Identifying and Explaining Map Imperfections Through Knowledge Provenance Visualization

Nicholas Del Rio
University of Texas at El Paso
Computer Science, 500 W. University Ave.
El Paso, TX
ndel2@utep.edu

Paulo Pinheiro da Silva
University of Texas at El Paso
Computer Science, 500 W. University Ave.
El Paso, TX
paulo@utep.edu

ABSTRACT

Applications deployed on cyber-infrastructures often rely on multiple data sources and distributed compute resources to access, process, and derive results. When application results are maps, it is possible that non-intentional imperfections can get introduced into the map generation processes because of several reasons including the use of low quality datasets, use of data filtering techniques incompatible for the kind of map to be generated, or even the use of inappropriate mapping parameters, e.g., low-resolution gridding parameters. Without some means for accessing and visualizing the provenance associated with map generation processes, i.e., metadata about information sources and methods used to derive the map, it may be impossible for most scientists to discern whether or not a map is of a required quality.

Probe-It! is a tool that provides provenance visualization for results from cyber-infrastructure-based applications including maps. In this paper, we describe a quantitative user study on how Probe-It! can help scientists discriminate between quality maps and maps with known imperfections. The study had the participation of fifteen active scientists from five domains with different levels of expertise with regards to gravity data and GIS. The study demonstrates that a very small percentage of the scientists can identify imperfections using maps without the help of knowledge provenance. The study also demonstrates 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.

Categories and Subject Descriptors

D.2.5 [**Software Engineering**]: testing and Debugging—*debugging aids, diagnostics, tracing*; H.5 [**Information Interfaces and Presentation**]: General

General Terms

knowledge provenance visualization

Keywords

knowledge provenance, maps, cyberinfrastructure

1. INTRODUCTION

The use of maps is becoming more pervasive as geographical information system (GIS) technologies succeed in their goal of providing users with easier ways of accessing, combining and visualizing geo-spatial data. The commercial success of products like Google Earth and Microsoft Virtual Earth demonstrates that the use of maps can and will keep increasing in the future. Of particular interest in science is the generation of maps from the combined use of GIS technology and more readily available data provided by cyber-infrastructure communities [2] such as National Science Foundation (NSF) funded Geosciences Network (GEON) [1] and Circumarctic Environmental Observatories Network (CEON) [5, 6]. Scientists, who are not necessarily GIS experts, can now use their data along with data provided by these and many other cyber-infrastructure communities to create maps on demand. Maps, however, as any scientific product, are subject to imperfections, and most imperfections are too subtle to be identified by scientists whether they are subject matter experts (SME) (with respect to data used to generate maps), GIS experts, or just ordinary scientists with a specific need for a given map. For example, maps may be inaccurate because of: a faulty sensor in a collection of thousands of sensors used to generate a large geo-spatial dataset; incompatible ways of reading and storing measured geo-spatial data; services used to derive maps that are incompatible when combined; or even because of inappropriate use of parameters for any of the services used to derive a map. GIS and cyber-infrastructure, thus, may provide a context for the creation and proliferation of maps that one could label as inaccurate if one could know more about how they were generated.

Knowledge provenance (KP) is meta-information about how products, which can be maps, are generated. KP often includes meta-information about the following: original datasets used to derive products; executions of processes, i.e., traces of workflow executions and composite services execution; methods called by workflows and composite services, i.e., services, tools, and applications; intermediate datasets generated during process executions; and any other information sources used. In a GIS context, *knowledge* *provenance visualization* provides map users, e.g., scientists, with the capability of visualizing maps together with KP about how the maps were generated.

On-demand creation of maps from scientific data by non-GIS-expert scientists is beneficial for those who can use maps to visualize spatial data that they could not fully understand otherwise. Thus, science should not be stopped by the undesirable side-effect of having maps that were created by scientists when successfully using GIS and cyber-infrastructure technologies and that may include imperfections. Instead, a new habit of keeping KP about maps should take place together with the habit of using visualization tools for KP.

Probe-It![1] is our knowledge provenance visualization tool that is being validated by scientists involved in NSF-funded GEON and CEON cyberinfrastructures. In this paper we describe a comprehensive user study base on Probe-It! where we analyze how scientists with different levels of expertise on gravity data for geophysics and on GIS can differentiate contour maps with known imperfections from contour maps assumed to be correct, and to explain the reasons of identified imperfections. Scientist explanations of map imperfections are used to quantify their level of understanding on the reasons for a map to be classified as imperfect, if any. The goals of this study are to: (1) verify how much knowledge provenance is needed, if any, for scientists to identify map imperfections; (2) verify which kind of KP visualization support is more useful and needed for scientists to identify map imperfections; (3) identify the main requirements to improve Probe-It! from a geoscientist's perspective.

