Lecture Notes in Computer Science

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
  Lancaster University, UK
Takeo Kanade
  Carnegie Mellon University, Pittsburgh, PA, USA
Josef Kittler
  University of Surrey, Guildford, UK
Jon M. Kleinberg
  Cornell University, Ithaca, NY, USA
Alfred Kobsa
  University of California, Irvine, CA, USA
Friedemann Mattern
  ETH Zurich, Switzerland
John C. Mitchell
  Stanford University, CA, USA
Moni Naor
  Weizmann Institute of Science, Rehovot, Israel
Oscar Nierstrasz
  University of Bern, Switzerland
C. Pandu Rangan
  Indian Institute of Technology, Madras, India
Bernhard Steffen
  TU Dortmund University, Germany
Madhu Sudan
  Microsoft Research, Cambridge, MA, USA
Demetri Terzopoulos
  University of California, Los Angeles, CA, USA
Doug Tygar
  University of California, Berkeley, CA, USA
Gerhard Weikum
  Max Planck Institute for Informatics, Saarbruecken, Germany
Provenance and Annotation of Data and Processes

4th International Provenance and Annotation Workshop, IPAW 2012
Santa Barbara, CA, USA, June 19-21, 2012
Revised Selected Papers
Preface

“Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.”

This quotation is from the W3C Provenance Incubator Group Final Report (http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/).

2012 is a watershed year for provenance/annotation research. Under the stewardship of the World Wide Web Consortium, the global community of provenance practitioners is converging on standardized definitions, models, representations, and protocols for provenance. An infrastructure may soon be in place that could potentially support universal access to the provenance of online artifacts. The time is ripe to explore the implications of ubiquitous provenance.

Provenance is understood to be a critical component of information trustworthiness. Provenance is also increasingly understood to be essential to scientific reproducibility – the provenance and annotation of a digital scientific artifact often fulfills the same function that a paper notebook did for earlier laboratory experiments. In many cases provenance offers the only coherent picture of ad-hoc digital workflows. Provenance is also a requirement for long-term preservation of digital information.

The spread of automatic systems for provenance capture and management will allow provenance to be associated with digital artifacts whose complexity (e.g., social networks) or volume (e.g., environmental satellite data) would make manual annotation prohibitive. Furthermore, the availability of large corpora of provenance records is enabling research into automatic exploration of and reasoning about provenance.

The Fourth International Provenance and Annotation Workshop (IPAW 2012) built on the success of previous workshops held in Troy (2010), Salt Lake City (2008), Chicago (2006, 2002), and Edinburgh (2003). IPAW 2012 was held in Santa Barbara, California at the Bren School of Environmental Science and Management at the University of California, Santa Barbara. The 50 attendees represented both academia and industry, and came from the US, the UK, the Netherlands, Brazil, and Germany.

In response to our call for papers, we received 49 full paper, poster, and demo submissions. Full papers received a minimum of 3 reviews and poster and demo papers received at least 2 reviews. After review, 14 full papers, 4 demo papers, and 12 poster papers were accepted. Many papers covered classic themes of the provenance literature including research on provenance for workflow systems, databases, the web, and applications to science. However, new themes emerged
including the application of network analysis techniques to provenance, as well as investigating the ability to reconstruct or recreate provenance traces.

In addition to the papers, posters and demos, the workshop had a session providing updates on related provenance events. Philip E. Bourne from the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego gave an outstanding keynote, The Provenance Divide, on the gap between fundamental provenance research and the demand for provenance in the biomedical and scientific domains. He encouraged the community to close that gap.

As with prior IPAW workshops, there were additional events surrounding the core workshop. A tutorial on the W3C Provenance Working Group’s emerging specifications for interchanging was attended by 28 participants. Likewise, the Data Observation Network for Earth (DataONE) organized a meeting on provenance and scientific workflow. Finally, the W3C Provenance Working Group held their third face-to-face meeting after the conclusion of the workshop. IPAW has become a nexus in the community not just for communicating results but also for starting and maintaining collaborations.

