Provenance and Trust

José Manuel Gómez Pérez

INCO-TRUST Workshop
4th–5th 05/2010
Why provenance is important

- **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 2010*

- **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."
Provenance is...

- Records of
  - Sources of information, including entities and processes, involved in producing or delivering an artifact
  - History of subsequent owners (change of custody)
Provenance is...

- Evidence of authenticity, integrity, and information quality
- Certifies products of good process
Provenance is...

- Valuable
- Hard to collect and verify
- Necessary to assign credit
- ...and blame
- i.e. to establish Trust
Provenance models define
- Types of provenance elements
- Relationships between them

Provenance represented as a graph
- Nodes: Pieces of provenance information (Artifacts, Processes, Agents)
- Edges: relate provenance elements to each other

Additionally: roles, temporal, dependency, causal dependency

The Open Provenance Model (OPM)
Provenance-related vocabularies

**General**
- DC – Dublin Core Metadata Terms
- FOAF – Friend of a Friend
- SIOC – Semantically-Interlinked Online Communities
- SWP – Semantic Web Publishing vocabulary
- VOID – Vocabulary of Interlinked Datasets

**Provenance-specific**
- OPM - Open Provenance Model
- Provenir
- PRV - Provenance Vocabulary
- PML - Proof Markup Language
- DC - Dublin Core
- PREMIS
- WOT – Web Of Trust schema
- PAV - SWAN Provenance Ontology
- SWP - Semantic Web Publishing Vocabulary
- CS - Changeset Vocabulary
## A snapshot on provenance

### Uses

- Meaningful **data integration**
- Establishing **trust** when information sources are diverse and of varying quality e.g. in the Web
- Providing **justification** to the conclusions of reasoning processes
- **Attribution** (who did what)
- Enabling comparison and reproducibility of processes

### Research areas

- **Scientific workflows**
- **Databases** (where and why provenance)
- **Security** (Information flow and trust)
- Justification and Argumentation in **KRR**
- **Information Retrieval** (analysis of information sources)

### Domains

- **Science**
- Open Government and Linked Data
- Web search & use
- Cultural heritage
- Licensing and attribution
- Manufacturing
The Linked Data paradigm

How can we exploit all the available data?

- Data reuse and remix
- Common flexible and usable APIs
- Standard vocabularies to describe interlinked datasets
- Tools
- Realize the Semantic Web vision

Tim Berners Lee, 2006 (Design Issues)

1. Use URIs to identify things
   - Anything, not just documents
2. Use HTTP URIs for people to lookup such names
   - Globally unique names
   - Distributed ownership
3. Provide useful information in RDF upon URI resolution
4. Include RDF links to other URIs
   - Enable discovery of related information
Provenance is key in Linked Data

- **Data Lineage**
  - The origins of data
  - Related artifacts and actors

- **Information quality**
  - Timeliness
  - (Semantic) consistency between datasets
  - Stable and meaningful data links

- **Trust**
  - Data authenticity
  - Reliability
Provenance is key in Open Government

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 we face when publishing government data as linked data for 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
Large amounts of data and processes

- In 2006, the size of the digital universe was estimated in 161 exabytes
- 3 million times the information in all books ever written
- In 2010, expected to turn 988 exabytes
- …and all this data is potentially published online

W3C Provenance Incubator Group
It’s all about producing and consuming information...

Automated means to evaluate Information Quality and Trust are needed!

W3C Provenance Incubator Group
Provenance and Trust in the Web: A real-life example

Reusing web data without the means that allow contrasting its provenance can be harmful, especially in sensitive domains.

- Two fake web sites
- A fake Wikipedia entry
- Fake California public safety phone numbers
- Fake local TV station

The hoax caused a 1000-word tome on Frankfurter Allgemeine Zeitung… and public apologies from DPA

- Trust on Wikipedia misled DPA
- In a provenance-aware world, DPA would have had means based on data provenance to evaluate trust
  - Bluewater did not exist
  - The Berlin Boys do not exist
Trust evaluation: Useful provenance questions

- Who created that content (author/attribution)?
- Was the content ever manipulated, if so by what processes/entities?
- Who is providing that content (repository)?
- Can any of the answers to these questions be verified e.g., through e-signatures?
Immediate need for provenance at W3C

Trust-driven

- **Web of trust**
  - Making trust judgments based on provenance
- **Social trust**
  - Attribution, authority, propagation
- **Social web**
  - Privacy and use policies of sensitive (personal) data

Others

- **Reasoning**
  - Attribution of assertions from diverse sources
- **Linked data**
  - Use of conflicting data of varying degrees of quality
- **Life sciences and e-Science at large**
  - Reproducibility of scientific results
W3C Provenance Group: Charter and goals of the incubator group

- 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

Provide state-of-the-art understanding and develop a roadmap for development and possible 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](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
- 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)
Provenance information needs to be:

- Represented
- Captured and recorded
- Stored and secured, queried, and reasoned about
- Visualized and browsed
Scientific and technical challenges of provenance (II)

- **Vocabularies** for representation of provenance content
  - Need representations of processes (workflows), entities, roles, data collections, meta-assertions, etc.
  - The Open Provenance Model (OPM): [http://twiki.ipaw.info/bin/view/OPM](http://twiki.ipaw.info/bin/view/OPM)

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

- **Provenance evaluation for Information Quality and Trust Management**

- **Evolution and updates**
  - Shelf timeliness of data
    - Determine when data becomes obsolete based on provenance info
  - Versioning of data sources
    - Relate updates of data based on provenance info
  - Reproducibility of processes

- **Provenance-aware visualization, navigation, and resource consumption**
Scientific and technical challenges of Provenance & 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
How to incentivize provenance take-up in the Web architecture so that content and service providers move into a provenance-aware paradigm?
- Generation of provenance metadata by content providers
- Consumption of provenance metadata by infrastructure and applications like search engines, browsers, etc.

Objective evidence of trust in online data or service allows
- Ranking increase in search engines
- Attracting internet traffic
- Increasing automation in data and service consumption

But, can you can trust such provenance information itself?
- Non-forgery of provenance metadata
  - Authoritative agencies
Thanks for your attention!