Open Provenance Model

An overview of provenance and its application to eBooks

Luc Moreau
l.moreau@ecs.soton.ac.uk
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

Open Provenance Model

- 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

Open Provenance Model

Part I
Provenance 101

Provenance Overview

Open Provenance Model

- A motivating example
- Definition

Scenario: BlogAgg

Open Provenance Model

Provenance Problem

Open Provenance Model

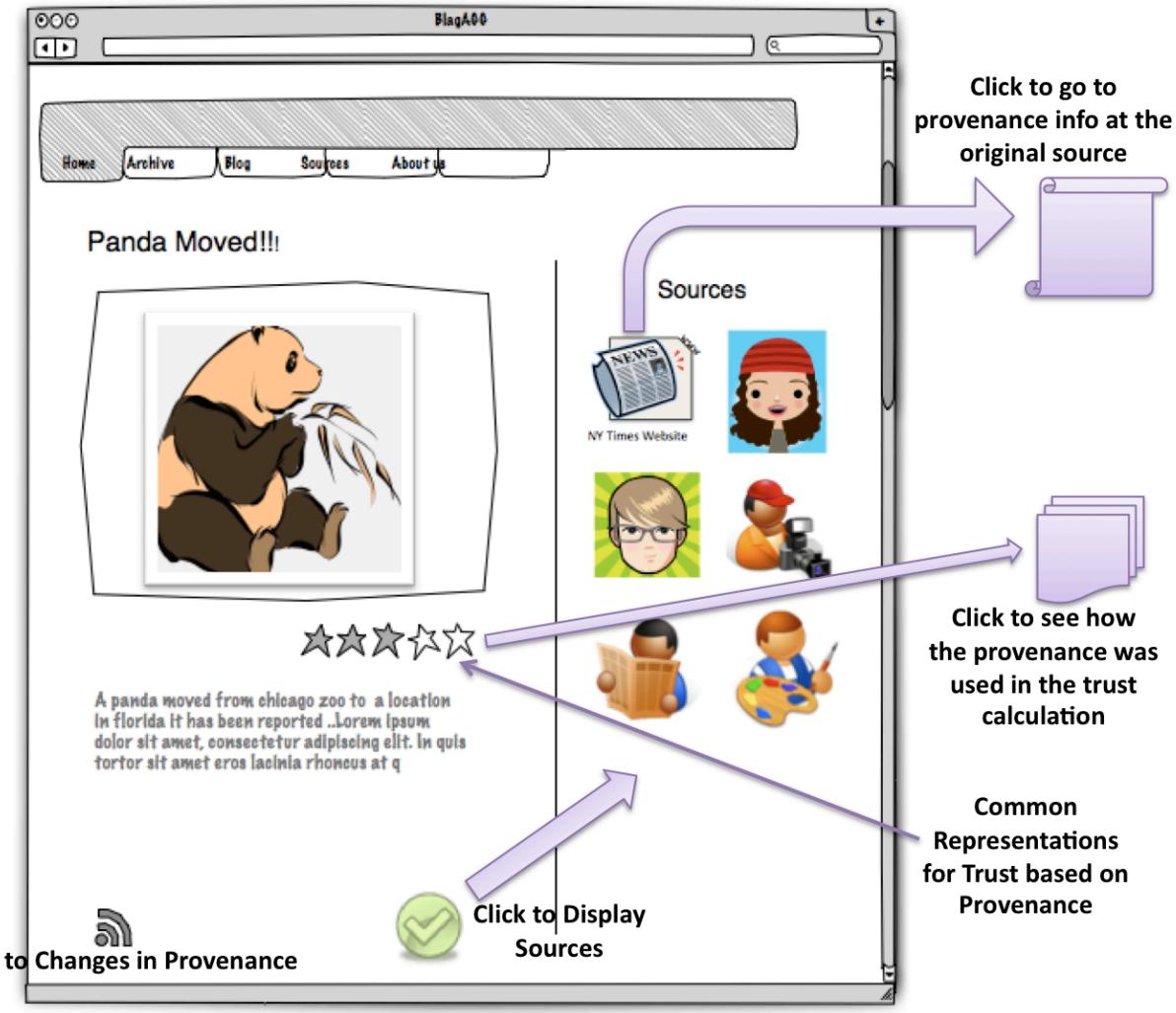
