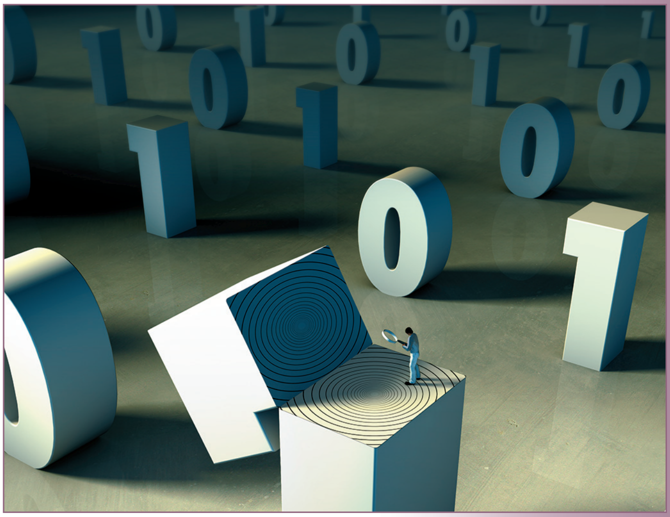

# Provenance in Web Applications

**Geetika T. Lakshmanan and Francisco Curbera**
*IBM T.J. Watson Research Center*

**Juliana Freire**
*University of Utah*

**Amit Sheth**
*Wright State University*

The Web has completely changed the way in which we share data, rapidly shifting us from a world of paper documents to a world of digital objects that include online documents, videos, photos, artwork, and databases. This shift has also made data management an increasingly complex problem as applications take advantage of loosely coupled resources brought together by distributed computing systems and abundant storage capacity. It's now easier than ever to modify documents, particularly with the help of general-purpose specifications such as XML, and extract data from documents or databases through the use of technologies such as query languages, REST interfaces, and Web service interconnectivity.

It's likewise easier to modify and update digital objects, and to do so collaboratively, via social collaboration platforms such as YouTube, Flickr, Facebook, Second Life, and Many Eyes (http://manyeyes.alphaworks.ibm.com). But as increasing volumes of data are shared and modified, it's crucial to track their provenance. Stemming from the French word *provenir* ("to come from"), provenance means the origin, or the source, of something, or the history of an object's ownership or location. A digital object's provenance (also referred to as *audit trail* and *lineage*) contains information about both the process and data used to derive the object. Provenance also provides documentation that's vital to preserving data, determining the data's quality and authorship, and reproducing as well as validating results. From the ability to reproduce digital objects to assessing data quality to enabling the enforcement of intellectual property rights and composite licensing, the provenance of digital items published and exchanged over the Web is exceedingly important.[1,2]

## Some Application Areas

Provenance has many different and compelling applications.

### Business Provenance

Geographically dispersed businesses have to manage data aggregated from different parts of the enterprise into a

1089-7801/11/$26.00 © 2011 IEEE

data warehouse. Business provenance gives the flexibility to selectively capture information required to address a specific compliance or performance goal.[3] Additionally, correlation mechanisms built on top of provenance stores can yield a representation of end-to-end operations that puts each business artifact into the right context. Execution traces of end-to-end business operations generated by provenance can capture an enterprise's operational aspects, enable modeling and predictive analytics for the business process represented by the traces,[4] and measure compliance to business rules and regulations.[3]

### Provenance for Science

Provenance is essential in science. Because reproducibility is the cornerstone of the scientific process, detailed provenance must be captured so that researchers can reproduce and validate results. Provenance is particularly important when computationally intensive science is carried out in highly distributed network environments using Internet-based collaboration tools.[5,6] More recently, with the emergence of open science in which data is widely shared and social tools are available that allow scientists to collaboratively explore data and solve problems (see myexperiment.org, www.crowdlabs.org, and www.nanohub.org), provenance is needed for tracking how experimental data is exchanged and contributed to by many different people over potentially long periods of time. Lately, the issue of publishing reproducible research has started to receive attention in the scientific community.[7]

### Provenance for Social and Sensor Networks

Provenance is vital from a social networking and Web 2.0 perspective as well.[8] Relationship discovery and community detection can be achieved on the basis of information aggregated from blogs, social bookmarking tools (such as IBM's Dogear and Delicious.com), and social networking sites. While tracking the provenance of a user's tagging behavior can give insight into his or her relationships, tracking how social networks evolve can potentially shed light into how people interact in the digital world.

