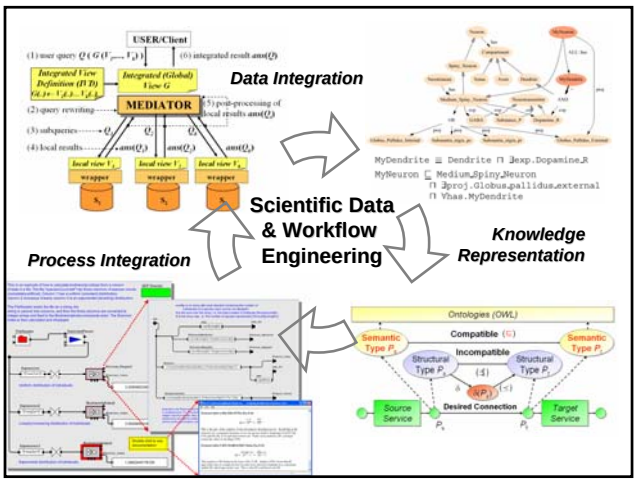
# ECS289F Winter'05
# Scientific Data Management


**Bertram Ludäscher**

**Associate Professor**
**Dept. of Computer Science & Genome Center**
**University of California, Davis**
**ludaesch@ucdavis.edu**

UC DAVIS Department of Computer Science
UC DAVIS genome center
San Diego Supercomputer Center SDSC

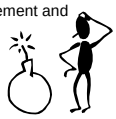




## Logistics

- Class:
  - MWF, 4:10-5pm, 244 Olson
  - http://www.sdsc.edu/~ludaesch/ECS289F-W05.html
  (or Google with "Bertram ECS"; then navigate)
  - Notes will be posted there (& hand-outs available)
- Office hours:
  - General:
    - Wednesday & Friday, 11am-noon, room 3051, Kemper Hall
  - Extra:
    - email ludaesch@ucdavis.edu, subject ECS289F
- Grading:
  - Project 50%, presentation 30%, homework 20%



## Projects

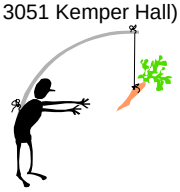


- **Implementation Projects (Hands-on)** with scientific data management and workflow tools:
  - KEPLER scientific workflow system
  - SDSC Storage Resource Broker
  - e.g. run a number of compute-intensive jobs (say from cheminformatics) on a cluster computer via KEPLER, or
  - access a real-time seismic sensor network client application using KEPLER and SRB
- **Research Projects (Theory):**
  - Readings in **data integration**: understand how a **database mediator** works i.e., *query rewriting*
  - Readings in **knowledge representation**: understand how *ontology* languages are used to capture and reason with domain knowledge

- **Combined Projects** (Theory & Hands-on)
  - Implementation of algorithms for query rewriting

## Research Assistant (RA)-ships

- Openings at the **CS department** and new **Genome Center** available; option for **summer internship** at the **San Diego Supercomputer Center**!
- **Mix of theory and practice** ideal:
  - Databases (& Information Systems) background helps a lot
  - No fear of ontologies or workflows (that's why we have this class ;-)
  - Problem-solving (thinking!) and Programming skills
- Come to my office hours (WF, 11am-noon, 3051 Kemper Hall) for details





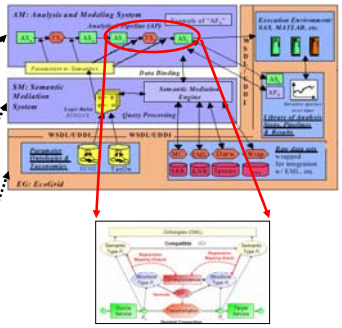
## Today's Outline

- Semantics & Scientific Data Integration
- Semantics & Scientific Workflow Management
- Conclusions



## Anatomy of the Science Environment for Ecological Knowledge (SEEK) Collaboratory

- **Domain Science Driver**
  - Ecology (LTER), biodiversity, …
- **Analysis & Modeling System**
  - Design & execution of ecological models & analysis ("scientific workflows")
  - **{application,upper}-ware**
  - ➔ **Kepler system**
- **Semantic Mediation System**
  - Data Integration of hard-to-relate sources and processes
  - Semantic Types and Ontologies
  - **upper middleware**
  - ➔ **Sparrow Toolkit**
- **EcoGrid**
  - Access to ecology data and tools
  - **{middle,under}-ware**
  - ➔ **unified API to SRB/MCAT, MetaCat, DiGIR, … datasets**




*sample CS problem [DILS'04]*

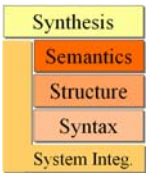


## Common Collaboratories / Distributed Science / Cyberinfrastructure Pieces

- Seamless and uniform data access ("Data-Grid")
  - data & metadata registry
- distributed and high performance computing platform ("Compute-Grid")
  - service registry
- User-friendly workbench / problem-solving environment
  - scientific workflow system
- A common problem:
  - **integrating** (or at least **linking**) data from multiple sites, investigators, communities, …, scales, …, species, …

➔ Federated, integrated, mediated databases
  - often use of semantic extensions (e.g. ontologies)

## Interoperability & Integration Challenges




- ***System aspects: "Grid" Middleware***
  - distributed data & computing, SOA
  - web services, WSDL/SOAP, WSRF, OGSA, …
  - *sources = functions, files, data sets …*
- ***Syntax & Structure:***
  ***(XML-Based) Data Mediators***
  - wrapping, restructuring
  - (XML) queries and views
  - *sources = (XML) databases*
- ***Semantics:***
  ***Model-Based/Semantic Mediators***
  - conceptual models and declarative views
  - Knowledge Representation: ontologies, description logics (RDF(S),OWL ...)
  - *sources = knowledge bases (DB+CMs+ICs)*
- ***Synthesis: Scientific Workflow Design & Execution***
  - Composition of declarative and procedural components into larger workflows
  - *(re)sources = services, processes, actors, …*
  - *Semantic extensions needed here as well!*

- **reconciling $S^5$ heterogeneities**
- **"gluing" together resources**
- **bridging information and knowledge gaps computationally**

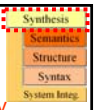


## Information Integration Challenges: *$S^4$ Heterogeneities*


- **System** aspects
  - platforms, devices, data & service distribution, APIs, protocols, …
  - ➔ **Grid middleware technologies**
  - + e.g. single sign-on, platform independence, transparent use of remote resources, …
- **Syntax** & **Structure**
  - heterogeneous data formats (*one for each tool* ...)
  - heterogeneous data models (*RDBs, ORDBs, OODBs, XMLDBs, flat files, …*)
  - heterogeneous schemas (*one for each DB* ...)
  - ➔ **Database mediation and warehousing technologies**
  - + XML-based data exchange, integrated views, transparent query rewriting, …
- **Semantics**
  - descriptive metadata, different terminologies, implicit assumptions & hidden semantics ("context") of experiments, simulations, observation, …
  - ➔ **Knowledge representation & semantic mediation technologies**
  - + "smart" data discovery & integration
  - + e.g. ask about ***X*** ('**mafic**'); find data about ***Y*** ('**diorite**'); be happy anyways!



