Scientific Workflow Tools

Daniel Crawl and Ilkay Altintas

San Diego Supercomputer Center

UC San Diego

eScience Today

- Increasing number of Cyberinfrastructure (CI) technologies
  - Data Repositories: Network File Systems, Databases, Web Services, SRB/iRODS
  - Job Execution: Cloud Computing, Grid, Cluster, Ad-hoc
  - Domain-specific analysis tools
- Difficult to orchestrate CI components to conduct eScience

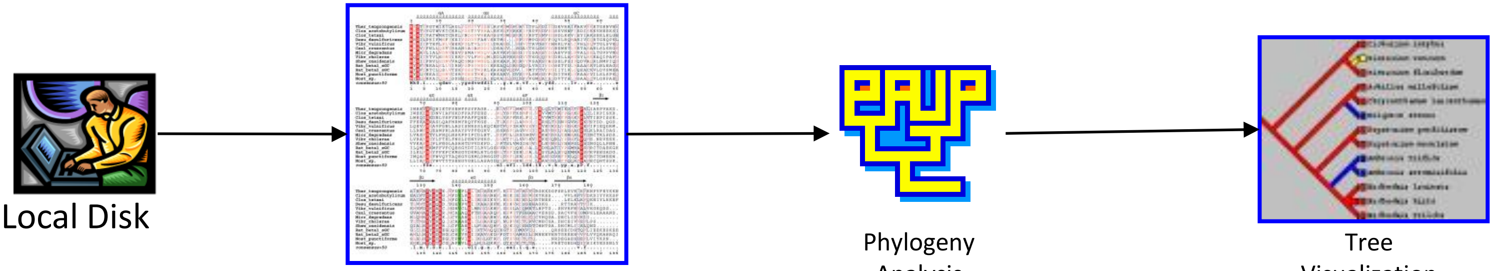
Scientific workflows emerged as an answer to the need to combine multiple Cyberinfrastructure components in automated process networks.

So, what is a scientific workflow?

The Big Picture: Supporting the Scientist

From *“Napkin Drawings”* to *Executable Workflows*

Phylogeny Analysis Workflow

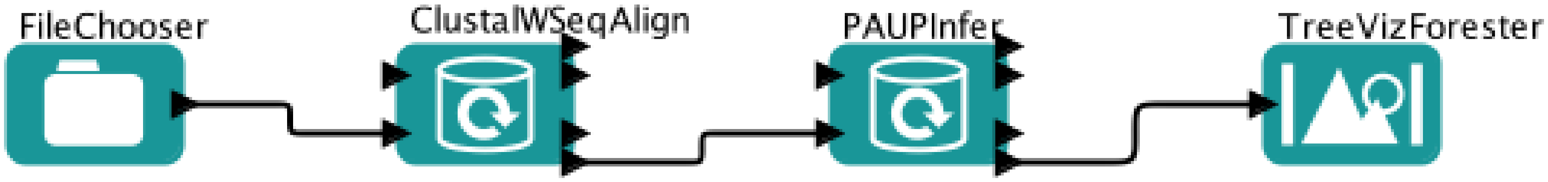


Local Disk

Multiple Sequence Alignment

Phylogeny Analysis

Tree Visualization



1. Read the Nexus file as input.

2. Do the multiple sequence alignment on the input data matrix using ClustalW;

3. Construct the phylogenetic tree using PAUP;

4. Read the tree and display it to the user using Forester.

Advantages of Scientific Workflow Systems

- Formalization of the scientific process
- Easy to share, adapt and reuse
  - Deployable, customizable, extensible
- Management of complexity and usability
  - Support for hierarchical composition
  - Interfaces to different technologies from a unified interface
  - Can be annotated with domain-knowledge
- Tracking provenance of the data and processes
  - Keep the association of results to processes
  - Make it easier to validate/regenerate results and processes
  - Enable comparison between different workflow versions
- Execution monitoring and fault tolerance
- Interaction with multiple tools and resources at once

