Provenance Explorer – A Tool for Viewing Provenance Trails and Constructing Scientific Publication Packages

Kwok Cheung[1], Jane Hunter[2]

[1]AIBN, The University of Queensland
St Lucia, Queensland, Australia
kwokc@itee.uq.edu.au
[2]ITEE, The University of Queensland
St Lucia, Queensland, Australia
jane@itee.uq.edu.au

Abstract. This paper presents *Provenance Explorer*, a secure provenance visualization tool, designed to dynamically generate customized views of scientific data provenance that depend on the viewer's requirements and/or access privileges. Using RDF and graph visualizations, it enables scientists to view the data, states and events associated with a scientific workflow in order to understand the scientific methodology and validate the results. Initially the Provenance Explorer presents a simple, coarse-grained view of the scientific process or experiment. However the GUI allows permitted users to expand links between nodes (input states, events and output states) to reveal more fine-grained information about particular sub-events and their inputs and outputs. Access control is implemented using Shibboleth to identify and authenticate users and XACML to define access control policies. The system also provides a platform for publishing scientific results. It enables users to select particular nodes within the visualized workflow and drag-and-drop them into an RDF package for publication or e-learning. The direct relationships between the individual components selected for such packages are inferred by the rule-inference engine.

Keywords: eScience, Provenance, Visualization, Inferencing

1 Introduction and Objectives

Provenance is essential within science because it provides a history or documentation of the steps taken during the scientific discovery process. Understanding the source of data or how scientific results were arrived at, is essential in order to verify or trust that data and to enable its re-use and comparison. A record of the complete scientific discovery process enables peers to review the method of conducting the science as well as the final conclusions. Precise, authenticated provenance data reduces duplication and insures against data loss because the additional contextual and provenance information ensures the repeatability and verifiability of the results [1]. It also enables precise attribution of individual credit during collaborations involving teams of scientists.

Ideally provenance collection systems are in place that are capable of recording both the domain-specific steps in the physical world (e.g., the laboratories or processing plants) as well as the data derivation steps in the digital domain. Increasingly, e-Laboratory notebooks and workflow systems are being developed specifically to relieve the effort required by scientists to capture the precise provenance metadata required to validate scientific results and enable their duplication. Assuming appropriate metadata is being captured at each stage in the workflow associated with scientific discovery process, then many of the relationships between the individual components are either explicitly captured or can be inferred later, as required. This is particularly true of systems that record the sequence of events, inputs and outputs in machine-processable descriptions represented using RDF graphs and domain-specific OWL ontologies.

We are interested in those workflow and e-Lab notebook systems that are based on RDF. Recentris' Collaborative Electronic Research Framework (CERF)[1] and the SmartTea [2] and MyTea [3] systems are examples of RDF-based laboratory notebook systems. RDF-based workflow systems that support the capture of provenance information include Kepler [4], Taverna[5] and Triana [6]. Our objective is to take the output from such systems (i.e., the RDF instances that describe the sequence of events and data products recorded during the execution of a scientific workflow) and apply reasoning across these sets of records to infer new relationships between indirectly related data products. These inferred relationships can be used to generate alternative but still correct views of the data provenance. Alternative views of provenance are required for a number of reasons. Simplified views of highly complex workflows may be required for teaching or publication purposes. Restricted views which hide certain information or details are required to protect the intellectual property associated with particular scientific processes. This is particularly important within collaborating teams of scientists to protect individual IP but still enable controlled sharing and validation of the overall process. Hence our objectives are to leverage existing RDF-based workflow tools and the captured provenance data and metadata in order to:

- generate visualizations of the lineage of the data and its products i.e., the relationships between the different derivative products generated during the scientific process;
- dynamically infer customized views of provenance depending on the user's requirements and privileges;
- restrict access to specific data or processing steps (using Shibboleth [7] to authenticate users and XACML [8] to define policies) - in order to protect intellectual property and maintain competitive advantage;
- streamline the construction of publication or e-learning packages (that link the raw data to its derivatives and traditional scholarly publications).

The remainder of this paper is structured as follows: Section 2 describes related work; Section 3 describes the case study we used for evaluation and testing; Section 4 describes the system architecture and components; Section 5 describes the implementation and user interface and Section 6 concludes with an evaluation, discussion and future work plans.

[1] http://www.rescentris.com/

2 Related Work

Our aim is to take the output from existing RDF-based provenance capture systems and to develop a visualization tool that dynamically generates customized views of the provenance trail. For example, Kepler [4] is a scientific workflow system designed for multiple disciplines which enables scientists to design and execute workflows. Recently, Kepler embedded a new provenance recording component that collects data and workflow provenance at runtime. Similarly, CERF provides a unified electronic record-keeping environment for scientists, in particular for biologists, to capture, curate, annotate, and archive their data, and to integrate the data into electronic lab notebook-like pages. Either of these two systems could integrate seamlessly into Provenance Explorer because they are both java-based applications. Furthermore, Protégé-OWL Plugin API can be used as the interface between either system and Provenance Explorer.