The rest of this paper is organized as follows. Section 2 provides background information on KP, KP visualization, and Probe-It! Section 3 describes a user study evaluation on how scientists identify and explain map imperfections. Section 4 shows results on how much KP is needed for scientists to identify and explain map imperfections. Section 5 shows results on how much KP visualization scientists need. Section 6 discusses the use of Probe-It! when compared with other tools. Section 7 summarizes the main results of our user study.

2. BACKGROUND

2.1 Knowledge Provenance

Knowledge provenance [14] includes *provenance meta-information*, which is a description of the origin of a piece of knowledge, and *process meta-information*, which is a description of the reasoning process used to generate the answer. We have used the phrase "knowledge provenance" instead of data provenance intentionally. Data provenance [4, 7] may be viewed as the analog to knowledge provenance aimed at the database community. That community's definition typically includes both a description of the origin of the information and the process by which it arrived in the database. Knowledge provenance is essentially the same except that it includes proof-like information about the process by which knowledge arrives in the knowledge base. In this sense, knowledge provenance broadens the notion of data derivation that can be performed before data is inserted into a database or after data is retrieved from a database. Nevertheless, data provenance and knowledge provenance have the same concerns and motivations.

[1]http://trust.utep.edu/probeit

The use of reasoning is not a requirement for using a knowledge provenance infrastructure. For instance, Inference Web [11] is a KP infrastructure and many of its components such as Proof Markup Language (PML) justifications [13] are used to provide simple source justification for answers that are simply retrieved or for answers that have been obtained using complex reasoning and, more typically, it can be used when the results are derived using a combination of both. A typical scenario includes using knowledge sources where information is available in a format appropriate for machine processing e.g., OWL [12]. If a knowledge base was built using a particular source, for example CNN, then Inference Web would store CNN as the original source of the knowledge. Additional information may be stored about knowledge sources such as the source's authoritativeness, URL, contributors, date of input and update, etc. If some of the information in a knowledge base is from another source, for example the AP news wire, then Inference Web may be used to store that certain assertions came from another source.

2.2 Knowledge Provenance Visualization

We see knowledge provenance visualization as a flexible framework suitable for visualizing both an application result and its associated KP. KP visualization is neither KP nor visualization but about the challenge of combining KP and visualization techniques so that users can have a single navigation model to understand results by understanding how they are derived. In this case, we see the visualization of process execution traces as one of the key components for knowledge provenance visualization.

Using the definition above, there are only a few systems that support knowledge provenance visualization. Most provenance systems (not necessarily knowledge provenance systems) focus on capturing and managing provenance information, while most visualization system focus only on providing an accurate rendering of some product, but not provenance. Furthermore, there are few systems that manage KP associated with distributed artifacts, such as Web services and publicly available datasets; many systems manage provenance associated only with locally available artifacts such as scripts and locally stored datasets, or with artifacts that are immediately associated with a local artifact such as a workflow specification.

Knowledge provenance dissemination refers to the means by which KP is displayed or presented to the user; textually, graphically, or on-demand such as a provenance querying API, which allows users to query for provenance information [16]. The scope of this paper is limited to dissemination via visualization, however discussion of query based dissemination will follow in Section 6.

2.3 Probe-It!

Probe-It! is an application for viewing both products, i.e., results from scientific workflow executions, and associated KP. Probe-It!'s navigation model is composed of two primary views: *process execution trace view* and *result view*.

2.3.1 Process Execution Trace View

In a GIS context, every aspect of a map generation trace is supposed to be visualized through a directed acyclic graph (DAG) that represents the execution trace of a workflow that generates the map. An example of a workflow execu-

