2008

A Semiotics Framework for Analyzing Data Provenance Research

Jun Liu

Sudha Ram

A Semiotics Framework for Analyzing Data Provenance Research

Sudha Ram and Jun Liu
Department of MIS
Eller College of Management
430J McClelland Hall
University of Arizona, Tucson, AZ 85718
{sram, jliu}@email.arizona.edu

Received 27 June 2008: Accepted 29 August 2008

Data provenance is the background knowledge that enables a piece of data to be interpreted and used correctly within context. The importance of tracking provenance is widely recognized, as witnessed by significant research in various areas including e-science, homeland security, and data warehousing and business intelligence. In order to further advance the research on data provenance, however, one must first understand the research that has been conducted to date and identify specific topics that merit further investigation. In this work, we develop a framework based on semiotics theory to assist in analyzing and comparing existing provenance research at the conceptual level. We provide a detailed review of data provenance research and compare and contrast the research based on a semiotics framework. We conclude with an identification of challenges that will drive future research in this field.

Categories and Subject Descriptors: Database Management [**Heterogeneous Databases**]

General Terms: Algorithm and Experiment

Additional Key Words and Phrases: Data Provenance, Data Lineage, Semiotics, Provenance-based Information Systems

1. INTRODUCTION

Provenance, also called lineage or pedigree, is a well-established concept in the art world where it can help to determine the authenticity of a work, to establish the historical importance of a work, and to determine the legitimacy of current ownership [Tan 2004]. It is of equal importance in present data-rich environments ranging from computational biology, high energy physics, to data warehousing where people request data recorded in information sources owned by other people.

To ensure that data provided by other sources can be trusted and used appropriately, it is imperative that the provenance of the data be recorded and made available to its users. Data provenance refers to the lineage or history of data including information such as its origin and key events that occur over the course of its lifecycle. It is the background knowledge that enables a piece of data to be interpreted and used correctly within context. Tracking provenance has several uses, including in-depth data analysis, data replication and reuse, data security management, and others as outlined in [Simmhan et al. 2005].

The need to capture data provenance has been addressed by both researchers and practitioners. During the past two decades, significant research has yielded designs and prototype software systems for preserving and retrieving data provenance in various areas including e-science [Frew and Bose 2002; Greenwood et al. 2003; Pancerella 2003], homeland security [Ding et al. 2005; Ceruti et al. 2006], and data warehousing and business intelligence [Buneman et al. 2001; Cui and Widom 2003]. In order to further advance the research on data provenance, however, one must first understand the research that has been conducted to date and identify specific topics that merit further investigation. The objectives of this paper are to develop a framework that can be used to analyze existing research on data provenance and to identify important future research directions. We draw upon research such as [Stamper 1991; Barron et al. 1999] to develop a framework based on the semiotics theory to assist us in analyzing and comparing existing provenance research at the conceptual level. We provide a detailed review of data provenance research literature, compare and contrast the existing research based on the semiotics framework, and identify challenges to drive future research in this field.

This paper is organized as follows. Section 2 presents the semiotics framework for data provenance analysis and discusses various elements of the framework. Section 3 analyzes existing research on data provenance based on this framework and identifies unresolved research issues. In Section 4, we identify open research questions and challenges in the field. Section 5 summarizes and concludes the paper.

2. A SEMIOTICS FRAMEWORK FOR ANALYZING DATA PROVENANCE RESEARCH

Several frameworks are described in literature to analyze and compare data provenance research. Bose et al. review data lineage research according to the modes of data processing including script and program-based, workflow management system (WFMS)-based, query-based, and service-based data processing, with the goal of arriving at a meta-model that describes lineage retrieval as depending on the workflow and metadata modeled designed into systems [Bose and Frew 2005]. Simmhan et al. present a taxonomy that helps compare and contrast existing provenance techniques along five different dimensions including provenance application, subject of provenance, provenance representation, provenance storage, and provenance dissemination [Simmhan et al. 2005].

Although the above frameworks for analyzing and comparing provenance research have their own merits, we strive to improve our understanding of the distinction among various provenance techniques based upon a theoretical foundation. Our

analysis is different from the extant frameworks such as [Bose and Frew 2005; Simmhan et al. 2005], that are derived based on observation and intuitive understanding of existing provenance research. We propose to identify and analyze the essential properties and features of various provenance techniques based on a systematic theory, namely, the theory of semiotics, with the purpose of providing rigor and organization to the analysis and identifying issues that have not been sufficiently addressed by the existing research. To do so, we propose our semiotics framework consisting of ten elements for analyzing data provenance research.

2.1 Semiotics

Semiotics, "the science of the life of signs within society" as Saussure [Saussure 1966] defined it, describes the form-related, meaning-related, and use-related aspects of "signs". Semiotics has multiple branches. We focus on "computer semiotics", defined as "a branch of semiotics that studies the special nature of computer-based signs and how they function in the use situation" [Andersen 1991], where inputs and outputs of computer-based information systems are "signs" and information systems are seen as "sign-vehicles" [Andersen 1991] that supports the representation, storage and processing of the signs, as well as their uses and interpretations by human and automated agents. Semiotics is traditionally divided into three areas or levels: *Semantics* or the meaning of signs deals with the relationship of signs to what they stand for. *Syntactics* analyzes the structural relations between signs, and *pragmatics* the ways in which signs are interpreted and used. While providing his view on "computer semiotics", Stamper proposes a semiotics framework for information systems research, in which Stamper adds another level, namely *social level*, which deals with the social consequences of signs, in addition to syntactics, semantics and pragmatics [Stamper 1991]. Barron et al. extend Stamper's framework and apply the framework to the analysis and classification of information systems [Barron et al. 1999]. Information systems that harvest, maintain, and represent data provenance are also information systems.

Drawing upon [Barron et al. 1999], we provide a generic, high-level view of the major components of a provenance based information system as shown in Figure 1. The "sensor/observer" harvests data provenance by observing various events such as creation and modifications that happened to data. The harvested provenance will then be input into a "provenance knowledge base" as "signs". The intended users will then make an impact in an application domain by taking actions based on the output of the "provenance knowledge base". The semiotics framework consisting of four levels including *syntactics*, *semantics*, *pragmatics*, and *social level* proposed in [Stamper 1991] enables us to view such a provenance based information system from the perspective of semiotics. The *social level* of semiotics is concerned with the *perlocutionary* effects of signs [Barron et al. 1999], i.e., the impact achieved by the user performing actions using signs (i.e., provenance in this case). *Pragmatics* should consider the origin and effects of signs within the behavior in which they occur [Morris 1946]. In our case, pragmatics deals with both the acquisition or observation of data provenance and the *illocutionary effects* [Barron et al. 1999] of the output on

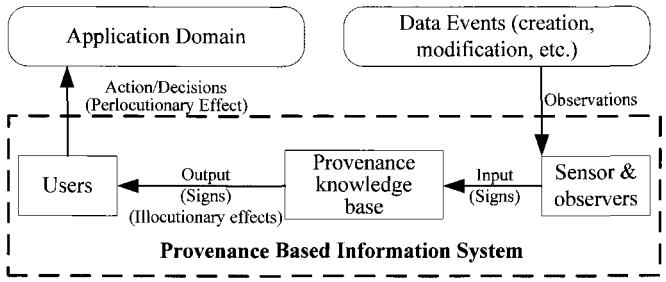


Figure 1. The major components of a provenance-based information system (adapted from [Barron et al. 1999]).

the users performing an intended task (see Figure 1). The input and output of a provenance based information systems are *signs*. *Syntactics* analyzes the structure and representation of the signs. *Semantics* deals with the relationship between the signs and the objects to which they are applicable [Morris 1946], i.e., the relationship between provenance and data events (e.g., creation and modifications) the provenance is intended to capture in our case. A semantics framework including these four levels thus helps us comprehensively analyze various aspects of a provenance based information system ranging from provenance acquisition, to the semantics and syntactics of provenance, to the use of provenance on the social level. Current research on data provenance usually focuses on certain aspects of provenance-based information systems. Analyzing the current research using a semiotics framework enables us to provide an overall picture of the current research on data provenance and identify future research directions.