Kepler Scientific Workflow System

http://www.kepler-project.org

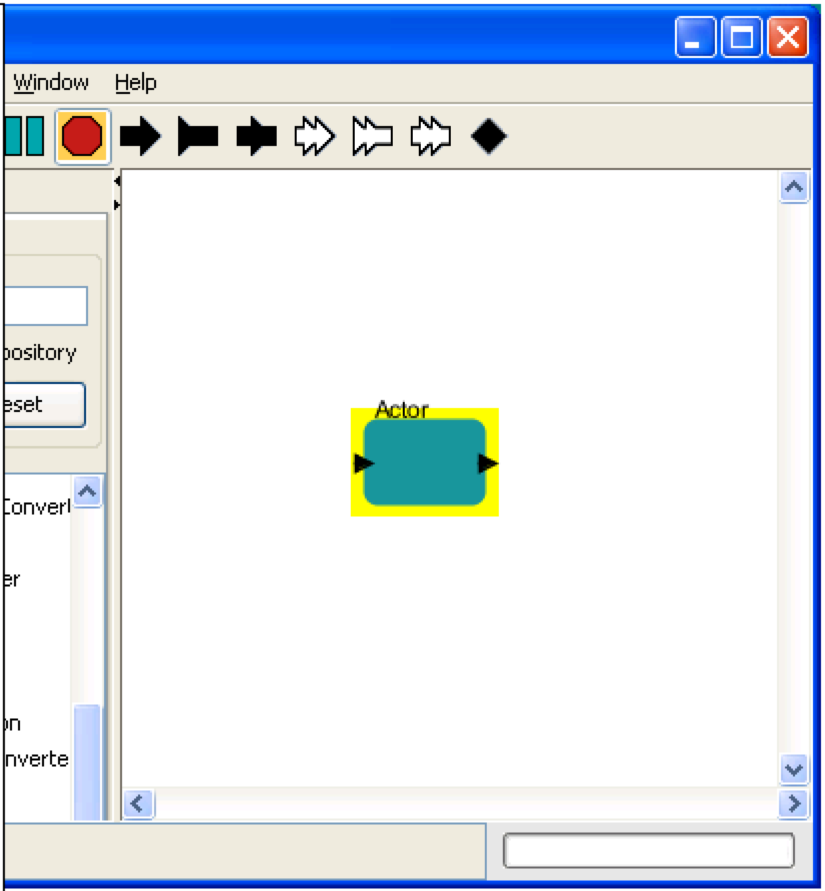
- Kepler is a cross-project collaboration
- Latest release available from the website
  - Kepler 2.1 released on 30 September 2010
- Builds upon the open-source Ptolemy II framework
- Vergil is the GUI, but Kepler also runs in non-GUI and batch modes.

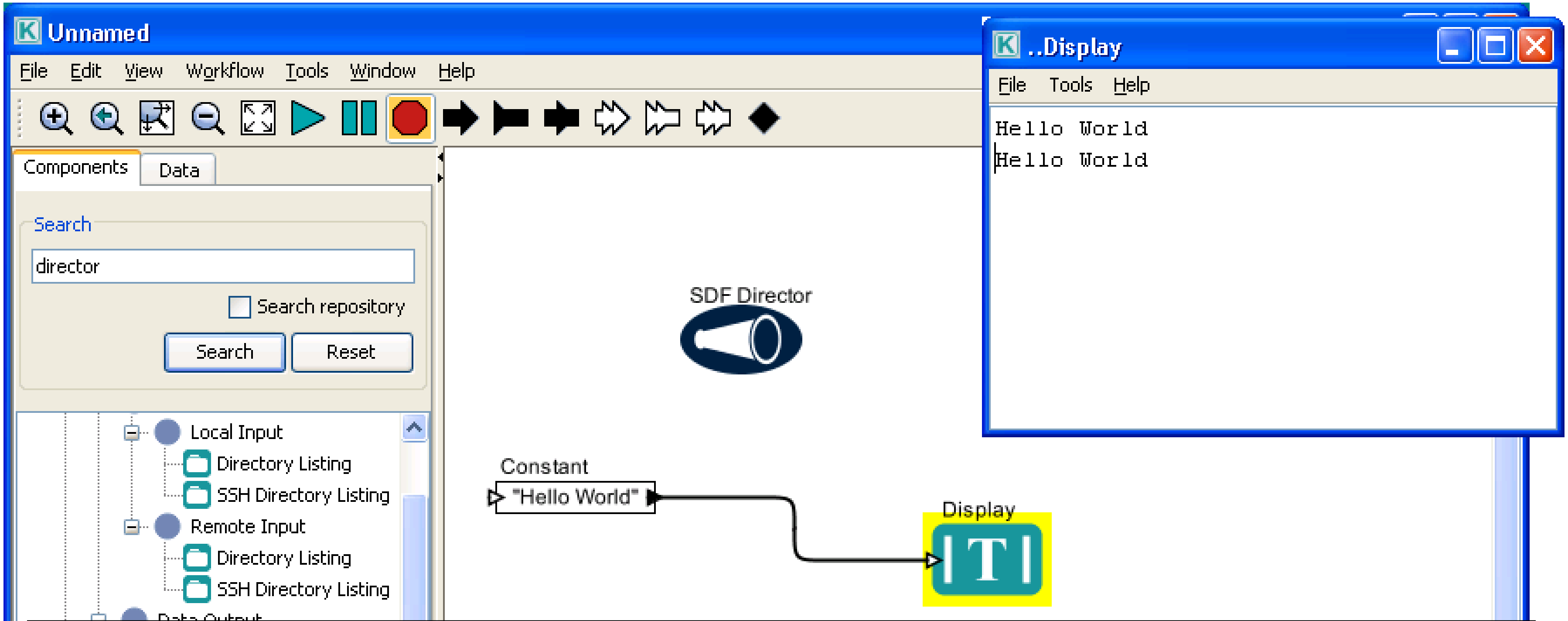
Ptolemy II: A laboratory for investigating design
KEPLER: A problem-solving support environment for Scientific Workflow development, execution, maintenance

