Geoscience Data Provenance: An Overview

Liping Di, *Senior Member, IEEE*, Peng Yue, *Senior Member, IEEE*,
Hampapuram K. Ramapriyan, *Senior Member, IEEE*, and Roger L. King, *Senior Member, IEEE*

***Abstract*—The advancement of Earth observing sensors, data, and information systems enhances significantly the capabilities to access and process large volumes of geoscience data, which are often consumed by scientific workflows and processed in a distributed information environment. Consequently, data provenance becomes important since it allows users to determine the usability and reliability of data products. Motivation for capturing and sharing provenance also comes from the distributed data and information infrastructure that has been benefiting the Earth science community in the past decade, such as spatial data and information infrastructure, e-Science, and cyberinfrastructure. This paper provides an overview of geoscience data provenance in supporting provenance-aware geoscience data and information systems by summarizing state-of-the-art technologies and methodologies of geoscience data provenance and highlighting key considerations and possible solutions for geoscience data provenance.**

***Index Terms*—Cyberinfrastructure, geoprocessing workflow, geoscience data provenance, geospatial service, lineage, preservation.**

I. Introduction

With the advancement of sensor and platform technologies, the capability for collecting geospatial data has increased significantly in recent years. More than 150 Earth observation satellites are currently on orbits measuring the state of the Earth system [1]. These satellites, together with countless air-, land-, and water-based sensors and monitoring systems, are generating large volumes of geospatial data. For example, the data managed by the National Aeronautics and Space Administration (NASA)'s Earth Observing System Data and Information System (EOSDIS) are multiple petabytes in volume and rapidly growing. In the meantime, data systems are evolving to support science data processing and production of high-level data products automatically and archiving and distribution to a diverse user community [2]. For example, from October 1, 2010 to September 30, 2011, the NASA EOSDIS alone served more than 13 TBs of data per day to a community of over 1.2 million users. The large volume and variety of Earth science data are often consumed by scientific workflows involving multiple complex geoprocessing steps in different contexts at different times. Consequently, the scientific and application communities are increasingly interested in the geoscience data provenance, which provides the lineage of a data product, the important information for users to determine the usability and reliability of the product. In the science domain, the data provenance is particularly important since scientists need to use such information to determine the scientific validity of a data product and to decide if such a product can be used as the basis for further scientific analysis. It can be further used to address a series of research issues, including transparency in data sharing and processing, proper credits to data and algorithm contributors, interoperability, and reproducibility and trustworthiness of scientific results.

Traditionally, Earth science data products are produced in the science data centers with predetermined processing procedures or workflows. In the distributed information infrastructure that has been benefiting the geoscience community in the past decade, such as spatial data and information infrastructure, e-Science, and cyberinfrastructure, sensor observation data and higher level derived products are generated, transformed, published, and disseminated frequently. There can be a mix of product generation using "standard" workflows and workflows generated "on the fly." In such a data-rich, network-based, and diverse production environment, provenance information is even more important since distributed services and inputs provided by diverse providers or sensors are engaged dynamically. It is essential to track and share provenance in such a distributed environment. This is further emphasized by fact that the U.S. National Science Foundation (NSF) "Earth Cube" initiative has identified provenance as a research priority in the development of the geocyberinfrastructure [3].

This paper provides an overview of concepts, technologies, and methodologies related to geoscience data provenance. It will not only contribute to the understanding of data provenance in the geospatial context but also provide possible solutions toward provenance-aware applications in the geoscience domain. The remainder of this paper is structured as follows. Section II provides the definition of geoscience data provenance; Section III outlines general considerations on provenance-aware applications and summarizes a number of major works in the literature; Section IV introduces related work on provenance in the geoscience domain; and Section V identifies some key

Manuscript received October 9, 2012; revised December 28, 2012; accepted December 31, 2012. This work was supported in part by the National Basic Research Program of China under Grant 2011CB707105, by the U.S. Department of Energy under Grant #DE-NA0001123, and by the National Natural Science Foundation of China (Project 41271397).

L. Di is with the Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA 22030 USA (e-mail: ldi@gmu.edu).

P. Yue is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: geopyue@gmail.com).

H. K. Ramapriyan is with the Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD 20771 USA (e-mail: hampapuram.k.ramapriyan@nasa.gov).

R. L. King is with the Center for Advanced Vehicular Systems, Mississippi State University, Starkville, MS 39762 USA (e-mail: rking@cavs.msstate.edu).

Digital Object Identifier 10.1109/TGRS.2013.2242478

considerations for geoscience data provenance and discusses possible solutions. Finally, Section VI discusses the conclusions of this paper.

II. What is Geoscience Data Provenance?

The term provenance means the origin or source of something [4], [5]. In the context of a database system, data provenance is defined as the description of the origins of a piece of data and process by which it arrived in the database [6]. Such provenance for relational views in the database system is referred to as a view data lineage problem, where the origin of the data is associated with the base data items or tables and the process is associated with the relational algebra operations that yield the database view data [7]. In the context of scientific workflow, data provenance records workflow processing steps and their inputs/outputs that contribute to the production of the final data products [8]. In the context of the Web, provenance is information about entities, activities, and agents that are involved in producing Web resources such as documents and ontologies [9], [10].

