



UvA-DARE (Digital Academic Repository)

Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

Publication date
2011

[Link to publication](#)

Citation for published version (APA):

Altıntaş, İ. (2011). *Collaborative provenance for workflow-driven science and engineering*.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

List of Figures

1.1	Typical components in a collaborative e-Science project (Not shown are the information flow, external data, service and computational infrastructure.) . . .	4
2.1	Parts of the SST MatchUp Workflow: (a) The Build Match Up workflow; (b) The Analysis workflow, (c) Viewing the results in Google Earth.	11
2.2	Steps in the life-cycle of workflow design, execution and experimentation. . .	13
2.3	(a) The configurable Kepler Provenance Recorder acts as a recording component of the Kepler Provenance Architecture. (b) Kepler Provenance Architecture is designed to collect provenance information in different formats in to a provenance store and provides querying API for accessing the collected information.	20
2.4	The reporting and workflow run manager modules in Kepler use the provenance information through a common querying API in Kepler.	21
2.5	Different Distributed Execution Pattern Techniques in Kepler are applied to numerous computational science challenges, including: (a) Map Reduce, and (b) Master-Slave, used in the Ecology domain for House Finch spatial stochastic birth-death process simulation.	22
2.6	Different Parameter Sweep Execution Techniques in Kepler are applied to numerous computational science challenges, including: (a) Nimrod/K used in the Chemistry domain for quantum chemical calculations; and (b) SSH tunneling used in the Physics domain for plasma fusion simulation.	23
2.7	A four-layer reference architecture for scientific workflow management systems. The <i>User Interaction Layer</i> facilitates the interaction of users with the workflow design, execution and monitoring interfaces along with the workflow repositories. The <i>Workflow Execution Layer</i> consist of workflow engines that support workflow scheduling, execution and failure management. The <i>Process Layer</i> supports the user interaction and workflow execution layers with background processes including listeners for provenance, data and task execution management. The <i>Physical Resource Layer</i> provides interfaces to utilize computational and data resources.	25

3.1	A simplified taxonomy of provenance system characteristics in scientific workflows.	30
3.2	Life-cycle of scientific workflow related provenance information from its collection to the analysis of different usages of the collected information.	32
3.3	(a) Edges in the Open Provenance Model (OPM) from effects to causes; and (b) An OPM graph for the “Atlas X Graphic Workflow” implemented for the First Provenance Challenge.	35
4.1	CAMERA 2.0 Architecture provides an adequate separation of concerns and allows external scientific developers to create workflows.	44
4.2	(a) The CAMERA Gamma Diversity Workflow; and (b) The portal interface to run the CAMERA Gamma Diversity workflow.	45
4.3	myExperiment enables researchers build social networks and share workflows.	47
5.1	Different observables of shared data, workflows and workflow executions (runs) in a typical scientific research project, where all (or part of) the output of workflow runs can be used as input to subsequent runs: (a) data $\{d_1, d_2, d_3, d_7\}$ published by users $\{u_3, u_5, u_6\}$; (b) workflows $\{wf_1 \dots wf_5\}$ published by users $\{u_2, u_4, u_5\}$; and (c) flow graph for workflow runs (customized through their parameters) and related provenance data $\{d_1 \dots d_8\}$ in user space $\{u_1, u_2, u_3\}$	51
5.2	Various collaborative provenance views: (a) data dependency view; (b) run dependency view; and (c) user collaboration view (based on Figure 5.1).	52
5.3	A collaborative provenance model where a user can share collaborations with other users when he performs a workflow execution and uses data and workflow published by other users (collaborations are shown with directed dashed lines).	54
5.4	Different user collaboration graph where attributes and edges show the direction and (a) nature (C_N); (b) weight (C_W); and (c) self-collaboration (C_S) aspect, of user collaborations based on the observables in Figure 5.1.	56
5.5	(a) A bottom up view over collaborative provenance attribute combinations from least informative to most informative. (b) Detailed views of possible collaborations between two users.	58
5.6	A collaboration graph where all the attributes of user collaborations are shown based on the observables for the scenario in Figure 5.1.	59
5.7	Acknowledgement view for data artifact d5 based on the observables for the scenario in Figure 5.1.	61
6.1	UML-based model for representing collaborative entities and their relationships (key attributes are underlined).	64