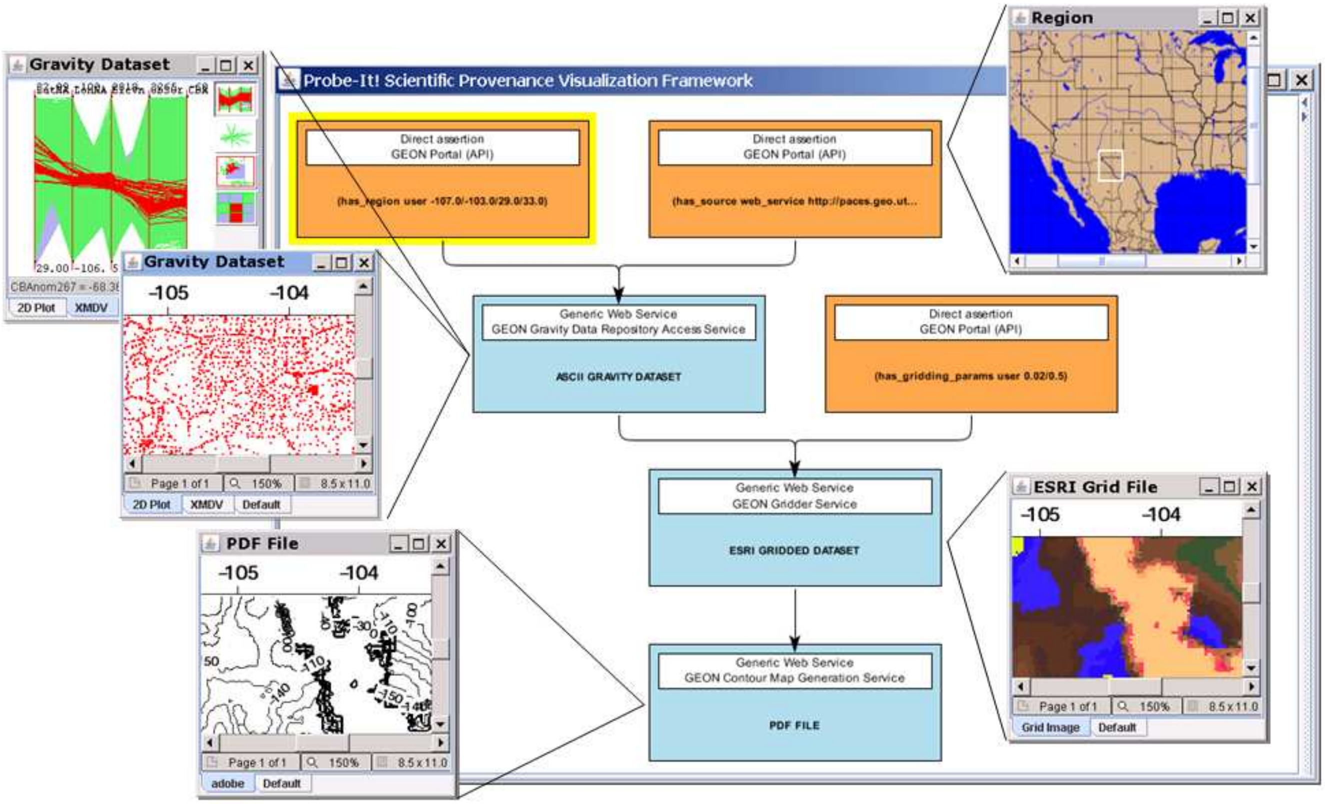

Figure 1: Probe-It! snapshot.

tion DAG can be found in Figure 1. In this DAG, data flow is represented by edges; the representation is such that data flows from the leaf nodes towards the root node of the DAG. The DAG contains two types of nodes, workflow inputs labeled as "direct assertions" and information transformation services labeled as "Generic Web Service". Workflow inputs may have been provided by a user, software agent, or data sink, and have no incoming edges into their nodes. Information transformation services are represented by the internal nodes of the DAG, and thus have one or more incoming edge representing data input. These services may have outgoing edges representing output data representing the fact that their results are consumed by other services. Each node contains a label indicating the name of the invoked service. The DAG root node represents the final service executed by the workflow, generating thus the workflow results. A dataset or product is associated with each service, which is the output of the service. The type of this dataset is made explicit by a label appended to each node; the data is described at a semantic level by labels such as "gravity dataset" rather than at a syntactic level by labels such as "ASCII table". Nevertheless, the KP encoding knows that gravity datasets can be encoded as ASCII tables. For example, in Figure 1, the "GEON Gridder Service" has the label "ESRI Gridded File" to indicate that the output of that service was and ESRI gridded dataset.

2.3.2 *Result View*

In addition to the visualizing the process execution trace, Probe-It! provides visualizations for datasets associated with the workflow services, whether the results are input data, intermediate results, or final results of workflow executions. Upon mouse clicking a DAG node, a visualization of the data associated with the node will be presented in a new window, such as the pop-up windows presented in Figure 1. Probe-It! will visualize the data according to its type and the set of visualization techniques available. Even the input parameters can be visualized; Figure 1 shows the "region parameters" visualized as a map of the United States with a footprint indicating the region requested by some user. Because the data is defined semantically, more powerful visualizations can be employed. At the bottom of each visualization pop-up is a tabbed menu, populated with all the different visualizations available for that particular type of data; every data type has at least one corresponding view or visualization, the default view, which always renders the data textually. Because different visualizations model the data from different perspectives, it is important to provide the scientists with as many views as possible. For example, gravity datasets have three associated visualizations: default textual view, location plot view and XMDV view. The default textual view is essentially a data table, the raw result from gravity database. Figure 1 shows a pop-up of both the location plot view and the XMDV view. The location plot visualization provides a 2D plot of the gravity reading in terms of latitude and longitude. XMDV provides a parallel coordinates view, a technique pioneered in the 1970's which has been applied to a diverse set of multidimensional problems. Additionally, the ESRI gridded dataset has associated colored image visualization, also presented in Figure 1 as well as a textual default view, another tabular structure. The pop-up visualization windows are useful when comparing provenance of different maps. Users can pop-up a visualization of some intermediate result, navigate to the provenance of a different map and pop-up the same type of intermediate result for comparison purposes. The pop-up windows contain not only the type of intermediate result that is benign viewed, but the ID of the map from which it is associated with. This allows users to pop-up many windows without worrying about loosing track of what map the visualization belongs to.