Some of the earliest works on geospatial data provenance can be traced back to the late 1980s or early 1990s. The work by Lanter [11] referred to data provenance as the lineage of map layers in a geographic information system (GIS). The Spatial Data Transfer Standard of the Federal Geographic Data Committee defined a lineage model for geospatial data. In the International Organization for Standardization (ISO) 19115 Geographic Information—Metadata (ISO 19115:2003) standard, the lineage of geospatial data is defined as "information about the sources and production processes used in producing a data set." The production process could be a single step of geoprocessing or a large aggregated geoprocessing. A detailed lineage metadata model was defined in ISO 19115:2003 and further expanded in ISO 19115-2:2009. Hereafter, in this paper, the terms lineage and provenance are used interchangeable, although lineage is often used in the context of ISO or other geospatial domain standards.

For the purposes of this paper, we define geoscience data provenance as the derivation history (lineage) of a geospatial data product. The lineage could be in the workflow context, Web context, or both. In conventional geoscience data archiving and distribution systems, scientific workflows are used extensively to produce Earth science data products. Examples of contents in the lineage include algorithms used, the process steps taken, the computing environment run, data source input to the processes, the organization/person responsible for the product, etc. From the emerging neogeography perspective [12], provenance information includes the citizens' map-making behavior and various location-related sources or volunteered geographic information involved in the crowdsourcing. In this context, provenance can help track how data with various quality can be synthesized to assure quality. In the Earth Observation Sensor Web environment [13], when rich streams of various sensor measurements are filtered, corrected, combined, and divided, the provenance information such as sensors or observation correction/processing chains can be used to enhance the fusion of multisensor data.

III. Provenance-Aware Applications—General Considerations and Literature

Provenance has been addressed actively in the e-Science or cyberinfrastructure in the past several years. The U.S. NSF task force report on grand challenges for cyberinfrastructure suggests that a robust persistent data infrastructure should include the data provenance component [14]. There are several well-known international forums on provenance, including the International Provenance and Annotation Workshop (since 2002), the Workshop on the Theory and Practice of Provenance (since 2009), the Provenance Challenge Workshop (2006–2010), and the International Workshop on the Role of Semantic Web in Provenance Management (since 2009). The World Wide Web Consortium (W3C) Provenance Working Group is working on defining a language for exchanging provenance information for Web resources [15]. In the Earth science domain, there has been also an increased interest in recent years on incorporating provenance support in geoscience data systems, in particular, the distributed data infrastructure. This is evidenced by studies presented in the 2010 American Geophysical Union Informatics Session on Encouraging and Enabling Transparency in Science Data and the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2011 special session on provenance in geoscience data. In addition, the Data Stewardship Committee of the U.S. Federation of Earth Science Information Partners (ESIP) is working toward proposing a Provenance and Context Content Standard (PCCS) enumerating items to be preserved along with data to ensure future understanding and reproducibility.

Here, we outline the general considerations on provenance-aware applications (Fig. 1). These considerations provide a basis for understanding similarities and differences of available provenance-aware applications and unify the terms used in the rest of this paper.

A. General Considerations

1) Provenance Representation: Provenance systems in different application domains have their own provenance representations tailored to their own specific needs. A representation includes the model for provenance and its implementation syntax. The model should allow dependence relations to be tracked and possibly derived, among data products and transformation processes. Some existing examples of provenance models include the Open Provenance Model [16], W3C PROV Data Model (PROV-DM) [9], and ISO 19115 lineage model.

2) Provenance Capturing: The provenance information could be captured and recorded manually or automatically. Instead of relying on manual work, it is particularly important in the distributed information infrastructure to automate the provenance capturing simply because of large volumes of data and frequent processing in the open environment. Automating provenance capture requires extending legacy applications with provenance capturing functions at either the workflow engine or individual services to add provenance annotations.

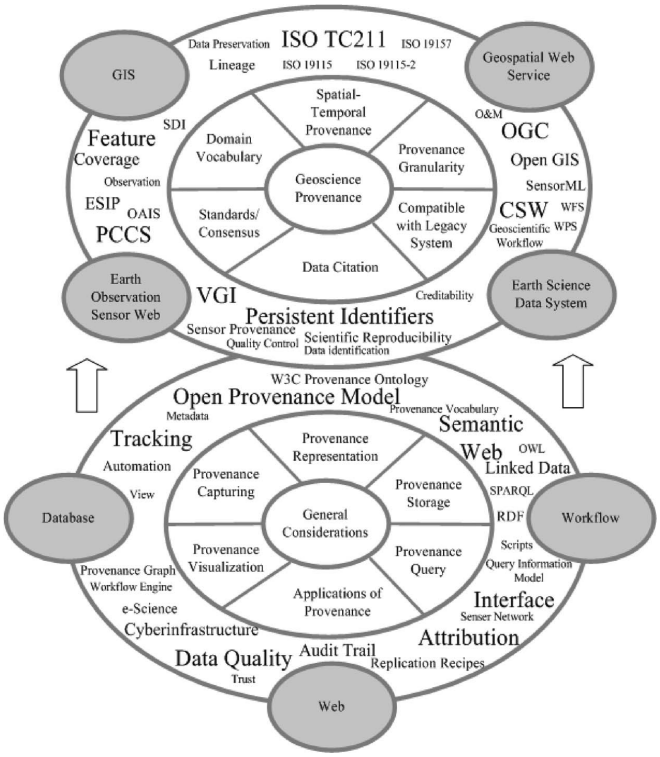

Fig. 1. Overview of geoscience data provenance.

3) Provenance Storage: Provenance information, usually considered as a part of metadata, can be tightly coupled with the rest of the metadata. An existing metadata catalogue can be used for storing and managing the provenance information and providing the information to users with the data products. It can also be managed using a separate storage system, which may be called the provenance store. Both approaches should support the distributed storage of provenance information.