## Information Integration Challenges: *$S^5$ Heterogeneities*




- **Synthesis** of applications, analysis tools, data & query components, … into "**scientific workflows**"
  - How to make use of these wonderful things & put them together to solve a scientist's problem?

➔ **Scientific Problem Solving Environments (PSEs)**

- ➔ **Portals,Workbench ("scientist's view", end user)**
- + ontology-enhanced data registration, discovery, manipulation
- + creation and registration of new data products from existing ones, …
- ➔ **Scientific Workflow System ("engineer's view", tool maker)**
- + for designing, re-engineering, deploying analysis pipelines and scientific workflows; ***a tool to make new tools …***
- + e.g., creation of new datasets from existing ones, dataset registration, …

Not discussed here: the "6th S": **Social challenges** …



## Our Focus

- Scientific Data Integration:
  - need DB/DI + KR ("semantic mediation")
- Automation of Scientific Data Analysis, Process & Application Integration
  - need for scientific workflow systems
  - need for semantic extensions
- But first:
  - Some data & information integration problems

## An Online Shopper's Information Integration Problem

*El Cheapo: "Where can I get the cheapest copy (including shipping cost) of Wittgenstein's Tractatus Logicus-Philosophicus within a week?"*

***addall.com***

***Mediator (virtual DB)***
*(vs. Datawarehouse)*
*NOTE: non-trivial data engineering challenges!*

***"One-World" Mediation***

*amazon.com* *barnes&noble.com* *half.com* *A1books.com*



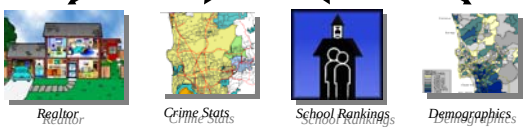
## A Home Buyer's Information Integration Problem

***What houses for sale under $500k have at least 2 bathrooms, 2 bedrooms, a nearby school ranking in the upper third, in a neighborhood with below-average crime rate and diverse population?***

**?**

**Information Integration**

***"Multiple-Worlds" Mediation***



*Realor* *Crime Stats* *School Rankings* *Demographics*



## Information Integration from a Database Perspective

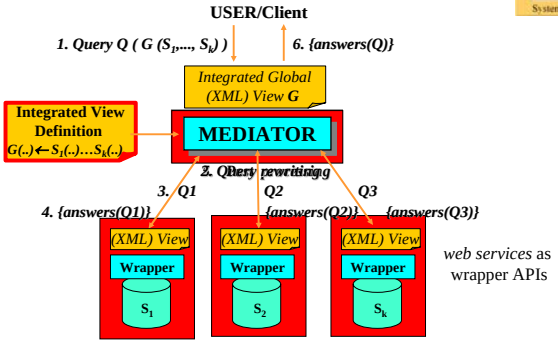
Synthesis
Semantics
Structure
Syntax
System Integ.

- Information Integration Problem
  - **Given:** data sources $S_1$, ..., $S_k$ (databases, web sites, ...) and user questions $Q_1$,..., $Q_n$ that can –in principle– be answered using the information in the $S_i$
  - **Find:** the answers to $Q_1$, ..., $Q_n$
- The Database Perspective: **source = "database"**
  - ⇒ $S_i$ has a **schema** (relational, XML, OO, ...)
  - ⇒ $S_i$ **can be queried**
  - ⇒ define virtual (or materialized) **integrated (or global) view G** over local sources $S_1$ ,..., $S_k$ using **database query languages** (SQL, XQuery,...)
  - ⇒ ***questions become queries*** $Q_i$ against $G(S_1,..., S_k)$



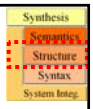
## Standard Mediator Architecture

Synthesis
Semantics
Structure
Syntax
System Integ.




*web services* as wrapper APIs

## Query Planning in Data Integration




- **Given**:
  - **Declarative user query Q**: ***answer(…) ← …G …***
  - … & { G ← … S … } *global-as-view* (GAV)
  - … & { S ← … G … } *local-as-view* (LAV)
  - … & { ic(…) ← … S … G… } *integrity constraints* (ICs)
- **Find**:
  - equivalent (or minimal containing, maximal contained) **query plan Q'**: ***answer(…) ← … S …***
    - ➔ query rewriting (logical/calculus, algebraic, physical levels)
- **Results**:
  - A variety of results/algorithms; depending on classes of queries, views, and ICs: **P**, **NP**, … , **undecidable**
  - hot research area in core CS (database community)



## A Neuroscientist's Information Integration Problem

Biomedical Informatics Research Network
http://nbirn.net

*What is the cerebellar distribution of rat proteins with more than 70% homology with human NCS-1? Any structure specificity? How about other rodents?*

?

**Information Integration**

***"Complex Multiple-Worlds" Mediation***

***Inter-source links:***
- *unclear for the non-scientists*
- *hard for the scientist*

*protein localization (NCMIR)*

*sequence info (CaPROT)*

*morphometry (SYNAPSE)*

*neurotransmission (SENSELAB)*



## The Problem: Scientific Data Integration or: … *from Questions to Queries* …

*What is the distribution and U/Pb zircon ages of A-type plutons in VA?*
*How about their 3-D geometry?*
*How does it relate to host rock structures?*

*domain knowledge*

*Knowledge Representation: ontologies, concept spaces*

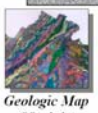
*Database mediation*
*Data modeling*

*raw data*

?

**Information Integration**

***"Complex Multiple-Worlds" Mediation***

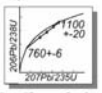


*Geologic Map (Virginia)*

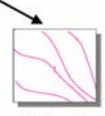
| | |
|---|---|
| SiO2 | 72.22 |
| CaO | 0.62 |
| K2O | 4.88 |
| Ga | 21.1 |
| Sr | 72.6 |

*GeoChemical*

*GeoPhysical (gravity contours)*




*GeoChronologic (Concordia)*



*Foliation Map (structure DB)*

GEON CYBERINFRASTRUCTURE FOR THE GEOSCIENCES www.geongrid.org 2

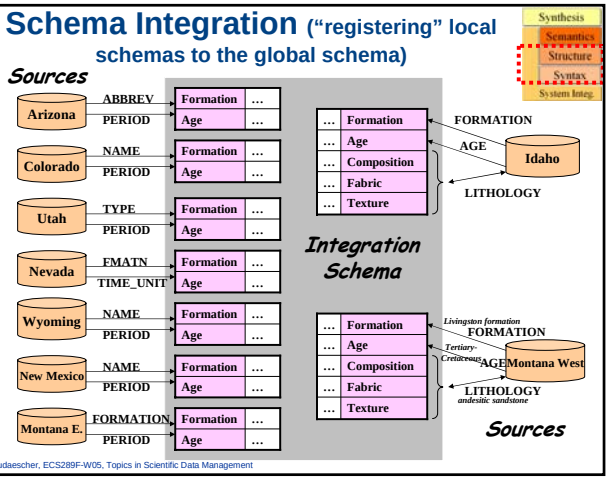
## Scientific Data Integration using Semantic Extensions

