

# Probe-It! Visualization Support for Provenance

Nicholas Del Rio and Paulo Pinheiro da Silva

University of Texas at El Paso  
500 W. University Ave, El Paso, Texas, USA

**Abstract.** Visualization is a technique used to facilitate the understanding of scientific results such as large data sets and maps. Provenance techniques can also aid in increasing the understanding and acceptance of scientific results by providing access to information about the sources and methods which were used to derive them. Visualization and provenance techniques, although rarely used in combination, may further increase scientists' understanding of results since the scientists may be able to use a single tool to see and evaluate result derivation processes including any final or partial result. In this paper we introduce Probe-It!: a visualization tool for scientific provenance information that enables scientists to move the visualization focus from intermediate and final results to provenance back and forth. To evaluate the benefits of Probe-It!, in the context of maps, this paper presents a quantitative user study on how the tool was used by scientists to discriminate between quality results and results with known imperfections. The study demonstrates that only a very small percentage of the scientists tested can identify imperfections using maps without the help of knowledge provenance and that most scientists, whether GIS experts, subject matter experts (i.e., experts on gravity data maps) or not, can identify and explain several kinds of map imperfections when using maps together with knowledge provenance visualization.

## 1 Introduction

In complex virtual environments like cyber-infrastructure, scientists rely on visualization tools to help them understand large amounts of data that are generated from experiments, measurements obtained by sensors, or a combination of measurements and applied derivations. Instead of tediously tracing through datasets, scientists view results condensed as a graph or map, and draw conclusions from these projected views. However, in order for scientists to fully understand and accept artifacts generated on the cyber-infrastructure, scientists may need to know which data sources and data processing services were used to derive the results and which intermediate datasets were produced during the derivation process. In fact, scientists may need to have access to provenance information, which in this paper is described as meta-information about the final results and how they were generated. Provenance information includes: *provenance meta-information*, which is a description of the origin of a piece of knowledge (e.g., names of organizations, people, and software agents who played some role in the generation of an artifact), *process meta-information*, which is a description of

the reasoning process used to generate an answer, such as a proof or execution trace, and *intermediate* or partial results.

Provenance visualization capabilities are expected to be more sophisticated than those required for the visualization of only final results. For example, in addition to the visualization of results, provenance visualization should include capabilities for visualizing intermediate or partial results, derivation processes, and any information regarding used sources. In this paper, we report our progress on Probe-It!, a general-purpose, provenance visualization prototype that has been used to visualize both logical proofs generated by inference engines and workflow execution traces. Additionally, the paper reports on an ongoing user study that confirms the need for provenance information in the tasks of identifying and explaining imperfections in maps generated by cyber-infrastructure applications.

## 2 Scientific Knowledge Provenance Visualization

Probe-It! is a browser suited to graphically rendering provenance information associated with results coming from inference engines and workflows. In this sense, Probe-It! does not actually generate content (i.e. logging or capturing provenance information); instead it is assumed that users will provide Probe-It! with end-points of existing provenance resources to be viewed. The task of presenting provenance in a useful manner is difficult in comparison to the task of collecting provenance. Because provenance associated with results from small workflows can become large and incomprehensible as a whole, Probe-It! consists of a multitude of viewers, each suited to different elements of provenance. Decomposing provenance into smaller more comprehensible chunks, however, raises the following questions:

1. How do scientists navigate back and forth between the visualizations of final and intermediate results (i.e., datasets and scientific artifacts such as maps) and information about the generation of such results (i.e., meta-data about the applied sources, methods, and sequencing regarding the execution of those methods)?
2. How do scientists define relevance criteria for distinct provenance information and how can tools use relevance criteria to improve scientist experiences during the visualization of scientific provenance?
3. How can scientists instruct tools to present scientific provenance by defining and selecting preferences?

The following sections describe how Probe-It! addresses these concerns.

### 2.1 Queries, Results, Justifications, and Provenance

Probe-It! consists of four primary views to accommodate the different kinds of provenance information: queries, results, justifications, and provenance, which refer to user queries or requests, final and intermediate data, descriptions of the

generation process (i.e., execution traces), and information about the sources respectively.

In a highly collaborative environment such as the cyberinfrastructure, there are often multiple applications published that provide the same or very similar function. A thorough integrative application may consider all the different ways it can generate and present results to users, placing the burden on users to discriminate between high quality and low quality results. This is no different from any question/answer application, including a typical search engine on the Web, which often uses multiple sources and presents thousands of answers back to users. The *query view* visually shows the links between application requests and results of that particular request. The request and each corresponding result is visualized as a node similar to the nodes in the justification view presented later.