Following the approach first proposed by Stamper [Stamper 1991], we present a semiotics framework consisting of four semiotics levels including syntactics, semantics, pragmatics, and social level for analyzing existing provenance research and classifying various provenance techniques. In their extension of Stamper's framework, Barron et al. identify 10 semiotics features based on an analysis of the four semantics levels as being the most common and appropriate in analyzing information systems' properties [Barron et al. 1999]. Drawing upon this work, we establish 10 semiotics elements based on these four semiotics levels: 1) application domain, 2) data processing architecture, 3) action complexity, 4) social consequences, 5) acquisition complexity, 6) acquisition scope, 7) trust and security, 8) usability, 9) semantics of provenance and 10) representation of provenance. The summary of our semiotics framework adapted from [Barron et al. 1999] is shown in Figure 2. The "input usability" and "output usability" in [Barron et al. 1999] are combined into "usability" since both the input and output of a provenance-based system are provenance, and data provenance is historic information and therefore often remains unchanged after being input into the system. Also, we add an element called "data processing architecture". According to the theory of semiotics, pragmatics should consider the "origin" of signs [Morris 1946]. As shown in Figure 1, a provenance-based information system harvests data provenance by observing the creation or modifications of data in a data processing architecture. Hence, analyzing the data

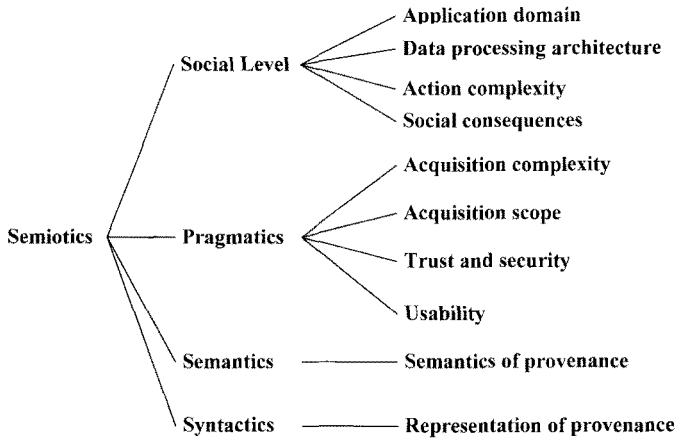


Figure 2. Summary of the Semiotics Framework.

processing architecture, where data is created and modified, helps us understand the "origin" of data provenance.

2.2 Taxonomy of the Framework

2.2.1 Social Level

Provenance information captured by various techniques can be viewed as signs stored in the systems. No sign can be fully understood without regard for its potential and actual social consequences. At the social level, we are concerned with the actual and *perlocutionary* effects of the signs [Barron et al. 1999]. We are interested in finding out how data provenance can be used in various application domains and how the users perform actions and decisions, i.e., perlocutionary acts, using provenance. Three elements of our framework are established on the social level.

2.2.1.1 Application domain

The application domain is concerned with the scope, boundary and actual perlocutionary effects of "signs" [Barron et al. 1999]. Data provenance that is captured by various provenance based systems is designed to solve problems, support activities and tasks, and make changes within one or more domains. As a result, we investigate the application of provenance in various domains.

2.2.1.2 Data processing architecture

This element is unique for data provenance research. Data provenance is meta-level information that describes the source and transformations of data products. Data provenance techniques are therefore often tailored to various data processing architectures that provide different means for generating, consuming data products, and bringing about transformations of the data. We consider the architecture for data processing an important means for categorizing data provenance techniques.

2.2.1.3 Social consequence
The social consequence element is adopted from [Barron et al. 1999] and refers to the effect of captured data provenance on the actions and decisions performed by the system users on the application. Researchers have recognized many potential social consequences from tracking data provenance. It helps users determine the quality and trustworthiness of data [Lynch 2001; Goble 2002; Prat and Madnick 2007], and it also enables users to share, discover, and reuse the data, thus streamlining collaborative activities and reducing the possibility of repeating dead ends [Ram and Liu 2007].

2.2.1.4 Action complexity
Action complexity refers to the nature of actions or decisions that users can perform by using the output of an information system [Barron et al. 1999]. An action can be structured, semi-structured, or unstructured. A structured or semi-structured action is an action for which a process is readily available and can be automatically or semi-automatically performed, while an unstructured action requires human expertise and judgment [Barron et al. 1999].

2.2.2 Pragmatics
Pragmatics is concerned with relationships between signs and behaviors of the users of an information system in a particular context [Barron et al. 1999]. In the context of data provenance research, pragmatics should consider the origin, i.e., the source of provenance information, how provenance can be captured, and uses of provenance information. At this level, we are concerned with the *illocutionary* potential of provenance [Barron et al. 1999], i.e., the usability of provenance, whereas we are interested in the actual and perlocutionary effects of provenance at the social level. There are four elements defined at the pragmatic level.

2.2.2.1 Acquisition complexity
Acquisition concerns the nature of the processing required to acquire provenance knowledge. We consider two aspects associated with provenance acquisition: *complexity* and *scope*. *Acquisition complexity* primarily refers to the level of automation of various provenance acquisition techniques. We also consider costs and overhead incurred during the process of capturing data provenance.

2.2.2.2 Acquisition scope
Acquisition scope refers to the range of sources for the acquisition of provenance information. While most of the existing research focuses on capturing data provenance in a single controlled environment (e.g., a database or a scientific workflow), data may move among databases, thus requiring provenance knowledge to be acquired from multiple sources.

2.2.2.3 Trust and security
This dimension is similar to the "justification" feature proposed in [Barron et al. 1999]. The focus of this element is *whether* the acquired provenance knowledge

should be trusted to be uncompromised and free from error, in contrast to acquisition's focus on *how* data provenance has been acquired. In some cases, knowing the latter can also suffice for trusting the captured provenance. As an example, automated provenance recording helps enhance the trustworthiness of the acquired provenance by eliminating the human errors.

2.2.2.4 Usability

The *usability* of provenance is concerned with whether the provenance that has been collected by the system is useful with regard to satisfying various provenance-related inquiries. The harvested provenance has a high usability when it meets the users' information requirements and generates more informative and pragmatic impacts on the users for the intended usage.

2.2.3 Semantics

Semantics deals with the meaning of signs. Then, the meaning can be considered as the mapping of a sign to reality. The semantic properties of signs deal with meaning in the special sense of how signs relate to reality, how they represent, designate and signify a real world phenomena [Barron et al. 1999]. We investigate the semantics or meaning of provenance as understood by researchers studying the provenance issue in various domains.

2.2.3.1 Semantics of provenance

Semantics or meaning can be considered as functions from signs to reality and may be different for different people. Provenance can be collected about various resources (e.g. inputs, output, and processes) present in data processing. It may be more applicable to capture provenance about certain types of data than on others, and provenance techniques may focus on certain aspects of provenance (e.g. the source of data) while ignoring others depending on the nature and importance of data. As a result, investigating the content of provenance in terms what aspects of data provenance are captured in existing research is critical for understanding the status of current provenance research.

2.2.4 Syntactics

The syntactic level concerns the form or representation of signs rather than their meaning and potential uses. According to [Barron et al. 1999], it is only concerned with the formal representation and relationships of signs, and the operations and processes to which they may be subjected.

2.2.4.1 Representation of provenance

Data provenance can be represented in different ways, often depending on the underlying data provenance systems. Representation is expressed relative to a particular language for representing and manipulating provenance. Examples of the languages are database languages such as SQL and data and knowledge representation languages such as XML and RDF.

Table I. Semiotics Framework.

Semiotics Level	*Element*	*Description*
Social level	Application domain	The scope and boundary within which a provenance system is designed to be used.
	Data processing architecture	Different architectures or modes of data processing that create the objects of provenance capture.
	Social consequences	The types of consequences that can result from the actions or decision performed by the users of data provenance.
	Action complexity	The nature of the actions. The actions can be structured, semi-structured, or unstructured.
Pragmatics	Acquisition complexity	The level of automation of provenance acquisition techniques as well as costs and overhead incurred.
	Acquisition scope	The range of sources for the acquisition of provenance information.
	Trust and security	The extent to which the captured data provenance can be trusted to be uncompromised and free from error.
	Usability	The extent to which the captured provenance is useful with regard to satisfying various provenance-related inquiries.
Semantics	Semantics of provenance	The meaning of provenance and the aspects of data provenance that are captured.
Syntactics	Representation of provenance	The schemes and languages used for representing provenance.

Our semiotics framework consisting of ten elements for analyzing and comparing provenance research is summarized in Table I.

3. ANALYSIS OF DATA PROVENANCE RESEARCH

The semiotics framework is employed in this section to analyze existing provenance research. This section is divided into subsections reflecting the ten elements of the framework.