KEPLER = "Ptolemy II + X" for **Scientific Workflows**

Actor-Oriented Modeling

- Actors
  - Single component or task
  - Well-defined interface (signature)
  - Given input data, produces output data
  - Configured with parameters
  - “Composite” actor for sub-workflows
- Ports
  - Each actor has a set of input and output ports
  - Denote the actor’s signature
  - Produce/consume data (a.k.a. tokens)
  - Can be semantically annotated with domain-specific concepts





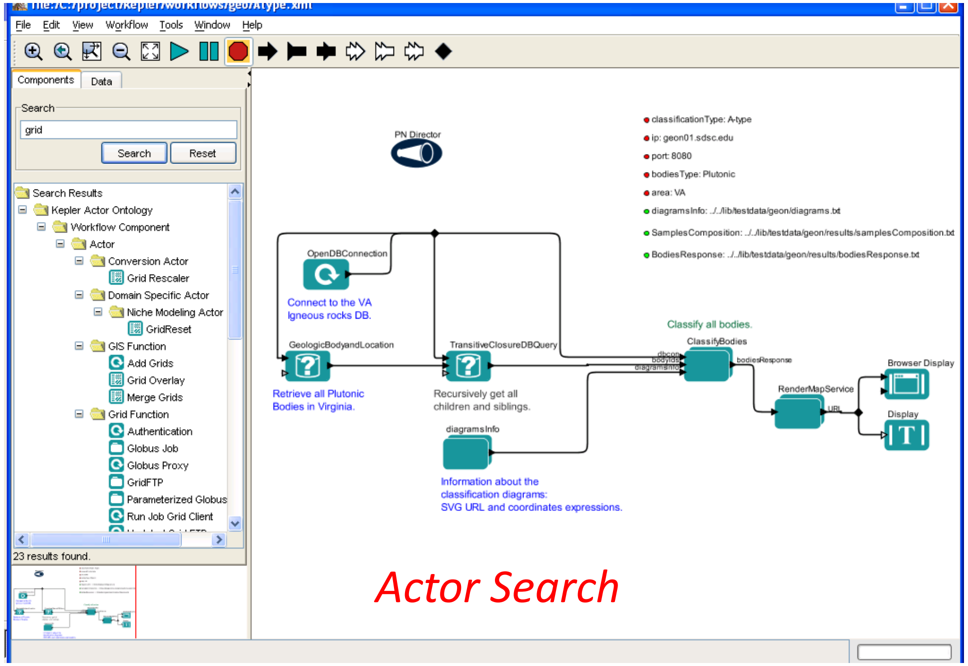
```
Hello World
Hello World
```

- Dataflow Connections
  - Actor "communication" channels
  - Directed edges
  - Connect output ports with input ports
- Directors
  - Execution models, define the execution semantics of workflow graphs
  - Executes workflow graph (some schedule)
  - Sub-workflows may have different directors

Kepler Actors

- Generic Web Service Clients- SOAP, REST, MS .Net
- Customizable RDBMS query and update
- Command Line wrapper tools (local, ssh, scp, ftp, etc.)
- Grid actors: Globus, GridFTP, Proxy Certificate Generator
- SRB and iRODS
- R and Matlab
- Interaction with MapReduce
- Communication with streaming data buffers- DataTurbine, ORB
- Imaging, Gridding, Viz Support
- Textual and Graphical Output
- Specialized actor for fault tolerance
- …additional generic and domain-oriented actors…

Vergil is the GUI for Kepler



23 results found.

