

# Knowledge Annotations in Scientific Workflows: An Implementation in Kepler

Aída Gándara<sup>1</sup>, George Chin<sup>2</sup>, Paulo Pinheiro da Silva<sup>1</sup>,  
Terence Critchlow<sup>2</sup>, Chandrika Sivaramakrishnan<sup>2</sup>,  
Signe White<sup>2</sup>

<sup>1</sup>The University of Texas at El Paso  
<sup>2</sup>Pacific Northwest National Laboratory

# PNNL-UTEP Research

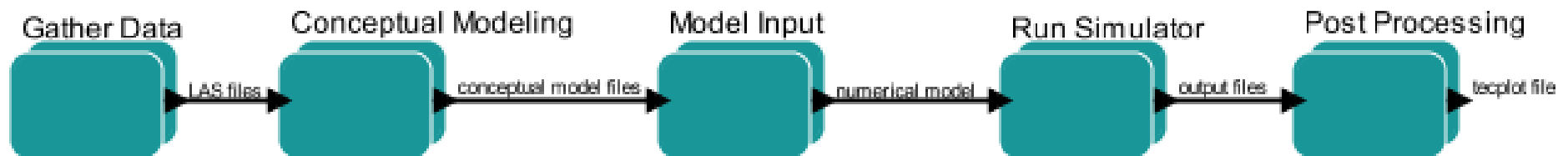
- Collaborative Team:
  - SciDAC Scientific Data Management Center at Pacific Northwest National Laboratory
  - Cyber-ShARE Research Center at UTEP
  - August – October 2010
- Collaboration Purpose
  - To help groundwater scientists at PNNL manage collaborative data that is traditionally generated during a research effort but not preserved as part of the research effort

# Research Goals

- **Generic Goals**
  - Understand collaborative research processes before developing a workflow for it
  - Understand needs for documenting research collaboration
- **Specific Goal**
  - Use the Kepler Scientific Workflow System as a way of understanding a research process at PNNL

# Case Study

- Subsurface Flow and Transport Analysis
  - Typically members include: project manager and several team members.
  - Each step requires expertise, e.g., groundwater scientists use STOMP and other software
  - Collaboration between steps



