Knowledge Annotations in Scientific Workflows: An Implementation in Kepler

Aída Gándara[1], George Chin[2],
Paulo Pinheiro da Silva[1], Terence Critchlow[2],
Chandrika Sivaramakrishnan[2], Signe White[2]

[1]The University of Texas at El Paso
[2]Pacific Northwest National Laboratory

CYBER-ShARE Center of Excellence
Sharing Resources to Advance Research and Education through Cyberinfrastructure
THE UNIVERSITY OF TEXAS AT EL PASO
UTEP

SciDAC
Scientific Discovery through Advanced Computing

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 after the effort is completed

Research Goals

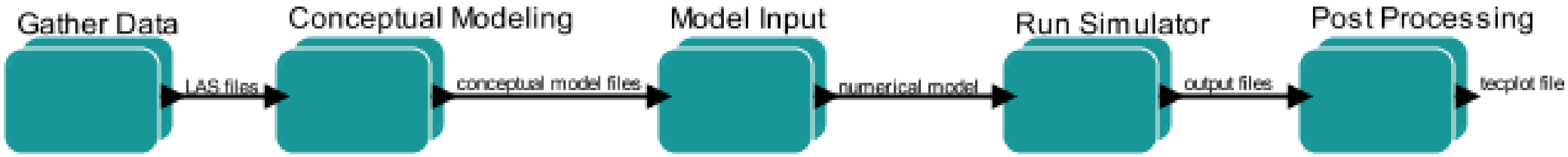
- Generic Goals
  - Understand collaborative research processes before developing a workflow for it
