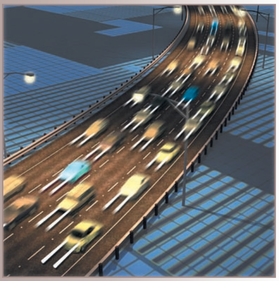
2008

Semantic Provenance for eScience: Managing the Deluge of Scientific Data

Satya S. Sahoo
Wright State University - Main Campus

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Cory Andrew Henson
Wright State University - Main Campus



Editors: M. Brian Blake • mb7@georgetown.edu
Michael N. Huhns • huhns@sc.edu

Semantic Provenance for eScience

Managing the Deluge of Scientific Data

Provenance information in eScience is metadata that's critical to effectively manage the exponentially increasing volumes of scientific data from industrial-scale experiment protocols. Semantic provenance, based on domain-specific provenance ontologies, lets software applications unambiguously interpret data in the correct context. The semantic provenance framework for eScience data comprises expressive provenance information and domain-specific provenance ontologies and applies this information to data management. The authors' "two degrees of separation" approach advocates the creation of high-quality provenance information using specialized services. In contrast to workflow engines generating provenance information as a core functionality, the specialized provenance services are integrated into a scientific workflow on demand. This article describes an implementation of the semantic provenance framework for glycoproteomics.

Satya S. Sahoo,
Amit Sheth,
and Cory Henson
Kno.e.sis Center,
Wright State University

eScience, also known as cyberinfrastructure, represents a paradigm shift in scientific research that lets scientists harness Web-based computing and data resources to achieve their objectives faster, more efficiently, and on an industrial scale. Using remote software and experimental equipment, scientists can not only access but also generate and process data from distributed sources. The resulting data deluge demands computing solutions that can use high-quality metadata — specifically, domain-specific provenance information — to automatically interpret, integrate, and process data. Such solutions bring real value to scientists by answering domain-specific queries effectively to support knowledge discovery over large volumes of scientific data. But creating provenance information of the requisite quality in the heterogeneous, distributed, and high-throughput environment of eScience is a daunting challenge.

We argue that incorporating domain knowledge and ontological underpinning in provenance using expressive domain-specific provenance ontologies is an approach equal to the challenge. This *semantic provenance* imposes a

1089-7801/08/$25.00 © 2008 IEEE

formally defined domain-specific conceptual view on scientific data (domain semantics), mitigates or eliminates terminological heterogeneity, and enables the use of reasoning tools for knowledge discovery. Furthermore, we define a "two degrees of separation" approach for creating semantic provenance using specialized software tools. Unlike many prevalent workflow-engine-centric approaches, these tools refer to domain-specific provenance ontologies to create provenance information and are integrated into a scientific workflow on demand.

We combine the essential aspects of high-quality provenance — characteristics, a representation model, the creation process, and usage — into a single semantic provenance framework. This framework will pave the way for software agents to interpret experimental data unambiguously for effective management of eScience data. We also describe an implementation of this framework — Spade (*s*emantic *p*rovenance *a*nnotation of *d*ata in prot*e*omics).

"Meaningful" Provenance for eScience

The available worldwide infrastructure of computing and data resources of eScience let scientists collaborate in virtual laboratories.[1,2] Examples of such large-scale eScience projects include the Biomedical Informatics Research Network (www.nbirn.net), myGrid (www.mygrid.org.uk), and TeraGrid (www.teragrid.org). The exponential increase in the scale and complexity of experiments made possible by this infrastructure has resulted in a corresponding increase in the amount of scientific data generated; see, for example, https://cabig.nci.nih.gov/inventory/inventory/data_resources and www.nbirn.net/bdr/index.shtm.

Figure 1 illustrates a high-throughput scientific workflow for processing and analyzing proteomics data that generates hundreds of files per sample run (described later in detail). The rapidly increasing volume of data raises important issues such as

- How can we leverage the data for critical insights that will in turn drive future research?
- How can we seamlessly manage (compare, integrate, and process) large volumes of data generated by hundreds of distributed laboratories using heterogeneous starting materials, equipment, protocols, and parameters?

