Lecture Notes in Computer Science 12839

Founding Editors

Gerhard Goos
Karlsruhe Institute of Technology, Karlsruhe, Germany

Juris Hartmanis
Cornell University, Ithaca, NY, USA

Editorial Board Members

Elisa Bertino
Purdue University, West Lafayette, IN, USA

Wen Gao
Peking University, Beijing, China

Bernhard Steffen
TU Dortmund University, Dortmund, Germany

Gerhard Woeginger
RWTH Aachen, Aachen, Germany

Moti Yung
Columbia University, New York, NY, USA

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

Boris Glavic · Vanessa Braganholo ·
David Koop (Eds.)

Provenance and Annotation of Data and Processes

8th and 9th International Provenance
and Annotation Workshop, IPAW 2020 + IPAW 2021
Virtual Event, July 19–22, 2021
Proceedings

Springer

Editors
Boris Glavic
Illinois Institute of Technology
Chicago, IL, USA

Vanessa Braganholo
Fluminense Federal University
Niterói, Brazil

David Koop
Northern Illinois University
DeKalb, IL, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-80959-1 ISBN 978-3-030-80960-7 (eBook)
https://doi.org/10.1007/978-3-030-80960-7

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

© Springer Nature Switzerland AG 2021, corrected publication 2021
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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains the proceedings of the 8th and 9th *International Provenance and Annotation Workshop* (IPAW), held as part of ProvenanceWeek in 2020 and 2021. Due to the COVID-19 pandemic, ProvenanceWeek 2020 was held as a 1-day virtual event with brief teaser talks on June 22, 2020. In 2021, ProvenanceWeek again co-located the biennial IPAW workshop with the annual *Workshop on the Theory and Practice of Provenance* (TaPP). Together, the two leading provenance workshops anchored a 4-day event of provenance related activities that included a shared poster and demonstration session, the first *Workshop on Provenance for Transparent Research* (T7), and the first *Workshop on Provenance and Visualization* (ProvViz). The events were held virtually during July 19–22 2021. At IPAW 2021, authors from both 2020 and 2021 presented and discussed their work.

This collection constitutes the peer reviewed papers of IPAW 2020 and 2021. These include eleven long papers which report in depth on the results of research around provenance and twelve short papers that were presented as part of the joint IPAW/TaPP poster and demonstration session. The final papers and short papers accompanied by poster presentations and demonstrations were selected from a total of 31 submissions. All full-length research papers received a minimum of three reviews.

The IPAW papers provide a glimpse into state-of-the-art research and practice around the capture, representation, querying, inference, and summarization of provenance. Papers also address applications of provenance such as security, reliability, and trustworthiness. The papers discussing provenance capture focus on templates and explore Artificial Intelligence scenarios, focusing on capturing provenance of Deep Neural Networks. Provenance representation papers include work on evidence graphs and a new JSON serialization for PROV. Several papers focus on provenance queries and inference. In particular, they explore provenance type inference, the use of provenance for query result exploration, and provenance inference of computational notebooks.

Provenance itself is meaningless if not used for a concrete purpose. The proceedings also cover papers reporting on real-world use cases of provenance. Application scenarios explored in the papers include health care and, especially, COVID-19.

We would like to thank the members of the Program Committee (PC) for their thoughtful and insightful reviews along with Dr. Thomas Moyer (local chair) and his team for their excellent organization of both IPAW and ProvenanceWeek 2020/2021. We also want to thank the authors and participants for making IPAW the stimulating and successful event that it was.

