W3C Provenance Incubator Group: An Overview

DRAFT

March 10, 2010

Thanks to Contributing Group Members

- <list here all group members that have contributed so far>
The Importance of Provenance

The Importance of Provenance

For the Web Architecture
- "At the toolbar (menu, whatever) associated with a document there is a button marked ‘Oh, yeah?’. You press it when you lose that feeling of trust. It says to the Web, ‘so how do I know I can trust this information?’. The software then goes directly or indirectly back to metainformation about the document, which suggests a number of reasons.” -- Tim Berners-Lee, W3C Chair, Web Design Issues, September 1997.

For Linked Data
- "Provenance is the number one issue we face when publishing government data as linked data for data.gov.uk” -- John Sheridan, UK National Archives, data.gov.uk, February 2009

For science:
- "We need a paradigm that makes it simple […] to perform and publish reproducible computational research. […] A Reproducible Research Environment (RRE) […] provides computational tools together with the ability to automatically track the provenance of data, analyses, and results and to package them (or pointers to persistent versions of them) for redistribution.” -- Jill Mesirov, Chief Informatics Officer of the MIT/Harvard Broad Institute, in Science, January 2010

For researchers in all areas of computer science
- "The number of publications on provenance is […] a total of 425 […] The first publication dates back to 1986, […] with about half the papers published in the last two years.” -- Luc Moreau, University of Southampton, in The Foundations of Provenance on the Web, November, 2009
What is Provenance

- **Provenance**: Sources of information, including entities and processes, involved in producing or delivering an artifact
- Some uses of provenance:
  - Meaningful integration of data (comparing apples with apples)
  - Making trust judgments when information sources are diverse and of varying quality (the Web!)
  - Providing justifications for conclusions
  - Establishing attribution
  - Enabling comparison and reproducibility of processes
- Provenance is ubiquitous:
  - Business practice, cultural artifacts, science applications, etc
  - Data integration, licensing and attribution, accountability

Outline

- What is Provenance
- Need for provenance
  - Open government
  - e-Science
  - Web: Linked Open Data
- W3C Provenance Group: What the community understands about provenance
  - Key dimensions of provenance
  - Requirements for provenance
- State-of-the-art
- Research challenges of provenance
The Need for Provenance is Ubiquitous

- Business practice
  - Manufacturing processes and providers of a given product
- Cultural artifacts
  - Origins, owners, processes
- Science applications
  - How new results were obtained: from assumptions to conclusions and everything in between
- Licensing and attribution
  - For a document/software that combines permissions and rights
- Web search/use
  - Making trust judgments on what web content to trust

Provenance is Key for Scientists
Provenance is Key in Open Government

data.gov, data.gov.uk

- **Open Government Initiative** (http://www.whitehouse.gov/open)
  - Release quickly, improve later
  - NSF to develop plan (http://www.nsf.gov/open):
    “NSF is developing an Open Government Plan, which will serve as the roadmap for our plans to improve transparency, better integrate public participation and collaboration into our core mission, and become more innovative and efficient.”

- **Similar initiative in the UK (data.gov.uk)**
  - “Provenance is the number one issue that we face when publishing government data in data.gov.uk” -- John Sheridan, UK National Archives

- **Why is provenance so important**
  - Government data comes from very diverse data sources
  - Varying quality
  - Different scope
  - Different assumptions

Provenance is Key in Linked Open Data

http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

- “The goal of the W3C SWEO Linking Open Data community project is to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources.”

- Adopted by many: data.gov.uk, NYT, etc

- Growing virally and very fast

As of May ’09, it included over 4.7 billion RDF triples, which are interlinked by around 142 million RDF links
Provenance is Key for Information Integration

Provenance questions useful for information integration:
- Who created that content (author/attribution)?
- Was the content ever manipulated, if so by what processes/entities?
- Who is providing that content (repository)?
- What is the timeliness of that content?
- Can any of the answers to these questions be verified (e.g., e-signatures)?

