Discovery Workflow: (ServiceFlow)
Programming the Grid

Prof. Yike Guo
Imperial College London
**Discovery Net**

**Goal:** Constructing the World’s First Infrastructure for Global Wide Knowledge Discovery Services

**Funding:**
- One of the Eight UK National e-Science Projects (£2.4 M)

**Key Features:**
- Allow Scientists to Construct, Share and Execute Complex Knowledge Discovery Procedures & Services
- Allow Institutions to Integrate, Manage and Utilise its Intellectual Property

**Applications:**
- Life Science
- Environmental Modelling
- Geo-hazard Prediction
**Business Processes Management:** Workflow describes interaction and collaboration between different entities (inter/intra organization)
- Workflow co-ordinates events and actions
- Business Process Re-engineering/ Business Process Management

**Application Integration:** Workflow provides Glue to integrate distributed applications.
- Workflow describes composition of individual programs/components/applications
- Leverage distributed software resources
- Mechanisms for moving data and results
- A natural model for service composition

**Data Integration and Analysis:** Workflow describes how a particular data results (value, table, ..aggregation, cluster and predicative model) have been generated and related
- Workflow defines a virtual schema
- Workflow for planning distributed query
- Workflow defines a series of data transformation and analysis tasks

**Resource Planning:** Workflow describes the action logic of computation
- Workflow plans a computational task
- Workflow provides audit trial of a complex process
- Workflow models a communicating (control/data flow) protocol of a complex system (Petri Net).
<table>
<thead>
<tr>
<th>Business WORKFLOWS</th>
<th>Web Service WORKFLOWS</th>
<th>Analytical WORKFLOWS</th>
<th>HPC Job Scheduling WORKFLOWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Workflow</td>
<td>WSFL</td>
<td>Clementine</td>
<td>Unicore</td>
</tr>
<tr>
<td>Staffware</td>
<td>XLANG</td>
<td>SAS</td>
<td>LSF</td>
</tr>
<tr>
<td>Carnot</td>
<td>BPEL4WS</td>
<td>Kwiz</td>
<td>TurboWorx</td>
</tr>
<tr>
<td>Maximum</td>
<td>BPML</td>
<td>Scitegic</td>
<td></td>
</tr>
<tr>
<td>COSA Workflow</td>
<td>SCUFL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eastman EW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InConcert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InTempo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MQSeries Workflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERflowware</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TeamWARE Flow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dolphin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Workflow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WFX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BizFlow</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discovery Workflow Design: Presenting Process Knowledge

- Capture of theories, analytical processes in science and business and the information relationships among processes
- Integrating of data and applications in a process
- Representing integrative knowledge
- Sharing and reuse of workflows through templates
- Organizing and planning complex processes in science and business
- Capture provenance and auditable history of processes
- Management and deployment of workflow for sharing process knowledge
Discovery Workflow Technology: Towards Compositional Services

Workflow Authoring
Composing Services

Workflow Warehousing

Workflow Management
Collaborative Knowledge Management

Resource Mapping

Workflow Execution
A Compositional GRID

Workflow Deployment:
Grid Service and Portal
Discovery Workflow: Language Issues

Language for informatics process---rich data model
- Data Access
- Data Cleaning/Transformation/Processing
- Data Analysis

Language for open informatics process---integration capacity
- Integration of data resource
- Integration of applications
- Integration of services

Language for open informatics process management---integrative knowledge representation and management
- Workflow for integrative knowledge representation
- Rich meta data model for process provenance
- Rich operations over workflow for abstraction, composition, storing, searching and deploying workflow

Language for distributed service computing---support grid computing model
- Composing distributed services
- Mapping a workflow to distributed resources
Support easy end user access
- Powerful visual workflow editor and visual analysis
- Automatic workflow capturing technology

Support collaborative work
- Groupware support for workflow construction

Support enterprise infrastructure
- J2EE compliant workflow middleware
- Oracle based workflow management
- Web service integration and deployment

Support grid computing model
- Automatic mapping workflow to grid resources
- Compositing grid services using workflow
- Optimal scheduling of workflow execution on a grid environment
Workflow for Information Process

