Towards Recovering Provenance with Experiment Explorer

Delmar B. Davis, Hazeline U. Asuncion
Computing and Software Systems
University of Washington, Bothell
Bothell, WA, USA
{davisdb1, hazeline}@u.washington.edu

Ghaleb M. Abdulla, Christopher W. Carr
Lawrence Livermore National Laboratory
Livermore, CA, USA
{abdulla1, carr19}@llnl.gov

***Abstract*—In this work, we present Experiment Explorer (EE), a framework for recovering provenance that uses provenance-compatible research processes and a lightweight, user-friendly metadata search tool. EE also captures recovered provenance along with file relationships and incorporates new files to support provenance recovery over time. Our case study at a research laboratory suggests that EE is effective in connecting distributed provenance information and in increasing the accessibility of related experiment files. Our scalability analysis also indicates that EE's tool support can scale to hundreds of thousands of heterogeneous files.**

Keywords-data provenance; metadata; provenance recovery; information management.

I. Introduction

More scientific research now involves the generation and analysis of heterogeneous data sets. Multiple instruments, analysis tools, and scripts are used to collect and analyze data. The origin of a data set or the processing applied to a data set is referred to as data provenance. Data provenance is necessary in assessing a data set's integrity and in supporting repeatability of analyses or experiments. However, obtaining provenance is a difficult task, especially for data sets that have already been reduced or aggregated from their raw form—some of the context of the data may have been lost. Data provenance may be obtained from experimental conditions and data processing or reduction. Deficiencies in data provenance related to experimental conditions may arise from mundane technical reasons (e.g., a loose cable) or fundamental reasons (e.g., not monitoring a critical parameter because its importance is not known at the time of the experiment). For example, the growth of sites exposed to laser was studied for many decades before it was discovered that the temporal pulse shape was important. In experiments conducted prior to this discovery, only the duration of the pulse was recorded.

Current provenance techniques often capture provenance *during* an analysis or experiment run by logging the steps that were followed or the analysis modules that were invoked [1, 2, 3]. Many of these techniques use scientific workflows where researchers pre-specify the experiment design as a workflow and a log of the workflow execution provides the provenance of the data produced at the end of the execution [1, 4]. Other techniques record commands or events while the dataset is being processed [2, 3, 5] .

To support the use case of recovering provenance where processing logs may not be available, we present Experiment Explorer (EE). EE allows users to recover provenance by incorporating provenance-compatible research processes and by enabling researchers to search for experiment-related information, possibly scattered across heterogeneously-represented files, that provide clues to the provenance of a data set. EE leverages a lightweight and user-friendly metadata search tool to aid researchers in uncovering provenance, which requires minimal training time and usage overhead from them. EE also allows researchers to capture the recovered provenance and to incrementally support provenance recovery over time. Moreover, EE can be used in conjunction with recording provenance techniques in cases where incomplete provenance has been captured (e.g., recording took place only in some parts of the entire data processing). We previously introduced EE as a metadata provenance search [6]. In this paper, we elaborate on how EE can be used to support the recovery of data provenance.

The contributions of this paper are as follows: 1) a technique for recovering provenance, 2) a means of increasing the accessibility of related experiment files, and 3) a means of capturing recovered provenance and file relationships for future reference. In addition, our case study at a research lab suggests the effectiveness of our technique in locating distributed provenance information and in raising the visibility of relevant experiment files.

The rest of the paper is organized as follows. In the next section, we compare our work with existing techniques. Section 3 covers challenges in providing provenance support to research in general, with an example of support for the optics inspection and analysis group at Lawrence Livermore National Laboratory (LLNL). We then introduce our approach in Section 4. We provide details regarding our tool support in Section 5 and discuss evaluation results in Section 6. We close the paper with future work.

II. Related Work

Provenance techniques generally record provenance as the data is being processed. These techniques are often associated with workflows [1, 4, 7]. Other recording techniques include capturing using interactions [2, 5] or listening to system-level events [3]. Different levels of provenance may also be recorded [8]. Recovering provenance complements both prospective provenance, i.e., specification of the process as in a workflow, and retrospective provenance, e.g., log of execution [9].

