Report on the International Provenance and Annotation Workshop (IPAW'06) 3-5 May 2006, Chicago

Rajendra Bose
University of Edinburgh

Ian Foster
Computation Institute
University of Chicago and
Argonne National Laboratory

Luc Moreau
University of Southampton

1. BACKGROUND

The *provenance* of a data item refers to its origins and processing history, while *annotation* is a term that refers to the process of adding notes or data to an existing structure. Because these terms are broad, and are used in slightly different ways by different communities, confusion is rampant. For example, consider that (1) annotating a data set with its provenance information, and (2) finding the provenance of a specific data annotation are both perfectly reasonable concepts.

To help clarify these issues and advance techniques to capture data provenance and facilitate annotation, the International Provenance and Annotation Workshop (IPAW'06) was held May 3-5, 2006 at the University of Chicago's Gleacher Center in downtown Chicago; it was co-chaired by Luc Moreau (University of Southampton) and Ian Foster (University of Chicago and Argonne National Laboratory) and included roughly 45 participants, representing about 25 organizations or projects. The workshop provided some continuity to two earlier events, the Workshop on Data Derivation and Provenance organized by Peter Buneman and Ian Foster in Chicago in 2002, and the Workshop on Data Provenance and Annotation organized by Dave Berry, Peter Buneman, Michael Wilde, and Yannis Ioannidis in Edinburgh in 2003; see Section 4 for brief notes on these previous meetings.

The single-track set of sessions during IPAW'06 [1] consisted primarily of presentations of a selection of papers refereed by the program committee, which will be published as Lecture Notes for Computer Science (LNCS) Volume 4145. The program also included two keynote talks, a discussion regarding the pros and cons of standardizing approaches for capturing and managing data provenance, an entertaining "Gong Show" to foster new and original ideas, and a wrap-up discussion about future meetings and collaborative efforts of this new and growing research community.

2. KEYNOTES AND DISCUSSIONS

In the first keynote presentation, Roger Barga discussed research at Microsoft aimed at supporting scientific workflow creation, featuring the automatic capture of provenance information and the ability to retrieve this information at different levels of granularity. In the second keynote, Juliana Freire (University of Utah) described the Vistrails system for creating versioned visual pipelines to construct scientific visualizations. Derived pipeline versions are connected by "trails"—trees that can be queried and displayed graphically.

Jim Myers (NCSA, University of Illinois, Urbana-Champaign) and Luc Moreau debated and sought audience input on whether the time was right to discuss standard models of data provenance or standard interfaces for recording, querying, and administering provenance stores. The wrap-up discussion at the end of the workshop took up this thread again, with general agreement voiced by the audience to begin a mailing list and continue the growth of this community by keeping the workshop an annual event. Participants also agreed on tentative steps to set up a "Provenance Challenge," which would include a data provenance-tracking scenario and some evaluation measures; these will enable different groups to test and compare their approaches for this scenario during the next several months (See [2]).

A mid-workshop diversion was provided by the Gong Show of outlandish, "outside-the-box" ideas, chaired by Ian Foster. Highlights included presentations tentatively linking data provenance to shoe shopping, horoscope consultation based on the time data was created, the social communication traits of 14-year old girls, divining research funds, selecting breakfast cereal, eschewing junk mail, and "date provenance."

3. SUMMARY OF SESSIONS

The three-day workshop included presentation sessions on applications and systems, semantics, workflow, and models of provenance, annotations and processes. The following sections present brief encapsulations of the presentation topics for each day, and are intended to provide a short overview of the workshop and direct interested readers to the full papers.

3.1 Day One Presentations

In the *Applications* session, Dimitri Bourilkov (University of Florida) spoke about a project to facilitate automatic logging and reuse of data analysis sessions at the UNIX command line by combining the functionality of data analysis software for high energy physics, CVS, and his CODESH UNIX shell with virtual logbook capabilities. Javier Vazquez-Salceda (Universitat Politecnica De Catalunya) discussed applying the Provenance Aware Service Oriented Architecture (PASOA) and EU Provenance projects to the domain area of distributed medical applications, using the example of organ transplant management. Guy Kloss (German Aerospace Center (DLR)) explained how to implement “provenance-awareness” for aerospace engineering simulations. Nithya Vijayakumar (Indiana University) spoke about provenance tracking for near-real time stream filtering within the Calder data stream processing system. Miguel Branco (CERN/University of Southampton) discussed implementing the PASOA model on the Grid for the high energy physics experiment results delivered by the future ATLAS detector at CERN.

3.2 Day Two Presentations

During the *Semantics1* session, Joe Futrelle (NCSA, University of Illinois, Urbana-Champaign) described a system that harvests provenance in the form of RDF triples augmented with actor and timestamp information. Tara Talbott (Pacific Northwest National Laboratory) described how a parser for extracting XML scientific data format descriptions, as well as the data itself, assists with recording provenance. Jennifer Golbeck (University of Maryland) first presented colleagues’ work on a web portal for managing images that can be annotated with semantic descriptions; these semantic annotations can help track the provenance of images. She also discussed her project on exploring how annotations in social networks on the web can help record and infer levels of trust. Ewa Deelman (University of Southern California) presented work on augmenting existing metadata catalogs with semantic representations; she described a prototype that allows queries on temporal attributes expressed in OWL.

