Semantic Provenance Registration and Discovery using Geospatial Catalogue Service

Peng Yue[1], Jianya Gong[1], Liping Di[2], Lianlian He[3], Yaxing Wei[4]
[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079
[2] Center for Spatial Information Science and Systems (CSISS), George Mason University, 4400 University Drive, MS 6E1, Fairfax, VA 22030, USA
[3] Department of Mathematics, Hubei University of Education, Nanhuan Road 1, Wuhan, Hubei, China, 430205
[4] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6407, USA

***Abstract* – A geospatial catalogue service allows geospatial users to discover appropriate geospatial data and services in a Web-based distributed environment. Metadata for geospatial data and services is organized structurally in catalogue services. Provenance for geospatial data products, as a kind of metadata describing the derivation history of data products, can be managed in a same way as other kinds of metadata using metadata catalogue services, thus keeping consistency and interoperability with existing metadata catalogue services. Meanwhile, Semantic Web technologies have shown considerable promises for more effective connection, discovery, and integration of provenance information. This paper addresses how geospatial catalogue services can be enriched with semantic provenance. Semantic relationships defined in provenance ontologies are registered in an OGC standard-compliant CSW service by extending ebRIM elements. The work illustrates that such a semantically-enriched CSW can assist in the discovery of data, service, and knowledge level of geospatial provenance.**

Keywords: Data Provenance, Lineage, GIS, CSW, ebRIM, Geospatial Web Service

I. INTRODUCTION

The advancement of Earth observing technologies has significantly increased the capability for collecting geospatial data. The National Aeronautics and Space Administration (NASA)'s Earth Observing System (EOS) alone is generating 1000 terabytes annually [1]. Significant efforts have been devoted to make full use of the data and derive useful information from the raw data. The Open Geospatial Consortium (OGC)'s Web Service technologies such as the Web Feature Service (WFS), Web Map Service (WMS), and Web Processing Service (WPS) [2] have been widely used in geospatial domain to facilitate the open discovery of, access to, and processing of distributed geospatial data. A geospatial catalogue service allows geospatial users to discover appropriate geospatial data and services in a Web-based distributed environment. Metadata for geospatial data and services is organized structurally in catalogue services. The OGC's Catalogue Services for the Web (CSW) is a domain consensus regarding an open, standard interface for geospatial catalogue service [3].

Provenance for geospatial data products records the derivation history of the data products. In a service-oriented information infrastructure, geoprocessing steps in deriving a data product are usually implemented by chaining multiple geoprocessing services together. To derive useful data products from large volumes of raw data, the integration of geoprocessing services become more and more frequent. Provenance provides important context information to help end users make decisions about the quality of the derived data products. Semantic Web technologies provide ways to connect Web resources together and allow semantics of Web resources to be machine-understandable, thus enabling more effective discovery, automation, integration, and reuse of resources. Semantic provenance, provenance information represented using Semantic Web technologies, therefore, can provide more informed understanding and effective usage of provenance information.

In the geospatial domain, provenance information has been regarded as part of metadata describing data quality information in the International Organization for Standardization (ISO) 19115 geospatial information—metadata standard. Similar to other kinds of geospatial metadata managed using metadata catalogue services, provenance information can be registered and discovered in the metadata catalogue services to keep consistency and interoperability with legacy geographic information system (GIS) applications. The registration of provenance information in the catalogue services requires the specification of the registration information model. OGC has recommended the ebXML Registry Information Model (ebRIM) for registration of geospatial information, the so-called ebRIM profile of CSW [4]. However, the existing standard does not address the registration of provenance information.

This paper explores the use of OGC CSW for registration and query of semantic provenance. To make use of semantics for provenance discovery in CSW, semantic relationships defined in provenance ontologies are registered in an OGC standard-compliant CSW service by extending ebRIM elements. The work illustrates that such semantically-enriched CSW can assist in the discovery of data, service, and knowledge level of geospatial provenance. The rest of the

paper is organized as follows. Section 2 introduces the semantic representation of provenance for geospatial data products. Section 3 describes the ebRIM-based information model in CSW, and Section 4 presents the registration of semantic provenance. Section 5 describes the provenance discovery using semantically-enriched CSW. The work is compared with related work in Section 6, and conclusions and pointers to future work are given in Section 7.

