Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

Citation for published version (APA):

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.
# Contents

1 Background and Problem Formulation .................................................. 1  
  1.1 Scientific Method and The Influence of Technology .......................... 1  
  1.2 The Need for Collaboration .............................................................. 3  
  1.3 Problem Statement and Research Contributions ............................... 4  
    1.3.1 Problem Definition ................................................................. 4  
    1.3.2 Contributions .................................................................................. 5  
    1.3.3 Research Roadmap ........................................................................ 6  
  1.4 Overview of the Thesis ........................................................................ 7  

2 Scientific Workflows ............................................................................. 9  
  2.1 Example: Sea Surface Temperature MatchUp Workflow ....................... 10  
  2.2 Requirements for Scientific Workflows ........................................... 11  
  2.3 Life-cycle of Scientific Workflows .................................................... 13  
  2.4 Advantages and Limitations of Scientific Workflows .......................... 14  
  2.5 Scientific Workflow Systems .............................................................. 17  
    2.5.1 Kepler Scientific Workflow Environment .................................. 19  
    2.5.2 A Reference Architecture for Scientific Workflow Management Systems .............................................................. 26  
  2.6 Usages and Current Challenges .......................................................... 26  

3 Provenance Tracking for Scientific Data and Process ............................. 29  
  3.1 Life-cycle of Scientific Workflow Provenance ................................... 31  
  3.2 Modeling and Storing Scientific Workflow Provenance ....................... 33  
    3.2.1 Open Provenance Model ............................................................. 34  
  3.3 Querying and Browsing Provenance .................................................. 37  
    3.3.1 Query Language for Provenance ................................................. 38  
  3.4 Comparing Different Scientific Workflow Provenance Approaches ......... 38  

4 Scientific Research and Collaboration Environments ............................ 41  
  4.1 Virtual Laboratories ............................................................................ 42  
    4.1.1 Virolab .......................................................................................... 42
4.1.2 The Virtual Laboratory for e-Science ................................. 42
4.2 Scientific Portals ................................................................. 43
   4.2.1 Community Cyberinfrastructure for Advanced Marine Microbial
         Ecology Research and Analysis ........................................... 43
   4.2.2 The Geosciences Network .................................................. 46
4.3 Social Networking and Sharing Environments .......................... 47
   4.3.1 myExperiment ............................................................... 47
   4.3.2 crowdLabs ................................................................. 48

5 Collaborative Provenance: A Definition ................................. 49
   5.1 Collaborative Provenance ...................................................... 49
   5.2 Collaborative Scenario ....................................................... 50
   5.3 Building Collaborative Provenance Views ............................... 53
   5.4 Analyzing User Collaborations ............................................. 54
      5.4.1 Nature of Collaboration ................................................. 55
      5.4.2 Weight of Collaboration ............................................... 55
      5.4.3 Self Collaboration ....................................................... 57
   5.5 Combining User Collaborations Attributes ............................. 57
   5.6 Example Collaborative Query Use cases ................................. 60
      5.6.1 Acknowledgement List for Collaborators ............................. 60
      5.6.2 Usage Trail of a Data Artifact ........................................ 61
   5.7 Advantages of the Collaborative Provenance Approach ............. 62

6 Modeling and Querying Collaborative Provenance .................... 63
   6.1 Collaborative Provenance Schema ....................................... 63
   6.2 Motivating Use case Schema ............................................... 65
   6.3 Generating Collaborative Provenance Views ............................. 65
      6.3.1 Data Dependency View ................................................. 66
      6.3.2 Run Dependency View .................................................. 67
      6.3.3 User Collaboration View .............................................. 69
      6.3.4 Querying for Combinations of Collaborative Attributes ......... 73
   6.4 Expressing Collaborative Queries in QLP .............................. 74
      6.4.1 Filtering Collaborative Provenance Views using QLP ............. 76
   6.5 Relation Between the Collaborative Model and OPM ................. 77

7 Collaborative Provenance Use cases ...................................... 81
   7.1 Virolab Virtual Patient Experiment Scenario ........................... 81
      7.1.1 Components of the Virtual Patient Experiment .................... 81
      7.1.2 Collaborative Provenance for VPE .................................. 83
   7.2 Collaborative Metagenomics in CAMERA ............................... 86
      7.2.1 Scientific Workflow-Driven Science in CAMERA ................... 86
8 Collaborative Provenance Database Implementation and Evaluation

8.1 Database Implementation ........................................ 97
8.1.1 CAMERA Workflows and Provenance Database ......... 97
8.1.2 Preparation of Collaborative Provenance Experimental Dataset .... 98
8.1.3 Implementation ............................................. 101
8.2 Evaluation ...................................................... 101

9 Addressing Interoperability in Collaborative Provenance

9.1 Interoperability of Scientific Workflows and Their Provenance .... 109
9.2 Interoperability Scenarios based on Provenance Challenges ...... 110
9.2.1 PC3 Usecase .................................................. 111
9.2.2 PC1 Usecase .................................................. 112
9.3 QLP-based Interoperable Query Framework for Provenance .... 114

10 Conclusions and Future Directions

10.1 Summary of Contributions ...................................... 117
10.2 Possible Extensions to the Model ............................... 119
10.3 Future Directions ............................................... 119
10.3.1 Interoperable Collaborative Provenance ................. 119
10.3.2 OPM Profile for Collaborative Provenance ............ 120
10.3.3 Restricted User Spaces ...................................... 120
10.3.4 Optimization of Collaborative Query Evaluation and Visualization .... 120
10.3.5 Semantic Collaborative Provenance Analysis using RDF .... 121
10.3.6 Social Network Analysis using Collaborative Provenance .... 121
10.3.7 Going Beyond Scientific Workflows and Data ........... 121

List of Figures .................................................... 123
List of Tables ...................................................... 127
Bibliography ......................................................... 129
Samenvatting (Dutch Summary) ..................................... 143