Automated Aggregator BlogAgg wants to

- 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

Open Provenance Model

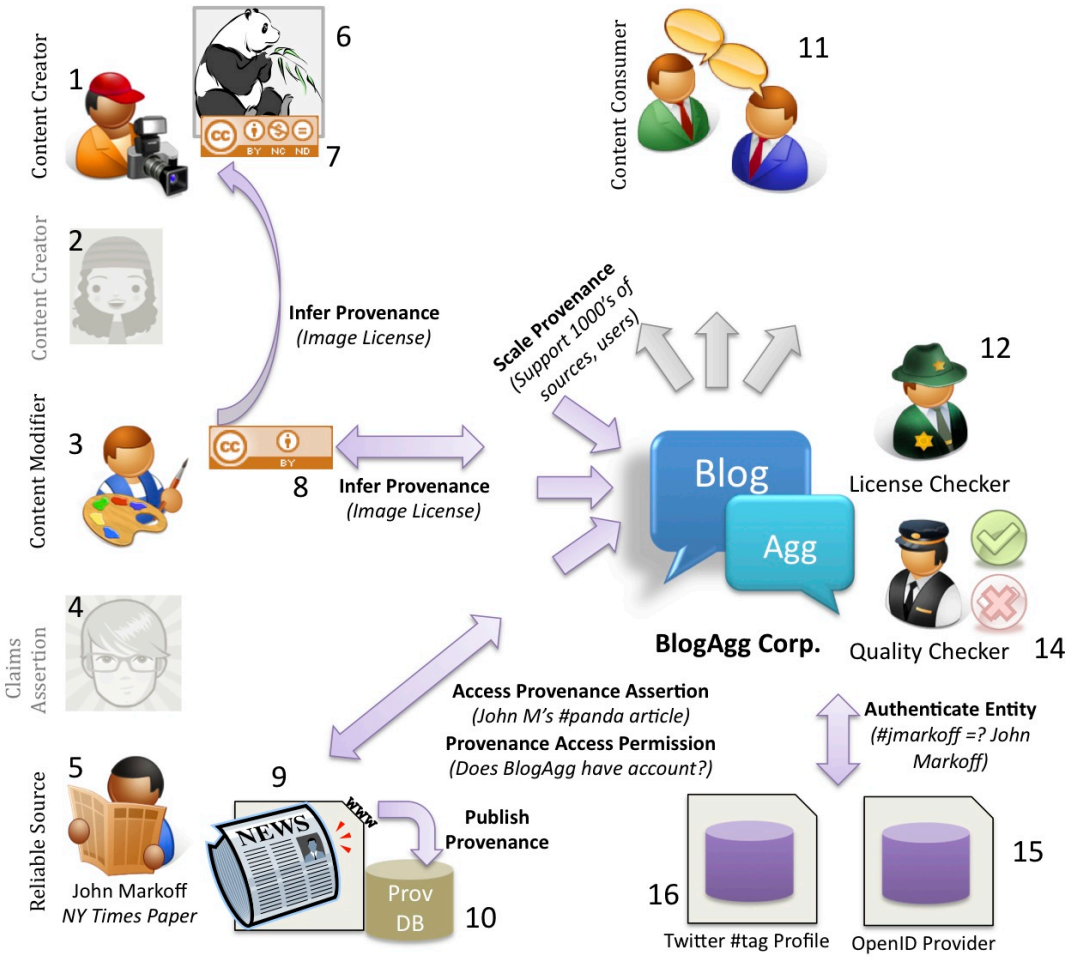
BLOGAGG : USE



Behind the scene

Open Provenance Model

BLOGAGG : MANAGEMENT



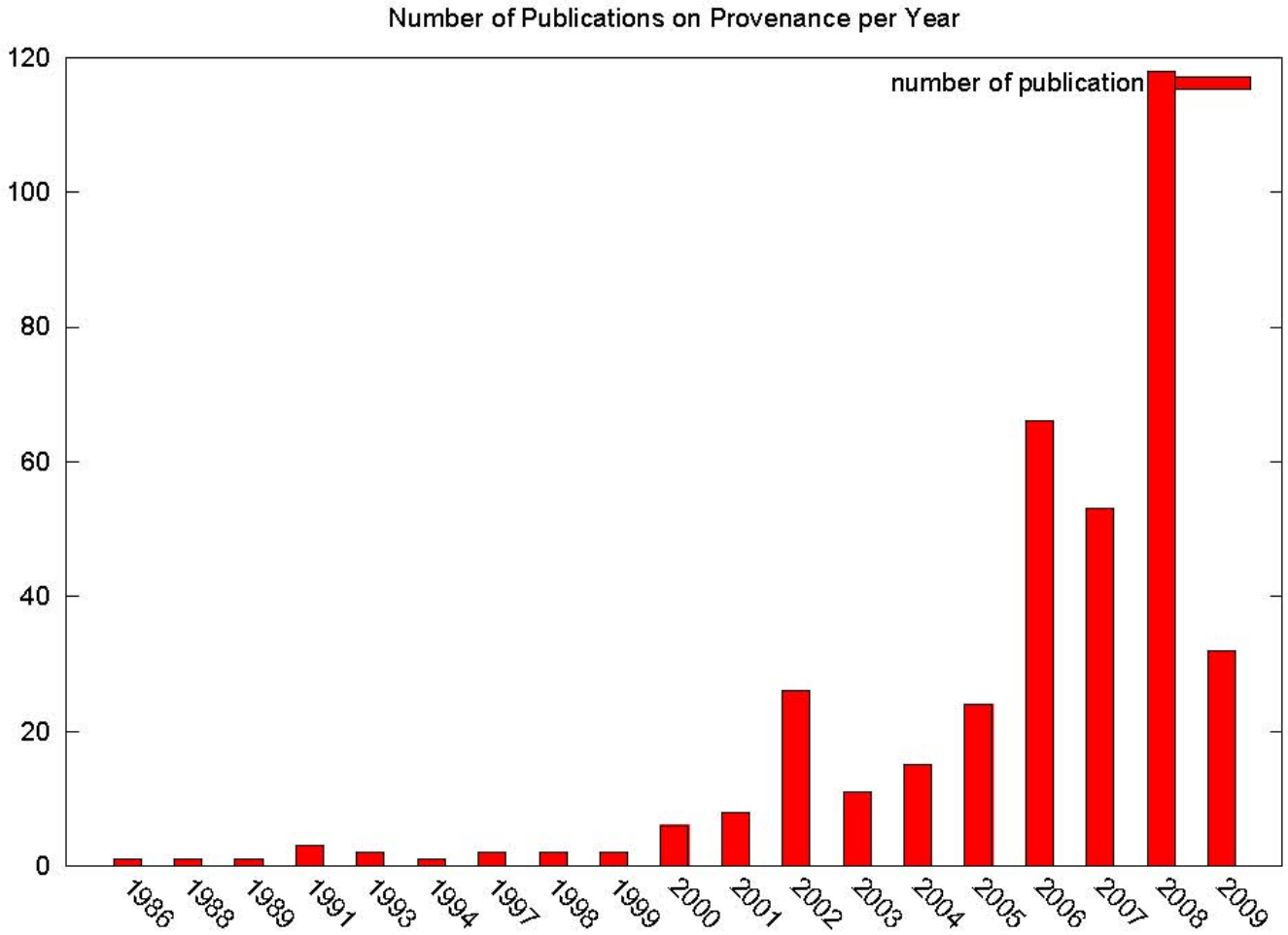
Definition of provenance

Open Provenance Model

- 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

Open Provenance Model



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

Open Provenance Model

