

# **Provenance:**

## **An Introduction to PROV**

# Synthesis Lectures on the Semantic Web: Theory and Technology

Editors

**James Hendler**, *Rensselaer Polytechnic Institute*

**Ying Ding**, *Indiana University*

Synthesis Lectures on the Semantic Web: Theory and Application is edited by James Hendler of Rensselaer Polytechnic Institute. Whether you call it the Semantic Web, Linked Data, or Web 3.0, a new generation of Web technologies is offering major advances in the evolution of the World Wide Web. As the first generation of this technology transitions out of the laboratory, new research is exploring how the growing Web of Data will change our world. While topics such as ontology-building and logics remain vital, new areas such as the use of semantics in Web search, the linking and use of open data on the Web, and future applications that will be supported by these technologies are becoming important research areas in their own right. Whether they be scientists, engineers or practitioners, Web users increasingly need to understand not just the new technologies of the Semantic Web, but to understand the principles by which those technologies work, and the best practices for assembling systems that integrate the different languages, resources, and functionalities that will be important in keeping the Web the rapidly expanding, and constantly changing, information space that has changed our lives.

Topics to be included:

- Semantic Web Principles from linked-data to ontology design
- Key Semantic Web technologies and algorithms
- Semantic Search and language technologies
- The Emerging “Web of Data” and its use in industry, government and university applications
- Trust, Social networking and collaboration technologies for the Semantic Web
- The economics of Semantic Web application adoption and use
- Publishing and Science on the Semantic Web
- Semantic Web in health care and life sciences

Provenance: An Introduction to PROV

Luc Moreau and Paul Groth

2013

Resource-Oriented Architecture Patterns for Webs of Data

Brian Sletten

2013

Aaron Swartz's A Programmable Web: An Unfinished Work

Aaron Swartz

2013

Incentive-Centric Semantic Web Application Engineering

Elena Simperl, Roberta Cuel, and Martin Stein

2013

Publishing and Using Cultural Heritage Linked Data on the Semantic Web

Eero Hyvönen

2012

VIVO: A Semantic Approach to Scholarly Networking and Discovery

Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding

2012

Linked Data: Evolving the Web into a Global Data Space

Tom Heath and Christian Bizer

2011

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2013

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Provenance: An Introduction to PROV

Luc Moreau and Paul Groth

ISBN: 978-3-031-79449-0      paperback

ISBN: 978-3-031-79450-6      ebook

DOI 10.1007/978-3-031-79450-6

A Publication in the Springer series

*SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY*

Lecture #7

Series Editors: James Hendler, *Rensselaer Polytechnic Institute*

Ying Ding, *Indiana University*

Series ISSN

Synthesis Lectures on the Semantic Web: Theory and Technology

Print 2160-4711    Electronic 2160-472X

# Provenance: An Introduction to PROV

Luc Moreau  
University of Southampton

Paul Groth  
VU University of Amsterdam

*SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND  
TECHNOLOGY #7*

## ABSTRACT

The World Wide Web is now deeply intertwined with our lives, and has become a catalyst for a data deluge, making vast amounts of data available online, at a click of a button. With Web 2.0, users are no longer passive consumers, but active publishers and curators of data. Hence, from science to food manufacturing, from data journalism to personal well-being, from social media to art, there is a strong interest in *provenance*, a description of what influenced an artifact, a data set, a document, a blog, or any resource on the Web and beyond. Provenance is a crucial piece of information that can help a consumer make a judgment as to whether something can be trusted. Provenance is no longer seen as a curiosity in art circles, but it is regarded as pragmatically, ethically, and methodologically crucial for our day-to-day data manipulation and curation activities on the Web.

Following the recent publication of the PROV standard for provenance on the Web, which the two authors actively help shape in the Provenance Working Group at the World Wide Web Consortium, this Synthesis lecture is a hands-on introduction to PROV aimed at Web and linked data professionals. By means of recipes, illustrations, a website at [www.provbook.org](http://www.provbook.org), and tools, it guides practitioners through a variety of issues related to provenance: how to generate provenance, publish it on the Web, make it discoverable, and how to utilize it. Equipped with this knowledge, practitioners will be in a position to develop novel applications that can bring openness, trust, and accountability.

## KEYWORDS

provenance, prov, audit trail, compliance, audit, traceability, semantic web

*To Ryan, Sammey, and Achika — L.M.*

*To Marieke — P.G.*

# Contents

<b>Preface</b> .....	<b>xiii</b>
<b>Acknowledgments</b> .....	<b>xv</b>
<b>1 Introduction</b> .....	<b>1</b>
1.1 The Case for Provenance .....	1
1.2 A Definition of Provenance .....	3
1.3 Provenance and the Web Architecture .....	4
1.4 The W3C PROV Standard .....	5
1.5 Online Extensions .....	6
<b>2 A Data Journalism Scenario</b> .....	<b>9</b>
2.1 Scenario: The employment report .....	9
2.1.1 Characters .....	10
2.1.2 Story Creation and Publication .....	10
2.1.3 Crunching Data .....	11
2.1.4 Reusing the Story .....	12
2.2 Provenance Use Cases .....	12
2.2.1 Quality Assessment .....	12
2.2.2 Compliance .....	14
2.2.3 Cataloging .....	15
2.2.4 Replay .....	15
2.3 A Brief Introduction to Expressing Provenance .....	16
2.4 Summary .....	19
<b>3 The PROV Ontology</b> .....	<b>21</b>
3.1 Overview .....	21
3.2 Qualified Relation Patterns .....	23
3.3 Data Flow View .....	24
3.3.1 Entity .....	24
3.3.2 Derivation .....	24
3.3.3 Revision .....	25

