

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, Lancaster, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Zürich, Switzerland

John C. Mitchell

Stanford University, Stanford, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Dortmund, Germany

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbrücken, Germany

More information about this series at <http://www.springer.com/series/7409>

Marta Mattoso · Boris Glavic (Eds.)

Provenance and Annotation of Data and Processes

6th International Provenance
and Annotation Workshop, IPAW 2016
McLean, VA, USA, June 7–8, 2016
Proceedings

Editors

Marta Mattoso
COPPE/UFRJ
Rio de Janeiro
Brazil

Boris Glavic
Illinois Institute of Technology
Chicago, IL
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-40592-6 ISBN 978-3-319-40593-3 (eBook)
DOI 10.1007/978-3-319-40593-3

Library of Congress Control Number: 2016941294

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

This volume contains the proceedings of the 6th International Provenance and Annotation Workshop (IPAW), held June 7–8, 2016, at The MITRE Corporation in McLean, Virginia, USA. Following the successful inception of ProvenanceWeek in 2014, this year’s installment again co-located the biennial IPAW workshop and the annual Workshop on the Theory and Practice of Provenance (TaPP). Together the two leading provenance workshops anchored ProvenanceWeek 2016, a full week of provenance-related activities that included a shared poster and demonstration session, and the PROV: Three Years Later and Provenance-based Security and Transparent Computing workshops.

This year’s installment of IPAW was able to honor the extraordinary achievements of IPAW’s authors through a best paper award sponsored by Springer. We would like to use this forum to again congratulate Wellington Moreira de Oliveira, Paolo Missier, Kary Ocaña, Daniel de Oliveira and Vanessa Braganholo, the authors of the paper titled “Analyzing Provenance Across Heterogeneous Provenance Graphs” for receiving this award.

This collection constitutes the peer-reviewed papers of IPAW 2016. These include 12 long papers that report in depth on the results of research around provenance and two short papers that discuss tools and services that were presented in the form of a system demonstration. Finally, we have included 14 short papers that were also presented as part of the joint IPAW/TaPP poster session. The final papers, demos, and short papers accompanied by poster presentations were selected from a total of 54 submissions. All full-length research papers received a minimum of three reviews.

The papers of IPAW 2016 provide a glimpse into state-of-the-art research and practice around the automatic capture, representation, and use of provenance. The papers discussing provenance capture exemplify the diversity of applications with provenance needs including operating systems, scripting environments, and distributed environments. While automated provenance capture is necessary for wide-spread provenance collection, analysis and visualization of provenance enable users to understand and make sense of the collected provenance. Several papers focus on this aspect describing tools for visualizing large provenance graphs, for creating understandable natural language descriptions from provenance graphs, and for analyzing provenance across multiple provenance graphs. Provenance itself is meaningless if not used for a concrete purpose. The proceedings also cover papers reporting on real-world use cases of provenance as well as how to model provenance for specific domains.

In closing, we would like to thank the members of the Program Committee for their thoughtful reviews, Dr. Adriane Chapman (local chair) and her team for their excellent organization of IPAW and ProvenanceWeek 2016 at MITRE, and—last not least—the authors and participants for making IPAW the stimulating and successful event that it was.

Organization

ProvenanceWeek Senior PC Chair

Marta Mattoso Federal University of Rio de Janeiro, Brazil

IPAW PC Chair

Boris Glavic Illinois Institute of Technology, USA

TaPP PC Chair

Sarah Cohen-Boulakia Université Paris-Sud, France

Local Chair

Adriane P. Chapman The MITRE Corporation, USA

Program Committee

Khalid Belhajjame	Université Paris-Dauphine, France
Shawn Bowers	Gonzaga University, USA
Vanessa Braganholo	Universidade Federal Fluminense, Brazil
Kevin Butler	University of Florida, USA
James Cheney	University of Edinburgh, UK
Susan Davidson	University of Pennsylvania, USA
Daniel de Oliveira	Fluminense Federal University, Brazil
Anton Dignös	Free University of Bozen-Bolzano, Italy
Mohamed Eltabakh	WPI, USA
Ian Foster	University of Chicago, USA
Daniel Garijo	Universidad Politécnica de Madrid, Spain
Ashish Gehani	SRI International, USA
Paul Groth	Elsevier Labs, The Netherlands
Torsten Grust	Universität Tübingen, Germany
Olaf Hartig	Hasso Plattner Institute, Germany
Melanie Herschel	Universität Stuttgart, Germany
Trung Dong Huynh	University of Southampton, UK
Grigoris Karvounarakis	LogicBlox, USA
Oliver Kennedy	SUNY Buffalo, USA
David Koop	University of Massachusetts at Dartmouth, USA
Bertram Ludäscher	University of Illinois at Urbana-Champaign, USA
Tanu Malik	University of Chicago, USA

Deborah McGuinness	Rensselaer Polytechnic Institute, USA
Paolo Missier	Newcastle University, UK
Beth Plale	Indiana University, USA
Ravi Ramamurthy	Microsoft Research, USA
Dan Suciu	University of Washington, USA
Justin Wozniak	Argonne National Laboratory, USA
Jun Zhao	Oxford e-Research Centre, UK

Contents

Provenance Capture

RecProv: Towards Provenance-Aware User Space Record and Replay	3
<i>Yang Ji, Sangho Lee, and Wenke Lee</i>	
Tracking and Analyzing the Evolution of Provenance from Scripts	16
<i>João Felipe Pimentel, Juliana Freire, Vanessa Braganholo, and Leonardo Murta</i>	
Trade-Offs in Automatic Provenance Capture	29
<i>Manolis Stamatiogiannakis, Hasanat Kazmi, Hashim Sharif, Remco Vermeulen, Ashish Gehani, Herbert Bos, and Paul Groth</i>	
Analysis of Memory Constrained Live Provenance	42
<i>Peng Chen, Tom Evans, and Beth Plale</i>	