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 processes 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

Open Provenance Model

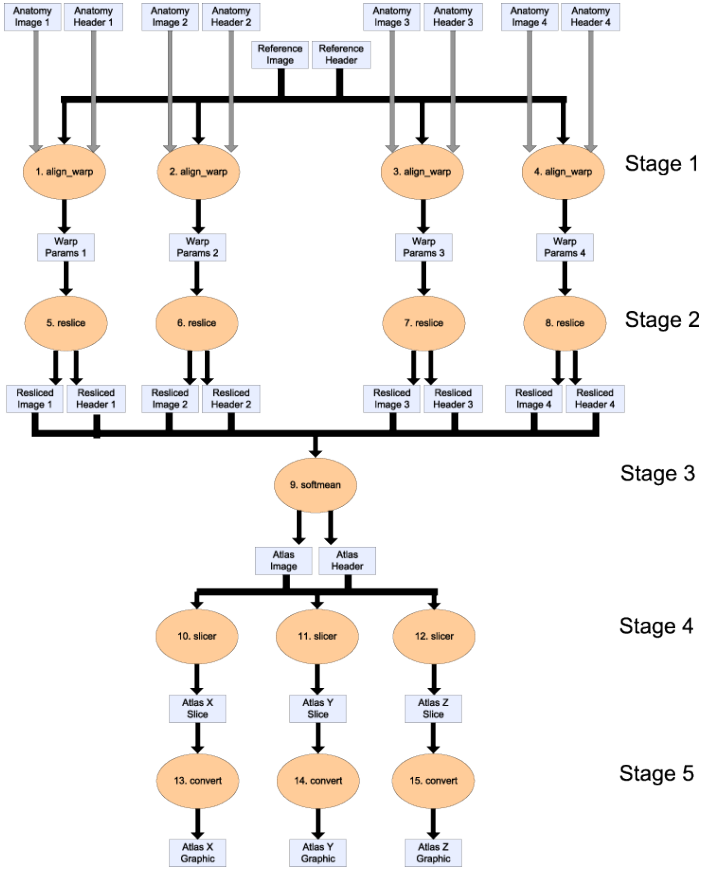
- Three use cases:
  - Audit:
    - The ability for users to find out how any result was derived
      - what input influenced a result,
      - what template was used, etc.
    - Provide an explanation of results
  - Reproducibility of results
    - “Animation” of execution
    - Reproducing execution, 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, …