4) Provenance Query: The design of the provenance query should take into consideration the query interface, language, and queryable and returnable provenance content. The query interface specifies protocols and operations for provenance queries. The query language, such as Structured Query Language (SQL), provides a set of predicate and data types. Queryable and returnable provenance content depends on the model and representation of provenance information. The implementation of the provenance query should support dispatching queries to distributed provenance stores and linking query results from multiple stores together before sending them back to clients.

5) Provenance Visualization: Provenance visualization allows general users to have a more informed understanding of provenance information. It should support easy navigation between provenance and data products. Visualization of scientific results can be combined with relevant provenance information to help scientific users discover anomalies and evaluate results.

6) Applications of Provenance: The application of provenance is diverse. Simmhan *et al.* suggested a list of applications of provenance information: data quality, audit trail, replication recipes, attribution, and informational [17]. Understanding data quality is the primary application of provenance in scientific domains. The transformations and base data included in the provenance information can assist users in evaluating the quality of the data based on specific quality metrics. Provenance can serve as a means to audit the trail of execution and help locate errors or exceptions. An entire workflow with detailed description of intermediate transformation steps, stored as the provenance information, can act as a recipe to reproduce a particular data product on demand instead of transporting or storing it. Attribution means that the intellectual property of contributors or copyright can be identified through provenance information. By interleaving provenance information and products together, exploitation and interpretation of geoscience products can be more informative.

B. Existing Works

A number of existing works have contributed to methods on provenance-aware applications [5], [17]–[19]. These methods can be classified into four major categories: database, scripts, services, and Semantic Web.

1) Database: Lineage in database concentrates on transformations, such as queries or functions, performed on the base data, which ultimately create a view, a table, or a data item in a database. Such transformations could be registered and inversed to trace the lineage from a final data product back to its source. For example, to update or delete a view, we can identify the source tables by using inversions. Inversion method is identified as a typical method for provenance applications in database [6], [7].

2) Scripts: Scripts are widely used by scientific community for data processing. Workflow scripts can compose executable commands or scripts together to perform complex data analysis functions. Their executions can be logged by scripting environments and extended to construct provenance information automatically for data products created by scripts. Example provenance systems in script-based data processing environments include the virtual data grid system Chimera [20] and Earth System Science Server [21].

3) Services: Service-oriented architecture (SOA) allows distributed resources and applications to work together for data processing and scientific discoveries. Individual services can be chained together, for example, by using an industry-wide service orchestration standard, the Web Services Business Process Execution Language (BPEL for short). The BPEL scripts can be executed by BPEL workflow engines. Provenance information can be acquired by generating provenance through workflow engines, aggregating provenance information generated by each service, or a combination of the previous two methods. A provenance-aware SOA has been proposed to provide a framework for provenance recording, storing, and querying [19].

4) Semantic Web: The emergence of Semantic Web technologies, including Resource Description Framework (RDF), Web Ontology Language, and SPARQL Protocol and RDF Query Language (SPARQL) and RDF Query Language,

provides a way to connect data for more effective discovery and integration and thus shows considerable promise for new approaches to data provenance [22]. For example, Chebotko *et al.* propose a design of a relational RDF store for provenance management [23]. Zhao *et al.* discuss how to answer provenance-related questions when combining workflow provenance, domain-specific annotations, and the Web of Data [24].

IV. Provenance-Aware Geoscience Applications

Some studies have been conducted for provenance-aware applications in the geospatial domain. Each of them has its own constraints and application context. We summarize them in the following categories (Fig. 1).

A. GIS

Provenance applications in geoscience can be traced back to some work on GIS in the early 1990s [11]. Lineage information was recorded when performing spatial analyses on vector data using commands in GIS software and can be used to support analysis on error propagation [25]. Another example is GeoOpera, a geospatial extension to the Open Process Engine for Reliable Activities, which provides lineage support to geospatial workflows [26]. Wang *et al.* propose a provenance-aware GIS architecture to record the spatial data provenance [27]. Yue *et al.* propose metadata tracking in geoprocessing workflows. Detailed metadata for geospatial data are generated and propagated through a workflow [28]. These metadata provide a context for the evaluation of the quality and reliability of the geoprocessing data product, thus contributing to the data provenance [29].

B. Earth Science Data System

The long-term preservation and curation of Earth science data have been an important goal of various Earth science data systems [30]–[32]. Provenance tracking is a key research issue in achieving this goal [33]. It also improves the creditability of data sets and ensures scientific reproducibility [34]. Frew and Bose added lineage-tracking support for remote sensing data processing in a script-based environment [35]. Tilmes and Fleig discussed some general concerns of provenance tracking for Earth science data processing systems [36]. Plale *et al.* described architectural considerations to support provenance collection and management in geosciences [37]. To define specific types of information that should be preserved along with Earth science observational data, the Data Stewardship Committee of ESIP is working toward proposing a PCCS, which includes a list of items to capture the provenance of products resulting from Earth science missions. They are grouped into eight categories: Preflight/Preoperation Calibration, Science Data Products, Science Data Product Documentation, Mission Data Calibration, Science Data Product Software, Science Data Product Algorithm Input, Science Data Product Validation, and Science Data Software Tools [38].