In sensor networks, we can combine raw data from heterogeneous sensors with background knowledge — and a variety of analytical and reasoning support — to deliver improved situational awareness to end users.[9] In the process, original data is transformed, merged, and process in myriad ways, so the provenance can be a key tool in addressing challenges such as trustworthiness of both data and decisions.

## Challenges in Provenance Management

A provenance management solution must deal with three main problems: how to capture provenance, which information to capture and how to model it, and how to store and efficiently access the information.

### Provenance Capture

Different provenance capture mechanisms are available, depending on the tools and environment in which digital objects are created.[10] For computational tasks specified as a workflow, the workflow engine can capture the tasks' steps, parameters, and data used; execution information; and user-specified annotations.[10,11] Workflow systems such as Taverna (http://taverna.sourceforge.net), Kepler (http://kepler-project.org), and VisTrails (www.vistrails.org) support provenance capture.

Process-based provenance capture mechanisms require each service or process involved in a computational task to document itself, with any information derived from autonomous processes pieced together to provide documentation for composite tasks. Operating system- (OS-) based mechanisms require no modification to existing scripts or programs. Instead, they rely on the OS environment's ability to transparently capture data and data process dependencies at the kernel (via the file system interface) or user levels (via the system call tracer). Because there's no formal specification associated with a task, in OS-based approaches, the provenance information is obtained by extracting relationships between system calls and tasks. When we consider social and sensor data, or citizen sensing reported via mobile devices (for example, a tweet report using a smartphone), we discover a large variety of interesting forms of metadata potentially relevant to provenance, such as user profile, device-collected metadata (location and GPS information), and time and sensor-related metadata (accelerometer information and the user's cultural background).

### Provenance Models

Different models support different kinds of provenance, including retrospective provenance, which represents the steps executed as well as information about the environment used to derive a specific data product (a detailed log of a computational task's execution), and prospective provenance, which captures the steps that must be followed to derive a particular type of data product. In essence, provenance is a graph that models data and process dependencies. For example, in scientific workflow systems, the provenance graph mirrors the workflow graph.

Despite a base commonality, provenance models tend to vary according to domain and user needs. Taverna, for instance, was developed to support the creation and management of workflows in the bioinformatics domain, so it provides an infrastructure that includes support for ontologies available in this domain.[12] VisTrails was designed to support exploratory tasks, such as simulations, data exploration, and visualization in which workflows are iteratively refined, and thus uses a model that treats workflow specifications as first-class data products and captures the provenance of workflow evolution.[13] Recently, there has been an effort to create an open model that allows provenance information to be freely exchanged across systems.[14] Provenir, a provenance ontology, advocates and supports the capture of semantic provenance — that is, domain-specific semantics (such as those specified using ontologies or domain models) — in addition to data and workflow provenance.[15]

**Storing, Accessing, and Querying Provenance.** A wide variety of provenance storage and retrieval systems have been proposed, ranging from specialized Semantic Web languages and XML dialects stored as files to tuples stored in relational database tables. One of the advantages of file system storage is that users don't need additional infrastructure to store provenance information. However, a relational database does provide centralized, efficient storage that a group of users can share. Recently, researchers have attempted to explore the utility of a cloud architecture for storing data provenance;[16] those supporting semantic provenance prefer to use RDF,[15] which is now a broadly adopted Semantic Web language.

The Linked Open Data (LOD) initiative has also increased the availability of massive amounts of datasets on the Semantic Web.[1] In particular, it promotes the publication of data in machine-accessible format and linking among heterogeneous data items. Linked data is represented in RDF and can be queried using SPARQL. This large-scale initiative already consists of billions of interlinked data items, including scientific datasets that now form a large graph for easy result navigation.

