GeoPWProv: Interleaving Map and Faceted Metadata for Provenance Visualization and Navigation

Ziheng Sun, Peng Yue, Lei Hu, Jianya Gong, Liangpei Zhang, and Xianchang Lu

Abstract—Visualization of geospatial data provenance aims to provide a user-friendly way for easy navigation and increased understanding of the derivation history of scientific results. Most existing work focuses on provenance modeling and management. This paper proposes to interleave map and faceted metadata for geospatial data provenance visualization and navigation. It shows how provenance in Web-geoprocessing workflows can be visualized at four levels: feature, dataset, service, and knowledge. A prototypical system, named GeoPWProv, is developed to demonstrate the applicability of the approach.

Index Terms—Faceted metadata, geoscientific workflow, geospatial data provenance, provenance visualization, web map.

I. INTRODUCTION

GEOSCIENTIFIC workflows are increasingly used in e-Science or cyberinfrastructure environments. Often, distributed data and geoprocessing services are used in scientific workflows to derive geospatial data products. Consequently, provenance has been identified as an important issue in understanding and scientific reproducibility of scientific results [1]–[3]. Most existing work focuses on the provenance modeling and management instead of provenance visualization [1], [4]. Provenance visualization aims to provide a user-friendly way for easy navigation and increased understanding of the derivation history. Although there are several workflow systems that can support the provenance visualization [5]–[7], they are intended to be general. Special efforts are still needed in the geospatial domain.

This paper proposes to interleave the map and faceted metadata for the geospatial data provenance visualization and navigation. Map has been identified as the primary way on the visual representation of geospatial data products. Geospatial features and feature types/datasets are rendered on maps. Map is used as the major entry point for geospatial data provenance visualization in this paper. It is then combined with the hierarchical faceted metadata [8] for provenance navigation. Metadata is hierarchically faceted when it is composed of orthogonal sets of categories in hierarchies. It is often used to facilitate information navigation and exploration [28]. Therefore, it is adopted in this paper to navigate the geospatial data provenance. Once specific source data items are located using the faceted metadata approach, they can be visualized on the map to improve the understanding of sources. Thus, the faceted metadata and map are interleaved to facilitate the provenance navigation and visualization. Such capabilities are coupled with a Web-geoprocessing workflow system in the paper for navigating and visualizing the geospatial data provenance at four levels: feature, dataset, service, and knowledge. A prototypical system, named GeoPWProv, is developed to demonstrate the applicability of the approach. It allows users to visualize sources continuously in the map and navigate freely among different levels of provenance. On the one hand, it offers an intuitive interface for understanding the origin, such as by comparing features and their sources in the map. On the other hand, it provides possibilities to allow users interact with the processing flow, for example, by changing the inputs or updating the geoprocessing models.

The rest of the paper is organized as follows. Section II introduces the related work. Section III describes a framework to support provenance navigation and visualization in Web-geoprocessing workflows. Section IV presents a prototypical system that implements the framework. Finally, conclusion and future work are provided in Section V.

II. RELATED WORK

Provenance has been studied in computer science for over two decades [1]. As the Web technologies have been widely used, provenance is even more important in the recent information infrastructures. One of the earliest studies on provenance in geospatial domain can be traced back to Lanter’s work in geographic information system (GIS) [9]. Yue et al. proposed a catalog service-based approach to organize and register geospatial data provenance by extending the ebXML registry information model (eBRIM) [10]. The semantic web techniques, like web ontology language (OWL) and resource description framework (RDF), could be used to represent semantic provenance [11]. By tracing geospatial metadata through semantic service chains, it is possible to provide the capability of automatic geospatial data provenance capturing [12]. Some cases, like the geological carbon sequestration [13], are experimented to support the spatial provenance.

