A Survey of Data Provenance in e-Science

Yogesh L. Simmhan Beth Plale Dennis Gannon
Computer Science Department
Indiana University, Bloomington, IN 47405
{ysimmhan, plale, gannon}@cs.indiana.edu

ABSTRACT

Data management is growing in complexity as large-scale applications take advantage of the loosely coupled resources brought together by grid middleware and by abundant storage capacity. Metadata describing the data products used in and generated by these applications is essential to disambiguate the data and enable reuse. Data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources.

In this paper we create a taxonomy of data provenance characteristics and apply it to current research efforts in e-science, focusing primarily on scientific workflow approaches. The main aspect of our taxonomy categorizes provenance systems based on why they record provenance, what they describe, how they represent and store provenance, and ways to disseminate it. The survey culminates with an identification of open research problems in the field.

1. Introduction

The growing number and size of computational and data resources coupled with uniform access mechanisms provided by a common Grid middleware stack is allowing scientists to perform advanced scientific tasks in collaboratory environments. Scientific workflows are the means by which these tasks can be composed. The workflows can generate terabytes of data, mandating rich and descriptive metadata about the data in order to make sense of it and reuse it. One kind of metadata is provenance (also referred to as lineage and pedigree), which tracks the steps by which the data was derived and can provide significant value addition in such data intensive e-science projects.

Scientific domains use different forms of provenance and for various purposes. Publications are a common form of representing the provenance of experimental data and results. Increasingly, Digital Object Identifiers (DOIs) [1] are used to cite data used in experiments so that the papers can relate the experimental process and analysis – which form the data's lineage – to the actual data used and produced. Some scientific fields go beyond this and store lineage information in a machine accessible and understandable form. Geographic information system (GIS) standards suggest that metadata about the quality of datasets should include a description of the lineage of the data product to help the data users to decide if the dataset meets the requirement of their application [2]. Materials engineers choose materials for the design of critical components, such as for an airplane, based on their statistical analysis and it is essential to establish the pedigree of this data to prevent system failures and for audit [3]. When sharing biological and biomedical data in life sciences research, presence of its transformation record gives a context in which it can be used and also credits the author(s) of the data [4]. Knowledge of provenance is also relevant from the perspective of regulatory mechanisms to protect intellectual property [5]. With a large number of datasets appearing in the public domain, it is increasingly important to determine their veracity and quality. A detailed history of the data will allow the users to discern for themselves if the data is acceptable.

Provenance can be described in various terms depending on the domain where it is applied. Buneman et al [6], who refer to data provenance in the context of database systems, define it as the description of the origins of data and the process by which it arrived at the database. Lanter [7], who discusses derived data products in GIS, characterizes lineage as information describing materials and transformations applied to derive the data. Provenance can be associated not just with data products, but with the process(es) that enabled their creation as well. Greenwood et al [8] expand Lanter's definition and view it as metadata recording the process of experiment workflows, annotations, and notes about experiments. For the purposes of this paper, we define data provenance as information that helps determine the derivation history of a data product, starting from its original sources. We use the term data product or dataset to refer to data in any form, such as files, tables, and virtual collections. The two important features of the provenance of a data product are the ancestral data product(s) from which this data product evolved, and the process of transformation of these ancestral data product(s), potentially through workflows, that helped derive this data product.

In this survey, we compare current data provenance research in the scientific domain. Based on an extensive survey of the literature on provenance [9], we have developed a taxonomy of provenance techniques that we use to analyze five selected systems. Four of the projects use workflows to perform scientific experiments and simulations. The fifth research work investigates provenance techniques for data transformed through queries in database systems. The relationship between

workflows and database queries with respect to lineage is evident[1]. Research on tracking the lineage of database queries and on managing provenance in workflow systems share a symbiotic relationship, and the possibility of developing cross-cutting techniques is something we expose in this study. We conclude this survey with an identification of open research problems. The complete version of this survey [9] reviews an additional four systems and also investigates the use of provenance in the business domain.

While data provenance has gained increasing interest recently due to unique desiderata introduced by distributed data in Grids, few sources are available in the literature that compare across approaches. Bose et al [10] survey lineage retrieval systems, workflow systems, and collaborative environments, with the goal of proposing a meta-model for a systems architecture for lineage retrieval. Our taxonomy based on usage, subject, representation, storage, and dissemination more fully captures the unique characteristics of these provenance systems. Miles et al [11] study use cases for recording provenance in e-science experiments for the purposes of defining the technical requirements for a provenance architecture. We prescribe no particular model but instead discuss extant models for lineage management that can guide future provenance management systems.

