Identifying and Explaining Map Quality Through Provenance: A User Study

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 high and low quality contour maps. The study had the participation of twenty active scientists from five domains with different levels of expertise with regards to gravity data and GIS. The study demonstrates that only 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 provenance to inspect maps.

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, cyber-infrastructure

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 geospatial 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 result, 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 or of low quality if one could know more about how they were generated.

Provenance, in the context of this study, is meta-information about how complex results, which can be maps, are generated. Provenance often includes meta-information about the following: original datasets used to derive results; 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 [19]. In a GIS context, *provenance visualization* provides map users, e.g., scientists, with the capability of visualizing maps together with any information available in the provenance including partial results, parameters, and name of information sources.

On-demand creation of maps from scientific data by non-GIS-expert scientists is beneficial for those scientists 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 provenance about maps should take place together with the habit of using provenance visualization tools for inspecting the quality of maps.

Probe-It![1] is our provenance visualization tool that is being validated by scientists involved in NSF-funded GEON and CEON cyber-infrastructures. In this paper we describe a comprehensive user study based on Probe-It! where we analyze how scientists with different levels of expertise on gravity data for geophysics and on GIS can differentiate between contour maps of high and low quality and to explain the reasons of identified qualities. Scientists' explanations are in turn used to quantify their level of understanding regarding the reasons why maps exhibit low quality features. The primary goal of this study is to verify that provenance is needed for scientists to identify and explain the quality of a map. Additionally, the long-term continuation of the study will provide us insight into what visualizations are most effective for viewing provenance and allow us to further improve Probe-It! by integrating these findings.

The rest of this paper is organized as follows. Section 2 provides background information on cyber-infrastructure based contour maps, provenance and provenance visualization, and the role played by Probe-It! in disseminating provenance. Section 3 describes a user study on how scientists identify and explain imperfections resident in gravity contour maps using Probe-It! Section 4 shows results providing strong evidence that provenance is needed for scientists to identify and explain map imperfections. Section 5 discusses the visualizations which were most popular among the study participants. Section 6 serves as related work and discusses the use of Probe-It! when compared with other tools. Finally, section 7 summarizes the main results of our user study.

2. BACKGROUND

2.1 Contour Map Generation on the Cyber-infrastructure

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. In a contour map, these anomalies are illustrated as a set of contour lines with very close proximity indicating a drastic change in gravity (or whatever data is being mapped). However these anomalies have the potential to be artificial and simply imperfections introduced during the map generation process. This process, which begins by scientists providing a region or interest or footprint, specified in terms of latitude and longitude, is defined by the sequence of tasks below:

1. gather raw point data (possibly from multiple sources) in the region
2. filter raw point data to remove duplicate data
3. uniformly distribute the filtered data using some gridding algorithm to generate a new *gridded* dataset with fixed intervals of data.
4. contour the gridded data

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

In a cyber-infrastructure setting, each one of the four tasks above can be realized by a web service. This set of web services is piped or chained together; the output of one service would be forwarded as the input to the next service specified in the *workflow*, such as in [8, 13].

In a highly distributed environment such as the cyber-infrastructure, 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 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 cyber-infrastructure, each based on a different gridding algorithm such as near-neighbor or minimum curvature [20]. In this case, the coordinating agent would construct two workflows, one using the near-neighbor gridding service and other using the minimum curvature service. In similar cases where scientists have no preferences regarding the use of particular services, the coordinating agent will always construct multiple workflows from the different combinations of services available, yielding different workflows all of which produce results of the same kind. A similar situation can result from the existence of multiple data sources, each of which are hosting datasets that overlap in terms of geospatial coverage. Again, the coordinating agent would construct multiple workflows in which each is utilizing either data coming from only a single source or data resulting from the combination of the different data sources available.

In these types of situations where multiple workflows can satisfy a single type of request, the set of results generated by each workflow are returned to the scientist. As in any question/answer scenario, it is up to the scientist to determine what result to use. However, 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 user to determine which answer is most appropriate.

