

RESEARCH STATEMENT

Satya Sanket Sahoo

The phenomenal growth in computing capabilities is transforming scientific research from being an experiment-driven discipline to a “data-driven” science. Scientists are harnessing distributed computing resources and high-throughput equipment to generate petabytes of scientific data and metadata. The primary goal of my research is to develop the theoretical foundation and practical tools to model, integrate and analyze this deluge of data and metadata (Figure 1).

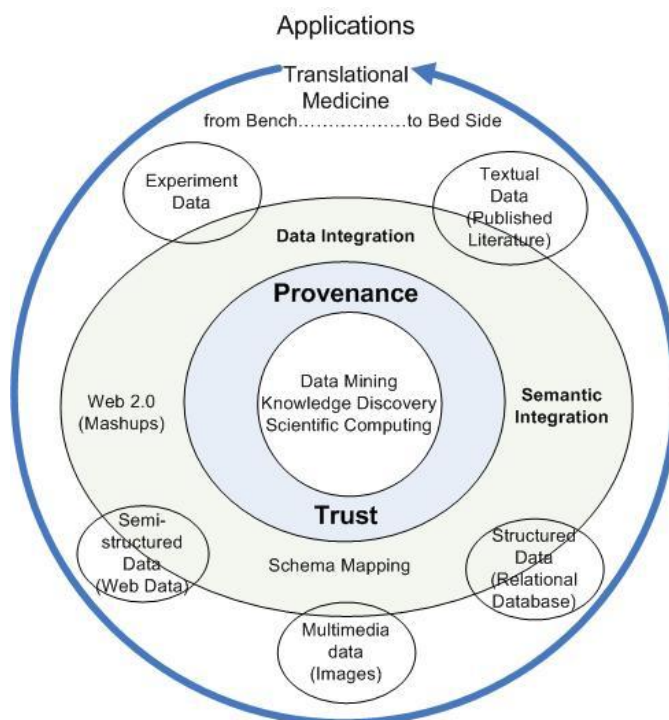


Figure 1: Research Overview

Data driven science is enabled by a variety of resources, such as “raw” experiment data, structured data, textual data, and computing resources including Web services. These resources have disparate levels of quality and trust; hence the metadata to describe the origin or lineage of each resource is critical to a range of computer science sub-disciplines such as data mining and knowledge discovery, and data integration. This category of metadata describing the history or lineage of an entity is called *provenance*. In addition to their central role in computer science, provenance and trust are also the key enablers for realizing broader objectives such as the National Institutes of Health (NIH) “Translational Research” roadmap that seeks to improve human health by translating “scientific discoveries” (*bench*) to “practical applications” (*bedside*).

Information provenance is a hard problem in computing science with many fundamental issues that are yet to be addressed. Further, there has been little or no work in understanding and leveraging the role of provenance as well as trust in data integration and data management. My research demonstrates the synergistic use of artificial intelligence, Semantic Web, and database theory to address three aspects of provenance and trust management namely, (a) **Modeling**, (b) **Querying**, and (c) **Scalable Implementations**, that have significant impact on the integration, analysis and management of data in inter-disciplinary science such as bioinformatics and computational biology.

PROVENANCE MANAGEMENT FRAMEWORK

Provenance tracking in scientific applications has traditionally involved hand written notes, ad hoc scripts or programs that captured only parts of the complete provenance information. This has created a significant challenge to ensure quality and trust in data management applications. For example, in high-throughput proteomics, researchers cannot accurately analyze mass spectrometry (ms) results without using provenance details to identify the enzyme used in digesting a protein sample or the parameter settings of the ms instrument. It is easy to extend this example scenario to many related research areas (for example, sensor networks and Web applications) and realize that without a provenance and trust layer the full potential of data driven science will not be realized.

In recent years, there has been a rapidly growing interest in defining the theory and designing practical provenance systems in science. In my Ph.D. thesis, I have investigated the use of description logic, context theory, and query optimization techniques to define a “provenance management framework”. This section outlines the components of this framework namely, (a) a foundational provenance model, (b) a provenance-specific query infrastructure, and (c) a scalable provenance query engine [1] [2]. This framework has enabled successful provenance and data management in the NIH-funded Semantic Problem Solving Environment for *Trypanosoma cruzi* (*T.cruzi* SPSE) project [3].