One key aspect to EE is the crafting of research processes that are provenance compatible. Others have also suggested

ISBN: 978-1-61208-254-7

process-centered approaches that revolve around the use of services and tools [10] or around collaboration steps [4]. Another technique collects provenance based on coordination points in distributed enterprise workflows [11].

Metadata has also been used to represent or link different types of data. These links between different types of data may be represented as a Resource Description Framework (RDF) graph [12]. Provenance metadata may also be represented as an RDF graph that can be queried [13]. Discovery techniques exist for discovering metadata via recommender systems [14] or discovering data using metadata [15]. Data provenance which has been registered to a service may also be later discovered [8]. Tools that capture metadata include XMC-Cat [16] and Taverna [17]. XMC-Cat relies on workflow cyber infrastructure to capture metadata as associated data is generated [16]. Taverna captures metadata by observing processing units and it imports metadata from existing entities [17]. In EE, we use metadata to search for experiments and their corresponding artifacts. EE can also generate metadata for experiment files.

There are also various techniques for searching for documents. These involve using string matching techniques [18] or incorporating user activity into the search for documents [19]. These techniques fall short of searching for files that do not have a text representation (e.g., image files). Other techniques use a database [20] or map-reduce [21] to increase the efficiency. As we demonstrate in our scalability analysis, EE's search tool is highly scalable. Document management systems like Placeless Documents allows users to search for documents according to metadata, such as file properties (e.g., file size = 100MB) or user-defined properties (e.g., priority = high) [22]. Instead of users specifying properties on a file-by-file basis, EE uses the information embedded in the directory structure and the provenance template to automatically assign provenance metadata to files. Configuration management systems also allow users to search for files based on commit records or change entries [23]. EE, on the other hand, allows users to search for files based on their relationship to an experiment or based on an experiment attribute.

Recovering provenance has been discussed outside the field of eScience. One technique discusses how provenance can be recovered from executable software [24].

III. Motivation

The utility of scientific data is ultimately limited by its provenance. Aggregating data into sets requires that the provenance of the individual data points is compatible. When data are collected at the same time, compatibility is typically ensured by good experimental hygiene, even if critical aspects of provenance are not documented, or even not known to exist as it is the same for each datum. However if data from different experiments or different experimenters are to be combined, then it becomes critical to ensure the provenance is sufficient and compatible. We now provide an overview of challenges encountered in scientific research which effective data provenance can assist.

Fast-paced and adaptive research environment: Research efforts in highly competitive fields or in support of a larger project must be highly adaptive. Though most programs have well defined deliverables with due dates many months or years in the future, situations emerge from time to time which must be addressed immediately. A researcher often responds to new project needs or publication of related discoveries by repurposing existing data and analysis. Such adaptability is greatly facilitated by provenance techniques that are lightweight, with minimal setup and training time. In addition to short term priority shifts, research labs are also highly adaptive to advances in current technology and changes to tools. Using new architectures and tools requires documenting the old and new environments as part of the metadata.

Numerous heterogeneous experiment files: Each experiment run produces hundreds of files and multiple experiments are conducted for each hypothesis. Individual researchers working on related topics conduct their own experiments with techniques and analysis algorithms optimized to isolate particular experimental parameters. Thus, the number and type of experiment files quickly grows, making it difficult to manually search through files. Using off-the-shelf search tool for some of these formats, such as binary or image files, is inadequate because these tools are designed to search across distributed and unrelated files and there is no utilization of the relationships between the different files and experimental objects.

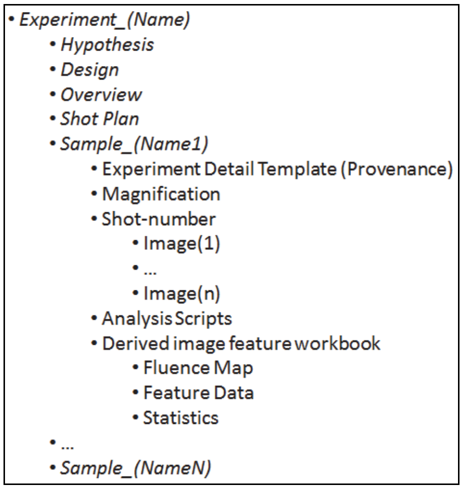
Accessing experiment files: Often researchers would like to amalgamate data from several data sets to create and test hypotheses. Compiling data sets without the aid of a provenance recovery tool requires researchers to search manually for the appropriate files. If the researchers are fortunate enough to be looking for an attribute that was used to categorize the experiment, they are able to gather a large data set and only incur the time it took to locate the data file. Otherwise, the experimental details must be manually retrieved from lab books, scripts, or correlated experiment files, which can be a time-consuming process.