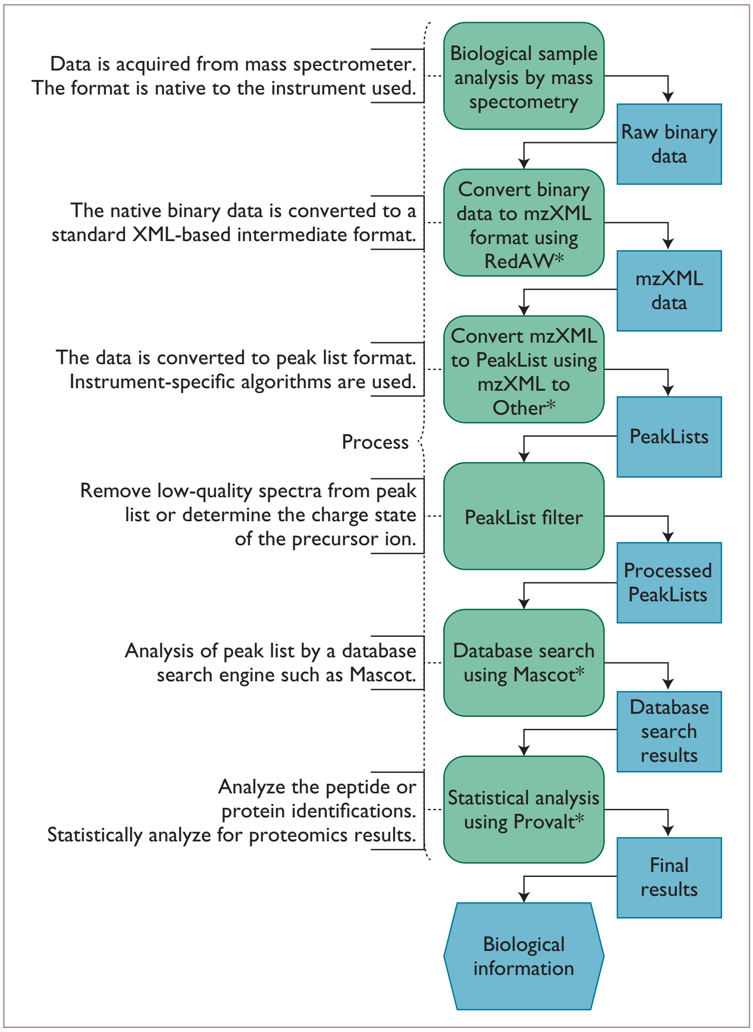

Data is acquired from mass spectrometer. The format is native to the instrument used.
Biological sample analysis by mass spectometry
Raw binary data
The native binary data is converted to a standard XML-based intermediate format.
Convert binary data to mzXML format using RedAW*
mzXML data
The data is converted to peak list format. Instrument-specific algorithms are used.
Convert mzXML to PeakList using mzXML to Other*
PeakLists
Process
Remove low-quality spectra from peak list or determine the charge state of the precursor ion.
PeakList filter
Processed PeakLists
Analysis of peak list by a database search engine such as Mascot.
Database search using Mascot*
Database search results
Analyze the peptide or protein identifications. Statistically analyze for proteomics results.
Statistical analysis using Provalt*
Final results
Biological information

Figure 1. Protocol for proteomics data analysis using a mass spectrometer. This high-throughput scientific workflow for processing and analyzing proteomics generates hundreds of files per sample run. (An asterisk indicates a third-party tool.)

It's precisely these issues that we'll address through the use of metadata — specifically, semantic provenance information.

Metadata and Provenance

Metadata's critical role in managing large volumes of data has long been understood in library management (www.loc.gov/standards), geography (www.opengeospatial.org/standards/gml), multimedia,[3] and the biological sciences.[4] The database community has extensively explored the use of metadata to exchange, share, and integrate data from heterogeneous information sources.[3] Because traditional metadata descriptions (such as electronic data-interchange formats) require manual interpretation, researchers have proposed using semantic meta-

Related Projects in Provenance

The myGrid project Pedro[1] was one of the earliest initiatives to create a process model (in UML) to capture domain details about a proteomics analysis protocol. The myGrid project has also identified provenance information as a platform for knowledge management in eScience.[2]

The Stanford Knowledge Provenance Infrastructure (KPI)[3] is an example of provenance architecture focused on providing computable provenance information related to Web data such as news feeds for use by both agents and humans. KPI's primary objective is to collect and provide the explanation associated with a piece of information,[4] which includes the source of data and any reasoning or inference processes applied to the data. The project doesn't use provenance information for data management, which is the focus of our article.

W.C. Tan discussed the classification of provenance information as fine-grained data provenance and coarse-grained workflow provenance.[5] Our definition of semantic provenance incorporates characteristics of both coarse- and fine-grained categories of provenance information.

References

1. P.N. Taylor et al., "A Systematic Approach to Modeling, Capturing, and Disseminating Proteomics Experimental Data," *Nature Biotechnology*, Mar. 2003, pp. 247–254.
2. R. Stevens, J. Zhao, and C. Goble, "Using Provenance to Manage Knowledge of *In Silico* Experiments," *Briefings in Bioinformatics*, vol. 8, no. 3, 2007, pp. 183–194.
3. P. Pinheiro da Silva, D.L. McGuinness, and R. McCool, "Knowledge Provenance Infrastructure," *IEEE Data Eng. Bulletin*, vol. 26, no. 4, 2003, pp. 26–32.
4. D.L. McGuinness and P. Pinheiro da Silva, "Explaining Answers from the Semantic Web: The Inference Web Approach," *J. Web Semantics*, vol. 1, no. 4, 2004, pp. 397–413.
5. W.C. Tan, "Provenance in Databases: Past, Current, and Future," *IEEE Data Eng. Bull.*, vol. 30, no. 4, pp. 3–12.

data to automate the integration of large-scale distributed data.