  - Understand needs for documenting research collaboration

- Specific Goal
  - Use the Kepler Scientific Workflow System as a way of understanding a research process at PNNL

CYBER-ShARE Center of Excellence
Sharing Resources to Advance Research and Education through Cyberinfrastructure
THE UNIVERSITY OF TEXAS AT EL PASO
UTEP
SciDAC
Scientific Discovery through Advanced Computing

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



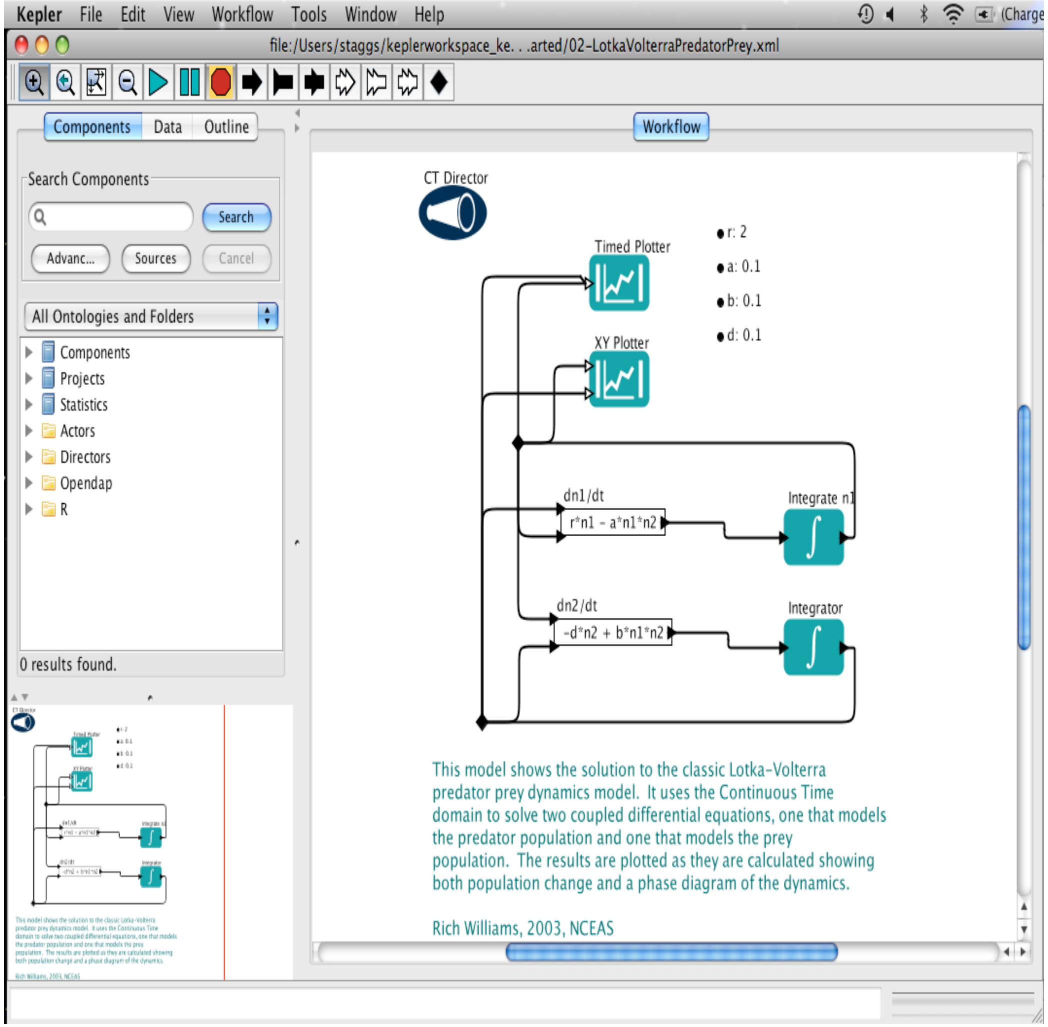
CYBER-ShARE Center of Excellence
Sharing Resources to Advance Research and Education through Cyberinfrastructure
THE UNIVERSITY OF TEXAS AT EL PASO UTEP
SciDAC
Scientific Discovery through Advanced Computing

