Getting a grip on the grid:

A knowledge base to trace grid experiments

Ammar Benabdelkader
ammarb@nikhef.nl

Mark Santcroos
m.a.santcroos@amc.uva.nl

Victor Guevara Masis
vguevara@nikhef.nl

Souley Madougou
souleym@nikhef.nl

Antoine van Kampen
a.h.vankampen@amc.uva.nl

Silvia Olabarriaga
S.D.Olabarriaga@amc.uva.nl

BiG Grid
the dutch e-science grid

Presentation Outlines

- Background, challenges and Focus
- Provenance: an overview
- Provenance API (Plier):
  - Database schema
  - Architecture & Implementation
- eBioCrawler
  - Abstract/concrete graph
  - Challenges
- Plier Toolbox:
  - Generic functionalities
  - Customized functionalities
- Scientific Impact
- Conclusion & future work

BiG Grid
the dutch e-science grid

amc

Big Grid (Dutch NGI)

- Founding partners: NCF, Nikhef and NBIC (2007-2011)
- Mission:

  To realise a fully operational world-class and resources-rich grid environment at the national level in the Netherlands to serve public scientific research, including particle physics, life sciences and all other disciplines, and to encourage actively general grid usage across all disciplines.

- Details:
  - Ca. 25% for “user support” and “application-specific support”
  - Ca. 50% for “hardware infrastructure”
  - Ca. 25% for “running costs”
- Focus:
  - Grid: networking, compute, storage (resources) , databases, sensors, backup, ...
  - e-science: conducting science, using all kinds of ICT infrastructure and opportunities

BiG Grid
the dutch e-science grid

amc

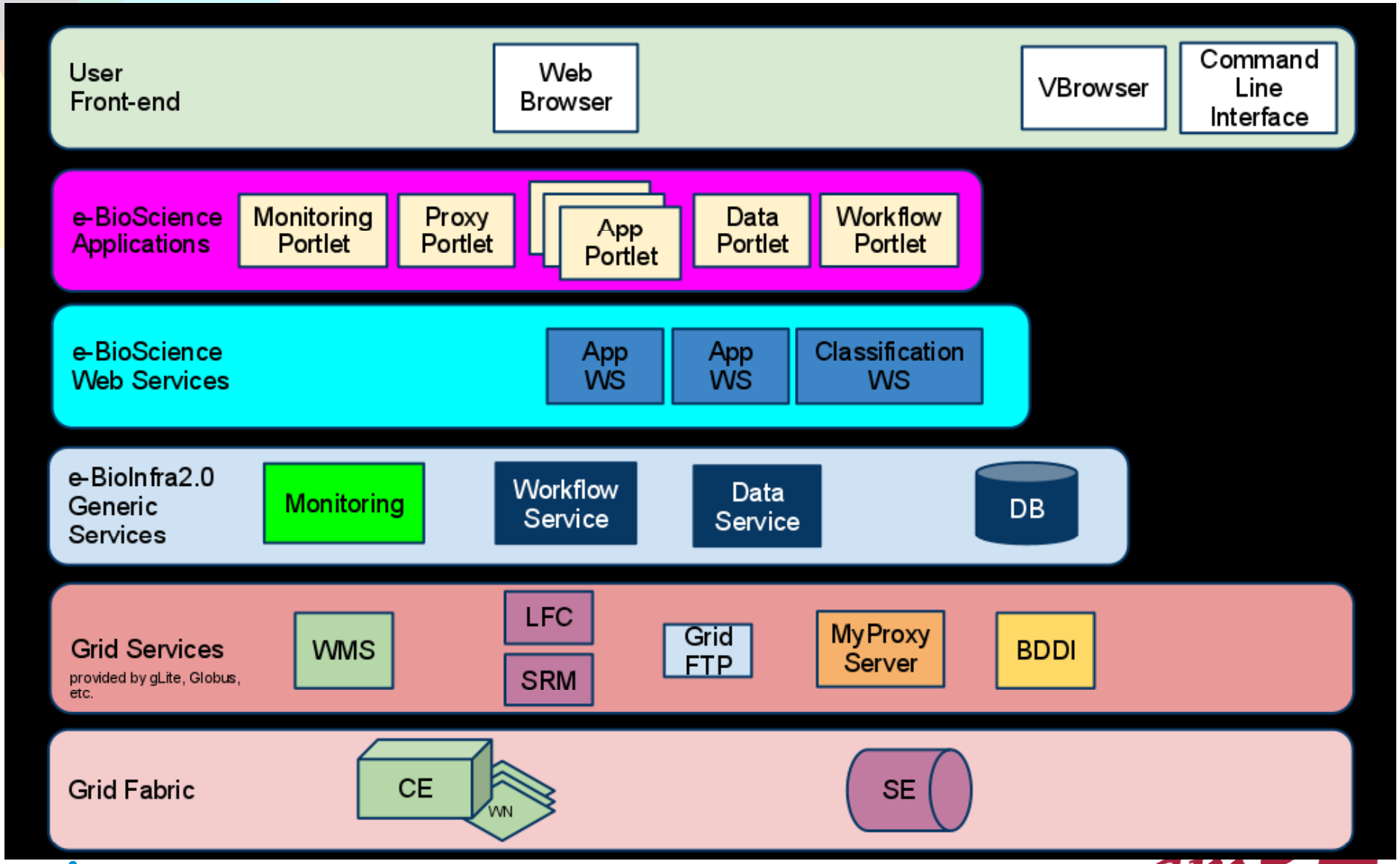
AMC: e-BioScience Group

- Bioinformatics Laboratory
  - Dept. Clinical Epidemiology, Biostatistics and Bioinformatics
  - Academic Medical Centre, University of Amsterdam
- Filling “gap” between medical researchers and the Dutch NGI
- Supporting a wide range of applications
  - Next Generation Sequencing
  - Medical Imaging
  - -Omics

BiG Grid
the dutch e-science grid

amc

e-BioScience Group: Layered Architecture



BiG Grid
the dutch e-science grid

Background

- To run their experiment, e-BioScience group deploys:
  - Moteur2/DIANE Workflow engine, and
  - GWENDIA (Grid Workflow Efficient Enactment for Data Intensive Applications)
- Most experiments are complex due to:
  - Iteration over input parameters of running experiments
    - Each job is instantiated several times according to the number input data links.
  - Re-trial of failing process
    - Each failing job gets re-tried until it succeeded (or reaches re-trial limit)
  - Each workflow experiment may consist of a large number of failed and succeeded jobs.

BiG Grid
the dutch e-science grid

amc

Challenges

- Hard to validate workflow experiments:
  - Identify whether an experiment succeeded or failed
  - Verify the validity of the output results
  - Identify the source of failure
- Hard to instrument and document experiments:
  - How to document validated experiments?
  - What to do with failed experiments?
  - How to keep track of the validation process?
  - How to preserve/publish the knowledge and expertise
- Hard to make use of the gained expertise:
  - How to prevent similar sources of failure?
  - How to spread the gained expertise?
  - How to better exploit the gained expertise?

BiG Grid
the dutch e-science grid

amc

Focus

Build a knowledge base to instrument scientific experimentations

- Start with …
  - Building a knowledge base to instruments scientific experimentations
  - Knowledge base should be flexible enough …
- Adopt the Open Provenance Model (OPM) …
  - Better suited to our case, since it provides history of occurrence of things (with flexiblity)
  - Implement tools to build and store OPM-compliant data objects related to scientific experimentations
- Build customized tools to explore the data
- Enhance the database and Toolbox whenever needed.