- classificationType: A-type
- ip: geon01.sdsc.edu
- port: 8080
- bodiesType: Plutonic
- area: VA
- diagramsInfo: ../../lib/testdata/geon/diagrams.txt
- SamplesComposition: ../../lib/testdata/geon/results/samplesComposition.txt
- BodiesResponse: ../../lib/testdata/geon/results/bodiesResponse.txt

Connect to the VA Igneous rocks DB.

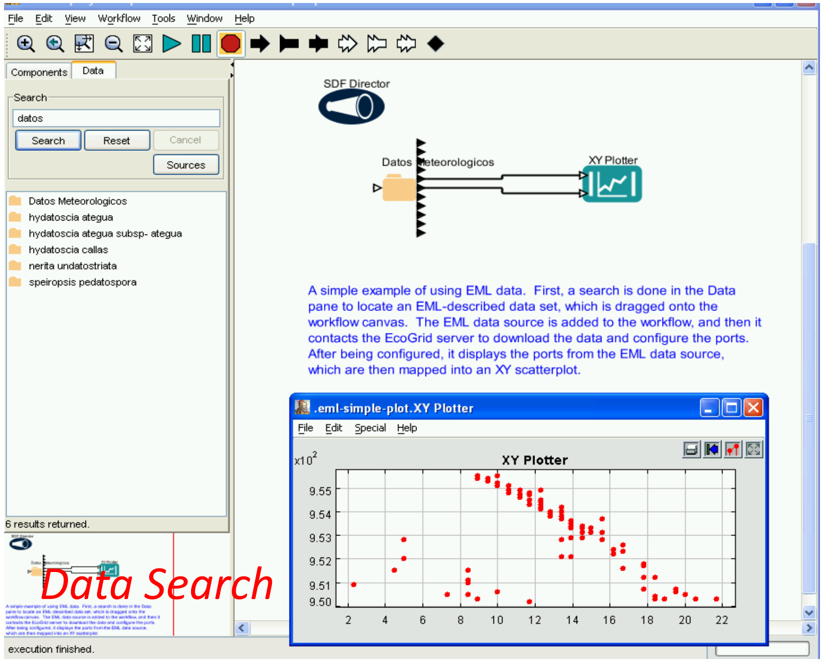
Retrieve all Plutonic Bodies in Virginia.

Recursively get all children and siblings.

Classify all bodies.

Information about the classification diagrams: SVG URL and coordinates expressions.

Actor Search



6 results returned.