The *Prototype Lineage Server* [9] allows users to browse lineage information by navigating through the sets of metadata that provide useful details about the data products and transformations in a workflow invocation. Web server scripts on the lineage server query the lineage database, and provide a Web browser interface, that allows navigation via HTML links. Views are restricted to parent and children metadata objects. Clicking on a parent object will move that link to the center of the screen and show that object's parents. Clicking on the metadata object link in the center of the screen will bring up the XML metadata for an object.

Pedigree Graph [10], one of tools in Multi-Scale Chemistry (MCS) portal from the Collaboratory for Multi-Scale Chemical Science (CMCS), is designed to enable users to view multi-scale data provenance. The portlet provides scientists with a two-dimensional visualization of a data object or file and all of its scientific pedigree relationships. The view is static, and rendered straight from GXL (Graphical eXchange Language) files but users are able to traverse the tree by clicking on links.

The MyGrid project renders graph-based views of RDF-coded provenances using Haystack [11]. This is used to visualize networks of semantic relationships among provenance resources associated with experiments. Haystack is a Semantic Web browser that enables developers to provide tailored views over RDF-metadata. The authors point out that Haystack is highly resource-consumptive because its execution is based on Adenine, a high level programming language developed on top of Java Programming Language. Hence the response time to user's instructions could be slow.

The *VisTrails* system [12] was developed by the University of Utah for building, storing, editing and visualizing workflows and interactively tracking workflow execution and evolution. Although it uses graphs to visualize workflows and provenance trails, it differs from the Provenance Explorer in that it is not designed to generate personalized views of provenance – adapted for publication or teaching purposes or to suit a user's interest or access permissions.

So although there are existing systems that enable visualization of RDF-encoded provenance graphs, the unique aspect of our Provenance Explorer system is its ability to generate personalized views of the provenance relationships automatically using a combination of user input, semantic reasoning and access policies.

3 Case Study

Within the University of Qld, materials scientists within the Australian Institute for Bioengineering and Nanotechnology are investigating the optimization of fuel cells – an alternative environment-friendly energy source to fossil fuels. Their efficiency depends on the internal structure of the fuel cell components and their interfaces. Electrolytes are one of the primary fuel-cell components. Figure 1 illustrates the complex set of steps involved in the process of manufacturing and testing electrolytes. Associated with each step in the workflow is a set of parameters, only some of which are controllable. The objective of the fuel-cell scientists is to determine the optimum combination of controllable parameters in order to attain the maximum strength, efficiency and longevity of the fuel cell for the minimum cost [13] .

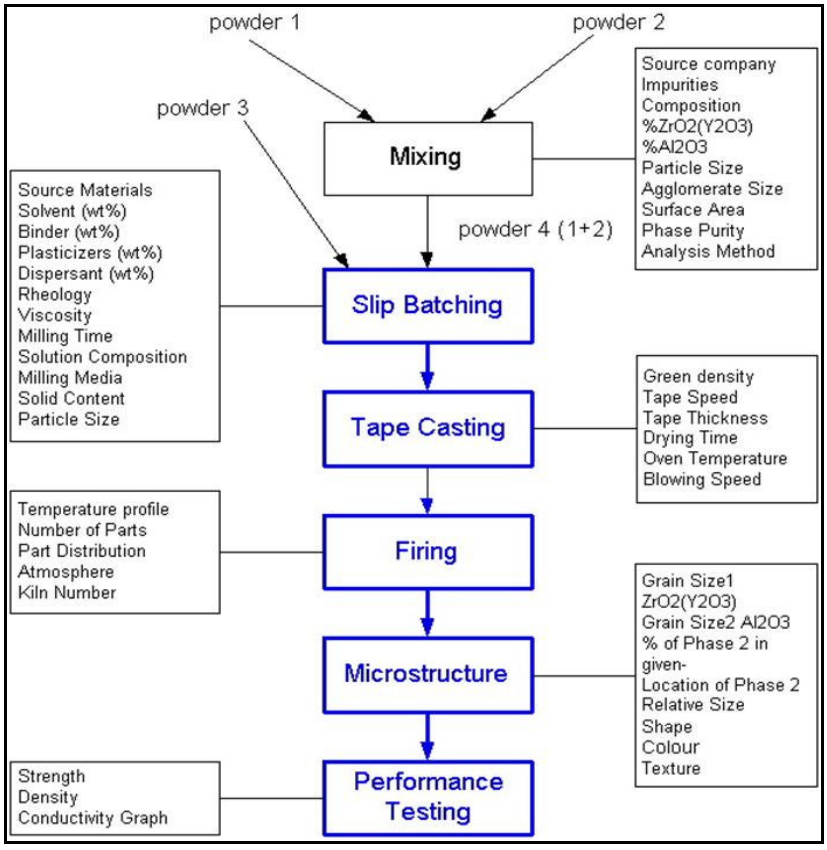

Fig. 1. A logical view of the manufacture and testing process of Fuel-Cell Electrolyte

Through the FUSION project [14] we have been collaborating with a team of fuel cell scientists on the development of an eScience workflow and provenance capture system that records the data associated with each of the steps in the electrolyte

manufacturing and testing process and enables its statistical analysis in order to generate new workflows [13]. Through this work we have access to data records from a series of manufacturing and testing experiments. Hence we decided to use this application as a case study for evaluating and attaining user feedback on the Provenance Explorer system. The first step involved modeling the workflow in Figure 1 and representing it in OWL. We decided to use the event-aware ABC ontology [15], developed within the Harmony project, to track the life cycle of digital objects. We first had to extend the ABC ontology to describe processing, simulation and experimental events. Given this extended ontology, we were able to represent the workflow instances corresponding to Figure 1 in OWL. This is illustrated in Figure 2.