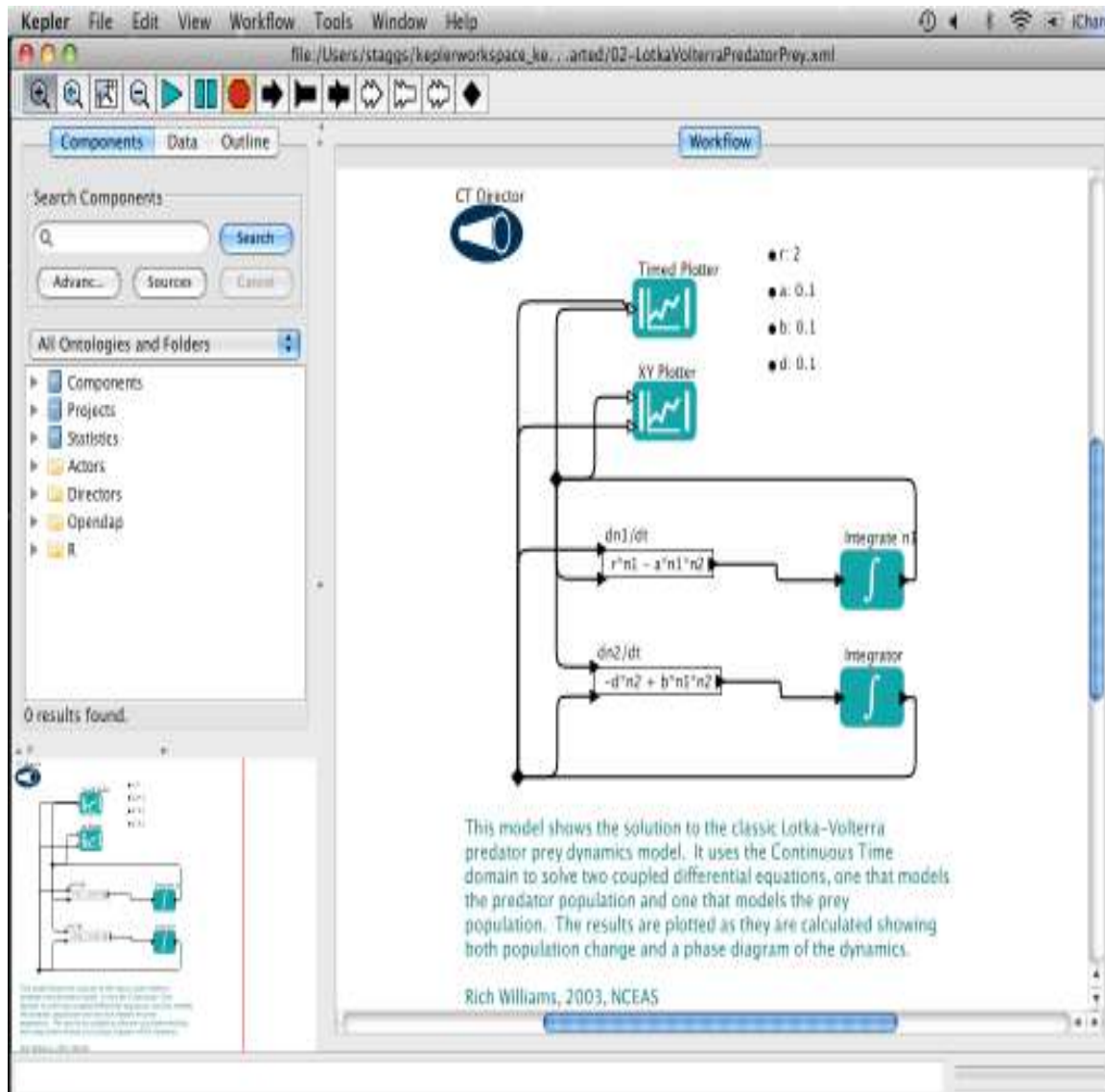
# Some Observations

- At some point, scientists seek to understand the “hows” and “whys” of scientific results
- Scientists keep journals and notes of what worked and what did not, e.g., decisions, assumptions and constraints
- Much of this information is needed for final reports and publications

**Scientists often need to capture their notes about ad hoc processes, not processes predefined in a workflow**

# Kepler Scientific Workflow System

- Collect sufficient information to document a scientific process
- Support reproducing results
- Help collect provenance



From Kepler getting started guide, the Lotka-Volterra Workflow

# Knowledge-Annotated Scientific Workflows :design principles

1. Scientists describe their research: build workflow from information
2. Align with scientific research process: reduce duplication and alteration of process
3. Leverage workflow to manage annotations: annotations relate to actors and connections in workflow

