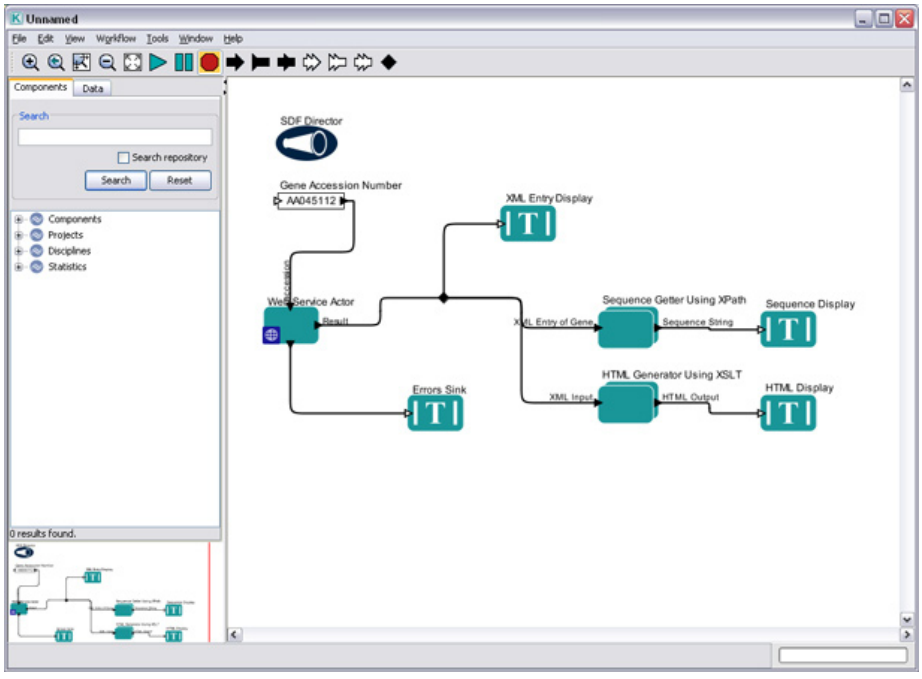
Kepler Project Newsletter

News from Kepler/CORE May 2008

Kepler 1.0.0 Released

The Kepler Project is pleased to announce the availability of Kepler 1.0.0, the first official release of the Kepler scientific workflow system. Representing the combined efforts of numerous individuals and projects, Kepler is a user-friendly, open-source application for analyzing, modeling, and sharing scientific data and analytical processes.

"This first release of Kepler will be of tremendous use to scientists from many disciplines because it provides an integrated environment for creating and executing analyses and models that were previously managed independently, " says Matthew Jones, Director of Informatics Research and Development at the National Center for Ecological Analysis and Synthesis (NCEAS) at UC Santa Barbara. "Kepler will facilitate significant science advances because it provides a platform for communicating precisely about the scientific process via the ability to share scientific workflows."



The Kepler workflow system is designed to help scientists with little background in computer science as well as analysts and computer programmers build models called scientific workflows. Workflows help solve the all too common problem of analyzing data stored in a variety of formats with software components deployed and invoked in different ways. By providing a sound infrastructure that permits users to easily integrate a wide diversity of data and analytical components, Kepler not only facilitates the execution of a specific analysis, but helps users share and reuse data, workflows, and components developed by the community to address common problems. Kepler workflows have been used to study the effect of climate change on species distribution, to simulate supernova explosions, to identify transcription factors, and to perform statistical analyses. The variety of applications is as broad as today's exciting range of scientific studies.

"We wanted to create an open, customizable, extensible and robust scientific workflow environment that is useful in solving problems in a variety of scientific disciplines through access to a diverse portfolio of technologies," says Ilkay Altintas, Lab Director, Scientific Workflow Automation Technologies (SWAT) San Diego Supercomputer Center(SDSC). "Kepler 1.0.0 is the first official release that realizes our goal. It provides a basis for many exciting features that are in the works and coming soon."