## Example: Geologic Map Integration

- **Given**:
  - Geologic maps from different state geological surveys (shapefiles w/ different data schemas)
  - Different ontologies:
    - Geologic age ontology (e.g. USGS)
    - Rock classification ontologies:
      - Multiple hierarchies (chemical, fabric, texture, genesis) from Geological Survey of Canada (GSC)
      - Single hierarchy from British Geological Survey (BGS)
- **Problem**:
  - Support **uniform queries** across all maps
  - … possibly using **different** ontologies
  - Support **registration** w/ ontology A, **querying** w/ ontology B

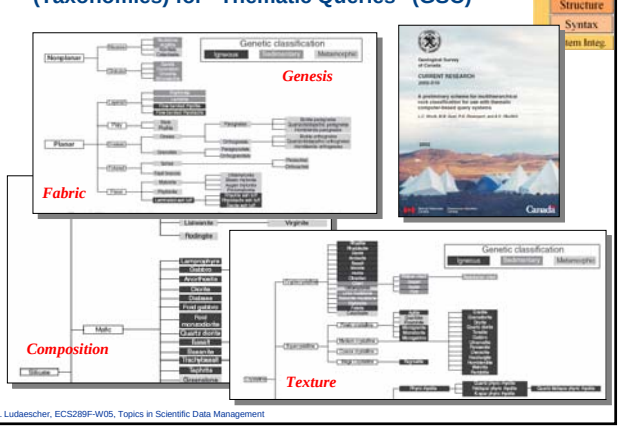


## Schema Integration ("registering" local schemas to the global schema)



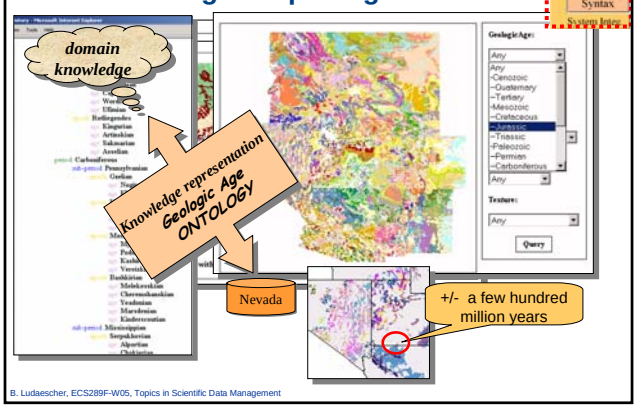



## Multihierarchical Rock Classification "Ontology" (Taxonomies) for "Thematic Queries" (GSC)

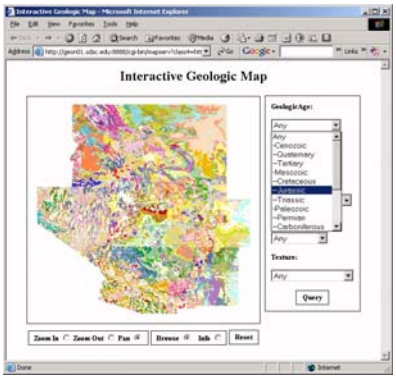





## Ontology-Enabled Application Example: Geologic Map Integration

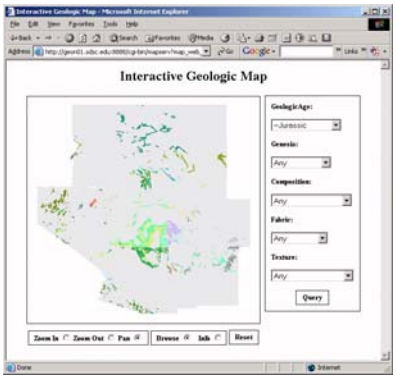


## Querying by Geologic Age …





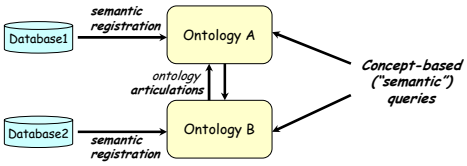

## Querying by Geologic Age: Results





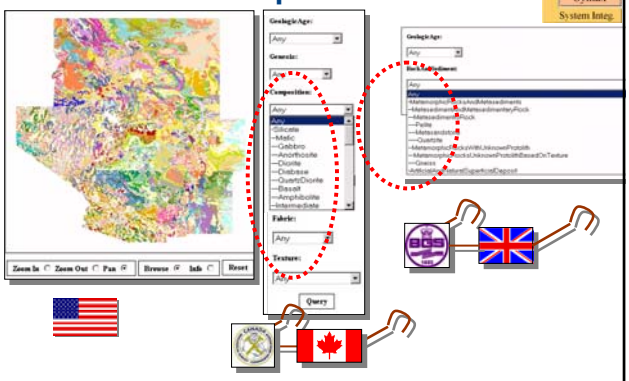

## Semantic Mediation (via "semantic registration" of schemas and ontology articulations)

- **Schema elements and/or data values are associated with concept expressions from the target ontology**
  - ➔ conceptual queries "through" the ontology
- Articulation ontology
  - ➔ source registration to A, querying through B
- Semantic mediation: query rewriting w/ ontologies

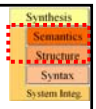


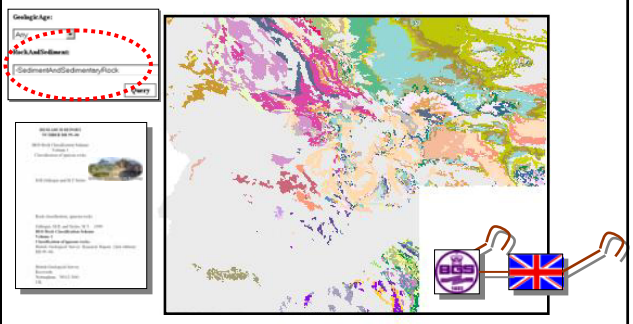



## Different views on State Geological Maps



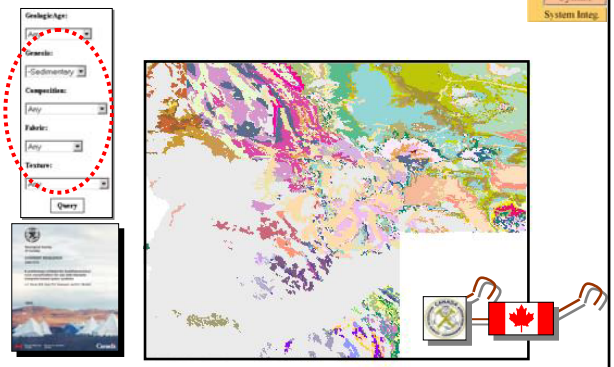
## Sedimentary Rocks: BGS Ontology






## Sedimentary Rocks: GSC Ontology




## Some Thoughts …

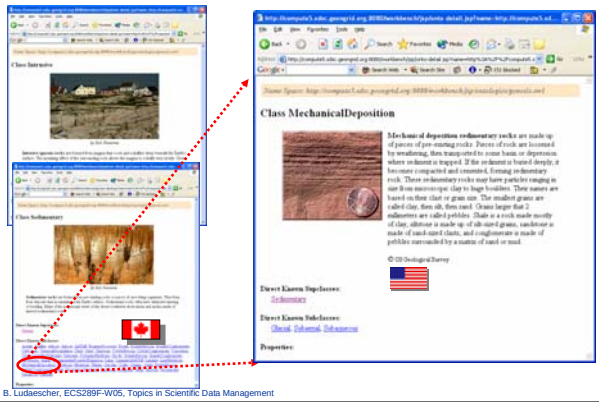
- Translate this idea of multiple conceptual (ontology) views to **your** domain!
  - e.g. datasets ⇔ biological pathways registration
- Your data is valuable (time & $$$ spent in producing it)