Upon accessing one of the answer nodes in the *query view*, Probe-It! switches over to the justification view associated with that particular result. Because users are expected to compare and contrast between different answers in order to determine the *best* result, all views are accessible by a menu tab, allowing users to navigate back to the query view regardless of what view is active.

The *results view* provides graphical renderings of the final and intermediate results associated with scientific workflows. This view is captured on the right hand side of Figure 1, which presents a visualization of a gridded dataset; this view is initiated by selecting one of the justification nodes, described in the next section. Because there are many different visualizations suited for gridded data and datasets in general, the *results view* is composed of a set of viewers, each implementing a different visualization technique suited to the data being viewed. The framework supporting this capability is described in Section 3.

The *justification view*, on the other hand, is a complimentary view that contains all the process meta-information associated with the execution trace, such as the functions invoked by the workflow, and the sequencing associated with these invocations. Probe-It! renders this information as a directed acyclic graph (DAG). An example of a workflow execution DAG can be found on the left hand side of Figure 1, which presents the justification of a contour map. From this perspective, Web services and sources (i.e., data sinks) are presented as nodes. Nodes contain a label indicating the name of a source or invoked service, as well as a semantic description of the resulting output data. In the justification view, data flow between services is represented by edges of the DAG; the representation is such that data flows from the leaf nodes towards the root node of the DAG, which represents the final service executed in the workflow. Users can access both provenance meta-information and intermediate results of the sources and services represented by the DAG nodes. In this sense, the justification DAG serves as a medium between provenance meta-information and intermediate results.

The *provenance view*, provides information about sources and some usage information e.g., access time, during the execution of an application or workflow. For example, upon accessing the node labeled *gravity database*, meta-information

about the database, such as the contributing organizations, is displayed in another panel. Similarly, users can access information transformation nodes, and view information about used algorithms, or the hosting organization.

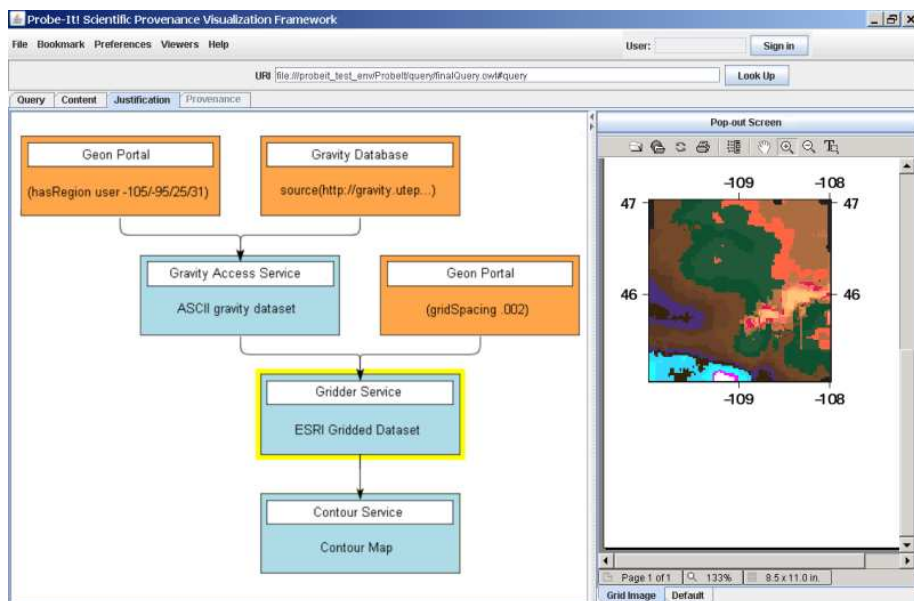


Fig. 1. Probe-It! justification view.

## 2.2 Result Viewers and Framework Support for Visualization Techniques

Different visualizations model data from different perspectives thus Probe-It! provides scientists with as many viewers as possible. For example, gravity datasets provided by the GIS center at the University of Texas at El Paso have three associated visualizations: default textual view, plot view, and XMDV view. The default textual view is a table; the raw ASCII result from gravity database. The location plot viewer provides a 2D plot of the gravity reading in terms of latitude and longitude. XMDV, on the other hand, provides a parallel coordinates view, a technique pioneered in the 1970's, which has been applied to a diverse set of multidimensional problems [9]. Figure 2 shows a pop-up of these three visualizations in their respective viewer windows. Upon selecting a node in a justification DAG, ProbeIt! is able to determine, based on a semantic description of the output data, which viewers are appropriate. This is similar to a Web browser scenario in which transmitted data is tagged with a MIME-TYPE