C. Geospatial Web Service

Conventional provenance applications in geoscience focus on provenance capture, representation, and usage in a standalone environment. They cannot support wide sharing and open access of provenance information in a distributed environment. In a service-oriented distributed environment, the data and processing utilities are becoming available as services, and Web service technologies can significantly reduce data and computing resources needed for the end-user to conduct Earth science research [39]. Managing and serving provenance information using the same service-oriented paradigm now shows great promise and consistency with the existing SOA. Di suggested the combination of ISO 19115 and ISO 19115-2 lineage information model for use in the Web service workflow environment [40]. Yue *et al.* proposed an approach to share geospatial provenance information using the Open Geospatial Consortium (OGC) Catalogue Services for the Web (CSW) standard [41]. These approaches fit well with the current service stack of the geoinformatic domain and facilitate the management of geospatial data provenance in an open and distributed service environment. The OGC Web Processing Service (WPS) Specification specifies the lineage element in the request message of the "Execute" operation. In the OGC Sensor Web Enablement standards, Sensor Model Language can provide an explicit description of the process by which an observation has been obtained (i.e., observation lineage). In order to provide a more comprehensive investigation, the OGC Web Services (OWS) Initiative—Phase 9 (OWS-9) included a task on the catalogue for provenance and provenance encodings in OGC services, such as Web Feature Service (WFS) and WPS [42].

D. Earth Observation Sensor Web

Data provenance is also a critical issue in a Sensor Web environment [43], [44]. Patni *et al.* presented a linked data approach to model and query provenance associated with the sensor data [45]. The sensor observations, encoded using the OGC Observations and Measurements specification, are converted into RDF and made available as the linked data. A sensor provenance ontology creates links between observed phenomena and the sensors involved and thus can answer provenance queries such as finding sensors that recorded observations. In sensor networks, Ledlie *et al.* used provenance to address the naming and indexing issues in sensor data storage [46]. Park and Heidemann proposed an approach to track sensornet data in sensornet republishing, a process of transforming sensor data [47].

V. Key Considerations for Geoscience Data Provenance and Possible Solutions

The related studies described earlier help identify the specific requirements of the geoscience domain that provenance-aware applications should satisfy. The first one is the geoscience-domain properties in data products. For example, modeling

the provenance for geospatial data products must take into consideration spatial and temporal properties and the relations between them. Before using a geospatial data set of interests in their applications, users may raise some questions, such as the following: "How was the data set generated? What are the source data and their spatial and temporal ranges? Is there an error in the source data?" Spatial and temporal elements in metadata standards such as ISO 19115 could provide a valuable reference for space-time conceptualization and recording in the provenance model.

Adopting a standard-based or community-consensus provenance information model is important in geoscience domain. In the Earth science domain, the ISO Technical Committee 211 has set metadata standards for geographic information, including ISO 19115:2003—Geographic information–Metadata and ISO 19115-2:2009–Geographic information–Metadata–Part 2: Extensions for imagery and gridded data. The data quality package of ISO 19115:2003 defines lineage information classes and subclasses. ISO 19115-2:2009 extends the ISO 19115:2003 lineage model and provides additional metadata classes needed for documenting provenance information in geoprocessing workflows. In addition, ISO 19130:2010—Imagery Sensor Models for Geopositioning and ISO 19130-2:2012–Imagery Sensor Models for Geopositioning–Part 2–Synthetic Aperture Radar (SAR), Interferometric Synthetic Aperture Radar (InSAR), Light Detection and Ranging (LiDAR), and Sound Navigation and Ranging (SONAR) define sensor metadata standards. The sensor observation lineage, such as the process by which an observation has been obtained, can be addressed by these sensor metadata standards. The combination of lineage models in these standards provides a comprehensive provenance information model needed for the geoscience domain. The Reference Model for an Open Archival Information System (OAIS) defines the types of information needed for a full understanding of digital data objects in general terms. The OAIS Reference Model has been followed by the U.S. National Oceanic and Atmospheric Administration (NOAA) in the development of the Comprehensive Large Array-Data Stewardship System [30]. However, it is still not enough to address the entire suite of information that must be preserved in order to ensure the long-term usability of Earth science data. The PCCS is now being addressed actively by the Data Stewardship Committee of the U.S. ESIP Federation [38], [48]. The European Space Agency's Long-Term Data Preservation Working Group has developed a document detailing content that should be preserved along with data and derived products resulting from Earth observations [49].

Provenance models in the general information domain could be extended in geoscience domain to provide an interoperable solution for the provenance representation of geoscience data products. The W3C Provenance Working Group is finalizing a generic provenance model that can accommodate different perspectives of provenance such as agent-centered, object-centered, and process-centered perspectives. The PROV-DM defines both core structures and extended structures [9]. The core structures record the essence of provenance that is commonly found by various domain-specific provenance descriptions, and the extended structures add more expressive capabilities to support advanced uses of provenance. An interoperable provenance representation for geoscience domain could be achieved by extending PROV. A community working group has been proposed in the 2012 NASA Earth Science Data Systems Working Group meeting to develop an Earth Science PROV Extension using extensibility points of PROV [50].

A key issue involved in the provenance representation is the data citation, which could include author(s), release date, title, version, archive (and/or distributor), locator/identifier, access data, and time [38]. Technologies for assigning persistent identifiers in data citation are available, such as Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), Extensible Resource Identifiers (XRIs), Handle System (Handles), Life Science Unique Identifiers (LSIDs), Object Identifiers(OIDs), Persistent Uniform Resource Locators (PURLs), Uniform Resource Identifiers/Names/Locators (URIs/URNs/URLs), and UUIDs (Universally Unique Identifiers). They can help resolve metadata, including provenance and context information, for a data product used in scientific research [51]. For example, the Distributed Active Archive Center (DAAC) at the Oak Ridge National Laboratory, one of 12 DAACs of NASA's EOSDIS, has been using DOIs to help users cite its data sets for a few years. The EOSDIS Project has recently undertaken the task of broadening the use of DOIs to most of the data sets held at the DAACs [38].

