Provenance in Earth Science Cyberinfrastructure

A White Paper for NSF EarthCube

Liping Di and Peng Yue
Center for Spatial Information Science and Systems (CSISS)
George Mason University
4400 University Drive, MS 6E1, Fairfax, VA 22030
Phone: 703-993-6114; Fax: 703-993-6127
Email: ldi@gmu.edu, http://csiss.gmu.edu

1. Introduction

Provenance, also called lineage, records the derivation history of a data product (Figure 1). The history could include the algorithms used, the process steps taken, the computing environment run, data sources input to the processes, the organization/person responsible for the product, etc. Provenance provides important information to data users for them to determine the usability and reliability of the product. In the science domain, the data provenance is especially 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 cyberinfrastructure-related issues, including transparency in data sharing and processing, proper credits to data and algorithm contributors, and reproducibility and trust-ability of scientific results.

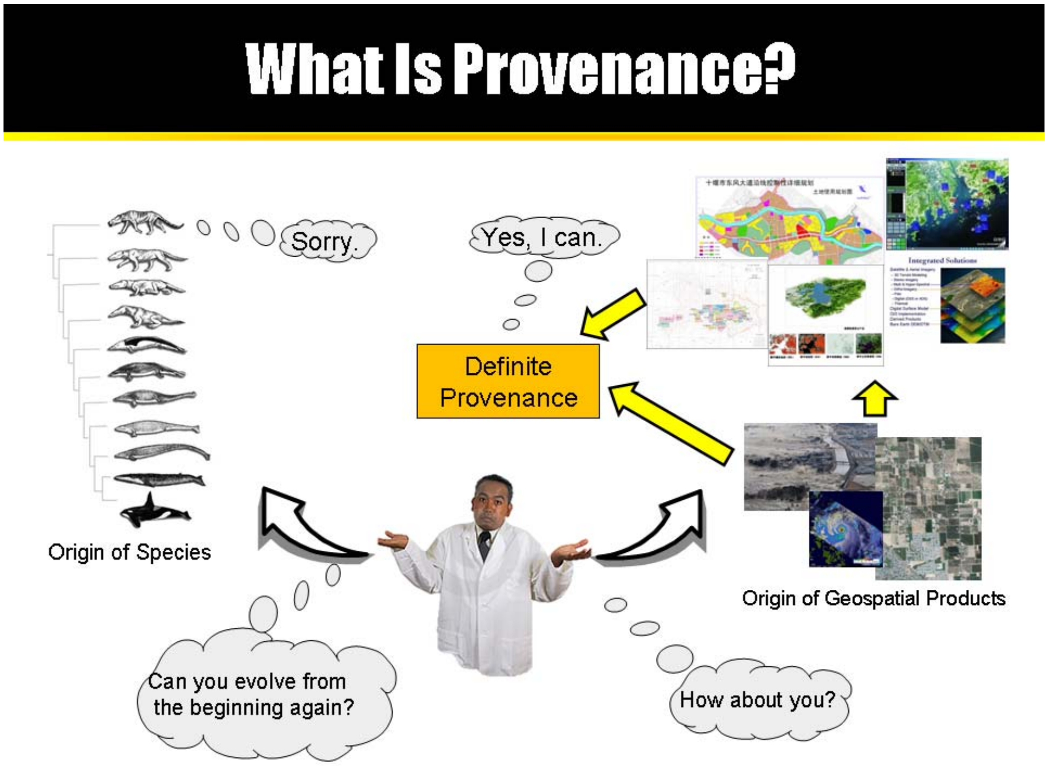

Figure 1. The concept of provenance

Traditionally, Earth science data products are produced in the science data centers with pre-determined processing procedures or workflows. In the distributed information cyberinfrastructure envisioned by EarthCube, sensor observations, data, and high-level information products are generated, transformed, published, and disseminated frequently over the Web. Foster (2005) used the term *Service-Oriented Science* to refer to the scientific research supported by distributed networks of interoperating services and data. In such a data-rich web-based production environment, provenance information is even more important since scientists rely on the information to understand and determine the reliability and usability of a scientific product generated from distributed services and inputs provided by diverse providers or sensors. The question is how to generate and manage provenance information in a distributed service environment of a CI.