IPAW 2012 was a fantastic event driven by an active and engaged community of provenance researchers facilitated by a beautiful and well-organized venue at the Bren School. We thank B.J. Danetra and her staff for their support during the conference, and Kim Fugate for handling conference registration and billing. We also thank the Program Committee for their thoughtful reviews.

July 2012

Paul Groth
James Frew
Organization

Program Committee

Ilkay Altintas University of California, San Diego
Eddy Banks Lawrence Livermore National Laboratory
Bruce Barkstrom SGA
Khalid Belhajjame University of Manchester
Shawn Bowers Gonzaga University
Remco Chang Tufts University
Adriane Chapman The MITRE Corporation
Paolo Ciccarese Harvard Medical School / Massachusetts General Hospital
Oscar Corcho Universidad Politécnica de Madrid
Helena Deus Digital Enterprise Research Institute, NUIG
Kai Eckert Mannheim University Library
Peter Edwards University of Aberdeen
Todd Elsethagen Pacific Northwest National Laboratory
Juliana Freire Polytechnic Institute of New York University
James Frew University of California Santa Barbara
Yolanda Gil Information Sciences Institute, University of Southern California
Jose Manuel Gomez-Perez Intelligent Software Components (iSOCO) S.A.
Paul Groth VU University Amsterdam
Olaf Hartig Humboldt-Universität zu Berlin
Jan Hidders Delft University of Technology
Jane Hunter University of Queensland
H.V. Jagadish University of Michigan
Qing Liu CSIRO ICT Centre
Shiyong Lu Wayne State University
Bertram Ludäscher UC Davis
Marta Mattoso COPPE – Federal Univ. Rio de Janeiro
Deborah L. McGuinness Tetherless World Constellation, Rensselaer Polytechnic Institute
Simon Miles King’s College London
Paolo Missier Newcastle University
James Myers CCNI/Rensselaer Polytechnic Institute
Edoardo Pignotti University of Aberdeen
Paulo Pinheiro Da Silva University of Texas at El Paso
Beth Plale Indiana University
VIII Organization

Satya Sahoo  Case Western Reserve University
Amit Sheth  Kno.e.sis Center, Wright State University
Eric Stephan  Pacific Northwest National Laboratory
Kerry Taylor  CSIRO ICT Centre
Curt Tilmes  NASA GSFC
Jan Van Den Bussche  Hasselt University and Transnational University of Limburg
Jun Zhao  University of Oxford

Additional Reviewers

Chen, Yuhui  Michaelis, James
Dey, Saumen  Nguyen, Vinh
Dias, Jonas  Oliveira, Daniel
Koehler, Sven  Palmer, Doug
Koop, David  Ritze, Dominique
Lebo, Timothy  Sarkar, Anandarup
McCusker, Jim
# Table of Contents

**Documents Databases**

SourceTrac: Tracing Data Sources within Spreadsheets ............... 1  
*Hazeline U. Asuncion*

Towards Integrating Workflow and Database Provenance ............. 11  
*Fernando Chirigati and Juliana Freire*

DEEP: A Provenance-Aware Executable Document System .......... 24  
*Huanjia Yang, Danius T. Michaelides, Chris Charlton,  
William J. Browne, and Luc Moreau*

**The Web**

Towards Unified Provenance Granularities ....................... 39  
*Timothy Lebo, Ping Wang, Alvaro Graves, and  
Deborah L. McGuinness*

Functional Requirements for Information Resource Provenance on the Web ......................................................... 52  
*James P. McCusker, Timothy Lebo, Alvaro Graves,  
Dominic Difranzo, Paulo Pinheiro, and Deborah L. McGuinness*

A PROV Encoding for Provenance Analysis Using Deductive Rules .... 67  
*Paolo Missier and Khalid Belhajjame*

**Reconstruction**

Declarative Rules for Inferring Fine-Grained Data Provenance from Scientific Workflow Execution Traces .................. 82  
*Shawn Bowers, Timothy McPhillips, and Bertram Ludäscher*