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

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

Kepler Scientific Workflow System



This model shows the solution to the classic Lotka-Volterra predator prey dynamics model. It uses the Continuous Time domain to solve two coupled differential equations, one that models the predator population and one that models the prey population. The results are plotted as they are calculated showing both population change and a phase diagram of the dynamics.

Rich Williams, 2003, NCEAS

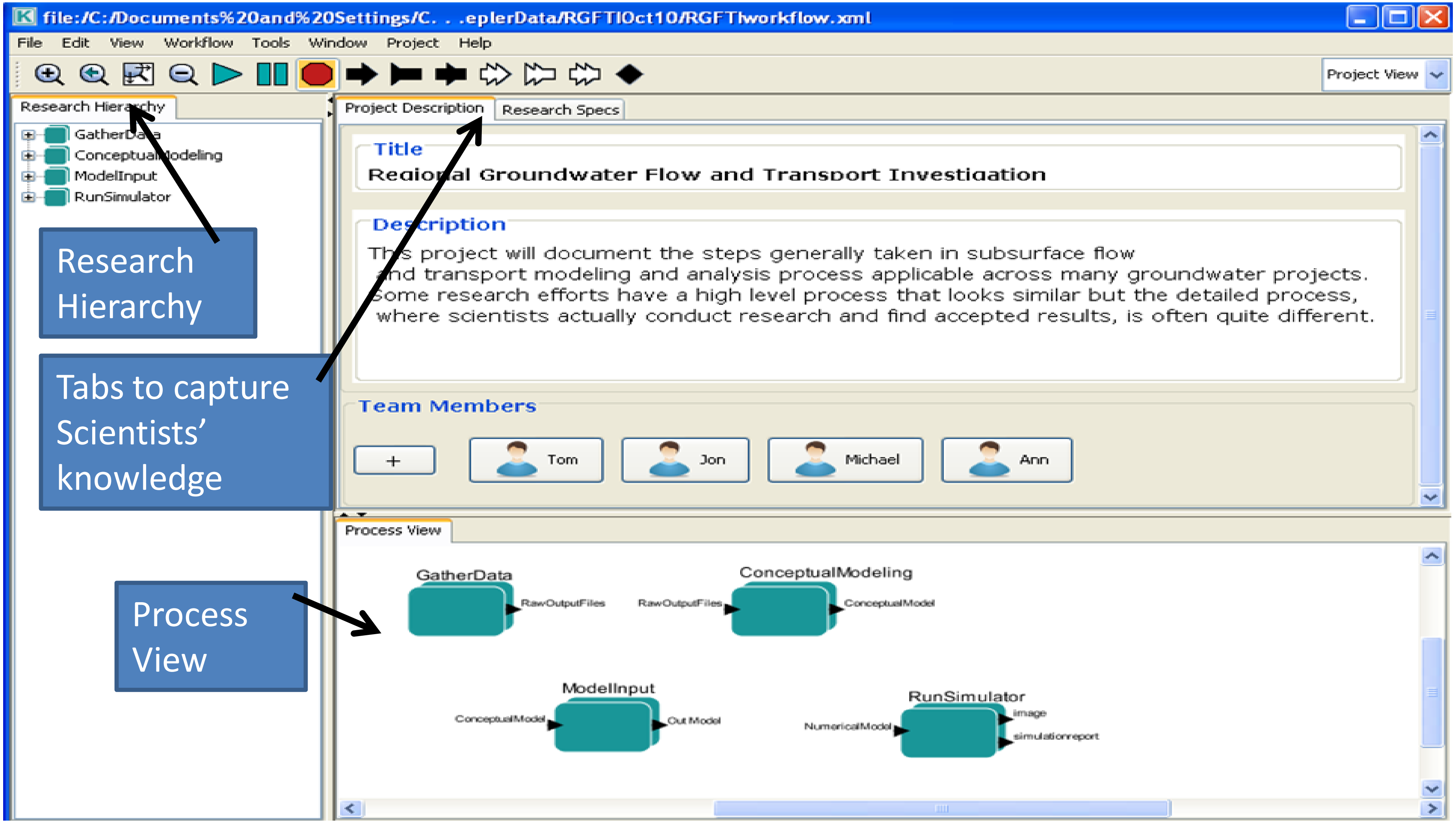
- Collect sufficient information to document a scientific process
- Support reproducing results
- Help collect provenance