Semantic metadata is "metadata that describes contextually relevant or domain-specific information about content (optionally) based on a[n]... ontology."[5] It not only mitigates terminological heterogeneity but also enables software applications to "understand" and reason over it. Specific motivating factors for using semantic metadata are

- to create a conceptual context to "capture domain knowledge and help impose a conceptual semantic view on the underlying data"[3] for accurate data interpretation; and
- to support interoperability: semantic metadata is effectively an instance of ontology concepts or relationships, so the outcome from extensive research in ontology mapping and merging will enable easy integration of semantic metadata that subscribes to different ontologies.

Metadata in the form of provenance information records the *how*, *where*, *what*, *when*, *why*, *which*, and *by whom*[6] of data generated in a scientific experiment. Scientists can manually record provenance information, or software tools can automatically generate it. Scientists traditionally use provenance information, along with (implicit) domain expertise, to interpret and evaluate data accurately. Provenance information also lets researchers verify and validate experimental procedures (see the "Related Projects in Provenance" sidebar).

In the eScience informatics community, sustained research in provenance has led to many models for creating, representing, storing, and querying provenance (see http://twiki.ipaw.info).[7] Most current eScience approaches to provenance creation center on a "workflow-engine perspective of the world." So, the *operations* (in the form of Web services or scripts) orchestrated by the workflow engine are the principle actors in the resulting provenance descriptions, along with information about the input and output files. This approach not only ignores the multiple domain-specific relationships that link the data, processes, and equipment but also imposes a system-level view on what is essentially a scientific procedure. We term this category of provenance information *system provenance*, also sometimes called *workflow provenance*.[8]

Semantic Provenance

To be used effectively for managing large and growing volumes of data in eScience, provenance information must be

- *Software-interpretable*: Human mediation is inadequate to process, analyze, integrate, store, and query the petabytes of data and associated metadata generated by the industrial-scale processes in eScience. For software agents to be able to use metadata – specifically, provenance information – to manage eScience data, they must be able to "compute" over it.[6]
- *Expressive*: Provenance information should be expressive enough to incorporate domain semantics of the data that will enable software agents to use the provenance informa-

tion to accurately interpret eScience data in the correct context.

To achieve these two objectives, we extend the notion of provenance information and combine it with two important attributes of semantic metadata — domain knowledge and ontological underpinning (see Figure 2). We thus define semantic provenance as "information created with reference to a formal knowledge model or an ontology that imposes a domain-specific provenance view on scientific data. It consists of formally defined concepts linked together using named relationships with a set of rules to represent domain constraints."

We illustrate the distinction between system provenance and semantic provenance using two types of queries. The first type is answered using system provenance; for example, "Find the original data from which result data X was derived." This query uses the workflow-centric provenance information that documents the invocation order of processes, the input data, and the output data for each process. So, using the links connecting a process's output data to its input data, a provenance-aware system could trace and identify the original data entity for result data X. Scientists typically use queries in this category to investigate the protocol that generated the data and to rerun a scientific workflow if needed for validation.

The second type of query is answered using semantic provenance. Queries in this category are complex and involve relationships that tie data, processes, and equipment parameters together using a domain-specific conceptual view. An example from the proteomics domain is, "Find proteins composed of peptides with N-glycosylation consensus sequence {*N[^P][S/T]*} identified in samples labeled with O18."

This query uses relationships between data entities that aren't modeled in a workflow view of provenance information such as "a peptide is derived from a protein" and "proteins are identified from a particular sample." Furthermore, the query constrains the samples (introduced in detection equipment such as a mass spectrometer) to be labeled with O18 (an isotope of oxygen), which is again a domain-specific relationship.

Note that in addition to incorporating domain-specific details, semantic provenance can also answer the first type of queries discussed.