IV. Approach & Application

To address the challenges discussed in the previous section, we present EE and apply it to the specific case of correlating laser induced damage data generated in a number of labs for optics inspection and analysis group at LLNL. Provenance recovery is a technique that approximates the provenance of a data set based on available information (e.g., researcher notes, scripts, hypotheses). EE facilitates the recovery of provenance by incorporating provenance-compatible research processes and enabling researchers to piece together distributed provenance information. While the provenance is an approximation, it may be "close-enough" for researchers to fill-in the missing gaps. Once provenance has been recovered, it can be captured for future reference. Additionally, EE supports incremental provenance recovery over time.

A. Incorporating Provenance-Compatible Research Practices and Conventions

The analysis process, which is part of the overall experiment process, is a collaborative effort, requiring the



- *Experiment_(Name)*
  - *Hypothesis*
  - *Design*
  - *Overview*
  - *Shot Plan*
  - *Sample_(Name1)*
    - Experiment Detail Template (Provenance)
    - Magnification
    - Shot-number
      - Image(1)
      - ...
      - Image(n)
    - Analysis Scripts
    - Derived image feature workbook
      - Fluence Map
      - Feature Data
      - Statistics
  - ...
  - *Sample_(NameN)*

Figure 1. Hierarchical relationship between experiment files

contributions of material science specialists, data analysts, and physicists. The workflows as well as the set of artifacts produced remain consistent across experiments.

Figure 1 shows the artifacts produced over the course of an experiment. (Note: an artifact is any file produced during a research process.) The process begins with a hypothesis that is used to derive an experiment design. Along with the design, a detail template, or provenance template, is created providing a central overview of the experiment attributes as data is analyzed. As scientists move through the workflow, raw experiment data are produced, labeled as image in the figure. When the images are analyzed, a derived image feature workbook is created, which contains information such as fluence map.

As part of the process, this provenance template is filled-in by researchers. Some of this information is obtained from the workflow and hypotheses, while others are derived from analysis. Attributes, which may be experiment design parameters (e.g., average fluence or average site separation), are also recorded in a provenance template.

All the artifacts produced for an experiment are stored in a folder labeled with the experiment sample name, to support provenance recovery. For example, hundreds or thousands of image files can result from one experiment. Each file is named in a way that captures metadata such as location of the damage image. In addition, a metadata file is generated for each experiment file, which includes author, date, experiment sample name, keywords, and category. This metadata provides a connection between each file and the associated experiment sample or among related files based on provenance-specific fields. Because the provenance template is stored at the top level sample name folder, the keywords, category, author, and date can be automatically extracted from the provenance template. This template can be manually used to locate experiments performed at certain dates, by specific scientists, or used fluence ranges between X and Y, etc.

B. Using a Lightweight Metadata Search

Once the metadata is created for every experiment file, researchers may now proceed with recovering provenance using a metadata search tool. First, we index the provenance template and the metadata associated with each file. The provenance template is an Excel workbook containing an experiment summary sheet, a high level data overview sheet, and the server location for the experiment sample. In order to search for experiment artifacts that span the entire data server, we first provide a means of relating the files. As we mentioned, all the experiment artifacts are stored within a folder with the experiment sample name. Thus, one way files can be related is by examining the path of an experiment file to obtain the experiment sample file name. Once the sample file name is obtained, the metadata for each file can include the experiment sample name. This metadata can then be re-indexed.