2.2 Quality Maps

Because of the highly collaborative and heterogeneous nature of the cyber-infrastructure, there are many factors that can contribute to the quality of resident maps. In this paper we define a *quality* map as an artifact maintaining the following properties:

- the source data used to derive the map has no known imperfections; the domain scientists, who have collected the data, are unaware of any erroneous values.
- any data aggregations or merges were performed correctly according to specifications provided by many experts in the field (e.g., eliminating duplicate data or ensuring that data is semantically compatible)

- all services used to generate the map are semantically compatible (for example, gridding or interpolation services are suitable for the type and density distribution of the source data)
- all interpolation parameters to the gridding services are suitable for the type and density distribution of the source data according to specifications provided by many experts in the field
- must have a high resolution

From our definition, quality maps are not only a factor of the quality or correctness of the source data or routines used, but they must also have been generated with a *reasonable* resolution for a particular task. In this paper, it is assumed that a map with higher resolution than another map is regarded as better quality.

2.3 Provenance

In this paper, provenance 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, which may include intermediate or partial results. In the context of map generation, provenance would include some description about the sources of spatial data that were used to generate the contour map as well as a description or execution trace of the underlying workflow that coordinated the execution of the services. Additionally, all data associated with the computation such as input datasets and the associated transformations of these datasets would be regarded as provenance. This notion of provenance is more comprehensive than the ones used in other communities such as in databases [4, 7] and scientific workflows [3] and is also referred as *knowledge provenance* [17]. 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.

The use of reasoning is not a requirement for using a knowledge provenance infrastructure. For instance, Inference Web [14] is a provenance infrastructure and many of its components such as Proof Markup Language (PML) justifications [16] 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 [15]. 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.

2.4 Provenance Visualization

We see provenance visualization as a framework suitable for visualizing both an application result and its associated provenance. Provenance visualization is focused around the challenge of combining provenance and visualization techniques so that users can have a single interaction and 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 provenance visualization.

Using the definition above, there are only a few systems that support provenance visualization as further discussed in Section 6. 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 result, but not provenance. Furthermore, there are few systems that manage provenance 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.

Provenance dissemination refers to the means by which provenance 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 [19]. The scope of this paper is limited to dissemination via visualization, however discussion of query based dissemination will follow in Section 6.

2.5 Probe-It!

Probe-It! is a browser suited to graphically rendering knowledge provenance information associated with results coming from software agents and scientific workflows, such as the workflows that generate our gravity contour maps. In this sense, Probe-It! does not actually generate content (i.e. logging or capturing provenance information); instead it is assumed that users or software agents 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 views, each suited to different elements of provenance.

2.5.1 Query View

In an environment such as the cyber-infrastructure, there are often multiple applications that provide the same or similar functionalities (e.g., generating elevation maps or seismic models). A thorough integrative application may consider different ways it can generate and present results to scientists, placing the burden on them to discriminate between high quality and low quality results. Probe-It!'s *query view* visually shows the links between application requests and results of that particular request. For example, Figure 1 shows a request for a contour map of gravity anomalies

```
(and  (istype ?x contourmap)
      (hasdatatype ?x cbanom)
      (hasregion ?x -110.0/-106.0/35.0/39.0))
```

