

An overview of provenance and its application to eBooks

Luc Moreau <u>I.moreau@ecs.soton.ac.uk</u> Work with Danius Michaelides and Huanjia Yang

Credits: The Open Provenance Tutorial at FIS'10, Berlin, 2010 by Luc Moreau, Paul Groth, Jun Zhao.

Contents



- Part I
 - Provenance 101
 - Provenance in eBooks
- Part II
 - Open Provenance Model Overview
 - Representation of eBook provenance in OPM
- Part III
 - Towards a W3C standard



Part I Provenance 101

Provenance Overview



- A motivating example
- Definition

Scenario: BlogAgg



NEW ITEM FINDER: tthew Lasar / Ars Technica:

Donald Melanson / Engadget

David / TmoNews:

Microsoft BizSpark: BizSpark Startup of the

Seesmic Blog

Day - Graphic.ly — The BizSpark startup of the day is

17 Find

Did Internet founders foresee future filled with paid, prioritized traffic? Find

iPad Wi-Fi Models Available in China on September

iPad headed to Target on October 3rd? Find

Hewlett-Packard to Buy Security-Software Maker ArcSight for \$1.5 Billion Find

myTouch HD Internal Doc Hints Toward FFC And

Graphic.ly, based in the US. Below you will find an

interview with Micah Baldwin, CEO of Graphic.ly.

America's Largest "4G" Network Find

TECHMEME SPONSOR POSTS:



BIG NEWS: Barack Obama | Fashion Week | Nature | Combat Sports | Silvio Berlusconi | Smarter Ideas | More...



Beal's Marketing Pilgrim, Engadget and Voices on All Things Digital

Jason Kincald / TechCrunch: It's Real: YouTube Debuts Live Streaming

Live on You Tibe Platform With Two-Day Test — The rumors were true: after months — years, even — of speculation, YouTube is

stream their video directly to users in real-time. Discussion: Music Ally, Mashable!, Telegraph, Black Web 2.0, CNET News and

Mark Zuckerberg opens up. - Mark Zuckerberg founded Facebook in his college dorm room six years ago. Five hundred million people have joined since, and eight hundred and seventy-nine of them are his friends. The site is a directory of the world's people, and a place for private citizens to create public identities.

Technologies

Daily, MediaMemo, Electricpig.co.uk and Pulse2, Thanks: atul

Erick Schonfeld / TechCrunch: Google Reported To Buy Quicksee For \$10 Million. ... Discussion: Google Maps Mania and I4U News

myTouch HD Internal Doc Hints Toward FFC And America's Largest "4G" Network - While most of the Android attention at T-Mobile these days is focused on the T-Mobile G2, the myTouch HD waits in the wings. We just received, via one of our ninjas, the above document showing off some early details on the myTouch HD. Discussion: Boy Genius Report. Android Phone Fans. MobileCrunch. IntoMobile



Some later remarked that those who had fallen had made one brave final decision to take









10:25 AM ET. September 13, 2010

About | Preferences

20 minutes ago

55 minutes ago

1½ hours ago

2½ hours ago

214 hours ago

» Extend timeline

Microsoft

BizSpark

Search



Rejecting the advice of his departing budget director Peter Orszag, Obama has insisted that the Bush era tax cuts, which expire this year, be extended for "only" about 98 percent of Americans, but not for households making over \$250,000 a year. Hard to argue with that, but watch the GOP try. The more the Republicans hold hostage this plan for tax relief for millionaires, the more voters appreciate whose side they are really on. Obama has belatedly proposed a \$50 billion infrastructure program, to put Americans back to work. He should have

FEATURED OBLOG POSTS

Lt. Dan Choi... Larry Flynt...

Harry Shearer...



MARK BLUMENTHAL Delaware GOP

Comments | Banks

Tea Party Candidate Could Pull Off Stunning Upset Comments (1,103) | 2010 Elections

Former Nevada Senate Candidate: GOP

LICK LES Concert Tickets

FIND The new YELLOWPAGES.COM

IF YOU HAVE SOMETHING TO SAY ... SAY IT ON HUFFPOST

ADVERTISE ON HUFFPOST



TOP ITEMS: Chris Dale / YouTube Blog:

Testing, testing...YouTube begins trial of new live streaming platform - From U2 to the Indian Premier League to the

White House to E3, we've worked closely with our partners to give you a front row seat to a wide array of live events. Today and tomorrow, tune in as we open a new chapter of YouTube live streaming.