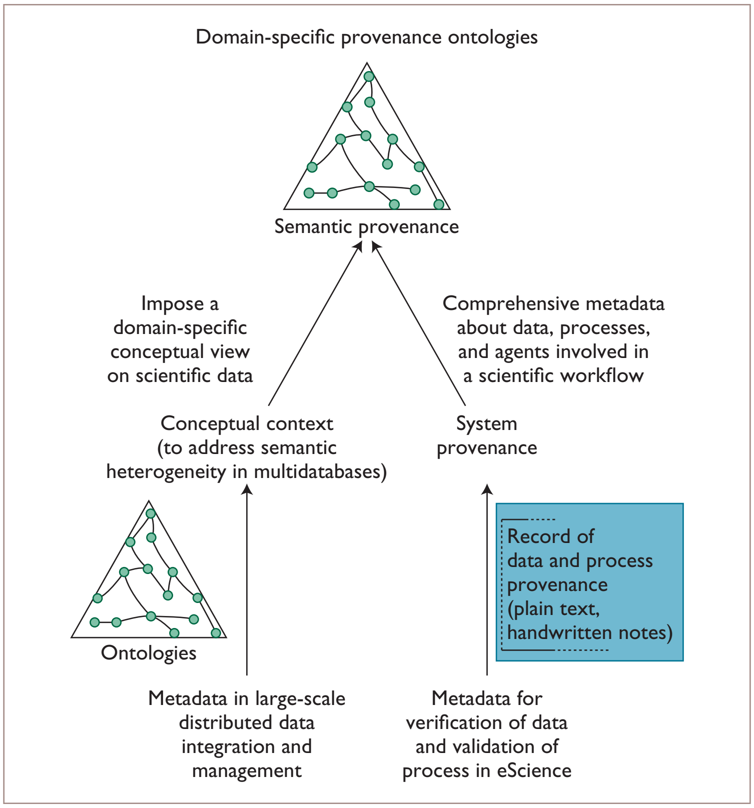

Domain-specific provenance ontologies
Semantic provenance
Impose a domain-specific conceptual view on scientific data
Comprehensive metadata about data, processes, and agents involved in a scientific workflow
Conceptual context (to address semantic heterogeneity in multidatabases)
System provenance
Ontologies
Record of data and process provenance (plain text, handwritten notes)
Metadata in large-scale distributed data integration and management
Metadata for verification of data and validation of process in eScience

Figure 2. The evolution of semantic provenance. This evolution can be traced to metadata's role in both integrating data in distributed environments and in verifying data and validating processing in eScience.

We thus define semantic provenance (*SemPro*) to be a superset of system provenance (*SysPro*): SemPro $\supset$ SysPro. Given the distinct limitations of the workflow-engine-centric view of provenance, we argue for a loosely coupled infrastructure for provenance creation using specialized services.

Two Degrees of Separation

Either the workflow engine or specialized annotation services can create provenance information.[9] As we've discussed, provenance created with a workflow-engine-centric approach can't answer queries that require use of domain semantics easily — if at all. Many teams participating in the Second International Provenance Challenge customized their provenance-collection systems to answer the challenge queries using ad hoc terms such as "Warp Params 2" to denote provenance information (see http://twiki.gridprovenance.org/bin/view/Challenge/WebHome).

We need a new strategy that decouples the task of generating high-quality semantic provenance from the core functionality of workflow

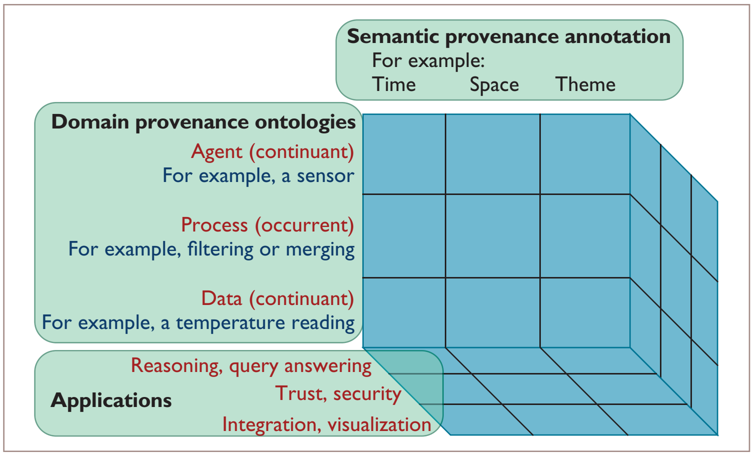

Semantic provenance annotation
For example:
Time
Space
Theme
Domain provenance ontologies
Agent (continuant)
For example, a sensor
Process (occurrent)
For example, filtering or merging
Data (continuant)
For example, a temperature reading
Applications
Reasoning, query answering
Trust, security
Integration, visualization

Figure 3. The three dimensions of the semantic provenance framework. One dimension represents semantic provenance annotations, another represents domain provenance ontologies, and the third dimension describes the different categories of usage.

engines. The task of semantic provenance creation should be managed by specialized services that refer to one or more domain-specific provenance ontologies and can be integrated into scientific workflows on demand. Then, a workflow engine, instead of providing native support for provenance creation, would feature a set of services and a suite of domain-specific provenance ontologies as resources that could be flexibly incorporated into a scientific workflow according to user needs. This service-oriented architecture (SOA) represents a scalable, adaptable, and workflow-engine-agnostic solution for eScience provenance. We term this approach "two degrees of separation" between the provenance information and the workflow engine.

The two-degrees-of-separation approach is also founded on the Component-Based Software Engineering principle and on recent developments in service-oriented computing (SOC). The CBSE approach is based on reusable, loosely coupled, independent components for software system development.[10] The Web-services-based SOA approach realizes the CBSE approach's objectives. Provenance-generation tools implemented as specialized Web services take advantage of the extensive and comprehensive Web services ecology already in place featuring representation schema, communication standards, and a registry standard.