3.1 Application Domains

The need for data provenance has been widely acknowledged and is evident in various application domains such as e-science [Frew and Bose 2002; Greenwood et al. 2003; Pancerella 2003], homeland security [Ding et al. 2005; Ceruti et al. 2006], and data warehousing and business intelligence [Buneman et al. 2001; Cui and Widom 2003]. Much of the research into provenance recording has come in the context of domain specific applications, while there also exist several provenance techniques designed to provide a general mechanism for recording provenance for use with multiple applications across domains.

3.1.1 Domain specific applications

Provenance finds its significant use in various science domains. Some of the first

significant research on provenance was conducted in the area of geographic information systems (GIS). Provenance is critical in GIS because it allows one to determine the quality of derived map product [Lanter 1991]. Lanter developed a meta-database for tracking the process of workflows in GIS [Lanter 1991; Lanter and Essinger 1991]. Another GIS system that includes provenance tracking is Geo-Opera [Alonso and Hagen 1997], which extends the approach to recording provenance from GOOSE [Alonso and El Abbadi 1993] and uses data attributes to record the inputs/outputs of data transformations. Related to GIS is the satellite image processing domain. The Earth System Science Workbench (ESSW) [Bose 2002] is a metadata management system for earth science researchers. It captures provenance by recording the sequence of invocations of the data transformation scripts in the form of a DAG. In chemistry, the Collaboratory for Multi-scale Chemical Science (CMCS) captures data provenance using Dublin Core elements such as *Creator* and *Date* [Myers, Chappell et al. 2003a]. Another domain where provenance tools are extremely important is bioinformatics. The myGrid project provides middleware in support of *in silico* experiments [Greenwood et al. 2003; Zhao et al. 2003]. In myGrid, provenance is captured about workflow executions and stored in user's personal repository. In addition, significant research has been carried out on describing the provenance of scientific data in domains such as high energy physics [Cavanaugh et al. 2002], astronomy [Mann 2002], material science [Romeu 1999], etc. In these scientific domains, the data generating processes in the form of workflows are the primary entities for which provenance are collected. Such workflow provenance is captured for validating experiments and ensuring data quality and reliability as the scientific fields is moving towards more collaborative research.

Business users often work with an organized schema and interact with trusted partners. Yet, identifying the source of data enables an analyst to check the origins of suspect or anomalous data to verify the reality of the sources or even repair the source data. In business intelligence and data warehousing applications, provenance is used to trace a view data item back to the source from which it was generated. Cui and Widom present lineage tracing algorithms to identify the exact set of base data that produced a given view data item [Cui et al. 2000]. Tracing the origins of a view data item enables users to detect the source of bad data.

3.1.2 Domain independent applications

Based the observation that data provenance means differently for different people, Ram et al. attempt to formally define the semantics of provenance that can be agreed upon by people from different domains [Ram and Liu 2007]. There are also several systems that can be used to capture provenance across domains. The design for the Chimera Virtual Data System matches the scope and ambition of the Grid, targeting invocations of data transformation in a "distributed, multi-use, multi-institutional environment" [Foster et al. 2002]. Its virtual data language (VDL) provides various commands for extracting derivation and transformation definitions. In Szomszor and Moreau [Szomszor and Moreau 2003], the authors argue for infrastructure support for recording provenance in Grids and presented a prototype system based around a workflow enactment engine submitting provenance data to a

provenance service. The provenance data submitted is information about the invocation of various web services specified by the workflow scripts. The Provenance Aware Service Oriented Architecture (PASOA) [Groth et al. 2004; Groth et al. 2005] builds a provenance infrastructure for recording, storing, and reasoning over provenance. The researchers introduce the provenance recording protocol (PReP) which specifies the messages that actors can exchange with the provenance store in order to record data provenance in the form of the interaction and actor state p-assertions. Provenance Recording Services (PReServ) is a web service implementation of the PReP protocol that stores submitted data provenance in various storage devices [Groth et al. 2005].

3.1.3 Analysis

Many provenance systems are developed to address domain specific provenance needs with their own proprietary methods for recording data provenance, and are therefore unable to be used outside their specific domain. There are several systems aimed to provide a general mechanism for recording data provenance in applications across domains and beyond the confines of a local machine. A common thread connecting these three domain independent provenance systems described above is that they are workflow-centric. They are developed to capture provenance associated with workflows executed in grid computing or web services environments. These provenance systems are "domain independent" but "architecture dependent". They capture provenance for data created in service oriented architectures. There are also provenance capture methods that are designed for database or file systems. We discuss difference modes or architectures of data processing that create the objects of provenance or lineage retrieval below.

3.2 Data Processing Architecture

Simmhan et al. categorizes provenance solutions in terms of database oriented, service-oriented and "other" [Simmhan et al. 2005]. Following [Muniswamy-Reddy et al. 2006], we extend database oriented architecture to include file-systems oriented approaches and name "others" the environment architectures.

3.2.1 Database and file system architectures

Database oriented provenance systems focus on identifying the source (an important aspect of provenance) of data. Significant research on tracking data source started in the early 1990s. Wang and Madnick propose the polygen model and algebra where source attrition can be carried along by results of database queries as a form of provenance annotation [Wang and Madnick 1990]. Woodruff and Stonebraker propose a method to support fine-grained data lineage [Woodruff and M. Stonebraker 1997]. Rather than relying on annotations, their approach computes data lineage by *weak inversion*. Cui et al. studies the problem of computing provenance by analyzing the operations of the relational algebra [Cui et al. 2000]. Buneman et al. develop algorithms for identifying the "where" and "why" provenance [Buneman et al. 2001]. The Trio project [Widom 2005] applies data provenance to probabilistic databases.

It was shown that the provenance of tuples can help correctly capture the set of possible instances in the result of a probabilistic query. A recent work by Buneman et al. proposes a copy-paste model for recording fine-grained provenance in manually curated databases [Buneman et al. 2006]. Different from the above database oriented systems, the Provenance Aware Storage System (PASS) is targeted toward file systems and tracks provenance of data files [Muniswamy-Reddy et al. 2006].

3.2.2 Service-oriented architectures

Recently, various e-sciences applications use provenance systems designed for grid or web service environments since provenance facilitates scientific verification, reproducibility and collaboration. Many of the provenance systems discussed in Section 3.1 such as Chimera [Foster et al. 2002], PASOA [Groth et al. 2004; Groth et al. 2005], and myGrid [Greenwood et al. 2003; Zhao et al. 2003] are based on service-oriented architectures. Most of these systems use a directed acyclic-graph to describe workflows and represent provenance. They include tools that capture provenance during workflow executions and transmit it to a grid provenance service.

3.2.3 Environment architectures

Some collaborative environments have the function of tracking work and recording provenance [Muniswamy-Reddy et al. 2006]. Collaborative information repositories such as Wikipedia automatically capture provenance, i.e., the edit history of data. In scientific domains, the Collaboratory for Multi-scale Chemical Science (CMCS) is an environment for chemists [Myers et al. 2003b], and the Earth System Science Workbench (ESSW) is an environment for earth scientist [Bose 2002]. As long as a user modifies data in one of these environments, the environment can effectively track provenance using various techniques. For example, ESSW captures provenance using custom application programming interface (API) commands within Perl scripts to construct lineage, while CMCS uses an annotation schema to capture provenance and associate it with the files.

3.2.4 Analysis

Different data processing architectures obviously require distinctive mechanisms for capturing data provenances. As a result, most of the current provenance systems are architecture dependent, i.e., they are tailored to capture provenance for data created in a specific processing mode. These systems tend to focus on different aspects of provenance and represent provenance in different formats depending on the data processing architecture they are designed for. As an example, provenance systems designed for databases often capture the source of data in the form inverse queries while in service oriented systems, data provenance often refers to the derivation history of data represented in XML. These differences between province systems designed for different data processing architectures make provenance interoperability and sharing critical issues.

3.3 Social Consequences

Social consequences are concerned with the impact of data provenance on the users' actions and decisions. Goble presents some uses of provenance in various decision making situations, as summarized below [Goble 2002].

- *Data quality:* Given a derived dataset, we need to measure its credibility/quality by investigating its provenance. This is particularly important for data produced in scientific information systems.
- *Audit trail:* Provenance provides an accurate historical record of the source and method of an experiment. In some situations, it will show why certain derivations have been made.
- *Reproducibility & repeatability:* A derivation path is not just a record of what has been done, but also a means by which others can repeat and validate the experiment.
- *Attribution:* Provenance provides a trusted source from which we can procure who the data product belongs to and precisely when and how it was created.

Below, we review research conducted on these application of data provenance.