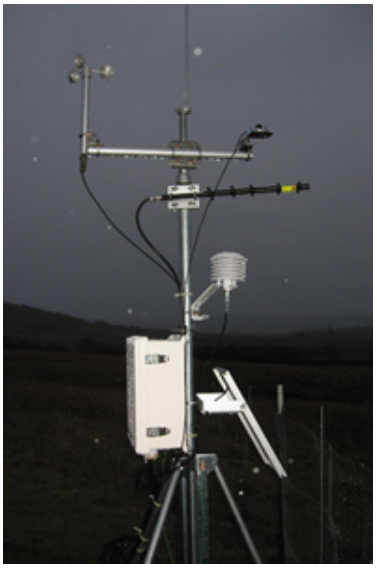
Workflows can leverage the computational power of grid technologies, and/or take advantage of Kepler's native support for parallel processing. Once created, workflows and customized components can be saved and reused, or shared with colleagues using the Kepler archive format (KAR). KAR (see Release on p. 4)

Kepler and Streaming Sensor Data: The REAP Project

The Kepler and the Real-time Environment for Analytical Processing (REAP) projects have been making great strides in a collaborative effort to create a near real-time environment for the analytical processing of data streams from both oceanic and terrestrial sensor networks. The project unites oceanographers, marine and terrestrial ecologists, technologists, and computer science researchers in the goal of constructing, testing and deploying scientific workflow systems that access, monitor, and analyze sensor data.

Project investigators are combining the real-time data grid being constructed through other projects (Data Turbine, OPeNDAP, EarthGrid) and the Kepler scientific workflow system to provide a framework for designing and executing scientific models that use sensor data. To this end, project collaborators are extending Kepler to access sensor data in workflows, monitor, inspect and control sensor networks, and simulate the design of new sensor networks.

Several new Kepler components have already emerged from the collaboration and are included in the Kepler 1.0.0 release: an OPeNDAP actor that can be used to access and output any Data Access Protocol (DAP) 2.0 compatible data source, and a new and improved MATLAB actor that incorporates the powerful data processing and visualization functionality of MATLAB ("MATrix LABoratory"). The new MATLAB actor responds to the presence or absence of local system resources to invoke MATLAB in the most efficient way possible. If the required Java libraries are present, the MATLAB actor will interface directly with the MATLAB engine; otherwise, MATLAB will be invoked via the command-line, and the application will be started each time the actor executes. Investigators are currently developing prototype workflows that use the technologies emerging from the collaboration to address the challenges of accessing and integrating environmental data. In one use case, researchers studying a large-scale invasion of non-native grasses occurring in the western United States are modeling the susceptibility of plants to an aphid-vectored pathogen. REAP investigators are prototyping a wide network (ranging from Canada to Mexico) of real-time land-based environmental sensors that stream data to a DataTurbine server, an open source streaming data middleware software that provides a robust and generic interface for accessing data from a diverse set of sensors, that is accessed by Kepler and provides data to answer questions about the pathogen-host community.