II. SEMANTIC PROVENANCE FOR GEOSPATIAL DATA PRODUCTS

In the context of this paper, we focus on the provenance in a service-oriented environment in which geospatial data products are generated by executing geoprocessing service chains. In the general information domain, service chaining is a hot research topic in the Web Service area and can be called service composition. Approaches for service composition generally follow a three-phase procedure [5-7]: (1) process modeling, which generates an abstract process model consisting of the control flow and data flow among process nodes; (2) process model instantiation, where the abstract process model is instantiated into an executable service chain; and (3) workflow execution, where the chaining result is executed in the workflow engine to generate the required data product. The information involved in the three phases, therefore, can contribute to the provenance of the data products.

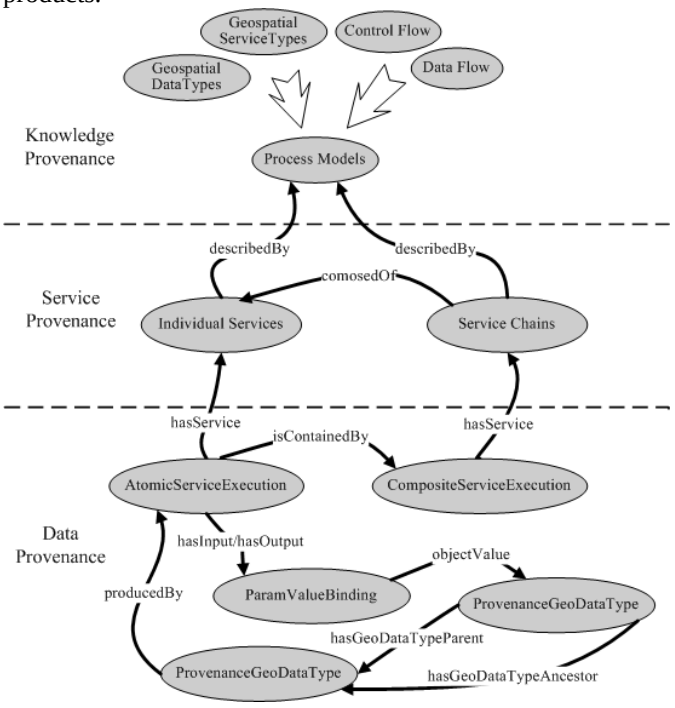

Figure 1. Semantic provenance for geospatial data products.

A three-level view of semantic provenance is adopted for the geospatial data products generated based on the three-phase procedure of service composition (Fig. 1). The first level is the knowledge level provenance, which contains process model ontologies as a knowledge base to support generation of complex process models. The process model ontologies are formulated by linking geospatial domain DataType, ServiceType, and workflow ontologies together. Examples of process model ontologies are atomic and composite process models for geoprocessing services described using the process model ontologies in the Web Ontology Language (OWL) Service Ontology (OWL-S). The second level is the service level provenance, which includes the individual services and service chains. Both can be represented using the service ontologies in OWL-S. And the final level is the data level provenance, which contains the provenance information generated during the execution. Examples of provenance in this level include source, intermediate, and final data products, atomic service executions, and service chain executions.

The ontologies for the knowledge level provenance and service level provenance use the geospatial domain ontologies and OWL-S ontologies. The data level provenance includes classes and relationships for data products required or generated by execution (ProvenanceGeoDataType class), value bindings between parameters and their values (ParamValueBinding class), specific executions of services (AtomicServiceExecution class) and service chains (CompositeServiceExecution class). Example ontologies in OWL can be viewed online at http://www.laits.gmu.edu/geo/nga/landslideprovenance.html .

The three-level view of geospatial provenance corresponds to the three phases of automatic service composition. The knowledge level provenance records the process model knowledge used to derive geospatial data products in the process modeling phase. Using provenance at this level, users can check the correctness of the process model and try a different model when necessary. The service level provenance describes concrete service chains that can be executed to generate the geospatial data products. Using this information, it is possible for users to re-select services based on the performance of services. The data level provenance helps users to find dependencies among physically-existed data products and supports analysis applications such as error source identification and propagation.