3.3.1 Data quality

Data quality assessment is widely mentioned in literature as one of the most important uses of provenance [Goble 2002; Tan 2004; Simmhan et al. 2005]. Ceruti et al. even argue that a computational quality model should be an integral part of a provenance framework [Ceruti et al. 2006]. Lynch advocates the integration of trust and provenance into information retrieval systems but does not discuss how provenance can be used [Lynch 2001]. Li Ding et al. argue that the "where", "who", "why" provenance are crucial for determining the trustworthiness of messages; however, they only develop metrics based on *who-provenance* (i.e., the creator of data) [Ding et al. 2005]. The research conducted by Fox and Huang introduces knowledge provenance to create an approach to determining the origin and validity of web information [Fox and Huang 2005]. In addition to using *who-provenance* to determine data validity, the researchers also consider information dependency when accounting for the trustworthiness of derived propositions and temporal factors when the truth value of web propositions may change over time. In [Prat and Madnick 2007], Prat and Madnick propose a framework for estimating the believability of Wikipedia data based on provenance. They adopt a metric developed by Ballou, et al. [Ballou et al. 1998] for determining the temporal believability of data based on provenance information such as when the data is created. They also develop measures for deriving the believability of output data from that of its inputs based on data quality research such as [Ballou and Pazer 1985].

3.3.2 Audit trail

Various e-science applications track data provenance in the form of a workflow for scientists to verify the correctness of their own experiment, or to review the correctness of their peers' work. However, there is only one significant study that proposes a systematic approach to validating e-science experiments using provenance. Miles et al. develop a system for performing workflow validation based on provenance [Miles et al. 2007]. According to [Miles et al. 2007], each workflow consists of a list of

activities, and the details of these activities are recorded as provenance information in the provenance store. The system performs semantic reasoning over the properties of each activity to determine the validity of each activity. If all activities are proved to be valid, then the experiment is valid.

3.3.3 Replication recipes

Zhao et al. call all the aspects of the procedure or workflow used to create a data object the "recipe" for creating that data [Zhao et al. 2006]. Obviously, it is possible to repeat the data creation or transformation if the provenance is detailed enough with precise information on each activity carried, the parameters of the activity, and datasets passed to the activity. The derivation may be repeated to maintain the currency of derived data when then source data changes or if the processing modules were modified [Simmhan et al. 2005]. According to [Cui and Widom 2003], tracking data lineage is related to the well-known view update problem. When data in one database are views derived from underlying source tables, data provenance enables the users to identify the source of the data and update it when the source data changes [Buneman et al. 2001; Cui and Widom 2003].

3.3.4 Attribution

Although there is no significant research that has been conducted specifically on this application of data provenance, it has been well recognized that "a chain of owners" is an important part of data provenance [Tan 2004; Moreau et al. 2007]. Users can identify the creator or owner of data and verify its copyright [Bose 2002]. Also, data provenance acts as one form of citation when publishing scientific datasets to public databases such as GenBank and SWISS-PROT [Simmhan et al. 2005].

3.3.5 Analysis

We focus our analysis on data quality, since significant research has been conducted on using provenance to evaluate data. Data quality is a well established research field. Previous research on data quality such as [Wang and Strong 1996] develops quality model framework by identifying various data quality dimensions such as data accuracy, currency, believability, etc. Many of these dimensions such as currency and believability are related to provenance. Prat and Madnick develop a framework of data believability based on existing data quality research and define some quality metrics using bits and pieces of provenance information [Prat and Madnick 2007]. Although preliminary and not comprehensive enough, this research points out a promising direction for future research. It is necessary to develop a framework for mapping various aspects of provenance (e.g. the source of data or the time of data creation) to relevant quality dimensions and design a methodology for determining data quality based on data provenance in a systematic way.

3.4 Action Complexity

Action complexity refers to the nature of the users' actions performed using data provenance. It primarily deals with the level of automation of the actions. Existing

research focusing on the use of data provenance such as [Prat and Madnick 2007] and [Miles et al. 2007] provides a certain level of automation in using data provenance. Also, tracking provenance such as the source and processing procedures for data in the data warehousing environment allows automatic view update to maintain the currency of derived data when then source data changes or if the processing procedures were modified [Simmhan et al. 2005].

3.4.1 Analysis

As discussed above, the current research on the use of data provenance focuses on using data provenance to assess the quality of data. Various research such as [Goble 2002; Tan 2004; Simmhan et al. 2005, Ceruti et al. 2006] stresses the importance of provenance in data quality assessment but fails to develop a systematic way of assessing quality using provenance. Very little research has been done on other uses of data provenance such as attribution, audit trail, and replication, let alone providing a capacity of using provenance automatically. Hence, an important future research is to discover novel and automatic ways of using data provenance.

3.5 Acquisition Complexity

In this paper, we focus on the level of automation in data provenance acquisition. Automated provenance recording is essential since humans are unlikely to record all the necessary interactions manually [Frew and Bose 2002]. Moreover, unobtrusive information collecting is desirable so that current working practices are not disrupted [Frew and Bose 2002]. Most of the current projects provide a mechanism to automatically capture at least some of the provenance. In this section, we focus on categorizing different approaches to capturing provenance in an automatic or semi-automatic way. We modify and extend Braun et al.'s research [Braun et al. 2006] that classifies provenance systems into observation vs. specification based systems.

3.5.1 Observation based systems

The observation based approach records data provenance by observing a user's actions. GenePattern is an environment for computational biologists [Reich et al. 2006]. It automatically captures provenance for objects created in this environment by observing a user's actions creating a control file to derive the objects. The Provenance Aware Storage System (PASS) automatically tracks provenance at the file system level [Muniswamy-Reddy et al. 2006]. It discovers the components and environment required for the production of a specific data item by observing and tracking system calls made to generate the data item. Curated databases typically involve copying data from other databases and entering provenance manually using a web form. In [Buneman et al. 2006], Buneman et al. provides an automatic approach to capturing provenance in curated databases by developing a provenance-aware web environment that enables the user to import data from a source database and pass it into the target database. The system then observes and captures the user's operations including insertion, deletion, and copy.

3.5.2 Specification based systems

Most of the above-mentioned provenance systems such as myGrid [Greenwood, Goble et al. 2003; Zhao et al. 2003], Chimera [Foster et al. 2002], and PASOA [Groth et al. 2004; Groth et al. 2005] that captures data provenance by recording details about workflow executions fall into the category of specification based provenance systems. These systems normally require that the user provides a workflow specification and discloses explicitly the provenance intended to be captured. A workflow engine then executes the workflow based on the specification, and the metadata associated with the execution is automatically captured. The combination of the workflow specification and the metadata captured during the workflow execution creates provenance in these systems [Braun et al. 2006], which makes them different from observation based provenance systems that capture data provenance by observing the users' *ad hoc* actions. The existing specification based systems are different in entities responsible for capturing provenance. In the myGrid project, the workflow engine records provenance for each step in a workflow, including inputs, outputs and parameters. The PASOA project [Groth et al. 2004; Groth et al. 2005], on the other hand, introduces a burden on the participating service to generate provenance metadata and submit the provenance to a centralized provenance store.

3.5.3 Computation based systems

Provenance systems such as [Wang and Madnick 1990; Cui et al. 2000; Buneman et al. 2001; Cui and Widom 2003] that identify the source of data stored in databases are computation based provenance systems since data provenance, primarily the source of data, is derived via computations. Computation based provenance systems can be further classified since some of the systems compute provenance "lazily" while others compute provenance "eagerly". In [Buneman et al. 2001] and [Cui et al. 2000], a "reverse" query is generated in order to compute data provenance. The reverse query approach is called the "lazy" approach for computing provenance. A query is generated and executed to compute the provenance when needed. Projects such as [Wang and Madnick 1990] and [Bhagwat et al. 2005] propose an annotation based approach where annotations may be attached to a piece of data and are carried along as data is being transformed. In this annotation based approach, the provenance of data is "eagerly" computed and requires minimal computation.

3.5.4 Analysis

The major disadvantage of the current systems using the observation-based approach, according to [Braun et al. 2006], is that they can only capture provenance to which they are exposed, and this frequently produces provenance with less semantic meaning than specification based provenance systems. Specification based approach usually provides richer semantic knowledge than observed systems [Braun et al. 2006]. However, as described in previous sections, specification based systems are usually domain or architecture dependent and often unable to capture the complete provenance of data since the data resulting from a workflow execution may undergo transformations outside the workflow application and may be transferred from one system to another.