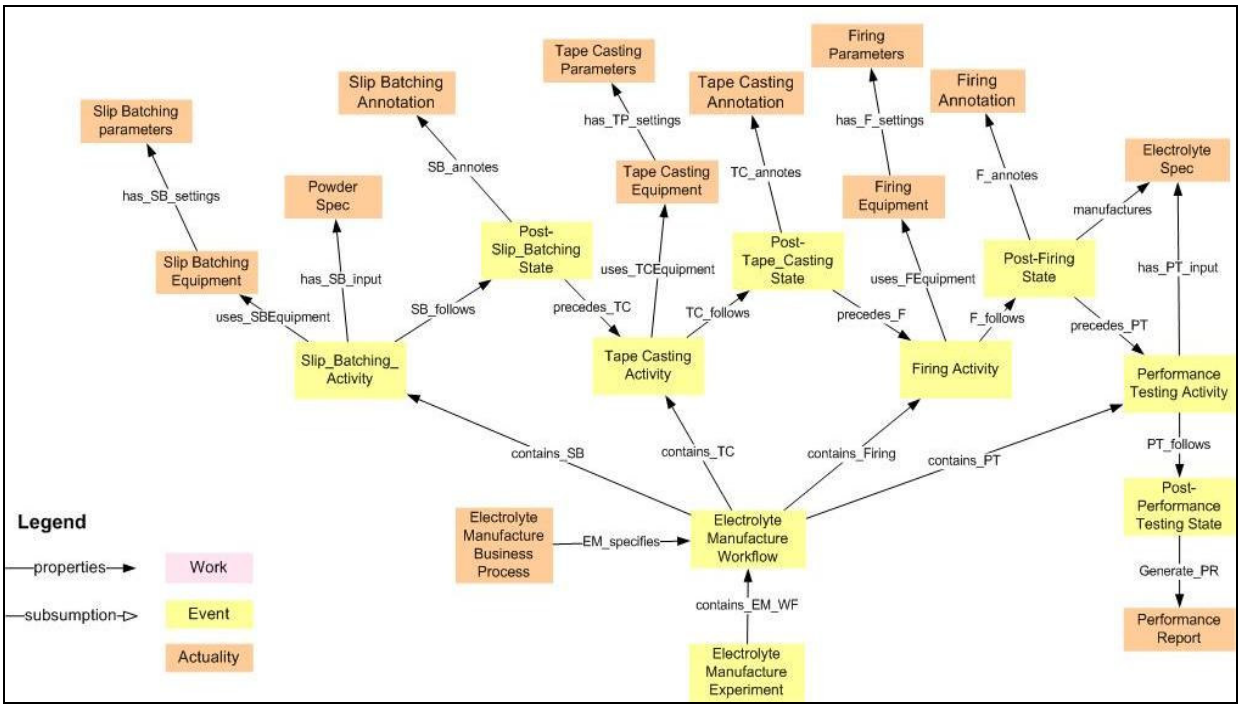

Fig. 2. Provenance Model of the Electrolyte Manufacture and Analysis Process

Given the OWL representation of the provenance data associated with the fuel cell manufacturing and testing process, the aim was to generate customized graphical visualizations of the data using the Provenance Explorer system – to satisfy the requirements of the scientists. In addition to the OWL instance data, we also had to develop rules for inferring relationships between entities that were not directly related and represent them in the Semantic Web Rule Language (SWRL)[16]. For example:

IF (Experiment A includes Workflow A) AND
(Workflow A contains Slip Batching A) AND
(Slip Batching A hasInput Powder A)
THEN (Experiment A hasInput Powder A)

4 System Architecture

Figure 3 illustrates the overall system architecture and its key components. The three key components of the system are:

- The knowledge base which consists of SWRL.OWL files that contain the provenance instance data and metadata and the inference rules.
- the Provenance Visualizer and
- Algernon, a rule-inference engine.

The SWRL.OWL files are input to both the Provenance Visualizer and Algernon. Jena and Protégé-OWL Plugin act as the interface between the Provenance Visualizer and the SWRL.OWL files, and between Algernon and the SWRL.OWL files, respectively. Jena [17], developed by HP Labs, provides the programmatic environment for RDF, RDFS and OWL. Jena supports SPARQL[18] which is used to query the SWRL.OWL files. The Protégé-OWL Plugin was used to generate the SWRL.OWL files and to retrieve the rules from the SWRL.OWL files for Algernon to process at runtime. Algernon[19] is a rule-inference engine that supports both forward and backward chaining rules of inference, and implements Access-Limited Logic. However because Algernon does not support the inference of subsumption between properties or comply with the SWRL rule format, the rules retrieved from SWRL.OWL files by Protégé-OWL Plugin APIs had to be transformed to the Algernon-compliant rules before being imported to Algernon at runtime.