BiG Grid
the dutch e-science grid

amc

Open Provenance Model (1)

http://openprovenance.org/

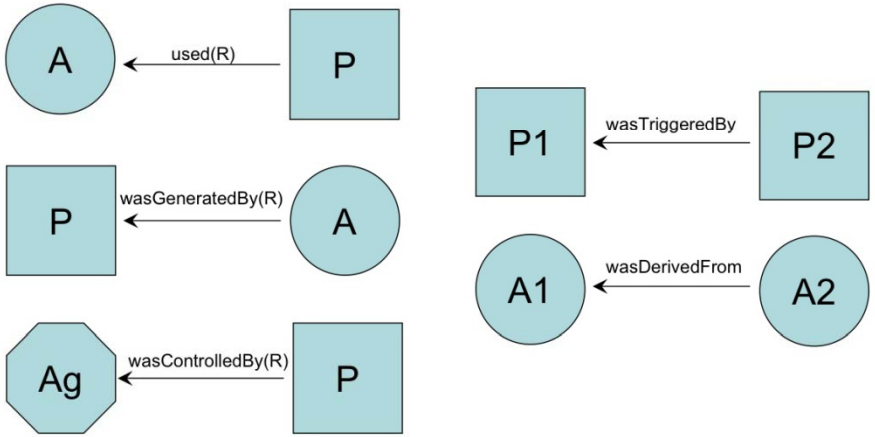
- Allow us to express all the causes of an item
  - e.g., provenance of a scientific experiment includes:
    - Processes composing the experiment
    - Where did the processes run
    - What input they used
    - What results it generates, when and where
    - Who did launch and monitor the experiment
    - Etc.

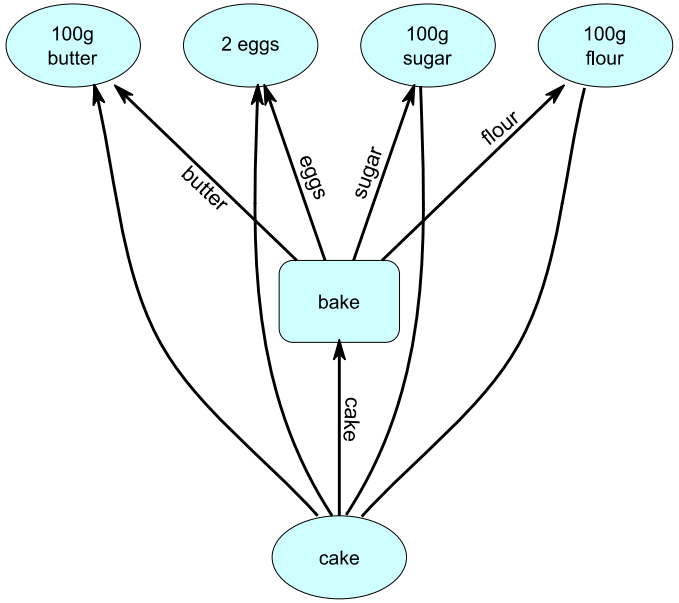
- Allow for process-oriented and dataflow oriented views

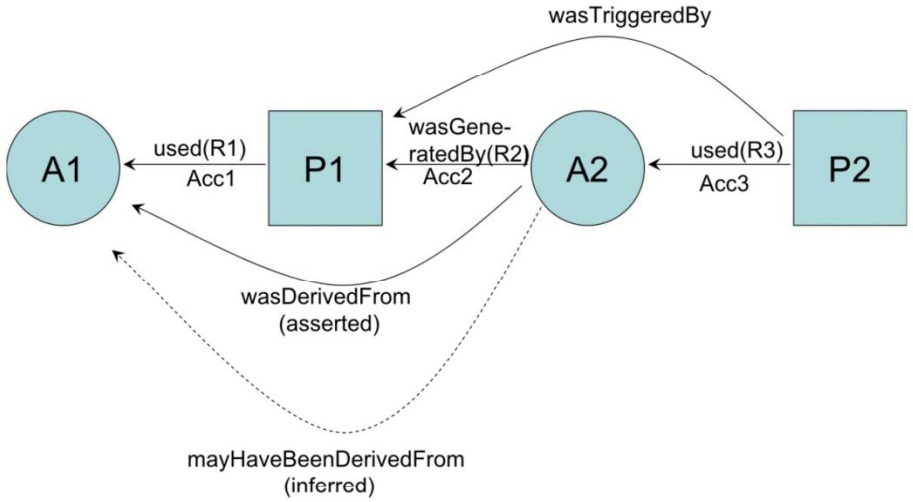
- Based on a notion of annotated causality graph

Open Provenance Model (2)

http://openprovenance.org/







BiG Grid
the dutch e-science grid

PLIER Development

The *Provenance Layer Infrastructure for E-science Resources (PLIER)* provides an implementation of the Open Provenance Model (OPM)

Four main components constitutes the Plier development:

1. Implementing the most optimum OPM-compliant relational database schema
2. Developing the Plier Core API:

   Java-based API to build and store OPM graphs
3. Developing the eBioCrawler:

   Java-based agents that crawls the input/output data for each experiments and stores it into the knowledge base.
4. Developing the Plier Toolbox:

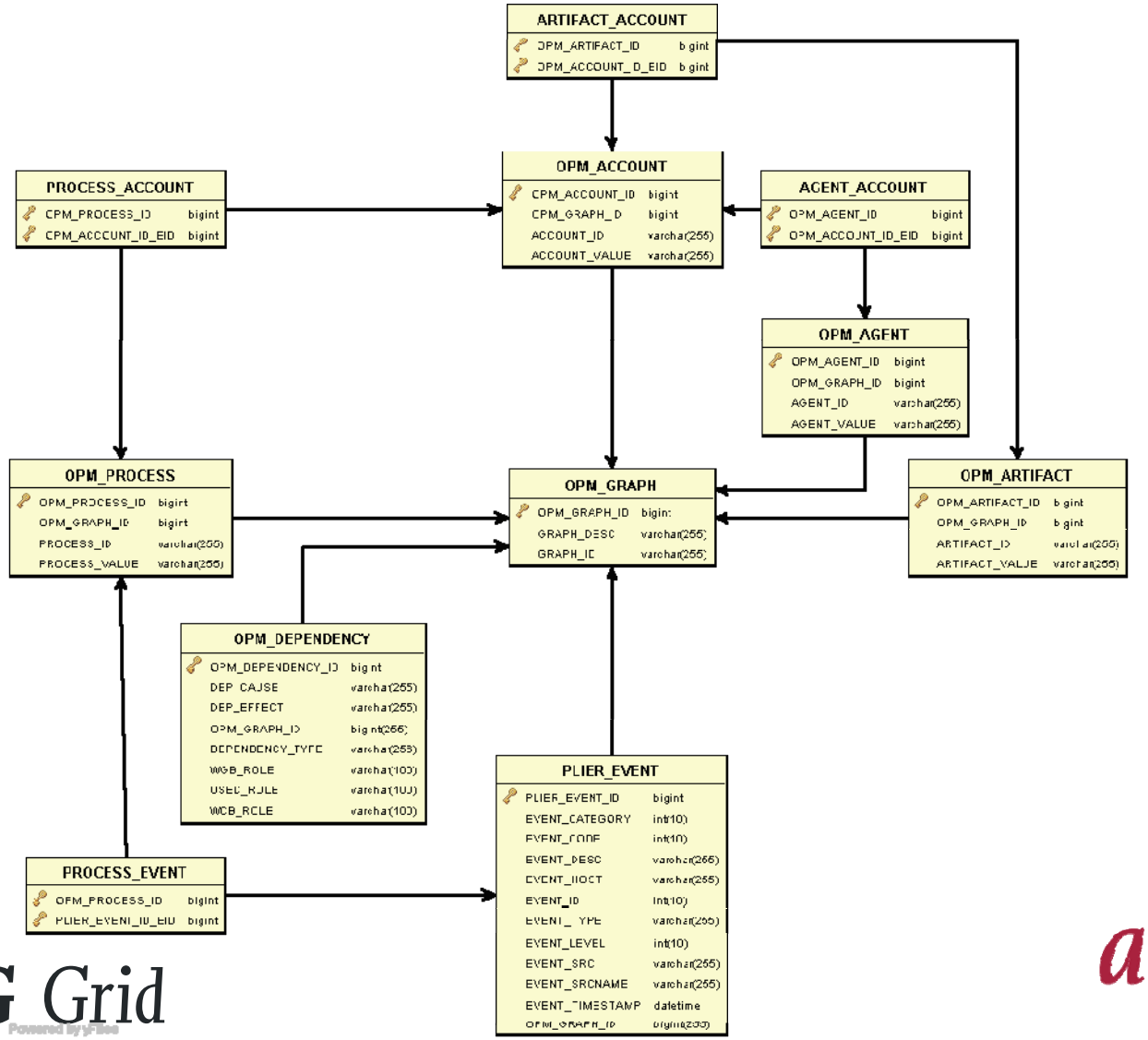
   Java-based UI to visualize, search, and share OPM graphs

