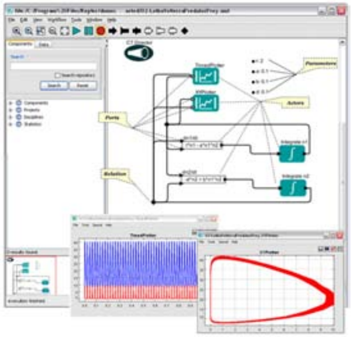
Kepler/CORE: A Comprehensive, Open, Reliable and Extensible Scientific Workflow Automation Framework

kepler-project.org

Kepler/CORE Project and Team

Background

- Kepler is a general purpose, multi-disciplinary, and open-source environment for modeling and executing scientific workflows
- Kepler grew out of an informal collaboration between researchers and engineers funded under the NSF/SEEK and DOE/SDM projects.
- Kepler builds upon the open-source Ptolemy II system originally developed as a modeling and design tool for the electrical engineering community.
- Originally developed as a single, integrated product, Kepler has been adopted and extended by a broad range of projects with a great diversity of distinct needs.
- Prior to Kepler/CORE no project had been funded specifically to coordinate development of Kepler.



Kepler/CORE Objectives

- Coordinate development of the Kepler base system and core feature set.
- Enhance Kepler with the features required for wide adoption and long-term sustainability.
- Gather, understand, and satisfy fundamental stakeholder and user requirements for Kepler.
- Implement new development infrastructure for supporting collaborative and independent development of Kepler.

Institutions and Personnel

University of California, Davis
Bertram Ludaescher (PI), Shawn Bowers (co-PI), Timothy McPhillips (co-PI), David Welker (software engineer), Sean Riddle (software engineer)

University of California, Santa Barbara
Matthew Jones (PI), Mark Schildhauer (co-PI), Aaron Schultz (software engineer), Chad Berkley (software engineer)

University of California, San Diego
Ilkay Altintas (PI), Jianwu Wang (postdoc)

Challenges & Strategies

The Challenge of Diversity

- **Enormous diversity of domains**

  Astrophysics, nuclear fusion research, geoinformatics, ecology, systematics, genomics, bioinformatics, data mining, environmental monitoring, dynamic systems modeling, simulation, …

- **Broad range of technical problems addressed via Kepler**
  - Facilitating ***workflow design***
  - ***Sharing*** actors, workflows, and system extensions between users and across projects
  - ***Distributing execution*** across heterogeneous resources
  - ***Moving data*** between computers over a broad spectrum of protocols
  - ***Integrating*** local applications, web services, and native actors within a single workflow
  - Supporting a variety of ***computational models***
- **Users with different backgrounds and responsibilities**
  - ***Scientists*** automating and sharing their analyses
  - ***Software engineers*** developing systems around Kepler
  - ***Computer scientists*** doing basic research in scientific workflows
- **Kepler applied in many different deployment contexts**
  - ***Desktop application*** for modeling, running, and monitoring workflows interactively
  - Non-interactive ***back end*** for web-based applications
  - ***Embedded workflow engine*** for other applications
  - Actor and workflow ***development platform***
  - Building and running workflows via **web browsers**.

Kepler/CORE is founded on the premise that no single product called 'Kepler' can possibly solve everyone's problems right out of the box.

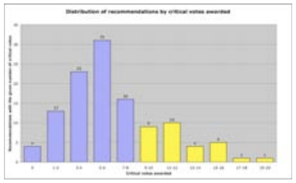
Kepler
Your Science. Enabled.
Developer Forum
A forum for the various contributors to Kepler teams, interest groups, and task groups to discuss issues in Kepler design and development.
Framework Team
Build and Release Team
User Representation Team

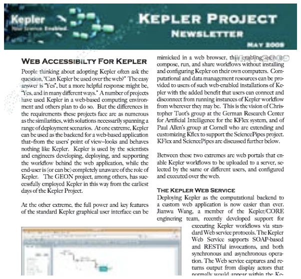
Kepler/CORE Strategy

- Clearly define a ***kernel*** of capabilities applicable to all projects.
- Identify and ***develop new*** critical ***core features.***
- Facilitate the application of Kepler to diverse scientific domains and deployment contexts by providing ***well-defined extension points***.
- Ensure core system ***stability*** by rigorous software ***testing***.
- Deliver and support ***regular software releases*** of the core system and base system modules.
- ***Train*** future system end-users, workflow engineers, and Kepler extension developers.
- ***Disseminate*** documentation and training materials to the broader scientific community.
- Evaluate approaches for ***sustaining*** development and maintenance of the common core of all Kepler-based systems.

