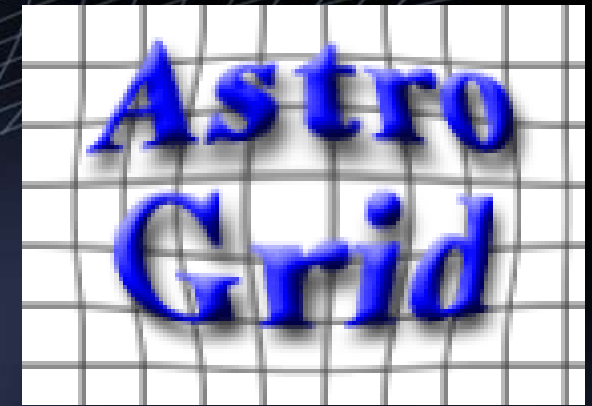


1st AstroGrid Science Workshop



AstroGrid Workflow
Noel Winstanley
Jodrell Bank



Definition

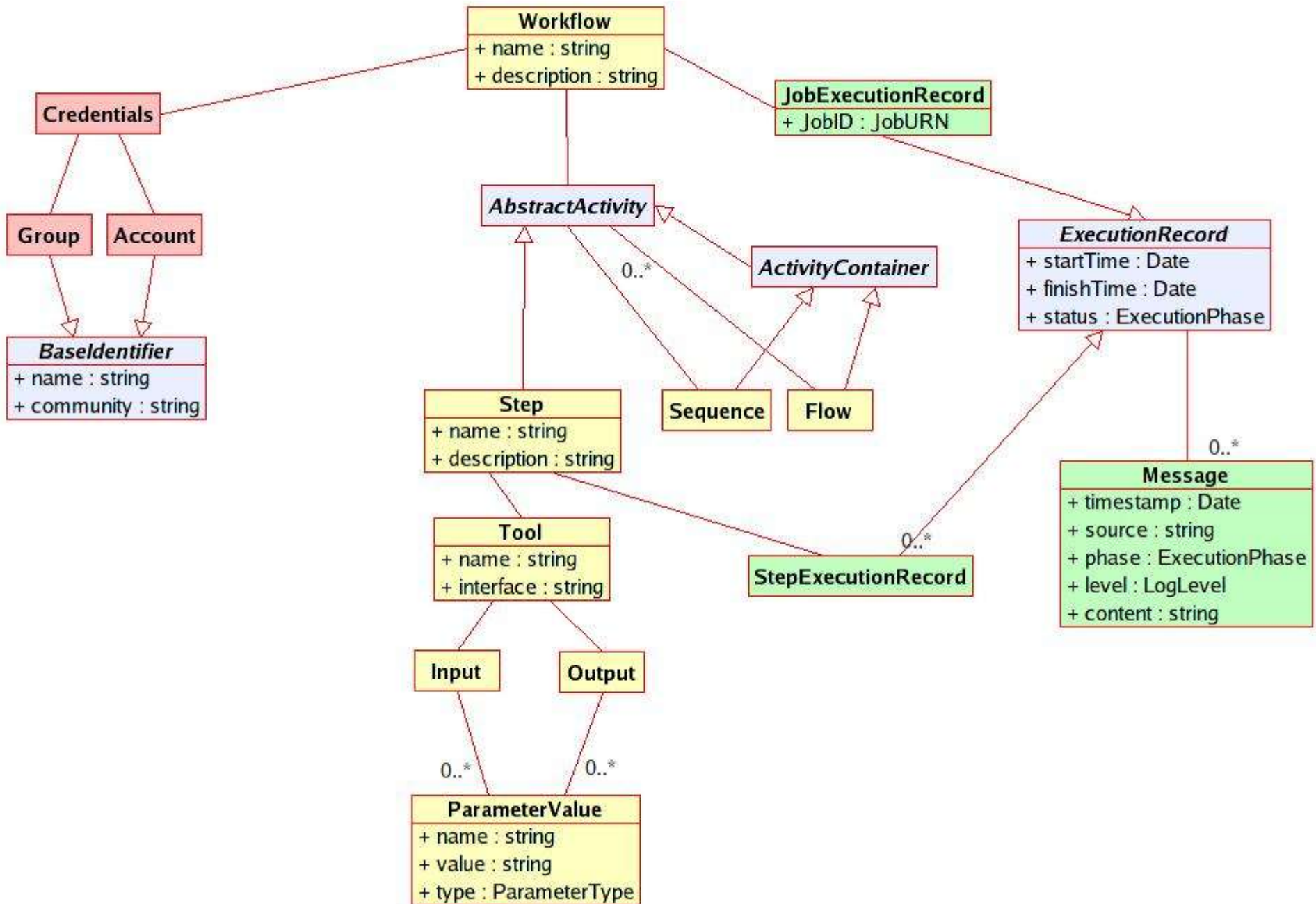
- a workflow performs a complex piece of work
- work requires more than one step
- each step invokes a separate computer program
 - distributed execution of a CEA Application
- step has input and output parameters.
 - inputs – control and astronomical data
 - outputs – final results, or intermediate results for input to a subsequent step
- parameter values can be
 - inline (direct parameters)
 - references to external resources (indirect parameters)
 - `http://...`, `ftp://...`, `ivo://...`
 - `not file://`