BiG Grid
the dutch e-science grid

amc

PLIER: Database Schema

OPM compliant database schema used by Plier:



BiG Grid
the dutch e-science grid

amc

PLIER: Core API (1)

Plier API is implemented using most recent standards and mechanisms:

1. JDO 3.1 is used as a java-centric API to access persistent data,
2. DataNucleus is used as a reference implementation of the JDO API,
3. MySQL is used as a back-end database to store provenance data

Plier Core API provides means to build OPM-compliant data objects and store them into the knowledge base

BiG Grid
the dutch e-science grid

amc

PLIER: Core API (2)

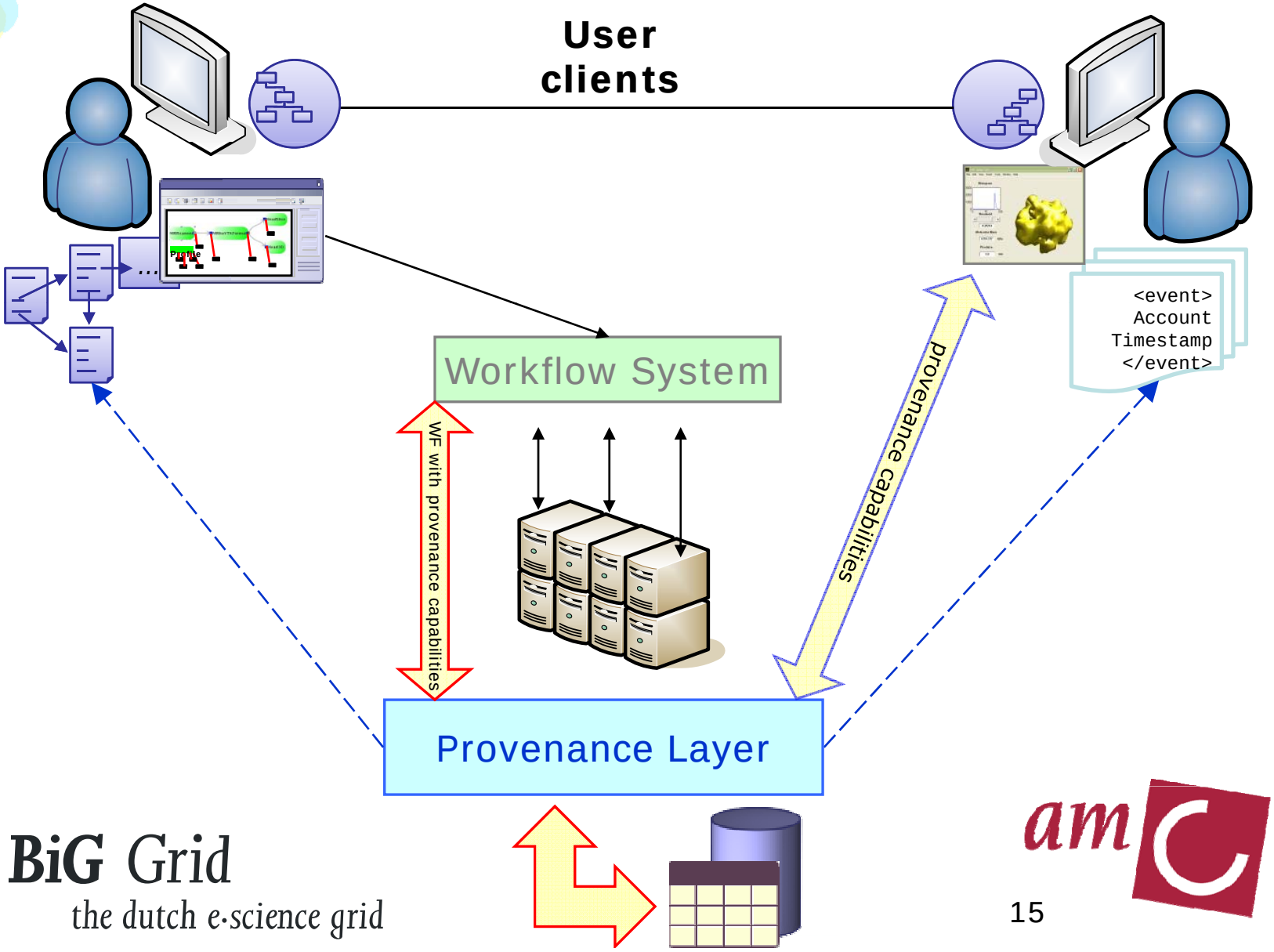
Plier API can be used in two manners:

1. Integrated within the workflow management system (WF with data provenance capabilities)
   - Scientists only need to enable the data provenance capabilities from the WF.
   - WF developers need to implement the DPC inside the workflow engine.
2. Implement the provenance data based on the input/output used/generated by the workflow system:
   - No need to change the workflow engine.
   - You may risk to build incomplete OPM graphs

BiG Grid
the dutch e-science grid

amc

PLIER: Core API (3)



BiG Grid
the dutch e-science grid

eBioCrawler

Java-based agents that crawls the input/output data for each experiment and stores it into the knowledge base.

- Uses GWENDIA workflow description to build the abstract model of the experiment.
- Uses other input/output/log files to build the concrete model of the experiment.
- Workflow experiment data available through secure https server

RISK:

of not being able to collect/extract the required minimum data set of each experiment.

