Provenance and Scientific Workflows: Challenges and Opportunities

Susan B. Davidson
University of Pennsylvania
3330 Walnut Street
Philadelphia, PA 19104-6389
susan@cis.upenn.edu

Juliana Freire
University of Utah
50 S. Central Campus Dr, rm 3190
Salt Lake City, UT 84112
juliana@cs.utah.edu

ABSTRACT

Provenance in the context of workflows, both for the data they derive and for their specification, is an essential component to allow for result reproducibility, sharing, and knowledge re-use in the scientific community. Several workshops have been held on the topic, and it has been the focus of many research projects and prototype systems. This tutorial provides an overview of research issues in provenance for scientific workflows, with a focus on recent literature and technology in this area. It is aimed at a general database research audience and at people who work with scientific data and workflows. We will (1) provide a general overview of scientific workflows, (2) describe research on provenance for scientific workflows and show in detail how provenance is supported in existing systems; (3) discuss emerging applications that are enabled by provenance; and (4) outline open problems and new directions for database-related research.

Categories and Subject Descriptors

H.2 [**Database Management**]: General

General Terms

Documentation,Experimentation

Keywords

provenance, scientific workflows

1. IMPORTANCE OF PROVENANCE FOR WORKFLOWS

Computing has been an enormous accelerator to science and has led to an information explosion in many different fields. To analyze and understand scientific data, complex computational processes must be assembled, often requiring the combination of loosely-coupled resources, specialized libraries, and grid and Web services. These processes may generate many final and intermediate data products, adding to the overflow of information scientists need to deal with. Ad-hoc approaches to data exploration (e.g., Perl scripts) have been widely used in the scientific community, but have serious limitations. In particular, scientists and engineers need to expend substantial effort managing data (e.g., scripts that encode computational tasks, raw data, data products, and notes) and recording provenance information so that basic questions can be answered, such as: Who created this data product and when? When was it modified and by whom? What was the process used to create the data product? Were two data products derived from the same raw data? Not only is the process time-consuming, but also error-prone.

Workflow systems have therefore grown in popularity within the scientific community [25, 41, 31, 42, 43, 45, 16, 17, 27, 38]. Not only do they support the automation of repetitive tasks, but they can also capture complex analysis processes at various levels of detail and systematically capture provenance information for the derived data products [15]. The provenance (also referred to as the audit trail, lineage, and pedigree) of a data product contains information about the process and data used to derive the data product. It provides important documentation that is key to preserving the data, to determining the data's quality and authorship, and to reproduce as well as validate the results. These are all important requirements of the scientific process.

Provenance in scientific workflows is thus of paramount and increasing importance, as evidenced by recent specialized workshops [6, 15, 21, 32, 33] and surveys [18, 14, 7, 36]. While provenance in workflows bears some similarity to that of provenance in databases (which was the topic of a tutorial in SIGMOD'2007 [10] and a recent survey [40]), there are important differences and new challenges for the database community to consider.

Our objective in this tutorial is to give an overview of the problem of managing provenance data for scientific workflows, illustrate some of the techniques that have been developed to address different aspects of the problem, and outline interesting directions for future work in the area. In particular, we will present techniques for reducing provenance overload as well as making provenance information more "fine-grained." We will examine uses of provenance that go beyond the ability to reproduce and share results, and will demonstrate how workflow evolution provenance can be leveraged to explain difference in data products, streamline exploratory computational tasks, and enable knowledge re-use. We will also discuss a new applications that are enabled by provenance, such as social data analysis [19], which have the potential to change the way people explore data and do science.

SIGMOD'08, June 9–12, 2008, Vancouver, BC, Canada.
ACM 978-1-60558-102-6/08/06.

2. TUTORIAL OUTLINE

2.1 Overview of Scientific Workflows

We motivate the need for scientific workflows using real applications as examples, in particular within genomics, medical imaging, environmental observatories and forecasting systems. We also introduce basic concepts for scientific workflows that are related to provenance.

Workflow and workflow-based systems have emerged as an alternative to ad-hoc approaches for constructing computational scientific experiments [25, 39, 41, 45, 31]. Workflow systems help scientists conceptualize and manage the analysis process, support scientists by allowing the creation and reuse of analysis tasks, aid in the discovery process by managing the data used and generated at each step, and (more recently) systematically record provenance information for later use. Workflows are rapidly replacing primitive shell scripts as evidenced by the release of Apple's Mac OS X Automator, Microsoft's Workflow Foundation, and Yahoo! Pipes.