This white paper describes the motivations for capturing provenance in EarthCube, and recommends technical solutions. It concludes that by developing a standard-based provenance information model and provenance services, we can provide a provenance-awareness information environment to EarthCube, in which a more informed understanding of resulted Earth science products can be achieved.

2. Description of the current problems

In the Earth science domain, a data product is normally generated from input sources such as satellite sensors through a series of processing steps. The series of interconnected processing steps forms a workflow. Traditionally, Earth science data products are produced in the scientific data centers with pre-determined workflows. The products as well as the algorithms and workflows used to generate them are strictly validated before the products are distributed to the scientific user community. The scientific community generally trusts the products from those sources. Provenance information for such products is normally carried as a part of the product metadata. However, such a model of data production has severe limitations, such as inflexibility, high costs, difficult to reuse, and lack of interoperability.

Recent development in service technology has moved the production of Earth science products gradually to the Web-based service environment. In the environment, a processing step becomes a standard compliant, chainable service component, and a workflow used to generate a product could be dynamically constructed on demand by users, and the services as well as the inputs to the production workflow are dynamically discovered from the distributed information cyberinfrastructure. This practice requires the generation of provenance in the loosely-coupled service environment. The lack of proper provenance information model for Earth science also prevents interdisciplinary geoscientific data provenance from being rapidly integrated and analyzed, and makes the long-term preservation and understanding of Earth science data problematic. If comprehensive provenance information model standards are available in Earth science domain and provenance for Earth science products are all described by following the standards, provenance information for the product will be easily interoperable and a general provenance-aware cyber service system can be developed to navigate among multidisciplinary products connected by the standard-based provenance. ISO 19115 (ISO, 2003) and ISO 19115-2 (ISO, 2009) are the example of such standards that can potentially be used for Earth science provenance.

The second problem is how to capture provenance information in the sensor Web environment where the Earth observation sensors are dynamically and lively connected to each other and to the Earth science models. Modern observational tools and advanced computational models will be connected through a solid cyber infrastructure to meet the challenges in geoscience, as stated in the NSF GEO Vision report (page 31) (NSF, 2009). The integrated sensor-simulation systems, and verification, validation, and reproducibility of simulations, are also research issues in advanced computational methods to enable the solutions to grand challenges in Cyberinfrastructure (page 34-35) (NSF, 2011). The simulation results can be enhanced by optimization of sensor networks, and inclusion of complementary sensors or observation correction/processing chains. For example, geopositioning of sensor observations is a key step for making the observation usable and for fusing multi-sensor data. In the dynamic sensor web environment, such infusion requires the automatic geopositioning, which can be accomplished if the physical sensor models (sensor and platform description with the associated physical and geometric information) are provided in standard way. Such information, if provided in provenance, can be used to analyze the errors in the final product caused by the geometric mis-registration and calculate the associated propagated errors. When rich streams of various sensors measurements are filtered, corrected, combined, and divided in the sensor Web environment, results reproductions and troubleshooting become a difficult job. The provenance information model and management, therefore, should have sufficient capabilities to support Sensor Web provenance.