To address the challenge of searching through heterogeneous artifacts, we index different types of data using artifact-specific indexing components. We use artifact-specific indexing components to extract the metadata. The metadata follows a uniform format, enabling the metadata from various artifact types to be indexed in a similar fashion. Consider the following example:

Shot plans are documents requesting a particular set of laser exposures including the sample to be exposed, the location on the sample, and the number and type of laser exposures. A new file is detected by the indexing component as a result of comparing the past directory structure with the current directory structure. The artifact

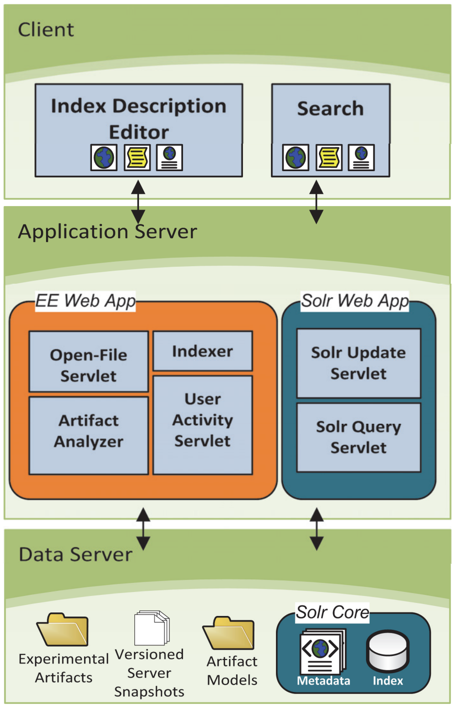


Figure 2. Architecture of EE's search tool

analyzer consults a set of models that reflect the conventions of the researchers. The file location, file naming pattern, extension, and structure match those of a shot plan.

The analyzer extracts the experiment name and the indexer searches for it. In this case, one result is expected because a shot plan file belongs to only one experiment. Upon success a link to this file is added to the experiment index document and the experiment is re-indexed.

We can also obtain the provenance of derived features. As the derived feature image workbook is modified on the server, it is re-indexed as described in our example above. By utilizing the file hierarchy shown in Figure 1, features from the workbook can be related to the experiment, which can then be related to the provenance template.

C. *Capturing Recovered Relationships and Provenance*

After the metadata for various files has been generated and indexed, researchers may search for experiment files for a given experiment sample name or a given attribute or even files created on the same date as the experiment. When search results are returned, researchers may directly access the experiment file by following a link from EE's search result, making the relevant files accessible to researchers.

In addition, since the search and access to files are integrated, it becomes straightforward to capture the search terms used and the files opened by users. The search terms can be captured to log how data sets are correlated. Capturing the correlations and the provenance of experiment files allows researchers to reflect upon their past search activities. This record can help them redo their past search activities or identify unexamined correlations.

D. *Supporting Provenance Recovery Over Time*

As time goes on, new hardware and software tools may be used to process the data. EE can accommodate these changes over time since it depends on the research processes and the metadata generated for each file, not the technology used for processing or analyzing data. New software tools may also produce new file formats that EE must accommodate. This can be handled by EE by creating a new component to index the new file format.

Over time, new experiment files will also be added into the file system. A version control system [25] can be used to track new files within the server folder. This way, new or modified artifacts can be indexed.

V. TOOL SUPPORT

We now describe the design and current implementation.

A. *Recovering Provenance with EE*

In order to match the distributed system and collaborative work within the optics inspection and analysis group, Experiment Explorer is comprised of components following a 3-tier client server architecture, as shown in Figure 2. (Note: an earlier version of this design was described in [6]). Two web applications are integrated with the data server to form a lightweight metadata search tool. The data server contains experimental artifacts and metadata describing them, as well as an index of the metadata and a set of version controlled server snapshots. Experimental artifacts are found on the lab server, but accessible by the components shown in the application server. The metadata that describes these artifacts are stored as XML documents formatted as Apache Solr update instructions [26]. Fields within the metadata documents are recorded to match the Solr schema, enabling directed indexing. The Solr core folder contains the index as well the metadata. The version controlled snapshots are used to detect changes on the server and trigger indexing of new or modified experiment associated files.

At the client, two pages allow users to manually index and search for experiments. In order to maintain extensibility and adapt to the addition of new types of experiment related files, both pages are constructed dynamically based on indexed fields. When new fields are added to the index they become available for user input. Currently, those fields corresponding to an experiment overview are the only ones indexed, as shown in Figure 3. In order to keep newly added fields from overrunning the page, a selection drop down will provide access to the additional fields.

When a user clicks the search button, scripts gather the field-specific input and format a query that is posted to the Solr query servlet of the application server shown in Figure 2. Results are indicated by the fields that best differentiate experiments from one another. Some of the fields point to files located on the server. These server locations are formatted into links. When clicked, they call a servlet responsible for delivering the specified file to the client.