2. Taxonomy of Provenance Techniques

Different approaches have been taken to support data provenance requirements for individual domains. In this section, we present a taxonomy of these techniques from a conceptual level with brief discussions on their pros and cons. A summary of the taxonomy is given in **Figure 1**. Each of the five main headings is discussed in turn.

2.1 Application of Provenance

Provenance systems can support a number of uses [12, 13]. Goble [14] summarizes several applications of provenance information as follows:

- *Data Quality*: Lineage can be used to estimate data quality and data reliability based on the source data and transformations [4]. It can also provide proof statements on data derivation [15].
- *Audit Trail*: Provenance can be used to trace the audit trail of data [11], determine resource usage [8], and detect errors in data generation [16].
- *Replication Recipes*: Detailed provenance information can allow repetition of data derivation, help maintain its currency [11], and be a recipe for replication [17].
- *Attribution*: Pedigree can establish the copyright and ownership of data, enable its citation [4], and determine liability in case of erroneous data.
- *Informational*: A generic use of lineage is to query based on lineage metadata for data discovery. It can also be browsed to provide a context to interpret data.

2.2 Subject of Provenance

Provenance information can be collected about different resources in the data processing system and at multiple levels of detail. The provenance techniques we surveyed focus on data, but this data lineage can either be available explicitly or deduced indirectly. In an explicit model, which we term a *data-oriented* model, lineage metadata is specifically gathered about the data product. One can delineate the provenance metadata about the data product from metadata concerning other resources. This contrasts to a *process-oriented*, or indirect, model where the deriving processes are the primary entities for which provenance is collected, and the data provenance is determined by inspecting the input and output data products of these processes [18].

The usefulness of provenance in a certain domain is linked to the *granularity* at which it is collected. The requirements range from provenance on attributes and tuples in a database [19] to provenance for collections of files, say, generated by an ensemble experiment run [20]. Increasing use of abstract datasets [17, 18] that refer to data at any granularity or format allows a flexible approach. The cost of collecting and storing provenance can be inversely proportional to its granularity.

2.3 Representation of Provenance

Different techniques can be used to represent provenance information, some of which depend on the underlying data processing system. The manner in which provenance is represented has implications on the cost of recording it and the richness of its usage. The two major approaches to representing provenance information use either annotations or inversion. In the former, metadata comprising of the derivation history of a data product is collected as *annotations* and descriptions about source data and processes. This is an *eager* form [21] of representation in that provenance is pre-computed and readily usable as metadata. Alternatively, the *inversion* method uses the property by which some derivations can be inverted to find the input data supplied to them to derive the output data. Examples include queries and user-defined functions in databases that can be inverted automatically or by explicit functions [19, 22, 23]. In this case, information about the queries and the output data may suffice to identify the source data.

While the inversion method has the advantage of being more compact than the annotation approach, the information it provides is sparse and limited to the derivation history of the data. Annotations, on the other hand, can be richer and, in addition to the derivation history, often include the parameters passed to the derivation processes, the versions of the workflows that

[1] Workflows form a graph of processes that transform data products. Database queries can form a graph of operations that operate on tables.

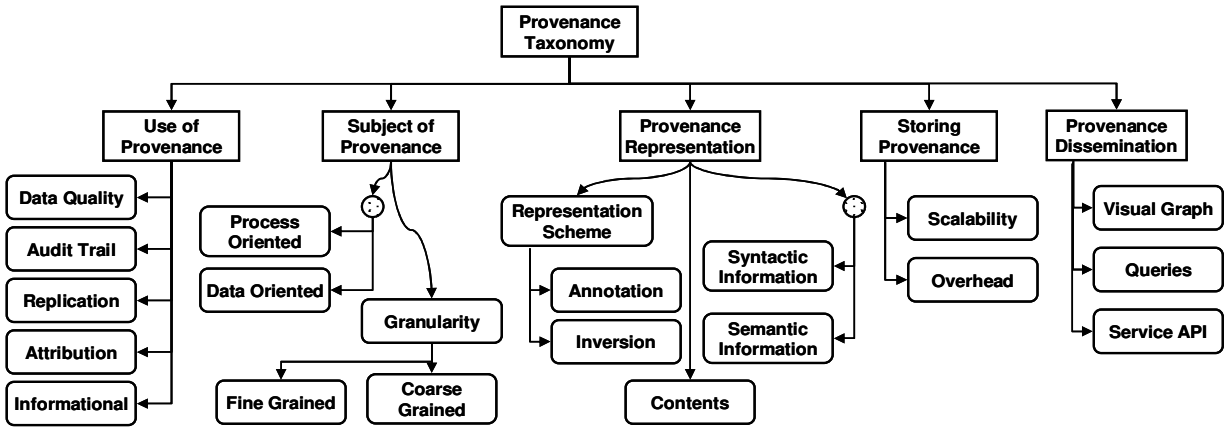

Figure 1 Taxonomy of Provenance

will enable reproduction of the data, or even related publication references [24].