Kepler Stakeholders

- Are projects and individuals whose work ***depends critically on the effectiveness of Kepler***.
- Likely to greatly ***extend Kepler*** and use Kepler within their own systems.
- Need to deliver the software systems they develop to ***their own community of users***.
- Must deliver their software systems according to ***their own release schedules*** as determined by their research and funding programs.
- Have different requirements that will conflict in the absence of mechanisms for enabling ***independent extension and deployment*** of Kepler-based systems.
- ***Require recognition*** for the contributions they make to Kepler as well as for their own systems based on Kepler.
- ***Know*** better than us what they need from Kepler.





KEPLER PROJECT NEWSLETTER

MAY 2009

WEB ACCESSIBILTY FOR KEPLER

People thinking about adopting Kepler often ask the question, "Can Kepler be used over the web?" The easy answer is "Yes", but a more helpful response might be, "Yes, and in many different ways." A number of projects have used Kepler in a web-based computing environment and others plan to do so. But the differences in the requirements these projects face are as numerous as the similarities, with solutions necessarily spanning a range of deployment scenarios. At one extreme, Kepler can be used as the backend for a web-based application that--from the users' point of view--looks and behaves nothing like Kepler. Kepler is used by the scientists and engineers developing, deploying, and supporting the workflow behind the web application, while the end-user is (or can be) completely unaware of the role of Kepler. The GEON project, among others, has successfully employed Kepler in this way from the earliest days of the Kepler Project.

At the other extreme, the full power and key features of the standard Kepler graphical user interface can be mimicked in a web browser, thus enabling users to compose, run, and share workflows without installing and configuring Kepler on their own computers. Computational and data management resources can be provided to users of such web-enabled installations of Kepler with the added benefit that users can connect and disconnect from running instances of Kepler workflow from wherever they may be. This is the vision of Christopher Tuot's group at the German Research Center for Artificial Intelligence for the KFlex system, and of Paul Allen's group at Cornell who are extending and customizing Kflex to support the SciencePipes project. KFlex and SciencePipes are discussed further below.

Between these two extremes are web portals that enable Kepler workflows to be uploaded to a server, selected by the same or different users, and configured and executed over the web.

THE KEPLER WEB SERVICE

Deploying Kepler as the computational backend to a custom web application is now easier than ever. Jianwu Wang, a member of the Kepler/CORE enginering team, recently developed support for executing Kepler workflows via standard Web service protocols. The Kepler Web Service supports SOAP-based and RESTful invocations, and both synchronous and asynchronous operation. The Web service captures and returns output from display actors that normally would appear within the Ke-

Engaging Users & Stakeholders

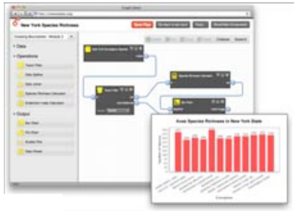
- Built a comprehensive model of the ***problem-solution landscape*** for Kepler, modeling and relating discussions, feature requests, and design suggestions.
- Held a ***stakeholders meeting*** using the initial requirements model as a guide for the agenda.
- ***Analyzed recommendations*** made by stakeholders and prioritized those with strongest support.
- Hosted ***Kepler developers meetings*** including representatives of stakeholder projects.
- Reported to stakeholders periodically through ***project newsletters***.

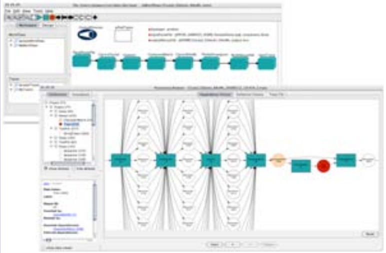
Featured Stakeholder Projects*

Science Pipes

- Environment in which students, educators, citizens, resource managers, and scientists can ***create and share analyses and visualizations*** of biodiversity data.
- Provides a ***simple web-based interface*** to Kepler.
- Allows analysis results and visualizations to be dynamically incorporated into web sites.