Some workflow engines' use of Web Services Description Language (WSDL)-based descriptions to create provenance is already constrained by the ambiguous data-typing of parameters (often as a "string" data type). Furthermore, the SOC community is rapidly adopting "lightweight" representational state transfer (REST) services as an alternative to a "heavyweight" WSDL-based architecture. Unlike a predefined contract in the WSDL-based approach, consisting of precondition, postcondition, and I/O parameters, REST services have minimal textual descriptions. Workflow engines relying purely on WSDL descriptions to derive provenance information might not be a sustainable approach.

The Semantic Provenance Framework

We describe the semantic provenance framework for eScience along three fundamental dimensions (see Figure 3):

- semantic provenance annotation,
- domain provenance ontologies, and
- usage.

The first dimension involves a set of specialized tools plugged into a scientific workflow on demand to create semantic-provenance information. Extracting comprehensive metadata from multiple sources, such as generated scientific data and Web forms (for parameter specifications, equipment details, project details, and so forth) is another important element of this dimension.

The second dimension uses domain-specific provenance ontologies to model scientific processes, data (including temporal information), and agents as formally defined concepts linked together using named relationships.

In the third dimension, software agents use reasoning tools to process the semantic-provenance information and answer complex domain queries. They can also use semantic-provenance information to compare, integrate, retrieve, and visualize scientific data.

This semantic-provenance framework achieves the important requirements identified by the proposed Open Provenance Model (OPM), part of the international provenance challenge (see http://twiki.ipaw.info/bin/view/Challenge/OPM). It also addresses many nonfunctional requirements using the rich set of publicly available resources that the Semantic Web research community has created.

Semantic provenance addresses four OPM requirements. The first is *provenance information interoperability*. Using ontology schema mapping and merging techniques (www.ontologymatching.org), semantic provenance from different workflows will be interoperable, so the eScience community can share and integrate them.

The second requirement addressed is *ease of application development*. The wide availability of tools for Semantic Web resources, such as the Jena toolkit (http://jena.sourceforge.net/) and Sesame (http://openRDF.org), make it easier to develop applications.

The well-defined semantics of the Resource Description Framework (RDF) model,[11] and expressive formal-logic-based OWL language, address the next requirement, *precise description of provenance information*.

The final OPM requirement addressed is *inference capability and digital representation of provenance*. Software applications can use tools such as Racer (www.sts.tu-harburg.de/~r.f.moeller/racer), Pellet (http://pellet.owldl.com), and FaCT++ (http://owl.man.ac.uk/factplusplus) to perform reasoning over semantic provenance. Because digital representation is a foundational characteristic of the Semantic Web, semantic provenance supports digital representation of provenance information.

Semantic provenance also addresses three nonfunctional requirements. The first is *publicly available ontologies*. The set of publicly available ontologies listed on open biomedical ontologies (OBO) at the National Center for Biomedical Ontologies (NCBO; www.bioontology.org) represent a tremendous research effort and should be reused for life sciences domain provenance. Many other domains that use the eScience platform are also developing high-quality ontologies such as in geospatial sciences (www.w3.org/2005/Incubator/geo/XGR-geo-ont) and environmental sciences (http://sweet.jpl.nasa.gov).

The next nonfunctional requirement addressed is *storage and querying resources*. The SPARQL query language (www.w3.org/TR/rdf-sparql-query) has been accepted as a W3C recommendation for querying RDF resources. There are multiple storage solutions available for Semantic Web resources including Oracle 11g (www.oracle.com/technology/products/database/oracle11g), Kowari (www.kowari.org), Virtuoso RDF (http://virtuoso.openlinksw.com/wiki/main/Main/VOSRDF), and Jena.

Finally, semantic provenance supports *visualization tools for Semantic Web resources*. Many open source applications have been developed for visualization and browsing Semantic Web data. Some examples projects include Welkin (http://simile.mit.edu/welkin), multiple plug-in tools for the Protégé environment (http://protege.stanford.edu), and Semantic Analytics Visualization (SAV; http://lsdis.cs.uga.edu/projects/semvis).

Spade

Here, we describe a realization of the semantic provenance framework in the glycoproteomics domain.

Background

Mass spectrometry (ms) is an analytical procedure for proteomics data to study protein structure and posttranslational modifications. Software tools analyze raw data produced by a mass spectrometer in a multistep process that yields a list of identified entities and their quantification. The protocol that scientists at the Complex Carbohydrate Research Center (CCRC) follow for protein identification from ms data (Figure 1) is typical in proteomics research. This high-throughput process might generate more than 500 data files from a single sample.

Scientists originally conducted this analytical procedure manually by transferring data across distributed systems and then invoking software tools. The scientists, who were responsible for keeping track of each result file across multiple projects, often spent frustratingly long hours searching for a previous result or trying to correlate results using handwritten notes. We completely automated this analytical process as a scientific workflow using Semantic Web services (Web services annotated with ontological concepts) orchestrated using the Taverna workflow engine (http://taverna.sourceforge.net).