There is no metadata standard for lineage representation across disciplines, and due to their diverse needs, it is a challenge for a suitable one to evolve [25]. Many current provenance systems that use annotations have adopted XML for representing the lineage information [11, 18, 25, 26]. Some also capture semantic information within provenance using domain ontologies in languages like RDF and OWL [18, 25]. Ontologies precisely express the concepts and relationships used in the provenance and provide good contextual information.

2.4 Provenance Storage

Provenance information can grow to be larger than the data it describes if the data is fine-grained and provenance information rich. So the manner in which the provenance metadata is stored is important to its *scalability*. The inversion method discussed in section 2.3 is arguably more scalable than using annotations [19]. However, one can reduce storage needs in the annotation method by recording just the immediately preceding transformation step that creates the data and recursively inspecting the provenance information of those ancestors for the complete derivation history.

Provenance can be tightly coupled to the data it describes and located in the same data storage system or even be embedded within the data file, as done in the headers of NASA Flexible Image Transport System files. Such approaches can ease maintaining the integrity of provenance, but make it harder to publish and search just the provenance. Provenance can also be stored with other metadata or simply by itself [26]. In maintaining provenance, we should consider if it is immutable, or if it can be updated to reflect the current state of its predecessors, or whether it should be versioned [14]. The provenance collection mechanism and its storage repository also determine the trust one places in the provenance and if any mediation service is needed [11].

Management of provenance incurs *costs* for its collection and for its storage. Less frequently used provenance information can be archived to reduce storage overhead or a demand-supply model based on usefulness can retain provenance for those frequently used. If provenance depends on users manually adding annotations instead of automatically collecting it, the burden on the user may prevent complete provenance from being recorded and available in a machine accessible form that has semantic value [18].

2.5 Provenance Dissemination

In order to use provenance, a system should allow rich and diverse means to access it. A common way of disseminating provenance data is through a derivation graph that users can browse and inspect [16, 18, 25, 26]. Users can also search for datasets based on their provenance metadata, such as to locate all datasets generated by a executing a certain workflow. If semantic provenance information is available, these query results can automatically feed input datasets for a workflow at runtime [25]. The derivation history of datasets can be used to replicate data at another site, or update it if a dataset is stale due to changes made to its ancestors [27]. Provenance retrieval APIs can additionally allow users to implement their own mechanism of usage.

3. Survey of Data Provenance Techniques

In our full survey of data provenance [9], we discuss nine major works that, taken together, provide a comprehensive overview of research in this field. In this paper, five works have been selected for discussion. A summary of their characteristics, as defined by the taxonomy, can be found in **Table 1**.

3.1 Chimera

Chimera [27] manages the derivation and analysis of data objects in collaboratory environments and collects provenance in the form of data derivation steps for datasets [17]. Provenance is used for on-demand regeneration of derived data ("virtual data"), comparison of data, and auditing data derivations.

Chimera uses a process oriented model to record provenance. Users construct workflows (called

derivation graphs or DAGs) using a Virtual Data Language (VDL) [17, 27]. The VDL conforms to a schema that represents data products as abstract typed *datasets* and their materialized *replicas*. *Datasets* can be files, tables, and objects of varying granularity, though the prototype supports only files. Computational process templates, called *transformations*, are scripts in the file system and, in future, web services [17]. The parameterized instance of the transformations, called *derivations*, can be connected to form workflows that consume and produce replicas. Upon execution, workflows automatically create *invocation* objects for each derivation in the workflow, annotated with runtime information of the process. Invocation objects are the glue that link input and output data products, and they constitute an annotation scheme for representing the provenance. Semantic information on the dataset derivation is not collected.