III. CSW-EBRIM PROFILE

CSW specification provides a framework for the implementation of application profiles. The core elements in an OGC catalogue service are the information model, the query language, and the interface [3]. The information model describes information structures and semantics of information resources. Therefore, the information model of catalogue services should address the content, syntax, and semantics of geospatial resources. The ebRIM standard has been defined by the Organization for the Advancement of Structured Information Standards (OASIS) and selected by OGC as the information model for specifying how catalogue content is structured and interrelated.

Fig. 2 shows the ebRIM-based catalogue information model. The core metadata class is the RegistryObject. Most other metadata classes in the information model are derived from this class. An instance of RegistryObject may have a set of zero or more Slot instances that serve as extensible attributes for this RegistryObject instance. An Association instance represents an association between a source RegistryObject and a target RegistryObject. Each association has an associationType attribute that identifies the type of that association. A Classification instance classifies a RegistryObject instance by referring to a node defined within a ClassificationScheme instance. A ClassificationScheme instance in the ebRIM model defines a tree structure made up of nodes that can be used to describe a taxonomy.

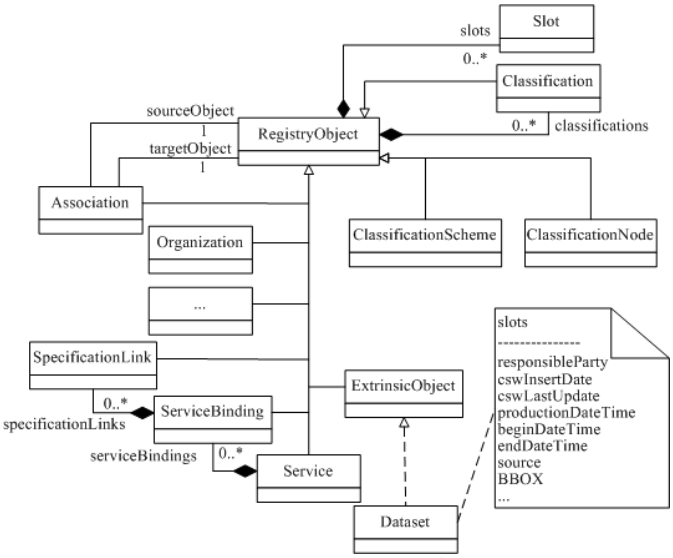

Figure 2. The ebRIM-based catalogue information model.

The ebRIM provides a general and standard metadata registration information model. However, it needs to be extended with some extension elements to meet common requirements in the geospatial domain. Under the guidelines of the ebRIM profile for CSW, the CSW implementation[1], developed and maintained by Laboratory for Advanced Information Technology and Standards (LAITS) from George Mason University [8], has extended ebRIM using international geographic standards: ISO 19115 Geographic Information — Metadata (including part 2: Extensions for imagery and gridded data) and ISO 19119 Geographic Information — Services.

The ebRIM is extended with ISO 19115 and ISO 19119 in two ways. The first is by importing new classes into the ebRIM class tree, deriving new metadata classes from existing ebRIM classes. The new Dataset class is used to describe geographic datasets. Many new attributes are added to the Dataset class based on ISO 19115 and its part 2. The second way to extend ebRIM is to use Slots to extend an existing class. The Service class included in ebRIM can be used to describe geographic services, but the available attributes in the class Service are not sufficient to describe geospatial Web services. New attributes derived from ISO 19119 are added to the Service class through Slots.

[1] Online services are available at http://geobrain.laits.gmu.edu/ .

IV. SEMANTIC PROVENANCE REGISTRATION

The registration of semantic provenance in the CSW takes advantages of extensibility points in ebRIM. Such extensibility points include new kinds of classes, associations, classifications, and additional slots to record OWL classes, properties and related axioms. Some efforts have already addressed the registration of OWL-based ontologies in ebRIM [9-12]. In this study, we focus on the application and extension of ebRIM in the provenance registration. In particular, the paper explores how to register the OWL-based semantic provenance in the ebRIM-based catalogue information model to support the provenance discovery.