The index description editor allows researchers to manually index artifacts. Many of the functions found on the search page are also used in the editor. Additionally the editor is able to load fields from an indexed experiment and those found in an experiment overview located on the server.

Components in the application server represent components called by the client and an indexing component that responds to changes on the data server. As mentioned above, EE integrates two web applications. The first, Apache Solr, is used to manipulate and access the index. The second is composed of the remaining java servlets, which are responsible for manipulating and accessing the data server. One exception to this is the servlet responsible for logging user actions, such as queries issued and files requested. A record of files opened from links in the client is used to direct priority in the indexer.

The indexer is used to find and index new or modified artifacts on the server and is called at regular intervals. To handle frequent changes to the server, files accessed from the client are given the highest priority when responding to differences between versions of the server snapshots. In addition, a queue of outstanding un-indexed files is maintained in order to throttle server access by the indexer. This queue is produced by making comparisons between the different version controlled server snapshots found in the data server area of Figure 2. The server snapshots do not contain actual files but file system information starting at the highest level folder designated for experiment files.

As we discussed, the process of indexing new artifacts relies on conventions followed by the researchers in formatting and naming data on the server. These conventions are specified by the models in the data server area of Figure 2. The models are used by the indexer to link the artifact to the experiment overview by sample name.

B. *State of Implementation*

Currently, EE's search tool supports the following: searching for experiment metadata (e.g., provenance), linking experiment files included in the search results, and capturing search terms and files opened by researchers. We plan to implement the following functionality: relating files to the experiment sample name with an indexer, displaying captured relationships between experiments' files based on artifact models, and integrating a version control system to automate the capture of future artifact metadata.

VI. Evaluation

We now discuss evaluation of EE's search tool through a case study with a group at LLNL and a scalability analysis.

A. *Case Study at the LLNL*

An evaluation was performed within the optics inspection and analysis group at LLNL. Five subjects participated in the study, including scientists and data analysts. The files indexed by EE were spreadsheets which contain the overview of the various experiments conducted in the lab and pointers to the locations of the various files. For the study, only a subset of the overview files was indexed in EE. Prior to the study, subjects were provided with training, including documentation and video tutorials on using the tool. During the study, subjects were asked to perform a search task that they might perform while conducting their research. Feedback was obtained via interviews.

We sought answers to the following research questions:

Q1: Is EE's search tool easy to use? Can it be incorporated into the research process at LLNL?

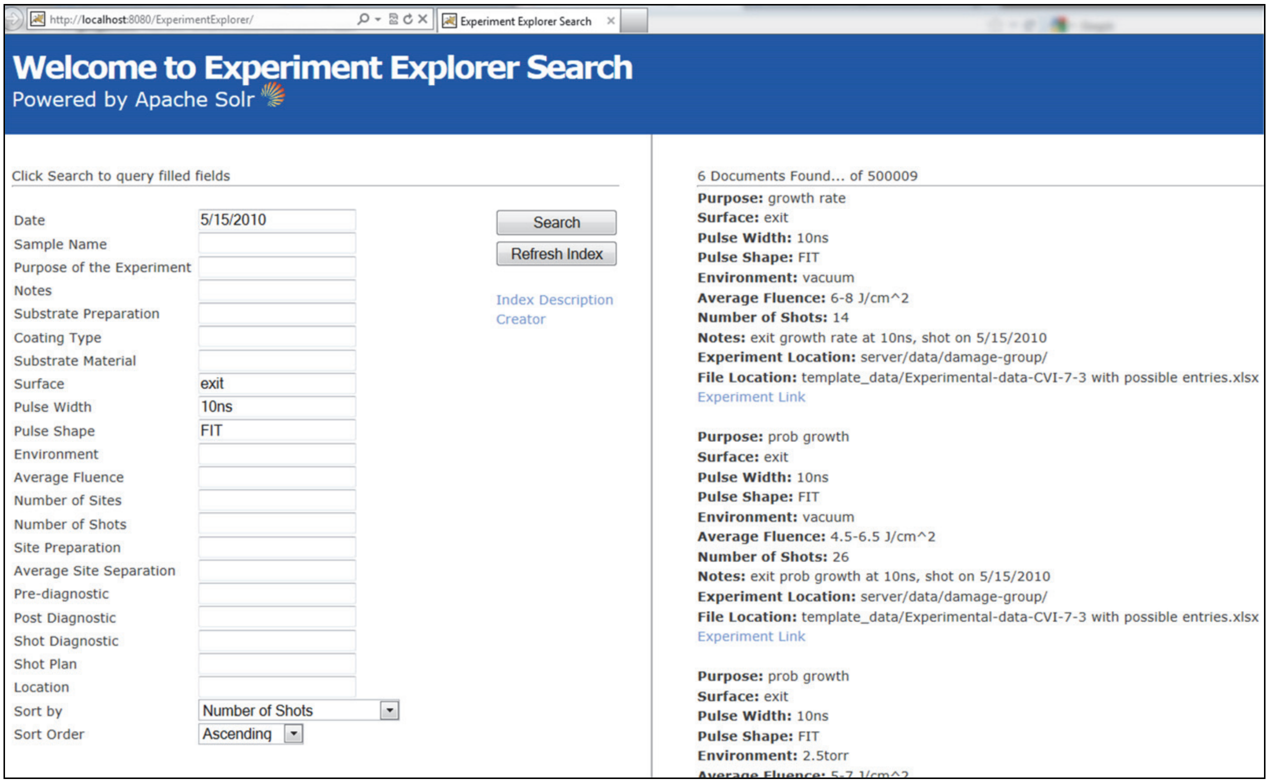
Q2: Does EE's search tool provide relevant experiment files?