Ivorns

- Uniquely define a resource in a VO Registry
- Used in workflow to refer to
 - CEA applications
 - Myspace files
 - also used for users / catalogues / services
- `ivo://<authority>/<name>#<more data>`
- `ivo://org.astrogrid/galaxev` – cea app
- `ivo://uk.ac.le.star/filestore-001` – myspace filestore
- `ivo://uk.ac.le.star/noelwinstanley` – me
- `ivo://uk.ac.le.star/noelwinstanley#votable/results.vot`
 - a file in my myspace

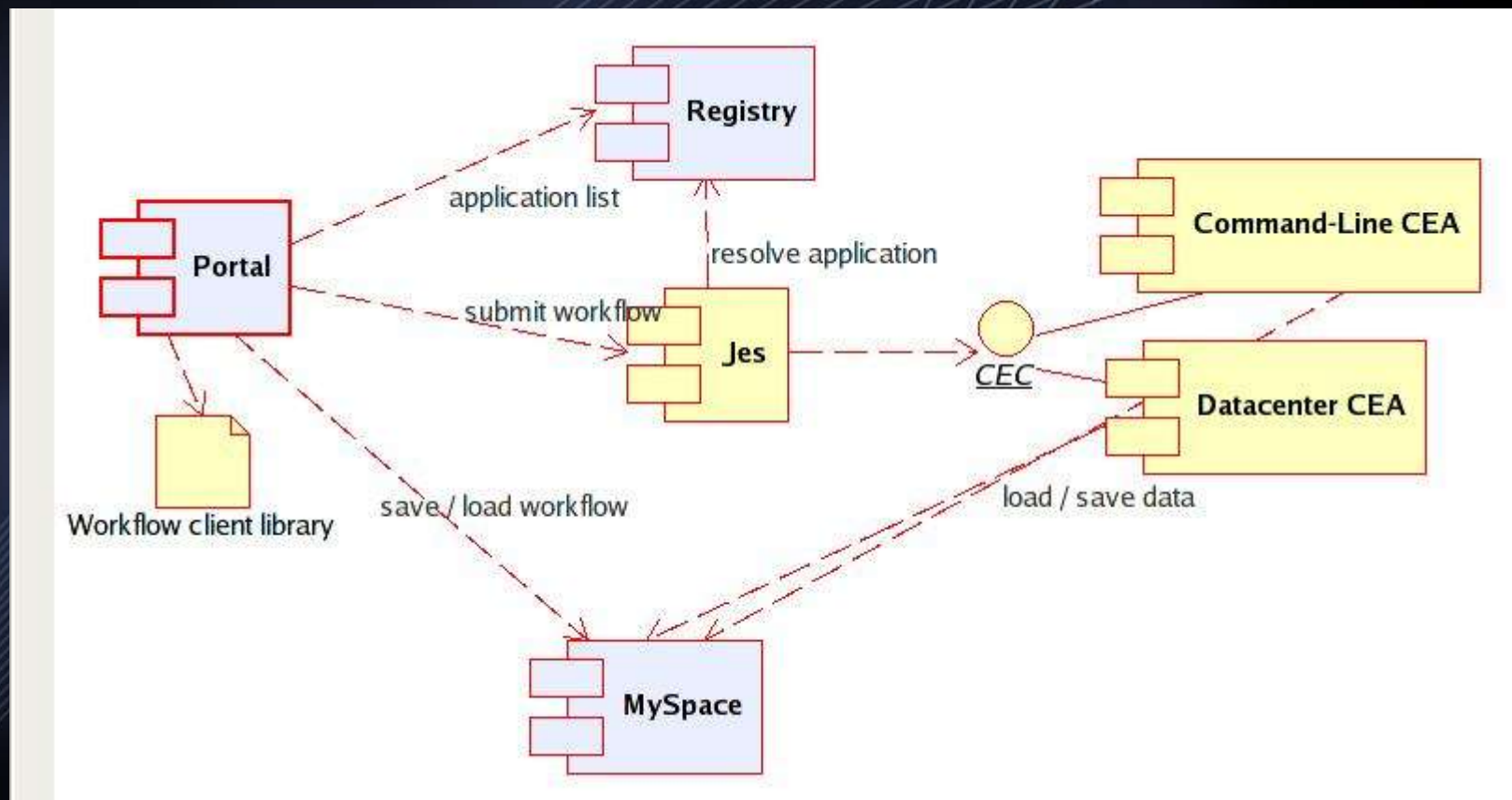
Structure

- xml – verbose but rigourous
- descriptive info
- steps to execute
- control structures
- scripts
- execution traces
 - logs generated from each step
 - results of steps (if required)



Executing Workflows

- JES – Job Execution System
- submitting a workflow to JES creates a job
- globally unique job ID is returned
- JES schedules & executes the workflow
 - decides which CEA servers to place application calls on
 - records log & results of application execution
 - executes workflow control structures to decide which steps to run
 - executes workflow scripts.



Structures

- Astrogrid Workflows can be as complex as needed
- Transformation of intermediate results
- Iteration and Conditionals
- Decisions based on intermediate results
- Expression language
- Embedded scripting

Why so complex?

- maybe linear workflow isn't enough to do real science
- even linear workflows may need to transform data
 - extract columns from intermediate results
 - arithmetic
 - mangle formats of dates, etc

Workflow Variables

- named variable that refers to an expression
- can be referenced from other bits of the workflow using \$ {..} notation
- can be assigned-to from results of CEA application executions.