Scientific workflows systems often adopt simple computational models, in particular a dataflow model, where the execution order of workflow modules is determined by the flow of data through the workflow. This is in contrast to business workflows which provide expressive languages (such as the Business Process Execution Language, BPEL [9]) to specify complex control flows [1]. In addition, unlike business workflows, scientific workflows are often used to perform data intensive tasks.

Workflow systems have a number of advantages for constructing and managing computational tasks compared to programs and scripts. They provide a simple programming model whereby a sequence of tasks is composed by connecting the outputs of one task to the inputs of another. Furthermore, workflow systems often provide intuitive visual programming interfaces, which make them more *suitable for users who do not have substantial programming expertise*. Workflows also have an *explicit structure*. They can be viewed as graphs, where nodes represent processes (or modules) and edges capture the flow of data between the processes. The benefits of structure are well-known when it comes to exploring data. A program (or script) is to a workflow what an unstructured document is to a (structured) database.

2.2 Managing Provenance

We first describe different kinds of provenance that can be captured for scientific workflows. Then, we discuss the three key components of a provenance management solution: the capture mechanism; the data model for representing provenance information; and the infrastructure for storing, accessing, and querying provenance. Last, but not least, we present different approaches used for each of these components and classify the different workflow systems based on a set of dimensions along which their treatments of the issues differ.

Information represented in provenance. In the context of scientific workflows, provenance is a record of the derivation of a set of results. There are two distinct forms of provenance [11]: prospective and retrospective. *Prospective provenance* captures the specification of a computational task (i.e., a workflow)—it corresponds to the *steps that need to be followed* (or a recipe) to generate a data product or class of data products. *Retrospective provenance* captures the *steps that were executed* as well as information about the execution environment used to derive a specific data product—a detailed log of the execution of a computational task. Figure 1 illustrates these two kinds of provenance.

An important piece of information present in workflow provenance is information about *causality*: the dependency relationships among data products and the processes that generate them. Causality can be inferred from both prospective and retrospective provenance and it captures the sequence of steps which, together with input data and parameters, caused the creation of a data product. Causality consists of different types of dependencies. Data-process dependencies (e.g., the fact that `head-hist.png` was derived by the sub-workflow on the left in Figure 1) are useful for documenting data generation process, and they can also be used to reproduce or validate the process. For example, it would allow new histograms to be derived for different input data sets. Data dependencies are also useful. For example, in the event that the CT scanner used to generate the input file `head.120.vtk` is found to be defective, results that depend on the scan can be invalidated by examining data dependencies.

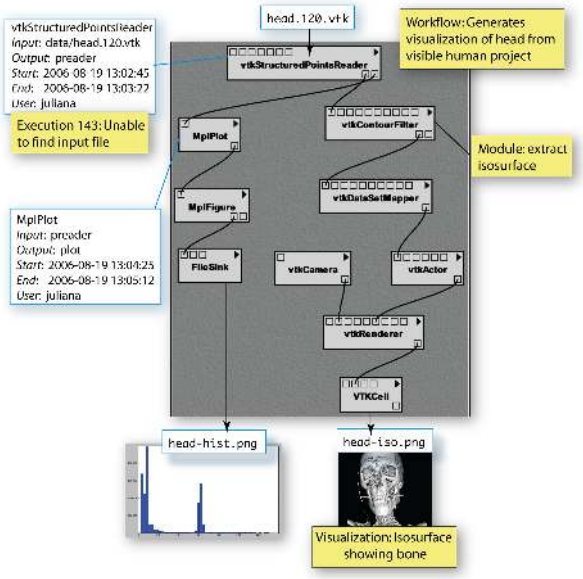

Figure 1: Prospective versus retrospective provenance. The workflow generates two data products: a histogram of the scalar values of a structured grid data set; and a visualization of an isosurface of the data set. The workflow definition provides prospective provenance, a recipe to derive these two kinds of data products. On the left, we show some of the retrospective provenance that was collected during a run of this workflow. This figure also illustrates user-defined provenance in the form of annotations, shown in yellow boxes.

Another key component of provenance is *user-defined information*. This includes documentation that cannot be automatically captured but records important decisions and notes. This data is often captured in the form of annotations. As Figure 1 illustrates, annotations can be added at different levels of granularity and associated with different components of both prospective and retrospective provenance (e.g., for modules, data products, execution log records).