The lineage in Chimera is represented in VDL that maps to SQL queries in a relational database, accessed through a virtual data catalog (VDC) service [27]. Metadata can be stored in a single VDC, or distributed over multiple VDC repositories with inter-catalog references to data and processes, to enable scaling. Lineage information can be retrieved from the VDC using queries written in VDL that can, for example, recursively search for derivations that generated a particular dataset. A virtual data browser that uses the VDL queries to interactively access the catalog is proposed [27]. A novel use of provenance in Chimera is to plan and estimate the cost of regenerating datasets. When a dataset has been previously created and it needs to be regenerated (e.g. to create a new replica), its provenance guides the workflow planner in selecting an optimal plan for resource allocation [17, 27].

3.2 myGrid

myGrid provides middleware in support of *in silico* (computational laboratory) experiments in biology, modeled as workflows in a Grid environment [18]. myGrid services include resource discovery, workflow enactment, and metadata and provenance management, which enable integration and present a semantically enhanced information model for bio-informatics.

myGrid is service-oriented and executes workflows written in *XScufl* language using the *Taverna* engine [18]. A provenance log of the workflow enactment contains the services invoked, their parameters, the start and end times, the data products used and derived, and ontology descriptions, and it is automatically recorded when the workflow executes. This process-oriented workflow derivation log is inverted to infer the provenance for the intermediate and final data products. Users need to annotate workflows and services with semantic descriptions to enable this inference and have the semantic metadata carried over to the data products. In addition to contextual and organizational metadata such as owner, project, and experiment hypothesis, ontological terms can also be provided to describe the data and the experiment [8]. XML, HTML, and RDF are used to represent syntactic and semantic provenance metadata using the annotation scheme [14]. The granularity at which provenance can be stored is flexible and is any resource identifiable by an LSID [18].

The myGrid Information Repository (mIR) data service is a central repository built over a relational database to store metadata about experimental components [18]. A number of ways are available for knowledge discovery using provenance. For instance, the semantic provenance information available as RDF can be viewed as a labeled graph using the Haystack semantic web browser [18]. COHSE (Conceptual Open Hypermedia Services Environment), a semantic hyperlink utility, is another tool used to build a semantic web of provenance. Here, semantically annotated provenance logs are interlinked using an ontology reasoning service and displayed as a hyperlinked web page. Provenance information generated during the execution of a workflow can also trigger the rerun of another workflow whose input data parameters it may have updated.

3.3 CMCS

The CMCS project is an informatics toolkit for collaboration and metadata-based data management for multi-scale science [24, 25]. CMCS manages heterogeneous data flows and metadata across multi-disciplinary sciences such as combustion research, supplemented by provenance metadata for establishing the pedigree of data. CMCS uses the Scientific Annotation Middleware (SAM) repository for storing URL referenceable files and collections [25].

CMCS uses an annotation scheme to associate XML metadata properties with the files in SAM and manages them through a Distributed Authoring and Versioning (WebDAV) interface. Files form the level of granularity and all resources such as data objects, processes, web services, and bibliographic records are modeled as files. Dublin Core (DC) verbs like *Has Reference*, *Issued*, and *Is Version Of* are used as XML properties for data files and semantically relate them to their deriving processes through XLink references in SAM [24]. DC elements like *Title* and *Creator*, and user-defined metadata can provide additional context information. Heterogeneous metadata schemas are supported by mapping them to standard DC metadata terms using XSLT translators. Direct association of provenance metadata with the data object makes this a data-oriented model.

There is no facility for automated collection of lineage from a workflow's execution. Data files and their metadata are populated by DAV-aware applications in workflows or manually entered by scientists through a portal interface [25]. Provenance metadata properties

Table 1 Summary of characteristics of surveyed data provenance techniques

	Chimera	myGRID	CMCS	ESSW	Trio
Applied Domain	Physics, Astronomy	Biology	Chemical Sciences	Earth Sciences	None
Workflow Type	Script Based	Service Oriented	Service Oriented	Script Based	Database Query
Use of Provenance	Informational; Audit; Data Replication	Context Information; Re-enactment	Informational; update data	Informational	Informational; up date propagation
Subject	Process	Process	Data	Both	Data
Granularity	Abstract datasets (Presently files)	Abstract resources having LSID	Files	Files	Tuples in Database
Representation Scheme	Virtual Data Language Annotations	XML/RDF Annotations	DublinCore XML Annotations	XML/RDF Annotations	Query Inversion
Semantic Info.	No	Yes	Yes	Proposed	No
Storage Repository/ Backend	Virtual Data Catalog/ Relational DB	mIR repository/ Relational DB	SAM over DAV/ Relational DB	Lineage Server/ Relational DB	Relational DB
User Overhead	User defines derivations; Automated WF trace	User defines Service semantics; Automated WF Trace	Manual: Apps use DAV APIs; Users use portal	Use Libraries to generate provenance	Inverse queries automatically generated
Scalability Addressed	Yes	No	No	Proposed	No
Dissemination	Queries	Semantic browser; Lineage graph	Browser;Queries; GXL/RDF	Browser	SQL/TriQL Queries