Automatic Discovery of High-Level Provenance Using Semantic Similarity ................................................................. 97  
*Tom De Nies, Sam Coppens, Davy Van Deursen,  
Erik Mannens, and Rik Van de Walle*

Transparent Provenance Derivation for User Decisions ............ 111  
*Ingrid Nunes, Yuhui Chen, Simon Miles, Michael Luck, and Carlos Lucena*
Science Applications

Detecting Duplicate Records in Scientific Workflow Results

Khalid Belhajjame, Paolo Missier, and Carole A. Goble

The Xeros Data Model: Tracking Interpretations of Archaeological Finds

Michael O. Jewell, Enrico Costanza, Tom Frankland, Graeme Earl, and Luc Moreau

Using Domain-Specific Data to Enhance Scientific Workflow Steering Queries

João Carlos de A.R. Gonçalves, Daniel de Oliveira, Kary A.C.S. Ocaña, Eduardo Ogasawara, and Marta Mattoso

Networks

Network Analysis on Provenance Graphs from a Crowdsourcing Application

Mark Ebden, Trung Dong Huynh, Luc Moreau, Sarvapali Ramchurn, and Stephen Roberts

Modelling Provenance Using Structured Occurrence Networks

Paolo Missier, Brian Randell, and Maciej Koutny

Demonstrations

DEMO: ourSpaces — A Provenance Enabled Virtual Research Environment

Peter Edwards, Chris Mellish, Edoardo Pignotti, Kapila Ponnampерuma, Thomas Bouttaz, Alan Eckhardt, Kate Pangbourne, Lorna Philip, and John Farrington

SOLE: Linking Research Papers with Science Objects

Quan Pham, Tanu Malik, Ian Foster, Roberto Di Lauro, and Raffaele Montella

DEMO: Managing the Provenance of Crowdsourced Disruption Reports

Milan Markovic, Peter Edwards, David Corsar, and Jeff Z. Pan

Designing a Provenance-Based Climate Data Analysis Application

Emanuele Santos, David Koop, Thomas Maxwell, Charles Doutriaux, Tommy Ellqvist, Gerald Potter, Juliana Freire, Dean Williams, and Cláudio T. Silva
Posters

Quality Assessment, Provenance, and the Web of Linked Sensor Data... 220
  Chris Baillie, Peter Edwards, and Edoardo Pignotti

Integrating Text and Graphics to Present Provenance Information .... 223
  Thomas Bouttaz, Alan Eckhardt, Chris Mellish, and Peter Edwards

Exploring Provenance in a Linked Data Ecosystem ................. 226
  David Corsar, Peter Edwards, Nagendra Velaga, John Nelson, and Jeff Z. Pan

Enabling Re-executions of Parallel Scientific Workflows Using Runtime
  Provenance Data .......................................... 229
  Flávio Costa, Daniel de Oliveira, Kary A.C.S. Ocaña,
  Eduardo Ogasawara, and Marta Mattoso

Access Control for OPM Provenance Graphs ....................... 233
  Roxana Danger, Robin Campbell Joy, John Darlington, and Vasa Curcin

Improving the Understanding of Provenance and Reproducibility of a
  Multi-Sensor Merged Climate Data Record ................. 236
  Hook Hua, Brian Wilson, Gerald Manipon, Lei Pan, and Eric Fetzer

Provenance Tracking in R ................................... 237
  Andrew Runnalls and Chris Silles

The Provenance Store prOOST for the Open Provenance Model ...... 240
  Andreas Schreiber, Miriam Ney, and Heinrich Wendel

A Comprehensive Model for Provenance ........................... 243
  Salmin Sultana and Elisa Bertino

Provenance Representation in the Global Change Information System
  (GCIS) ....................................................... 246
  Curt Tilmes

Integrating Provenance into an Operational Data Product Information
  System ......................................................... 249
  Stephan Zednik, James Michaelis, and Peter Fox

On Presenting Apropos Provenance for Situation Awareness and Data
  Forensics .................................................... 250
  Jing Zhao, Yogesh Simmhan, and Viktor Prasanna

Author Index .................................................. 255