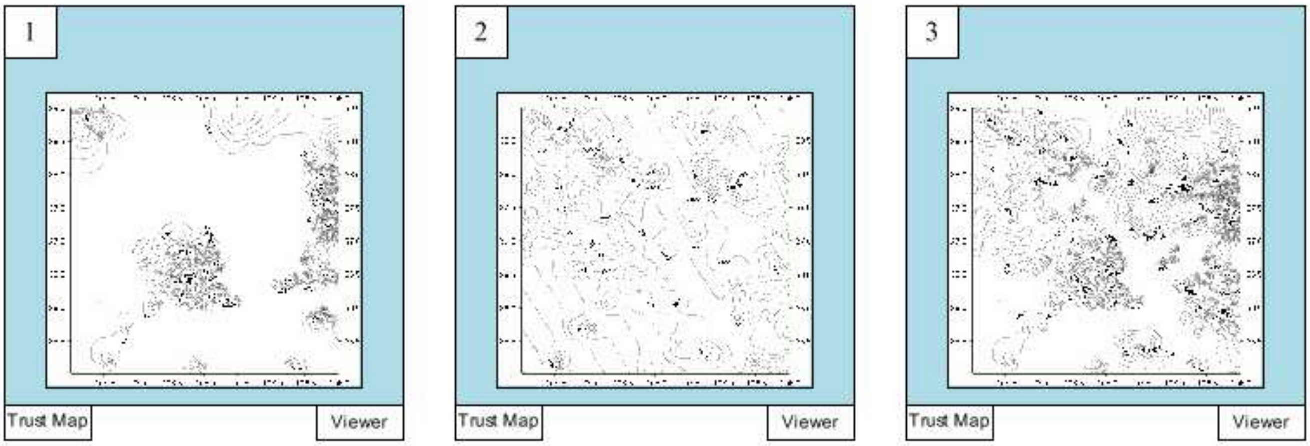

Figure 1: Probe-It! query view snapshot.

for the region -110/-106/35/39 and the corresponding maps that could be generated for that region, provided that the data to be mapped is of type *cbanom* (complete Bouguer anomaly).

The query view serves as both an entry point to the provenance of a map and as a mechanism to support easy comparison among results. Upon selection of a answer, the query view will initiate the justification view described in Section 2.5.2, which will present the process meta-information associated with that result. In order to facilitate comparison between results and between the provenance of those results, each result is tagged with a unique ID. The function of this ID is discussed in Section 2.5.3.

Additionally Scientists can use multiple strategies to investigate each map using a set of visualizations as tools for diagnosing the quality of each map. For example, upon accessing the *viewer* button at the bottom of a result, a content aware pop-up viewer is opened with a set of suitable visualizations for the type of the result. Details on how Probe-It! associates specific viewers with a particular data type is discussed in the Section 2.5.3.

2.5.2 *Justification View*

Once a scientists has selected a answer in the query view, a directed acyclic graph (DAG) can be used to visualize the execution trace of a workflow result that generated a map. We call this DAG a justification because it provides information on how a result was generated. An example of such a DAG can be found in Figure 2, which shows a potential justification for one of the resultant maps displayed in the query view. 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. A justification DAG representing an execution trace of a workflow contains only two types of nodes, workflow inputs labeled as "direct assertions" and information transformation services labeled as "Generic Web Service", each of which have their respective output rendered inside the DAG nodes. Workflow inputs may have been provided by a user, software agent, or data sink, and have no incoming edges into their nodes such as requests or queries. 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, illustrating the fact that their results are consumed by other services. For example, the output dataset of the *GEON Gridder Service* is consumed by the *GEON Contour Map Generation Service* which generates contoured renderings of gridded datasets, thus an edge connects these services exemplifying this concept. The DAG root node represents the final service executed by the workflow, generating thus the workflow result. The output of all other transformation services other than the root service are regarded as intermediate or partial results, as they represent the final result in some earlier state.

2.5.3 *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 accessing the view button of 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 3. Probe-It! will visualize the data according to its type, defined semantically, and the set of visualization plug-ins registered with Probe-It! Because the data is defined semantically, more suitable visualizations can be employed. For example, gravity datasets have three associated visualizations: default textual view, plot view, and XMDV view. The raw content view is a table; the ASCII result from gravity database formatted as a table. The location plot visualization provides a 2D plot of the gravity reading in terms of latitude and longitude. XMDV, on the other

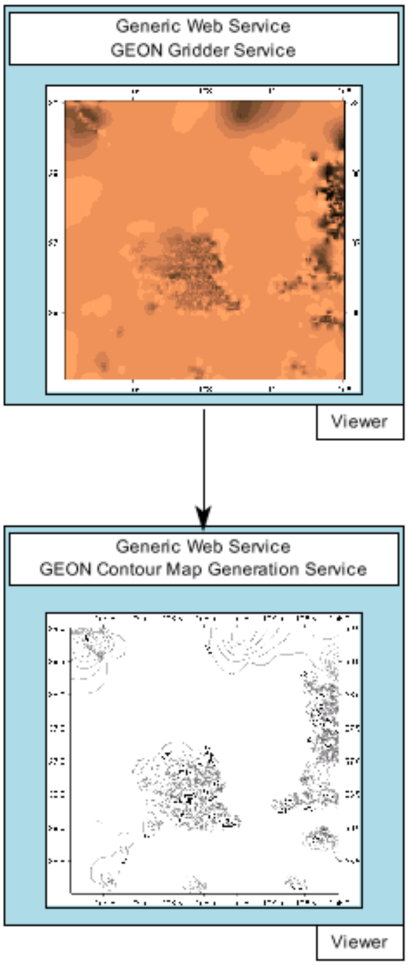

Figure 2: Probe-It! justification view snapshot.