Capturing, modeling, storing and querying provenance. One of the major advantages to using workflow systems is that they can be easily instrumented to automatically capture provenance — this information can be accessed directly through system APIs. While early workflow systems (e.g., Taverna [41] and Kepler [25]) have been *extended* to capture provenance, newer systems, such as VisTrails [45] have been *designed* to support provenance.

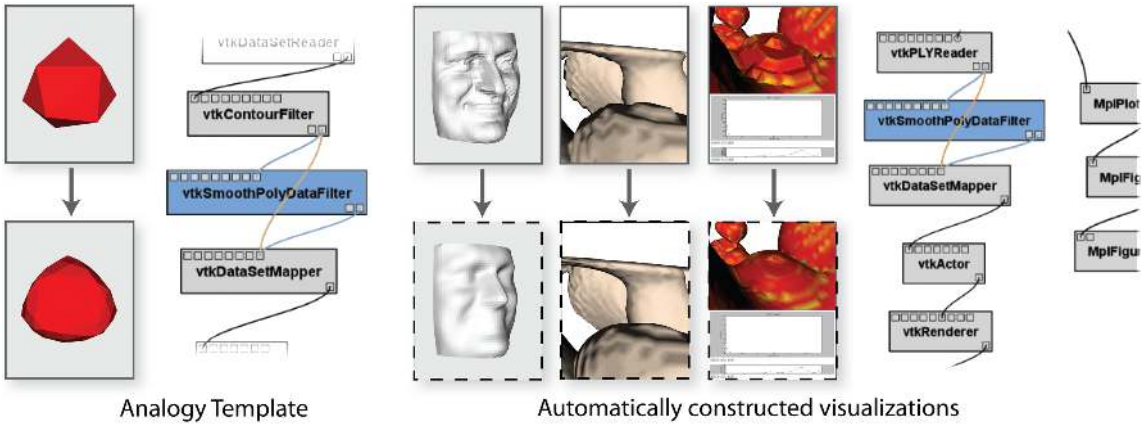

Figure 2: Usable interface to refine workflows by analogy. The user chooses a pair of data products to serve as an analogy template. In this case, the pair represents a change to a workflow that downloads a file from the Web and creates a simple visualization, into a new workflow where the resulting visualization is smoothed. Then, the user chooses a set of other workflows to apply the same change automatically. The workflow on the left reflects the original changes, and the one on the right reflects the changes when translated to the workflow used to derive the last visualization on the right. The workflow components to be removed are shown in orange, and the ones to be added, in blue. Note that the surrounding modules do not match exactly: the system identifies out the most likely match. Image from [34].

Several provenance models have been proposed in the literature [37, 28, 12, 2, 46, 26, 11, 20, 22]. All of these models support some form of retrospective provenance and many also provide the means to capture prospective provenance as well as annotations. Although these models differ in many ways, including the use of different structures and storage strategies, they all share an essential type of information: process and data dependencies. In fact, a recent exercise to explore interoperability issues among provenance models has shown that it is possible to integrate different provenance models [33].

While several approaches have been proposed to capture and model provenance, only recently has the problem of storing, accessing, and querying provenance started to receive attention. Besides allowing users to explore and better understand results, the ability to query the provenance of workflows enables knowledge re-use. For example, users can identify workflows that are suitable and can be re-used for a given task; compare and understand differences between workflows; and refine workflows by analogy (see Figure 2). Provenance information can also be associated with data products (e.g., images, graphs), allowing structured queries to be posed over these unstructured data.

A common feature across many of the approaches to querying provenance is that their solutions are closely tied to the storage models used. A wide variety of data models and storage systems have been used ranging from specialized Semantic Web languages (e.g., RDF and OWL) and XML dialects that are stored as files and to tuples stored in relational database tables. Hence, they require users to write queries in languages like SQL [3], Prolog [8] and SPARQL [46, 26, 22]. While such standard languages can be useful if users are already familiar with their syntax, none of them have been designed for provenance. For that reason, simple queries can be awkward and complex. We will discuss these approaches and contrast them to recent work on intuitive visual interfaces to query workflows [4, 34].

Provenance systems. We survey approaches to provenance adopted by scientific workflow systems. We present and compare different proposals for capturing, modeling, storing and querying provenance (e.g., [34, 13, 8, 18, 20, 29, 36, 32, 33]).