July 2021

Vanessa Braganholo
David Koop
Boris Glavic

Organization

Organizing Committee

Boris Glavic (ProvenanceWeek 2020/2021 Senior PC Chair)	Illinois Institute of Technology, USA
Vanessa Braganholo (IPAW 2020/2021 PC Chair)	Fluminense Federal University, Brazil
Thomas Pasquier (TaPP 2020 PC Chair)	University of Bristol, UK
Tanu Malik (TaPP 2021 PC Co-chair)	DePaul University, USA
Thomas Pasquier (TaPP 2021 PC Co-chair)	University of Bristol, UK
David Koop (2020/2021 Demos/Poster Chair)	Northern Illinois University, USA
Thomas Moyer (2020/2021 Local Chair)	UNC Charlotte, USA

IPAW 2020 Program Committee

Andreas Schreiber	German Aerospace Center (DLR), Germany
Barbara Lerner	Mount Holyoke College, USA
Beth Plale	Indiana University, USA
Daniel de Oliveira	Fluminense Federal University, Brazil
Daniel Garijo	University of Southern California, USA
David Corsar	Robert Gordon University, UK
Dong Huynh	King's College London, UK
Fernando Chirigati	New York University, USA
Grigoris Karvounarakis	LogicBlox, USA
Hala Skaf-Molli	Nantes University, France
Ilkay Altintas	San Diego Supercomputer Center, USA
Jacek Cala	Newcastle University, UK
James Cheney	University of Edinburgh, UK
James Frew	University of California, Santa Barbara, USA
James Myers	University of Michigan, USA
Jan Van Den Bussche	Universiteit Hasselt, Belgium
João Felipe Pimentel	Fluminense Federal University, Brazil

Luc Moreau	King's College London, UK
Luiz M. R. Gadelha Jr.	LNCC, Brazil
Paolo Missier	Newcastle University, UK
Paul Groth	University of Amsterdam, Netherlands
Pinar Alper	University of Luxembourg, Luxembourg
Shawn Bowers	Gonzaga University, USA
Seokki Lee	IIT, USA
Simon Miles	King's College London, UK
Tanu Malik	DePaul University, USA
Timothy Clark	University of Virginia, USA

IPAW 2021 Program Committee

Andreas Schreiber	German Aerospace Center (DLR), Germany
Adriane Chapman	University of Southampton, UK
Bertram Ludascher	University of Illinois at Urbana-Champaign, USA
Cláudia Bauzer Medeiros	UNICAMP, Brazil
Daniel de Oliveira	Fluminense Federal University, Brazil
Daniel Garijo	University of Southern California, USA
David Corsar	Robert Gordon University, UK
Eduardo Ogasawara	CEFET, Brazil
Grigoris Karvounarakis	LogicBlox, USA
Hala Skaf-Molli	Nantes University, France
Jacek Cala	Newcastle University, UK
James Cheney	University of Edinburgh, UK
James McCusker	Rensselaer Polytechnic Institute, USA
James Myers	University of Michigan, USA
Jan Van Den Bussche	Universiteit Hasselt, Belgium
João Felipe Pimentel	Fluminense Federal University, Brazil
Luc Moreau	King's College London, UK
Luiz M. R. Gadelha Jr.	LNCC, Brazil
Marta Mattoso	Universidade Federal do Rio de Janeiro, Brazil
Paolo Missier	Newcastle University, UK
Paul Groth	University of Amsterdam, Netherlands
Pinar Alper	University of Luxembourg, Luxembourg
Seokki Lee	University of Cincinnati, USA
Timothy Clark	University of Virginia, USA
Vasa Curcin	King's College London, UK

Contents

Provenance Capture and Representation

A Delayed Instantiation Approach to Template-Driven Provenance for Electronic Health Record Phenotyping 3
Elliot Fairweather, Martin Chapman, and Vasa Curcin

Provenance Supporting Hyperparameter Analysis in Deep Neural Networks . . . 20
Débora Pina, Liliane Kunstmann, Daniel de Oliveira, Patrick Valduriez, and Marta Mattoso

Evidence Graphs: Supporting Transparent and FAIR Computation, with Defeasible Reasoning on Data, Methods, and Results 39
Sadnan Al Manir, Justin Niestroy, Maxwell Adam Levinson, and Timothy Clark