hand, provides a parallel coordinates view, a technique pioneered in the 1970's, which has been applied to a diverse set of multidimensional problems [22]. 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.

The pop-up windows facilitate easier comparison of provenance between different results. 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 being 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.

2.5.4 Provenance View

The *provenance view*, provides information about sources and some usage information e.g., access time, during the execution of an application or workflow. Every node in the justification DAG has an associated provenance description. This information, usually textual, is accessible by selecting any of the aforementioned nodes. For example, upon accessing the input node labeled *gravity database*, meta-information about the database, such as the responsible organization, is displayed in another panel. Similarly, users can access information transformation nodes, and view information about used algorithms.

3. PROVENANCE USER STUDY

The goal of our study is to determine if scientists need provenance to correctly assess the quality of maps. There are many factors that may effect the outcome of our study, such as the level of expertise or familiarity with domain. Since our sample was relatively small, we choose to address two basic hypotheses:

1. Scientists with access to provenance will identify and explain the quality of maps more accurately than scientists without access to provenance
2. Scientists with access to provenance and the ability to compare among alternative maps will identify and explain the quality of maps more accurately than scientists with access to provenance but without the ability to compare

This user study is ultimately bootstrapping a dataset, from which we will be able to infer more interesting conclusions, such as what kind and how much provenance is needed by scientists to accurately determine the quality of maps given their experience or familiarity with the field. From these conclusions, new requirements for both provenance and visualization can be obtained and used to enhance Probe-It! Additionally, because this user study is an ongoing project, the study itself will be enhanced as we gather more data and better understand the limitations of Probe-It! and provenance visualization in general.

The following sections present our subject demographics, discuss our evaluation cases and how the experiment was initiated, and finally the procedure used to elicit and record results.

3.1 Demographics

The 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 provenance visualization, most scientist can identify and explain the quality of some map, regardless of the services and domains of the datasets used to generate the map, granted the scientists has some understanding of what can go wrong in the generation. The user study presented in this paper includes the participation of twenty scientists from various fields including geophysics, geology, biology, environmental sciences, and physics. These scientists are affiliated to various organizations located in Alaska, Arizona, California, Oklahoma, Texas and Brazil.

Table 1 shows the percentage of GIS experts (GISE) and subjects who are not familiar with GIS (NGISE), segmented by education level.

Table 1: Subject demographics in percent.

Education	GISE	NGISE
Complete PhD	30%	30%
Graduate Student	20%	20%

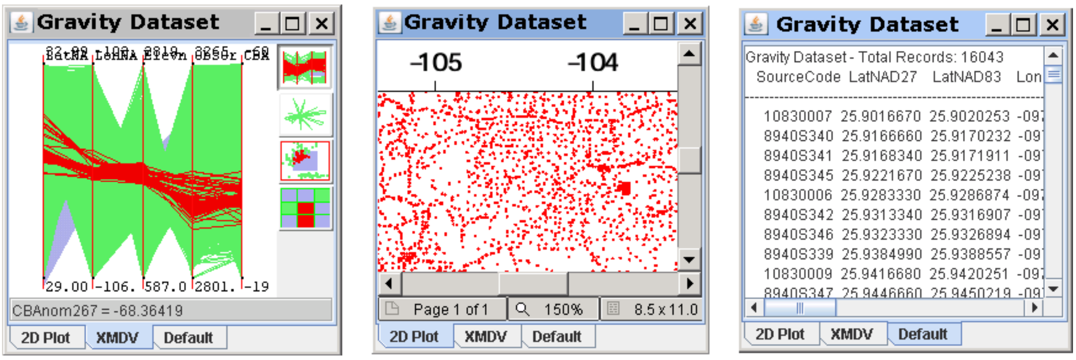

Figure 3: Probe-It! result viewers snapshot.