Provenance Analysis and Visualization

Analyzing Provenance Across Heterogeneous Provenance Graphs	57
<i>Wellington Oliveira, Paolo Missier, Kary Ocaña, Daniel de Oliveira, and Vanessa Braganholo</i>	
Prov Viewer: A Graph-Based Visualization Tool for Interactive Exploration of Provenance Data	71
<i>Troy Kohwalter, Thiago Oliveira, Juliana Freire, Esteban Clua, and Leonardo Murta</i>	
Intermediate Notation for Provenance and Workflow Reproducibility.	83
<i>Danius T. Michaelides, Richard Parker, Chris Charlton, William J. Browne, and Luc Moreau</i>	
Towards the Domain Agnostic Generation of Natural Language Explanations from Provenance Graphs for Casual Users.	95
<i>Darren P. Richardson and Luc Moreau</i>	

Provenance Models and Applications

Versioning Version Trees: The Provenance of Actions that Affect Multiple Versions	109
<i>David Koop</i>	

Enabling Web Service Request Citation by Provenance Information 122
Nicholas John Car, Laura S. Stanford, and Aaron Sedgmen

Modelling Provenance of Sensor Data for Food Safety Compliance
Checking 134
Milan Markovic, Peter Edwards, Martin Kollingbaum, and Alan Rowe

Modelling Provenance Collection Points and Their Impact on Provenance
Graphs. 146
David Gammack, Steve Scott, and Adriane P. Chapman

System Demonstrations

Yin & Yang: Demonstrating Complementary Provenance from
noWorkflow & YesWorkflow. 161
*João Felipe Pimentel, Saumen Dey, Timothy McPhillips,
Khalid Belhajjame, David Koop, Leonardo Murta, Vanessa Braganholo,
and Bertram Ludäscher*

MPO: A System to Document and Analyze Distributed Heterogeneous
Workflows 166
*Kesheng Wu, Elizabeth N. Coviello, S. M. Flanagan, Martin Greenwald,
Xia Lee, Alex Romosan, David P. Schissel, Arie Shoshani,
Josh Stillerman, and John Wright*

Joint IPAW/TaPP Poster Session

PROV-JSONLD: A JSON and Linked Data Representation for Provenance . . . 173
Trung Dong Huynh, Danilus T. Michaelides, and Luc Moreau

Provenance as Essential Infrastructure for Data Lakes 178
Isuru Suriarachchi and Beth Plale

Provenance-Based Retrieval: Fostering Reuse and Reproducibility Across
Scientific Disciplines 183
*Lucas Augusto Montalvão Costa Carvalho, Rodrigo L. Silveira,
Caroline S. Pereira, Munir S. Skaf, and Claudia Bauzer Medeiros*

Addressing Scientific Rigor in Data Analytics Using Semantic Workflows . . . 187
*John S. Erickson, John Sheehan, Kristin P. Bennett,
and Deborah L. McGuinness*

Reconstructing Human-Generated Provenance Through Similarity-Based
Clustering 191
Tom De Nies, Erik Mannens, and Rik Van de Walle

Social Media Data in Research: Provenance Challenges	195
<i>David Corsar, Milan Markovic, and Peter Edwards</i>	
Fine-Grained Provenance Collection over Scripts Through Program Slicing. . . .	199
<i>João Felipe Pimentel, Juliana Freire, Leonardo Murta, and Vanessa Braganholo</i>	
Prov2ONE: An Algorithm for Automatically Constructing ProvONE Provenance Graphs	204
<i>Ajinkya Prabhune, Aaron Zweig, Rainer Stotzka, Michael Gertz, and Juergen Hesser</i>	
Implementing Unified Why- and Why-Not Provenance Through Games	209
<i>Seokki Lee, Sven Köhler, Bertram Ludäscher, and Boris Glavic</i>	
SisGExp: Rethinking Long-Tail Agronomic Experiments	214
<i>Sergio Manuel Serra da Cruz and José Antonio Pires do Nascimento</i>	
Towards Provenance Capturing of Quantified Self Data	218
<i>Andreas Schreiber and Doreen Seider</i>	
A Review of Guidelines and Models for Representation of Provenance Information from Neuroscience Experiments.	222
<i>Margarita Ruiz-Olazar, Evandro S. Rocha, Sueli S. Rabaça, Carlos Eduardo Ribas, Amanda S. Nascimento, and Kelly R. Braghetto</i>	
Tracking and Establishing Provenance of Earth Science Datasets: A NASA-Based Example.	226
<i>Hampapuram K. Ramapriyan, Justin C. Goldstein, Hook Hua, and Robert E. Wolfe</i>	
DataONE: A Data Federation with Provenance Support	230
<i>Yang Cao, Christopher Jones, Víctor Cuevas-Vicentín, Matthew B. Jones, Bertram Ludäscher, Timothy McPhillips, Paolo Missier, Christopher Schwalm, Peter Slaughter, Dave Vieglais, Lauren Walker, and Yaxing Wei</i>	
Erratum to: Trade-Offs in Automatic Provenance Capture	E1
<i>Manolis Stamatogiannakis, Hasanat Kazmi, Hashim Sharif, Remco Vermeulen, Ashish Gehani, Herbert Bos, and Paul Groth</i>	
Author Index	235