➤ Existing systems for scientific provenance - ad hoc and not useful in real world

➤ Defined a novel provenance management framework

➤ Provenance framework: (a) foundational ontology, (b) query infrastructure, (c) scalable query engine

Provenir ontology: Towards a common model of provenance

A primary challenge for provenance research is defining a consistent modeling paradigm that reduces terminological heterogeneity, incorporates *domain semantics*, and facilitates provenance interoperability. Ontologies defined using knowledge representation languages are a suitable modeling choice and ontologies also support reasoning-based knowledge discovery. But, provenance varies across different domains and creation of a monolithic provenance ontology is not a feasible solution. To address this, I proposed a *modular approach* to provenance modeling that uses a foundational provenance ontology as the core component that can be extended to create a suite of interoperable domain-specific provenance ontologies.

As the first step to realize this approach, I created an upper-level ontology called Provenir (provenance is derived from the French word “provenir”) that extends primitive philosophical ontology concepts and links them using a set of foundational relations [4]. The Provenir ontology has been described as “significant” model by the Open Biomedical Ontologies (OBO), where it is currently under review for listing as a reference model for scientific provenance.

In the next step, I have led the creation of three domain-specific provenance ontologies by extending the Provenir ontology. In collaboration with biomedical researchers, Brent Weatherly (University of Georgia) and Dr. Flora Logan (The Wellcome Trust Sanger Institute, Cambridge, UK), I have created the *Parasite Experiment ontology* to model provenance of gene knockout and strain creation experiments [2]. Further, in collaboration with Dr. William York, at the University of Georgia Complex Carbohydrate Research Center (CCRC), I developed the *ProPreO ontology* (WWW’06)[5] to model provenance in proteomics experiments. These two ontologies have been released for public use through the Stanford University National Center for Biomedical Ontologies (NCBO). I created the third ontology, called *Trident*, during my research internship at Microsoft Research, Redmond in 2008, to model provenance in the oceanography domain.

➤ Primary challenge in provenance management: (a) consistent modeling, (b) use of domain semantics

➤ Innovation: A modular and flexible approach to provenance modeling
➤ Created foundational Provenir ontology

➤ Collaborations: (a) University of Georgia, (b) Microsoft Research, (c) Wellcome Trust Sanger Institute

Provenance Analytics: Query classification and query operators

In addition to modeling, real world applications require a well-defined mechanism to query and analyze provenance information. The provenance literature feature a large variety of provenance queries, but there has been no exhaustive study of these queries and most previous work used ad-hoc or generic query mechanisms. I have defined a novel classification scheme for provenance queries with three broad categories; to our knowledge this is the first work to systematically categorize provenance queries. Further, I used this classification scheme to define a set of specialized query operators for provenance.

The four primary provenance query operators are, (a) *provenance*(): to retrieve the provenance of an entity, (b) *provenance_context*(): to retrieve data entities using a “context structure” defined over provenance, (c) *provenance_compare*() and (d) *provenance_merge*(): to compare and merge provenance information respectively. I used the Provenir ontology to define the formal and functional semantics of the query operators; hence they can be used seamlessly with any domain-specific provenance ontologies that extend the Provenir ontology.

In addition to supporting provenance queries, the *provenance_context*() query operator makes a unique contribution to data integration. Traditionally, data integration has dealt with three types of heterogeneities (structural, syntactic, semantic); in the data driven science, I have identified a fourth source of heterogeneity called *trust heterogeneity*. The *provenance_context*() operator uses the query input to dynamically construct a context structure, a form of “problem solving context” for query answering, called “provenance context” and uses it to identify data entities with similar trust values [4]. In ongoing research, I am investigating the cost and performance of the query operator to simultaneously reconcile all four types of data heterogeneity.

- Defined novel classification scheme for provenance queries
- Used query classification scheme to define the first set of specialized query operators for provenance
- Identified new *trust heterogeneity* in data integration
- Used *provenance_context*() query operator to reconcile trust heterogeneity

Materialized Provenance Views: Optimizing Provenance Queries

To demonstrate the practical use of the provenance query operators, I implemented a provenance query engine over a Resource Description Framework (RDF) data store. Provenance queries are computationally expensive graph-based path operations for fixed paths, recursive pattern-based paths and neighborhood retrieval. The evaluation results of a straight-forward implementation of the query engine showed that provenance queries over very large real world datasets (~308 million RDF triples) took six days to complete!