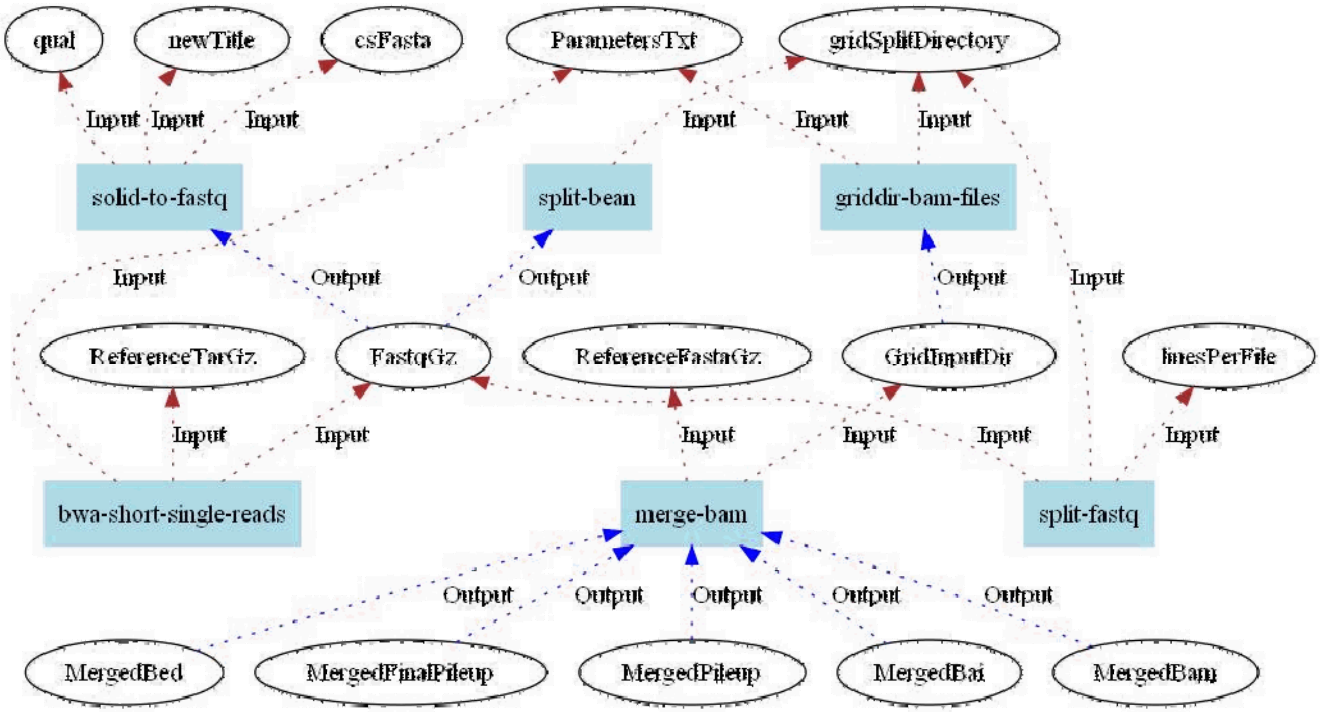
BiG Grid
the dutch e-science grid

amc

eBioCrawler: Abstract Graph

Extracted from the workflow description (GWENDIA XML format)

→ Straight forward process



workflow-0b846ce3 (Abstract)

BiG Grid
the dutch e-science grid

eBioCrawler: Concrete Graph

Extracted from the different input/output/log ,
used/generated by the workflow engine

complex process …

For each workflow experiments

- Users and host machines are modelled as AGENTs
- Executed Jobs are modelled as PROCESSes
- Input files/parameters are modelled as ARTIFACTs
- Output results are also modelled as ARTIFACTs
- Nodes are linked using CAUSAL DEPENDENCYies

BiG Grid
the dutch e-science grid

eBioCrawler: Concrete Graph

Major issues, we faced:

- Re-tried processes causes data duplication, mainly with input files, which results in heavy graphs
- It was hard to identify input files/parameters for each job (values and order)
- Output results were hard to link to their corresponding processes

→ Most of the issues were solved by dedicating more programming efforts into eBioCrawler

BiG Grid
the dutch e-science grid

amc

eBioCrawler: a success!

The approach was very successful:

- At first, eBioCrawler was able to collect about 70% of the required information
- With additional programming efforts into eBioCrawler we were able to collect more than 95% of the required information

→ This work is being used as a proof of concept to validate the suitability of the OPM model to our case

BiG Grid
the dutch e-science grid

amc

PLIER Toolbox

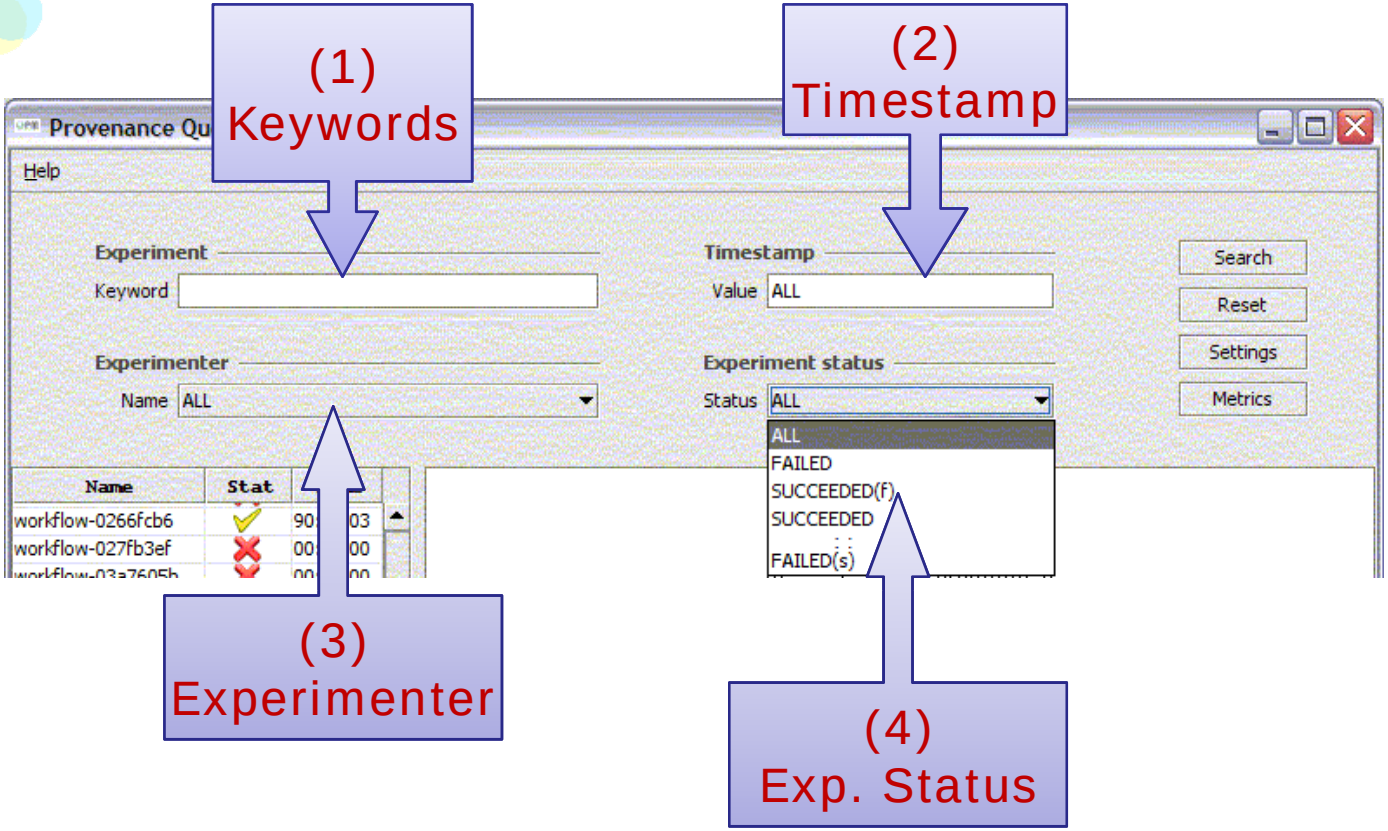
Java-based UI to visualize, search, and share OPM graphs

1. Provide general functionalities like; summary of experiments with their status, execution time, etc.
2. Provide search functionalities based on keywords, user, date/time, status of experiment, etc.
3. Provide detailed information about each experiment/graph (e.g. input/output parameters, events, processes, etc.)
4. Provide OPMX (OPM XML format) and DOT (graphviz) data related to each experiment
5. Customized functionalities could be added to the interface (e.g. detailed report, analysis of output data, etc.)