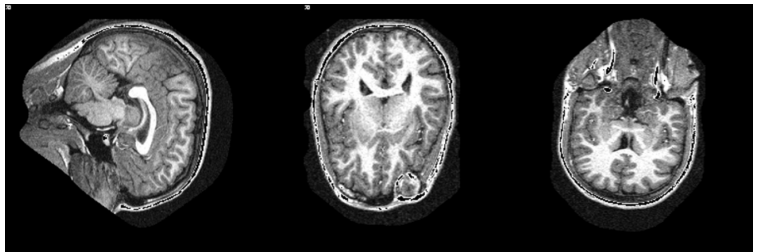
Open Provenance Model

Part II
Open Provenance Model Overview

Provenance Challenges 1, 2, 3

Open Provenance Model





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

Provenance Challenge

Outcomes

Open Provenance Model

- 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

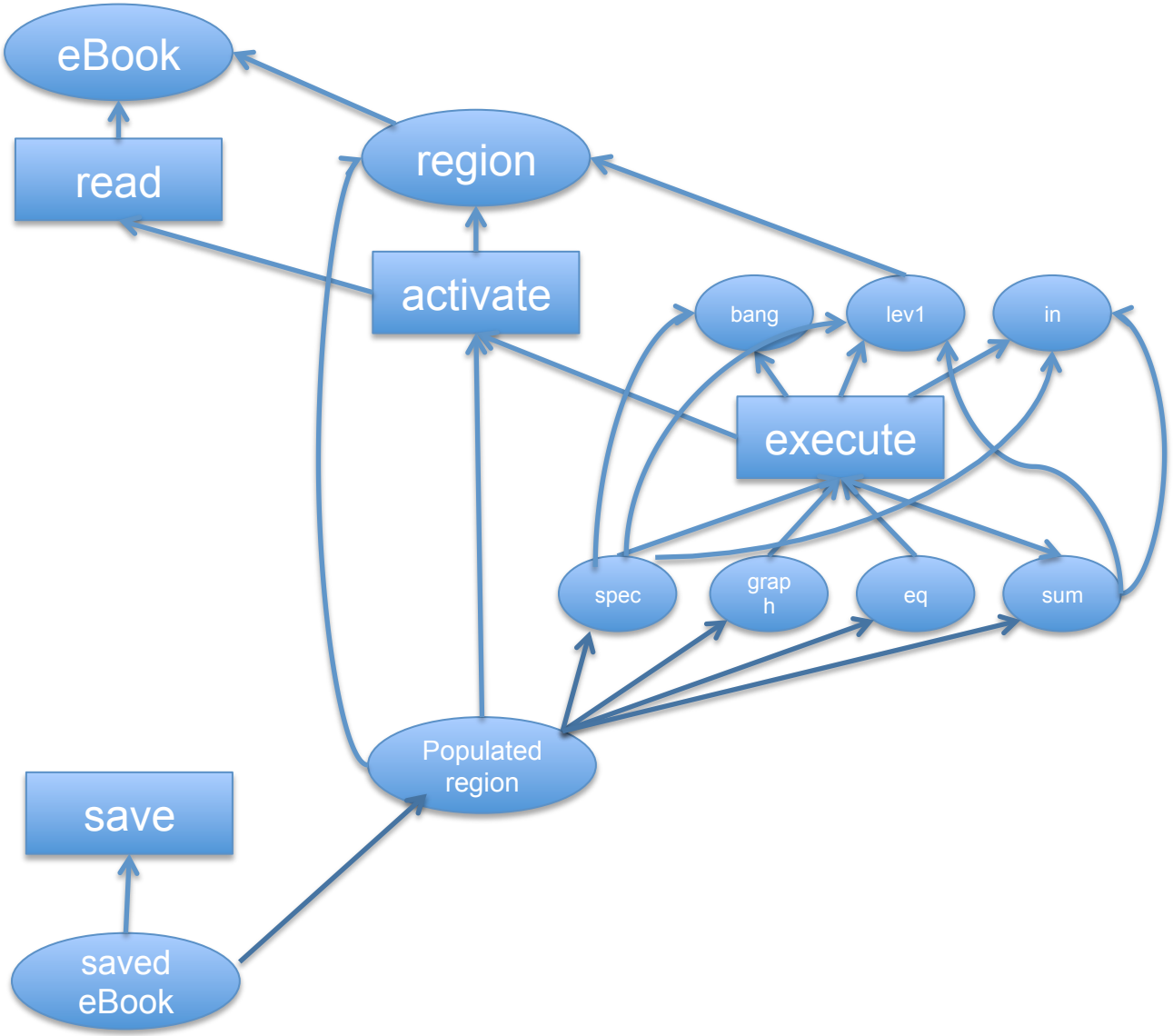
eBook scenario

Open Provenance Model

- 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

Open Provenance Model

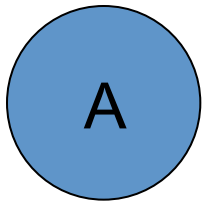


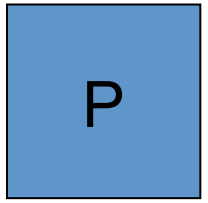
THE OPEN PROVENANCE MODEL (OPM)

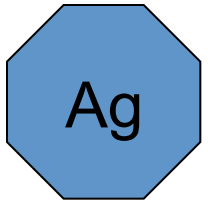
Nodes

Open Provenance Model

- 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.