Q3: Does EE enable researchers to determine which experiments were related to which files?

Q1: Subjects were asked to compare EE's search tool with previous search techniques. Subjects were asked to rate EE on a scale of 1 to 5, with 1 as exceptional and 5 as unacceptable. On average researchers rated the acceptability of time spent learning the software at 1.6, the time spent using the software at 1, and the time spent finding relevant files at 1.2. Three subjects even pointed out that the feature they like about the tool is ease of use.

The subjects were generally pleased with EE's search capability. Three of the five subjects said that they would use the tool to perform their research, while one user said that he would use the tool if his suggested changes were incorporated. Previous search techniques involved manually traversing folders on a server or asking a colleague regarding the location of files. Since there are numerous files, these previous techniques were time-consuming.

Q2: On the average, subjects found the experiment they were searching 84% of the time. When asked how often links to the actual overview files were missing, they answered 0%.


http://localhost:8080/ExperimentExplorer/
Experiment Explorer Search
Welcome to Experiment Explorer Search
Powered by Apache Solr
Search
Refresh Index
Index Description
Creator

Click Search to query filled fields

Date	5/15/2010
Sample Name	
Purpose of the Experiment	
Notes	
Substrate Preparation	
Coating Type	
Substrate Material	
Surface	exit
Pulse Width	10ns
Pulse Shape	FIT
Environment	
Average Fluence	
Number of Sites	
Number of Shots	
Site Preparation	
Average Site Separation	
Pre-diagnostic	
Post Diagnostic	
Shot Diagnostic	
Shot Plan	
Location	
Sort by	Number of Shots
Sort Order	Ascending

6 Documents Found... of 500009

Purpose: growth rate
Surface: exit
Pulse Width: 10ns
Pulse Shape: FIT
Environment: vacuum
Average Fluence: 6-8 J/cm^2
Number of Shots: 14
Notes: exit growth rate at 10ns, shot on 5/15/2010
Experiment Location: server/data/damage-group/
File Location: template_data/Experimental-data-CVI-7-3 with possible entries.xlsx
Experiment Link

Purpose: prob growth
Surface: exit
Pulse Width: 10ns
Pulse Shape: FIT
Environment: vacuum
Average Fluence: 4.5-6.5 J/cm^2
Number of Shots: 26
Notes: exit prob growth at 10ns, shot on 5/15/2010
Experiment Location: server/data/damage-group/
File Location: template_data/Experimental-data-CVI-7-3 with possible entries.xlsx
Experiment Link

Purpose: prob growth
Surface: exit
Pulse Width: 10ns
Pulse Shape: FIT
Environment: 2.5torr
Average Fluence: 5-7 J/cm^2

Figure 3. Experiment Explorer Search Tool

Q3: While the entire experiment artifacts were not available for the study, the overview files which were indexed, contains pointers to the experiment artifacts. According to one subject, once she obtains the experiment overview from EE's search tool, she can find a given experiment artifact at least twice as fast without the tool.