BiG Grid
the dutch e-science grid

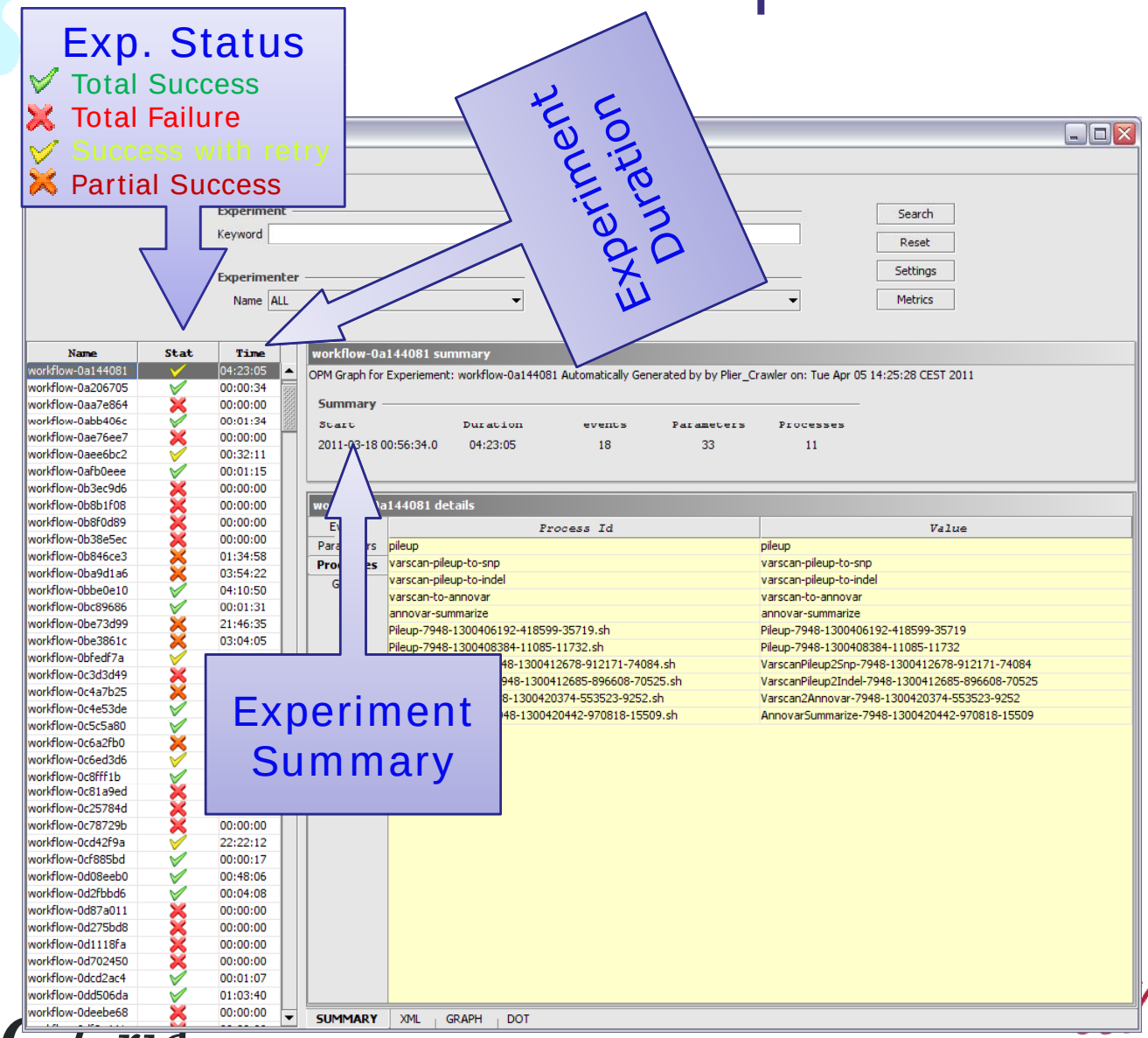
amc

PLIER Toolbox- Search Menu



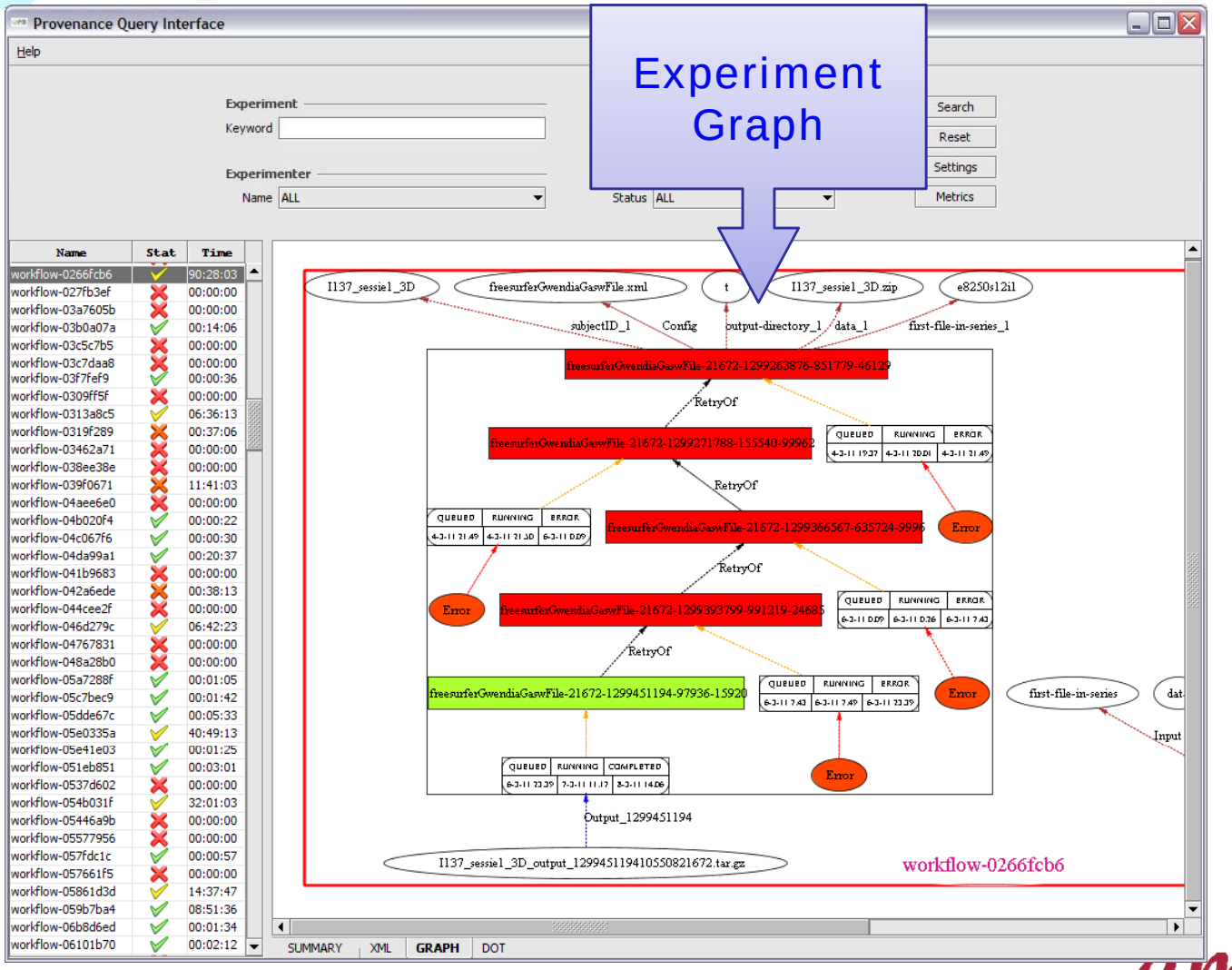
BiG Grid
the dutch e-science grid

PLIER Toolbox- Exp. Summary



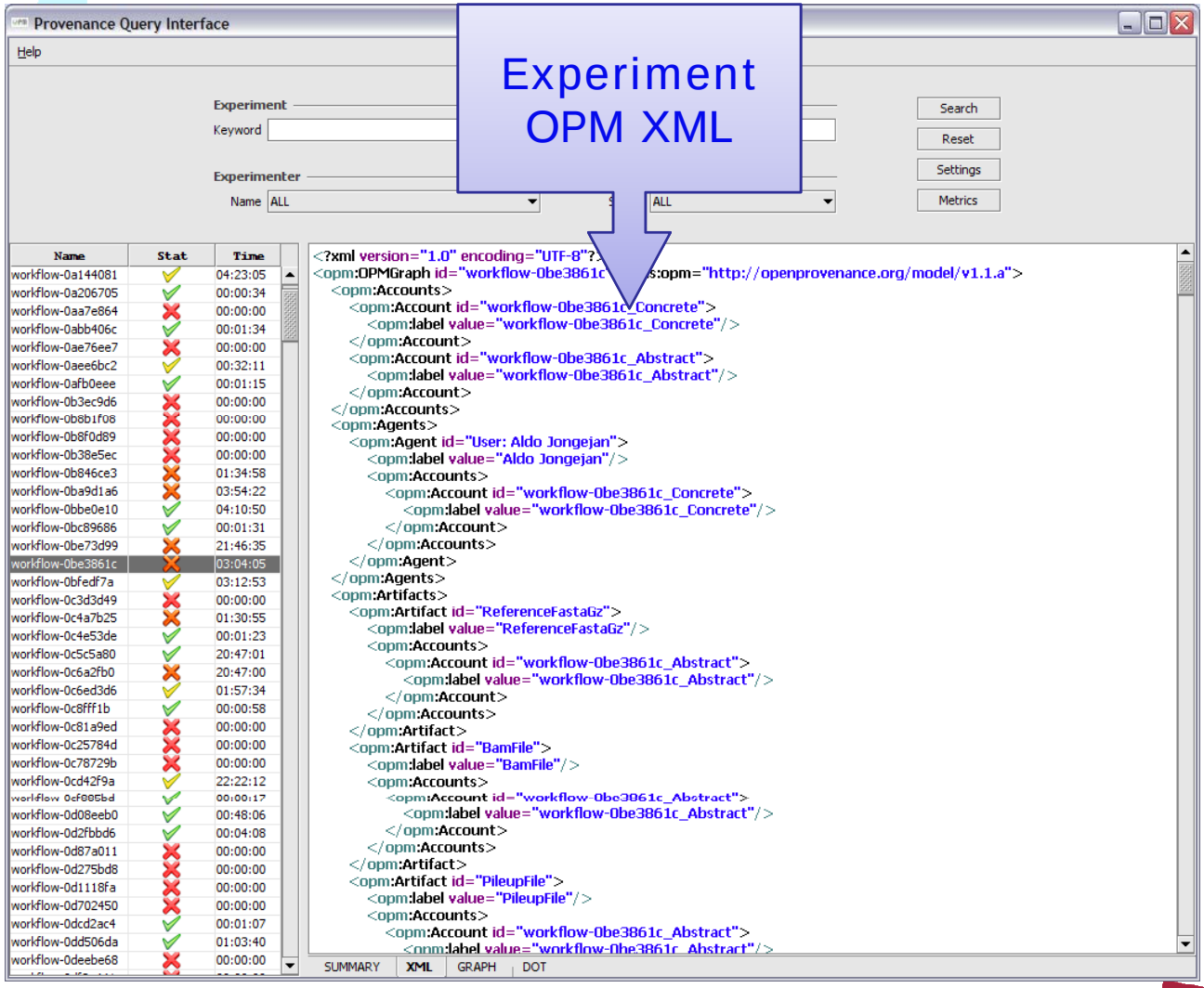
BiG Grid
the dutch e-science grid

PLIER Toolbox – Exp. Graph



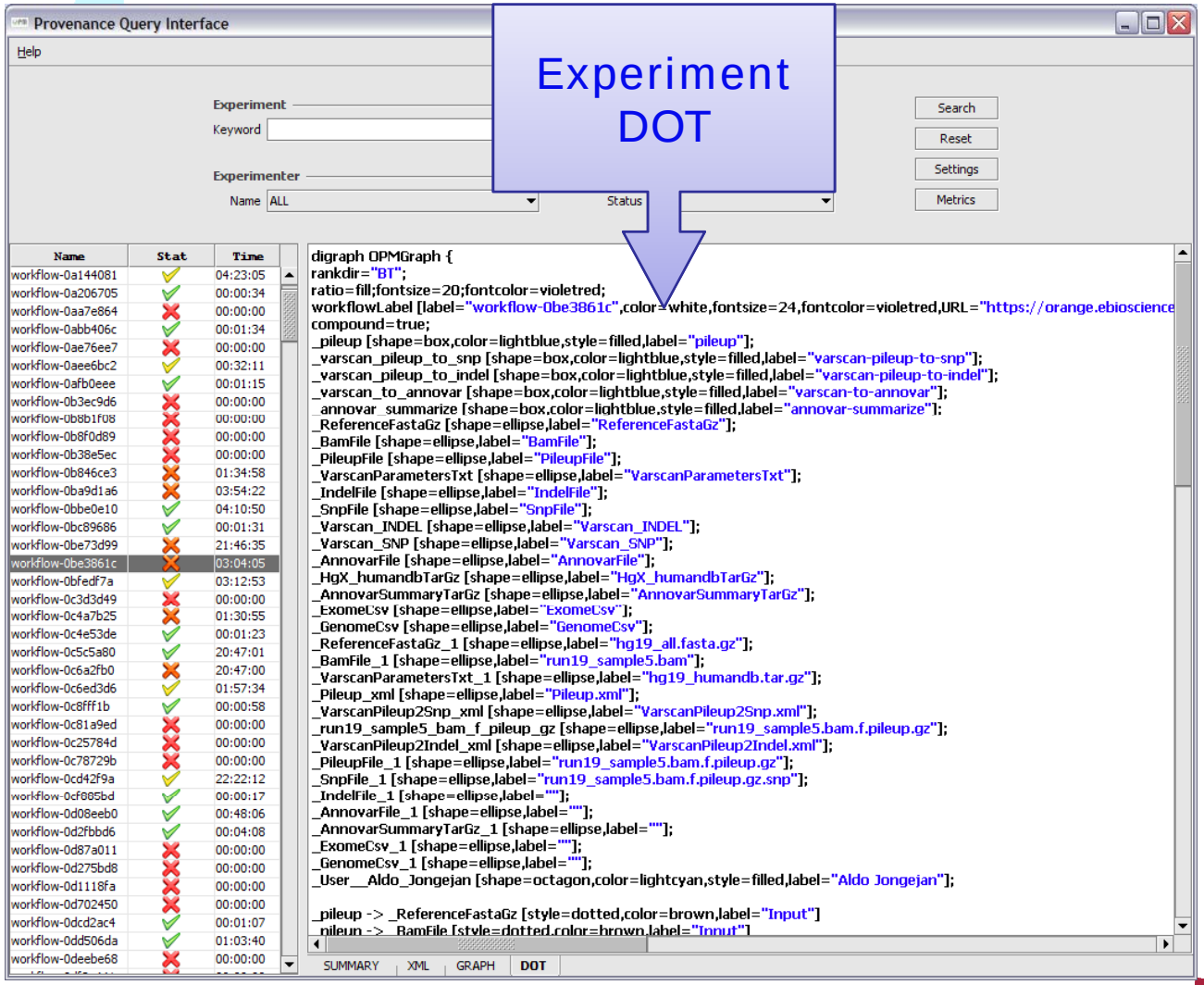
BiG Grid
the dutch e-science grid

PLIER Toolbox – Exp. OPMX



BiG Grid
the dutch e-science grid

PLIER Toolbox – Exp. DOT

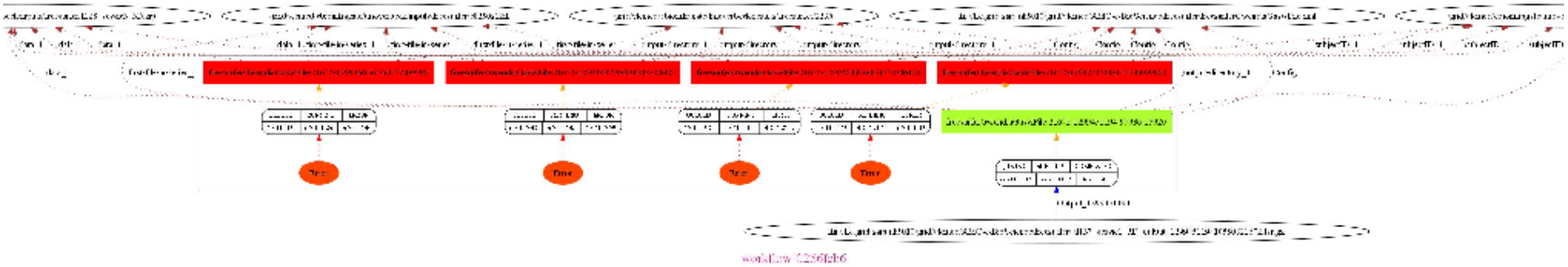


BiG Grid
the dutch e-science grid

Graph Customization

Initial Graph

Long
path names

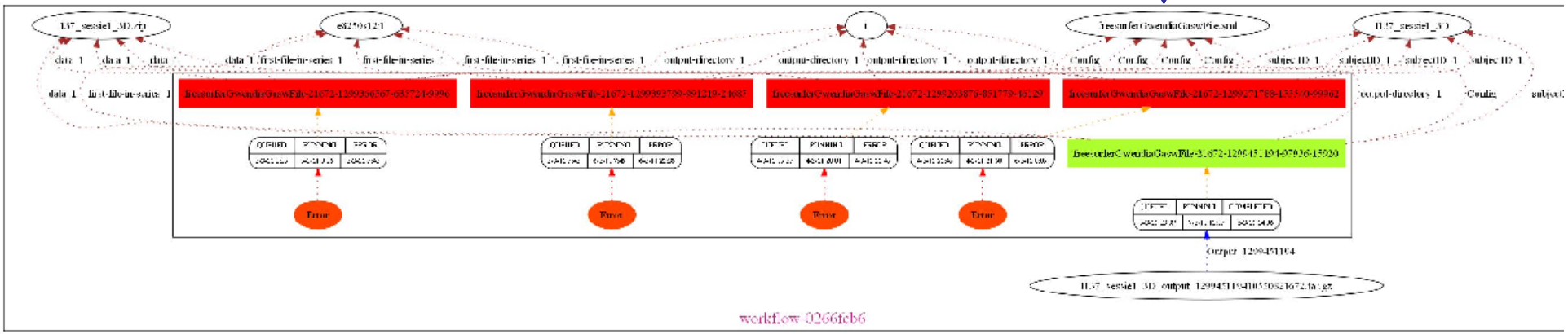