For the knowledge level and service level provenance, we adopt the previous approach on registration of OWL/OWL-S [13]. A new type of ExtrinsicObject, named ProcessModel, is created in the ebRIM model to describe process models. Geospatial DataType and ServiceType ontologies are recorded using two new ClassificationScheme instances, which can be used to classify the ProcessModel and Dataset instances. The Service class in the ebRIM model can be used to describe both services and service chains, since a service chain as a whole can act as a service. The semantics for inputs, outputs, preconditions and effects (i.e. IOPE semantics) are recorded by using slots.

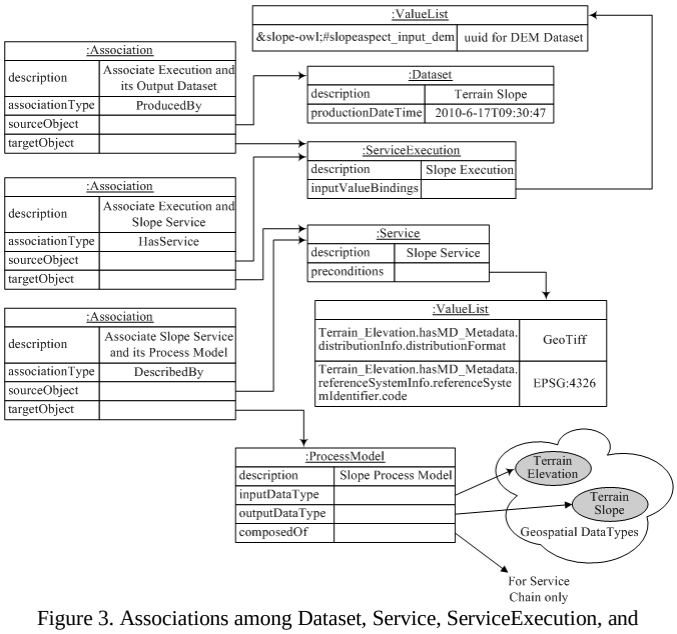

Figure 3. Associations among Dataset, Service, ServiceExecution, and Process Model.

For the data level provenance, a new type of ExtrinsicObject, ServiceExecution, which can support the registration of both atomic and service chain execution, is created. ProvenanceGeoDataType in OWL is mapped to the existing class Dataset. Individuals of ParamValueBinding in

provenance ontologies are recorded using the slots of the ServiceExecution. The relationships among AtomicServiceExecution, CompositeServiceExecution, and ProvenanceGeoDataType in provenance ontologies are registered using associations in the ebRIM.

Fig. 3 shows an execution of slope computation service, which generates terrain slope data from the digital elevation model (DEM) data. The knowledge level provenance is recorded by using instances of ProcessModel whose slots specifies the input Geospatial DataType (Terrain Elevation) and output Geospatial DataType (Terrain Slope). The service level provenance is recorded using instances of Service. DescribedBy association connects a service with its process model. Some individual geospatial services have their own metadata constraints on the input data and this can be recorded using slots. For example, the slope computation service in Fig. 3 specifies that the input terrain elevation data should be in the GeoTIFF data format with the EPSG:4326 geographic coordinate reference system. Data level provenance includes the registration of ServiceExecution and Dataset. A ServiceExecution is linked to the service executed using the HasService association. The Terrain slope dataset generated by the specific ServiceExecution is described using the ProducedBy association. More kinds of associations can be registered such as the HasGeoDataTypeAncestor relationship between datasets.

V. PROVENANCE DISCOVERY

Based on the semantic content registered in the CSW, three types of provenance discoveries are achieved using CSW queries:

A. Discovery for data level provenance

The discovery is based on provenance associations at the data level. Examples of CSW queries includes: collecting descendant or ancestor datasets to a specific dataset; finding service executions to generate a specific dataset; retrieving parameters and values involved when conducing a specific service execution.

B. Discovery for service level provenance

One discovery is to locate services or service chains used to generate a specific geospatial data product. The query is based on the HasService association between service executions and services. Additional discovery includes query on the preconditions of a specific service. The results from this query can help check preconditions of the service to find whether input data is semantically valid. For example, does the input DEM data have a valid spatial projection?