In the *Workflow* session, Ian Wootten (Cardiff University) spoke about using a prototype to explore how to capture actor state assertions during the enactment of a process according to PASOA ideas. Ilkay Altintas (San Diego Supercomputing Center/University of California, San Diego) discussed a provenance framework for the open source-based Kepler scientific workflow system; this framework includes a provenance “listener” utility to save information about the details of workflow executions. Bertram Ludascher (University of California, Davis) noted the importance of accounting for different models of computation, such as the directed acyclic graph (DAG), process network (PN), and synchronous data-flow (SDF) models, in constructing user-oriented provenance for pipelined scientific workflows. Michael Wilde (University of Chicago/Argonne National Laboratory) discussed provenance collection for large-scale workflow execution in the Virtual Data System (VDS), and work on providing the ability to query virtual data relationships, annotation and workflow patterns.

Finally, in the *Models of Provenance, Annotations and Processes* session, James Cheney (University of Edinburgh) presented a formal model for the process of manually curating databases with cut and paste operations. Margo Seltzer (Harvard University) described the idea of storage systems that automatically collect complete, low-level, and query-able data transformation details. Simon Miles (University of Southampton) discussed the idea of using a filter to provide the proper scope for provenance queries on potentially large directed acyclic graphs. Rajendra Bose (University of Edinburgh) presented work on a prototype system that allows individual research groups to create annotations over existing, distributed catalogues of astronomy data; these annotations record a group’s assertions of matching entries across the different catalogues.

3.3 Day Three Presentations

The *Systems* session began with Victor Tan (University of Southampton) speaking about security issues within the PASOA model; he discussed approaches to achieving access control on potentially sensitive combinations of assertions within a provenance store. Jane Hunter and Imran Khan (University of Queensland) described a system architecture that combines existing Semantic Web annotation (Annotea) and security (Shibboleth, XACML) components. Yogesh Simmhan (Indiana University) presented a quantitative performance comparison of two methods of recording provenance for scientific workflow execution: the Karma framework and the Provenance Recording Protocol (PReP) from the PASOA project. Christine Reilly (University of Wisconsin) discussed how to achieve provenance functionality for a distributed job execution system like Condor. Last, Ludek Matyska (CESNET) discussed how to enhance basic Grid job tracking with more detailed job provenance in the gLite middleware developed as part of the EU Enabling Grids in E-SciencE (EGEE) project.

During the *Semantics2* session, Carole Goble (University of Manchester) discussed the "identity crisis" caused by the assignment of multiple identifiers to the same entities within bioinformatics workflows; she proposed using sets of IDs to manage this potential problem for determining an entity's provenance. Hugo Mills (University of Southampton) spoke on behalf of David De Roure about the Combechem project which uses an RDF store to capture the provenance and semantics behind human-driven experimental chemistry, including the precise environmental conditions during an experiment. In the final talk, Paul Groth (University of Southampton) expounded on two principles for high quality documentation of provenance, which he explained were supported by PReP: only recording facts that (1) can be verified and (2) possess correct attribution.

4. PREVIOUS WORKSHOPS

We close this event report with some brief notes on previous related workshops that helped to set the scene for IPAW'06.

The Workshop on Data Derivation and Provenance organized by Peter Buneman and Ian Foster in Chicago in 2002 [3] was an important first venue for comparing and contrasting the definitions, expectations and requirements of data provenance and annotation from those involved with data management across various scientific domains. Most participants submitted position papers, and a number of short presentations were given [4]. The Workshop on Data Provenance and Annotation organized by Dave Berry, Peter Buneman, Michael Wilde, and Yannis Ioannidis in Edinburgh in 2003 [5] included surveys of provenance topics selected by the organizing committee with short presentations [6].

The 2002 provenance workshop had about 40 participants representing roughly 15 organizations or projects, and the 2003 provenance workshop had about 50 participants representing roughly 20 organizations or projects. About 10 of the same organizations/projects attended both workshops. The topic of workflow was considered in the 2002 provenance workshop, while the next year a separate e-Science Workflow Workshop [7] with 75 participants immediately followed the 2003 provenance workshop, with some participants attending both meetings.

5. ACKNOWLEDGMENTS

IPAW'06 was sponsored by Springer and Microsoft, and endorsed by the Global Grid Forum. Acknowledgment and thanks are due to the IPAW'06 program committee, and the hosts and meeting coordinators at the University of Chicago and Argonne National Laboratory.

6. WEBSITE REFERENCES

[1] http://www.ipaw.info/ipaw06

[2] http://twiki.ipaw.info/bin/view/Challenge

[3] http://www-fp.mcs.anl.gov/~foster/provenance

[4] http://people.cs.uchicago.edu/~yongzh/position_papers.html

[5] http://www.nesc.ac.uk/esi/events/304/

[6] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=304

[7] http://www.nesc.ac.uk/esi/events/303/