can be queried from SAM using generic WebDAV clients. Special portlets allow users to traverse the provenance metadata for a resource as a web page with hyperlinks to related data, or as a labeled graph represented in the Graphics eXchange Language (GXL). The provenance information can also be exported to RDF that semantic agents can use to infer relationships between resources. Provenance metadata that indicate data modification can generate notifications that trigger workflow execution to update dependent data products.

3.4 ESSW

The Earth System Science Workbench (ESSW) [28] is a metadata management and data storage system for earth science researchers. Lineage is a key facet of the metadata created in the workbench, and is used for detecting errors in derived data products and in determining the quality of datasets.

ESSW uses a scripting model for data processing i.e. all data manipulation is done through scripts that wrap existing scientific applications [26]. The sequence of invocation of these scripts by a master workflow script forms a DAG. Data products at the granularity of files are consumed and produced by the scripts, with each data product and script having a uniquely labeled metadata object. As the workflow script invokes individual scripts, these scripts, as part of their execution, compose XML metadata for themselves and the data products they generate. The workflow script links the data flow between successive scripts using their metadata ids to form the lineage trace for all data products, represented as annotations. By chaining the scripts and the data using parent-child links, ESSW is balanced between data and process oriented lineage.

ESSW puts the onus on the script writer to record the metadata and lineage using templates and libraries that are provided. The libraries store metadata objects as files in a web accessible location and the lineage separately in a relational database [26]. Scalability is not currently addressed though it is proposed to federate lineage across organizations. The metadata and lineage information can be navigated as a workflow DAG through a web browser that uses PHP scripts to access the lineage database [28]. Future work includes encoding lineage information semantically as RDF triples to help answer richer queries [26].

3.5 Trio

Cui and Widom [22, 29] trace lineage information for view data in data warehouses. The Trio project [23] leverages some of this work in a proposed database system which has data accuracy and data lineage as inherent components. While data warehouse mining and updation motivates lineage tracking in this project, any e-science system that uses database queries and functions to model workflows and data transformations can apply such techniques.

A database view can be modeled as a query tree that is evaluated bottom-up, starting with leaf operators having tables as inputs and successive parent operators taking as input the result of a child operator [22]. For ASPJ (Aggregate-Select-Project-Join operator) views, it is possible to create an inverse query of the view query that operates on the materialized view, and recursively moves down the query tree to identity the source tables in the leaves that form the view data's lineage [22].

Trio [23] uses this inversion model to automatically determine the source data for tuples created by view

queries. The inverse queries are recorded at the granularity of a view tuple and stored in a special *Lineage* table. This direct association of lineage with tuples makes this a data-oriented provenance scheme. Mechanisms to handle (non-view) tuples created by insert and update queries, and through user-defined functions are yet to be determined. Lineage in Trio is simply the source tuples and the view query that created the view tuple, with no semantic metadata recorded. Scalability is not specifically addressed either. Other than querying the *Lineage* table, some special purpose constructs will be provided for retrieving lineage information through a Trio Query Language (TriQL).

4. Conclusion

In this paper, we presented a taxonomy to understand and compare provenance techniques used in e-science projects. The exercise shows that provenance is still an exploratory field and several open research questions are exposed. Ways to federate provenance information and assert its truthfulness need study for it to be usable across organizations [12]. Evolution of metadata and service interface standards to manage provenance in diverse domains will also contribute to a wider adoption of provenance and promote its sharing [11]. The ability to seamlessly represent provenance of data derived from both workflows and databases can help in its portability. Ways to store provenance about missing or deleted data (phantom lineage [23]) require further consideration. Finally, a deeper understanding of provenance is needed to identify novel ways to leverage it to its full potential.

5. References

[1] J. Brase, "Using Digital Library Techniques - Registration of Scientific Primary Data," in *ECDL*, 2004.

[2] D. G. Clarke and D. M. Clark, "Lineage," in *Elements of Spatial Data Quality*, 1995.

[3] J. L. Romeu, "Data Quality and Pedigree," in *Material Ease*, 1999.