BiG Grid
the dutch e-science grid

Graph Customization

Hiding artifact Path

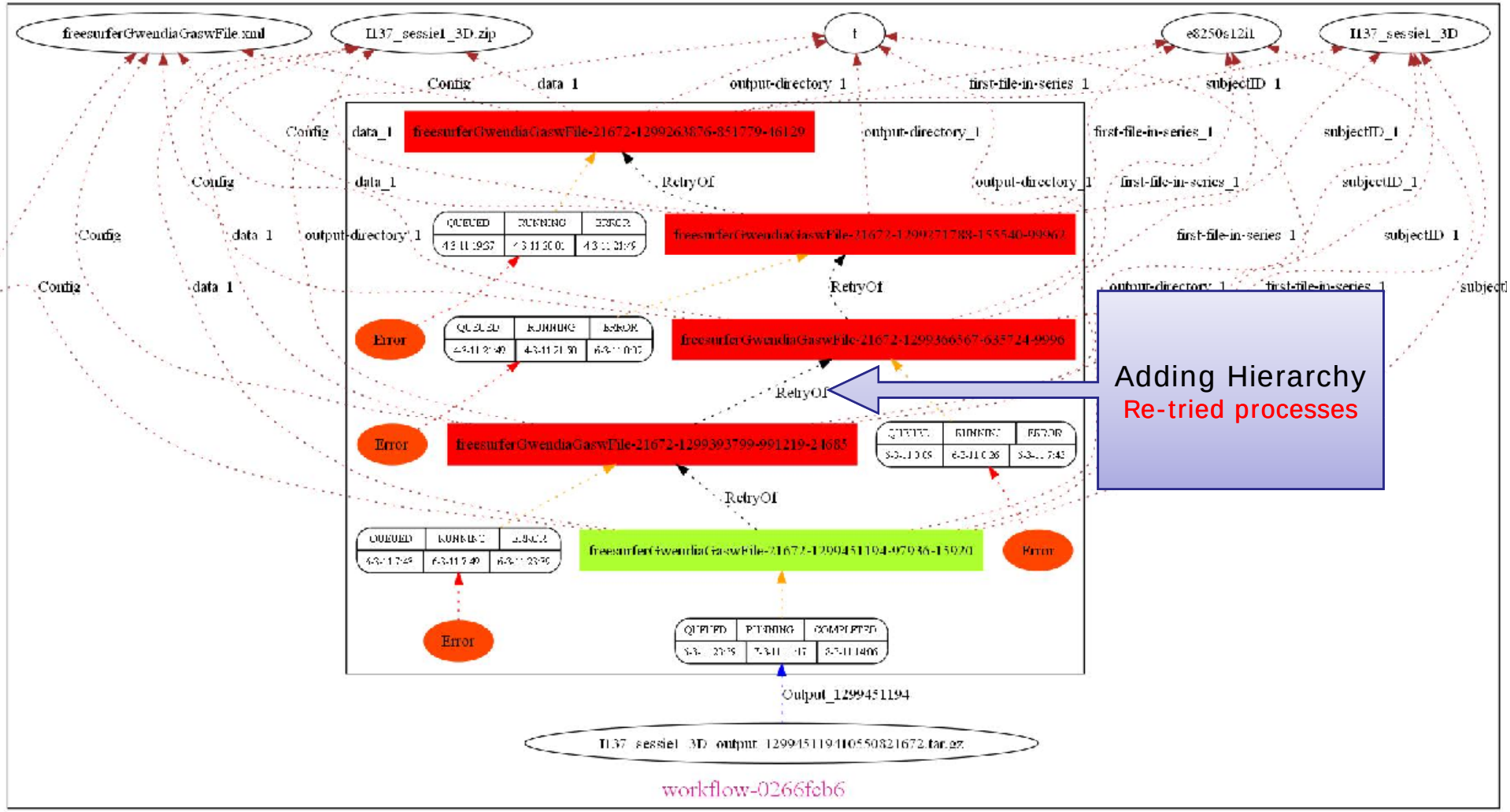
Hiding the path name



BiG Grid
the dutch e-science grid

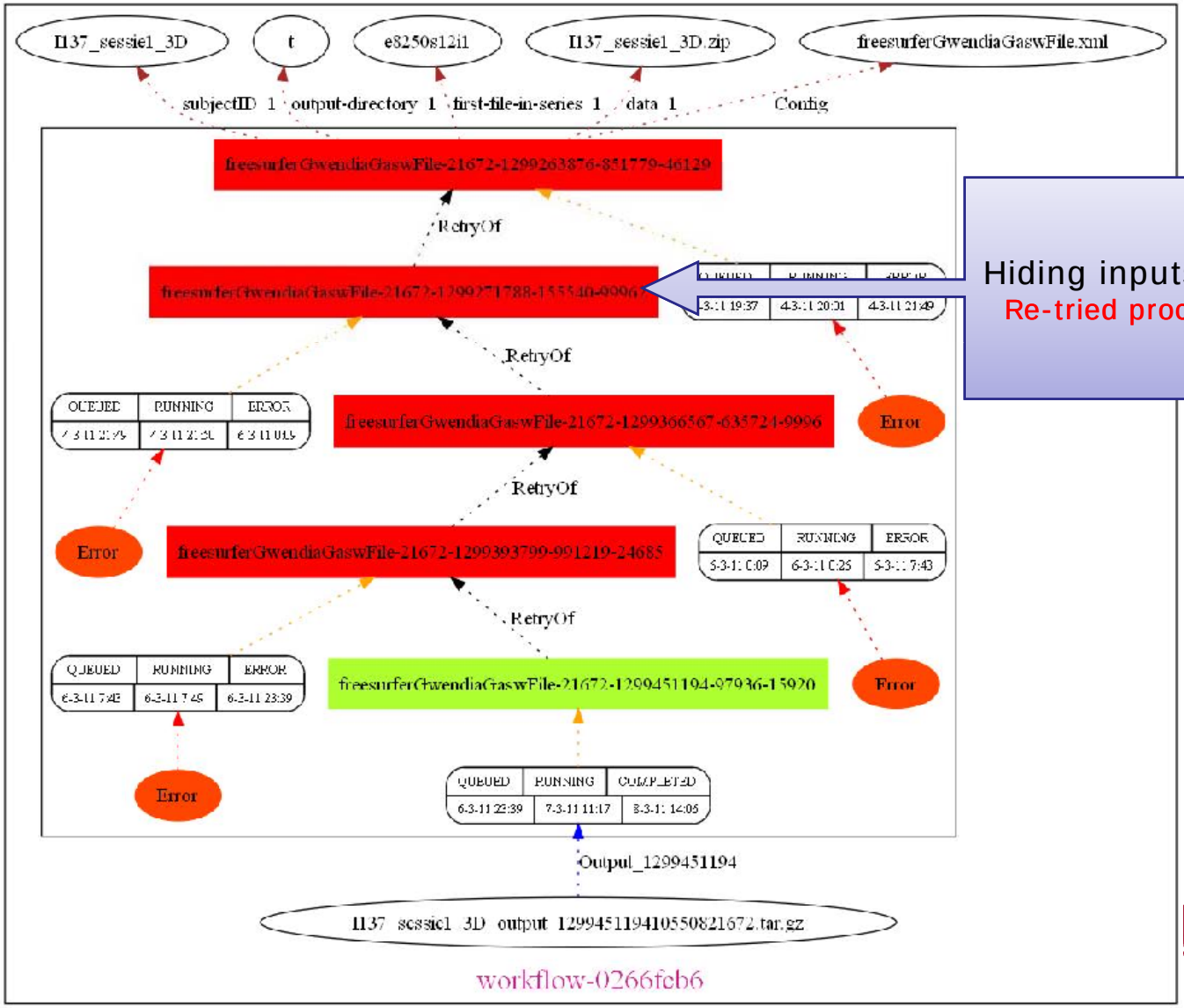
Graph Customization

Adding process hierarchy



Graph Customization

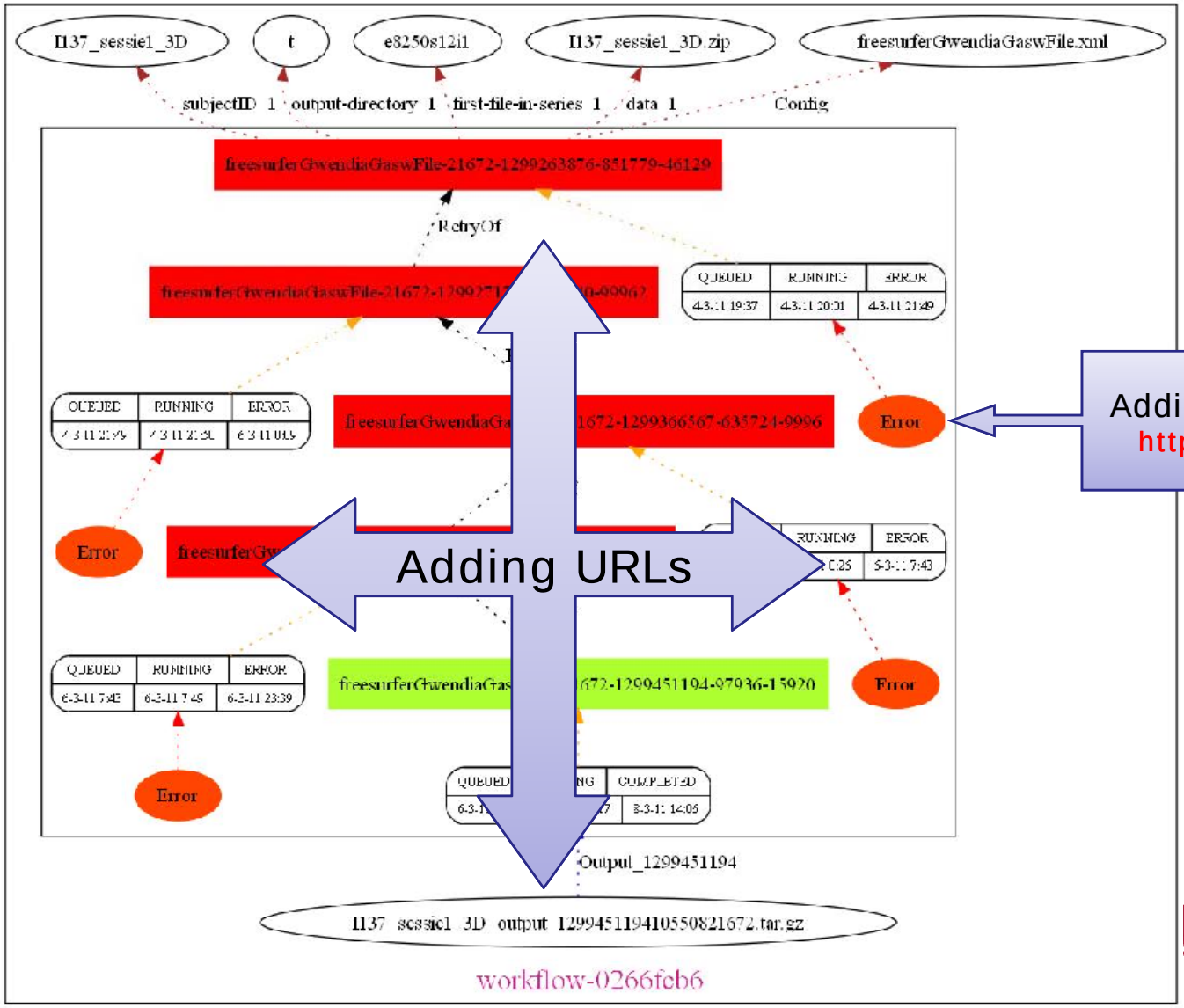
Hiding duplicate inputs



Hiding inputs from
Re-tried processes

Graph Customization

Adding URL Links



Conclusion and Future Work

Usefulness - we were able to :

- identify final status of experiments (5 status)
- easily trace the source of error
- identify the reason for error occurrence
- decide what to do with failing jobs