Discussion: PC World, Beet.TV, Fortune, Inquirer, 9 to 5 Mac, Techland, eWeek, USA Today, Newlaunches.com, Softpedia News, SlashGear, The Next Web, 14U News, Techie Buzz, Download Squad, NewTeeVee, V3.co.uk, SocialTimes.com, Google Operating System, Andy



Electricpig.co.uk

Jose Antonio Vargas / New Yorker:

Discussion: Silicon Alley Insider, BoomTown, The Huffington Post, Anil Dash, FM Blog and UMBC ebiguity

Google buying second Israeli startup: Quiksee - Google is buying its second Israeli startup: Quiksee, based in Or Yehuda. The deal is estimated at \$10 million. Both Google Israel and Quiksee refused to comment. - Quiksee, also known as MentorWave







Discussion: VatorNews, Screenwerk, Search Engine Land, Fortune, Softpedia News, Tech Trader

RELATED:

David / TmoNews









Some Held Hands as They Jumped — Some held hands as they jumped. Others went alone.





ELLYJEAN /FLICKR Guy Grimland / Haaretz:

Provenance Problem



Automated Aggregator BlogAgg wants to

- determine the correct originator of an item
- ascertain if it can reuse an image
- establish if the image was modified
- provide confidence to the end-user

TBL's "Oh yeah?" button





Behind the scene





Definition of provenance



- 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.

A fairly recent community





Source Luc Moreau. The foundations for provenance on the web. *Foundations and Trends in Web Science*, 2(2-3):99-241, November 2010.

W3C Incubator Group Use Cases



- Result Differences
- Anonymous Information
- Information Quality Assessment for Linked Data
- Timeliness
- Simple Trustworthiness Assessment
- Ignoring Unreliable Data
- Answering user queries that require semantically annotated provenance
- Provenance in Biomedicine
- Closure of Experimental Metadata
- Provenance Tracking in the Blogosphere
- Provenance of a Tweet
- Provenance and Private Data Use
- Provenance of Decision Making in Emergency Response
- Provenance of Collections vs Objects in Cultural Heritage
- Provenance at different levels in Cultural Heritage
- Locating Biospecimens With Sufficient Quality

- Identifying attribution and associations
- Determining Compliance with a License
- Documenting axiom formulation
- Evidence for public policy
- Evidence for engineering design
- Fulfilling Contractual Obligations
- Attribution for a versioned document
- Provenance for Environmental Marine Data
- Crosswalk Maintenance
- Metadata Merging
- Mapping Digital Rights
- Computer Assisted Research
- Handling Scientific Measurement Anomaly
- Human-Executed Processes
- Semantic disambiguation of data provider identity
- Hidden Bug
- Using process provenance for assessing the quality of Information products

http://www.w3.org/2005/Incubator/prov/wiki/Use_Cases

Dimensions Summary



Category	Dimension	Description				
Content	Object	The artifact that a provenance statement is about.				
	Attribution	The sources or entities that contributed to create the artifact in question.				
	Process	The activities (or steps) that were carried out to generate or access the artifact at hand.				
	Versioning	Records of changes to an artifact over time and what entities and proces were associated with those changes.				
	Justification	Documentation recording why and how a particular decision is made.				
	Entailment	Explanations showing how facts were derived from other facts.				
Management	Publication	Making provenance available on the Web.				
	Access	The ability to find the provenance for a particular artifact.				
	Dissemination	Defining how provenance should be distributed and its access be controlled				
	Scale	Dealing with large amounts of provenance.				
Use	Understanding	How to enable the end user consumption of provenance.				
	Interoperability	Combining provenance produced by multiple different systems.				
	Comparison	Comparing artifacts through their provenance.				
	Accountability	Using provenance to assign credit or blame.				
	Trust	Using provenance to make trust judgments.				
	Imperfections	Dealing with imperfections in provenance records.				
	Debugging	Using provenance to detect bugs or failures of processes.				

Provenance in eBook scenario



• Three use cases:

- Audit:
 - The ability for users to find out how any result was derived
 - what input influenced a result,
 - what template was used, etc.
 - Provide an explanation of results
- Reproducibility of results
 - "Animation" of execution
 - Reproducing execution, to check whether same results are reproduced
 - Re-execution with different configurations
 - Different inputs, different software package, etc
 - Returns previously computed results (static eBooks)
- Provenance of stats results shared on the Web
 - Ability to track the provenance of data beyond a given eBook
 - Ebook Import/export capabilities
 - Files: raw data, processed data, ...