Outline

- What is Provenance
- Need for provenance
  - Open government
  - e-Science
  - Web: Linked Open Data
- W3C Provenance Group: What the community understands about provenance
  - Key dimensions of provenance
  - Requirements for provenance
- State-of-the-art
- Research challenges of provenance
Immediate Need for Provenance at W3C

- Web of trust
  - Making trust judgments based on provenance
- Reasoners
  - Attribution of assertions from diverse sources
- Linked data
  - Use of conflicting data of varying degrees of quality
- Social trust
  - Attribution, authority, propagation
- Social web
  - Privacy and use policies of sensitive (personal) data
- Life sciences and e-Science at large
  - Method capture and reproducibility of scientific results

The Role of Provenance in Web Layer Diagram
W3C Chartered a New Provenance Group

Provenance is a pressing issue in many contexts
- Linked Data and the semantic web (linkedopendata.org)
- Open government (data.gov, data.gov.uk)

Noone understands how to approach provenance
Yet many people are asking for some sort of standard that they can use immediately
Prior work scattered in many areas of computer science and library sciences research
- Databases, scientific workflow, graph structures, etc.

Resources:
- http://www.w3.org/2005/Incubator/prov/
- http://www.w3.org/2005/Incubator/prov/wiki/

W3C Provenance Group: Charter and Goals of the Incubator Group

Provide state-of-the-art understanding and develop a roadmap for development and possible standardization
Articulate requirements for accessing and reasoning about provenance information
- Develop use cases
Identify issues in provenance that are direct concern to the Semantic Web
- Articulate relationships with other aspects of Web architecture
Report on state-of-the-art work on provenance
Report on a roadmap for provenance in the Semantic Web
- Identify starting points for provenance representations
- Identifying elements of a provenance architecture that would benefit from standardization
W3C Provenance Group: Products of the Group to Date

- Group formed in September 2009
  - All information is public: http://www.w3.org/2005/Incubator/prov/wiki/
- Developed a set of key dimensions for provenance (11/09)
  - Grouped into three major categories: content, management, use
- Developed use cases for provenance (12/09)
  - More than 30 use cases, most were improved and curated
  - Many relevant to data integration and intelligence analysis
- Developed requirements for provenance that arise from the use cases (1/10)
  - User requirements: what is the purpose/use of the provenance information
  - Technical requirements: derived from the user requirements
- Currently developing state-of-the-art report (expected 6/10)

W3C Provenance Group: Use Cases Developed to Date

1. Result Differences
2. Anonymous Information
3. Information Quality Assessment for Linked Data
4. Timeliness
5. Simple Trustworthiness Assessment
6. Ignoring Unreliable Data
7. Answering user queries that require semantically annotated provenance
8. Provenance in Biomedicine
9. Closure of Experimental Metadata
10. Locating Biospecimens With Sufficient Quality
11. Using process provenance for assessing the quality of Information products
12. Provenance Tracking in the Blogosphere
13. Provenance of a Tweet
14. Provenance and Private Data Use
15. Provenance of Decision Making in Emergency Response
16. Provenance of Collections vs Objects in Cultural Heritage
17. Provenance at different levels in Cultural Heritage
18. Identifying attribution and associations
19. Determining Compliance with a License
20. Documenting axiom formulation
21. Evidence for public policy
22. Evidence for engineering design
23. Fulfilling Contractual Obligations
24. Attribution for a versioned document
25. Provenance for Environmental Marine Data
26. Crosswalk Maintenance
27. Metadata Merging
28. Mapping Digital Rights
29. Computer Assisted Research
30. Handling Scientific Measurement Anomaly
31. Human-Executed Processes
32. Semantic disambiguation of data provider identity
33. Hidden Bug
W3C Provenance Group:
Major Dimensions of Provenance

Content
- Attribution - provenance as the sources or entities that were used to create a new result
  - Responsibility - knowing who endorses a particular piece of information or result
  - Origin - recorded vs reconstructed, verified vs non-verified, asserted vs inferred
- Process - provenance as the process that yielded an artifact
  - Reproducibility (e.g. workflows, mashups, text extraction)
  - Data Access (e.g. access time, accessed server, party responsible for accessed server)
- Evolution and versioning
  - Republishing (e.g. retweeting, reblogging, republishing)
  - Updates (e.g. a document with content from various sources and that changes over time)
- Justification for decisions - Includes argumentation, hypotheses, why-not questions
- Entailment - given the results to a particular query, what axioms or tuples led to those results