Visualization is an important way for users to understand scientific results such as large images and maps [14]. The combination of visualization and provenance techniques can increase scientists’ understanding of results [15]. The visualization of provenance has been studied in several application domains. Ram et al. [16] have developed a system called PROVISIA to visualize the provenance in Wikipedia from the following five aspects: what, where, how, why, and who.
Cheung et al. [17] implement a secure provenance visualization tool, Provenance Explorer, to dynamically generate customized views of scientific data provenance according to viewers’ requirements and authority levels. Silva et al. [18]–[20] present an open-source provenance-management system, named Vistrails, which provides a framework for the data exploration and visualization. Probe-it! is a provenance-aware system and can be used to visualize knowledge provenance for map imperfections [21]. Porras et al. [22] describe a caching system in Probe-it! to enhance the performance of the tool when browsing proofs including images. CI-Browse-It is another tool for visualizing scientific provenance using semantic web techniques [15]. Karma allows visualization of the provenance collected from workflows [23], [24]. Kunde [25] discusses concepts of visualization in provenance and implements a software for validation of them. However, most work is on the general information domain or other application domain such as biological domain. The work in this paper focuses on the provenance in the geospatial domain. Map and features are concepts specific to the geospatial domain. In addition, the work allows the provenance navigation and visualization at four levels, which has not been addressed before.

Metadata of both geoscientific data and derivation workflows could be linked together to provide an informed understanding of provenance for the final data products. In the Web-based scientific computation domain, metadata has been widely used for the Web information search, and the faceted metadata approach is proposed to make the searching style more flexible [26]. In the approach, metadata is faceted according to orthogonal sets of categories [8]. For example, in the geospatial domain, possible facets might be the file formats (GeoTIff, ESRI Shape, HDF, IMG, etc.), themes (fire, landslide, earthquake, etc.), owners’ names, owners’ organizations, time periods, and so on. Kauppinen et al. [27] used an ontology model and metadata schema, which can be used to record historical changes of municipalities in Finland to support the development of a semantic cultural heritage portal. The portal allows the semantic faceted search of content and visualizes historical divisions of different eras on overlapping maps. Nagel et al. [28] propose a framework for browsing information interactively by selecting and filtering various metadata facets through tangible geovisualization of maps. However, none of them have been addressed in the context of provenance and geoprocessing workflows. In our paper, the geospatial data provenance is organized in orthogonal categories of several themes aiming at flexible visualization and navigation of provenance through metadata facets.

III. FRAMEWORK TO SUPPORT GEOSPATIAL DATA PROVENANCE NAVIGATION AND VISUALIZATION

This section proposes a framework to support navigation and visualization for geospatial data provenance by integrating map and faceted metadata into an existing Web-geoprocessing system [8] (see Fig. 1). The system leverages geoprocessing Web services and workflows for geoprocessing modeling, workflow execution, and provenance tracking. The framework for the extended provenance-aware system is composed of four modules: provenance recorder, provenance storage, provenance finder, and provenance exhibitor. They are introduced as follows.
be captured by respective collectors, organized as the faceted metadata, and delivered to a provenance store.

B. Provenance Storage

The provenance collected from provenance recorder is stored in a provenance store. The provenance store could be implemented using RDF stores. RDF and its extension OWL provide capabilities for description and modeling of concepts and information contained in Web resources. The provenance model follows the design in [12]. Table II shows an example of encoding using RDF.

The provenance can be stored into a database by RDF tools. Many RDF tools support automatically creating tables and indexes for RDF, storing RDF files persistently, and querying them using SQL/SPARQL. Provenance databases can be published through Web servers.

C. Provenance Finder

This module is responsible for processing users requests. It contains three parts: provenance manager, RDF database manager, and datasets transformer. The RDF database manager is used to communicate with the RDF database, such as creating a connection to the database, transferring provenance through the connection, and parsing the information into RDF objects. It is invoked by the provenance manager that processes all requests from browsers. The datasets transformer retrieves datasets using resources in RDF (e.g., a URL address to an ESRI Shapefile) and transforms them into formats that can be acceptable by the Web map. Thus, feature-level provenance in fields of each feature could be extracted from datasets, and transferred to Web maps through the Provenance Manger.