3. PROVENANCE VISUALIZATION USER STUDY

3.1 Gravity Map Scenario

A Gravity Map scenario is based on a cyberinfrastructure application, which generates contour maps from gravity readings associated with a 2-D geospatial region. In this scenario, scientists request the generation of gravity maps by providing a geospatial footprint defined by upper and lower latitude and longitudes. In order to generate a contour gravity map, as is generated in the gravity map scenario, the following sequence of tasks must be performed:

1. Gather raw point data (possible from multiple sources) specified in the region
2. Filter the data (remove unlikely point values)
3. Smooth the filtered data using some gridding algorithm
4. Contour the smoothed data

Each one of the four tasks above is realized by a web service. This set of web services would be piped or chained together; the output of one service would be forwarded as the input to the next service specified in the workflow.

In a highly distributed environment such as the cyberinfrastructure, many times the workflows that generate such maps are constructed dynamically, based on the demands of the requesting scientists. Assuming that there were multiple services that provide the same functionality for some of the workflow activities or steps described above, some coordinating agent would be responsible for deciding which services to use and which to exclude based on the preferences of the requesting scientist. For example, in step (3), multiple gridding services may be published on the cyberinfrastructure, each based on a different gridding algorithm such as near neighbor or block mean. In the case where the scientist has no preferences regarding the use of particular services, the coordinating agent will construct multiple workflows from the different combinations of services available; essentially the Cartesian product of all service types will be calculated, yielding a number of different workflows all of which produce the products of the same kind. Using the previous example, the coordinating agent would construct two workflows, one using the near neighbor gridding service and other using the block mean service. In these types of situations, multiple maps of the same region will be produced, each constructed by a different process. A similar situation can result from existence of multiple data sources, each of which hosting datasets that overlap in terms of geospatial coverage. Again, the coordinating agent would construct multiple workflows, each using different datasets or some complex aggregation. In the case where multiple workflows can satisfy a single request, the set of results generated by the workflows are presented to the user. As in any question/answer scenario, it is up to the user to determine what result (answer) to use. This situation is no different from how users interact with Web search engines. A single query often yields thousands of results, yet the burden is placed on the use to determine which answer is most appropriate.

3.2 Creating Evaluation Cases

Scientists generate contour maps from gravity data to get a rough idea of the subterranean features that exist within some region. Geoscientists are often only concerned with anomalies, or spikes in the data, which often indicate the presence of a water table or oil reserve. These anomalies have the potential to be artificial, in that they are simply imperfections introduced during the map generation process. Without KP, it may be impossible to evaluate the quality of the map generation process and thus determine whether the anomalies are naturally occurring or simply errors introduced in the generation process. Table 1 shows the evaluation cases used to compose the user study. The scientists, who are the subjects in our study, were required to classify the maps produced in each evaluation case as either correct or with some inaccuracy. For maps identified as imperfect, the study further requires that the subjects explain why they believe the maps contain some imperfection. Subjects are supposed to perform this classification task by using Probe-It! to visualize the maps and their associated KP.

Table 1: Evaluation cases used in user study.

Condition	Code	Evaluation Case
without comparison	C0	No knowledge provenance
	C1	Single source wrong grid parameters
with comparison	C2	Single source correct
	C3	Dual source correct
	C4	Single source random data skew
	C5	Single source wrong grid parameters
	C6	Dual source uniform data skew

In the gravity map scenario, execution control is transferred between four remote services. If an unexpected map were to be generated, e.g., a map suspected to be incorrect, it would be difficult to identify the source of a potential error. It would be difficult to identify if there is a problem with the process, methods used in the process, or in/out data. For example, the unexpected map may be the result of a faulty sensor used to gather data in step (1) of the Gravity Map scenario in Section 3.1 or the use of a filter in step (2) that is unsuitable for the kind of data produced in step (1). Without KP, scientists would have a difficult time deciding whether an experimental result is correct, because they would have very little understanding of the workflow that derived it.