Provenance could be captured at different levels of granularity. Some may require the provenance at the level of a single data set (or called data granule). Others may need the provenance at the level of data set collections. The representation of provenance may also have a level of granularity. The common characteristics could be shared at a high level, while the specifics would be represented at a low level. In certain applications, provenance information may be required at the level of pixels in image products. For example, in vegetation time-series studies for the contiguous U.S., long-term monthly normalized difference vegetation index (NDVI) maximum value composite images can be generated from the following two sources: the NOAA's global Advanced Very High Resolution Radiometer (AVHRR) NDVI (1981–2006) from the Global Inventory Modeling and Mapping Studies data set at a 15-day time step with a spatial resolution of 8 km and the monthly global Moderate Resolution Imaging Spectroradiometer (MODIS) NDVI composites (2000–2012) from NASA at a spatial resolution of 0.05°. The AVHRR NDVI could be processed in a consistent and quantitatively comparable manner with the MODIS NDVI based on the overlapping period of available data (2000–2006) for long-term time-series studies. The final monthly composite NDVI data set generated for U.S. thus has its spatial and temporal provenance from both sources. People might even be interested in knowing the per-pixel provenance, i.e., where the maximum value is from.

In the feature (vector)-based GIS, provenance could exist at both the feature-type and feature-instance levels. In the OGC OWS-9 testbed, conflation service is investigated, which can combine geospatial features from different sources into an integrated result. For example, it could compare two features

from different data sources and update one feature by adopting a specific conflation action such as adding attributes or updating values from the other feature. Thus, at the feature-type level, provenance includes the feature types from two data sources, while at the feature-instance level, both source features and conflation actions can be tracked as the provenance. Both levels can be recorded with the ISO lineage model. In the ISO 19115:2003, lineage is described using the combination of sources and process steps. ISO 19115-2:2009 extends the lineage model in ISO 19115:2003 to add support on recording intermediate sources, processing chains, parameters, and algorithms. The ISO 19139:2007 defines the implementation XML schema, where the role value of scope in lineage encodings could be "data set," "featureType," or "feature," respectively. At the feature-type level, sources are encoded as citations to feature types, and process steps are citations to specific geoprocessing services. In the feature-instance level, sources are citations to specific WFS features, and process steps include specific conflation parameters applied.

Most of the existing legacy data systems do not support data provenance. Therefore, another consideration is how to support data provenance in the legacy systems in a systematic and scalable way, e.g., how to support provenance in the current NASA EOSDIS systems. Another example is the OGC services and standards. Is there any way to incorporate the provenance support in the OWS standards stack? A review and framework for provenance-aware SOA is available in [19], showing considerable promise for provenance-aware geospatial applications in the cyberinfrastructure. Taking OWS as an example, provenance could be supported and interchanged among various existing services in the following way. Provenance is captured using the geoprocessing services (WPS). It is stored in a data-providing service such as WFS. Data set/feature-type level provenance is registered in a metadata catalogue (CSW), and feature level provenance is managed by WFS. The provenance tracking clients interact with the catalogue service using the CSW interface. The catalogue service processes lineage requests at the data set/feature-type level through the lineage associations such as ancestor relations stored in the metadata database [41]. Once provenance tracking clients are interested in the feature-level provenance, it interacts with a WFS to locate the lineage element in a feature. Thus, the catalogue focuses on the provenance management at the data set/feature-type level, while WFS is used to manage the lineage information at the feature-instance level. The catalogue supports the lineage tracking to its ancestors, such as the parent of parent, while a WFS only supports the feature lineage tracking to its parent. This means that if a client wants to find the parent of parent of a feature instance, it will combine the distributed WFSs and CSW to accomplish this goal. The interchange of provenance information among these services and clients follows the lineage information model specified by ISO.

In a Sensor Web environment, data provenance is particularly difficult [43]. First, the volumes of sensor data could be very large, and it is sometimes not practical to store all of them. Some data are typically thrown away after initial fusion or aggregation. Second, the sensor data often go through complex steps of filtering, correction, aggregation, and division, making provenance tracking to the origins nearly impossible. When extended with the social dimension such as citizens as sensors [52], the provenance tracking and quality assurance are even more complicated. More investigations are needed in this area.

VI. Conclusion

Geoscience data provenance is an essential part of geoscience data management and processing. It has become a fundamental issue in establishing the trustworthy geoscience information infrastructure. This paper discusses the current state of the art in geoscience data provenance, including the concept of geoscience data provenance, application contexts of provenance, general considerations and technologies in developing provenance-aware applications, and methods and applications of geoscience data provenance. This paper suggests that, by adopting standard-based provenance information model, it is possible to achieve the interoperability among provenance for scientific products in geoscience disciplines. It also reveals that granularity of provenance must be considered for specific geoscience application contexts and in existing systems. The key considerations discussed in this paper offer a guideline and direction for the future study of this subject.

References

[1] A. J. Tatem, S. J. Goetz, and S. I. Hay, "Fifty years of Earth observation satellites," *Amer. Sci.*, vol. 96, no. 5, pp. 390–398, Sep./Oct. 2008.

[2] H. K. Ramapriyan, J. Behnke, E. Sofinowski, D. Lowe, and M. A. Esfandiari, "Evolution of the Earth Observing System (EOS) Data and Information System (EOSDIS)," in *Standard-Based Data and Information Systems for Earth Observation*. Berlin, Germany: Springer-Verlag, 2010, pp. 63–92.