All of the interviewees agreed that the experiment overview was important information regardless of what type of exploration they performed. In terms of direct links from the results page, three of the five researchers felt that the addition of the data overview for an experiment would support a more efficient exploration. The fifth wanted all of the artifact links embedded directly into the overview file.

Discussion: In general the researchers were pleased with EE's tool support. One of the researchers said, "Like this idea of organizing and indexing my experiments. Productivity and sharing is much easier this way." Another researcher said, "It was very fast and the fields were relevant to the searches that we normally make." In addition, the results show that the tool is easy to use (with average ratings of less than 2.0). The results also indicate that all (except for one subject) would use the tool. This suggests that the tool would benefit the fast-paced adaptive research environment of the optics inspection and analysis group at LLNL.

Areas for improvement within the search software included support for range queries, richer logical support for interpreting user input, and limiting possible input for fields that have a small set of possible field data. All of the researchers expected to have the system show a preview of the most likely input as they typed, indicating that support for fast incremental exploration be supported directly from the search page. The users would also like to be able to directly link to the experiment files once an experiment sample name has been obtained from EE.

B. *Scalability Analysis*

A scalability analysis was also performed. Over five hundred thousand metadata documents were created and indexed, twenty five thousand documents at a time, on a machine with an i7 processor and 6 GB of RAM. Indexing time at each step took around 5 minutes. However, even at five hundred thousand documents, the search still performed in near real-time, taking 40-70ms to deliver results. Restarting the server increased the search time to 120ms and leveled back down to 40-70ms after 3 searches. Searching with two integer fields and a date field resulted in an average of 12 documents being returned from over 500,000 possibilities, which are the correct documents. This was done over 20 runs with different known numbers.

VII. Conclusion and Future Work

In this paper, we provided a technique for recovering provenance for data sets that have already been analyzed or processed. This technique, referred to as EE, incorporates a provenance-compatible process with a lightweight metadata search tool. EE complements existing techniques which record provenance while a data set is being analyzed. We conducted two types of evaluation to assess EE's tool support: a case study at LLNL and a scalability analysis. The case study suggests that EE is effective in connecting information which provides clues regarding the provenance of a given experiment file. The scalability analysis reveals that the tool can easily handle large sets of experiment files.

In the future, we will investigate incorporating ontologies into the metadata generation and search to enable sharing research files with researchers outside the optics inspection and analysis group. We plan to provide a visualization to help researchers more efficiently find related data sets. We will also examine other automated techniques for determining the related experiment sample name for a given experiment file. Finally, we will investigate how EE can be applied to other settings, such as recovering the provenance of software development files or the provenance of business reports.

Acknowledgment

We thank David Cross, Zhi Liao, Raluca Negres, Mary Norton, and Rajesh Raman at LLNL for valuable insights. This work is partly funded by University of Washington's Royalty Research Fund No. A65951.

References

[1] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance collection support in the Kepler Scientific Workflow System," in *Proc of Int'l Provenance and Annotation Workshop (IPAW)*, 2006.

[2] D. Bourilkov, "The CAVES project - Collaborative Analysis Versioning Environment System, the CODESH project — COllaborative DEvelopment SHell," *International Journal of Modern Physics*, vol. A20, no. 16, pp. 3889–3892, 2005.

[3] D. A. Holland, M. I. Seltzer, U. Braun, and K.-K. Muniswamy-Reddy, "PASSing the provenance challenge," *Concurrency and Computation: Practice and Experience*, vol. 20, pp. 531–540, 2008.

[4] P. Missier, B. Ludascher, S. Bowers, S. Dey, A. Sarkar, B. Shrestha, I. Altintas, M. Anand, and C. Goble, "Linking multiple workflow provenance traces for interoperable collaborative science," in *Workshop on Workflows in Support of Large-Scale Science*, 2010.

[5] H. U. Asuncion, "*In Situ Data* provenance capture in spreadsheets," in *Proc of the Int'l Conf on e-Science*, 2011.