2.3 Using Provenance for Reproducibility and Beyond

We will also discuss a number of emerging applications for workflow provenance and discuss the challenges they pose to database research. Some of these applications are described below.

Provenance and scientific publications. A key benefit for maintaining provenance of computational results is reproducibility: a detailed record of the steps followed to produce a result allows others to reproduce and validate these results. Recently, the issue of publishing reproducible research has started to receive attention in the scientific community. In 2008, SIGMOD has introduced the "experimental repeatability requirement" to "help published papers achieve an impact and stand as reliable reference-able works for future research".[1] A number of journals are also encouraging authors to make their publications reproducible, including, for example the IEEE Transactions on Signal Processing[2] and the Computing in Science and Engineering (CiSE) magazine[3]. Provenance management infrastructure and tools will have the potential to transform scientific publications as we know them today. However, for these to be widely adopted, they need to be usable and within reach for scientists that do not have computer science training.

Provenance and data exploration. Provenance can also be used to simplify exploratory processes. In particular, we present mechanisms that allow the flexible re-use of workflows; scalable exploration of large parameter spaces; and comparison of data products as well as their corresponding workflows [20, 35]. In addition, we show that useful knowledge is embedded in provenance which can be re-used to simplify the construction of workflows [34].

Social analysis of scientific data. Social Web sites and Web-based communities (e.g., Flickr, Facebook, Yahoo! Pipes), which facilitate collaboration and sharing between users, are becoming increasingly popular. An important benefit of these sites is that they en-

[1] http://www.sigmod08.org/sigmod_research.shtml
[2] http://ewh.ieee.org/soc/sps/tsp
[3] http://www.computer.org/portal/site/cise/index.jsp

able users to leverage the wisdom of the crowds. In the (very) recent past, a new class of Web site has emerged that enables users to upload and collectively analyze many types of data (e.g., Many Eyes [44]). These are part of a broad phenomenon that has been called "social data analysis". This trend is expanding to the scientific domain where a number of collaboratories are under development. As the cost of hardware decreases over time, the cost of people goes up as analyses get more involved, larger groups need to collaborate, and the volume of data manipulated increases. Science collaboratories aim to bridge this gap by allowing scientists to share, re-use and refine their workflows. We discuss the challenges and key components that are needed to enable the development of effective social data analysis (SDA) sites for the scientific domain [19]. For example, usable interfaces that allow users to query and re-use the information in these collaboratories are key to their success. We will present recent work that has addressed usability issues in the context of workflow systems and provenance (see Figure 2).

Provenance in education. Teaching is one of the killer applications of provenance-enabled workflow systems, in particular, for courses which have a strong data exploration component such as data mining and visualization. Provenance can help instructors to be more effective and improve the students' learning experience. By using a provenance-enabled tool in class, an instructor can keep detailed record of all the steps she tried while while responding to students questions; and after the class, all these results and their provenance can be made available to students. For assignments, students can turn the detailed provenance of their work, showing all the steps they followed to solve a problem.

2.4 Open Problems

We discuss a number of open problems and outline possible directions for future research, including:

- *Information management infrastructure.* With the growing volume of raw data, workflows and provenance information, there is a need for efficient and effective techniques to manage these data. Besides the need to handle large volumes of heterogeneous and distributed data, an important challenge that needs to be addressed is *usability*: Information management systems are notoriously hard to use [23, 24]. As the need for these systems grows in a wide range of applications, notably in the scientific domain, usability is of paramount importance. The growth in the volume of provenance data also calls for techniques that deal with information overload [5].

- *Provenance analytics and visualization.* The problem of mining and extracting knowledge from provenance data has been largely unexplored. By analyzing and creating insightful visualizations of provenance data, scientists can debug their tasks and obtain a better understanding of their results. Mining this data may also lead to the discovery of patterns that can potentially simplify the notoriously hard, time-consuming process of designing and refining scientific workflows.

- *Interoperability.* Complex data products may result from long processing chains that require multiples tools (e.g., scientific workflows and visualization tools). In order to provide detailed provenance for such data products, it becomes necessary to integrate provenance derived from different systems and represented using different models. This was the goal of the Second Provenance Challenge [33], which brought together several research groups with the goal of integrating provenance across their independently developed workflow systems. Although the preliminary results are promising and indicate that such an integration is possible, there needs to be more principled approaches to this problem. One direction currently being investigated is the creation of a standard for representing provenance [30].