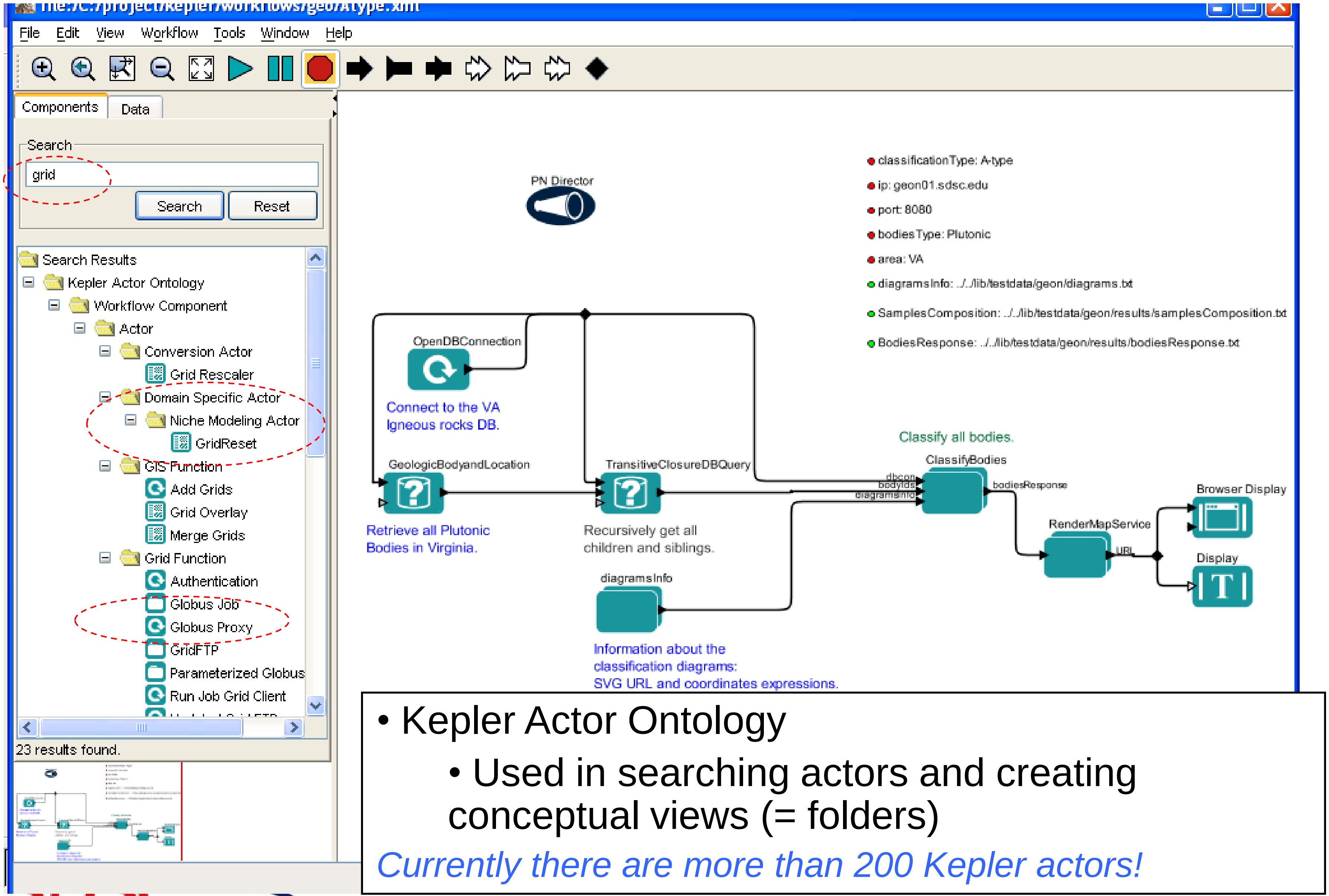
A simple example of using EML data. First, a search is done in the Data pane to locate an EML-described data set, which is dragged onto the workflow canvas. The EML data source is added to the workflow, and then it contacts the EcoGrid server to download the data and configure the ports. After being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

execution finished.

Data Search

- Actor ontology and semantic search for actors
- Search -> Drag and drop -> Link via ports
- Metadata-based search for datasets

Actor Search



- classificationType: A-type
- ip: geon01.sdsc.edu
- port: 8080
- bodiesType: Plutonic
- area: VA
- diagramsInfo: ../../lib/testdata/geon/diagrams.txt
- SamplesComposition: ../../lib/testdata/geon/results/samplesComposition.txt
- BodiesResponse: ../../lib/testdata/geon/results/bodiesResponse.txt

Connect to the VA Igneous rocks DB.

Retrieve all Plutonic Bodies in Virginia.

Recursively get all children and siblings.

Classify all bodies.