The PROV-JSONLD Serialization: A JSON-LD Representation for the PROV Data Model 51
Luc Moreau and Trung Dong Huynh

Security

Proactive Provenance Policies for Automatic Cryptographic Data Centric Security 71
Shamaria Engram, Tyler Kaczmarek, Alice Lee, and David Bigelow

Provenance-Based Security Audits and Its Application to COVID-19 Contact Tracing Apps 88
Andreas Schreiber, Tim Sonnekalb, Thomas S. Heinze, Lynn von Kurnatowski, Jesus M. Gonzalez-Barahona, and Heather Packer

Provenance Types, Inference, Queries and Summarization

Notebook Archaeology: Inferring Provenance from Computational Notebooks 109
David Koop

Efficient Computation of Provenance for Query Result Exploration 127
Murali Mani, Naveenkumar Singaraj, and Zhenyan Liu

Incremental Inference of Provenance Types 145
David Kohan Marzagão, Trung Dong Huynh, and Luc Moreau

Reliability and Trustworthiness

Non-repudiable Provenance for Clinical Decision Support Systems 165
Elliot Fairweather, Rudolf Wittner, Martin Chapman, Petr Holub, and Vasa Curcin

A Model and System for Querying Provenance from Data Cleaning Workflows 183
Nikolaus Nova Parulian, Timothy M. McPhillips, and Bertram Ludäscher

Joint IPAW/TaPP Poster and Demonstration Session

ReproduceMeGit: A Visualization Tool for Analyzing Reproducibility of Jupyter Notebooks 201
Sheeba Samuel and Birgitta König-Ries

Mapping Trusted Paths to VGI 207
Bernard Roper, Adriane Chapman, David Martin, and Stefano Cavazzi

Querying Data Preparation Modules Using Data Examples 211
Khalid Belhajjame and Mahmoud Barhamgi

Privacy Aspects of Provenance Queries 218
Tanja Auge, Nic Scharlau, and Andreas Heuer

ISO 23494: Biotechnology – Provenance Information Model for Biological Specimen And Data 222
Rudolf Wittner, Petr Holub, Heimo Müller, Joerg Geiger, Carole Goble, Stian Soiland-Reyes, Luca Pireddu, Francesca Frexia, Cecilia Mascia, Elliot Fairweather, Jason R. Swedlow, Josh Moore, Caterina Strambio, David Grunwald, and Hiroki Nakae

Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles 226
Sheeba Samuel, Frank Löffler, and Birgitta König-Ries

ProvViz: An Intuitive Prov Editor and Visualiser 231
Ben Werner and Luc Moreau

Curating Covid-19 Data in Links 237
Vashti Galpin and James Cheney

Towards a Provenance Management System for Astronomical Observatories 244
Mathieu Servillat, François Bonnarel, Catherine Boisson, Mireille Louys, Jose Enrique Ruiz, and Michèle Sanguillon

Towards Provenance Integration for Field Devices in Industrial IoT Systems . . . 250
Iori Mizutani, Jonas Brütsch, and Simon Mayer

COVID-19 Analytics in Jupyter: Intuitive Provenance Integration Using *ProvIt* 256
Martin Chapman, Elliot Fairweather, Asfand Khan, and Vasa Curcin

CPR-A Comprehensible Provenance Record for Verification Workflows in Whole Tale 263
Timothy M. McPhillips, Thomas Thelen, Craig Willis, Kacper Kowalik, Matthew B. Jones, and Bertram Ludäscher

Correction to: ISO 23494: Biotechnology – Provenance Information Model for Biological Specimen And Data C1
Rudolf Wittner, Petr Holub, Heimo Müller, Joerg Geiger, Carole Goble, Stian Soiland-Reyes, Luca Pireddu, Francesca Frexia, Cecilia Mascia, Elliot Fairweather, Jason R. Swedlow, Josh Moore, Caterina Strambio, David Grunwald, and Hiroki Nakae

Author Index 271