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



Title

Regional Groundwater Flow and Transport Investigation

Description

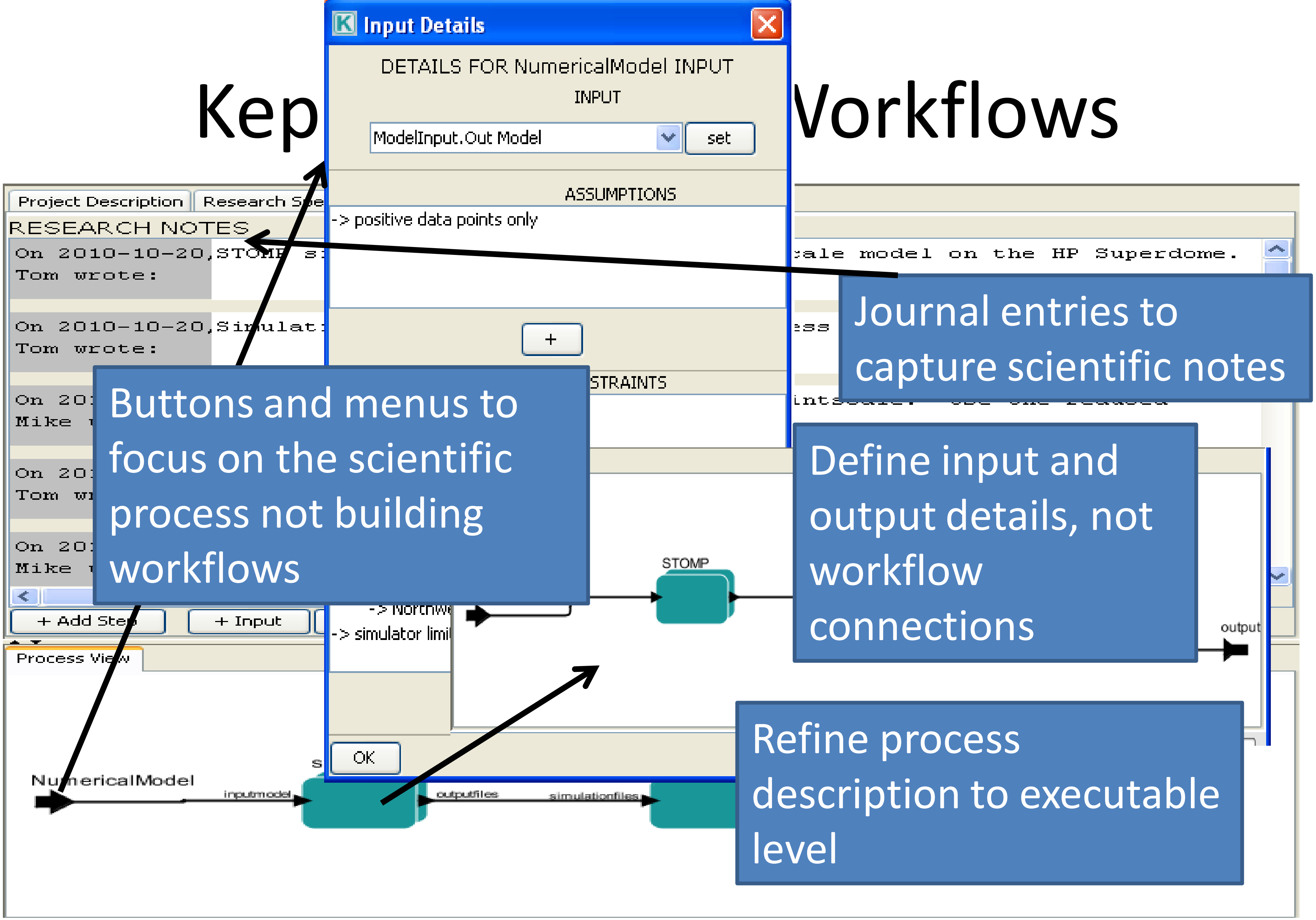
This project will document the steps generally taken in subsurface flow and transport modeling and analysis process applicable across many groundwater projects. Some research efforts have a high level process that looks similar but the detailed process, where scientists actually conduct research and find accepted results, is often quite different.

Research Hierarchy

Tabs to capture Scientists' knowledge

Process View

Kep Vorkflows



Input Details

DETAILS FOR NumericalModel INPUT

INPUT

ModelInput.Out Model [set]

ASSUMPTIONS

-> positive data points only

STRAINTS

-> Northw

-> simulator limi

RESEARCH NOTES

On 2010-10-20,STOMP s
Tom wrote:
cale model on the HP Superdome.

On 2010-10-20,Simulat
Tom wrote:
ess

On 20
Mike
ints

On 20
Tom wr

On 20
Mike

Journal entries to capture scientific notes

Buttons and menus to focus on the scientific process not building workflows

Define input and output details, not workflow connections

Refine process description to executable level

CYBER-ShARE Center of Excellence
Sharing Resources to Advance Research and Education through Cyberinfrastructure
THE UNIVERSITY OF TEXAS AT EL PASO

SciDAC
Scientific Discovery through Advanced Computing

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

CYBER-ShARE Center of Excellence
Sharing Resources to Advance Research and Education through Cyberinfrastructure
THE UNIVERSITY OF TEXAS AT EL PASO
SciDAC Scientific Discovery through Advanced Computing

Contact

Aída Gándara
The University of Texas at El Paso
agandara1@miners.utep.edu

George Chin
Pacific Northwest National Laboratory
George.Chin@pnnl.gov