- *Connecting database and workflow provenance.* In many scientific applications, database manipulations co-exist with the execution of workflow modules: Data is selected from a database, potentially joined with data from other databases, reformatted, and used in an analysis. The results of the analysis may then be put into a database and potentially used in other analyses. To understand the provenance of a result, it is therefore important to be able to connect provenance information across databases and workflows. Combining these disparate forms of provenance information will require a framework in which database operators and workflow modules can be treated uniformly, and a model in which the interaction between the structure of data and the structure of workflows can be captured.

3. ABOUT THE PRESENTERS

Susan B. Davidson received a B.A. degree in Mathematics from Cornell University in 1978, and a Ph.D. degree in Electrical Engineering and Computer Science from Princeton University in 1982. Dr. Davidson joined the University of Pennsylvania in 1982, and is now the Weiss Professor and Department Chair of Computer and Information Science. She is an ACM Fellow, a Fulbright scholar, and a founding co-Director of the Center for Bioinformatics at UPenn (PCBI). Dr. Davidson's research interests include database systems, database modeling, distributed systems, and bioinformatics. Within bioinformatics she is best known for her work in data integration, XML query and update technologies, and more recently provenance in workflow systems.

Juliana Freire joined the faculty of the School of Computing at the University of Utah in July 2005. Before, she was member of technical staff at the Database Systems Research Department at Bell Laboratories (Lucent Technologies) and an Assistant Professor at OGI/OHSU. She received Ph.D. and M.S. degrees in Computer Science from the State University of New York at Stony Brook, and a B.S. degree in Computer Science from Universidade Federal do Ceará, Brazil. Dr. Freire's research has focused on extending traditional database technology and developing techniques to address new data management problems introduced by the Web and scientific applications. She is a co-creator of VisTrails (www.vistrails.org), an open-source scientific workflow and provenance management system.

4. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation and the Department of Energy.

5. REFERENCES

[1] W. Aalst and K. Hee. *Workflow Management: Models, Methods, and Systems*. MIT Press, 2002.

[2] I. Altintas, O. Barney, and E. Jaeger-Frank. Provenance collection support in the kepler scientific workflow system. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pages 118–132, 2006.

[3] R. S. Barga and L. A. Digiampietri. Automatic capture and efficient storage of escience experiment provenance.

Concurrency and Computation: Practice and Experience, 20(5):419–429, 2008.

[4] C. Beeri, A. Eyal, S. Kamenkovich, and T. Milo. Querying business processes. In *VLDB*, pages 343–354, 2006.

[5] O. Biton, S. Cohen-Boulakia, S. Davidson, and C. Hara. Querying and managing provenance through user views in scientific workflows. In *Proceedings of ICDE*, 2008. To appear.

[6] R. Bose, I. Foster, and L. Moreau. Report on the International Provenance and Annotation Workshop. *SIGMOD Rec.*, 35(3):51–53, 2006.

[7] R. Bose and J. Frew. Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys*, 37(1):1–28, 2005.

[8] S. Bowers, T. McPhillips, and B. Ludaescher. A provenance model for collection-oriented scientific workflows. *Concurrency and Computation: Practice and Experience*, 20(5):519–529, 2008.

[9] Business Process Execution Language for Web Services. http://www.ibm.com/developerworks/library/specification/ws-bpel/.

[10] P. Buneman and W.Tan. Provenance in databases. In *Proceedings of ACM SIGMOD*, pages 1171–1173, 2007.

[11] B. Clifford, I. Foster, M. Hategan, T. Stef-Praun, M. Wilde, and Y. Zhao. Tracking provenance in a virtual data grid. *Concurrency and Computation: Practice and Experience*, 20(5):565–575, 2008.

[12] S. Cohen, S. C. Boulakia, and S. B. Davidson. Towards a model of provenance and user views in scientific workflows. In *DILS*, pages 264–279, 2006.

[13] S. Cohen-Boulakia, O. Biton, S. Cohen, and S. Davidson. Addressing the provenance challenge using zoom. *Concurrency and Computation: Practice and Experience*, 20(5):497–506, 2008.

[14] S. B. Davidson, S. C. Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44–50, 2007.

[15] E. Deelman and Y. Gil. NSF Workshop on Challenges of Scientific Workflows. Technical report, NSF, 2006. http://vtcpc.isi.edu/wiki/index.php/Main_Page.

[16] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming Journal*, 13(3):219–237, 2005.