The Provenance Visualizer, is the graphical user interface (GUI) powered by JGraph [20] (an extension of Java Swing GUI Component to support directed graphs). The Provenance Visualizer GUI is divided into three panels horizontally:

1) The Provenance View, in the upper panel, presents a graphical view of the provenance process modeled using RDF graphs.
2) The Publishing Interface, in the central panel, enables users to construct packages for publishing scientific results. The users can drag and drop selected components from the upper panel into an RDF package. Any two can be linked manually with the relationships inferred automatically by Algernon.
3) Finally, the Provenance data, in the bottom panel, displays the provenance details (metadata) for the object highlighted in the upper panel.

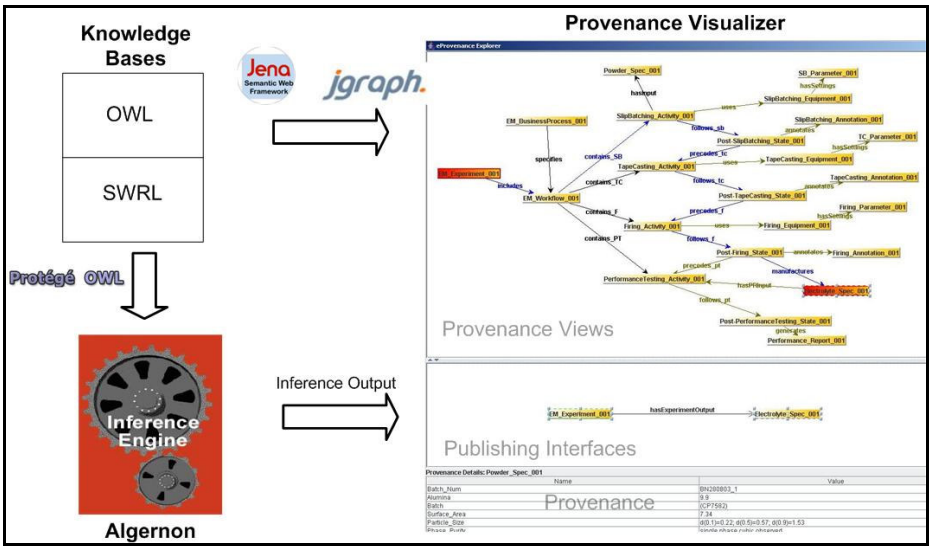

Fig. 3. System Architecture

There are access controls imposed on the upper panel's graphical view. The granularity of the displayed view depends on the user privileges and access policies, enforced and defined by Shibboleth and XACML

To enforce the inter-institutional authentication and access control, Shibboleth , a centralized identity and authorization mechanism developed by the NSF Middleware Initiative, was adopted and incorporated within the Provenance Explorer. Shibboleth is standards-based, open source middleware software which provides Web Single SignOn (SSO) across or within organizational boundaries. Figure 4 demonstrates the two primary components of Shibboleth: the Identity Provider (IdP) and Service Provider (SP). The IdP maintains user credentials and attributes. Upon request the IdP will assert authentication and attribute statements to requesting parties, specifically SPs. The SP then uses predefined-XACML policies to control access to the Provenance Explorer and fine-grained provenance views on the upper panel.

XACML complements Shibboleth to address fine-grained access control on the resources. XACML, the Extensible Access Control Markup Language, provides a vocabulary for expressing the rules needed to define fine-grained and machine-readable policies and make authorization decisions. In this system we use Sun's XACML[2] implementation which includes an XACML engine and an API for easy integration.

Initially, authenticated users of Provenance Explorer are presented with the coarsest view of provenance. When a user attempts to retrieve finer-grained views by clicking on links between entities, a request is generated, the XACML engine compares the request with the policies on these entities and makes the authorization decision.

[2] http://sunxacml.sourceforge.net/

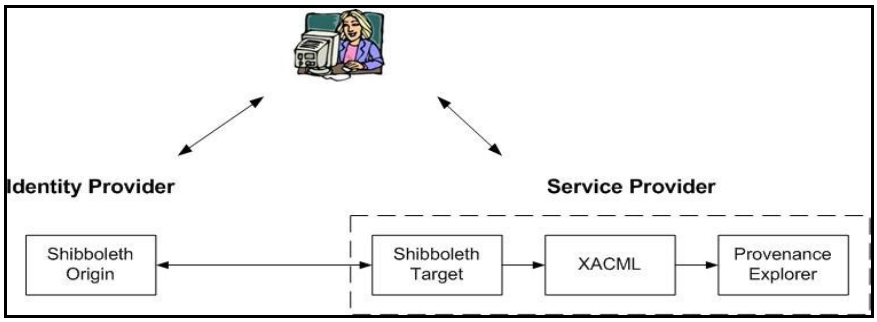

Fig. 4. Authentication and Authorization System Architecture

5 Demonstration and User Interface

Within the FUSION project, members of the "Virtual Organization" (those users collaborating on the project and sharing different aspects of the data) can be categorized into three main role types with three different levels of access:

1. the fuel-cell researcher from the AIBN (also the project leader)
2. the technicians from the fuel-cell manufacturing company
3. post-graduate students from the University of Qld and Monash University,

The fuel-cell researcher designed the original workflows, over-saw the entire process, developed new hypotheses and models, designed new experiments, and wrote publications describing the results and conclusions. The technicians carried out the manufacturing (slip batching, tape casting, firing) and performance testing activities. Finally, the post-graduate students working on specific aspects of fuel cells were entitled to view different components of the process to different levels. The fuel-cell researcher had the highest privileges and was entitled to explore the complete set of provenance records. He/she was also able to select provenance components to incorporate within publication or e-learning packages. The technicians had modest privileges – they were able to access to the provenance associated with each of their own activities, whereas the students had the minimum privileges with restricted access to provenance details. In the following section we describe the system from the point of view of each of these user types.