D. Provenance Exhibitor

The module is the key component for provenance visualization and navigation. Most existing approaches use text-table and workflow diagrams as major modes for provenance visualization. Maps are an intuitive way for geospatial data visualization. Some Web map providers such as Google Map and Bing Map have provided user-friendly maps for information Mashup [31]. They can be used as the major entry point for visualizing geospatial data provenance. Each geospatial dataset can be rendered as a layer overlapping on the base map layer in the Web map. The Web map’s render mechanisms are applied to render geospatial features using various colors or symbols. Each feature could have its own provenance. The Web map’s event mechanisms are used to add the Click event for geospatial features. If a feature is clicked, a table window listing its provenance will pop up. Considering the dataset-level provenance, when users are interested in one of its ancestor
datasets, they can click the corresponding View button, which will add the ancestor dataset as a new layer overlapped on the same Web map. The comparison between the dataset and its ancestors will become easier in such an intuitive way. The potential of Web maps could be further explored. If users click the View button related to the feature-level provenance, the Web map API can be called to highlight the source features in several times. Thus, the interface is intuitive in discovering ancestor features and help track the history in feature processing. We also use markers to label locations of servers on which geospatial Web services are deployed. When users click a server marker on the Web map, the metadata of the server and deployed services in the service-level provenance will be displayed in a table window. This provides potential for service location trace in the derivation history.

The Web map method provides an intuitive environment to geoscientific users for understanding geospatial data products and their ancestors, allowing them to explore data and origin in a familiar environment. Meanwhile, faceted metadata can facilitate the provenance navigation and workflow diagrams showing that geoprocessing model is intuitive in showing the production pipeline. Therefore, panels for graphical geoprocessing modeling and faceted metadata are provided and linked to Web maps. This allows users to navigate freely among the faceted metadata, workflow modeling, and scientific results’ visualization. For example, considering the knowledge-level provenance, after clicking the View button for the related process model in the popup table listing the provenance, a graphical geoprocessing modeling interface showing the process model will appear. The interface is from an existing Web geoprocessing system [40]. It allows users to edit, instantiate, and execute geoprocessing models. These functions make the selection of alternative models possible. On the other side, the faceted-metadata panel interacts with the Web map continuously. It assists users in locating sources or data products in a facet-directed way, and could load data into the map as additional layers. In addition, the panel also provides text-table-based information showing details of the selected dataset.

IV. IMPLEMENTATION

The framework is implemented in a prototypical system, named GeoPWProv. A case is introduced and helps to demonstrate the system.

A. Use Case

Suppose that John is an employee of the Bureau of Land Management in the United States. In a project on building a new Central Business District (CBD) in an old town, he needs to create a reconstruction map of the region of interest. He has a vector file about some buildings in that area. However, the data are outdated since it was collected in 2005. The file’s format is ESRI Shapefile and named “Building1.”

He got a new vector file (named “Building2”) that is digitized from a satellite image in 2010, Building2 has some updated features yet less complete than Building1. He wants to conflate Building1 and Building2 to generate an updated building file (named “Vector1”). Meanwhile, the project-funding agency provides him with the polygon, covering the area planned for CBD. The polygon is stored in an ESRI Shapefile (named “PlanningCBD”). The data have the spatial projection ESPG 900913. He must transform the PlanningCBD’s projection to ESPG 32618, which is the same as Building1 and Building2.

The reprojection result of the PlanningCBD is “Vector2.” Then he will do a spatial buffer analysis on Vector2 using a distance (50 m). The buffer result is named “Vector3.”

Buildings in the buffer will be reconstructed for the CBD. The “Select-Overlaps” process works on Vector1 and Vector3 to select buildings in the buffer. The selection result will overlay on the road network to generate the final reconstruction map. In a service-oriented environment, geoprocessing services such as spatial analysis services from GeoPW [32] can be used. Fig. 2 shows the process model for the workflow.

Once the workflow helps John to get the processing work done, John would like to evaluate results and is interested to know the provenance of data, in particular some specific features. GeoPWProv can provide such capabilities to him in an intuitive and user-friendly way. Section IV-B briefly describes the implementation of GeoPWProv, and then how GeoPWProv can help John solve his problem is illustrated in Section IV-C.

B. System Development

GeoPWProv is an extension of an existing Web geoprocessing system [39], with ontologies from [31], [40]. The MySQL [33] is used as the RDF database. We apply Jena API [34] to encode the collected provenance’s traces\(^1\) using RDF, save them into a MySQL database, and query the database for provenance discovery. OpenLayers [35] is an open source and professional Web map that is already widely used for various academic and commercial purposes. We use OpenLayers as the client Web map tool and Google

\(^1\)Since the paper focuses on the provenance visualization and navigation, it is assumed in the implementation that geoprocessing services can provide feature level provenance in the output dataset. An investigation is still needed in the future on how the internal implementation of services can provide feature-level provenance.
street map as the base layer. GDAL/OGR [36] is a translator among various formats of raster and vector geospatial datasets. It is used in GeoPWProv to transform geospatial datasets into formats like Geography Markup Language (GML) that OpenLayers can accept and load. The client functions are realized by JavaScript. The layout, tables, and styles of pages are designed using Hypertext Markup Language (HTML) [37] and Cascading Style Sheet (CSS) [38]. An example provenance model in OWL is published online (http://geopw.whu.edu.cn:8090/ontologies/tasks/provenance.owl). Users can import it to describe their own geospatial data provenance. In the ontology, the knowledge provenance applies geospatial process models as task models to describe the problem-solving knowledge.