Suppose that in step (2), the search radius parameter is set to a value that is inappropriate for the dataset it is gridding. The search radius parameter specifies how many neighboring points to consider in thus producing a map with too coarse a resolution. This occurs because the gridding routine would not consider enough points in the interpolation. Ultimately, the gridding routine will generate a low quality data, which will exclude many features present in the region. Scenarios C0, C1, and C5 in Table 1 capture the case where a gridding parameter is specified incorrectly for the density of the data points being gridded.

Additionally, suppose that the contour map was derived from a pair of disjoint datasets, one of which is uniformly skewed by 10 percent. This can be the result of the different instrument's precision or configuration that recorded the data. The resultant contour map usually contains a very prominent fault line where the region covered by each dataset meets. This particular case is captured in C6.

Errors can also arise even when using only a single dataset that happens to have been derived from faulty sensors. In this example, random points or records in the dataset contain values that are drastically higher or lower in value than the neighboring points resulting in a dense contour map. In fact, depending upon the severity of the faulty points, the contour map can become unreadable. C4 represents such a case.

Our set of evaluation cases also contain maps that are supposed to be correct because they were generated using reliable sources of data, using compatible methods, and parameters considered to be correct. C2 is a case where a correct map is generated from a single source. C3 represents the case where the map is generated from two reliable data sources.

3.3 Demographics

The only requirement for participation in the user study is that the subjects are active researchers in some scientific field. Although the scenario is based on gravity contour maps, the claim is that with KP visualization, most scientist can identify and explain the quality of some product, regardless of the services and domains of the datasets used to generate the product. The user study presented in this paper includes the participation of fifteen scientists from various fields including geophysics, geology, biology, environmental sciences, and physics. These scientists are affiliated to various organizations located in Alaska, Arizona, Oklahoma and Texas. Table 2 shows the percentage of SME on the use of gravity data in geophysics, GIS experts (GISE), subjects who are SME and GISE (SME+GISE), and subjects who are not familiar with either gravity data or GIS. Additionally, these groups of subjects are further segmented by their level of education.

Table 2: Subject demographics.

Education	SME	GISE	SME+GIS	NE
Complete PhD	20.0	53.3	13.3	33.3
Graduate Student	6.6	6.6	0	0

4. KNOWLEDGE PROVENANCE NEED

Each subject usually required about 45 minutes to complete the evaluation. This time included a description of key concepts such workflows and web services, a brief tutorial on how to use Probe-It!, and an explanation of the evaluation tasks, and the actual completion of the tasks. The main results of the evaluation are described below in this section.

4.1 Identifying Map Imperfections

Access to KP enables scientists to assess the quality of maps in several ways. For example, the identification of dataset sources may help scientists establish whether they believe a map to be of the required quality, by assessing how much they trust the sources. Our study elaborates on this claim by providing qualitative evidence supporting the need for provenance in assessing the quality of maps generated on the cyberinfrastructure.

Table 3: Percentage of correct identifications of map imperfections.

	(%) Without Comparison		(%) With Comparison				
Experience	C0	C1	C2	C3	C4	C5	C6
SME	50	100	100	100	100	75	100
GISE	11	78	89	89	100	89	89
SME+GISE	50	100	100	100	100	100	100
NE	0	75	100	100	75	100	100
all users	13	80	93	93	93	86	93

Table 3 contains the percentage of scientists who correctly identified the maps in each evaluation case as either correct or with fault. Cases C0 and C1 tested whether provenance was needed in order to correctly assess the map; both cases are based on the same map containing the same error with the ability to access KP in case C1 being the only difference. By visually examining the provenance information of each map, every category of scientist correctly identified at least two times as many maps as they did without. Prior to the use of provenance, many scientists were unable to determine whether the map contained any imperfections at all, in which case their responses were regarded as unsuccessful. Without provenance, the scientists were unable to determine either the sources of data used to generate the map or how the map was generated. After the scientists were able to access the provenance, both their accuracy and confidence in determining the quality of the map improved. These positive results transcended to the non-expert users, who initially could not correctly identify the map imperfections in C0. Surprisingly, 75 percent of the non-experts were then able to correctly identify that same map provided provenance and the corresponding visualizations, despite their unfamiliarity with both gravity and GIS systems in general.