➔ **data (re-)usability**

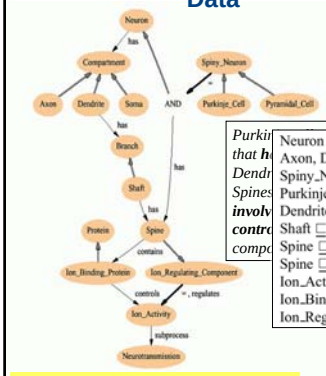
- Metadata helps to discover, localize, assess relevant data sets, given particular scientific questions & queries
- Does your system "understand" what to do with the metadata?
- **Capturing more semantics** of a data set in a way that humans and systems can exploit it is an **investment in reusability**
  - "We are producing more and more data"
  - Today "we can store everything!"
  - But can we **use anything?** (i.e., is anyone looking at the data after the initial creation?)
- Design system, interfaces, data and metadata models with reusability in mind (think archives and "time capsules")
- This may even be pushed to the experiment/simulation/workflow design…



## Data Semantics and Ontologies should be useful for Humans and "The Machine"



## Example: Domain Knowledge to "glue" SYNAPSE & NCMIR Data



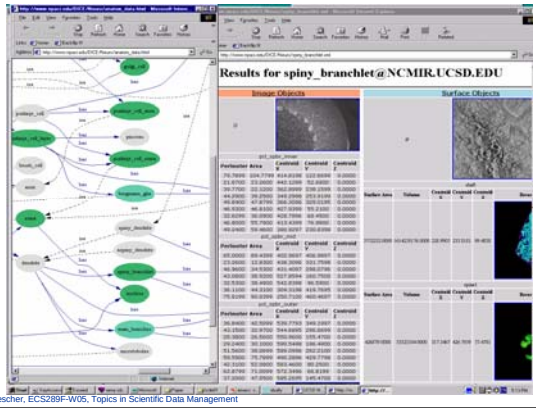

Neuron ⊑ ∃has.Compartment
Axon, Dendrite, Soma ⊑ Compartment
Spiny_Neuron ≡ Neuron ⊓ ∃has.Spine
Purkinje_Cell, Pyramidal_Cell ⊑ Spiny_Neuron
Dendrite ⊑ ∃has.Branch
Shaft ⊑ Branch ⊓ ∃has.Spine
Spine ⊑ ∃contains.Ion_Binding_Protein
Spine ⊑ Ion_Regulating_Component
Ion_Activity ⊑ ∃subprocess_of.Neurotransmission
Ion_Binding_Protein ⊑ Protein ⊓ ∃controls.Ion_Activity
Ion_Regulating_Component ≡ ∃regulates.Ion_Activity

*Made usable for the **system** using Description Logic*

*formalized as domain map/ontology*

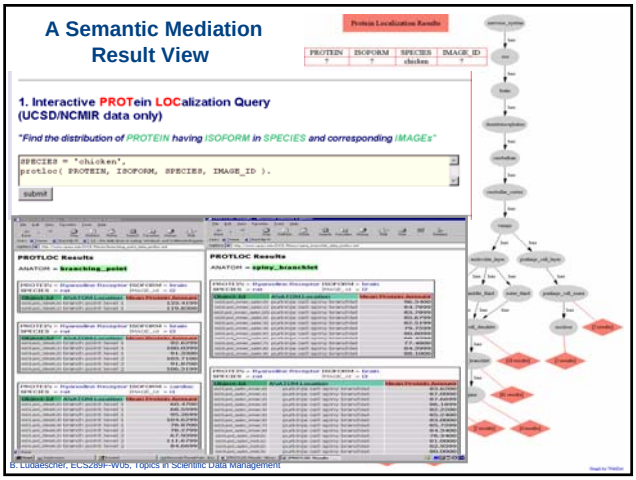
## "Semantic Source Browsing":

**Domain Maps/Ontologies (left) & conceptually linked data (right)**






## A Semantic Mediation Result View

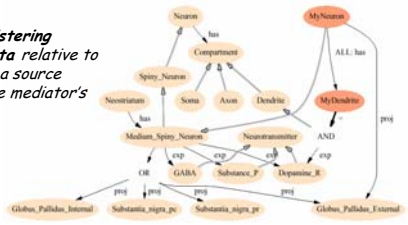





## Source Contextualization through Ontology Refinement

Biomedical Informatics Research Network

*In addition to **registering** ("hanging off") **data** relative to existing concepts, a source may also **refine** the mediator's domain map…*




MyDendrite ≡ Dendrite ⊓ ∃exp.Dopamine_R
MyNeuron ⊑ Medium_Spiny_Neuron
⊓ ∃proj.Globus_pallidus_external
⊓ ∀has.MyDendrite

⇒ *sources can register **new concepts** at the mediator …*
⇒ *increase your **data usability***

## Outline

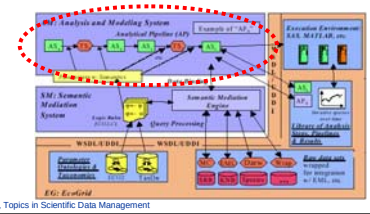
- Semantics & Scientific Data Integration
- Semantics & Scientific Workflow Management
- Conclusions



## What is a Scientific Workflow (SWF)?

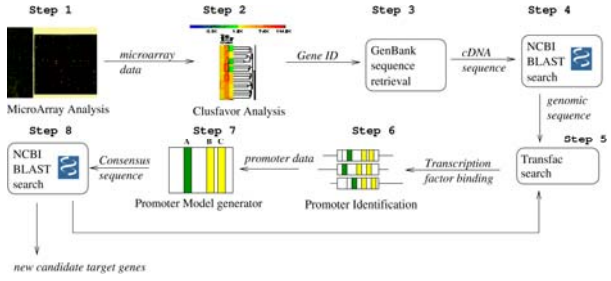
- **Goals**:
  - automate a scientist's repetitive data management and analysis tasks
  - typical phases:
    - data access, scheduling, generation, transformation, aggregation, analysis, visualization

➔ design, test, share, deploy, execute, reuse, … SWFs





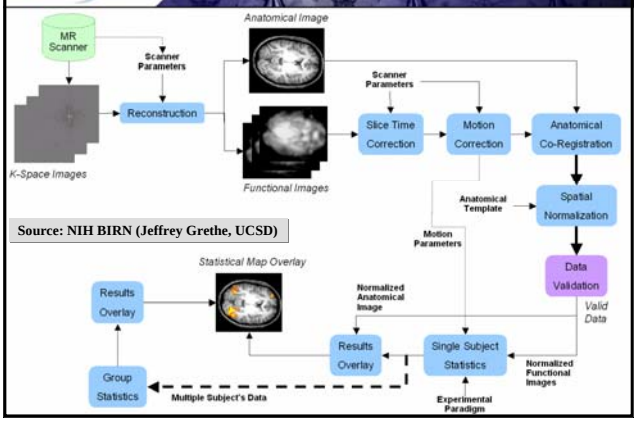
## Promoter Identification Workflow




Source: Matt Coleman (LLNL)