3.2 Creating Evaluation Cases

Without provenance, it may be impossible for scientists to evaluate the quality of the map generation process and thus determine whether the observed anomalies are naturally occurring or simply errors introduced in the generation process. Our study aims at determining whether or not scientists require provenance in order to correctly identify and explain map quality, thus our user study is focused on generating maps of varying quality and asking subjects to both classify and explain them under different conditions. In order to test our hypotheses, we require at least three different test conditions or cases when subjects perform the identification and explanation tasks:

1. when subjects have only the maps to examine, denoted as case *map*
2. when subjects have both provenance visualized through Probe-It! and the maps to examine, denoted as case *map+p*
3. and when subjects have provenance visualized through Probe-It!, the maps to examine, and as well as the ability to compare provenance visualization among different maps, denoted as case *map+p+c*

These three cases form the foundation of our experiment. The following sections describe the different types of maps that were used for each evaluation case.

3.2.1 Evaluation Map Descriptions

Table 2: Evaluation cases comprising the study.

Cases	Code	Map Type
map	M0	Single source wrong grid params
map+p	M0	Single source wrong grid params
map+p+c	M0	Single source wrong grid params
	M1	Single source correct
	M2	Dual source correct
	M3	Single source random data skew
	M4	Dual source uniform data skew

Table 2 presents the three evaluation cases and the corresponding maps M0 – M4. M0 represents a map that has been generated with poor gridding parameters with respect to the density of the data being gridded. Our employed near-neighbor gridding algorithm requires two parameters, the *search radius* and the *grid spacing* which determines the amount of interpolation needed to derive missing points and resolution of the data respectively. If not specified correctly, these parameters have the potential to either hide many features present in a region or cloud the data with unrealistic points resulting from poor interpolation. For M0 we hard-coded parameters that were unsuitable for the density distribution of the gravity data hosted at the University of Texas at El Paso; this data was used to generate all the evaluation maps. The hard-coded parameters were elicited by scientists who are experts with the subject matter data and were able to provide us with both *good* and poor parameters.

Additionally, maps generated on the cyber-infrastructure have the potential to be derived from datasets coming from multiple different sources. Map M4 was generated by a pair of disjoint datasets, one of which is uniformly skewed by 10 percent, which can be the result of 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.

Errors can also arise even when using only a single dataset that happens to have been derived from faulty sensors, which is exemplified in M3. In this case, random points in the dataset were adjusted to values that are drastically higher or lower in value than the neighboring points effectively simulating faulty sensors. The resultant contour map is usually very dense and full of apparent anomalies. In fact, depending upon the severity of the faulty points, the contour map can become very difficult to read.

Our set of maps also contain those that are supposed to be correct because they were generated using reliable sources of data, using compatible methods, and parameters considered to be correct. M1 is a case where a correct map is generated from a single source. M2 represents the case where the map is generated from two reliable data sources.

3.2.2 Evaluation Case Descriptions

Each evaluation case aims at isolating the effects of our independent variables, which in this study are the ability to browse provenance and the ability to compare results and to compare provenance of different results. The following describes how each evaluation case was designed to test the effects of provenance and comparisons.

Evaluation case *map* consists only a single map M0, which tests whether a scientist can assess the quality of a single map without provenance. This scenario is closer to what you might experience with current cyber-infrastructure based portals such as GEON, where only the map and very shallow meta-information such as the map's author is provided. Similarly, evaluation case *map+p* consists of only M0 as well, except that the user is provided with access to provenance through Probe-It! Together, results from case *map* and *map+p* will allow us to measure the effect provenance has on identifying and explaining map imperfections; provenance is the only differing factor between these two cases.