The computation based approach is used for identifying the source of data stored in databases. The major disadvantage of the inverse query approach or the "lazy" approach is that it requires a reverse query to be created every time the provenance of output data is sought for. Hence, if the provenance of a large number of output data is required, this may not be the optimal way to compute provenance. The annotation approach or the "eager" approach, on the other hand, trades space for time. However, the cost for storing the annotations can be significant. The size of provenance or annotations can easily exceed that of data's.

When analyzing acquisition complexity, we focus on comparing various provenance acquisition approaches with respect to the level of automation provided by the approach. However, factors other than the level of automation in acquiring provenance including social factors such as copyright and security constraints, data features such as the type, dynamism, and movement of the data, as well as computational resource constraints have a significant impact on acquisition complexity. As an example, the above mentioned observation-based approach can only be applied to capture provenance for data generated and processed in a single controlled environment. As a result, the selection of the most appropriate approach to provenance acquisition relies on considering these social factors and data features.

3.6 Acquisition Scope

Acquisition scope refers to the range of sources for the acquisition of provenance information. Most of the existing approaches to provenance management typically concern provenance collected from a single source. They focus on situations in which all the interactions take place in a single controlled environment such as a database [Buneman et al. 2001; Widom 2005; Buneman 2006] or in which new data is only constructed from existing data using nondestructive mechanisms such as scientific workflows with a workflow engine collecting all the provenance [Foster et al. 2002; Zhao et al. 2003]. Provenance interoperability, i.e., integrating and sharing the provenance provided by different systems when data moves among databases or grid resources, is still a largely unsolved issue. In [Muniswamy-Reddy et al. 2006], researchers find it challenging to collect provenance, when data originates from a non-PASS source such as a user or another computer. They are developing provenance-aware network protocols so that provenance can be atomically transmitted with data [Muniswamy-Reddy et al. 2006]. To address the issue of provenance interoperability, [Groth et al. 2004] proposes the Provenance Recording Protocol for recording and querying provenance through a set of messages exchanged between participating services and a provenance server, thus enabling applications that run under the control of different runtime systems at separate locations to contribute provenance data.

3.6.1 Analysis

Data provenance is extremely useful when data is created in distributed applications and travels among disparate systems. Current provenance solutions are effective for capturing provenance from a single source in a controlled environment. Nonetheless,

tracking the provenance for data that moves among databases or Grid resources is still challenging because there is no one system that can capture all of the actions involved [Buneman 2006]. Instead, many systems must cooperate in order to maintain a network of provenance systems. The incompatibility of current provenance systems, however, prevents provenance from being integrated and shared, which makes provenance interoperability the uttermost important issue that needs to be addressed in future research. We agree with Simmhan et al. [Simmhan et al. 2005] that the work by PASOA is in the right direction, since it promotes federated collection of provenance from systems across organizations instead of a centralized approach where, say, a workflow engine is solely responsible for recording provenance. Protocols like the one proposed by PASOA that controls interactions between various provenance collection actors and the centralized provenance store are undoubtedly useful. However, provenance interoperability first requires the captured provenance to be "semantically" interoperable. As a result, we need a provenance model that defines the semantics of provenance and specifies how provenance can be represented and queried. This provenance model should be generic enough to satisfy the provenance requirements in various application domains. We will discuss the semantic issue of provenance in more detail in Section 3.7.

3.7 Trust and Security

This element concerns whether the captured provenance should be trusted to be free of error and uncompromised. As discussed previously, the various mechanisms for automatically capturing data provenance help avoid human errors. However, the data provenance can still be compromised while it resides in the database. In response, the PASS project provides access controls for provenance by designing a security model for provenance [Braun and Shinnar 2006]. The issue of trust and security is critical in service-oriented architectures where provenance is tracked in a non-reputable manner. In [Tan et al. 2006], Tan et al. develop a trust framework for actors and provenance stores in PASOA, establishing liability for creation of p-assertions and sensitivity of information in p-assertions. The authors also describes some of the basic security issues of enforcing accessing control over provenance [Tan et al. 2006]. The myGrid project also investigates security issues and solutions but in an application dependent manner [Zhao et al. 2003].

3.7.1 Analysis

While a service-oriented approach to collecting provenance is promising since it enables the capture of provenance from multiple heterogeneous systems, effort is required to allow users to place their trust in the provenance submitted by various unknown applications. The trust issue, however, has been largely ignored by existing research and can be a source of research opportunities.

Provenance often requires access controls different from the data it describes [Muniswamy-Reddy et al. 2006]. As an example, the author of a research paper is allowed to view the peer review result but not the provenance of the review. Sometimes, some parts of the provenance are readable or rewritable by a certain

group of users while others are not. As a result, in many applications, we need to define a security model for data provenance separate from that defined for the data the provenance describes.

3.8 Usability

The *usability* of data provenance is assessed based on whether it generates informative and pragmatic impact on a user for the intended usage. As an example, in order for a scientist to determine the currency of some data, the provenance of the data is considered *usability-high* only when the time of the creation of data and its subsequent modifications is captured as a part of the provenance. Since data provenance is often intended for future users who may use it for different purposes, data provenance is more usable when it is *semantically rich*. Ram and Liu define an ontology of data provenance called the W7 model that captures the semantics of data provenance by recording various elements of data provenance, including *what*, *when*, *where*, *who*, *how*, *which*, and *why*, and the relationships between them [Ram and Liu 2007]. With the development of semantic web technologies, there are more projects that capture semantic information within provenance using ontology languages like RDF and OWL [Fileto et al. 2003; Bose and Frew 2004; Zhao et al. 2004] to improve the usability of data provenance. In [Bose and Frew 2004], RDF is used to elaborate on the elementary parent/child relationship between metadata objects. An OWL ontology has been created in [Zhao et al. 2004] to represent the semantics within provenance.

3.8.1 Analysis

We believe that capturing the semantics within provenance using ontologies is a research direction that merits attention. Provenance ontologies clearly define the concepts and relations related to data provenance, thus allowing an enhanced use of provenance and helping to reason about and provide proof statements about the lineage of data [Simmhan et al. 2005]. Moreover, annotating provenance information using domain ontologies can also help enhance the usability of data provenance, since it enables the users to navigate among provenance documents or even from provenance to semantically related information such as the personal website of the data creator or the design specification of an experiment that are useful for the users' actions and decisions.

3.9 Semantics of Provenance

We notice that data provenance means differently for different people. Some researchers define provenance as the origin of data and the process by which it arrived at the database [Buneman et al. 2000], while others view it as metadata recording the process of experiment workflows, annotations and notes about experiment [Frew and Bose 2001]. This distinction results in the classification of *data* vs. *process* provenance.

3.9.1 Data vs. process provenance

Data provenance refers to provenance specifically gathered about the data product. The provenance systems used in a database architecture such as [Buneman et al. 2001; Cui and Widom 2003] focus on deriving data provenance, more specifically the source of data. Data provenance is also captured in various scientific domains. In the domain of biology, provenance has been captured as annotations attached to data stored in genetics database such as SWISSPROT and OMIM. Examples of systems that capture data rather than process provenance also include CMCS [Pancerella 2003] in Chemistry in and LIP [Lanter 1991] in GIS. Various domain specific provenance schemas have been developed in these projects to capture data provenance.

Process provenance, i.e., details about the procedures used for processing the data, describes the derivation path of the data in the form of the workflow of an experiment. Process provenance is effective in describing the processes that derives and transforms the data. Representative examples of systems capturing process provenance include Chimera [Foster et al. 2002], myGrid [Greenwood et al. 2003; Zhao et al. 2003], and Earth System Science Workbench (ESSW) [Frew and Bose 2001].

3.9.2 Coarse-grain vs. fine-grain provenance

An alternative classification scheme is to classify provenance into the *coarse-grain* provenance and *fine-grain* provenance [Buneman and Tan 2007]. The coarse-grain provenance refers to the derivation history of some data set. Most of the above mentioned systems such as Chimeria, myGrid, CMCS, POSOA, and ESSW capture coarse-grain provenance. Fine-grain provenance refers to the derivation of *part* of the resulting data set [Buneman and Tan 2007]. Buneman et al. makes a distinction of the so-called *why* and *where* provenance [Buneman et al. 2001]. The former normally refers to tuples in the source databases that had some influence on the existence of the target data; the latter specifies the exact source element where data in the target is copied from. In essence, this classification deals with the granularity of data the provenance describes. Data products that are subsets of a parent dataset may inherit some provenance from the parent as well as share their provenance with their parent. Provenance inheritance or sharing among data at different granularities is an issue that demands further research.

