supported by:

The National Science Foundation under Grant Numbers 9980154, 9904777, 0131178, 9905838, 0129792, and 0225676.

The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.

The Andrew W. Mellon Foundation.

PBI Collaborators: NCEAS, University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research), UC Davis Genome Center

Kepler contributors: SEEK, Ptolemy II, SDM/SciDAC, GEON, RoadNet, EOL, Resurgence