Steve Kelling, Rick Bonney, and Paul Allen. See SciencePipes.org or *contact* paul.e.allen@cornell.edu *for more information. NSF DUE-0734857.*


New York Species Richness
Area Species Richness in New York State



pPOD/COMAD

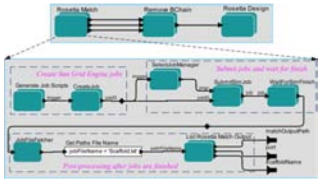
- Collection-oriented modeling and design (COMAD) computational model simplifies workflow design and management of nested data collections.
- Applied to phylogenetics in *pPOD* for AToL researchers, and to genomics applications in *Chip2*.
- Interactive provenance browser *(left)* enables researchers to explore data dependencies of workflow products.

NSF IIS-0612326 (Chip2), IIS-0630033 (pPOD)
Contact Bertram Ludaescher (ludaesch@ucdavis.edu)

Kepler in UCGrid

- Capabilities for resource consolidation, parallelism, provenance tracking, and fault tolerance.
- Computational processes automated, pipelined, efficient, extensible, stable, and easy to use.
- Applied to enzyme design processes *(right)* among others.

Jianwu Wang, Ilkay Altintas (SDSC) ; Prakashan Korambath, Seonah Kim, Scott Johnson (UCLA)
www.ucgrid.org

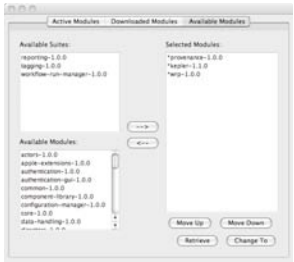


Open Architecture, Open Project, Open Source

- Kepler now has an ***extremely open architecture***.
- ***Restructured code repository*** to manage independently developed modules contributed by different projects operating on different release schedules.
- Revised ***custom build system*** to support building ***sets of modules*** as determined by developers.
- Implemented a ***run-time module manager*** to enable users to download and upgrade Kepler modules in the field.
- New ***web site***, new ***developer forums***, and new project organization support both top-down and grass-roots development teams.
- Kepler system is 100% ***open source*** with a liberal license; BSD license used for code managed in Kepler repository.
- Third-party libraries with conflicting licenses were exhaustively identified and either removed or renegotiated at time of Kepler 1.0 release.

New Development Infrastructure for Extension Developers

- Structured ***source code repository*** to support ***multiple modules***, both for the core system modules and for modules contributed by stakeholders.
- ***Revised build system*** to support developers working with ***different sets of modules***.
- Included features that accelerate development and enable new engineers to set up their development environments quickly.
- ***Addressed*** need for supporting ***conflicting*** feature and ***third-party library requirements***.



- ***Run-time module manager*** for browsing, downloading, and installing add-on modules to Kepler at run-time is nearly complete.
- Build system and module manager will be used to ***roll out patches, upgrades,*** and replacements to existing modules.
- Kepler 2.0 release will enable stakeholder projects to ***distribute specialized code*** running on a common base system used by all Kepler users.

- Redesigned the nightly build system to utilize the ***NMI Build and Test framework***.
- Result was a new level of automation and nightly builds on a greater variety of different build permutations.
- Current NMI ***tests run on five operating*** systems (Ubuntu 5.10, Red Hat Enterprise Linux 4, Fedora Core 5, Mac OS 10.4, and Windows XP) ***and three versions of Java*** (JDK 1.4, 1.5, 1.6).
- The build system is configured to build installers for each of these platforms, with each nightly build producing snapshot installers that represent the current state of the code and greatly ease testing for non-developers.
- The build system also ***supports NMI tests*** to be run ***over distinct suites of modules***, not just the core set of base system modules.

Ilkay Altintas[3], Shawn Bowers[2], Matt Jones[4], Bertram Ludäscher[1], Timothy McPhillips[1], and Mark Schildhauer[4]

[1]UC Davis Genome Center & Department of Computer Science, UC Davis; [2]Department of Computer Science, Gonzaga University;
[3]San Diego Supercomputer Center (SDSC), UC San Diego; [4]National Center for Ecological Analysis and Synthesis (NCEAS), UC Santa Barbara

NSF award
OCI-0722079

**See also* kepler-project.org/users/projects-using-kepler