[3] EarthCube, *EarthCube Capabilities*, Dec. 2011. [Online]. Available: http://earthcube.ning.com/

[4] G. T. Lakshmanan, F. Curbera, J. Freire, and A. Sheth, "Guest editors' introduction: Provenance in Web applications," *IEEE Internet Comput.*, vol. 15, no. 1, pp. 17–21, Jan./Feb. 2011.

[5] L. Moreau, "The foundations for provenance on the Web," *Found. Trends Web Sci.*, vol. 2, no. 2/3, pp. 99–241, Feb. 2010.

[6] P. Buneman, S. Khanna, and W. C. Tan, "Why and where: A characterization of data provenance," in *Proc. ICDT*, 2001, pp. 316–330.

[7] Y. Cui, J. Widom, and J. L. Wiener, "Tracing the lineage of view data in a warehousing environment," *ACM Trans. Database Syst.*, vol. 25, no. 2, pp. 179–227, Jun. 2000.

[8] A. Chebotko, Y. Simmhan, and P. Missier, "Guest editorial: Scientific workflows, provenance and their application," *Int. J. Comput. Appl.*, vol. 18, no. 3, pp. 130–132, Jan. 2011.

[9] L. Moreau and P. Missier, "PROV-DM: The PROV Data Model," PROV-DM, 2012. [Online]. Available: http://www.w3.org/TR/prov-dm/

[10] P. Groth and Y. Gil, "Using provenance in the Semantic Web," *J. Web Semantics*, vol. 9, no. 2, pp. 147–148, Jul. 2011.

[11] D. P. Lanter, "Design of a lineage-based meta-data base for GIS," *Cartogr. Geogr. Inf. Syst.*, vol. 18, no. 4, pp. 255–261, Oct. 1991.

[12] M. F. Goodchild, H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, M. Ehlers, J. van Genderen, D. Jackson, A. J. Lewis, M. Pesaresi, G. Remetey-Fülöpp, R. Simpson, A. Skidmore, C. Wang, and P. Woodgate, "Next-generation digital Earth," *Proc. Nat. Acad. Sci. USA*, Jun. 21, 2012, to be published.

[13] L. Di, K. Moe, and T. L. van Zyl, "Earth Observation Sensor Web: An overview," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 3, no. 4, pp. 415–417, Dec. 2010.

[14] "Task Force on Grand Challenges, NSF Advisory Committee for Cyberinfrastructure," Arlington, VA, USA, p. 116, 2011.

[15] W3C, *W3C Provenance Working Group*, 2011. [Online]. Available: www.w3.org/2011/prov/

[16] L. Moreau, B. Plale, S. Miles, C. Goble, P. Missier, R. Barga, Y. Simmhan, J. Frutelle, R. E. McGrath, J. Myers, P. Paulson, S. Bowers, B. Ludaescher, N. Kwasnikowska, J. Van den Bussche, T. Ellkvist, J. Freire, and P. Groth,