Firstly consider the researcher/project leader. He/she logs onto the Shibboleth Service provider where the Provenance Explorer service is installed. Initially, the user is redirected to Shibboleth's Identity Provider for authentication and authorization. Once authenticated, the user's attributes are returned back to the Service Provider and the user is granted access to the Provenance Explorer. The researcher searches for provenance of Batch Number 280818. Initially the researcher is presented with the basic view of the experiment provenance. This is the default view for all users with access privileges to the FUSION project's Provenance Explorer service. Figure 5 demonstrates the default expandable view. The pink arrows indicate relationships that can be expanded to reveal further fine-grained information about the sub-activities.

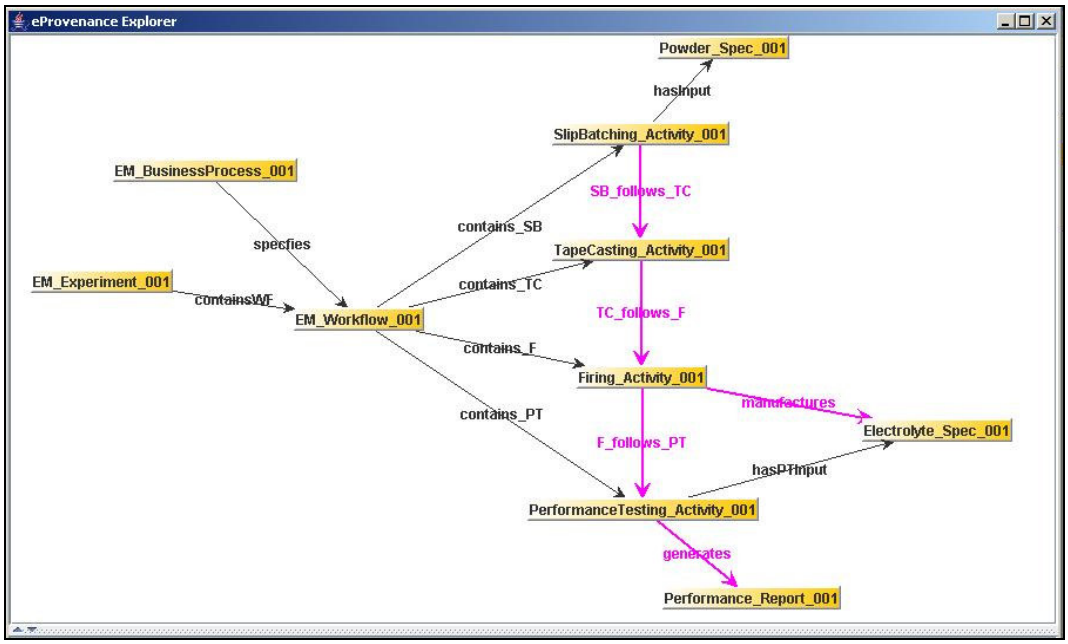

Fig. 5. A standard basic view

When the researcher clicks on a pink arrow, a request for additional information is generated and submitted to the XACML engine. The XACML engine compares the request with the policy and makes an authorization decision accordingly. Figure 6 demonstrates the policy and request.

Fuel-Cell Researchers (AIBN ="researcher") Read All Views= "Permit"	Subject AIBN="researcher" Resource Resource="http://www.owl-ontologies.com/EM_ScientificProcess.ow/#SBViews" Action Action-id ="read"

Fig. 6. Example policies and requests

Eventually by interactively drilling down via the links, the researcher is presented with the complete view. Figure 7 illustrates the complete view in the upper panel. The dark green arrows indicate parts of the expanded view and can be collapsed manually back to the original view i.e., the pink expandable links. If an individual node on the upper panel is selected, the complete provenance metadata for this node is displayed in the bottom panel. Figure 7 demonstrates this feature. Node *Powder_Spec_001* is highlighted in a red circle on the upper panel, and the associated provenance information is displayed in the bottom panel.