3.3.4	Quotation	25
3.3.5	Primary Source	25
3.4	Process Flow View	25
3.4.1	Activity	26
3.4.2	Generation	26
3.4.3	Usage	27
3.4.4	Invalidation	27
3.4.5	Start	28
3.4.6	End	29
3.4.7	Communication	30
3.5	Responsibility View	30
3.5.1	Agent	30
3.5.2	Attribution	31
3.5.3	Association	31
3.5.4	Delegation	32
3.6	Alternates View	33
3.6.1	Specialization	33
3.6.2	Alternate	33
3.7	Bundles	34
3.8	Miscellaneous	34
3.8.1	Collection and Membership	34
3.8.2	Refined Derivation	35
3.8.3	Further Properties	36
3.9	Ontology Structure	36
3.10	Summary	37
<b>4</b>	<b>Provenance Recipes</b>	<b>39</b>
4.1	Modeling	39
4.1.1	Iterative Modeling	39
4.1.2	Identify, Identify, Identify!	40
4.1.3	From Data Flow to Activities	41
4.1.4	Plan for Revisions	42
4.1.5	Modeling Update and Other Destructive Activities	43
4.1.6	Modeling Message Passing	44
4.1.7	Modeling Parameters	45
4.1.8	Introduce the Environment	46
4.1.9	Modeling Sub-activities	46

4.2	Organizing	48
4.2.1	Stitch Provenance Together	48
4.2.2	Use Content-Negotiation when Exposing Provenance	48
4.2.3	Bundle Up and Provide Attribution to Provenance	49
4.2.4	Embedding Provenance in HTML	50
4.2.5	Embedding Provenance in Other Media	51
4.2.6	When all Else Fails, add Provenance to HTTP Headers	53
4.2.7	Embedding Provenance in Bundles: Self-Referential Bundles	54
4.2.8	When Displaying Provenance, Adopt Conventional Layout	55
4.3	Collecting	56
4.3.1	Use Structured Logs to Collect Provenance	56
4.3.2	Collect in a Local Form, Expose as PROV	56
4.4	Anti-patterns	57
4.4.1	Activity but No Derivation	57
4.4.2	Association but No Attribution	58
4.4.3	Specify Responsibility First, What a Prov:Agent is Will Follow	58
4.5	Summary	59
<b>5</b>	<b>Validation, Compliance, Quality, Replay</b>	<b>61</b>
5.1	Validation Use Cases	62
5.2	Principles of Validation	62
5.2.1	Events and Their Ordering	63
5.2.2	Simultaneous Events	66
5.2.3	Nested Intervals and Specialization	67
5.2.4	Use Cases Revisited	67
5.3	Utilizing Provenance	68
5.3.1	Provenance-Based Compliance	70
5.3.2	Provenance-Based Quality Assessment	71
5.3.3	Provenance-Based Cataloging	71
5.3.4	Provenance-Based Replaying	72
5.4	Implementation Techniques for Provenance Analysis	72
5.4.1	Finding Ancestors	73
5.4.2	Deep Traversal	73
5.4.3	Pattern Detection for Policy Compliance	73
5.4.4	Time Comparison	74
5.4.5	Trust-Based Filtering	74
5.4.6	Finding External Ancestor Resources	76

5.4.7	Replay Technique .....	77
5.5	Summary .....	79
<b>6</b>	<b>Provenance Management .....</b>	<b>81</b>
6.1	Exposing Provenance .....	81
6.1.1	Embedding Provenance in HTML with RDFa .....	81
6.1.2	Provenance Services .....	85
6.2	Provenance Management Tools .....	89
6.2.1	ProvToolbox .....	89
6.2.2	ProvPy .....	89
6.2.3	provconvert and ProvTranslator .....	90
6.2.4	ProvStore .....	90
6.2.5	ProvValidator .....	90
6.2.6	Browser PROV Extractor .....	90
6.2.7	ProvVis: Interactive Visualizations for PROV .....	92
6.3	Provenance Management on www.provbook.org .....	93
6.3.1	Directories .....	93
6.3.2	URI Schemes for Entities, Agents, and Activities .....	93
6.3.3	The PROV Book Ontology .....	95
6.3.4	Data Journalism Provenance .....	95
6.3.5	Exposing Provenance .....	95
6.4	Summary .....	98
<b>7</b>	<b>Conclusion .....</b>	<b>99</b>
7.1	Toward Provenance Self Certification: A Checklist .....	99
7.2	Applying Provenance in the Wild .....	100
7.3	Open Issues .....	101
7.3.1	Provenance Enabling Systems .....	101
7.3.2	Fundamentals of Provenance .....	102
7.3.3	Provenance Analytics .....	102
7.3.4	Securing Provenance .....	103
7.4	Final Words .....	103
	<b>Bibliography .....</b>	<b>105</b>
	<b>Authors' Biographies .....</b>	<b>109</b>
	<b>Index .....</b>	<b>111</b>