When scientists were granted the capability to visually compare different maps and their corresponding provenance, as in cases C4 – C6, they were able to better decide which maps were correct than if they were only presented a single map. Although the increase is relatively small, it suggests that providing users with alternatives increases their accuracy in this type of decision process. Essentially, the task then becomes, given a set of maps and their corresponding provenance, identify which maps are correct and which are not. Identifying which maps contain imperfections from this perspective allows users to easily identify the features shared by all maps and identify the "odd" maps which do not share the characteristics of the majority. These odd maps were quickly regarded as incorrect by the scientists. For example, the map used in case C6 was derived from a pair of disjoint datasets. This map, because of its unique fault line as described in Section 3.2, was usually the first map to be regarded as incorrect. Once the more obvious maps were identified as incorrect and disregarded from the comparison, the scientists then more thoroughly analyzed the provenance of the remaining maps until they could identify the more subtle differences and further narrow their set of candidate maps. The case which most exemplifies this notion is C4. As described in Section 3.2, case C4 presents subjects with a contour map generated by a single dataset containing randomly skewed data points. Essentially C4 represents the case when a map is generated using bad data. Without the ability to compare this map, and its associated KP, non-subject matter experts are unable to identify whether the map is correct or not since they cannot know what the appropriate values for data points are without some prior KP of both gravity and the particular region.

4.2 Explaining Map Imperfections

Table 4: Percentage of corrent explanations of map imperfections.

	(%) Without Comparison		(%) With Comparison				
Experience	C0	C1	C2	C3	C4	C5	C6
SME	25	100	75	75	75	75	100
GISE	11	78	89	89	67	78	78
SME+GISE	50	100	100	100	100	100	100
NE	0	75	75	75	25	75	75
all users	6	80	80	80	53	73	80

Table 4 presents the percentage of subjects that were able to correctly explain why the map was of low or high quality. Explaining why a map has been generated incorrectly was regarded as the most difficult task of the user study; less scientists were able to correctly explain the map defects than were able to identify whether the map was low quality. Explaining why a map is of low quality entails that the subjects understand the factors in the map generation process that lead to the maps being imperfect. Thus the more experience a scientist has generating maps and using GIS, the more accurate their explanations were regarding the correctness of the maps. The scientists with GIS experience were able to explain more of the maps that they correctly identified in Table 3.

In general however, every group of scientist showed an improvement in their explanations with the use of provenance as exemplified in cases C0 and C1. With provenance, the scientists were able to browse the intermediate results that they believed to be the most indicative of the quality of a map and formulate an explanation for why the map is correct or not. The non-expert group again performed very well with a 75 percent increase in the number of scientists who were able to complete the task. Of course, even better results were yielded by the scientists who have experience with both gravity data and GIS.

Positive results were also yielded when scientists were able to compare between the different maps and their provenance, as in cases C2 – C6. Except for only a couple of scientists who were very knowledgeable about gravity data, the majority employed a similar method to formulate their explanations as when they were asked to identify the perfect/imperfect maps. Each scientist would attempt to find the common features between the maps and isolate the differing maps. Usually, the scientists were able identify some artifact in the provenance that was vastly differ from the majority of the maps and label this differing artifact as the cause of some error. From the results of the experiment, this technique proved to be very productive.

In almost all cases, scientists used KP related both to the process or execution trace of the workflow and some intermediate result. The results from case C6 support this claim. The map generated for case C6 usually contained a prominent fault line where the two datasets were merged. This feature was most noticeable in the ESRI gridded data file, however non-subject matter experts were initially leery as to whether this was a natural or some artificial feature introduced in the generation process. Upon inspecting the provenance related to the execution trace, subjects could identify that the map was constructed from dual sources. Almost instantly the subjects concluded that one of the datasets must have higher or lower values. In this case subjects relied on both the ESRI gridded file and the execution DAG to formulate a belief and thus an explanation for the imperfection.

5. PROVENANCE VISUALIZATION NEED

Every aspect of the KP is supposed to be visualized in some way, including the execution trace and the intermediate results. The raw viewers (the viewers that provide the minimum level of transformation of data) are available as a control mechanism for users to see the raw data in case they cannot understand or believe the more elaborate visualizations. From an evaluative point of view, the use of the raw viewers provides a test to determine whether the more elaborate visualizations are needed at all; the gravity dataset was available in its raw tabular form as well as the ESRI gridded dataset. This section quantifies the need for provenance visualization by presenting the percentage of users who relied on some visualization, other than the default view, to formulate their answers.

Table 5: Probe-It! feature usage (* indicates raw viewer).