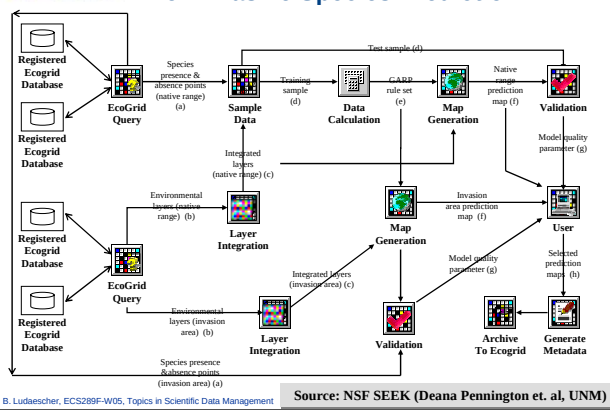


## Functional MRI Analysis Workflow




Source: NIH BIRN (Jeffrey Grethe, UCSD)

## Ecology: GARP Analysis Pipeline for Invasive Species Prediction



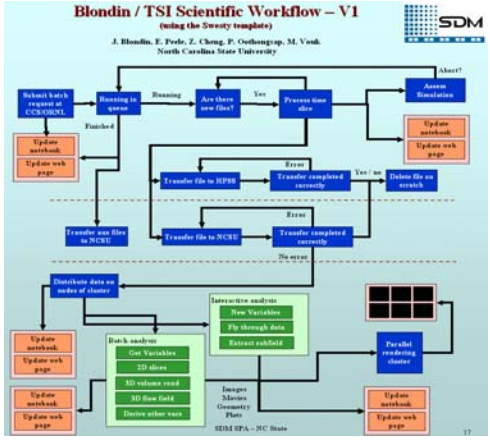

Source: NSF SEEK (Deana Pennington et. al, UNM)



## Blondin / TSI Scientific Workflow – V1

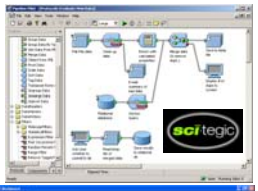
(using the Sweety template)

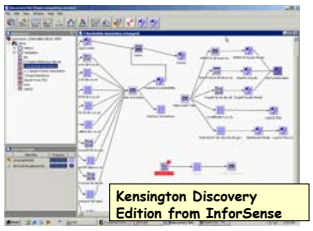
J. Blondin, E. Peele, Z. Cheng, P. Oesthongrap, M. Vouk
North Carolina State University



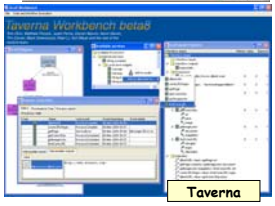



## Commercial & Open Source Scientific "Workflow" (often **Dataflow**) Systems



Kensington Discovery Edition from InforSense

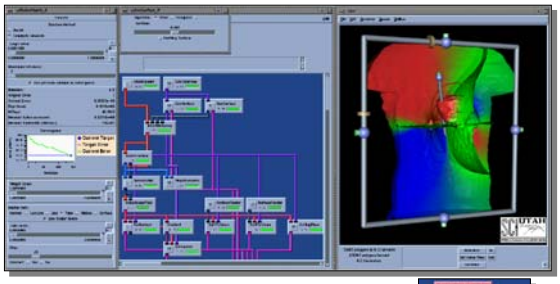



Taverna

Triana



## SCIRun: Problem Solving Environments for Large-Scale Scientific Computing

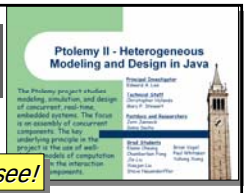


- SCIRun: PSE for interactive construction, debugging, and steering of large-scale scientific computations and visualizations
- Component model, based on generalized dataflow programming

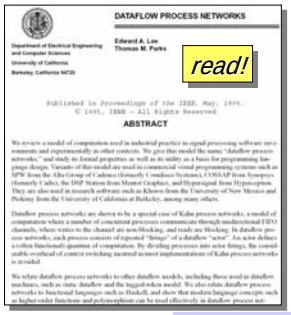
Steve Parker (cs.utah.edu)

# Ptolemy II

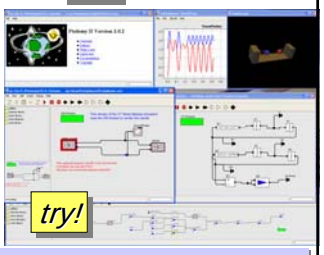





see!




read!



try!

**Source: Edward Lee et al. http://ptolemy.eecs.berkeley.edu/ptolemyII/**



## Why Ptolemy II (and thus KEPLER)?

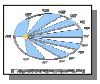
- Ptolemy II Objective:
  - "The focus is on **assembly of concurrent components**. The key underlying principle in the project is the use of **well-defined models of computation** that govern the interaction between components. A major problem area being addressed is the use of **heterogeneous mixtures of models of computation**."
- Dataflow Process Networks w/ natural support for **abstraction**, **pipelining** (streaming) **actor-orientation,** actor **reuse**
- **User-Orientation**
  - Workflow design & exec console (Vergil GUI)
  - **"Application/Glue-Ware"**
    - excellent modeling and design support
    - run-time support, monitoring, …
    - **not** a middle-/underware (we use someone else's, e.g. Globus, SRB, …)
    - but middle-/underware is conveniently accessible through actors!
- PRAGMATICS
  - Ptolemy II is mature, continuously extended & improved, well-documented (500+pp)
  - open source system
  - Ptolemy II folks actively participate in KEPLER



## KEPLER/CSP: Contributors, Sponsors, Projects

(or loosely coupled Communicating Sequential Persons ;-)

Ilkay Altintas *SDM, Resurgence*
Kim Baldridge *Resurgence, NMI*
Chad Berkley *SEEK*
Shawn Bowers *SEEK*
Terence Critchlow *SDM*
Tobin Fricke *ROADNet*
Jeffrey Grethe *BIRN*
Christopher H. Brooks *Ptolemy II*
Zhengang Cheng *SDM*
Dan Higgins *SEEK*
Efrat Jaeger *GEON*
Matt Jones *SEEK*
Werner Krebs, *EOL*
Edward A. Lee *Ptolemy II*
Kai Lin *GEON*
Bertram Ludaescher *SDM, SEEK, GEON, BIRN, ROADNet*
Mark Miller *EOL*
Steve Mock *NMI*
Steve Neuendorffer *Ptolemy II*
Jing Tao *SEEK*
Mladen Vouk *SDM*
Xiaowen Xin *SDM*
Yang Zhao *Ptolemy II*
Bing Zhu *SEEK*
…

www.kepler-project.org

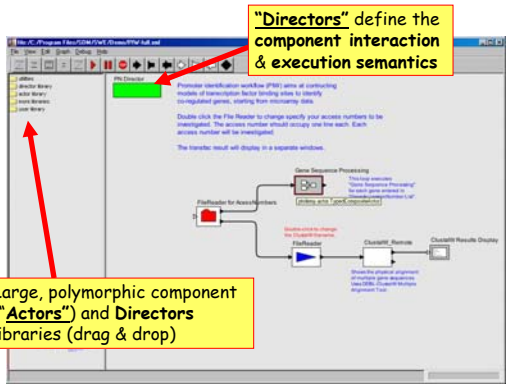
ROADNet
SEEK
GEON The Geosciences Network
EOL
BIRN Biomedical Informatics Research Network
SciDAC Scientific Discovery through Advanced Computing
SDM
DARPA



## KEPLER: An Open Collaboration

- Initiated by members from DOE SDM/SPA and NSF SEEK; now several other projects (GEON, Ptolemy II, EOL, Resurgence/NMI, …)
- Open Source (BSD-style license)
- Intensive Communications:
  - Web-archived mailing lists
  - IRC (!)
  - Meetings, Hackathons
- Co-development:
  - via shared CVS repository
  - joining as a new co-developer (currently):
    - get a CVS account (read-only)
    - local development + contribution via existing KEPLER member
    - be voted "in" as a member/co-developer
- Software & social engineering
  - How to better accommodate new groups/communities?
  - How to better accommodate different usage/contribution models (core dev … special purpose extender … user)?