6.2	An example scenario for a typical scientific research project: (a) data ($\{d_1, d_2, d_3\}$) published by users in $\{u_1, u_6\}$.; (b) ready to run workflows ($\{wf_1.. wf_5\}$) published by users in $\{u_2, u_4, u_5\}$.; and (c) flow graph for published workflow runs (customized through their parameters) and related data ($\{d_1.. d_{10}\}$) in user spaces ($\{u_1, u_2, u_3\}$), separated by horizontal lines.	66
6.3	(a) Combined workflow run graph that shows the flow of data through different workflow runs, and (b) the complete data dependency view, based on the scenario in Figure 6.2.	67
6.4	Relational tables of the provenance schema corresponding to the example in Figure 6.2.	68
6.5	Run dependency view based on the scenario in Figure 6.2.	69
6.6	RUN-DEP view can be generated by joining ddep and produces tables as shown.	69
6.7	User collaboration view, based on the scenario in Figure 6.2.	70
6.8	Generation of (a) C-WF, (b) C-Data, and (c) C-Run views based on the scenario in Figure 6.2.	71
6.9	The C_{NSW} graph, based on the scenario in Figure 6.2.	72
6.10	The C_{NWS} graph, based on the scenario in Figure 6.2.	73
6.11	The entities and edges in the standard OPM model was the extended by <i>Workflow</i> (<i>WF</i>) entity, and <i>wasPublishedBy</i> and <i>wasExecutedIn</i> edges in the collaborative provenance model.	79
6.12	An abstract model of collaborative provenance nodes and dependencies using the extended Open Provenance Model.	79
7.1	Virtual Patient Experiment (VPE): (a) Combined Workflow and (b) Data Flow	82
7.2	The detailed view for a VPE collaboration scenario where users Annemieke, Charles and Peter work together on different parts of the same combined workflow and depend on each others workflow executions. The figure also shows user actions along with data used and produced by these workflows.	84
7.3	Simpler views over the VPE collaboration scenario showing: (a) Users Annemieke, Charles and Peter working on their parts of the combined workflow by running workflows and using data from the system and saving the provenance of their workflow runs.; (b) Run dependency view for Figure 7.2.; (c) Data dependency view for Figure 7.2.; (d) User collaboration view for Figure 7.2.	85
7.4	A typical scenario for different observables of shared data, workflows and workflow executions (runs) in CAMERA: (a) data $\{d_1, d_2, d_5\}$ published by users $\{u_2, u_3, u_5\}$; (b) workflows $\{QCF, Asbly, Taxon, Annot, Comp\}$ published by users $\{u_2, u_4, u_5\}$; and (c) flow graph for workflow runs (customized through their parameters, p_1) and related provenance data $\{d_1 \dots d_7\}$ in user space $\{u_1, u_2, u_3\}$	87
7.5	A possible end-to-end analysis stream in CAMERA with raw sequencing data.	88

7.6	Collaborative provenance views in CAMERA: (a) data dependency; (b) run dependency; and (c) user collaboration (based on Figure 7.4).	90
8.1	An example workflow execution; A_i illustrates Kepler's processing components (Actors), and d_i illustrates data.	98
8.2	A snapshot of the tables in the Kepler Provenance Database Schema.	99
8.3	Query execution time cost for data dependency view, run dependency view and collaboration view in: (a) linear time scale, and (b) logarithmic time scale. 5, 10, 25 and 50 indicate the number of run dependencies in the dataset.	106
8.4	Query execution time cost for evaluation queries 1 through 7 in: (a) linear time scale, and (b) logarithmic time scale. 5, 10, 25 and 50 indicate the number of run dependencies in the dataset.	106
9.1	Conceptual process for the Provenance Challenge 3	111
9.2	DToL system architecture, implementing workflow run provenance interoperability between Kepler and Taverna: data is published by copying from local stores S_T, S_K to the public store S_P , generating public identifiers in the process. Local provenance records are mapped to a common model and published, replacing local data references with global ones in the common provenance store, CPS.	113
9.3	Architecture for answering collaborative queries	114