Management
- Publication - Making provenance information available (expose, distribute)
- Access - Finding and querying provenance information
- Dissemination control - Track policies specified by creator for when/how an artifact can be used
  - Access Control - incorporate access control policies to access provenance information
  - Licensing - stating what rights the object creators and users have based on provenance
  - Law enforcement (e.g. enforcing privacy policies on the use of personal information)
- Scale - how to operate with large amounts of provenance information

Use
- Understanding - End user consumption of provenance.
  - abstraction, multiple levels of description, summary
  - presentation, visualization
- Interoperability - combining provenance produced by multiple different systems
- Comparison - finding what's in common in the provenance of two or more entities (e.g. two experimental results)
- Accountability - the ability to check the provenance of an object with respect to some expectation
  - Verification - of a set of requirements
  - Compliance - with a set of policies
- Trust - making trust judgments based on provenance
  - Information quality - choosing among competing evidence from diverse sources (e.g. linked data use cases)
  - Incorporating reputation and reliability ratings with attribution information
- Imperfections - reasoning about provenance information that is not complete or correct
  - Incomplete provenance
  - Uncertain/probabilistic provenance
  - Erroneous provenance
  - Fraudulent provenance
- Debugging
Outline

- What is Provenance
- Need for provenance
  - Open government
  - e-Science
  - Web: Linked Open Data
- W3C Provenance Group: What the community understands about provenance
  - Key dimensions of provenance
  - Requirements for provenance
- Research challenges of provenance
- State-of-the-art

Challenges of Representing and Managing Provenance: 1) Content Representation

- Need representations of processes (workflows), entities, roles, data collections, meta-assertions, etc.
Challenges of Representing and Managing Provenance: 2) Granularity

- Granularity of provenance record keeping
  - How much detail is useful, manageable/scalable in practice?
    - Size of provenance can be orders of magnitude larger than base data


Challenges of Representing and Managing Provenance: 3) Policies for Information Quality and Trust

- Policies based on provenance information
  - Association-based policies
    - Source is NYT, source cites NYT
    - Source is cited in Wikipedia
  - Bias-based policies
    - Source is an oil company
  - Distrust policies
    - Source is a blog

- Policies may be restricted to a context
  - Topic of search, topics of page, tags of page

- Trust policies may be shared across users
  - Like bookmarks in del.icio.us
Challenges of Representing and Managing Provenance: 4) Evolution and Updates

- Shelf time of data
  - Determine when data becomes obsolete based on provenance info
- Evolution and versioning of data sources
  - Relate updates of data based on provenance info
- Reproducibility of processes (workflow)

Challenges of Representing and Managing Provenance: 5) Many Unexplored Research Areas

- Managing separation of provenance (metadata) and base data
- Assigning provenance to data collections
  - Subsets and regrouping
- Scalable provenance querying
- Incomplete/incorrect provenance
  - Inferred provenance by a 3rd party
- Verification of provenance information
  - Electronic signature, deception
- Presentation to end user
  - Abstractions and visualizations appropriate to understand provenance
Outline

- What is Provenance
- Need for provenance
  - Open government
  - e-Science
  - Web: Linked Open Data
- W3C Provenance Group: What the community understands about provenance
  - Key dimensions of provenance
  - Requirements for provenance
- Research challenges of provenance
- State-of-the-art

Prior Research on Provenance

Provenance group is currently working on an overview of the state-of-the-art in provenance, covering:

- Workflow Systems
  - Reproducibility
- Databases
  - Query derivation, aggregations of data, streaming
- Knowledge representation and reasoning
  - Justification and explanation of reasoning
- Argumentation
  - What is taken into account to make a judgment
- Information retrieval
  - Q/A when documents are contradictory or complementary
- Security: electronic signatures