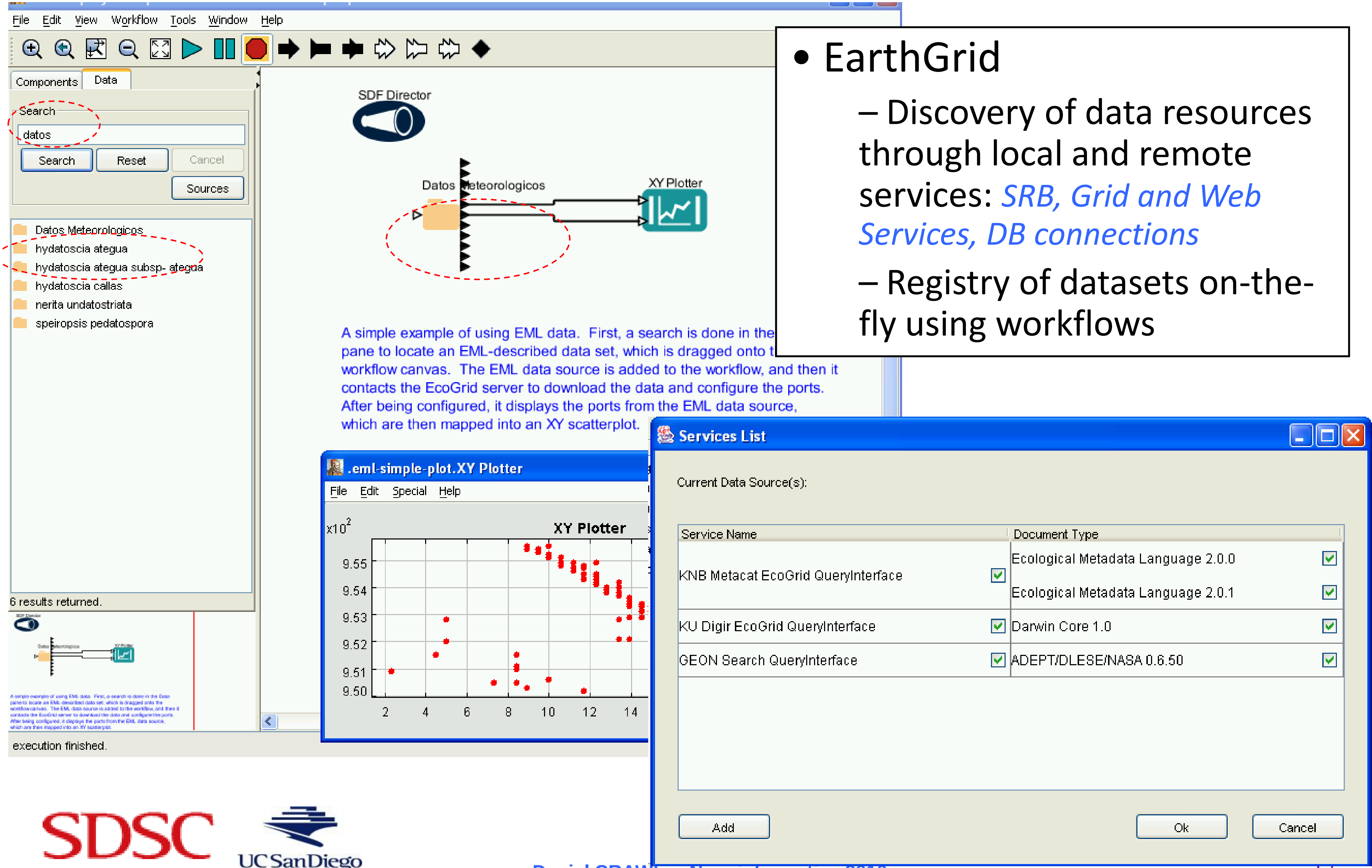
Information about the classification diagrams: SVG URL and coordinates expressions.

- Kepler Actor Ontology
  - Used in searching actors and creating conceptual views (= folders)

Currently there are more than 200 Kepler actors!

Data Search and Usage of Results

- EarthGrid
  - Discovery of data resources through local and remote services: *SRB, Grid and Web Services, DB connections*
  - Registry of datasets on-the-fly using workflows



A simple example of using EML data. First, a search is done in the pane to locate an EML-described data set, which is dragged onto t 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.

Current Data Source(s):