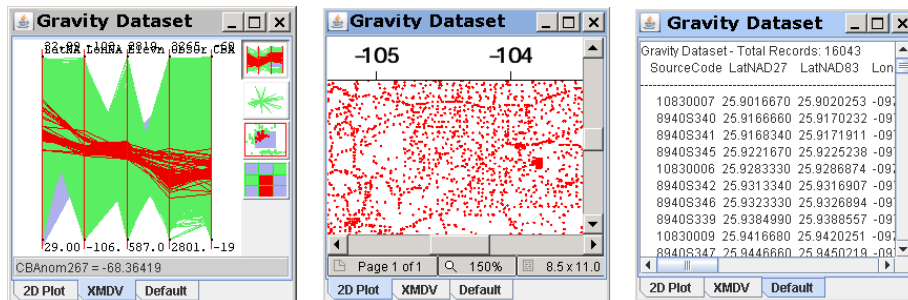


Fig. 2. Three different viewers for gravity data sets.

that is associated with a particular browser plug-in. Probe-It! should be flexible enough to support a wide array of scientific conclusion formats just as Web browsers can be configured to handle any kind of data, but also leverage any semantic descriptions of the data. For example, XMDV is a viewer suited to any N dimensional data; the data rendered by XMDV need only be in a basic ASCII tabular format, as shown on the right hand side of Figure 2, with a few additional headers. Because gravity datasets are retrieved in an ASCII tabular format, XMDV can be used to visualize them. However, this kind of data is also semantically defined as being *gravity point data*, in which case Probe-It! is configured to invoke the more appropriate 2D spatial viewer, as shown in the center of Figure 2. The semantic capabilities provided by the Probe-It! viewer framework compliments the MIME tables used in typical Web browsers, which only indicate the format or syntax of the data.

In order to manage the many relationships between a kind of data and the appropriate viewer, Probe-It relies on a MIME-like table to store these mappings. This table contains all the known data types, their semantic descriptions, and their respective renderers. Thus, an appropriateness of a particular renderer is based on both the data's format and semantic description. The property that makes this MIME-like table so desirable for the Probe-It is its extensibility; scientists can register new mappings on request, keeping Probe-It! up-to-date with the scientists' needs.

### 2.3 Comparing Knowledge Provenance: Pop-up Viewers

In many cases, scientists may need to compare the provenance associated with different results in order to decide which result best fits their needs. To facilitate such comparisons, results of a workflow can be popped out in separate windows. The pop-up capability provided by the tool is useful when comparing both final and intermediate results of different maps. Users can pop-up a visualization of intermediate results associated with one map, navigate to the justification of a different map and pop-up a window for the corresponding results, i.e., results of the same type, for comparison purposes. In addition to the result that is

being viewed, pop-up windows contain the ID of the artifact from which it is associated. This allows users to pop-up several windows without losing track of what artifact the pop-up window belongs to.

### 3 Underlying Technologies

#### 3.1 Proof Markup Language (PML) and the Inference Web

Provenance browsed by ProbeIt! is encoded in the Proof Markup Language (PML) [6] provided by the Inference Web [3]. Inference Web leverages PML for encoding and publishing provenance information on the web as well as providing a set of tools and services for handling these documents. PML is an OWL [5] based language that can be used for encoding provenance. PML consists of a specification of terms for encoding collections of justifications for computationally derived results. Depending upon the application domain, users may view each justification as an informal execution trace or as a proof describing the inference steps used by an inference engine, e.g., theorem prover or web service, to derive some conclusion. From this perspective, PML node sets (i.e., the topmost element in the language) represent the invocation of some service; the node set conclusion serves as the output of the service while the inference step represents the provenance meta-information associated with the function provided by a service or application. For example, the inference step proof elements *antecedent*, *rule*, and *inference engine* can be used to describe the applications inputs, function, and name respectively.

IW-Base is a repository of provenance elements that can be referenced by PML documents [4]. In order to support interoperability when sharing provenance among Inference Web tools and between Inference Web tools and other Semantic Web tools in general, provenance elements such as *sources* are stored and made publicly available in IW-Base. For example, PML *direct assertions* can be linked to provenance elements in IW-Base to indicate the agent, person, or organization responsible for the information being asserted. Since a single source can contribute to the generation of a number of different artifacts, IW-Base alleviates PML provenance loggers from always re-generating files describing sources that can otherwise be shared from the database.