3. Related work

Provenance, in general, has being addressed actively in the e-Science or Cyberinfrastructure in the past several years. The NSF task force report on grand challenges for Cyberinfrastructure suggests that a robust persistent data infrastructure should include the data provenance component (page 58) (NSF, 2011). The well known international forums on provenance include the Provenance Challenge workshop (Since 2006), the International Provenance and Annotation Workshop (IPAW) (Since 2002), and the International Workshop on the role of Semantic Web in Provenance Management (SWPM) (Since 2009). The W3C Provenance Working Group is working on defining a language for exchanging provenance information for Web resources (W3C, 2011). In the Earth science domain, there is also an increased interest in recent years on incorporating provenance support in geoscientific data systems, in particular the distributed data infrastructure, such as studies shown in the 2010 AGU Informatics Session on Encouraging and Enabling Transparency in Science Data, and the IEEE IGARSS 2011 special session on Provenance in Geoscience Data.

The general considerations in designing provenance-aware system include:

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 dependency relations to be tracked, and possibly derived, among data products and transformation processes.

2) Provenance capturing: The provenance information could be captured and recorded manually or automatically. It is important in the Cyberinfrastructure to automate the provenance capturing rather than rely on manual work, simply because of large volumes of data and frequent processing in the Cyberinfrastructure. Provenance information can be captured by tracing the execution of the workflow engine or aggregating provenance information generated by distributed services in a workflow when the workflow is executed.

3) Provenance storage: Provenance information can be tightly coupled with metadata, using existing metadata catalogue for storage and management and providing that to users with the data products. It could also be managed using a separated storage system, or called the provenance store. Both of them should support the distributed storage of provenance information

4) Provenance query: The design of the provenance query should take into consideration of the query interface, language, and queryable and returnable provenance content. The query interface specifies protocols and operations for provenance queries. The query language provides a set of predicate and data types such as SQL. Queryable and returnable provenance content depends on the model and representation of provenance information. 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 navigation between provenance and data products with a more comfortable user experiences. Visualization of scientific results can be combined with relevant provenance information to help scientific users discover anomalies and evaluate results.

6) Application 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*. Data quality issue is the primary application of provenance in scientific domains. In geospatial domain, ISO 19115 defines lineage as a part of data quality information. 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. Storing entire workflow with detailed description as intermediate transformation steps 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. Interleaving provenance information and products together, discovery and interpretation of Earth science products can be more informational.

A number of existing works have contributed to methods on provenance-aware applications (Simmhan et al., 2005; Miles et al., 2007; Moreau, 2010). These methods can be classified into four 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.

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.

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 and executed using 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.

4) Semantic Web: The emergence of Semantic Web technologies, including Resource Description Framework (RDF), Web Ontology Language (OWL), and SPARQL Protocol And RDF Query Language (SPARQL), provides a way to connect data for more effective discovery and integration, and thus shows considerable promise for new approaches to data provenance.

Provenance investigation in geoscience can be traced back to Lanter's work in the early 1990's. Lineage information was recorded when performing spatial analyses on vector data using commands in Geographic Information Systems (GIS) software (Lanter, 1991), and can be used to support analysis on error propagation (Veregin and Lanter, 1995). Geo-Opera, a geospatial extension to the Open Process Engine for Reliable Activities (OPERA), provided lineage support for geospatial workflows (Alonso and Hagen, 1997). Frew and Bose (2001) added lineage-tracking support for remote sensing data processing in a script-based environment. Wang et al. (2008) proposed a provenance-aware architecture to record the lineage of spatial data. Tilmes and Fleig (2008) discussed some general concerns of provenance tracking for Earth science data processing systems. Plale et al. (2010) described architectural considerations to support provenance collection and management in geosciences.

Conventional provenance applications in geoscience focus on provenance capture, representation, and usage in a stand-alone 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 (Di and McDonald, 1999). Managing and serving provenance information using the same service-oriented paradigm now shows great promise and consistency with the existing service-oriented architecture. Di (2011) suggests the combination of ISO 19115 and ISO 19115-2 lineage information model for use in the Web service workflow environment. Yue et al. (2011) proposes an approach to share geospatial provenance information using the Open Geospatial Consortium (OGC) Catalogue Services for the Web (CSW) standard. These approaches fit well the current service stack of the geoinformatic domain and facilitate the management of geospatial data provenance in an open and distributed service environment.