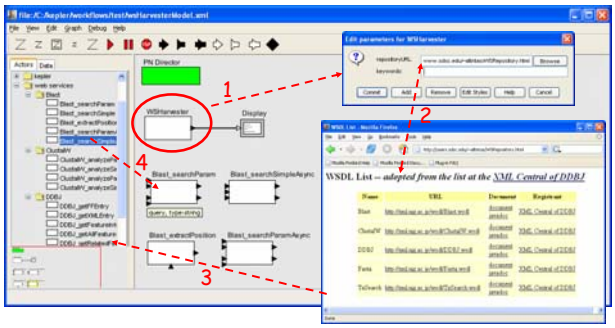
## Ptolemy II/KEPLER GUI (Vergil)






## Web Services ➔ Actors

(WS Harvester)



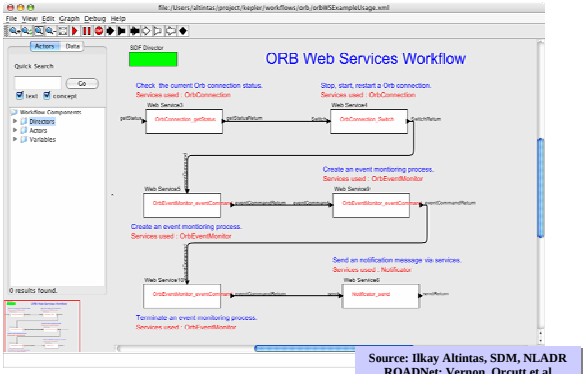

➔ "Minute-made" (MM) WS-based application integration

- Similarly: MM workflow design & sharing w/o implemented components



## Rapid Web Service-based Prototyping

(Here: ROADNet Command & Control Services for LOOKING Kick-Off Mtg)

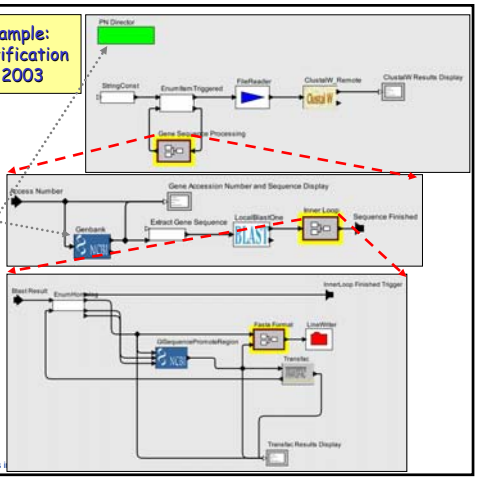


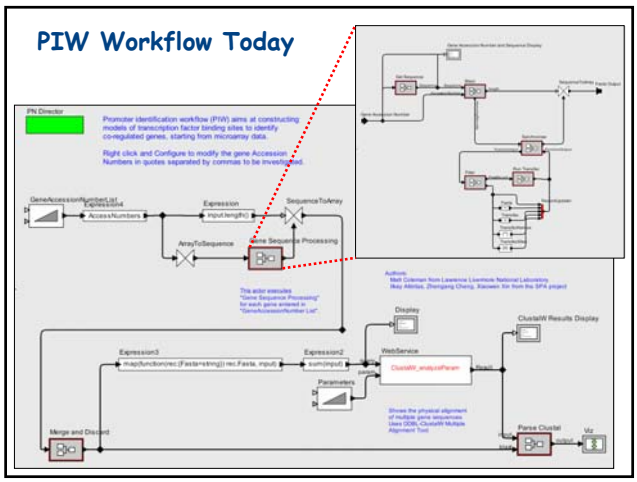

Source: Ilkay Altintas, SDM, NLADR
ROADNet: Vernon, Orcutt et al
Web services: Tony Fountain et al

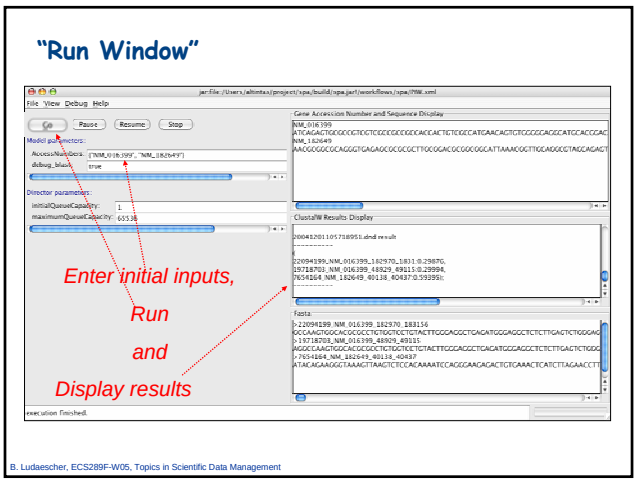


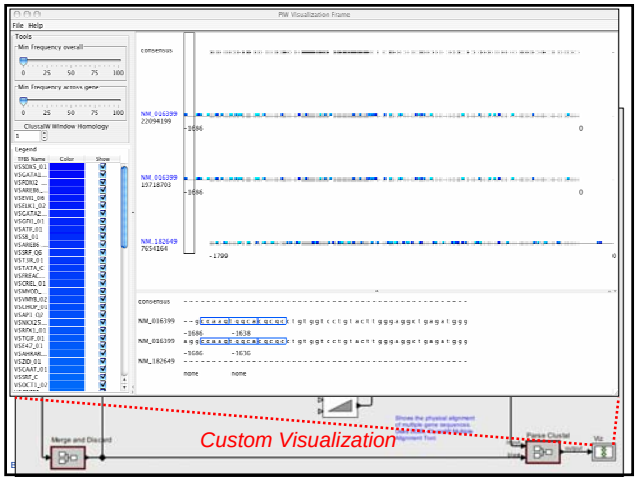
## An "early" example: Promoter Identification SSDBM, AD 2003