#### 3.2 PML Service Wrapper (PSW)

PML Service Wrapper (PSW) is a general-purpose wrapper that logs knowledge provenance associated with workflow executions as a set of PML documents. Since workflows can be composed entirely of Web services, PSW logs workflows at the level of service invocations and transactions. Thus, information such as the input/output of each service and meta-information regarding the used algorithm are all logged by PSW.

In a cyber-infrastructure setting, functionality or reasoning is often supported by Web services that can be considered “black boxes” hard to be instrumented at

source-code level to generate PML. This is the primary reason why PSW, a sort of external logger, must be deployed to intercept transactions and record events generated by services instead of modifying the service and workflows themselves to support logging.

## 4 Evaluation

The effectiveness of provenance visualization in the task of understanding complex artifacts was verified by a user study described below. The context of the user study is presented first, following with a brief discussion of how provenance and visualization aided scientists in the evaluation tasks.

### 4.1 Gravity Map Scenario

Contour maps generated from gravity data readings serve as models from which geophysicists can identify subterranean features. In particular, geophysicists are often concerned with data anomalies, e.g., spikes and dips, because these are usually indicative of the presence of some subterranean resource such as a water table or an oil reserve. The Gravity Map scenario described in this section is based on a cyber-infrastructure application that generates such gravity contour maps from the Gravity and Magnetic Dataset Repository<sup>1</sup> hosted at the Regional Geospatial Service Center at the University of Texas at El Paso. In this scenario, scientists request the generation of contour maps by providing a footprint defined by latitude and longitude coordinates; this footprint specifies the 2D spatial region of the map to be created. The following sequence of tasks generate gravity data contour maps in this scenario:

1. *Gather Task*: Gather the raw gravity dataset readings for the specified region of interest
2. *Filter Task*: Filter the raw gravity dataset readings (remove unlikely point values)
3. *Grid Task*: Create a uniformly distributed dataset by applying a gridding algorithm
4. *Contour Task*: Create a contoured rendering of the uniformly distributed dataset

The gravity map scenario is thus based on a workflow in which each activity is implemented as an independent Web service. The following Section describes how this scenario served as a test-bed to evaluate the effectiveness of Probe-It! in aiding scientists to both identify and explain imperfect maps generated by the gravity map workflow.

---

<sup>1</sup> <http://irpsrvgis00.utep.edu/repositorywebsite/>

## 4.2 Results

The premise of our work is that scientific provenance is a valuable resource that will soon become an integral aspect of all cyber-infrastructure applications. The use of provenance is still being researched and its various applications are still being explored, thus a widespread adoption of provenance has yet to take place. A previous study of ours has indicated that providing scientists with visualizations of provenance helps them to both identify and explain map imperfections [7]. This study was composed of seven evaluation cases all derived from the different possible errors that can arise in the gravity map scenario; each case was based on a gravity contour map that was incorrectly generated. The subjects were each asked to identify the map as either correct or with imperfections. Additionally, they were asked to explain why they identified the map as such, usually by indicating the source of error. Table 1 shows the subjects accuracy in completing the identifying and explaining tasks with a contour map that was generated using a grid spacing parameter that was too large with respect to the density of data being mapped; this causes a loss of resolution hiding many features present in the data. The results are presented in terms of the classification of the subjects such as: subject matter experts (SME), Geographic Information Systems Experts (GISE), and non experts (NE).

**Table 1.** Percentage of correct identifications and explanations of map imperfections introduced by the inappropriate gridding parameter. [No Provenance (NP), Provenance (P)]

Experience	(% ) Correct Identifications		(% ) Correct Explanations	
	NP	P	NP	P
SME	50	100	25	100
GISE	11	78	11	78
NE	0	75	0	75
all users	13	80	6	80

## 5 Related Work

VisTrails, a provenance and data exploration system provides an infrastructure for systematically capturing provenance related to the evolution of a workflow [1]. Provenance information managed by VisTrails refers to the modifications or history of changes made to particular workflow in order to derive a new workflow; modifications include, adding, deleting or replacing workflow processes. VisTrails renders this *history of modifications* as a treelike structure where nodes represent a version of some workflow and edges represent the modification applied to a workflow in order to derive a new workflow. Upon accessing

a particular node of the provenance tree, users of VisTrails are provided with a rendering of the scientific product which was generated as a result of a particular workflow associated with the node. In this sense, VisTrails provides excellent support for visualizing both process meta-information and intermediate results but may not provide a rich description of provenance meta-information as defined in this paper. ProbeIt! is an attempt to visualize all aspects of provenance, including rich information about the sources used.