Evaluation case *map+p+c* consists of maps M0 – M4, which subjects were able to both access and compare the map's associated provenance. This evaluation case best represents what we envision the cyber-infrastructure to provide in the near future, an environment that returns multiple answers for a single request and provides a justification for each answer, leaving the users to discriminate among a set of possibly high and low quality answers. In this evaluation case users have the power of comparison to help them form beliefs about the quality of certain maps (based on similarities between maps for example) and provenance information to validate those beliefs. Together, both cases *map+p* and *map+p+c* will help us determine the effect of allowing users to compare among alternatives.

3.3 Gravity Map Clients: Software Support for Evaluation Cases

The different kinds of contour maps in table 2 are each generated by a specific workflow. Since we wanted to make the experiment as real as possible, we developed three separate client applications, each tailored to configuring the environment associated with an evaluation case. The client applications were meant to simulate a real-world Web application published on the cyber-infrastructure, allowing scientists to request generation of maps on-the-fly by providing a geospatial footprint based on latitudes and longitudes. Although it might have been simpler to pre-generate and store a set of maps based on some hard-coded region, we believed that is was important that the scientist be granted the full functionality they might expect from a scientific portal. Additionally, having each subject specify their own regions might better prevent subjects from sharing answers.

Aside from requesting the generation of the maps comprising an evaluation case, the client applications also controlled the experimental environment (i.e., Probe-It!). Depending on the evaluation case, a client application would limit the available views of Probe-It! For example, upon completion of the workflow, the client application associated with evaluation case *map* invokes Probe-It!, which displays the resultant map with only the query view enabled. On the other hand, the client application tailored to case *map+p* allowed subjects to access every view provided by Probe-It! Finally, the client application associated with case *map+p+c* allowed subjects to access all Probe-It! views as well, except that there are multiple maps that can be compared and contrasted between.

3.4 Procedure

The following section describes how the experiment was conducted, including how the subjects were introduced to the concepts, how the subjects interacted with the client applications, and how responses were collected.

3.4.1 Providing Context

Before completing the evaluation tasks, subjects were provided with background information about cyber-infrastructure and the role of provenance, discussion about the goals of the study, a tutorial on Probe-It!, and finally a description of the tasks to be performed. Because our subjects ranged widely in background and experience with GIS it was necessary to discuss the possible factors that can contribute to the quality of maps that have been generated on environments such as the cyber-infrastructure. Had our sample group only be composed of GIS experts this phase might have been eliminated. However, because of the large percentage of non-experts taking part in our study, we literally gave our subjects a crash course on map making. We felt this was an attempt to *level the playing field* and thus effectively eliminate the experience factor from our experiment.

3.4.2 Running the Experiment

Following the introduction, subjects engaged in the experiment by using each of the three client applications one-by-one to generate the evaluation case maps. For each evaluation case, the subjects were asked to use a unique region and thus generate unique set of maps in an attempt to keep the subjects from leveraging learned knowledge from a previous evaluation case in subsequent cases. For example, if the same map M0 used in evaluation case *map* was used in case *map+provenance*, then subjects would most likely reuse their response.

3.4.3 Collecting Responses

The subjects were asked to complete the identification and explanation tasks of each evaluation case using the speak aloud method. Because of the diversity of our subjects and ultimately their vocabulary, it was necessary to devise a list of correct answer *synonyms* or alternatives especially for the explaining-task of this experiment. To illustrate this necessity, the phrases “wrong grid spacing”, “wrong grid parameters”, “low resolution”, and “pixelated” are all regarded as correct explanations for the imperfection of map M0, and thus could be used interchangeably by our subjects.

All experiments ended in an open discussion in which participants provided the experimenter with comments ranging from noted difficulties with the tasks to possible enhancements to Probe-It!, such as more complex visualizations, that might better facilitate the identification and explanations tasks.

4. KNOWLEDGE PROVENANCE NEED

Each subject usually required about 45 minutes to complete the evaluation. This time included the introduction, completion of evaluation cases, and open discussion. The following subsections describe how scores were computed for each subject followed by the main result of the study.