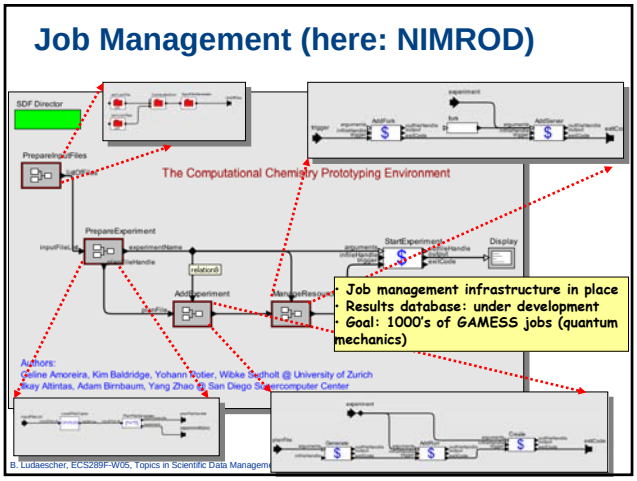
- Scientist models application as a "workflow" of connected components ("actors")
- If all components exist, the workflow can be automated/ executed
- Different **directors** can be used to pick appropriate execution model (often "pipelined" execution: PN director)



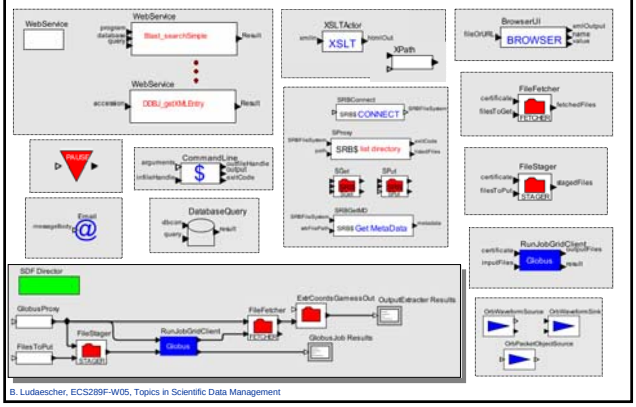

PIW Workflow Today
PN Director
Promoter identification workflow (PIW) aims at constructing models of transcription factor binding sites to identify co-regulated genes, starting from microarray data.
Right click and Configure to modify the gene Accession Numbers in quotes separated by commas to be investigated.
GeneAccessionNumberList
AccessNumbers
Expression
SequenceToArray
ArrayToSequence
Gene Sequence Processing
Display
ClustalW Results Display
Expression3
Expression2
WebService
Parameters
Merge and Discard
Parse Clustal
Viz


"Run Window"
File View Debug Help
Go
Pause
Resume
Stop
Model parameters:
Director parameters:
Gene Accession Number and Sequence Display
ClustalW Results Display
Fasta
execution finished.
Enter initial inputs,
Run
and
Display results
B. Ludaescher, ECS289F-W05, Topics in Scientific Data Management


PIW Visualization Frame
File Help
Tools
Legend
Merge and Discard
Parse Clustal
Viz
Custom Visualization

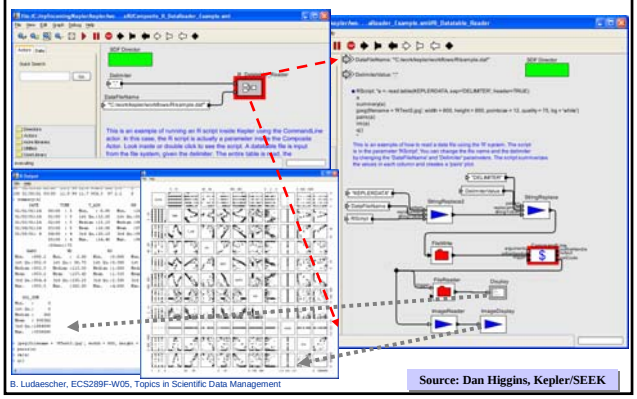

Job Management (here: NIMROD)
SDF Director
PrepareInputFiles
The Computational Chemistry Prototyping Environment
PrepareExperiment
AddExperiment
MergeResults
StartExperiment
Display
Job management infrastructure in place
Results database: under development
Goal: 1000's of GAMESS jobs (quantum mechanics)
Authors:
Claline Amoreira, Kim Baldridge, Yohann Nobier, Wibke Sudholt @ University of Zurich
Ilkay Altintas, Adam Birnbaum, Yang Zhao @ San Diego Supercomputer Center
B. Ludaescher, ECS289F-W05, Topics in Scientific Data Management

## Some Recent Actor Additions





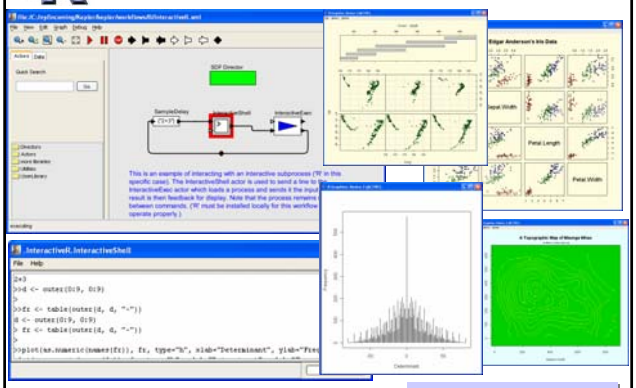

## in KEPLER (w/ editable script)



Source: Dan Higgins, Kepler/SEEK



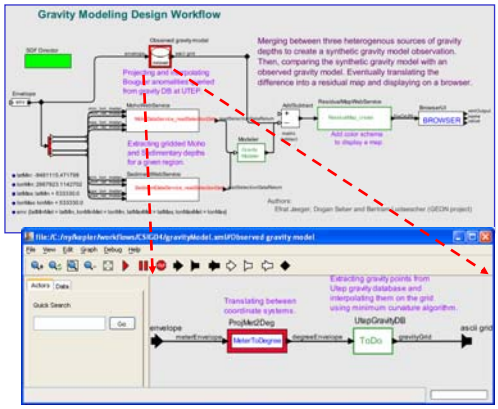
## in KEPLER (interactive session)



Source: Dan Higgins, Kepler/SEEK



## Blurring Design (ToDo) and Execution



## Some Scientific Workflow Challenges

- **Typical Features**
  - data-intensive and/or compute-intensive
  - plumbing-intensive (consecutive web services won't fit)
  - dataflow-oriented
  - distributed (remote data, remote processing)
  - user-interaction "in the middle", …
  - … vs. (C-z; bg; fg)-ing ("detach" and reconnect)
  - advanced programming constructs (map(f), zip, takewhile, …)
  - logging, provenance, "registering back" (intermediate) products…

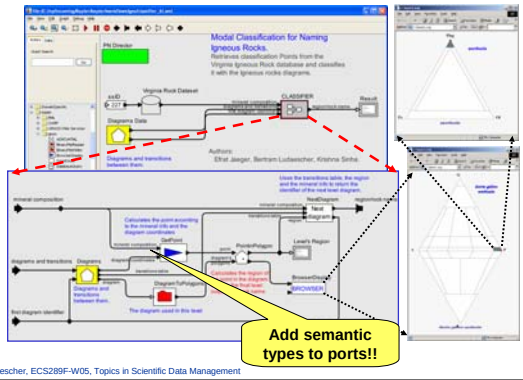


## Scientific Workflows & Semantics

- Registering data to ontologies: semantic types (in addition to structural data types)
- Smarter data set discovery & integration
- Now also:
  - Smarter workflow design
  - More "intelligent" (semantics-aware) component composition
  - Improved (re-)usability of data, services (actors), and workflows
  - Given semantic type of my input ports, what other data sets / actors produce such input