[4] H. V. Jagadish and F. Olken, "Database Management for Life Sciences Research," in *SIGMOD Record*, vol. 33, 2004.

[5] "Access to genetic resources and Benefit-Sharing (ABS) Program," United Nations University, 2003.

[6] P. Buneman, S. Khanna, and W. C. Tan, "Why and Where: A Characterization of Data Provenance," in *ICDT*, 2001.

[7] D. P. Lanter, "Design of a Lineage-Based Meta-Data Base for GIS," in *Cartography and Geographic Information Systems*, vol. 18, 1991.

[8] M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn, "Provenance of e-Science Experiments - experience from Bioinformatics," in *Proceedings of the UK OST e-Science 2nd AHM*, 2003.

[9] Y. L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance Techniques," in *Technical Report TR-618*: Computer Science Department, Indiana University, 2005.

[10] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: a survey," in *ACM Comput. Surv.*, vol. 37, 2005.

[11] S. Miles, P. Groth, M. Branco, and L. Moreau, "The requirements of recording and using provenance in e-Science experiments," in *Technical Report, Electronics and Computer Science, University of Southampton*, 2005.

[12] D. Pearson, "Presentation on Grid Data Requirements Scoping Metadata & Provenance," in *Workshop on Data Derivation and Provenance, Chicago*, 2002.

[13] G. Cameron, "Provenance and Pragmatics," in *Workshop on Data Provenance and Annotation, Edinburgh*, 2003.

[14] C. Goble, "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics," in *Workshop on Data Derivation and Provenance, Chicago*, 2002.

[15] P. P. da Silva, D. L. McGuinness, and R. McCool, "Knowledge Provenance Infrastructure," in *IEEE Data Engineering Bulletin*, vol. 26, 2003.

[16] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita, "Improving Data Cleaning Quality Using a Data Lineage Facility," in *DMDW*, 2001.

[17] I. T. Foster, J. S. Vöckler, M. Wilde, and Y. Zhao, "The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration," in *CIDR*, 2003.

[18] J. Zhao, C. A. Goble, R. Stevens, and S. Bechhofer, "Semantically Linking and Browsing Provenance Logs for E-science," in *ICSNW*, 2004.

[19] A. Woodruff and M. Stonebraker, "Supporting Fine-grained Data Lineage in a Database Visualization Environment," in *ICDE*, 1997.

[20] B. Plale, D. Gannon, D. Reed, S. Graves, K. Droegemeier, B. Wilhelmson, and M. Ramamurthy, "Towards Dynamically Adaptive Weather Analysis and Forecasting in LEAD," in *ICCS workshop on Dynamic Data Driven Applications*, 2005.

[21] D. Bhagwat, L. Chiticariu, W. C. Tan, and G. Vijayvargiya, "An Annotation Management System for Relational Databases," in *VLDB*, 2004.

[22] Y. Cui and J. Widom, "Practical Lineage Tracing in Data Warehouses," in *ICDE*, 2000.

[23] J. Widom, "Trio: A System for Integrated Management of Data, Accuracy, and Lineage," in *CIDR*, 2005.

[24] C. Pancerella, J. Hewson, W. Koegler, D. Leahy, M. Lee, L. Rahn, C. Yang, J. D. Myers, B. Didier, R. McCoy, K. Schuchardt, E. Stephan, T. Windus, K. Amin, S. Bittner, C. Lansing, M. Minkoff, S. Nijsure, G. v. Laszewski, R. Pinzon, B. Ruscic, Al Wagner, B. Wang, W. Pitz, Y. L. Ho, D. Montoya, L. Xu, T. C. Allison, W. H. Green, Jr, and M. Frenklach, "Metadata in the collaboratory for multi-scale chemical science," in *Dublin Core Conference*, 2003.

[25] J. Myers, C. Pancerella, C. Lansing, K. Schuchardt, and B. Didier, "Multi-Scale Science, Supporting Emerging Practice with Semantically Derived Provenance," in *ISWC workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, 2003.

[26] R. Bose and J. Frew, "Composing Lineage Metadata with XML for Custom Satellite-Derived Data Products," in *SSDBM*, 2004.

[27] I. T. Foster, J.-S. Vöckler, M. Wilde, and Y. Zhao, "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation," in *SSDBM*, 2002.

[28] J. Frew and R. Bose, "Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products," in *SSDBM*, 2001.

[29] Y. Cui and J. Widom, "Lineage tracing for general data warehouse transformations," in *VLDB Journal*, vol. 12, 2003.