C. Evaluation

John uses the search function in GeoPWProv and finds the final product dataset in HTML pages. He then clicks the View-in-map button. The datasets will be downloaded to the server, and the GDAL/OGR will transform it from ESRI Shapefile into GML. The browser will receive a URL address of the generated GML and create a GML layer for it in OpenLayers. OpenLayers will load the GML to the browser automatically and render it using predesigned styles. Fig. 4(b) and (c) show the interface of GeoPWProv for browsing the provenance at the feature level. Fig. 4(e) can be used to understand the provenance at the knowledge level. The dialog in Fig. 4(h) shows the provenance including parameter values for each node at the service/execution level. Fig. 4(a) shows the interface for users to query geospatial metadata and provenance using facets. The visualization in Web maps provides an intuitive interface for users to track the lineage of features.

Fig. 4 shows a general process on using GeoPWProv to guide users in viewing provenance from different perspectives and levels. Fig. 4(a) illustrates that the final reconstruction map can be found using the faceted metadata. After clicking the “View in map” button, the reconstruction map will be overlapped on Google map in OpenLayers [see Fig. 4(b)]. If a feature in the new layer is selected, a table-listing provenance will popup. In the table, the ancestors can be accessed in the View buttons. As Fig. 4(c) shows, once ancestor data are added, users can contrast features of the final map with its ancestors to find changes. Once users want to know the workflow model used in deriving the final map, the table allows him to jump to the graphical modeling page showing the process model that includes process nodes and input/output parameters [shown in Fig. 4(e)]. Fig. 4(f)–(h) show that users could also check which service or service chain was used and what data and parameter values were used in the execution. The navigation among different levels of provenance provides possibilities for reproducing the geoprocessing workflows, e.g., selecting different geoprocessing services [Fig. 4(g)] or choosing alternative process models. For example, John could choose to do the selection process before the conflations process, which could improve the performance of the workflow when the size of the data is large.

V. CONCLUSION

Provenance visualization can help better understand the derivation history of scientific results. This paper proposed a framework for visualization and navigation of provenance in Web geoprocessing workflows by interleaving Web maps and faceted metadata. The visualization of geospatial data provenance can happen at the feature, dataset, service, and knowledge levels. The Web map provides an intuitive and user-friendly interface for visualizing the feature-level provenance, and serves as the entry point for geospatial data provenance navigation. Faceted metadata show different aspects of provenance, and allow flexible navigation of provenance-related information. Interleaving faceted metadata and map together allowed users to locate sources or datasets effectively and then visualize them on the map to digest source information in an intuitive way. The coupling of provenance visualization with the Web-geoprocessing workflows provided possibilities for reproducing these workflows using different levels of provenance and creation of a more flexible system for the Web-geoprocessing workflows. The framework was implemented in an existing Web-geoprocessing workflow system.

Future work will develop a robust tool that fully supports provenance-aware geoprocessing workflow reproduction. Such a tool could track provenance automatically. It runs in Web browsers and would allow users to change data or parameter values to rerun the workflows, select different services when necessary, choose alternative process models to assess their fitness, and visualize provenance and geoprocessing results on the fly. More data formats including raster-based data formats will also be supported in the tool in the future.
ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers for their valuable comments.

REFERENCES


Zhiheng Sun is currently pursuing the Ph.D. degree with Wuhan University, Wuhan, China.

Peng Yue is a Professor with Wuhan University, Wuhan, China.

Lei Hu is currently pursuing the Ph.D. degree Wuhan University, Wuhan, China.

Jianya Gong is a Professor with Wuhan University, Wuhan, China.

Liangpei Zhang is a professor with Wuhan University, Wuhan, China.

Xianchang Lu is currently pursuing the M.S. degree with Wuhan University, Wuhan, China.