MyGrid, from the e-science initiative, tracks data and process provenance of some workflow execution. Authors of MyGrid draw an analogy between the type of provenance they record for cyber-infrastructure type applications and the kind of information that a scientist records in a notebook describing where, how and why results were experimental results were generated [10]. From these recordings, scientists can achieve three primary goals: (i) debugging, (ii) validity checking, and (iii) updating, which refer to situations when, a result is unexpected, when a result is novel, or a workflow component is changed respectively. The Haystack application displays the provenance as a labeled directed graph, tailored to a specific user; only relevant provenance elements related to the role of a user are rendered in order to reduce data overloading on the screen. In this scenario, links between resources are rendered allowing users to realize the relationships between provenance elements such as inputs/outputs and applied processes thus realizing the execution trace.

In contrast to graphically displaying scientific provenance, the Kepler [2] workflow design and execution tool provides an interface for querying recorded provenance associated with workflow execution via a set of predefined operators. In this case, provenance is queried, with the result of the query being some relation. Similarly, Trio, a management system for tracking data resident in databases, tracks data as it is projected and transformed by queries and operations respectively [8]. Because of the controlled and well understood nature of a database, lineage of some result can many times be derived from the result itself by applying an inversion of the operation that derived it. These inverse transformations of the data are stored in special table and made available via querying capabilities.

## 6 Conclusions and Future Work

The research presented in this paper is based on the fundamental problem of developing tools and methods that can help scientists understand complex scientific products (e.g., datasets, reports, graphs, maps) derived from complex software systems (e.g., applications and services) deployed on a distributed and heterogeneous environments such as cyber-infrastructures. We have developed Probe-It!, a tool that visualizes all aspects of provenance, to address these concerns. The work has been developed in the context of a realistic scenario based on ongoing cyber-infrastructure efforts in the fields of Earth Sciences. A user study driven by this scenario verified the effectiveness of provenance visualization provided by Probe-It! in helping scientists understand complex artifacts, strengthening

the notion that provenance should be maintained by all cyber-infrastructure applications, and available on demand in some useful representation.

Since the effectiveness of provenance has been demonstrated, the strategy will be to present scientists with the most effective ways of browsing such information. The current evaluation approach was based on the suitability of provenance in decision making scenarios, rather than the usability of the tool itself. Usability is based on the evaluation of many dimensions including learnability, understandability, and handling ability. Each of the aforementioned aspects refer to the amount of time of necessary training before independent use of a system is possible, ability of users to correctly draw conclusion from display, and the speed of a trained user respectively. The next step is to develop a more formal model of how users interact with provenance visualizations in order to improve the usability of Probe-It!

## References

1. Juliana Freire, Claudio T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo. Managing Rapidly-Evolving Scientific Workflows. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, 2006. (to appear).
2. B. Ludašcher and et al. Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice & Experience*, 2005. Special Issue on Scientific Workflows.
3. Deborah L. McGuinness and Paulo Pinheiro da Silva. Explaining Answers from the Semantic Web. *Journal of Web Semantics*, 1(4):397–413, October 2004.
4. Deborah L. McGuinness, Paulo Pinheiro da Silva, and Cynthia Chang. IW-Base: Provenance Metadata Infrastructure for Explaining and Trusting Answers from the Web. Technical Report KSL-04-07, Knowledge Systems Laboratory, Stanford University, 2004.
5. Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. Technical report, World Wide Web Consortium (W3C), February 10 2004. Recommendation.
6. Paulo Pinheiro da Silva, Deborah L. McGuinness, and Richard Fikes. A Proof Markup Language for Semantic Web Services. *Information Systems*, 31(4-5):381–395, 2006.
7. N. Del Rio and P. Pinheiro da Silva. Identifying and Explaining Map Imperfections Through Knowledge Provenance Visualization. Technical report, The University of Texas at El Paso, June 2007.
8. Jennifer Widom. Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In *Proceedings of the Second Biennial Conference on Innovative Data Systems Research*, pages 262–276, Asilomar, CA, January 2005.
9. Zaixian Xie. Towards Exploratory Visualization of Multivariate Streaming Data. <http://davis.wpi.edu/>.
10. J. Zhao, C. Wroe, C. Goble, R. Stevens andq D. Quan, and M. Greenweid. Using Semantic Web Technologies for Representing E-science Provenance. In *Proceedings of the 3rd International Semantic Web Conference*, pages 92–106, November 2004.