Many prior efforts have automated scientific protocols, and workflow-based automation in itself isn't novel; what's new is the support for semantic provenance. To help scientists manage the large volumes of data using provenance information, we developed the ProPreO proteomics provenance ontology (described in the next section).[11] Next, we implemented a set of semantic-provenance creation services that are plugged in at each intermediate step of the workflow (see Figure 4). This infrastructure is Spade.

Spade creates semantic provenance in two phases. The first phase is entity extraction. Relevant descriptions for creating provenance in-

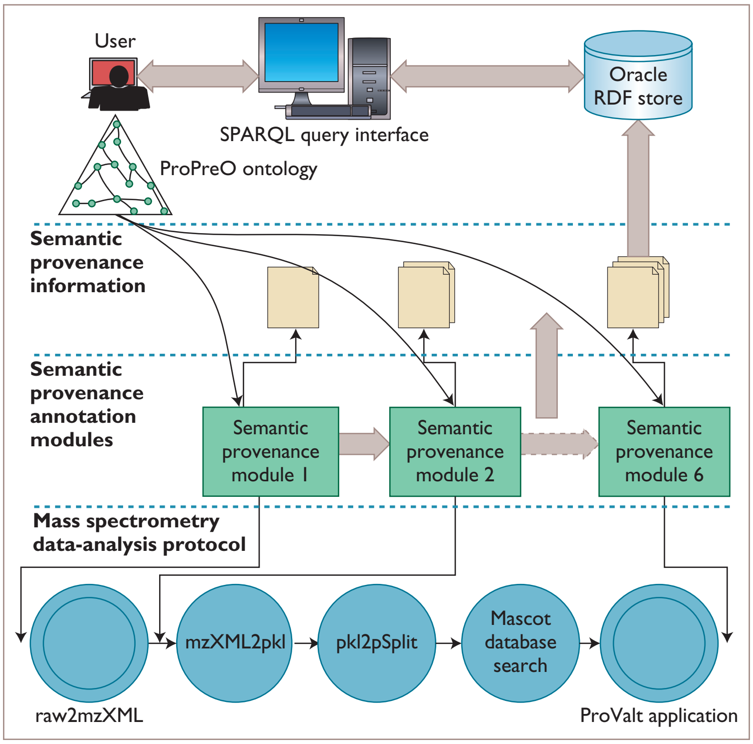

User
SPARQL query interface
Oracle RDF store
ProPreO ontology
Semantic provenance information
Semantic provenance annotation modules
Semantic provenance module 1
Semantic provenance module 2
Semantic provenance module 6
Mass spectrometry data-analysis protocol
mzXML2pkl
pkl2pSplit
Mascot database search
raw2mzXML
ProValt application

Figure 4. The Semantic Provenance for Data in protEomics (Spade) architecture. Spade encompasses the scientific workflow, the ProPreO ontology, provenance creation service, and generated semantic provenance.

formation — such as parameter details, project descriptions, and identified biological entities (for example, protein groups) — are extracted either from Web forms that users fill out at the start of the workflow or from data files generated during the sample run. These entities are categorized as instances of ProPreO ontology classes using class membership relations based on a set of heuristic rules. The entity extraction and classification at each step of the workflow results in an aggregated list of ProPreO ontology class instances at the end of the workflow.

During the second phase, the provenance-creation services assert named relationships that apply between two entities (categorized as instances of ProPreO classes in the previous step), using the ProPreO ontology schema as reference. We use Jena to traverse the ontology schema and identify the correct relationship between two entities.

The semantic provenance thus created during each sample run is represented as RDF triples and is loaded after conversion to Notation3 (N3) format (using Jena) into the Oracle 10g database (www.oracle.com/technology/software/products/database/oracle10g). We currently use the SPARQL query interface supported by the Oracle 10g (Release 2) database to query the semantic provenance, but we're developing a more intuitive, graphical query interface for scientists.

The ProPreO ontology

The ProPreO ontology is the central resource that underpins semantic provenance in Spade. ProPreO is a large domain-specific provenance ontology[12] with three primary concepts to model proteomics data analysis: data, tasks, and agents (which initiate or participate in task execution). ProPreO currently has approximately 490 classes and 35 named relationships with 145 constraints, such as *class-level* restrictions. It's also populated with 3.1 million instances of `ProPreO:tryptic peptide`. (Program-code font indicates ontological terms here and throughout this section.) ProPreO has been released for community use and is listed at the OBO at NCBO.

We describe the CCRC proteomics data-analysis procedure as modeled in the ProPreO ontology (see Figure 5) to illustrate the expressiveness of semantic provenance in Spade. The analysis procedure yields a set of peptides and protein groups (`ProValt_output_data`) generated by the algorithm ProValt (`ProValt`), which statistically analyzes the peptide or protein identifications made by the Mascot database search engine in the previous step. The search engine performs tasks (`data_classification` and `data_correlation`) to analyze peak-list data (`ms-ms_peak_list`) to identify peptides or proteins that are represented as records in a protein database (`protein_sequence_database`). The database search engine and ProValt each use a set of operating parameters to generate the data sets. These parameter sets (`input_operating_parameter_collection`) are related to the computational tasks via the named relationship (`has_input_operating_parameter_collection`). The original peak lists are created by a task implemented by peak-list extraction algorithms that use data (`MS_raw_data_native_format`) recorded by a specific category of mass spectrometer (`Micromass_QTOF_2_quadrupole_time_of_flight_mass_spectrometer`). The pattern we see emerging is a rich, interconnected graph that logically correlates data sets, processes, and instruments, as Figure 5 illustrates.