The Open Provenance Model (v1.01). Southampton, U.K.: University of Southampton, 2008.
[17] Y. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-Science," *SIGMOD Rec.*, vol. 34, no. 3, pp. 31–36, Sep. 2005.
[18] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: A survey," *ACM Comput. Surveys*, vol. 37, no. 1, pp. 1–28, Mar. 2005.
[19] S. Miles, P. Groth, and M. Branco, "The requirements of using provenance in e-Science experiments," *J. Grid Comput.*, vol. 5, no. 1, pp. 1–25, Mar. 2007.
[20] I. Foster, J. Vockler, M. Wilde, and Y. Zhao, "Chimera: A virtual data system for representing, querying, and automating data derivation," in *Proc. 14th Int. Conf. SSDBM*, Edinburgh, Scotland, 2002, pp. 37–46.
[21] J. Frew, D. Metzger, and P. Slaughter, "Automatic capture and reconstruction of computational provenance," *Concurr. Comput., Practice Exp.*, vol. 20, no. 5, pp. 485–496, Apr. 2007.
[22] S. S. Sahoo, A. Sheth, and C. Henson, "Semantic provenance for eScience: Managing the deluge of scientific data," *IEEE Internet Comput.*, vol. 12, no. 4, pp. 46–54, Jul./Aug. 2008.
[23] A. Chebotko, S. Lu, X. Fei, and F. Fotouhi, "RDFProv: A relational RDF store for querying and managing scientific workflow provenance," *Data Knowl. Eng.*, vol. 69, no. 8, pp. 836–865, Aug. 2010.
[24] J. Zhao, S. S. Sahoo, P. Missier, A. Sheth, and C. Goble, "Extending semantic provenance into the Web of data," *IEEE Internet Comput.*, vol. 15, no. 1, pp. 40–48, Jan./Feb. 2011.
[25] H. Veregin and D. P. Lanter, "Data quality enhancement techniques in layer-based geographic information systems," *Comput. Environ. Urban Syst.*, vol. 19, no. 1, pp. 23–36, Jan. 1995.
[26] G. Alonso and C. Hagen, "Geo-Opera: Workflow concepts for spatial processes," in *Proc. 5th Int. SSD*, Berlin, Germany, 1997, pp. 238–258.
[27] S. Wang, A. Padmanabhan, D. J. Myers, W. Tang, and Y. Liu, "Towards provenance-aware geographic information systems," in *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. ACM GIS*, 2008, pp. 70–73.
[28] P. Yue, J. Gong, L. Di, and L. He, "Automatic geospatial metadata generation for Earth science virtual data products," *GeoInformatica*, vol. 16, no. 1, pp. 1–29, Jan. 2012.
[29] P. Yue, J. Gong, and L. Di, "Augmenting geospatial data provenance through metadata tracking in geospatial service chaining," *Comput. Geosci.*, vol. 36, no. 3, pp. 270–281, Mar. 2010.
[30] R. Rank and K. R. McDonald, "A NOAA/NASA pilot project for the preservation of MODIS data from the Earth Observing System (EOS)," in *Proc. PV, Ensuring Long-term Preserv. Adding Value Sci. Tech. Data*, pp. 1–9.
[31] S. Albani and D. Giaretta, "Long term data and knowledge preservation to guarantee access and use of the Earth science archive," in *Proc. PV, Conf. Ensuring Long-Term Preserv. Adding Value Sci. Tech. Data, ESA CESA*, Madrid, Spain, 2009, pp. 1–7.
[32] S. P. Morris and J. Tuttle, "Curation and preservation of complex data: The North Carolina geospatial data archiving project," in *Proc. NDIIPP, Lib. Congr.*, Washington, DC, USA, 2008, pp. 1–12.
[33] R. Bose and F. Reitsma, "Advancing geospatial data curation," in *Proc. PV Conf. Ensuring Long-term Preserv. Adding Value Sci. Tech. Data*, Edinburgh, U.K., 2005, pp. 1–12.
[34] C. Tilmes, Y. Yesha, and M. Halem, "Tracking provenance of Earth science data," *Earth Sci. Inf.*, vol. 3, no. 1/2, pp. 59–65, Jun. 2010.
[35] J. Frew and R. Bose, "Earth system science workbench: A data management infrastructure for Earth science products," in *Proc. 13th Int. Conf. SSDBM*, Fairfax, VA, USA, 2001, pp. 180–189.
[36] C. Tilmes and J. A. Fleig, "Provenance tracking in an Earth science data processing system," in *Proc. 2nd IPAW LNCS*, 2008, vol. 5272, pp. 221–228.
[37] B. Plale, B. Cao, C. Herath, and Y. Sun, "Data provenance for preservation of digital geoscience data," in *Proc. GSA*, 2011, pp. 125–137.
[38] H. K. Ramapriyan, L. Di, and G. Rochon, "Technical committees corner: Data Archiving and Distribution (DAD) Technical Committee (TC)," *IEEE Geoscience and Remote Sensing Newsletter*, no. 163, pp. 26–29, Jun. 2012.
[39] L. Di and K. McDonald, "Next generation data and information systems for Earth sciences research," in *Proc. 1st Int. Symp. Digit. Earth*, Beijing, China, 1999, vol. 1, pp. 92–101.
[40] L. Di, "Use of ISO 19115 and ISO 19115-2 lineage models for geospatial Web service provenance," in *Proc. IEEE IGARSS*, Vancouver, Canada, 2011, pp. 1–4.
[41] P. Yue, Y. Wei, L. Di, L. He, J. Gong, and L. Zhang, "Sharing geospatial provenance in a service-oriented environment," *Comput., Environ. Urban Syst.*, vol. 35, no. 4, pp. 333–343, Jul. 2011.
[42] "OGC Web Services, Phase 9," Wayland, MA, USA, 2012.
[43] M. Balazinska, A. Deshpande, M. Franklin, P. B. Gibbons, J. Gray, S. Nath, M. Hansen, M. Liebhold, A. Szalay, and V. Tao, "Data management in the worldwide sensor Web," *IEEE Pervasive Comput.*, vol. 6, no. 2, pp. 30–40, Apr.–Jun. 2007.
[44] C. Keßler and K. Janowicz, "Linking sensor data-why, to what, and how?" in *Proc. 3rd Int. Workshop Semantic Sensor Netw., CEUR-WS*, 2010, pp. 1–15.
[45] H. Patni, S. S. Sahoo, C. Henson, and A. Sheth, "Provenance aware linked sensor data," in *Proc. 2nd Workshop Trust Privacy Social Semantic Web*, Heraklion, Greece, May 31, 2010, pp. 1–12.
[46] J. Ledlie, C. Ng, D. A. Holland, K.-K. Muniswamy-Reddy, U. Braun, and M. Seltzer, "Provenance-aware sensor data storage," in *Proc. pNetDB*, Apr. 2005, p. 1189.
[47] U. Park and J. Heidemann, "Provenance in sensornet republishing," in *Proc. 2nd Int. Provenance Annotation Workshop*, Salt Lake City, UT, USA, Jun. 2008, pp. 208–292.
[48] H. K. Ramapriyan, J. F. Moses, and R. Duerr, "Preservation of data for Earth system science—Toward a content standard," in *Proc. IEEE IGARSS*, Jul. 2012, pp. 5304–5307.
[49] "Long Term Data Preservation: Earth Observation Preserved Data Set Content (LTDP/PDSC)," Eur. Space Agency, Paris, France, LTDP-GSEG-EOPG-RD-11-0003, Jul. 25, 2012.
[50] "PROV-ES Earth Science Extension to W3C PROV," Washington,DC, USA, 2012.
[51] R. Duerr, R. Downs, C. Tilmes, B. Barkstrom, W. C. Lendhardt, J. Glassy, L. E. Bermudez, and P. Slaughter, "On the utility of identification schemes for digital Earth science data: An assessment and recommendations," *Earth Sci. Inf.*, vol. 4, no. 3, pp. 139–160, Sep. 2011.
[52] M. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, Aug. 2007.

Liping Di (M'01–SM'13) received the Ph.D. degree in remote sensing/geographic information system (geography) from the University of Nebraska, Lincoln, USA, in 1991.