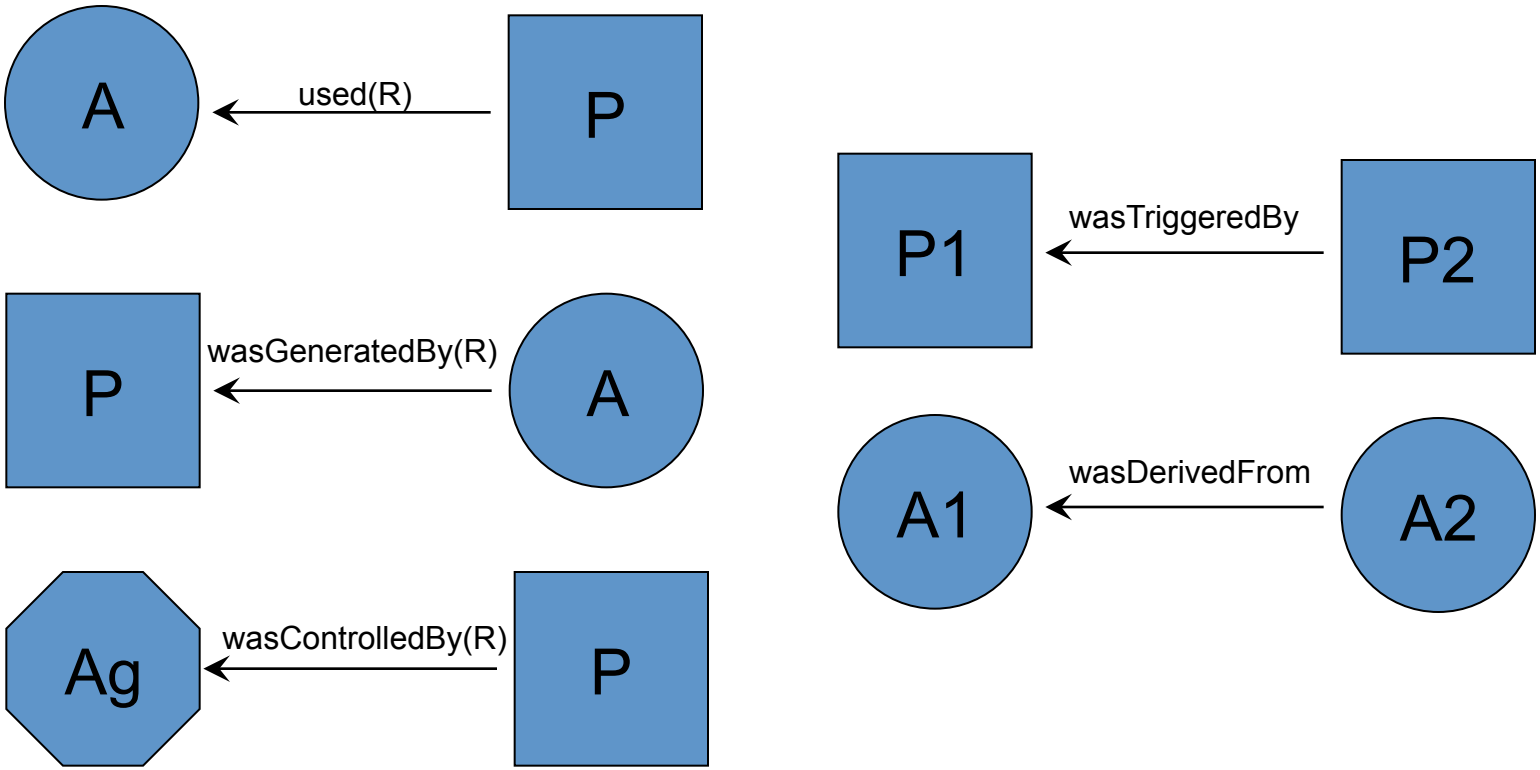






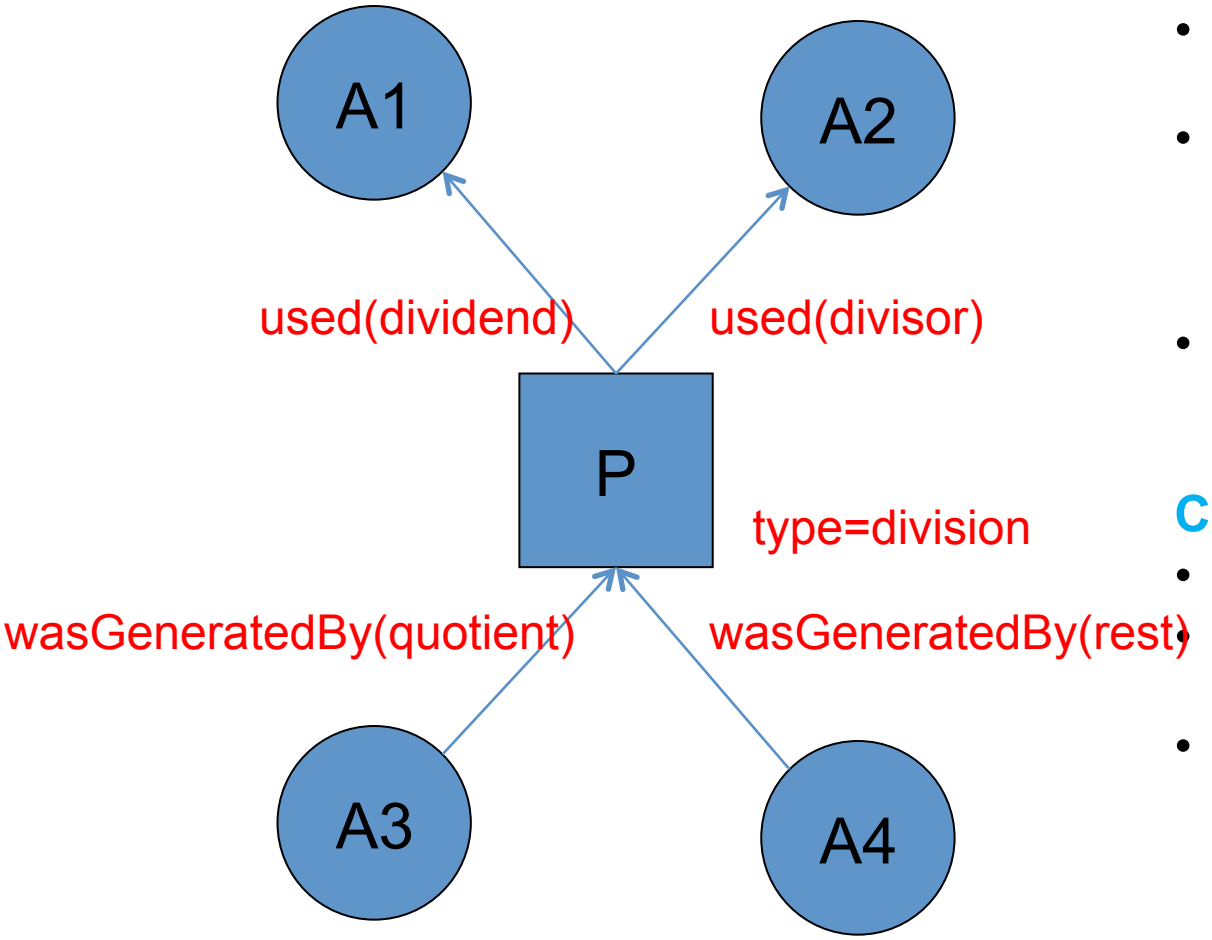
Edges

Open Provenance Model



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

Illustration



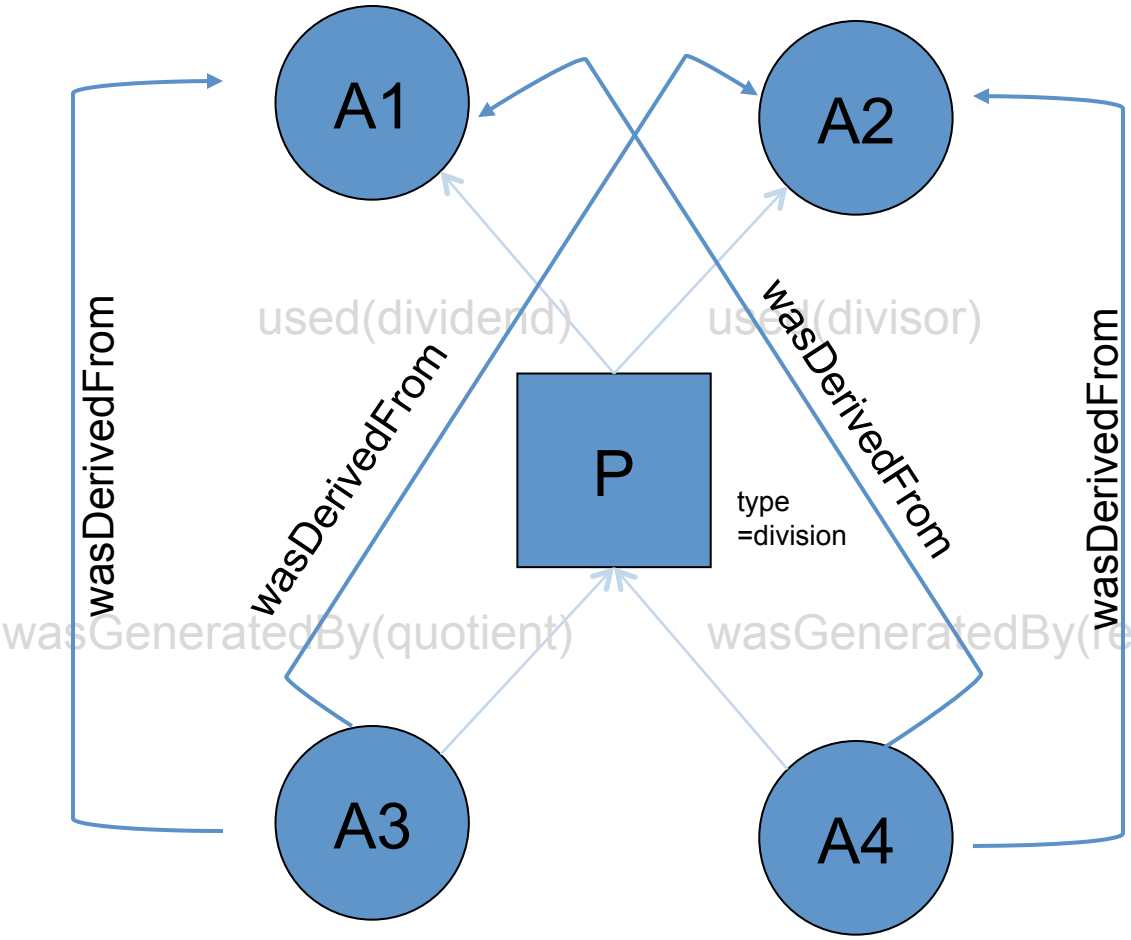
- Process “used” artifacts and “generated” artifact
- Edge “roles” indicate the function of the artifact with respect to the process (akin to function parameters)
- Edges and nodes can be typed