3.9.3 Analysis

An important design choice in developing provenance systems is to capture process vs. data provenance. In service-oriented environments such as myGrid [Greenwood et al. 2003; Zhao et al. 2003] and PASOA [Groth et al. 2004], where data are generated as the result of a workflow execution, it is a natural choice to capture process provenance by tracing the execution of the workflow and identifying the input and output data to each service. In contract, data provenance may have a costly overhead to execute a process related query [Simmhan et al. 2005]. However, data may travel outside the service-oriented environment. In such a situation, we need to derive the complete provenance of data and make it accessible for other systems. It is potentially costlier to extract data provenance from process provenance since this involves examining all process oriented provenance records in which this data appears [Simmhan et al. 2005]. As a result, in [Simmhan et al. 2006],

Simmhan et al. suggests that in grid or web services environments, process provenance and data provenance should be captured separately. This data provenance should be independent of the workflow engine and data storage system such that it will not be lost when data goes from one resource to another. Loose coupling with these external entities enables provenance to be collected in a heterogeneous grid environment [Simmhan et al. 2006].

As suggested by [Pearson 2002], data provenance needs to be captured with the hope that it is comprehensive enough to be useful in the future. However, current efforts on capturing data provenance focus on only one or two aspects of provenance. As an example, both "why" and "where" provenance proposed by Buneman et al. help tracing the source from which the data came from. Undoubtedly, locating the source data is important provenance knowledge especially for derived or imported data. Nevertheless, information such as how the data has been derived by whom with which program may be equally important. Moreover, data that resides in a database may be obtained in different ways than deriving it from existing sources. It may be the result of measurement, observation or surveys. Oftentimes, it may have undergone changes since it arrives in the database. Data provenance, therefore, means much more than what is captured in [Buneman et al. 2001], i.e., the source of data. It may include the creator of data, its history in terms of how the data was obtained and transformed, and the sequence of ideas leading to an experiment, just to name a few. Current practices on data provenance often focus only on some aspects of data provenance while ignoring others. As a result, provenance knowledge captured is often incomplete and can not be shared across applications. The absence of a provenance model that can satisfy the provenance requirements for various types of data is an obvious hindrance to promoting provenance interoperability, as discussed in II.5.1. In response to this problem, Ram and Liu define the W7 model, a generic model that captures the semantics of data provenance [Ram and Liu 2007]. The W7 model represents data provenance as a combination of seven interconnected elements including, "what" (i.e., events such as data creation and transformation), "when" (i.e., the time of the event), "where" (i.e., the location), "how" (i.e., the actions and processes), "who" (i.e., the agent initiating the event), "which" (i.e., the instrument or software), and "why" (i.e., the goal or reasons for the event). We believe the W7 model is an important step toward a standard provenance model that can be agreed upon by people in different application domains. Such a standard model is required so data provenance can be shared communicated reliably between systems.

3.10 Provenance Representation

The existing research efforts on data provenance represent data provenance in two forms: (1) inverse queries and (2) annotations.

3.10.1 Inverse queries

Inverse queries are often used to represent data provenance in the relational databases, where the provenance of a piece of data d in the output of a query Q is the

answer to the following question: which parts of the source database D contribute to d in the output according to Q. Techniques such as [Buneman et al. 2001; Cui and Widom 2003] compute the provenance of a piece of output data "eagerly" by analyzing the input database and output database, as well as the definition of Q, to arrive at the answer. These techniques record an inverse query Q' derived from the original query Q as a compact representation of data provenance without explicitly recording the provenance in the database. The inverse query Q' is applied to the output data to identify the provenance, i.e., the source data of the output.

3.10.2 Annotations

The annotation approach, in contrast, represents provenance as annotations that are attached to the data or provided upon request. As discussed in Section 3.4.3, techniques such as [Wang and Madnick 1990; Bhagwat et al. 2005] compute provenance "lazily". They carry data provenance as annotations to the output database. Scientific databases normally support data lineage using annotations [Cui et al. 2000]. In scientific workflows such as [Foster et al. 2002; Zhao et al. 2003; Simmhan et al. 2006], data provenance comprising of the derivation history of the final output of the workflows is often represented as annotations.

Most of the current provenance systems that represents provenance as annotations have adopted XML for representing data provenance [Myers et al. 2003b; Zhao et al. 2003; Groth et al. 2004]. The benefits of this are apparent given that many of them using service-oriented architectures where XML is the primary language [Simmhan et al. 2005].

3.10.3 Analysis

Representing provenance as inverse queries has limited applications since it is restricted to a certain class of relational queries. Moreover, the inverse approach is applicable to only one type of data provenance, i.e., the source data in the relational setting. The obvious advantage of this approach, compared to the annotation approach, is that simply recording a single inverse query provides a compact way of representing data provenance. Representing provenance as annotations, on the other hand, is being used by most of the existing provenance techniques. It eliminates the need to derive provenance "just-in-time" like the inversion approach. More importantly, it provides the flexibility to represent provenance information that is semantically rich.

So far we have discussed each semiotics level and each element that belongs to the level separately. However, the four semiotics levels as well as their elements are closely related. Stamper represents the semiotics framework in the form of a "semiotic ladder" [Stamper 1991], with the following steps from bottom to top: syntactics, semantics, pragmatics and social level. When we take the ladder from the bottom up, we move from syntactics and semantics to pragmatics and then arrive at the social level. Properties on higher levels of the semiotics ladder rely on those on lower levels. As an example, the usability of data provenance relies largely on the syntactics/semantics of provenance. Provenance is more usable when it is semantically rich and represented in an easily retrievable format. Also, the users' actions performed

based on provenance have significant social consequences when the provenance is usable, semantically rich and meaningful, and represented in an easily retrievable and understandable format. We can also move down the ladder from the social level. For instance, provenance is usable only when it is informative and pragmatic for intended usage.

Table II provides a summary of the existing research mentioned in Section 3. We classify the existing research on data provenance based on our semiotics framework.

4. DISCUSSION

In this paper, we propose a semiotics framework for analyzing and comparing provenance techniques discussed in the extant research. Existing surveys on provenance research such as [Bose and Frew 2005; Simmhan et al. 2005] adopt a literature review based approach that derives a classification framework for classifying and

Table II. Summary of the Existing Research Based on the Semiotics Framework.

Semiotic Level	*Element*	*Classification*	*Existing Research*
Social level	Application domain	Domain specific	(Lanter 1991; Lanter and Essinger 1991; Alonso and Hagen 1997) in GIS, (Myers et al. 2003b) in chemistry, (Bose 2002) in earth science, (Greenwood et al. 2003) in biology, (Cavanaugh et al. 2002) in physics, (Cui et al. 2000) in business intelligence.
		Domain independent	(Foster et al. 2002), (Szomszor and Moreau 2003), (Groth et al. 2004; Groth et al. 2005).
	Data processing architecture	Database & file system	(Wang and Madnick 1990), (Woodruff and M. Stonebraker 1997), (Cui et al. 2000), (Buneman et al. 2001), (Widom 2005), (Buneman et al. 2006), (Muniswamy-Reddy et al. 2006).
		Service-oriented	(Foster et al. 2002), (Groth et al. 2004; Groth et al. 2005), (Greenwood et al. 2003; Zhao et al. 2003).
		Environment	(Myers et al. 2003b), (Bose 2002).
	Social Consequences	Data quality	(Simmhan et al. 2005), (Ceruti et al. 2006), (Lynch 2001), (Ding et al. 2005), (Fox and Huang 2005), (Prat and Madnick 2007).
		Audit trail	(Miles et al. 2007)
		Replication & reproduction	(Zhao et al. 2006)
		Attribution	(Tan 2004), (Moreau et al. 2007)
	Action Complexity	N/A	(Prat and Madnick 2007), (Zhao et al. 2006).

Table II. Summary of the Existing Research Based on the Semiotics Framework (Continued).