4.1 Computing Subject Scores

There are two types of scores associated with each evaluation case: an identification score and an explanation score, both of which are used to assess the validity of our hypotheses. An identification score is computed as the average of points earned for correctly classifying the maps comprising an evaluation case. Similarly, the explanation score is computed as an average of points earned for correctly explaining the map imperfections. Because the measure of identifying and explaining maps is a binary value (e.g., 0 for incorrect answers and 1 for a correct answers), both types of evaluation scores are always between 0 and 1, inclusive. Table 3 illustrates this scoring technique for evaluation case *map+p+c*, in which a subject correctly classified map M0, M2, and M3, earning an identification score of 0.6. Additionally, the subject earned an explanation score of 0.4, due to only being able to correctly explain M0 and M3.

For evaluation cases *map* and *map+p*, only a single map M0 is presented to the user, thus no averages need to be taken; the scores are equal to the points earned for identifying and explaining the single map.

Table 3: A subject's score table for evaluation case 3

Map Code	IDs	Explanations
M0	1	1
M1	0	0
M2	1	0
M3	1	1
M4	0	0
Average	0.6	0.4

4.2 Statistical Validity

Significance of the the results were verified by a single-tail t-test at 95% confidence.

4.3 Identifying Map Imperfections

Table 4: Subjects' average accuracy in identifying map imperfections.

Evaluation Cases	*map*	*map+p*	*map+p+c*
Average Score	0.1	0.79	0.94

Table 4 contains the average accuracy of scientists in identifying maps quality. Condition cases *map* versus *map+p* tested whether provenance was needed in order to correctly assess maps; both cases are based on the same map containing the same error with the ability to access provenance in condition *map+p* being the only difference. 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 earning 0. After the scientists were able to access the provenance, both their accuracy and confidence in determining the quality of the map improved significantly given the number of test subjects.

When scientists were granted the capability to visually compare different maps and their corresponding provenance, as in case *map+p+c* there accuracy seemed to only improve marginally. However, given our small sample and the relatively small difference in accuracies between the means of *map+p* and *map+p+c* we cannot conclude at this time that providing users with both comparisons and provenance will allow them to better identify map imperfections than users with only provenance. However, from an observation point-of-view, it appeared as if subjects could more confidently complete the tasks given a set of maps rather than only a single map. 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 usually required that subjects identify the features shared by the majority and segregate the *odd* maps which do not share those characteristics. These odd maps were usually quickly regarded as incorrect by the scientists. For example, map M4 was usually the first map to be regarded of low quality because of its unique fault line as described in Section 3.2. 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 map M3, which represents a contour map generated by a single dataset containing randomly skewed data points, simulating the case when a map is generated using bad data. Without the ability to compare this map with others, and its associated provenance, 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 knowledge of both gravity data and the particular region being mapped.

4.4 Explaining Map Imperfections

Table 5: Subjects' average accuracy in explaining map imperfections.

Evaluation Cases	*map*	*map+provenance*	*map+provenance+cor*
Average Score	0.05	0.78	0.8

Table 5 presents the average accuracy of subjects in correctly explaining why a 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 should be regarding the correctness of the maps.

In general however, the data shows a major improvement in the explanations with the use of provenance as exemplified in the condition cases *map* and *map+provenance*. With provenance, the scientists were able to browse the intermediate results that they believed to have the most impact on the quality of a map and formulate an explanation for why the map is correct or not. Given our sample size, we can say that there is a significant difference between the two sample means, thus supporting the hypothesis that scientists can more accurately explain map imperfections, given provenance, than scientists without provenance.

Positive results were also yielded when scientists were able to compare between the different maps and their provenance, as in condition case *map+provenance+comp*. 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. Unfortunately, given our sample of subjects and the relatively small mean different between scientists who could compare and scientist who could not, we cannot substantially infer that the ability to compare against alternative maps yields better accuracy than when only evaluating single map.

In almost all cases, scientists used provenance related both to the process or execution trace of the workflow and some intermediate result to complete their tasks. The results recorded for map M4 support this claim the best; the map usually contained a prominent fault line where the two datasets were merged. This feature was most noticeable in the visualized form of the gridded data set, resulting from step 3 in Section 2.1. 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 provenance 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. Although the data presented below does not indicate that users need provenance visualization to identify and explain map imperfections, it does indicate that the use of the more elaborate visualizations were generally preferred.