# Preface

This book stems from the authors' decade of experience related to provenance, their leadership of the W3C Provenance Working Group, and several tutorials on provenance delivered at FIS'10,<sup>1</sup> IPAW'12,<sup>2</sup> ISWC'12,<sup>3</sup> and ESWC'13.<sup>4</sup> The PROV specifications provide normative and non-normative material about data models and protocols related to provenance, but offer very little guidance on how to design and deploy provenance. The purpose of this book is to address this concern, by providing a compact and practical guide to developing PROV-based applications, combined with a fully-deployed example that readers can inspect and study in detail.

This book is intended for developers aiming to make their systems provenance-aware. It can also be used as a textbook for an undergraduate course on provenance, or potentially part of a larger module on the Semantic Web or Information Assurance.

We assume that readers are familiar with core technologies of the Web and Semantic Web. Specifically, understanding of URIs, basic understanding of HTTP, basic notions of HTML, and the principles of RDF are prerequisites for using this book.

The rest of this book is structured as follows.

- Chapter 1 provides an introduction to provenance.
- Chapter 2, *Data Journalism Scenario*, introduces a data journalism example, which we use extensively across the book. It describes the publication of an article by a fictitious news agency, based on some government statistics recently published. This example is used to illustrate many use cases that can be addressed by means of provenance. The chapter also provides a brief introduction to how provenance can be represented.
- Chapter 3, *The PROV Ontology*, is a concise presentation of the PROV ontology, which can be used as reference material for this book.
- Chapter 4, *Provenance Recipes*, is concerned with methodological recipes on how to model provenance for specific problems, and how to deploy it in an inter-operable manner.
- Chapter 5, *Validation, Compliance, Quality, Replay*, first focuses on the notion of valid provenance. It then expands on various forms of utilization of provenance. A series of technical requirements is introduced, and SPARQL queries are used to illustrate how these can be implemented.

<sup>1</sup>OPM Tutorial: <http://openprovenance.org/tutorial/>

<sup>2</sup>PROV Tutorial at IPAW'12: [http://www.w3.org/2011/prov/wiki/IPAW\\_2012\\_Tutorial](http://www.w3.org/2011/prov/wiki/IPAW_2012_Tutorial)

<sup>3</sup>PROV Tutorial at ISWC'12: <http://www.w3.org/2011/prov/wiki/ISWCProvTutorial>

<sup>4</sup>PROV Tutorial at ESWC'13: <http://www.w3.org/2001/sw/wiki/ESWC2013ProvTutorial>

- Chapter 6, *Provenance Management*, is dedicated to techniques to manage provenance, and specifically to make it available, by means of `RDFa` embedded in `HTML` documents and provenance services. A series of libraries, services, and tools are then briefly discussed. Finally, a guided tour of <http://www.provbook.org> illustrates the provenance management techniques deployed on the website associated with the book.
- Chapter 7, *Conclusion*, summarizes the book with a checklist that developers can follow to check whether their provenance is properly structured and exposed. Open issues and future research directions are also discussed.

Luc Moreau and Paul Groth  
August 2013

# Acknowledgments

Luc Moreau wishes to thank Alex Fraser, Trung Dong Huynh, Mike Jewell, Amir Sezavar Keshavarz, and Danilus Michaelides for their work on the Southampton Provenance Tool Suite.

Paul Groth thanks Frank van Harmelen for supporting his involvement in the working group and Stefan Schlobach for putting up with constant telecons.

The authors thank the entire W3C Provenance Working Group for their efforts producing PROV, and Trung Dong Huynh for his comments on a draft of the book. Additionally, the authors thank Yolanda Gil for her work leading to the creation of the W3C Provenance Working Group.

Luc Moreau's work is supported under SOCIAM: The Theory and Practice of Social Machines; ORCHID: Human-Agent Collectives: From Foundations to Applications; SmartSociety: hybrid and diversity-aware collective adaptive systems: where people meet machines to build a smarter society; and PATINA: Personal Architectonics Through Interactions with Artefacts.

- The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1.
- The ORCHID Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/I011587/1.
- The SmartSociety Project is funded under FP7 Grant agreement number 600854.
- The PATINA Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/H042806/1.

Paul Groth's work is supported by the Data2Semantics project in the Dutch national program COMMIT as well as the EU IMI Open PHACTS project. Open PHACTS receives financial support provided by the IMI-JU, grant agreement number 115191.

Luc Moreau and Paul Groth  
August 2013