Rich Data Models: Table, text, sequence, stream, image, XML.
Power of Rich Data Model: Workflow for Multi-Modality Analysis

- Data mining
- Domain Maps
- Relationship Functions
- Metabonomic Analysis
- Spectrum data mining
- Text mining
- chemical/sequence data model
Workflow = Compositional Process by Dynamic Application/Service Integration
Integrating an Application: Action Abstraction

- Applications/Services
- Functional Abstraction (parameters+meta data)
- Provenance Abstraction (history+control protocol)
- Data Abstraction (data type mapping)
Deploying a Composed Application
Workflow Parameterisation

- **Functional Abstraction** (parameters + meta data)
- **Provenance Abstraction** (history + control protocol)
- **Data Abstraction** (data type mapping)

**Wizard**

**Super node**

**Web service**
An Oracle 10g Example
(The Lymphoma Example)

**Generate and Compile Code**
Automatically create the Java code needed to build analytical pipelines inside the database.

**Accuracy Testing**
Test the model on the data set of interest.

**Choose A Build**
Build a classification model.

**View Model**
Naïve Bayes has built a model that distinguishes DLBC from Follicular with 77% accuracy.

**Choose algorithm/parameter**
Select the classification model.

**Feature Selection**
Attribute Importance identifies genes correlated with Lymphoma subtypes.

**Sequence Search**
Access XML schema using XML Spy (XML editor) which connects to the database using WebDAV.
Workflow with Oracle 10g
Seamlessly Integrated

Pre-processing

Naïve Bayes Mining and Evaluation

Adaptive NB Mining and Evaluation

Naïve Bayes Mining, Evaluation & BLAST

Apply Classification
Workflow for Knowledge Integration = Dynamically construction of schema to organise related cross-domain analysis results and background knowledge

Towards a Knowledge Schema Framework for integrative knowledge

– **A Mechanism of indexing and cross-annotating related analytical results:** workflow as a schema for integrative knowledge: representing related knowledge by their generation process.

– **Workflow based knowledge management:** workflow indexing, workflow ontology, workflow provenance form the base for building a process knowledge base.
Workflow = execution plan of distributed service computing

Grid computing mapping = automatic mapping and scheduling workflow over distributed resources

Discovery workflow can be mapped to various scheduling systems: LSF, Sun GridEngine, Unicore and Condor (DAGman)

Discovery workflow can be deployed as OGSA compliant grid services
Workflow Deployment

- workflow parameterisation

- Define actions
- Deploy

Volcano plot

- Significance threshold
- Effect threshold
- Plot dot size
- View scatter plot
- View table of interesting genes

Sample size
Deploying Workflow as New Application/Service

Volcano plot process deployed for interactive web analysis.

**Parameters**
- Sample: 500
- Effect threshold: 0.4
- Significance threshold: 0.5
- Dot size: 2

**Actions**
- Interesting genes
- Volcano plot

**Result**
Volcano Plot
Discovery Workflow: Programming the Grids

Scientific Information
- Literature
- Databases
- Operational Data
- Images
- Instrument Data

Discovery Workflow
- Real Time Data Integration
- Discovery Services
- Dynamic Application Integration
- Integrative Knowledge Management

Using Distributed Resources

Scientific Discovery
Compositional Services for SARS Mutation Analysis

- 50 data resource, with scaling up to 1000’s
- > 200 software applications and services
- Designed on top of the Web service environment
- Used by more than 100 scientists for SARS analysis
Service Abstraction and Workflow Deployment
Service Abstraction by Parameterizing Workflow
Executing Deployed Service through Portal
Further Composing of Deployed Services
• Discovery Net is developing an advanced scientific workflow technology
• Discovery workflow is more than just a scientific workflow system. It offers much more:
  – Discovery workflow enables dynamic build new services by integrating data/application/service—a powerful EAI (enterprise application integration) tool
  – Discovery workflow is a uniform mean for integrative knowledge representation and management
  – *Discovery workflow provides a systematic mechanism of mapping compositional services over distributed resources*
• Discovery workflow: towards a language of programming the GRID