Part II Open Provenance Model Overview

Provenance Challenges 1, 2, 3





- Workflow
- Provenance Questions
- Implement and run workflow, produce provenance
- Export provenance
- Import provenance
- Run queries to answer questions

Provenance Challenge

Outcomes



- Over 30 teams participated to Challenges
- Designed OPM as "lingua franca" for provenance
- Open source governance model for OPM
- Promotion of "profiles" to specialize OPM to specific application domains
- Open Community work led to OPM1.1

eBook scenario



- A user loads an eBook, and begins reading it
- The user activates a region of interest
- The user provides some inputs and chooses a data set file
- The eBook reader initiates the execution of the template, and automatically produces a model, equations, graphs and a summary
- The user saves the eBook and quits the reading session
- Later, the user resumes the reading and wants to find out how the graphs were produced

Provenance of eBook results





THE OPEN PROVENANCE MODEL (OPM)



Nodes



- Artifact: Immutable piece of state, which may have a physical embodiment in a physical object, or a digital representation in a computer system.
- Process: Action or series of actions performed on or caused by artifacts, and resulting in new artifacts.
- Agent: Contextual entity acting as a catalyst of a process, enabling, facilitating, controlling, affecting its execution.













Edge labels are in the past to express that these are used to describe past execution

Illustration





- Process "used" artifacts and "generated" artifact
- Edge "roles" indicate the function of the artifact with respect to the process (akin to function parameters)
- Edges and nodes can be typed

Causation chain:

- P was caused by A1 and A2
 A3 and A4 were caused by P
- Does it mean that A3 and A4 were caused by A1 and A2?

Explicit Data Derivations





Revisiting eBook Provenance





Accessing provenance

Equation rendering:





After you submit the input of explanatory variables, you should find that Stat-JR has produced a nicely-formatted mathematical description of the model (in LaTeX code), and a variant of the model specification language associated with the WinBUGS package.

Introduction

Results: Summary and Graphs

Graph beta 1

Chapter 1: 1-level model

-Graph beta 0

Chapter 2: Multi-level models

Introduction, input and model

Model run summary

Results



Browsing Provenance



bout Resource	Content Information Provenance Export	~
template1-code template1-code template1-equations template1-formula_beta0 template1-formula_beta1 template1- formula_deviance template1-graph_beta0 template1-graph_beta1 template1-graph_deviance template1-model template1-output template1-summary urn:uuid:f7aeea40-	The queried object with URI urn:uuid:f7b35711-ddf0-11e0-9af6-003048d59cdd was generated by the process urn:uuid:f7b070e1-ddf0-11e0-8a27-003048d59cdd was generated Today at 10:13:03 was derived from the resource urn:uuid:f7aeea40-ddf0-11e0-b39c-003048d59cdd urn:uuid:dadcce30-be80-11e0-a765-003048d59cdd#1levelMod urn:uuid:dadcce30-be80-11e0-a765-003048d59cdd#bang	

for (i in 1:length(use)) {

Addressing the Use Cases



- Audit
- Reproducibility
 - Backend checks whether an execution has already run for some inputs, and avoids use



Part III Towards a Standard for Provenance

An old idea of mine ...

PROVENANCE Enabling and Supporting Provenance in Grids for Complex Problems

Contract Number: 511085



Standardisation of Provenance Systems in Service Oriented Architectures White Paper

Authors:Luc Moreau and John IbbotsonType:DeliverableVersion:1.0Version:March 29, 2006Status:public

Abstract

This White Paper presents *provenance* in computer systems as a mechanism by which business and e-science can undertake compliance validation and analysis of their past processes. We discuss an open approach that can bring benefits to application owners, IT providers, auditors and reviewers. In order to capitalise on such benefits, we make specific recommendations to move forward a standardisation activity in this domain.

Charter



Provenance Interchange Working Group Charter

The **mission** of the <u>Provenance Working Group</u>, part of the <u>Semantic Web Activity</u>, is to support the widespread publication and use of provenance information of Web documents, data, and resources. The Working Group will publish W3C Recommendations that define a language for *exchanging* provenance information among applications.

Join the Provenance Working Group.