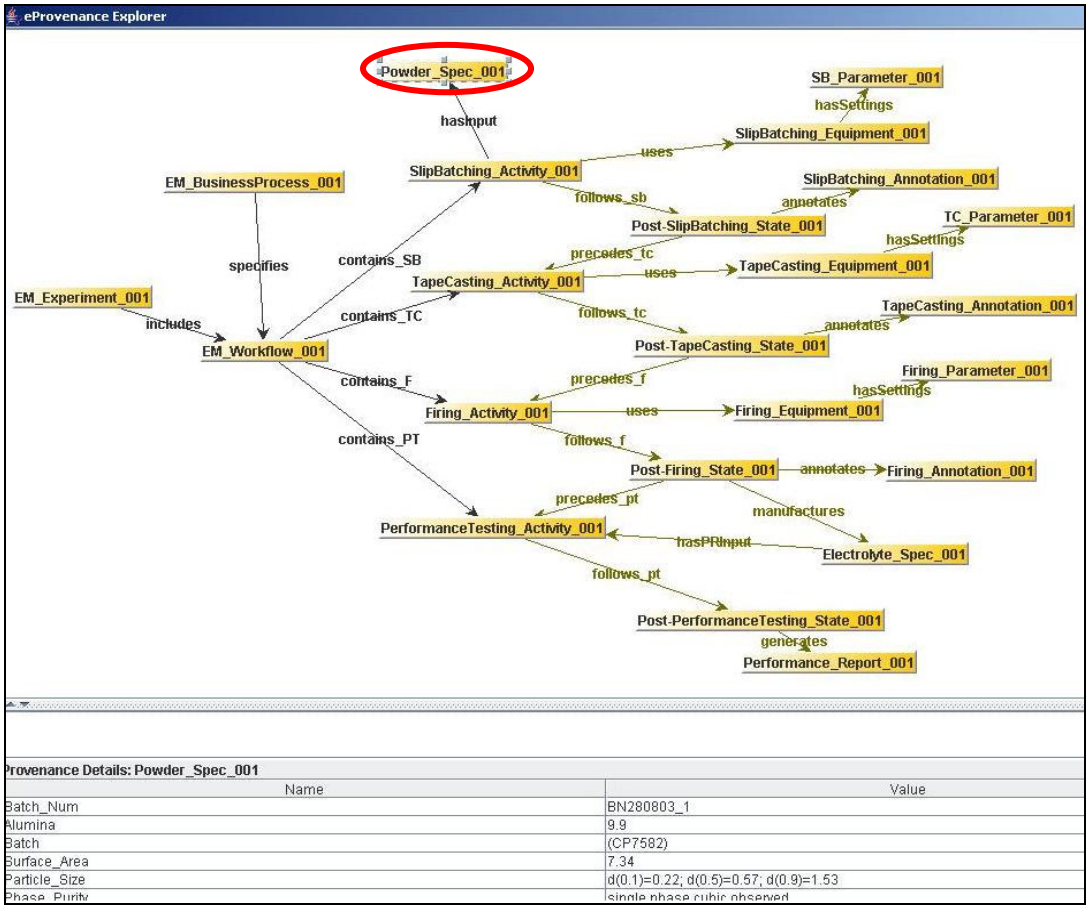

Fig. 7. An expanded complete provenance view for the Researcher/Project Leader

Furthermore, using this interface, the researcher is able to manually construct a package of related components for publication or dissemination. This is performed by selecting nodes in the top panel and dragging and dropping them into the middle panel. By linking them manually, the relationship between the nodes is inferred by the rule-inference engine. For example, Figure 8 demonstrates that the relationship inferred between the two selected nodes, Experiment_001 and Electrolyte_Spec_001 is *hasExperimentOutput*. The path used to infer this relationship is highlighted in blue (with the beginning and end nodes highlighted in red) in the upper panel. Figure 2 illustrates that in the ontology we define an experiment as comprising of a sequence of activities with particular post-event states. The inferencing rule states that any product generated by one of the activities in the sequence is an output of the experiment.