C. Discovery for knowledge level provenance

This is to discover process model knowledge used to derive geospatial data products. The CSW query uses DescribedBy association as a search condition. The process model, when obtained, can be rechecked and compared with alternative process models. Another query strategy is to add semantically-matched ServiceTypes in the search condition to find alternate process models for decision support. The semantic match is performed based on the subsumption reasoning in description logic.

```
<?xml version="1.0" encoding="UTF-8"?>
<csw:GetRecords …>
  <csw:Query typeNames="ServiceExecution Association
Dataset ClassificationNode">
    <csw:ElementSetName>full</csw:ElementSetName>
    <csw:ElementName>/ServiceExecution/</csw:ElementName
>
    <csw:Constraint version="1.0.0"><ogc:Filter><ogc:And>
     <!--temporal condition-->…
     <!--spatial condition-->…
     <!—ontological concept-->
      <ogc:PropertyIsEqualTo><ogc:PropertyName>/Dataset/@i
d</ogc:PropertyName>
        <ogc:PropertyName>/Classification/@classifiedObject</
ogc:PropertyName></ogc:PropertyIsEqualTo>
      <ogc:PropertyIsEqualTo><ogc:PropertyName>/Classificati
on/@classificationScheme</ogc:PropertyName>
        <ogc:PropertyName>/ClassificationScheme/@id</ogc:Pr
opertyName></ogc:PropertyIsEqualTo>
      <ogc:PropertyIsEqualTo>
        <ogc:PropertyName>/ClassificationScheme/Description/
LocalizedString/@value</ogc:PropertyName>
        <ogc:Literal>geospatial data type ontology</ogc:Literal>
      </ogc:PropertyIsEqualTo>
      <ogc:PropertyIsEqualTo><ogc:PropertyName>/Classificati
on/@classificationNode</ogc:PropertyName>
        <ogc:PropertyName>/ClassificationNode/@id</ogc:Prop
ertyName></ogc:PropertyIsEqualTo>
      <ogc:PropertyIsEqualTo><ogc:PropertyName>/Classificati
onNode/@code</ogc:PropertyName>
    <ogc:Literal>ETM_NDVI</ogc:Literal>
    </ogc:PropertyIsEqualTo>
    <!--provenance association-->
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>/Dataset/@id</ogc:PropertyName>
      <ogc:PropertyName>/Association/@sourceObject</ogc:Pr
      opertyName></ogc:PropertyIsEqualTo>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>/ServiceExecution/@id</ogc:Propert
yName>
      <ogc:PropertyName>/Association/@targetObject</ogc:Pro
pertyName></ogc:PropertyIsEqualTo>...
  </ogc:And></ogc:Filter></csw:Constraint></csw:Query>
  </csw:GetRecords>
```

Figure 4. Provenance query using CSW operation.

All queries are realized through CSW standard query operations. The query language is implemented using the OGC Filter specification. It supports comparison operators and spatial operators. An example provenance query is shown in Fig. 4. A Web client, e.g. HTML form, can submit queries using the GetRecords operation based on the request-response model of the HTTP protocol.

VI. RELATED WORK

A substantial research on provenance issue has been

conducted in the general information domain. Traditional data provenance issue focuses on the database systems [14-16]. With the advancement of service-oriented infrastructure in recent years, provenance for scientific workflows or service chains becomes an active research field [17, 18]. The international workshop on data derivation and provenance and its follow-up workshops, namely International Provenance and Annotation Workshop (IPAW), have been held five times and resulted in the "provenance challenge" activities. Within GIS domain, how to incorporate provenance support in geospatial services is still a challenge. The use of OGC CSW for serving geospatial provenance is compliant with existing service standards in geospatial domain can allows easy integration with legacy GIS applications.

Some efforts have been devoted to the use of Semantic Web technologies for representing and querying data provenance information [19-22]. Our approach differs from their approaches in that we use existing registry services for management of provenance. The registration of ontologies in ebRIM can support semantics-enhanced discovery of information resources in registries [9-12]. The work here extends this approach in the provenance research area and proposes the registration of semantic provenance in the ebRIM model.