Query example

The following example from the ms group at

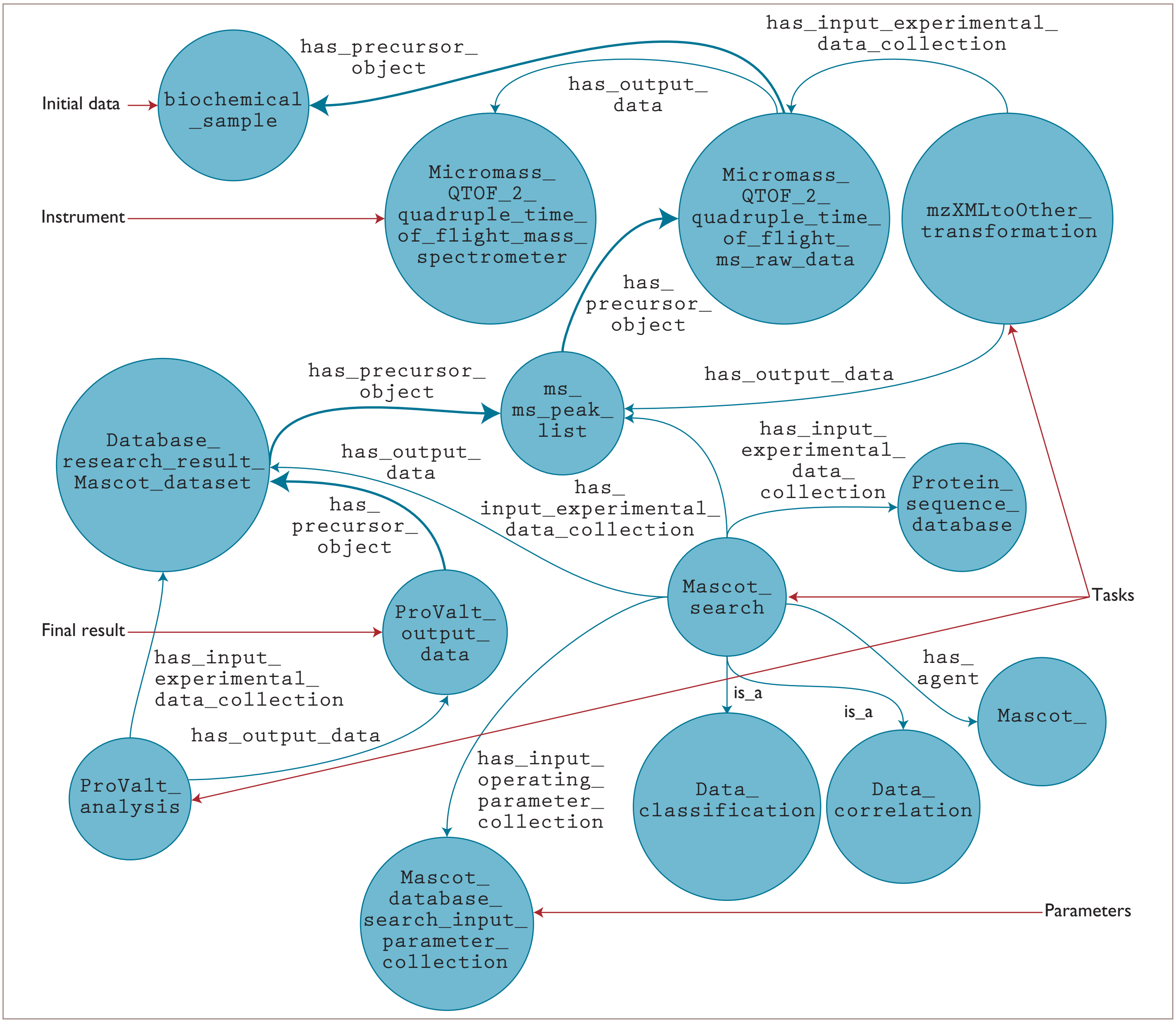


Figure 5. The proteomics data-analysis protocol. This schematic representation of the protocol shows how it uses ProPreO ontology concepts and the named relationships linking them.

CCRC illustrates real-world use of Spade: Jean, a new graduate student, is scheduled to make a presentation in the next group meeting. The presentation will let the group peer review Jean's research protocols by evaluating the quality of her experimental results.