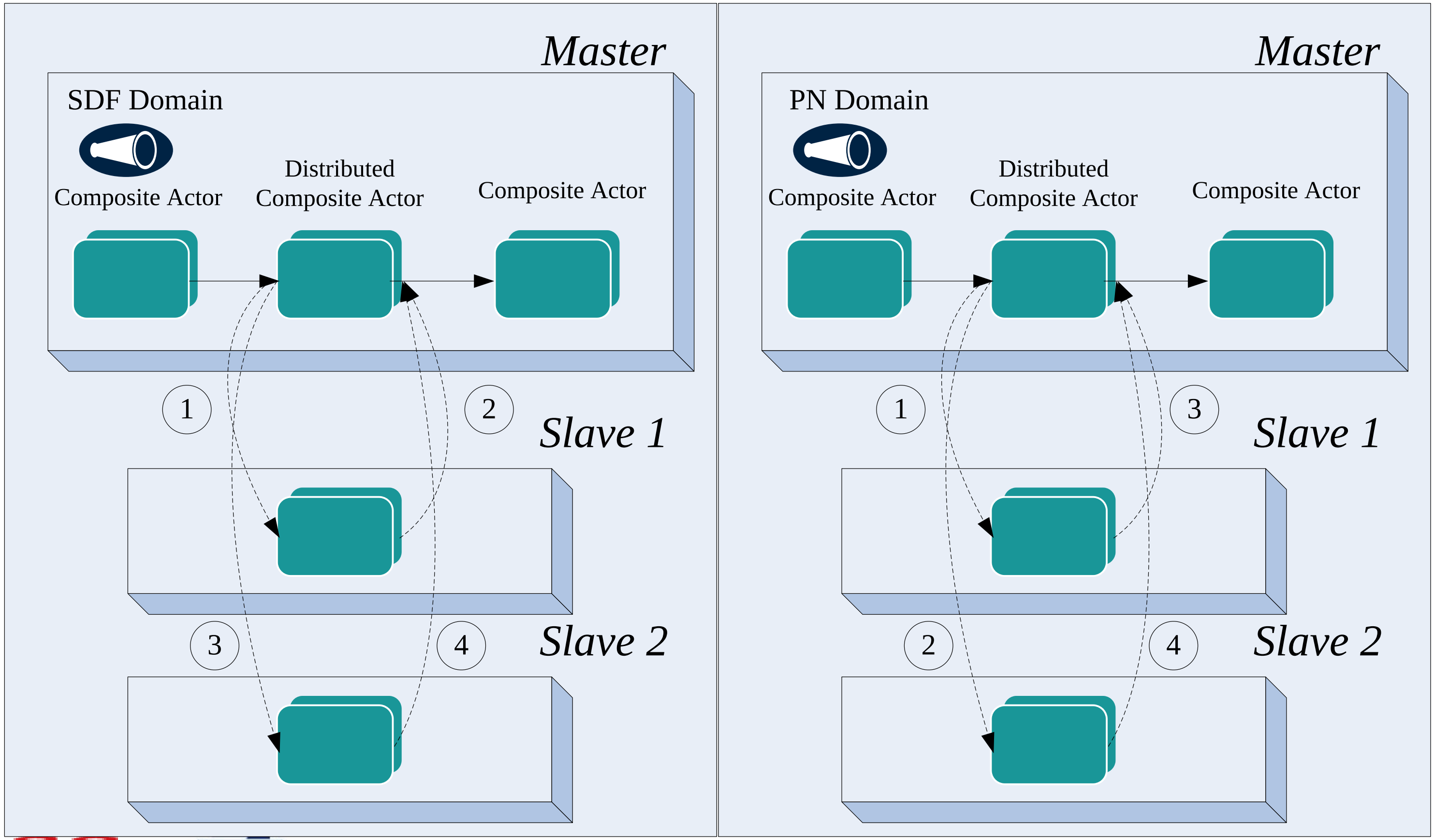
Service Name		Document Type	
KNB Metacat EcoGrid QueryInterface	☑	Ecological Metadata Language 2.0.0	☑
		Ecological Metadata Language 2.0.1	☑
KU Digir EcoGrid QueryInterface	☑	Darwin Core 1.0	☑
GEON Search QueryInterface	☑	ADEPT/DLESE/NASA 0.6.50	☑

Distributed Execution

- Master-Slave
  - Execute sub-workflows on slave nodes
- Map-Reduce
  - Map and reduce sub-workflows executed in Hadoop cloud
- Job actors for PBS, LSF, SGE, Globus, etc.
- Kepler web service

Master-Slave Framework

- Single workflow created, sub-workflows seamlessly run on other resources
- Data is automatically distributed to Slave nodes and results returned
- Different behavior with different computation models



Provenance of Workflow Related Data

- Provenance: A concept from art history and library
  - Inputs, outputs, intermediate results, workflow design, workflow run
- Collected information
  - Can be used in a number of ways
    - Validation, reproducibility, fault tolerance, etc…
  - Linked to the data resources
  - Viewable and searchable from outside Kepler

Kepler Provenance Framework

- What provenance is recorded:
  - Workflow Specification: actors, ports, connections, parameters, etc.
  - Workflow Evolution: parameter values that change over time, addition/removal of actors, ports, etc.
  - Workflow Execution:
    - Start/stop of workflow, individual actor executions
    - Data exchanged between actors: data lineage
- Where provenance recorded:
  - Modular interface supports saving to different output types.
  - SQL, XML, or Open Provenance Model

Kepler is a Team Effort

Some CI projects using Kepler:

- SEEK (ecology)
- SciDAC SDM (astrophysics, bio, ...)
- CPES (plasma simulation, combustion)
- GEON (geosciences)
- CiPRes (phylogenetics)
- ROADnet (real-time data)
- Processing Phylodata (pPOD)
- REAP (streaming data)
- Digital preservation (DIGARCH)
- COMET (environmental science)
- ITER (fusion)