Table 6: Probe-It! feature usage (* indicates raw viewer).

Feature	Viewer	Usage %
Process trace	DAG	85
Gravity dataset	ASCII tabular*	20
	XMDV	20
	2D Plot	40
Grid dataset	ESRI dataset*	0
	color grid	65
Contour map	postscript*	0
	contour image	90

Table 6 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 provenance 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 provenance 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 provenance 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, we believe that provenance is only as useful as the visualization overlaid on it.

6. DISCUSSION AND RELATED WORK

The uses of provenance 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 use fits all. 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 [9]. 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 result 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 workflow executions. 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 [23]. From these recordings, scientists are able to operate in three basic modes: (i) debug, (ii) check validity, and (iii) update mode, which refer to situations when, a result is of low quality and the source of error must be identified, when a result is novel and must be verified for correctness, or when a workflow has been updated and its previous versions are requested. Based on particular user roles, the appropriate dimension of provenance is presented, knowledge, organization, data, or process level [23]. MyGrid is yet another system that supports different tasks or uses of provenance, thus there are multiple "modes" that users can operate in that effectively show only provenance relevant for a particular task. We believe that all levels of provenance are required in order for scientists to identify the quality of complex results.

Currently, MyGrid RDF provenance is viewed using Haystack [18]. 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 the specific user browsing the provenance 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 [10]. Upon user requests, ESSW leverages a suite of Notebook tools that can display both the scientific result and the associated provenance. Stored scientific visualizations such as swaths [10] are rendered in HTML upon request; the request is in the form of a query. Additionally, ESSW leverages GraphViz [12] 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, similarly to how Probe-It! renders justifications.

Karma [19] 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 [21]. 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 result, 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 provenance 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.

This study is an attempt to verify the effectiveness of new methods for quality assessment other than the more traditional approaches such as uncertainty propagation or error-model development, which we believe are complimentary to provenance visualization. In 1991, the National Center for Geographic Information and Analysis (NCGIA), held a four day meeting, in which GIS and spatial data expert met and tried to come to a consensus on the definition spatial quality and the different factors that contribute to it. Additionally, the participants expressed their thoughts on how quality could be visualized; some of the participants even suggested that users of GIS systems, as well as being able to visualize target data, should also be provided with visualizations of the error model associated with that data. However, the majority felt that when applicable, datasets should be visualized as some graphic or image rather than in its raw tabular for, thus enabling researchers to identify aberrant data more rapidly [11].

7. CONCLUSIONS

This paper described the use of Probe-It!, a knowledge provenance visualization tool, to support a user study about provenance generated from a real end-to-end, cyber-infrastructure based application and the effect of this provenance in identifying and explaining map quality. The evaluation demonstrated that most scientists are unable to correctly classify a map as poor or of high quality if no knowledge provenance is provided with the maps. With the use of provenance, however, the study showed that most scientists performed significantly better in classifying the maps. Moreover, the study demonstrated that most scientists could identify the factors leading maps to be of low quality with the help of provenance information better than scientists who did not have access to provenance. Unfortunately, at this time, we are unable to claim anything about the effects of comparison in the tasks of identifying and explaining map imperfections. As our study and evaluation data mature, however, we believe that the benefits of comparison will be revealed along with several aspects of provenance covered in the experiment but not discussed in this paper due to the lack of statistical significance.

State-of-the-art cyber-infrastructure-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 as most of them are already aware of the necessity of cyber-infrastructure-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] E. Deelman and et al. Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming Journal*, 13(3):219–237, 2005.

[9] 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).

[10] 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.

[11] B. B. Kate Beard and W. Mackaness. Visualization of the Quality of Spatial Information: Closing Report. Technical report, University of Maine and University of New York, 1994.

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

[13] B. Ludašcher and et al. Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice & Experience*, 2005. Special Issue on Scientific Workflows.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] W. H. F. Smith and P. Wessel. Gridding with continuous curvature splines in tension. *Geophysics*, 55:293–305, 1990.

[21] 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.

[22] Z. Xie. Towards Exploratory Visualization of Multivariate Streaming Data. http://davis.wpi.edu/.

[23] 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.