4. Recommendations

This white paper proposes to add provenance support in EarthCube by

1) adopting standard-based provenance information model for EarthCube cyberinfrastructure.

Provenance is a kind of metadata. In the Earth science domain, the International Organization for Standardization (ISO) Technical Committee 211 (ISO TC 211) have set metadata standards for geographic information, including ISO 19115:2003-Geographic information-Metadata (ISO, 2003), and ISO 19115-2:2009-Geographic information-Metadata-Part 2: Extensions for imagery and gridded data (ISO, 2009). ISO 19115 defines lineage information classes and subclasses. ISO 19115-2 extends the lineage model in ISO 19115 and provides additional metadata classes needed for documenting provenance information in geoprocessing workflows. In addition, ISO 19130 - Imagery Sensor Models for Geopositioning, and ISO 19130-2, Imagery Sensor Models for Geopositioning—Part 2—SAR, InSAR, LIDAR, and 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 EarthCube.

The work needed to be performed will be a) develop a standard provenance information model based on the ISO standards to support provenance needs in the EarthCube, b) create XML schema for implementing the model in XML in a standard and consistent way, and c) develop a metadata editing tool for users to easily create a provenance metadata instance based on the schema. In addition, in order to achieve the maximum worldwide interoperability, we all should make the provenance information model and the XML schema ISO standards.

2) Implementing a provenance generation and service system in the EarthCube by using the standard provenance information model.

The system can be built by reusing provenance components and services in existing geospatial Web service systems, such as GeoBrain, a geospatial web service system developed by the Center for Spatial Information Science and Systems (CSISS), George Mason University (Deng and Di, 2010; Di, 2004). GeoBrain has been operational since 2005. In GeoBrain, more than 200 image processing functions have been implemented as web services (Li et al., 2010). The system has been linked to the on-line data sources from NASA, NOAA, and USGS. A prototype provenance generation and service system has been implemented in GeoBrain (Di, 2011).

The implementation work will include the following steps: a) create the static provenance metadata of the algorithm and running environment for each deployed web services, b) create the template for the dynamic provenance metadata of a deployed web service, c) for each type of source data in EarthCube, templates will be created for data-source provenance metadata, d) implement a provenance capture module that can provide a preview of provenance information by analyzing the workflow before the execution and can works with a geoscience workflow engine, such as BPELPower (Di et al., 2009), to create a complete provenance record during the workflow execution for the end product which the workflow generates. The provenance information is encoded in XML based on the schema defined in ISO 19139 (ISO, 2007). It can be either attached to the end-product file as a part of the metadata or stored in a separate metadata file for user to download or query.

5. Conclusion

“Data should be documented adequately enough to find it, interpret it, and understand its provenance” (page 58) (NSF, 2009). By adopting standard-based provenance information model in EarthCube, it is possible to achieve the interoperability among provenance for scientific products in all geoscience disciplines. We believe that Web service technologies will be widely used in the EarthCube. The capturing and sharing geospatial provenance in the Web service environment will also ensures the interworkability of the approach.

6. References

Alonso, G., & Hagen, C. (1997). Geo-Opera:workflow concepts for spatial processes. In Proceedings of the Fifth International Symposium on Spatial Databases (SSD’ 97), Berlin, Germany (pp. 238-258).

Di, L., 2011. Use of ISO 19115 and ISO 19115-2 lineage models for geospatial web service provenance. In: Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS11), 2011, Vancouver, Canada. 4pp.

Di L., Peisheng Zhao, Weiguo Han, Xiaoyan Li, Meixia Deng, 2009. “The Implementation of Geospatial Web Services and Workflows in GeoBrain”. In Proceedings of 2009 IEEE International Geoscience and Remote Sensing Symposium, July 12-17, 2009, Cape Town, South Africa.(Extended abstract).