Feature	Viewer	Usage %
Process trace	DAG	86
Gravity dataset	dataset*	26
	XMDV	26
	2D Plot	6
Grid dataset	ESRI dataset*	0
	grid image	66
Contour map	binary PDF*	0
	contour image	93

Table 5 shows the various visual features available in Probe-It! and the percentage of subjects who used each feature. The features are broken down into the following categories: process, gravity, grid, contour, which correspond to the features supporting visualizations for the execution trace, gravity datasets, ESRI gridded datasets, and contour PDF files respectively. The process category contains a single viewer, the DAG. If a subject used any information available in the DAG such as the input parameters or the dataflow itself, then that subject would contribute to the numbers shown the table. The other KP categories refer to the intermediate results that were generated by the workflow. Gravity for instance, has three corresponding viewers. Once again, if a subject relied on a particular view, then that was recorded as well. From the table, it is evident that the vast majority of the subjects relied on some viewer other than the raw viewer. This shows that KP is more useful to scientists if is presented in some visualization. This was especially evident with the ESRI gridded dataset, as every subject who accessed this map KP used the corresponding grid image viewer rather than the raw dataset. This is in part due to the size of the intermediate results which are rather large. In their raw form, intermediate results associated with map making are useless without some condensed view. In other words, the KP is only as useful as the visualization overlaid on it.

6. DISCUSSION AND RELATED WORK

Provenance views are dictated by the goals of the particular systems; because various dimensions of provenance can be used to achieve various goals, there is no one view fits all. Instead, various views or projections of provenance exist. For instance, a first category of provenance systems aim at providing users with a sort of "history of changes" associated with some workflow, thus their view of provenance differs from that of a second category of provenance systems, which aim at providing provenance for use of debugging, or understanding an unexpected result. A third category of provenance systems record events that are well suited for re-executing the workflow it is derived from. From this point of view, Probe-It! fits into the second category of provenance systems. Provenance systems representative of these categories are reviewed below.

VisTrails, a provenance and data exploration system, provides an infrastructure for systematically capturing provenance related to the evolution of a workflow [8]. VisTrails users edit workflows while the system records the various modifications being applied. In the context of this system, provenance information 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 provides a novel way to render this history of modifications. A treelike structure provides a representation for provenance where nodes represent a version of some workflow and edges represent the modification applied to a 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 the workflow associated with the node. In the context of VisTrails, only workflows that generate visualizations are targeted, however the authors describe how this system could be transformed to handle the general case as provided by Probe-It!; to provide a framework that can manage and graphically render any scientific result ranging from processed datasets to complex visualizations.

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 in-silico experiments and the kind of information that a scientist records in a notebook describing where, how and why experimental results were generated [18]. 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. Based on particular user roles, the appropriate dimension of provenance is presented, knowledge, organization, data, or process level [18]. Currently, MyGrid RDF provenance is viewed using Haystack [15]. Haystack displays the provenance log as a labeled directed graph tailored to the needs of a specific user; only relevant provenance elements related to the role of a user are rendered. In this scenario, connections between different resources are rendered allowing users to realize the relationships between provenance elements such as inputs/outputs and applied processes and thus realize the execution trace. MyGrid however is moving towards presenting provenance as a set of linked documents, which are browsed similarly to HTML documents on the Web. In this case, each provenance document is just a piece of the whole, thus providing users with local views of the provenance graph.

The Earth Science System Workbench (ESSW) is another effort at capturing and presenting scientific results to users [9]. Upon user requests, ESSW leverages a suite of Notebook tools that can display both the scientific product and the associated provenance. Stored scientific visualizations such as swaths [9] are rendered in HTML upon request; the request is in the form of a query. Additionally, ESSW leverages GraphViz [10] in order to graphically render the execution trace in the form of a directed graph, where nodes are data objects and edges define relationships between objects.

Karma [16] is a non-obtrusive provenance recorder for scientific results from Indiana University. Karma, unlike ESSW provides an in-house approach for rendering provenance; an algorithm accurately pieces together a directed acyclic graph that describes the data or process provenance. Karma is primarily targeted at capturing provenance associated with service oriented workflows, thus rich provenance associated with Web service invocations are captured by the system. The two efforts MyGrid and VisTrails both support graphical visualization of the tracked provenance. In contrast, the Kepler workflow design and execution tool provides an interface for querying recorded provenance via a set of predefined operators. In the scope of Kepler, only provenance related to functional aspects of the workflow are captured by default [3]. For example, the set of inputs that contributed to some intermediate result are recorded, however information such as timestamps and authors of services are deemed as non-functional and dismissed. Thus, scientists can only query about information related to the events triggered by a workflow, such as reading, writing and state-resetting [3]. The method for presenting intermediate results, which can be accessed by Kepler queries, is not addressed.

All these provenance system thus far track provenance related to workflows. Trio is a management system for tracking data resident in a database; provenance is tracked as the data is projected and transformed by queries and operations respectively [17]. Provenance related to some function is recorded in a lineage table with various fields such as the derivation-type, how-derived, and lineage-data. 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. Additionally, Trio provides the capability of querying the lineage table, thus allowing users to request provenance on demand.