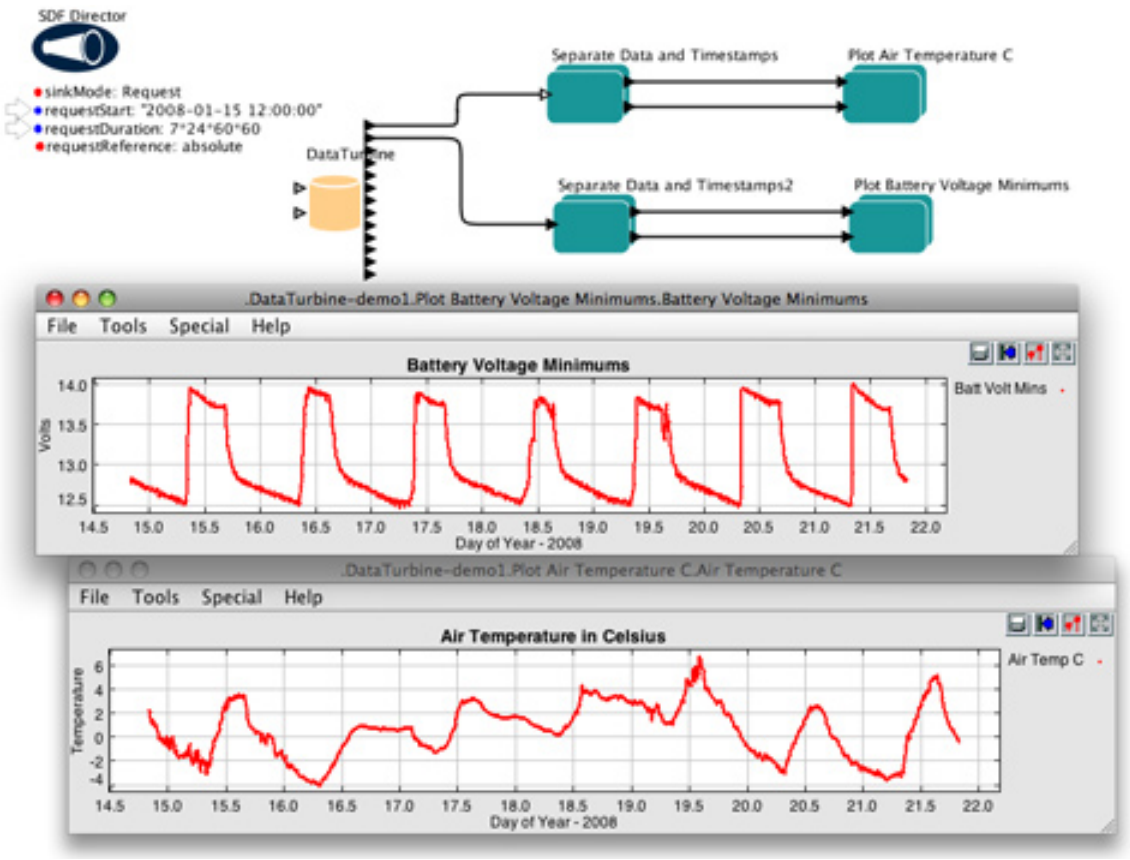


REAP weather station, deployed at the Baskett Slough National Wildlife Refuge.

In a second prototype, the new technologies are being used to compare and match-up existing remote-sensed images of sea surface temperature found in OPeNDAP archives. Users specify the time span, sampling rate, and geographical location of interest, and Kepler actors identify and return matching data sets based on information contained in metadata. The Kepler workflow will reduce the complexity of comparing and integrating data collected by a variety of satellite-borne, ship-board, and other in situ instruments that form today's bewildering array of sea surface temperature sensors.

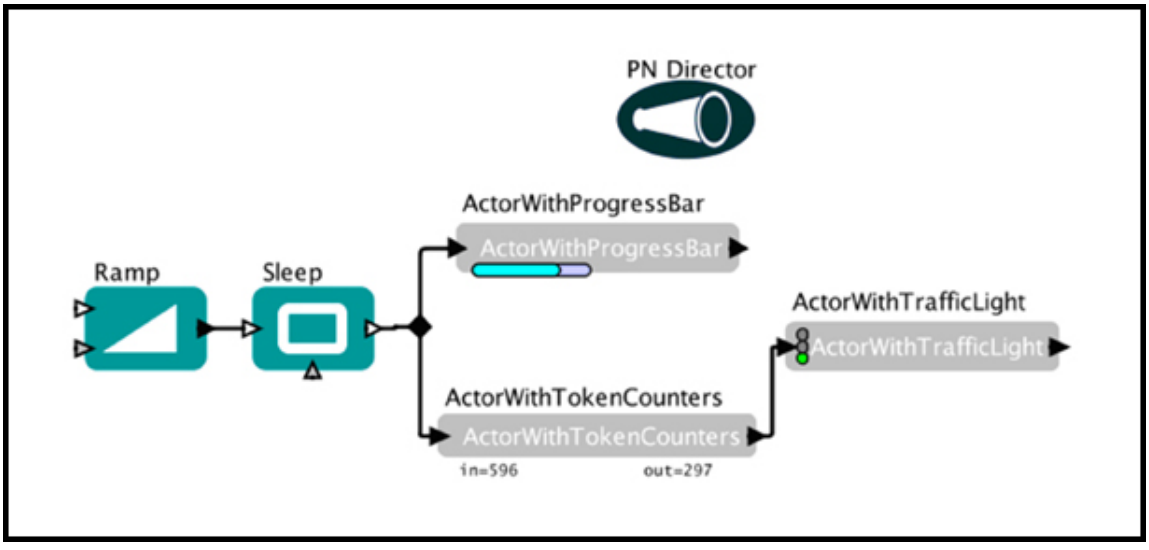
A cabled seafloor project, also under development, accesses, analyzes and plots data collected at the Kilo Nalu Nearshore Reef Observatory in Hawaii by a Workhorse Acoustic Doppler Current Profiler (ADCP) instrument. These data are sent to a DataTurbine server, and Kepler's forthcoming DataTurbine actor is used to grab requested data from the server, perform an analysis, and plot the results. The DataTurbine actor automatically assigns "nil" values to missing data points that arise when the sensors drop offline or other real-world problems occur so that time series are well-formed and can be more easily processed. Alternatively, the new Ensemble actor, which converts the ADCPs binary data format to numeric data, can be used to deal directly with the ADCP binary stream, bypassing conversions that occur at the DataTurbine server. REAP investigators are also looking into the feasibility and suitability of using a Sensor Web Enablement interface to aid in the discovery and accessibility of DataTurbine data.
(see REAP on p. 3)