- can be referenced /assigned-to from script blocks

Script

- manipulate results of steps
- language is groovy – (<http://groovy.codehaus.org>) javascript style, with access to java libraries.
- can call java libraries

Scripting Environment

- Predefined libraries to do common tasks
- read and write to myspace
- manipulate table formats
- all libraries of java available (networking, xml, etc)

Flow and Sequence

- Sequence defines a series of activities, that are executed one after another.
- Flow defines a set of activities that are executed simultaneously.
- Flow achieves speedup, but can't be data dependencies between the activities.

Conditional

- execute a block of activities only if a test is true
- may have 'then' and 'else' parts
- test is an script expression – may refer to workflow variables.
- show example

While Loop

- repeat a block of activities while a test is true
- again, test is a script expression
- example

For Loop

- repeat a block of activities for each item in a list
- for each iteration, a loop variable provides the value of the current item.
- possible to define a list, numerical sequence, or use a list stored in a workflow variable.

Parallel For

- execute a block of activities for each item in a list, simultaneously.
- introduces parallel execution
 - can't be any inter-dependencies between items on the list (e.g. shared workflow variables, shared resources (writing to same file))
- leads to concurrent cea applications running
 - same task may run (for different list items) on different cea servers.

Error-Handling

- future research.

References

- Workflow Documentation.
 - <http://wiki.astrogrid.org/bin/view/Astrogrid/BuildingWorkflow>