He is with the George Mason University, Fairfax, VA, USA, where he is a Professor and the Founding Director of the Center for Spatial Information Science and Systems and a Professor of the Department of Geography and Geoinformation Science. He has engaged in geoinformatics and remote sensing research for more than 25 years and has published over 300 publications. He has served as the Principal Investigator (PI) for more than $34 million in research grants and as a Co-PI for more than $8 million in research grants/contracts awarded by U.S. federal agencies and international organizations. His current research activities are mainly in the following three areas: remote sensing standards, Web-based geospatial information and knowledge systems, and remote sensing applications.

Dr. Di has actively participated in the activities of a number of professional societies and international organizations, such as IEEE Geoscience and Remote Sensing Society (GRSS), International Society for Photogrammetry and Remote Sensing (ISPRS), Committee on Earth Observation Satellites (CEOS), International Organization for Standardization (ISO) Technical Committee (TC) 211, Open Geospatial Consortium, InterNational Committee for Information Technology Standards (INCITS), and Group on Earth Observations (GEO). He served as the Cochair of the Data Archiving and Distribution Technical Committee of IEEE GRSS from 2002 to 2005, where he became the Chair from 2005 to 2009. He currently chairs INCITS/L1, a U.S. national committee responsible for setting U.S. national standards on geographic information and representing the U.S. at ISO TC 211.

Peng Yue (M'06–SM'13) received the B.S. degree in geodesy and surveying engineering from Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 2000, the M.S. degree in geodesy and survey engineering from the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, in 2003, and the Ph.D. degree in geographic information system (GIS) from Wuhan University in 2007. His doctoral research was conducted at George Mason University, Fairfax, VA, USA, from 2004 to 2007.

He was a Research Associate Professor with George Mason University from 2011 to 2013. He is a Professor with LIESMARS, Wuhan University. His research interests include Web GIS, GIService, geospatial data management and interoperability, distributed geoprocessing, and geospatial Semantic Web. He has been involved in research projects funded by the National Aeronautics and Space Administration, National Geospatial-Intelligence Agency (NGA), Department of Energy (DOE), and Open Geospatial Consortium. He is also a Principal Investigator (PI)/Co-PI for more than RMB 2 million in national research grants from the National Science Foundation of China and the Ministry of Science and Technology of China. He is the author or coauthor of over 50 publications in scientific journals, conference proceedings, and books. His representative paper "Semantics-based automatic composition of geospatial Web services chains" is the top ten cited paper of all computers and geoscience papers published between 2007 and 2012.

Hampapuram K. Ramapriyan (M'03–SM'03) received the B.Sc. degree from the University of Mysore, Mysore, India, the B.E. and M.E. degrees in electrical engineering from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, USA.

He is the Assistant Project Manager of the Earth Science Data and Information System (ESDIS) Project at the Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, MD, USA. He has over 40 years of managerial and technical experience in science data system development, image processing, remote sensing, parallel processing, algorithm development, science data processing, archiving, and distribution. The ESDIS Project develops and operates one of the largest civilian science data systems in the world—the Earth Observing System Data and Information System (EOSDIS) in support of NASA's Science Mission Directorate. The ESDIS Project has been instrumental in establishing, as a part of EOSDIS, a set of Distributed Active Archive Centers (DAACs) around the U.S. that manages NASA's Earth science data and provides convenient access to trillions of bytes of data in various scientific disciplines such as land processes, oceanography, hydrology, atmospheric sciences, cryospheric studies, etc. The project has also developed systems that facilitate "one-stop shopping" access to international data centers. His responsibilities in the project have ranged from supervising a group of technical professionals in the design and implementation of EOSDIS and managing the early development and operation of the DAACs to providing a customer focus by interfacing with the scientific customer community to understand their requirements and assuring that the system development accommodates their requirements. He has been involved in the study of the evolution of EOSDIS for the future decade and the implementation of its initial steps. His most recent focus is on data preservation and stewardship. He has developed NASA's Earth Science Data Preservation Content Specification.

Dr. Ramapriyan is a senior member of the IEEE Geoscience and Remote Sensing Society (GRSS). He was the Vice-Chair of the Technical Committee of the GRSS on Data Archiving and Distribution between 2005 and 2009, where he is currently the Chair. He is an active member of Data Stewardship Committee within the U.S. Earth Science Information Partners' Federation.

Roger L. King (M'73–SM'95) received the B.S. degree from West Virginia University, Morgantown, USA, in 1973, the M.S. degree in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 1978, and the Ph.D.degree in engineering from the University of Wales, Cardiff, U.K., in 1988.

He began his career with Westinghouse Electric Corporation but soon moved to the U.S. Bureau of Mines Pittsburgh Mining and Safety Research Center. Upon receiving the Ph.D. degree in 1988, he accepted a position with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, USA, where he holds the position of Giles Distinguished Professor and serves as the Director of the Center for Advanced Vehicular Studies (CAVS), Bagley College of Engineering. He also holds the CAVS Endowed Chair in Engineering.

Dr. King has been the recipient of numerous awards for his research including the Department of Interior's Meritorious Service Medal. Over the last 30 years, he has served in a variety of leadership roles with the IEEE Industry Applications Society, Power and Energy Society, and Geosciences and Remote Sensing. He has served for four years as the Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Data Archiving and Distribution Technical Committee and served as a member of the IEEE GRSS Administrative Committee (AdCom). He also served as the Cotechnical Chair for IGARSS 2009 in Cape Town, South Africa. He is a member of the European Image Information Mining Coordination Group, Tau Beta Pi, Phi Kappa Phi, Sigma Xi, and Eta Kappa Nu.