Di, L., 2004. GeoBrain-A Web Services based Geospatial Knowledge Building System. Proceedings of NASA Earth Science Technology Conference 2004. June 22-24, 2004. Palo Alto, CA, USA. (8 pages. CD-ROM).

Di, L., McDonald, K., 1999. Next generation data and information systems for earth sciences research, in: Proceedings of the First International Symposium on Digital Earth, vol. I., Science Press, Beijing, China, pp. 92–101.

Deng, M., and L. Di, 2010. Facilitating Data-intensive Research and Education in Earth Science - A Geospatial Web Service Approach. LAP LAMBERT Academic Publishing GmbH & Co. KG, Saarbrücken, Germany. ISBN: 978-3-8383-9714-6.

Frew, J., & Bose, R. (2001). 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 (SSDBM’01), Fairfax, Virginia, USA, IEEE Computer Society (pp. 180-189).

Foster, I., 2005. Service-oriented science, Science 308(5723): 814-817.

ISO, 2009. ISO 19115-2-Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data. http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229

ISO, 2008. ISO/DTS 19130-Geographic Information - Imagery Sensor Models for Geopositioning. International Organization for Standardization (ISO), Technical Committee 211. ISO/TC 211 Doc. N 2509, 164p. (Di, L., project chair and editing committee editor).

ISO, 2007. ISO/TS 19139:2007 Geographic information -- Metadata -- XML schema implementation. http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557

ISO, 2003. ISO 19115-Geographic Information-Metadata. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020.

Lanter, D. P. (1991). Design of a lineage-based meta-data base for GIS. Cartography and Geographic Information Systems, 18(4), 255-261.

Li, X., L. Di, W. Han, P. Zhao, and U. Dadi, 2010. Sharing geoscience algorithms in a Web service-oriented environment (GRASS GIS example). Computers & Geosciences. Volume 36, Issue 8, August 2010, Pages 1060-1068. doi:10.1016/j.cageo.2010.03.004

Miles, S., Groth, P., Branco, M., & Moreau, L. (2007). The requirements of using provenance in e-Science experiments. Journal of Grid Computing, 5(1), 1-25.

Moreau, L. (2010). The foundations for provenance on the web. Foundations and Trends® in Web Science, 2(2-3), 99-241.

NSF, 2009. GEO Vision report, NSF Advisory Committee for Geosciences. 44pp.

NSF, 2011. Task Force on Grand Challenges, NSF Advisory Committee for Cyberinfrastructure. 116pp.

Plale, B., Cao, B., Herath, C., & Sun, Y. (2010). Data provenance for preservation of digital geoscience data. Geological Society of America (GSA), Memoir Volume, 12/2010, 14pp. <http://www.cs.indiana.edu/~plale/papers/PlaleDataProvenancePreservationPreprint.pdf> (last date accessed: 13 July 2010).

Simmhan, Y.L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. SIGMOD Record, 34(3), 31-36.

Tilmes, C., Fleig, J. A. (2008). Provenance tracking in an earth science data processing system. In Proceedings of the Second International Provenance and Annotation Workshop (IPAW 2008), Salt Lake City, UT, USA, Lecture Notes in Computer Science (LNCS) 5272, Springer, Berlin, Germany (pp. 221-228).

Veregin, H., & Lanter, D. P. (1995). Data quality enhancement techniques in layer-based geographic information systems. Computers, Environment and Urban Systems, 19(1), 23-36.

Wang, S., Padmanabhan, A., Myers, D. J., Tang, W., & Liu, Y. (2008). Towards provenance-aware geographic information systems. In Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008), Irvine, California, USA, 4pp.

W3C, 2011. W3C Provenance Working Group , www.w3.org/2011/prov/

Yue, P., Wei, Y., Di, L., He, L., Gong, J., and Zhang, L., 2011. Sharing geospatial provenance in a service-oriented environment. Computers, Environment and Urban Systems, 35(4): 333-343.