A common feature across many approaches to querying provenance is that their solutions are closely tied to the storage models used. Hence, they require users to write queries in languages such as SQL, Prolog, and SPARQL. Although such general languages are useful to those already familiar with their syntax, they weren't designed specifically for provenance, which means simple queries can be awkward and complex to write. The VisTrails system uses a language specifically designed to query workflows and their provenance and includes a visual interface that lets users specify queries in the same environment they use to construct workflows.[10] Some provenance models use Semantic Web technology both to represent and query provenance information. Semantic Web languages such as RDF and OWL combined with SPARQL provide a natural way to model provenance graphs and the ability to represent complex knowledge, such as annotations and metadata. Recent work that demonstrates the scalability of Semantic Web infrastructures in handling large provenance stores is now emerging.[15]

## In This Issue

The four articles in this special issue address some of the challenges involved in constructing and using provenance today.

In the article "From Business Processes to Process Spaces," Hamid Reza Motahari-Nezhad, Boualem Benatallah, Fabio Casati, and Regis Saint-Paul propose a novel system architecture to capture business provenance by enabling the discovery and understanding of relationships between business or scientific process artifacts. They propose the "process space" as a new abstraction for process management in modern-day, dynamic, and distributed business process environments. Process-space management systems (PSMSs) will enable definition, analysis, and management of process spaces over process artifacts. Furthermore, they offer the notion of process views in a process space to represent the process execution from various perspectives (different systems, business functions, or users) and at various levels of abstractions (detailed or abstract).

Yannis Theoharis, Irini Fundulaki, Grigoris Karvounarakis, and Vassilis Christophides, in their article "On Provenance of Queries on Semantic Web Data," introduce abstract provenance models to capture the relationship between query results and source data by taking into account the query operators. This information can be recorded in the repository when the data is imported to compute appropriate annotations for different applications and users at a later time. They argue for the benefits of this approach in settings where data is materialized in repositories from various sources and there's a need to assess its quality afterward. Queries can combine data from different sources, some of which are trusted; multiple sources can be involved in alternative derivations of an item in the query result. To make trust judgments, more detailed provenance expressions are required that, in addition to provenance tokens, also record query operators involved in the derivation of a data item, thereby storing information on how input data items were combined to produce the resulting data item.

"Extending Semantic Provenance into the Web of Data," by Jun Zhao, Satya S. Sahoo, Paolo Missier, Amit Sheth, and Carole Goble, describes a single metadata architecture based on the Provenir upper-level provenance ontology that combines workflow provenance, semantics, domain-specific annotations, and LOD conventions to answer complex user queries in the context of a bioinformatics workflow. This article also describes Janus, a semantic and linked data-aware provenance infrastructure that operates on metadata produced by the Taverna workflow system. Janus demonstrates the use of semantic provenance to answer domain-specific user questions, the use of provenance query operators to implement those questions, and the use of semantics to expose provenance collected during workflow execution as part of the LOD cloud. It also demonstrates how LOD-aware provenance queries, not supported earlier in scientific workflows, can be answered.

Finally, "Papel: Provenance-Aware Policy Definition and Execution" by Christoph Ringelstein and Steffan Staab introduces a formal language that specifies the relationship between policy conditions and provenance information, based on the open provenance model. Existing policy languages aren't able to express policies that can make statements about the properties of data (and the flow of data), and this article seeks to fill this gap by enabling policy conditions to relate to provenance information.

The four articles in this special issue address only a handful of topics in provenance for Web applications. We anticipate that the growth of the Web, the increased sharing of scientific, social, and sensor data, and broad adoption of data sharing on the Web such as through the LOD initiative will fuel an explosion in the demand for provenance systems.

**Acknowledgments**

We express our gratitude to the authors, whose submissions made this special issue possible and showed how active and exciting this field is; the reviewers, who volunteered their time to read and provide constructive comments to help the authors improve the quality of their articles; the *IEEE Internet Computing* editorial staff members who helped us manage the submissions; and the magazine's editorial board for its guidance in the process.