Semiotic Level	*Element*	*Classification*	*Existing Research*
Pragmatics	Acquisition complexity	Observation-based	(Reich et al. 2006), (Muniswamy-Reddy et al. 2006), (Buneman et al. 2006)
		Specification-based	(Greenwood et al. 2003, Zhao et al. 2003), (Foster et al. 2002), (Groth et al. 2004, Groth et al. 2005), (Braun et al. 2006).
		Computation-based	(Buneman et al. 2001), (Cui and Widom 2003), (Bhagwat et al. 2005).
	Acquisition scope	Single controlled environment	(Buneman et al. 2001), (Widom 2005), (Buneman 2006), (Foster et al. 2002), (Zhao et al. 2003), (Muniswamy-Reddy et al. 2006)
		Distributed environment	(Groth et al. 2004, Groth et al. 2005)
	Trust and security	N/A	(Braun and Shinnar 2006), (Tan et al. 2006), (Zhao et al. 2003)
	Usability	N/A	(Fileto et al. 2003), (Bose and Frew 2004), (Zhao et al. 2004), (Ram and Liu 2007).
Semantics	Semantics of provenance	Data-oriented	(Lanter 1991), (Buneman et al. 2001), (Cui and Widom 2003), (Pancerella 2003), (Ram and Liu 2007).
		Process-oriented	(Foster et al. 2002), (Greenwood et al. 2003), (Zhao et al. 2003), (Frew and Bose 2001).
		Coarse-grained	(Greenwood et al. 2003), (Zhao et al. 2003), (Frew and Bose 2001).
		Fine-grained	(Buneman et al. 2001), (Buneman and Tan 2007).
Syntactics	Representation of provenance	Annotation	(Wang and Madnick 1990), (Bhagwat et al. 2005), (Myers et al. 2003b), (Zhao et al. 2003), (Groth et al. 2004).
		Inverse queries	(Buneman et al. 2001), (Cui and Widom 2003), (Buneman 2006)

comparing existing provenance research based on reviewing and intuitive understanding of the existing research. In our research, we adopt a theory based approach and propose to identify and analyze the essential properties and features of various provenance techniques based on a validated theory, namely, the theory of semiotics, with the purpose of providing rigor and organization to the analysis and identifying issues that have not been sufficiently addressed by the existing research. Our survey, together with other surveys on data provenance such as [Bose and Frew 2005; Simmhan et al. 2005], presents a comprehensive picture of the current status of research on data provenance and points out the direction of future research.

Analyzing existing research based on our semiotics framework helps us identify

interesting open research questions on provenance and challenges that need to be overcome.

First, an ontology that capture the semantics of data provenance is needed to support provenance capture and sharing between systems. Standardizing the semantics of data provenance will allow unambiguous interpretation of provenance, support sharing of data provenance between systems, and improve the usability of data provenance by enabling richer queries. The myGrid project is progressing along these lines by migrating to the Web Ontology Language (OWL) for describing their provenance [Zhao et al. 2003]. CMCS too has preliminary support for a semantic description of provenance that can be improved upon by using specific semantic terms to describe provenance instead of using overloaded Dublin Core verbs [Myers et al. 2003b]. Ram and Liu define the W7 model, a generic model that captures the semantics of data provenance [Ram and Liu 2007]. It is an important step toward a standard provenance model that can be agreed upon by people in different application domains. Such a standard model is required so that data provenance can be shared and communicated reliably between systems.

Secondly, we need to develop methods to federate data provenance collected from different sources. Data is increasingly being shared across organizations and it is essential for provenance to be shared along with the data, wherever the data goes. As discussed above, if data provenance is to be federated, we need to overcome the semantic heterogeneity and agree on the *semantics* of data provenance. Moreover, most of the research surveyed has its own proprietary protocols for managing and transferring provenance, and the absence of open standards for collecting, representing, transferring, and querying for provenance is an obvious hindrance to promoting provenance federation [Simmhan et al. 2005]. The work by PASOA [Groth et al. 2004; Groth et al. 2005] on defining a provenance recording API is in the right direction but needs further refinement on how provenance is represented, queried, and transferred as the data travels.

Thirdly, we need to further investigate the uses of data provenance. The focus of the current research on the uses of provenance is assessing data quality using provenance, but we still lack a systematic approach to using provenance to assess data quality metrics. Prat and Madnick proposed a framework for evaluating the believability of data, a dimension of data quality, based on bits and pieces of provenance information [Prat and Madnick 2007]. Although preliminary and not comprehensive enough, this research points out a promising direction for future research. Also, little research has been done to investigate other uses of data provenance such as audit trail, replication and reproduction, and attribution. According to [Simmhan et al. 2005], discovering novel ways to use provenance will drive more organizations to collect provenance. For this to happen, provenance needs to be fully understood and studied in the context of its potential use in various application domains.

Fourthly, we need to develop approaches to guaranteeing the security and trustworthiness of provenance. Using provenance for decision making largely depends upon its trustworthiness. Data provenance must be collected from trusted sources. Also, there should be a mechanism to ensure that the captured data provenance

remains uncompromised even after the provenance travels from one system to another. Signing provenance using digital signatures is a possible solution. In many applications, we also need to define a model of access control for data provenance separate from that defined for the data the provenance describes.

5. CONCLUSION

As a well-established theory describing sign-based communications, semiotics can be used to describe the form-, meaning- and use-related aspects of information systems. In this research, we develop a *semiotics framework* for analyzing data provenance research, consisting of ten elements. The framework provides a coherent way to distinguish among various types of provenance systems, thereby providing a clear view of the state-of-art research in this area. The framework helps us identify several issues that need to be addressed in future research. Highly significant among these are (1) a standard semantic model to support provenance capture and sharing between systems, (2) ways to federate provenance information collected from different sources, (3) a systematic framework for using data provenance to evaluate data quality, and (4) approaches to guaranteeing the security and trustworthiness of provenance. Finally, our semiotics framework is an attempt to ground data provenance research on a well established theoretical foundation. We are exploring the possibility of extending the framework and using it to provide a requirement specifications checklist for provenance system development.

REFERENCES

ALONSO, G. AND A. EL ABBADI. 1993. Goose: Geographic object oriented support environment. In *ACM Workshop on Advances in Geographic Information Systems*, Arlington, Virginia, 38–49.

ALONSO, G. AND C. HAGEN. 1997. Geo-opera: Workflow concepts for spatial processes. In *5th International Symposium on Spatial Databases*, Berlin, Germany, 238–258.

ANDERSEN, P. 1991. A semiotic approach to construction and assessment of computer systems. Information Systems research: Contemporary Approaches & Emergent Traditions. Nissen, Klein and Hirschhaim. North Holland, *Elsevier Science Publishers*, 465–514.

BALLOU, D., R. Y. WANG, et al. 1998. Modeling information manufacturing systems to determine information product quality. *Management Science*, 44(4):462–484.

BALLOU, D. P. AND H. L. PAZER. 1985. Modeling data and process quality in multi-input, multioutput information systems. *Management Science*, 31(2):150–162.

BARRON, T. M., R. H. L. CHIANG, et al. 1999. A semiotics framework for information systems classification and development. *Decision Support Systems*, 25:1–17.

BHAGWAT, D., L. CHITICARIU, et al. 2005. An annotation management system for relational databases. *VLDB JOURNAL*, 14(4):373–396.

BOSE, R. 2002. A conceptual framework for composing and managing scientific data lineage. In *14th International Conference on Scientific and Statistical Database Management*, 15–19.

BOSE, R. and J. FREW. 2004. Composing Lineage Metadata with XML for Custom Satellite-Derived Data Products. In *the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004)*, Greece, 275–284.

BOSE, R. and J. FREW. 2005. Lineage Retrieval for Scientific Data Processing: A Survey. *ACM Computing Surveys*, 37(1):1–28.

BRAUN, U., S. GARFINKEL, et al. 2006. Issues in Automatic Provenance Collection. Lecture Notes in Computer Science 4145. L. Moreau and I. Foster. Springer, 171–183.

BRAUN, U. and A. SHINNAR. 2006. A Security Model for Provenance, Harvard University.

BUNEMAN, P. 2006. Provenance management in curated database. SIGMOD, Chicago, Illinois, 539–550.

BUNEMAN, P., A. CHAPMAN, et al. 2006. A provenance model for manually curated data. LNCS 4145. L. Moreau and I. Foster. Berlin/Heidelberg, Springer, 162–170.

BUNEMAN, P., S. KHANNA, et al. 2000. Data Provenance: Some Basic Issues. FSTTCS, New Delhi, India, 87–93.

BUNEMAN, P., S. KHANNA, et al. 2001. Why and Where: A Characterization of Data Provenance. Lecture Notes in Computer Science 1973, Springer, 316–330.

BUNEMAN, P. and W. TAN. 2007. Provenance in Databases. SIGMOD, Beijing, China, 1171–1173.

CAVANAUGH, R., G. GRAHAM, et al. 2002. Satisfying the Tax Collector: Using Data Provenance as a way to audit data analyses in High Energy Physics. *Workshop on Data Derivation and Provenance.*

CERUTI, M., S. DAS, et al. 2006. Pedigree Information for Enhanced Situation and Treat Assessment. In *9th International Conference on Information Fusion (ICIF 2006)*, Florence, Italy.