- clean the grid storage (outputs of failing exp.)
- Etc.

BiG Grid
the dutch e-science grid

Conclusion and Future Work

- Enhance Plier API to OPM core specifications (v1.1)
- Implement the provenance model into the Moteur workflow engine
- Enhance the data management Toolbox with additional components:
  - Improve the search criteria
  - Documenting, annotating, reviewing and publishing experiments
  - Fully automate the process of validating experiments
- Extend the usage of the data management Toolbox to other groups

BiG Grid
the dutch e-science grid

amc

Acknowledgments

- **Big Grid**

  This work is part of the program of Big Grid, the Dutch e-Science Grid, which is financially supported by the Netherlands Organization for Scientific Research, NWO.

- **AMC**

  E-BioScience Group,

- **Modalis Team**

  Developers of MOTEUR WS

BiG Grid
the dutch e-science grid

amc

Useful Links

- Plier Core API:
  http://twiki.ipaw.info/bin/view/OPM/Plier
- Plier Toolbox:
  http://twiki.ipaw.info/bin/view/OPM/PlierToolBox
- eBioCrawler:
  http://bioinformaticslaboratory.nl/twiki/bin/view/EBioScience/EBioCrawler
- Open Provenance Model (OPM):
  http://openprovenance.org/
- Moteur:
  http://modalis.i3s.unice.fr/moteur2
- DIANE:
  http://it-proj-diane.web.cern.ch/it-proj-diane/

BiG Grid
the dutch e-science grid

amc