[17] I. Foster, J. Voeckler, M. Wilde, and Y. Zhao. Chimera: A virtual data system for representing, querying and automating data derivation. In *Proceedings of SSDBM*, pages 37–46, 2002.

[18] J. Freire, D. Koop, E. Santos, and C. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3), May/June 2008. To appear.

[19] J. Freire and C. Silva. Towards enabling social analysis of scientific data. In *CHI Social Data Analysis Workshop*, 2008. To appear.

[20] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo. Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10–18, 2006. Invited paper.

[21] D. Gannon et al. A Workshop on Scientific and Scholarly Workflow Cyberinfrastructure: Improving Interoperability, Sustainability and Platform Convergence in Scientific And Scholarly Workflow. Technical report, NSF and Mellon Foundation, 2007. https://spaces.internet2.edu/display/SciSchWorkflow.

[22] J. Golbeck and J. Hendler. A semantic web approach to tracking provenance in scientific workflows. *Concurrency and Computation: Practice and Experience*, 20(5):431–439, 2008.

[23] L. Haas. Information for people. http://www.almaden.ibm.com/cs/people/laura/ Information For People keynote.pdf, 2007. Keynote talk at ICDE.

[24] H. V. Jagadish. Making database systems usable. http://www.eecs.umich.edu/db/usable/ usability-sigmod.ppt, 2007. Keynote talk at SIGMOD.

[25] The Kepler Project. http://kepler-project.org.

[26] J. Kim, E. Deelman, Y. Gil, G. Mehta, and V. Ratnakar. Provenance trails in the wings/pegasus system. *Concurrency and Computation: Practice and Experience*, 20(5):587–597, 2008.

[27] Microsoft Workflow Foundation. http://msdn2.microsoft.com/en-us/netframework/aa663322.aspx.

[28] S. Miles, P. Groth, S. Munroe, S. Jiang, T. Assandri, and L. Moreau. Extracting Causal Graphs from an Open Provenance Data Model. *Concurrency and Computation: Practice and Experience*, 20(5):577–586, 2008.

[29] L. Moreau and I. Foster, editors. *Provenance and Annotation of Data - International Provenance and Annotation Workshop*, volume 4145. Springer-Verlag, 2006.

[30] L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, and P. Paulson. The open provenance model, December 2007. http://eprints.ecs.soton.ac.uk/14979.

[31] S. G. Parker and C. R. Johnson. SCIRun: a scientific programming environment for computational steering. In *Supercomputing*, page 52, 1995.

[32] First provenance challenge. http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge, 2006. S. Miles, and L. Moreau (organizers).

[33] Second provenance challenge. http://twiki.ipaw.info/bin/view/Challenge/SecondProvenanceChallenge, 2007. J. Freire, S. Miles, and L. Moreau (organizers).

[34] C. Scheidegger, D. Koop, H. Vo, J. Freire, and C. Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1560–1567, 2007. Papers from the IEEE Information Visualization Conference 2007.

[35] C. Silva, J. Freire, and S. P. Callahan. Provenance for visualizations: Reproducibility and beyond. *Computing in Science & Engineering*, 9(5):82–89, 2007.

[36] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005.

[37] Y. L. Simmhan, B. Plale, and D. Gannon. Karma2: Provenance management for data driven workflows. *International Journal of Web Services Research, Idea Group Publishing*, 5:1, 2008. To Appear.

[38] Y. L. Simmhan, B. Plale, D. Gannon, and S. Marru. Performance evaluation of the karma provenance framework for scientific workflows. In L. Moreau and I. T. Foster, editors, *International Provenance and Annotation Workshop (IPAW), Chicago, IL*, volume 4145 of *Lecture Notes in Computer Science*, pages 222–236. Springer, 2006.

[39] The Swift System. www.ci.uchicago.edu/swift.

[40] W. C. Tan. Provenance in databases: Past, current, and future. *IEEE Data Eng. Bull.*, 30(4):3–12, 2007.

[41] The Taverna Project. http://taverna.sourceforge.net.

[42] The Triana Project. http://www.trianacode.org.

[43] VDS - The GriPhyN Virtual Data System. http://www.ci.uchicago.edu/wiki/bin/view/VDS/VDSWeb/WebMain.

[44] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.

[45] The VisTrails Project. http://www.vistrails.org.

[46] J. Zhao, C. Goble, R. Stevens, and D. Turi. Mining taverna's semantic web of provenance. *Concurrency and Computation: Practice and Experience*, 20(5):463–472, 2008.