End date	1 October 2012				
Confidentiality	Proceedings are public				
Initial Chairs	Luc Moreau, University of Southampton Paul Groth, VU University Amsterdam				
Initial Team Contacts (FTE %: 20)	Sandro Hawke				
Usual Meeting Schedule	Teleconferences: Weekly Face-to-face: Once Annually				

Deliverables

- **D1. PIL Conceptual Model** (*W3C Recommendation*). This document consists of a natural language description and a graphical illustration of concepts involved in PIL. Such a document will help broaden the appeal and uptake of provenance beyond the community of technical experts.
- **D2. PIL Formal Model** (*W3C Recommendation*). The purpose of this document is to provide a normative formalization of the conceptual model, making use of the appropriate Semantic Web languages like RDFS or OWL.
- **D3. PIL Formal Semantics** (*W3C Note, optional*). This optional note consists of a mathematical definition of PIL. It will focus on facets of formalization that have not been captured in the formal model.
- **D4. Accessing and Querying Provenance** (*W3C Note*). This document specifies how provenance can be accessed or queried in embedded documents and from remote services. Specifically, it defines how to access provenance embedded in an HTML document using RDFa, how to access provenance from a service by means of HTTP, and how to query provenance through a SPARQL endpoint.
- **D5. PIL XML Serialization** (*W3C Note*). This document specifies an XML serialization of PIL.
- **D6. PIL Best Practice Cookbook** (*W3C Note*). This document includes a limited set of best practice profiles that link with other relevant models, such as Dublin Core provenance related concepts, licensing in Creative Commons, and the OpenId identity mechanism for people.
- **D7. PIL Primer** (*W3C Note*). This educational document provides users with an easy to understand description of the model.

Timetable

Specification	FPWD	LC	CR	PR	Rec or Note
D1 (PIL Conceptual Model, Recommendation)	T+6	T+9	T+12	T+15	T+18
D2 (PIL Formal Model, Recommendation)	T+6	T+9	T+12	T+15	T+18
D3 (PIL Formal Semantics, Optional WG Note)	T+12	T+18	n/a	n/a	T+18
D4 (Accessing and Querying Provenance, WG Note)	T+9	T+18	n/a	n/a	T+18
D5 (PIL XML Serialization, WG Note)	T+9	T+18	n/a	n/a	T+18
D6 (PIL Best Practice Cookbook, WG Note)	T+15	T+18	n/a	n/a	T+18
D7 (PIL Primer, WG Note)	T+12	T+18	n/a	n/a	T+18

Membership

- Group participants:
 - 40 participants from 27 organizations
 - 11 Invited Experts
- Participants:
 - Academic institutions: 18
 - Business: 6
 - Governmental organizations: 2
 - Various consortia: 6

First Public Working Drafts

- To be released at end of September
- Three FPWDs:
 - Data model for provenance
 - Mapping to OWL2 ontology language
 - Provenance Access and Query
- Primer in preparation



CONCLUSIONS



- R
 - produces provenance-like logs
 - Team at Kent generating OPM provenance for provenance

Conclusions

- Embracing an emerging standard
- Opportunity for EStat to set the guidelines for provenance in its community
- eBooks demonstrate a compelling set of use cases
- Questions:
 - How would you like to exploit provenance?
 - How would you like it to be rendered?
 - Can we help organize organize your research outputs, your logbooks?
 - Can we advise on how to generate provenance in some stats package?

Further reading

- Provenance incubator final report, 2010 <u>http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/</u>
- Luc Moreau. The foundations for provenance on the web. Foundations and Trends in Web Science, 2(2-3): 99-241, November 2010. <u>http://eprints.ecs.soton.ac.uk/21691/</u>
- Luc Moreau, Paul Groth, Simon Miles, Javier Vazquez, John Ibbotson, Sheng Jiang, Steve Munroe, Omer Rana, Andreas Schreiber, Victor Tan, and Laszlo Varga. The Provenance of Electronic Data. *Communications of the ACM*, 51(4):52-58, April 2008. <u>http://www.ecs.soton.ac.uk/~lavm/papers/cacm08.pdf</u>

Further pointers

- OPM http://openprovenance.org/
- OPM v1.1 specification: <u>http://eprints.ecs.soton.ac.uk/21449/</u>
- OPM Tutorial: <u>http://openprovenance.org/tutorial/</u>
- OPM Wiki (community process): <u>http://twiki.ipaw.info/bin/view/OPM/</u>

Further pointers

PROV WG <u>http://www.w3.org/2011/prov/wiki/</u> <u>Main_Page</u>