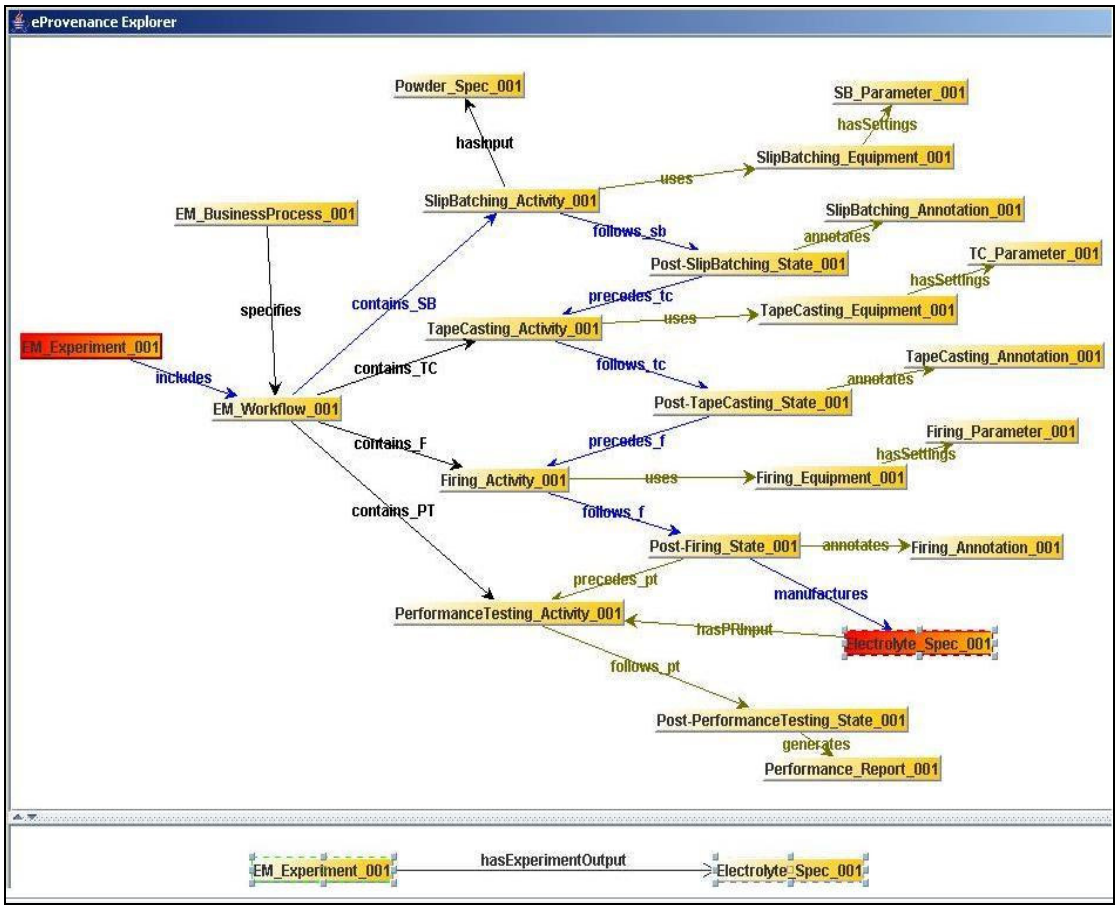

Fig. 8. Demonstration of Provenance Inferencing

Now consider the system from the point of view of the Slip-batching operator. After logging in and being authenticated, the operator/technician is presented with the default view. This is almost identical to Figure 5, except that there is just one expandable pink arrow *SB_follows_TC*, indicating that further expansion is restricted to the slip batching activity. Finally, the Post-graduate students were also entitled to access the default coarse-grained view of the experiment – but with no expandable pink arrows.

6 Discussion and Conclusion

6.1 User Feedback

Initial feedback from the fuel-cell scientists involved in the FUSION project has been very positive. The system enables them to quickly and intuitively understand

quite complex workflows and to compare different workflows. They are able to pinpoint problems within a particular workflow and to generate new experimental workflows accordingly. Users can understand the system very quickly because of its close analogy to the web – the use of hyperlinks for information exploration and navigation. Furthermore, with regard to the data's validity, the scientists can intuitively track the data's provenance via the complete graphical view of visualized scientific processes and the detailed metadata associated with any node. The users were also very positive about the security framework – in particular the advantages of the single sign-on capability of Shibboleth and the ability to hide certain steps or the details associated with certain steps in the process.

However, users did raise concerns with regard to scalability and searching. At this stage, our demonstration involves multiple instances of a single workflow. In reality, the scientists may need to search, retrieve and compare multiple experiments simultaneously and the experimental workflows may be very different. Moreover, the current methods by which scientists can discover and retrieve experimental workflows are limited. Currently the system only permits search and retrieval of experiments via a unique ID. Scientists would like to be able to search for experiments via particular attributes e.g., particular parameter values. The optimum methods for describing, indexing and discovering workflows require further investigation and direct input from the end-users.

6.2 Limitations and Future Work

The provenance metadata, graphical views and inferencing rules of the Provenance Explorer were all based on the provenance model in Figure 2. This model is an extension of the ABC model developed within the Harmony project - extended to support experiments in laboratories. This model provides the semantic underpinning of the system, and the ontology's robustness may become a significant issue if/when the system is expanded across domains and organizations. Colomb argues that formal ontologies, such as DOLCE[21] and BWW [22], provide a rich meta-vocabulary and abstract data types, and well-understood structural organizational principles, thereby technically enhancing the reliability of material ontologies [23] like our ontology. Thus, it may be worth carrying out further investigation on formal ontologies to determine how they can make the provenance model more reliable and rational in terms of the data structures.

To date, the workflows that we have considered have really only focused on the provenance data/metadata and inferencing rules associated with processing events in a laboratory or manufacturing/processing plant. We need to extend the underlying model and the inferencing rules to support the data processing activities in the digital domain e.g., reformatting, segmentation, normalization etc.

Currently the XACML access policies we use are defined manually and are manually associated with relationships between nodes in the RDF graphs. This is a relatively time-consuming process. We need to determine a more streamlined mechanism for defining access policies and associating them with provenance relationships. For example, the individual or type of participant who is responsible for a particular activity or set of activities should automatically have access to the provenance data associated with those activities and any sub-activities.

Another limitation of the current system is that it currently only supports expansion down one level of detail. Ideally users would be able to incrementally drill down to multiple levels of detail. For example one link can be expanded to two links, each of which can be further expanded. This may prove quite complex to implement because it involves multiple levels of inferencing rules and the specification of access policies associated with relationships at multiple levels.