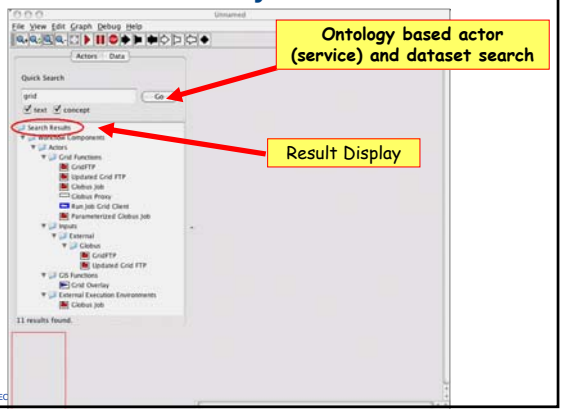


## Reengineering a Geoscientist's Mineral Classification Workflow






## Beginnings: Ontology-based Actor/Service Discovery



## Semantics & Scientific Workflows

Data comes from heterogeneous sources
- Real-world observations
- Spatial-temporal contexts
- Collection/measurement protocols and procedures
- Many representations for the same information (count, area, density)
- Schematically heterogeneous

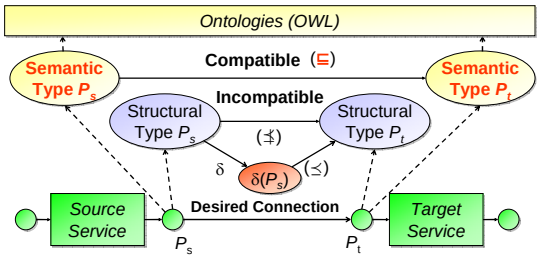
Data discovered and "synthesized" manually

Hard to reuse/repurpose existing analytical steps (another form of heterogeneity)



## A KR+DI+Scientific Workflow Problem

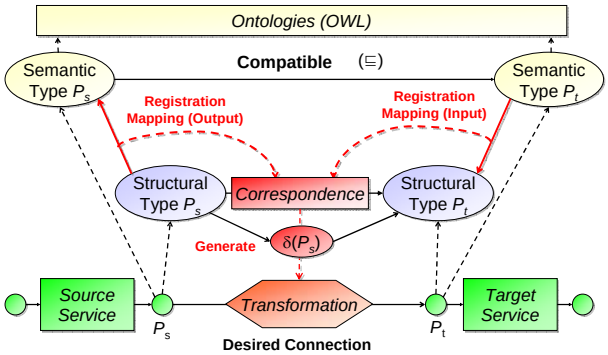
- **Services can be** *semantically compatible*, **but** *structurally incompatible*






Source: [Bowers-Ludaescher, DILS'04]

## Ontology-Informed Data Transformation ("Structure-Shim")





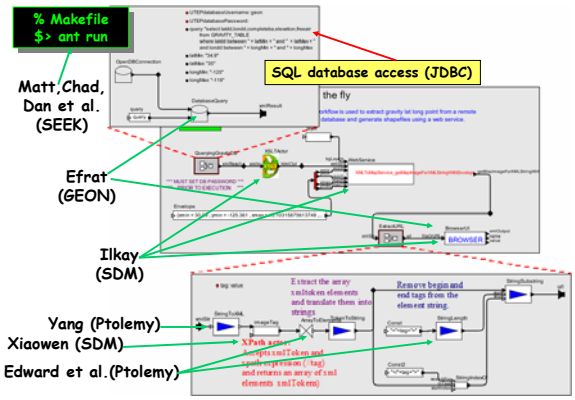

Source: [Bowers-Ludaescher, DILS'04]

## Outline

- Scientific Data Integration
- Scientific Workflow Management
- Musings & Conclusions

## GEON Dataset Generation & Registration

(a co-development in KEPLER)






## Further Reading

under review – available upon request from ludaesch@sdsc.edu



## Related Publications

- Semantic Data Registration and Integration
  - **On Integrating Scientific Resources through Semantic Registration, S. Bowers, K. Lin, and B. Ludäscher,** *16th International Conference on Scientific and Statistical Database Management* (SSDBM'04), 21-23 June 2004, Santorini Island, Greece.
  - **A System for Semantic Integration of Geologic Maps via Ontologies, K. Lin and B. Ludäscher. In** *Semantic Web Technologies for Searching and Retrieving Scientific Data* (SCISW), Sanibel Island, Florida, 2003.
  - **Towards a Generic Framework for Semantic Registration of Scientific Data, S. Bowers and B. Ludäscher. In** *Semantic Web Technologies for Searching and Retrieving Scientific Data* (SCISW), Sanibel Island, Florida, 2003.
  - **The Role of XML in Mediated Data Integration Systems with Examples from Geological (Map) Data Interoperability, B. Brodaric, B. Ludäscher, and K. Lin. In** *Geological Society of America (GSA) Annual Meeting*, volume 35(6), November 2003.
  - **Semantic Mediation Services in Geologic Data Integration: A Case Study from the GEON Grid, K. Lin, B. Ludäscher, B. Brodaric, D. Seber, C. Baru, and K. A. Sinha. In** *Geological Society of America (GSA) Annual Meeting*, volume 35(6), November 2003.
- Query Planning and Rewriting
  - **Processing First-Order Queries under Limited Access Patterns, Alan Nash and B. Ludäscher,** *Proc. 23rd ACM Symposium on Principles of Database Systems* (PODS'04) Paris, France, June 2004.
  - **Processing Unions of Conjunctive Queries with Negation under Limited Access Patterns, Alan Nash and B. Ludäscher.,** *9th Intl. Conference on Extending Database Technology* (EDBT'04) Heraklion, Crete, Greece, March 2004, LNCS 2992.
  - **Web Service Composition Through Declarative Queries: The Case of Conjunctive Queries with Union and Negation, B. Ludäscher and Alan Nash. Research abstract (poster),** *20th Intl. Conference on Data Engineering* (ICDE'04) Boston, IEEE Computer Society, April 2004.



## Related Publications

- Scientific Workflows
  - **Kepler: An Extensible System for Design and Execution of Scientific Workflows, I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock,** *16th International Conference on Scientific and Statistical Database Management* (SSDBM'04), 21-23 June 2004, Santorini Island, Greece.
  - **Kepler: Towards a Grid-Enabled System for Scientific Workflows, Ilkay Altintas, Chad Berkley, Efrat Jaeger, Matthew Jones, Bertram Ludäscher, Steve Mock,** Workflow in Grid Systems (GGF10), Berlin, March 9th, 2004.
  - **An Ontology-Driven Framework for Data Transformation in Scientific Workflows, S. Bowers and B. Ludäscher,** *Intl. Workshop on Data Integration in the Life Sciences* (DILS'04), March 25-26, 2004 Leipzig, Germany, LNCS 2994.
  - **A Web Service Composition and Deployment Framework for Scientific Workflows, I. Altintas, E. Jaeger, K. Lin, B. Ludaescher, A. Memon, In the** *2nd Intl. Conference on Web Services* (ICWS), San Diego, California, July 2004.