## References

1. C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data: The Story So Far," *Int'l J. Semantic Web and Information Systems*, vol. 5, no. 3, 2009, pp. 1–22.
2. L. Moreau et al., "The Provenance of Electronic Data," *Comm. ACM*, vol. 51, no. 4, 2008, pp. 52–58.
3. F. Curbera et al., "Business Provenance: A Technology to Increase Traceability of End-to-End Operations," *Proc. OTM 2008 Confederated Int'l Conf.*, vol. 1, Springer, 2008, pp. 100–119.
4. G. Lakshmanan et al., "A Heuristic Approach for Making Predictions for Semi-Structured Case Oriented Business Processes," to be published in *Proc. Workshop Traceability and Compliance for Semi-Structured Business Processes (TC4SP'10) at Business Process Management Conf.*, LNBIP, Springer, 2010.
5. S.S. Sahoo, A. Sheth, and C. Henson, "Semantic Provenance for eScience: Managing the Deluge of Scientific Data," *IEEE Internet Computing*, vol. 12, no. 4, 2008, pp. 46–54.
6. S.B. Davidson and J. Freire, "Provenance and Scientific Workflows: Challenges and Opportunities," *Proc. SIGMOD Conf.*, ACM Press, 2008, pp. 1345–1350.
7. Special Issue on Reproducible Research, *Computing in Science & Eng.*, S. Fomel and J.F. Claerbout, eds., vol. 11, no. 1, 2009.
8. J. Golbeck, "Combining Provenance with Trust in

Social Networks for Semantic Web Content Filtering,” *Proc. Int’l Provenance and Annotation Workshop*, LNCS 4145, Springer, 2006, pp. 101–108.

9. H. Patni, C. Henson, and A. Sheth, “Linked Sensor Data,” *Proc. Int’l Symp. Collaborative Technologies and Systems*, IEEE Press, 2010, pp. 362–370.
10. J. Freire et al., “Provenance for Computational Tasks: A Survey,” *Computing in Science & Eng.*, vol. 10, no. 3, 2008, pp. 11–21.
11. L. Yogesh et al., “A Framework for Collecting Provenance in Data-Centric Scientific Workflows,” *Proc. Int’l Conf. Web Services*, IEEE Press, 2006, pp. 427–436.
12. T.M. Oinn et al., “Taverna: Lessons in Creating a Workflow Environment for the Life Sciences,” *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, 2006, pp. 1067–1100.
13. J. Freire et al., “Managing Rapidly-Evolving Scientific Workflows,” *Proc. Int’l Provenance and Annotation Workshop* (IPAW), LNCS 4145, Springer, 2006, pp. 10–18.
14. L. Moreau et al., *The Open Provenance Model Core Specification* (v1.1), Future Generation Computer Systems, 2010; doi: 10.1016/j.future.2010.07.005.
15. S.S. Sahoo et al., “Provenance Context Entity (PaCE): Scalable Provenance Tracking for Scientific RDF Data,” *Proc. 22nd Int’l Scientific and Statistical Database Management Conf.*, M. Gertz and B. Ludascher, eds., Springer, 2010, pp. 461–470.
16. K.-K. Muniswamy-Reddy, P. Macko, and M.I. Seltzer, “Provenance for the Cloud,” *Proc. Conf. File and Storage Technologies*, Usenix Assoc., 2010, pp. 197–210.

**Geetika T. Lakshmanan** is a research staff member in the IBM T.J. Watson Research Center. Her research interests span business process management, distributed systems, and data and knowledge management. Lakshmanan has a PhD in computer science from Harvard University. She is a member of the ACM. Contact her at gtlakshm@us.ibm.com.

**Francisco Curbera** is a research staff member at the IBM T.J. Watson Research Center. His research interests include the use of component-oriented software in distributed computing systems. Curbera has a PhD in computer science from Columbia University. Contact him at curbera@us.ibm.com.

**Juliana Freire** is an associate professor at the School of Computing and a faculty member at the SCI Institute, at the University of Utah. An important theme in her work is the development of data management technology to address new problems introduced by emerging applications, including the Web and scientific applications. Freire is an active member of the database and Web research communities, having co-authored over 100 technical papers and holding four US patents. She is a recipient of an NSF CAREER and an IBM Faculty award. Contact her at juliana@cs.utah.edu.

**Amit Sheth** is a fellow of IEEE, LexisNexis Ohio Eminent Scholar, and director of the Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis) at Wright State University. His focus is on developing semantic approaches and background knowledge to process, integrate, analyze, understand, and make actionable a wide variety of sources, including scientific experimental and literature, social media, and sensors. Sheth has a BE from BITS-Pilani, India and an MS and a PhD from Ohio State University. Contact him via http://knoesis.org/amit or at amit@knoesis.org.

cn *Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*