ECS289F Winter’05
Scientific Data Management

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

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
Interoperability & Integration Challenges

- **System aspects**: “Grid” Middleware
  - distributed data & computing, SOA
  - web services, WSDDI/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
  - (res)ources = services, processes, actors, …
  - Semantic extensions needed here as well!

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, …

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

Not discussed here: the “6th S” Social challenges …
An Online Shopper’s Information Integration Problem

El Cheapo: “Where can I get the cheapest copy (including shipping cost) of Wittgenstein’s Tractatus Logico-Philosophicus within a week?”

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

“One-World” Mediation

Information Integration

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

“A Multiple-Worlds” Mediation

Information Integration from a Database Perspective

- 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

1. Query $Q(G(S_1, ..., S_k))$
2. Query rewriting
3. $S_1$ Wrapper
4. $S_2$ Wrapper
5. $S_k$ Wrapper
6. $\{\text{answers}(Q)\}$

web services as wrapper APIs
Query Planning in Data Integration

- **Given:**
  - Declarative user query \( Q \): \( \text{answer}(...) \leftarrow G ... \)
  - \( \cdots \& \{ G \leftarrow S \cdots \} \) global-as-view (GAV)
  - \( \cdots \& \{ S \leftarrow G \cdots \} \) local-as-view (LAV)
  - \( \cdots \& \{ \text{ic}(...) \leftarrow S \cdots G \cdots \} \) integrity constraints (ICs)

- **Find:**
  - equivalent (or minimal containing, maximal contained) query plan \( Q' \): \( \text{answer}(...) \leftarrow S \cdots \)
  - query rewriting (logical/calculus, algebraic, physical levels)

- **Results:**
  - A variety of results/algorithm; depending on classes of queries, views, and ICs: \( P, NP, \ldots \), undecidable
  - hot research area in core CS (database community)
Example: Geologic Map Integration

• **Given:**
  – Geologic maps from different state geological surveys (shapefiles with 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

**Multihierarchical Rock Classification “Ontology” (Taxonomies) for “Thematic Queries” (GSC)**

**Ontology-Enabled Application Example: Geologic Map Integration**
**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**
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”
Purkinje cells and Pyramidal cells have dendrites that have higher-order branches that contain spines. Dendritic spines are ion (calcium) regulating components. Spines have ion binding proteins. Neurotransmission involves ionic activity (release). Ion-binding proteins control ion activity (propagation) in a cell. Ion-regulating components of cells affect ionic activity (release). Domain expert knowledge made usable for the system using Description Logic.

Example: Domain Knowledge to “glue” SYNAPSE & NCMIR Data

“Semantic Source Browsing”: Domain Maps/Ontologies (left) & conceptually linked data (right)

Source Contextualization through Ontology Refinement

In addition to registering (“hanging off”) data relative to existing concepts, a source may also refine the mediator’s domain map...

⇒ 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)

BIRN Experimental Workflow

Source: NIH BIRN (Jeffrey Grethe, UCSD)
Ecology: GARP Analysis Pipeline for Invasive Species Prediction

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

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

Kensington Discovery Edition from InforSense

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)
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-Orientaton**
  - 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, …)
  - 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: 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)?
"Directors" define the component interaction & execution semantics

Large, polymorphic component ("Actors") and Directors libraries (drag & drop)

"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)

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)

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

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

Source: Ilkay Altintas, SDM, NLADR
PIW Workflow Today

Enter initial inputs, Run and Display results

“Run Window”

Job Management (here: NIMROD)

Job management infrastructure in place. Results database: under development. Goal: 1000’s of GAMESS jobs (quantum mechanics)
Some Recent Actor Additions

in KEPLER (w/ editable script)

Source: Dan Higgins, Kepler/SEEK

in KEPLER (interactive session)

Blurring Design (ToDo) and Execution

Source: Dan Higgins, Kepler/SEEK

Source: Dan Higgins, Kepler/SEEK
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

- Add semantic types to ports!!
- Ontology based actor (service) and dataset search
- Result Display
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

Outline

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

- Semantic Data Registration and Integration

- Query Planning and Rewriting

Further Reading

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