To address this issue, I defined a new class of materialized views using the Provenir ontology called *materialized provenance views* (MPV) that reduced the time for query execution by three orders of magnitude to a few minutes [2]. The MPV is a form of cost-based query optimization technique that does not require query rewriting and guarantees logical correctness of the query results. Further, unlike common materialized views, MPV is a single *logical* unit of provenance that is computed using a domain-specific provenance ontology.

- Straight forward implementation of provenance query engine not scalable - complexity of queries, large size of data
- New class of materialized views using Provenir ontology – *Materialized Provenance View* using domain semantics
- Reduced query time by three orders of magnitude, evaluation over real world dataset - 308 million RDF triples

SEMANTIC WEB and DATA INTEGRATION

During my research internships at the Lister Hill National Center for Biomedical Communications (NLM/NIH) in 2006 and 2007, I collaborated with Dr. Olivier Bodenreider towards the creation of a Biomedical Knowledge Repository (BKR) [6]. To achieve this, I defined a multi-ontology environment to integrate heterogeneous and distributed resources using Semantic Web standards such as RDF, Web Ontology Language (OWL), and a novel use of reasoning rules to reconcile ABox semantic heterogeneity [7].

In 2006, we converted the complete NCBI Entrez gene data, with information on two million genes, to RDF with about 400 million RDF triples. We integrated a section of this gene data with Gene Ontology and demonstrated the effectiveness of RDF as a semantic integration platform by answering a real-world biomedical query linking a specific molecular function, *glycosyltransferase*, to the *congenital muscular dystrophy* disorder [8].

Extending our work in 2007, we collaborated with researchers at the National Institute on Drug Abuse (NIDA/NIH) to identify links between a set of genes, identified in a study funded by NIDA, and biological pathways in context of nicotine dependence. Using two OWL ontologies, integrated at schema-level using named relations, data from five sources (NCBI Entrez Gene, HomoloGene, KEGG, BioCyc, and Reactome) were integrated and represented in RDF [7]. This knowledgebase supports complex biological queries including the identification of *hub genes*, common pathways across species and genes expressed in brain tissue [7].

- Created a multi-ontology environment for semantic integration of heterogeneous and distributed biomedical data
- Supported complex real world queries that involved reasoning based knowledge discovery
- Collaborations with the National Institute on Drug Abuse (NIDA/NIH) and National Library of Medicine (NLM/NIH)

GRANT WRITING, COMMUNITY ENGAGEMENT, and RESEARCH IMPACT

During my doctoral study, I had the opportunity to participate in writing multiple grant proposals to both the National Science Foundation (NSF) and the NIH. I played an integral role in the formulation and writing of the *T.cruzi* SPSE grant proposal that was primarily based on my research work and involved collaboration with Stanford University and the University of Georgia. This proposal was funded in 2008 by the National Heart, Lung, and Blood Institute (NHLBI/NIH) as a RO1 project (total funding \$1.5 million) [3].

I believe in active community engagement that allows me to learn from as well as influence ongoing research initiatives. I led the organization of a full day workshop on role of Semantic Web in Provenance Management at the International Semantic Web Conference (ISWC) 2009 in collaboration with Dr. Juliana Freire (University of Utah) and Dr. Paolo Missier (University of Manchester). I am also an active member of the W3C Semantic Web Health Care and Life Science Interest Group (BioRDF task force), W3C RDB2RDF Incubator Group – I led a survey to document the current state of the art in approaches to map relational data to RDF (published as a W3C technical report), and the recently announced W3C provenance incubator group. I have participated in the creation of a XML-based data exchange standard for glycomics called GLYDE, which is now accepted and used by multiple institutions such as EurocarbDB, Kyoto Encyclopedia for Genes and Genome (KEGG), and the Consortium for Functional Glycomics.

As a measure of research impact, my paper outlining the provenance framework [1] is listed as course material for both the “Advanced Semantic Web” course (CSCI 6965) at the Rensselaer Polytechnic Institute (RPI) and for the “Database Systems for Scientific Applications” course (CSC 875) at the San Francisco State University. A second paper describing our multi-ontology data integration work [7] is part of the course material for the “Semantic Web in Biomedicine” course (MEBI591C) at the University of Washington Seattle.