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

Open Provenance Model

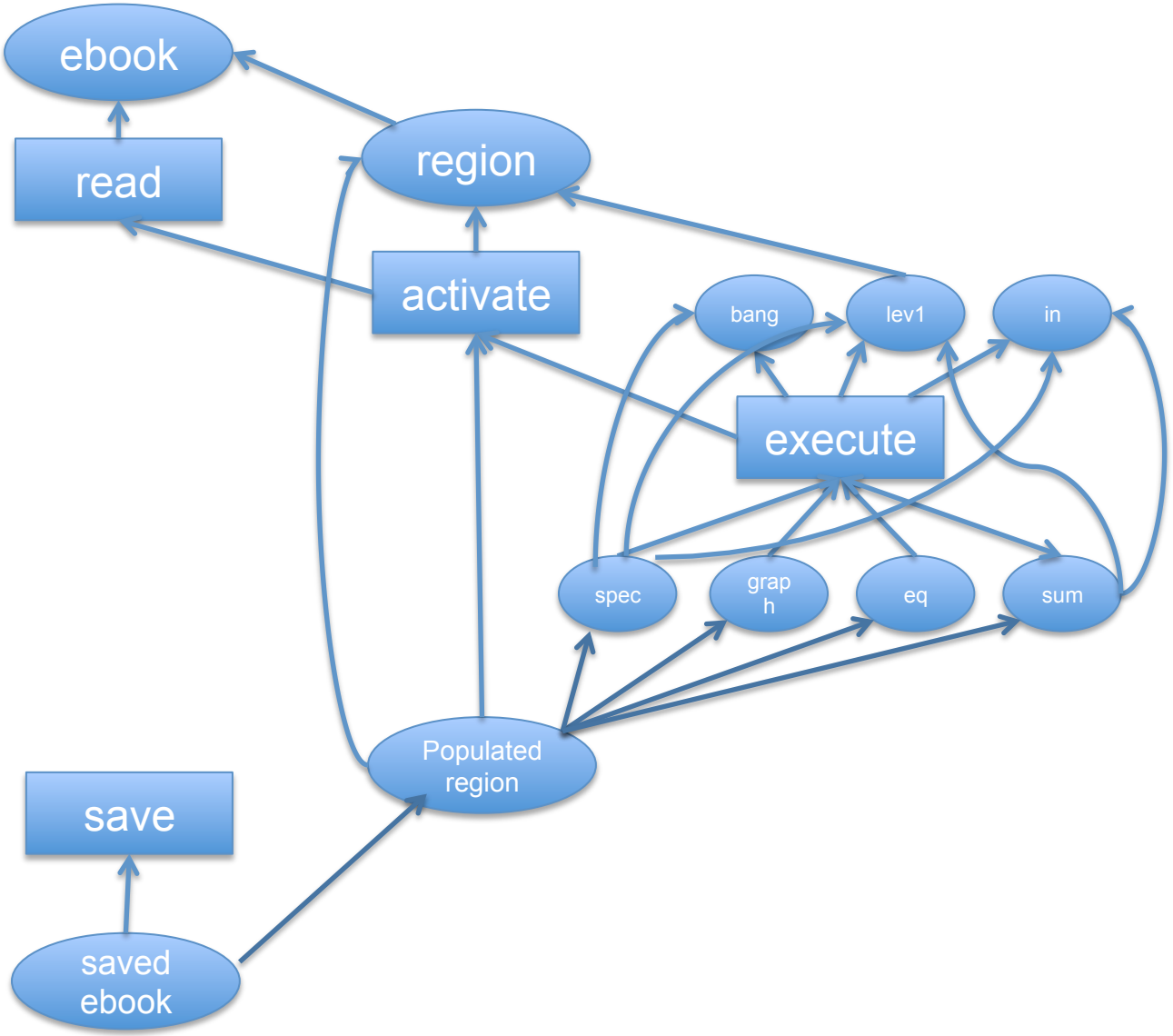


Causation chain:

- P was caused by A1 and A2
- A3 and A4 were caused by P
- A3 was caused by A1 and A2
- A4 was caused by A1 and A2