[6] D. B. Davis, H. U. Asuncion, and G. Abdulla, "Experiment explorer: Lightweight provenance search over metadata," in *Proc of the USENIX Workshop on the Theory and Practice of Provenance*, 2012.

[7] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo, "Managing rapidly-evolving scientific workflows," in *Proc of the IPAW*, 2006.

[8] P. Yue, J. Gong, L. Di, L. He, and Y. Wei, "Semantic provenance registration and discovery using geospatial catalogue service," in *Proc of the Int'l Workshop on the Role of Semantic Web in Provenance Mgmt*, 2010.

[9] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," in *Proc of Int'l Conf on Mgmt of Data*, 2008.

[10] L. Chen, X. Yang, and F. Tao, "A semantic web service based approach for augmented provenance," in *Proc of Int'l Conf on Web Intelligence*, 2006.

[11] M. Allen, A. Chapman, B. Blaustein, and L. Seligman, "Capturing provenance in the wild," in *Provenance and Annotation of Data and Processes*, vol. 6378, pp. 98–101, Springer Berlin / Heidelberg, 2010.

[12] U. Marjit, K. Sharma, and U. Biswas, "Provenance representation and storage techniques in linked data: A state-of-the-art survey," *Int'l Journal of Computer Applications*, vol. 38, no. 9, pp. 23–28, 2012.

[13] A. Chebotko, X. Fei, C. Lin, S. Lu, and F. Fotouhi, "Storing and querying scientific workflow provenance metadata using an RDBMS," in *Proc of the Int'l Conf on e-Science and Computing Grid*, 2007.

[14] M. S. Aktas, M. Pierce, G. C. Fox, and D. Leake, "A web based conversational case-based recommender system for ontology aided metadata discovery," in *Proc of the Int'l Workshop on Grid Computing*, 2004.

[15] S. M. S. Da Cruz, P. M. Barros, P. M. Bisch, M. L. M. Campos, and M. Mattoso, "A provenance-based approach to resource discovery in distributed molecular dynamics workflows," in *Proc of Int'l Conf on Resource Discovery*, 2010.

[16] S. Jensen and B. Plale, "Trading consistency for scalability in scientific metadata," in *Proc of Int'l Conf on e-Science*, 2010.

[17] K. Belhajjame, K. Wolstencroft, O. Corcho, T. Oinn, F. Tanoh, A. William, and C. Goble, "Metadata management in the Taverna workflow system," in *Int'l Symposium on Cluster Computing and the Grid*, 2008.

[18] C. Duda, D. Kossmann, and C. Zhou, "Predicate-based indexing for desktop search," *VLDB Journal*, vol. 19, no. 5, pp. 735–758, 2010.

[19] J. Chen, H. Guo, W. Wu, and W. Wang, "iMecho: an associative memory based desktop search system," in *Proc of Conf on Information and Knowledge Mgmt*, 2009.

[20] S. M. S. Da Cruz, P. M. Barros, P. M. Bisch, M. L. M. Campos, and M. Mattoso, "A provenance approach to trace scientific experiments on a grid infrastructure," in *Proc of Int'l Conf on e-Science*, 2011.

[21] E. Dede, Z. Fadika, C. Gupta, and M. Govindaraju, "Scalable and distributed processing of scientific XML data," in *Proc of Int'l Conf on Grid Computing*, 2011.

[22] P. Dourish, W. K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. B. Terry, and J. Thornton, "Extending document management systems with user-specific active properties," *ACM Transactions on Information Systems*, vol. 18, no. 2, pp. 140–170, 2000.

[23] "Git." http://git-scm.com/, Accessed Dec 13, 2012.

[24] N. Rosenblum, B. P. Miller, and X. Zhu, "Recovering the toolchain provenance of binary code," in *Proc of Int'l Symposium on Software Testing and Analysis*, 2011.

[25] J. Estublier, D. B. Leblang, G. Clemm, R. Conradi, A. van der Hoek, W. Tichy, and D. Wiborg-Weber, "Impact of the research community on the field of software configuration management," *Trans on Software Engineering Methodology (TOSEM)*, vol. 14, no. 4, pp. 383–430, 2005.

[26] Apache Software Foundation, "Apache Solr." http://lucene.apache.org/solr/, Accessed Dec 13, 2012.