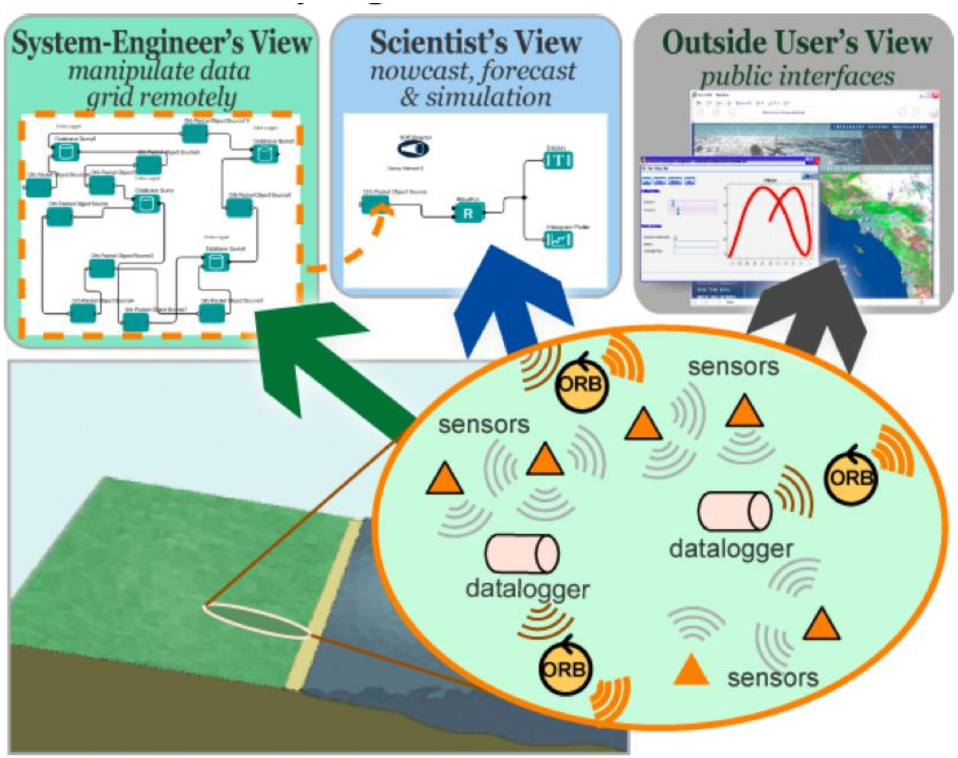
- OOI CI - ORION (ocean observing CI)
- LOOKING (oceanography)
- CAMERA (metagenomics)
- Resurgence (computational chemistry)
- ChIP-chip (genomics)
- Cheshire Digital Library (archival)
- Cell Biology (Scripps)
- DART (X-Ray crystallography)
- Ocean Life
- Assembling the Tree of Life project
- NEES (earthquake engineering)
- ...

Real-time Environment for Analytical Processing (REAP)

- Management and Analysis of Observatory Data using Kepler Scientific Workflows
- Overall goal: To bring together, *for the first time*, seamless access to sensor data from real- time data grids with analytical tools and sophisticated modeling capabilities of scientific workflow environments

- Funded 2006-2010
  - NSF CEO:P
    - Jones, Altintas, Baru, Ludaescher, Schildhauer
  - Partners:
    - UCSB, SDSC/UCSD, UCDavis, UCLA, OpenDAP, OSU
    - Lead institution: NCEAS/UCSB

http://reap.ecoinformatics.org/



Sea Surface Temperature (SST) Match-up Workflows

- Quantitative evaluation and integration of SST data sets
  - Allows researchers to find data sets for a given space-time window
  - Builds match-up data sets from various sources, e.g., NOAA, JPL, FSU, using OPeNDAP
  - Performs a variety of statistical comparisons and visualizations on match-ups using R and Matlab
- Collaborators:
  - Peter Cornillon, Univ. of Rhode Island
  - Nathan Potter, James Gallagher, OPeNDAP Inc.

Summary

- Scientific workflows help scientists manage diverse CI technologies
- Kepler is an open-source system and collaboration
  - Grows by application requirements from contributors
  - More information: http://kepler-project.org

- Acknowledgements:
  - NSF award 0619060 for Real-time Environment for Analytical Processing
  - NSF award 0941692 for Distributed Ocean Monitoring via Integrated Data Analysis of Coordinated Buoyancy Drogues
  - DOE award DE-FC02-01ER25486 for SDM Center

Thanks!
&
Questions…

Daniel Crawl
crawl@sdsc.edu

Ilkay Altintas
altintas@sdsc.edu