CUI, Y. and J. WIDOM. 2003. Lineage tracing for general data warehouse transformation. *VLDB Journal*, 12:41–58.

CUI, Y., J. WIDOM, et al. 2000. Tracing the Lineage of View Data in a Warehousing Environment. *ACM Transactions on Database Systems*, 25(2):179–227.

DING, L., P. KOLARI, et al. 2005. On Homeland Security and the Semantic Web: a Provenance and Trust Aware Inference Framework. *AAAI Spring Symposium on AI Technologies for Homeland Security*, Stanford University, CA, 1–8.

FILETO, R., C. B. MEDEIROS, et al. 2003. Using domain ontologies to help track data provenance. *LNCS*, 3806:84–98.

FOSTER, I. T., J. S. VOECKLER, et al. 2002. Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation. In *14th International Conference on Scientific and Statistical Database Management*, 37–46.

FOX, M. and J. HUANG. 2005. Knowledge Provenance in Enterprise Information. *International Journal of Production Research*, 43(20):4471–4492.

FREW, J. and R. BOSE. 2001. Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products. In *13th International Conference on Scientific and Statistical Database Management*, Fairfax, VA, 180–189.

FREW, J. and R. BOSE. 2002. Lineage issues for scientific data and information. *Data provenance/derivation workshop.*

GOBLE, C. 2002. Position statement: Musings on provenance, workflow and annotations for bioinformatics. In *Data provenance/derivation workshop.*

GREENWOOD, M., C. GOBLE, et al. 2003. Provenance of e-Science Experiments - experience from Bioinformatics. UK e-Science All Hands Meeting, Nottingham, UK.

GROTH, P., M. LUCK, et al. 2004. A protocol for recording provenance in service oriented grids. Lecture Notes in Computer Science 3544/2005, Springer, 124–139.

GROTH, P., S. MILES, et al. 2005. Recording and Using Provenance in a Protein Compressibility Experiment. High Performance Distributed Computing, HPDC-14. 201–208.

LANTER, D. 1991. Design of a Lineage-Based Meta-Data Base for GIS. *Cartography and Geographic Information Systems*, 18:255–261.

LANTER, D. and R. ESSINGER. 1991. User-centered graphical user interface design for GIS. *National Center for Geographic Information and Analysis*, UCSB: 91–96.

LYNCH, C. 2001. When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web. *Journal of the American Society for Information Science and Technology*, 52(1):12–17.

MANN, B. 2002. Annotation of special structures in astronomy. In *Workshop on Data Derivation*

and Provenance, Chicago, Illinois.

MILES, S., S. WONG, et al. 2007. Provenance-based validation of e-science experiments. Web Semantics: Science, Services and Agents on the World Wide Web 5(1):28–38.

MOREAU, L., P. GROTH, et al. 2007. The Provenance of Electronic Data. *Communications of the ACM*, 51(4):52–58.

MORRIS, C. W. 1946. Signs, Language and Behavior, Prentice-Hall, New York.

MUNISWAMY-REDDY, K., D. HOLLAND, et al. 2006. Provenance-Aware Storage System. *In the 2006 USENIX Annual Technical Conference*, Boston, MA: 4–4.

MYERS, J., A. CHAPPELL, et al. 2003a. Re-integrating the research record. *IEEE Computing in Science & Engineering*, 5(3):44–50.

MYERS, J., C. PANCERELLA, et al. 2003b. Multi-scale Science: Supporting Emerging Practice with Semantically-Derived Provenance. In *Semantic Web Technologies for Searching and Retrieving Scientific Data Workshop at the 2nd International Semantic Web Conference*, Sanibel Island, FL.

PANCERELLA, C. 2003. Metadata in the Collaboratory for Multi-scale Chemical Science. In *DC-2003: the 2003 Dublin Core Conference*, Seattle, Washington.

PEARSON, D. 2002. The Grid: Requirements for Establishing the Provenance of Derived Data. In *Workshop on Data Derivation and Provenance*, Chicago, Illinois.

PRAT, N. and S. MADNICK. 2007. Evaluating and Aggregating Data Believability across Quality Sub-Dimensions and Data Lineage. In *Seventeenth Annual Workshop on Information Technologies and Systems (WITS2007)*, Montreal, Canada.

RAM, S. and J. LIU. 2007. W7 Model: an Ontological Model for Capturing Data Provenance Semantics. Lecture Notes in Computer Science 4512. P. Chen, Springer: 17–29.

REICH, M., T. LIEFELD, et al. 2006. GenePattern 2.0. *Nature Genetics*, 38:500–501.

ROMEU, J. L. 1999. Data Quality and Pedigree, *Material Ease*, 1999.

SAUSSURE. 1966. Course in General Linguistics. *McGraw-Hill.*

SIMMHAN, Y., B. PLALE, et al. 2005. A Survey of Data Provenance Techniques. Technical Report IUB-CS-TR618, Indiana University.

SIMMHAN, Y., B. PLALE, et al. 2006. A Framework for Collecting Provenance in Data-Centric Scientific Workflows. *The IEEE International Conference on Web Services.* 427–436.

STAMPER, R. 1991. The Semiotic Framework for Information Systems Research. Information Systems Research: Contemporary Approaches and Emergent Traditions. H. Nissen, H. Klein and R. Hirschheim, *Elsevier Science Publishers*, 515–527.

SZOMSZOR, M. and L. MOREAU. 2003. Recording and reasoning over data provenance in web and grid services. Lecture Notes in Computer Science 2888, Springer, 603–620.

TAN, V., P. GROTH, et al. 2006. Security Issues in a SOA-Based Provenance System. LNCS 4145. L. Moreau and I. Foster. Springer, 203–211.

TAN, W. 2004. Research Problems in Data Provenance. *IEEE Data Engineering Bulletin*, 27(4):45–52.

WANG, R. and D. STRONG. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–30.

WANG, Y. R. and S. E. MADNICK. 1990. A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. In *the sixteenth international conference on Very large databases*, Brisbane, Australia, 519–533.

WIDOM, J. 2005. Trio: A System for Integrated Management of Data, Accuracy and Lineage. In *Biennial Conference on Innovative Data Systems Research (CIDR)*, 262–276.

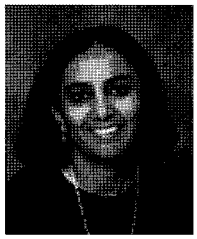
WOODRUFF, A. and M. STONEBRAKER. 1997. Supporting Fine-grained Data Lineage in a database Visualization Environment. In *13th International Conference on Data Engineering (ICDE)*, 91–102.

ZHAO, Y., M. WILDE, et al. 2006. Applying the Virtual Data Provenance Model. Lecture Notes in Computer Science 4145. L. Moreau and I. Foster. Springer, 148–161.

ZHAO, J., C. GOBLE, et al. 2003. Annotating, linking and browsing provenance logs for e-Science.

In *2nd Intl Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data*, Sanibel Island, FL.

ZHAO, J., C. WROE, et al. 2004. Using Semantic Web Technologies for Representing E-science Provenance. Lecture Notes in Computer Science 3298. Berlin/Heidelberg, Springer, 92–106.



Sudha Ram Sudha Ram is McClelland Professor of Management Information Systems in the Eller School of Management at the University of Arizona. She received her MBA from the Indian Institute of Management, Calcutta in 1981 and a Ph.D. from the University of Illinois at Urbana-Champaign, in 1985.

Dr. Ram has published in such journals as CACM, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Information Systems Research, Management Science, and MIS Quarterly. Dr. Ram's research deals with issues related to Enterprise Data Management, including, Interoperability in Heterogeneous Database Systems, Semantic Modeling, Data Provenance, and Biological Data Integration. Her research has been funded by SAP, IBM, Raytheon, US ARMY, NIST, NSF, NASA, and ORD (CIA). Dr. Ram serves as a senior editor for Information Systems Research, and many other journal editorial boards. She is a cofounder of the Workshop on Information Technology and Systems (WITS) and chair of the steering committee for the ER Conference. Dr. Ram is a member of ACM, IEEE Computer Society, INFORMS, and AIS. She is also the director of the Advanced Database Research Group at the University of Arizona.

Jun Liu Jun Liu is a doctoral student in Management Information Systems at the Eller School of Management of the University of Arizona. His research interests are in the areas of data provenance, data/information quality, knowledge sharing and coordination, data modeling, and intelligent agents for information resource management. He has an MS in Management Information Systems from the University of Arizona.