Jean issues a query against the semantic-provenance information associated with the ms data repository:

> List the protein groups identified with high confidence value — that is, protein groups with a Mascot score > 3500 — detected by the Mascot search engine against a T.cruzi database (Mascot search input parameter, Taxonomy = T.cruzi. The protein groups should contain at least one peptide fragment with a specific consensus sequence of {*N [^P] [S/T]*}.

This query seeks to identify the best-quality results from all the sample runs executed until the current date to identify and integrate data from multiple result files. In the proteomics data-analysis protocol, the Mascot database search engine assigns scores to protein groups that reflect the confidence value of the identification. Each of the identified protein groups is associated with its Mascot score using a named relationship that identifies protein groups with a Mascot score greater than 3500. The other constraint, described as the presence of the amino acid sequence {*N [^P] [S/T]*} in peptide fragments, is the N-glycosylation consensus sequence in peptides, which is of particular interest to glycobiologists. The peptide fragments in a protein group are associated with their amino acid sequence, again using a named relationship. The peptide fragments are related to the protein groups through the ProPreO ontology relationship `ProPreO:has_parent_protein`. A SPARQL query representing the user query is executed against the semantic provenance information to retrieve the relevant results from the data repository.

The semantic-provenance framework is a generic approach to building a provenance infrastructure in different domains by extending and adapting to the requirements of specific domains. We're implementing this framework to model provenance information of sensor data related to weather forecasting to demonstrate the use of semantic provenance information for data integration. We're also extending the ProPreO ontology to incorporate a Nuclear Magnetic Resonance (NMR)-based data-analysis protocol. This will let software applications use semantic provenance information to create an unambiguous context for comparing experimental data for toxicology metabolomics using ms-based and NMR-based data-analysis approaches.

Acknowledgment

This research was supported in part by the NIH-NCRR-funded Integrated Technology Resource for Biomedical Glycomics (P41 RR18502) project and by NSF-ITR Award #0714441, "SemDIS: Discovering Complex Relationships in the Semantic Web." We also thank Will York and James Atwood III for their valuable input during the development of the ProPreO ontology.

References

1. D. Atkins, *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, Nat'l Science Foundation, 2003.
2. T. Hey and A.E. Trefethen, "Cyberinfrastructure for e-Science," *Science*, May 2005, pp. 817–821.
3. V. Kashyap and A. Sheth, "Schematic and Semantic Similarities between Database Objects: A Context-Based Approach," *Very Large Databases J.*, vol. 5, no. 4, 1996, pp. 276–304.
4. E. Camon et al., "The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro.," *Genome Res.*, vol. 13, no. 4, 2003, pp. 662–672.
5. A. Sheth, "Semantic Meta Data for Enterprise Information Integration," *Data Management Rev.*, July 2003; www.dmreview.com/issues/20030701/6962-1.html.
6. C. Goble, "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics," *Proc. Workshop on Data Derivation and Provenance*, 2002; http://people.cs.uchicago.edu/~yongzh/papers/provenance_workshop_3.doc.
7. Y.L. Simmhan, A.B. Plale, and A.D Gannon, "A Survey of Data Provenance in e-Science," *SIGMOD Record*, vol. 34, no. 3, 2005, pp. 31–36.
8. W.C. Tan, "Provenance in Databases: Past, Current, and Future," *IEEE Data Eng. Bull.*, vol. 30, no. 4, 2007, pp. 3–12.
9. R. Stevens, J. Zhao, and C. Goble, "Using Provenance to Manage Knowledge of *In Silico* Experiments," *Briefings in Bioinformatics*, vol. 8, no. 3, 2007, pp. 183–194.
10. I. Sommerville, *Software Engineering*, Pearson Education, 2004.
11. P. Hayes, *RDF Semantics*, World Wide Web Consortium (W3C), B. McBride, ed., 2004; www.w3.org/TR/rdf-mt.
12. S.S. Sahoo et al., "Knowledge Modeling and its Application in Life Sciences: A Tale of Two Ontologies," *Proc. 15th Int'l Conf. World Wide Web* (WWW 06), ACM Press, 2006, pp. 317–326.

Satya S. Sahoo is a PhD student at the Kno.e.sis Center at Wright State University. His research interests include semantic provenance, knowledge representation, and information integration in biomedical and sensor Web domains. Sahoo has a masters in computer application from Goa University, India. Contact him at satyasahoo@gmail.com; http://knoesis.wright.edu/students/satya.

Amit Sheth is the LexisNexis Ohio Eminent Scholar and director of Kno.e.sis Center at Wright State University. His research interests include the Semantic Web and semantic applications, management of rich media content and information sources, and digital libraries. Sheth has a PhD in computer and information science from Ohio State University. He's an IEEE Fellow. Contact him at amit.sheth@wright.edu; http://knoesis.wright.edu/amit.

Cory Henson is a PhD student at the Kno.e.sis Center at Wright State University. His research is in the semantic sensor Web. Henson has a BS in computer science from the University of Georgia. Contact him at coryhenson@gmail.com; http://knoesis.wright.edu/students/cory.