REAP (CONTINUED FROM P. 2)



Example of a Kepler workflow that uses the DataTurbine actor to access and plot data from a REAP DataTurbine server.

Other developments emerging from the collaboration include a Kepler Execution Monitoring system that uses visual indicators to represent workflow execution progress: actors can have a progress bar, which indicates the amount of work remaining; a traffic light, which shows the actor's state (running, waiting for input data, idle, or error, for example); and/or a token counter, which displays the number of tokens read or written by an actor. A Kepler Provenance Framework that tracks and records the lineage of a workflow and associated data is also under development, as well as tools for System Engineers to monitor sensor networks and automatically alert users of troubles such as faulty sensors, low power, environmental conditions of concern, telemetry problems.



Kepler actors with traffic lights, progress bars, and token counters display monitoring information during workflow execution.

Release (continued from p. 1)

files can be emailed to colleagues, shared on web sites, or uploaded to the Kepler Component Repository.

Kepler provides a graphical user interface and a run-time engine that can execute workflows from within the graphical application interface or from a command line. In addition, workflows can be nested, allowing complex tasks to be composed of simpler components, and enabling workflow designers to build re-usable, modular components that can be saved and used for many different applications.

Kepler 1.0.0 ships with a searchable library containing over 350 ready-to-use processing components called actors that can be easily customized, connected and then run from a desktop environment to perform an analysis, automate data management, and integrate applications efficiently. In addition to generic mathematical, statistical, and signal processing components and components for data input, manipulation, and display, highlights include R and Matlab actors that easily integrate powerful statistical analyses into Kepler workflows; the WebService actor, which can access and execute WSDL-defined Web services and return execution results from within a workflow; a ReadTable actor, which can access legacy data stored in Excel files; and an ExternalExecution actor, which can execute any command line application from a workflow.

Direct access to scientific data is available via the EarthGrid, which can be searched from within the application. Using one of several actors specifically designed to ingest and output data, a wide variety of data sources can be accessed and used by workflows. Currently, Kepler has support for data described by Ecological Metadata Language (EML), data accessible using the DiGIR protocol, the OPeNDAP protocol, GridFTP, JDBC, SRB, and others.

Kepler 1.0.0 is the first stable and documented release of Kepler, and subsequent 1.X versions will be backward compatible. Kepler collaborators are committed to the continued development and maintenance of the Kepler system, and users are encouraged to contribute to the product by suggesting features that would be of use or by actively participating in development.

"We look forward to receiving feedback and buy in from the Kepler community and making Kepler even better in the future together," says Altintas. Work will continue under the leadership of the Kepler/CORE project, a team of researchers from UC Davis, UC Santa Barbara, and UC San Diego, who will continue to develop Kepler as a comprehensive, open, reliable, and extensible scientific workflow infrastructure.

The Kepler collaboration was originally founded in 2002 by researchers at the National Center for Ecological Analysis and Synthesis (NCEAS) at University of California Santa Barbara, the San Diego Supercomputer Center (SDSC) at University of California San Diego, and the University of California Davis as part of the Science Environment for Ecological Knowledge (SEEK) and Scientific Data Management (SDM) projects. It has since grown to include contributors from scores of research projects in many science disciplines, including ecology, biology, geosciences, physics, engineering, and chemistry, among others. The Kepler software extends the Ptolemy II system developed by researchers at the University of California Berkeley, which provides a mature platform for building and executing workflows, and supports multiple models of computation.

Kepler is freely available under the BSD License. To download the application or to learn more about Kepler, please see http://www.kepler-project.org.

UC DAVIS UNIVERSITY OF CALIFORNIA

UCSB

UCSD

NCEAS

Kepler/CORE and REAP are collaborative efforts of the University of California at Davis, Santa Barbara, and San Diego funded by the National Science Foundation under grant numbers 0722079 and 0619060, respectively.