Finally the packages of components that are able to be constructed provide a very efficient mechanism: for publishing and sharing scientific results; for teaching complex scientific concepts; and for the selective archival, curation and preservation of scientific data. Although we currently enable these packages to be saved, they are not indexed or able to be searched and retrieved. Tools are required to enable these RDF packages to be described, stored to institutional repositories and searched and retrieved for reuse.

6.3 Conclusions

In this paper, we have described the Provenance Explorer system that we have developed. It is a provenance visualization system that dynamically generates different graphical views of provenance trails depending on the user's requirements and access privileges. It enables users to search and retrieve the data provenance associated with scientific workflows or experiments, without compromising the security of the data. Even within the context of workflows that capture and share data across institutional boundaries, the system is able to authenticate users to enforce fine-grained, role-based access controls. The hypermedia user interface that we have developed enables easy drilling down from simple high-level views to detailed views of complex sub-activities by enabling links to be expanded or collapsed. This feature was easy to implement and can quickly be refined or customized because it is implemented using SWRL rules and the Algernon inferencing engine.

Finally scientists are under increasing pressure from funding organizations to publish their experimental and evidential data together with the related traditional scholarly publication(s). This system makes it easy for scientists to wrap related outputs into a single package for publication, peer-review, e-learning or selective preservation purposes – and to have the provenance trail between the components automatically inferred to enable validation and verification.

7 References

1. Goble, C. *Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics*. in *Workshop on Data Derivation and Provenance*. 2002.
2. schraefel, m.c., et al. *Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment*. in *CHI04*. 2004. Vienna, Austria.
3. Gibson, A., et al. *myTea: Connecting the Web to Digital Science on the Desktop*. in *World Wide Web Conference*. 2006. Edinburgh.

4. Altintas, I., O. Barney, and E. Jaeger-Frank. *Provenance Collection Support in the Kepler Scientific Workflow System*. in *International Provenance and Annotation Workship (IPAW'06)*. 2006. Chicago, Illinois, USA.
5. Oinn, T., et al., *Taverna: A tool for the composition and enactment of bioinformatics workflows.* Bioinformatics Journa, 2004. 20(17): p. 3045-3054.
6. Majithia, S., et al. *Triana: A Graphical Web Service Composition and Execution Toolkit*. in *IEEE International Conference on Web Services (ICWS'04)*. 2004: IEEE Computer Society.
7. Morgan, R.L.B., et al., *Federated Security: The Shibboleth Approach.* EDUCAUSE QUARTERLY, 2004. 4: p. 12 - 17.
8. Lorch, M., et al. *First Experiences Using XACML for Access Control in Distributed Systems*. in *ACM Workshop on XML Security*. 2003. Fairfax, Virginia.
9. Bose, R. and J. Frew. *Composing lineage metadata with XML for custom satellite-derived data products*. in *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*. 2004.
10. Myers, J.D., Pancerella, C., Lansing, C., Schuchardt, K.L. & Didier, B. *Multi-scale science: supporting emerging practice with semantically derived provenance*. in *ISWC 2003 Workshop: Semantic Web Technologies for Searching and Retrieving Scientific Data,*. 2003. Sanibel Island, Florida, USA.
11. Zhao, J., et al. *Using Semantic Web Technologies for Representing E-science Provenance*. in *Third International Semantic Web Conference*. 2004. Hiroshima, Japan.
12. Freire, J., et al. *Managing Rapidly-Evolving Scientific Workflows*. in *International Provenance and Annotation Workship (IPAW'06)*. 2006. Chicago, Illinois, USA.
13. Hunter, J. and K. Cheung. *Generating eScience Workflows from Statistical Analysis of Prior Data*. in *APAC'05*. 2005. Royal Pines Resort, Gold Coast.
14. Hunter, J., J. Drennan, and S. Little, *Realizing the Hydrogen Economy through Semantic Web Technologies.* IEEE Intelligent Systems Journal - Special Issue on eScience, 2004: p. 40-47.
15. Lagoze, C. and J. Hunter, *The ABC Ontology and Model*. Journal of Digital Information, 2001. 2(2).
16. Horrocks, I., et al., *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. 2004.
17. Carroll, J.J., et al., *Jena: implementing the semantic web recommendations*, in *Proceedings of the 13th international World Wide Web conference on Alternate track papers \& posters*. 2004, ACM Press: New York, NY, USA. p. 74-83.
18. McCarthy, P., *Search RDF data with SPARQL: SPARQL and the Jena Toolkit open up the semantic Web*, in *developerWorks*. 2005, IBM.
19. Crawford, J.M. and B.J. Kuipers, *Algernon - a tractable system for knowledge-representation.* SIGART Bull., 1991. 2(3): p. 35-44.
20. Alder, G., *The JGraph Swing Component*, in *Department of Computer Science*. 2002, Federal Institute of Technology ETH: Zurich, Switzerland.
21. Gangemi, A., et al. *Sweetening Ontologies with DOLCE*. in *13th International Conference on Knowledge Engineering and Knowledge Management*. 2002. Siguenza, Spain.
22. Weber, R., *Ontological Foundations of Information Systems*. 1997, Melbourne: Coopers & Lybrand Accounting Research Methodology.
23. Colomb, R.M., *Formal versus Material Ontologies for information Systems interoperation in the Semantic Web.* The Computer Journal, 2006. 49(1).