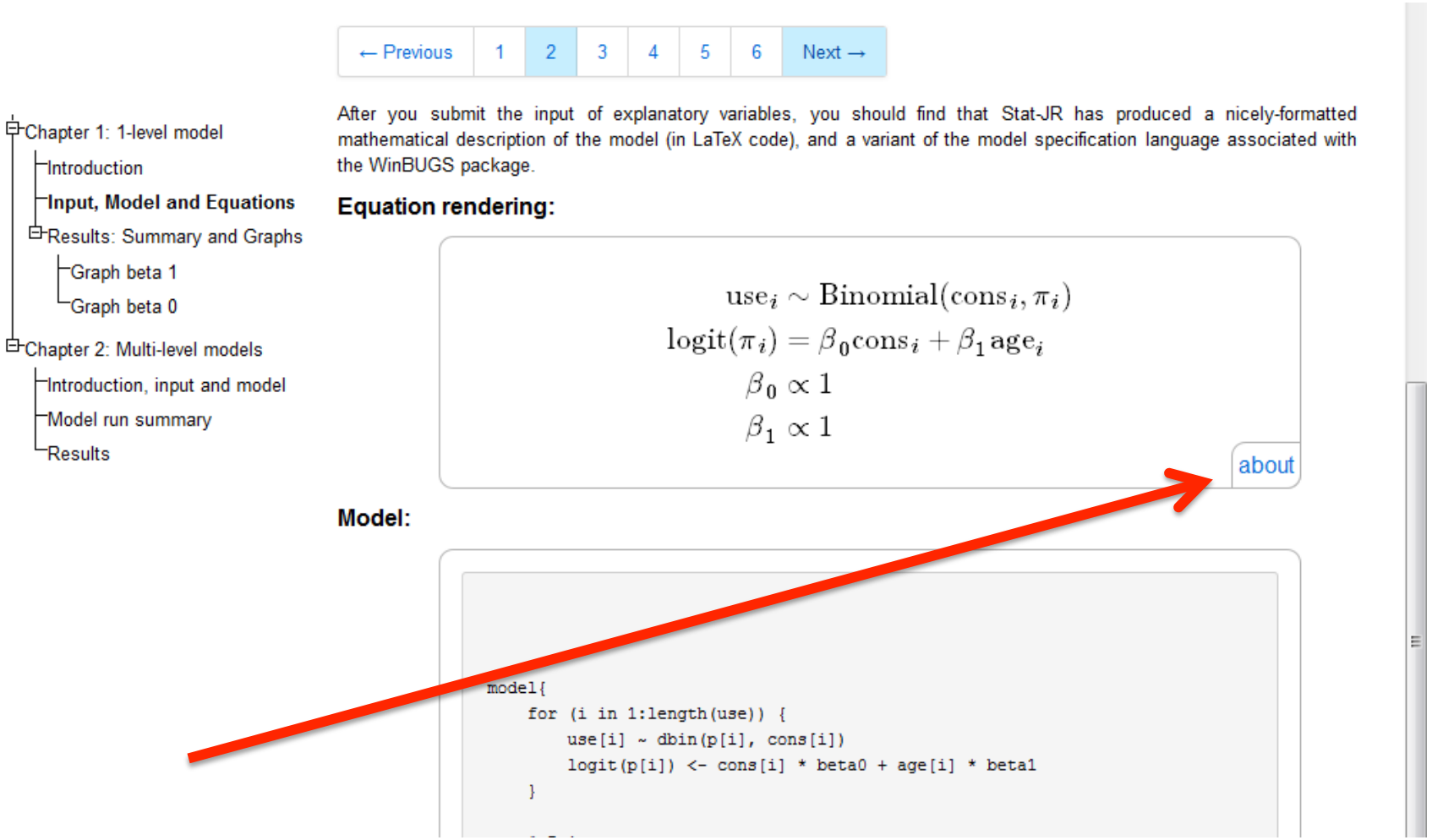
Revisiting eBook Provenance

Open Provenance Model



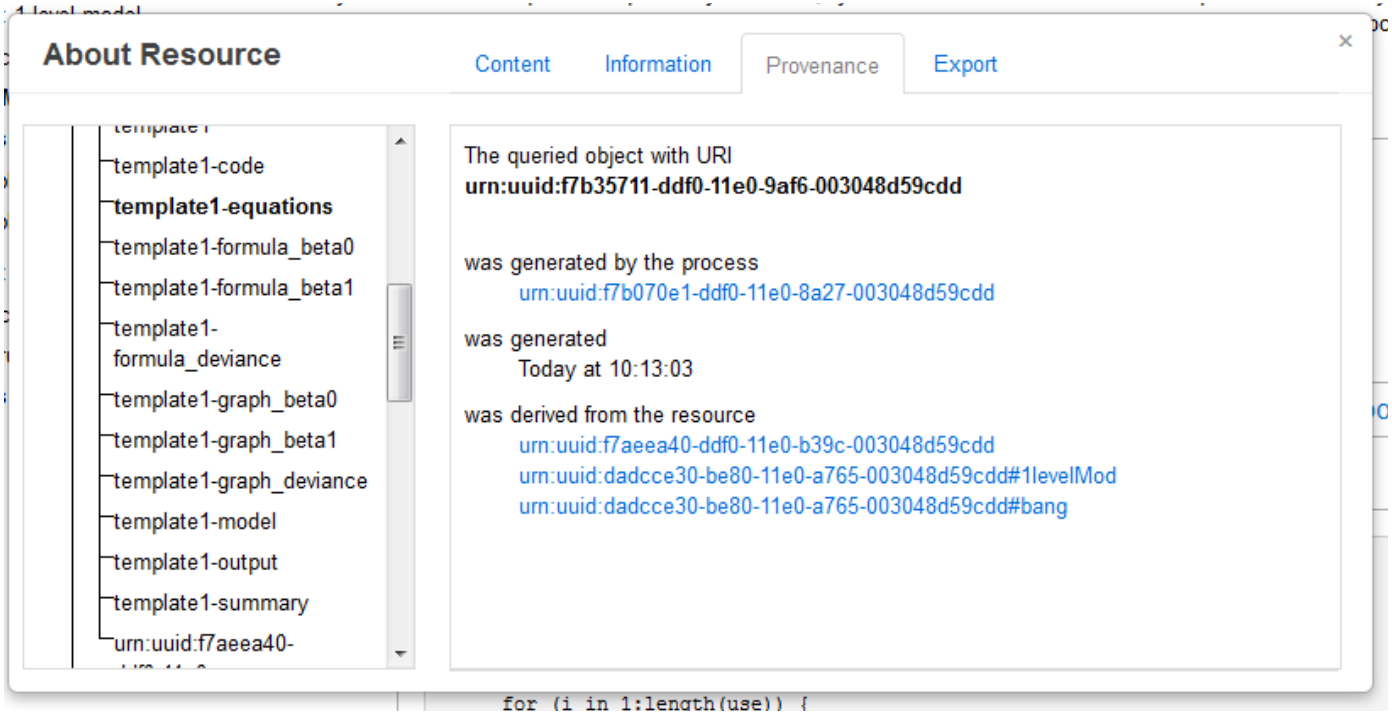
Accessing provenance

Open Provenance Model



Browsing Provenance

Open Provenance Model



Addressing the Use Cases

Open Provenance Model

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

Open Provenance Model

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 Ibbotson
Type: Deliverable
Version: 1.0
Version: March 29, 2006
Status: 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

W3C Technology and Society domain

Provenance Interchange Working Group Charter

The **mission** of the Provenance Working Group, part of the Semantic Web Activity, 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

Open Provenance Model

CONCLUSIONS

Open Provenance Model

- 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 http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/
- Luc Moreau. The foundations for provenance on the web. *Foundations and Trends in Web Science*, 2(2-3): 99-241, November 2010. http://eprints.ecs.soton.ac.uk/21691/
- 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. http://www.ecs.soton.ac.uk/~lavm/papers/cacm08.pdf

Further pointers

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

Further pointers

- PROV WG
  http://www.w3.org/2011/prov/wiki/Main_Page