# Knowledge-Annotated Kepler Workflow System

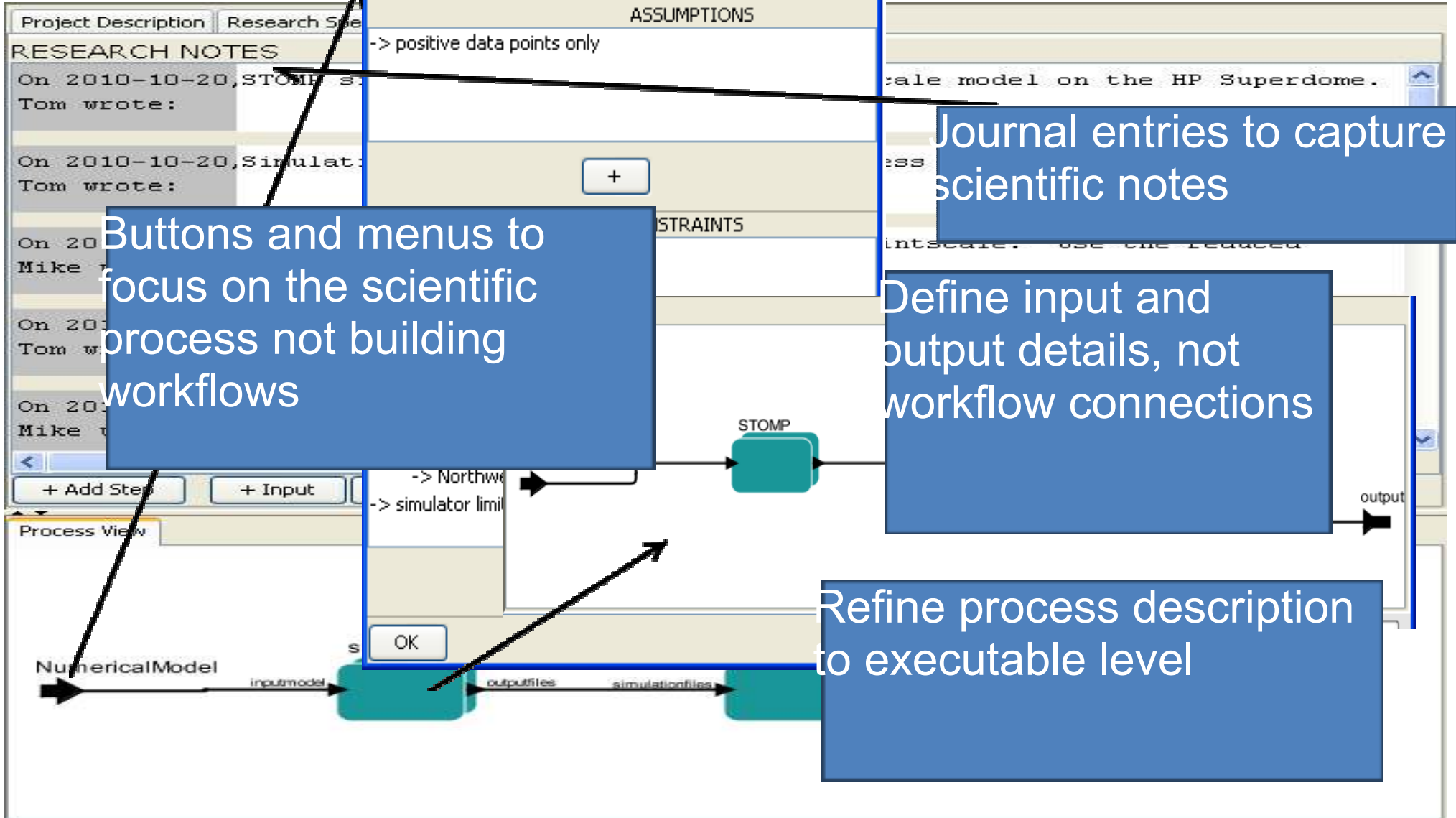
The screenshot displays the Kepler Workflow System interface. The window title is "file:/C:/Documents%20and%20Settings/C...eplerData/RGFT10Oct10/RGFTWorkflow.xml". The menu bar includes File, Edit, View, Workflow, Tools, Window, Project, and Help. The toolbar contains various icons for navigation and execution. The interface is divided into several panes:

- Research Hierarchy:** A tree view on the left showing a hierarchy of tasks: GatherData, ConceptualModeling, ModelInput, and RunSimulator. A blue box labeled "Research Hierarchy" has an arrow pointing to this pane.
- Project Description:** A central pane with tabs for "Project Description" and "Research Specs". It contains a "Title" field with the text "Regional Groundwater Flow and Transport Investigation", a "Description" field with a paragraph of text, and a "Team Members" section listing Tom, Jon, Michael, and Ann. A blue box labeled "Tabs to capture Scientists' knowledge" has an arrow pointing to the "Project Description" tab.
- Process View:** A bottom pane showing a workflow diagram with four main tasks: GatherData, ConceptualModeling, ModelInput, and RunSimulator. Data flows are indicated by arrows: GatherData outputs "RawOutputFiles" to ConceptualModeling; ConceptualModeling outputs "ConceptualModel" to ModelInput; ModelInput outputs "Out Model" to RunSimulator; RunSimulator outputs "Image" and "simulationreport". A blue box labeled "Process View" has an arrow pointing to this pane.



Keep

Workflows



# Results

- Scientists do not add workflow components
  - steps, journal entries, inputs/outputs, assumptions, constraints, comments ...
- Various views of the data:
  - Research summary report
  - Process traversal (forward and back through inputs/outputs)
  - Status of a step
  - RDF output that links to workflow information in SIOC

# Current Status

- kadm.jar with features identified in research
- Embedded in UTEP CyberShARE tools for use by environmental scientists and geoscientists
- Building RDF specific to research teams with annotations, workflows and data
- Evaluation of process and data

# Conclusions

- Workflows not always intuitive
- Some scientists feel workflows are too rigid
- This research has presented an alternative method for scientists to create and annotate an ad hoc scientific workflow

# Contact

Aída Gándara

The University of Texas at El Paso

[agandara1@miners.utep.edu](mailto:agandara1@miners.utep.edu)

George Chin

Pacific Northwest National Laboratory

[George.Chin@pnnl.gov](mailto:George.Chin@pnnl.gov)