FUTURE RESEARCH AGENDA

The pace of data generation and consumption in science has continued to accelerate and is expected to do so for the foreseeable future. This coupled with the increasing use of the ubiquitous World Wide Web is a unique opportunity for the research community to conduct transformational science, a goal overlapping with the NIH translational research and NSF cyberinfrastructure initiatives. My future research is geared to create an enabling computing environment with provenance and trust as integral layers to realize this vision. Through extensive collaborations with biology and biomedical researchers to solve real world problems, I have come to realize that scientific data integration and analysis are still open research problems. Further, the added complexity of provenance and trust in these applications requires innovative solutions that bring together traditional data analysis paradigms such as Online Analytical Processing (OLAP) with new initiatives such as the Semantic Web

and cloud computing. In this section, I weave together these themes to describe some of my near and medium term research objectives.

Provenance cube

Provenance is inherently multi-dimensional and is queried or analyzed from multiple aspects (for example, spatial, temporal or thematic). I am extending the graph OLAP framework [9] to define “Provenance Cube”, a coherent structure to support analytical provenance queries in high-dimensional space. Practical development of the provenance cube requires addressing a variety of issues including a new set of OLAP dimensions, and measures, defining the semantics of OLAP operations over provenance cube, and developing an efficient implementation.

Provenance in the Semantic Web layer cake

The fundamental aspect of my provenance framework is the modeling of both data and provenance as first class entities. This is in contrast to the traditional approach of modeling provenance separately in the Semantic Web through use of RDF reification and named graphs. The advantage of my approach is the ability to query both provenance and data together while avoiding the drawbacks of current approaches. I am collaborating with Dr. Olivier Bodenreider at NLM/NIH to extend my provenance framework for tracking of provenance in RDF triples extracted from biomedical literature (for example, PubMed) without using reification.

Biomedical Problem Solving Environment in the cloud

The new cloud computing paradigm combines the characteristics of distributed and parallel computing and offers computing infrastructure as a service. Leveraging the “pay as you go” pricing mechanism, I plan to use cloud computing as a cost effective platform to create a Biomedical Problem Solving Environment featuring a single, interface to access a vast range of integrated data and computational resources along with detailed provenance information. I believe such a resource would enable even small research groups in biology to take advantage of the state-of-the-art (and till now prohibitively expensive) computer science resources for their daily research.

References

- [1] S. S. Sahoo, Sheth, A., Henson, C., "Semantic Provenance for eScience: Managing the Deluge of Scientific Data," *IEEE Internet Computing*, vol. 12, pp. 46-54, 2008.
- [2] S. S. Sahoo, Weatherly, D.B., Muttharaju, R., Anantharam, P., Sheth, A., Tarleton, R.L., "Ontology-driven Provenance Management in eScience: An Application in Parasite Research," in *The 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 09)*, Vilamoura, Algarve-Portugal, 2009, pp. 992-1009.
- [3] "Semantics and Services enabled Problem Solving Environment for Teruzzi", <http://knoesis.wright.edu/trykipedia>
- [4] S. S. Sahoo, Barga, R.S., Goldstein, J., Sheth, A., "Provenance Algebra and Materialized View-based Provenance Management," Microsoft Research Technical Report, November 2008.
- [5] S. S. Sahoo, Thomas, C., Sheth, A., York, W. S., and Tartir, S., "Knowledge modeling and its application in life sciences: a tale of two ontologies," in *Proceedings of the 15th international Conference on World Wide Web (WWW)* Edinburgh, Scotland, 2006, pp. 317-326.
- [6] O. Bodenreider, Rindflesch, T.C., "Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications," LHCBC Communications, NLM, Bethesda, 2006.
- [7] S. S. Sahoo, Bodenreider, O., Rutter, J.L. Skinner, K.J., Sheth, A.P., "An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence," *Journal of Biomedical Informatics*, vol. 41, October 2008, pp. 752-765.
- [8] S. S. Sahoo, Zeng, K., Bodenreider, O., Sheth, A.P., "From "glycosyltransferase" to "congenital muscular dystrophy": Integrating knowledge from NCBI Entrez Gene and the Gene Ontology.," in *Medinfo*, Brisbane Australia, 2007, pp. 1260-1264.
- [9] C. Chen, Yan, X., Zhu, F., Han, J., Yu, P. S., "Graph OLAP: Towards online analytical processing on graphs.," in *ICDM*, 2008, pp. 103–112.