On the commercial side, ArcGIS from ESRI allows users to both develop and execute workflows (or "models" as called by ArcGIS). From a workflow, users have access to the final product, i.e., a map, intermediate results, and meta-data associated with the source data. Additionally, all these elements associated with a model can be visualized. ArcGIS tools draw no distinction between executable models and execution traces of models; no view of a model's execution trace is provided, only the model itself. Therefore, ArcGIS may not necessarily support KP visualization as defined in this paper. However the model provides certain features such as data point visualization which can be used to analyze final results and thus identify and explain map imperfections.

7. CONCLUSIONS

This paper described the use of Probe-It!, a knowledge provenance visualization tool, to support a comprehensive user study evaluation of how scientists can identify and explain map imperfections. The evaluation demonstrated that most scientists are unable to identify map imperfections if maps are provided with no knowledge provenance. With the use of knowledge provenance, however, the study showed that most scientists, including those who are neither GIS experts nor experts in the map's field, were capable or correctly differentiate correct maps from inaccurate maps. Moreover, the study demonstrated that most scientists could understand the factors leading maps to be inaccurate as demonstrated by their map imperfection explanations.

The evaluation confirmed the benefits of using knowledge provenance visualization especially in scenarios where more than a single map is provided by a single request. State-of-the-art cyberinfrastructure-based applications are getting close to a point where they will be able to generate large quantities of maps, probably several of those with one ore more imperfections. Probe-It! is moving towards the right direction as pointed by the evaluation results summarized above and indicated by the study's subjects, most of them are already aware of the necessity of cyberinfrastructure-based applications to support knowledge provenance.

8. REFERENCES

[1] R. Aldouri, G. Keller, A. Gates, J. Rasillo, L. Salayandia, V. Kreinovich, J. Seeley, P. Taylor, and S. Holloway. GEON: Geophysical data add the 3rd dimension in geospatial studies. In *Proceedings of the ESRI International User Conference 2004*, page 1898, San Diego, CA, August 2004.

[2] D. Atkins, K. Droegemeier, and et al. Revolutionizing Science and Engineering through Cyberinfrastructure. Technical report, National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January 2003.

[3] S. Bowers, T. McPhillips, B. Ludascher, S. Cohen, and S. B. Davidson. A Model for User-Oriented Data Provenance in Pipelined Scientific Workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS. Springer, 2006.

[4] P. Buneman, S. Khanna, and W.-C. Tan. Why and Where: A Characterization of Data Provenance. In *Proceedings of 8th International Conference on Database Theory*, pages 316–330, January 2001.

[5] Circumarctic Environmental Observatories Network (CEON). www.ceoninfo.org.

[6] CEON Internet Map Server. www.ceonims.org.

[7] Y. Cui, J. Widom, and J. L. Wiener. Tracing the Lineage of View Data in a Warehousing Environment. *ACM Trans. on Database Systems*, 25(2):179–227, June 2000.

[8] J. Freire, C. 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).

[9] J. Frew and R. Bose. Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management*, pages 180–189, Fairfax, VA, July 2001.

[10] A. R. Labs. AT&T Graphiz. http://www.graphviz.com.

[11] D. L. McGuinness and P. Pinheiro da Silva. Infrastructure for Web Explanations. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *Proceedings of 2nd International Semantic Web Conference (ISWC2003)*, LNCS-2870, pages 113–129, Sanibel, FL, USA, October 2003. Springer.

[12] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. Technical report, World Wide Web Consortium (W3C), December 9 2003. Proposed Recommendation.

[13] P. Pinheiro da Silva, D. L. McGuinness, and R. Fikes. A Proof Markup Language for Semantic Web Services. *Information Systems*, 31(4-5):381–395, 2006.

[14] P. Pinheiro da Silva, D. L. McGuinness, and R. McCool. Knowledge Provenance Infrastructure. *IEEE Data Engineering Bulletin*, 25(2):179–227, December 2003.

[15] D. Quan, D. Huynh, and D. Karger. Haystack: A platform for authoring end user semantic Web applications. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 738–753, 2003.

[16] Y. L. Simmhan, B. Pale, and D. Gannon. A Survey of Data Provenance Techniques. Technical Report IUB-CS-TR618, Computer Science Department, Indiana University, USA, 2005.

[17] J. 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.

[18] J. Zhao, C. Wroe, C. Goble, R. S. andq D. Quan, and M. Greenweed. Using Semantic Web Technologies for Representing E-science Provenance. In *Proceedings of the 3rd International Semantic Web Conference*, pages 92–106, November 2004.