Provenance investigation in GIS can be traced back to Lanter's [23] work on data lineage metadata. Frew et al. [24] provide lineage support for remote sensing data processing in a script-based environment. Wang et al. [25] proposed a provenance-aware architecture to record the lineage of spatial data. Tilmes and Fleig [26] discuss some general concerns of provenance tracking for Earth science data processing systems. Plale et al. [27] described architectural considerations to support provenance collection and management in geosciences. Yue et al. [28] propose provenance capture in geospatial service composition when instantiating a geoprocessing model into an executable service chain. How provenance can be integrated into existing service-oriented GIS applications has not been addressed in the literature. In addition, the arrangement of provenance in the CSW-ebRIM profile facilitates the query of data, service, and knowledge level of provenance by exploring the associations among provenance, data, services, and chains.

VII. CONCLUSION AND FUTURE WORK

The ontology approach for provenance representation provides a common vocabulary for provenance information and defines explicitly the meaning of the terms and the relations between them. Registration of provenance ontologies in CSW allows users to take advantage of that benefit in registries. This paper describes how semantic provenance can be registered into the ebRIM-based CSW. Such a semantically-enriched CSW provides support in discovery of data, service, and knowledge level of geospatial provenance. Future work includes developing user-friendly tools to facilitate provenance registration and visualization of query results, exploring the lifetime management of provenance information, and developing provenance-aware applications to demonstration advantages and usage of provenance.

ACKNOWLEDGEMENT

We are grateful to the anonymous reviewers for their comments. This work was funded jointly by Project 40801153 supported by NSFC, 863 Program of China (2007AA120501), LIESMARS and SKLSE (Wuhan University) Special Research Fundings.

REFERENCE

[1] D. Clery and D. Voss, "All for one and one for all," Science, 308 (5723), p. 809, 2005.

[2] OGC, Open Geospatial Consortium, www.opengeospatial.org, [Accessed 16 May, 2010].

[3] D. Nebert, A. Whiteside, and P. Vretanos (eds), OpenGIS® Catalog Services Specification, Version 2.0.2, OGC 07-006r1, Open GIS Consortium Inc. 218 pp, 2007.

[4] R. Martell (ed), CSW-ebRIM Registry Service—Part 1: ebRIM profile of CSW, Version 1.0.0, OGC 07-110r2, Open Geospatial Consortium, Inc., 57 pp, 2008.

[5] B. Srivastava and J. Koehler, "Web service composition - current solutions and open problems," in: Proceedings of International Conference on Automated Planning and Scheduling (ICAPS) 2003 Workshop on Planning for Web Services, Trento, Italy, pp. 28-35.

[6] J. Rao and X. Su, "A survey of automated web service composition methods," in: Proceedings of First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004), San Diego, California, USA, pp. 43-54.

[7] J. Peer, Web service composition as AI planning - a survey, Technical report, University of St.Gallen, Switzerland, 63 pp, 2005.

[8] Y. Wei, L. Di, B. Zhao, G. Liao, A. Chen, Y. Bai, and Y. Liu, 2005. "The Design and Implementation of a Grid-enabled Catalogue Service," 25th Anniversary IGARSS 2005, July 25-29, COEX, Seoul, Korea. pp. 4224-4227.

[9] A. Dogac (ed.), ebXML Registry Profile for Web Ontology Language (OWL), Version 1.5, regrep-owl-profile-v1.5-cd01, Organization for the Advancement of Structured Information Standards (OASIS). 76 pp, 2006.

[10] A. Dogac, Y. Kabak, G.B. Laleci, C. Mattocks, F. Najmi, and J. Pollock, "Enhancing ebXML registries to make them OWL aware," Distributed and Parallel Databases Journal, Springer-Verlag, 18(1), pp. 9-36, July 2005.

[11] W. Liu, K. He, and W. Liu, "Design and realization of ebXML registry classification model based on ontology." In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), 2005, pp. 809-814.

[12] A. Bechini, A. Tomasi, and J. Viotto, "Enabling ontology-based document classification and management in ebXML registries," In: Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil, 2008, pp. 1145-1150.

[13] P. Yue, J. Gong, L. Di, L. He, and Y. Wei, "Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure," GeoInformatica. 2009. DOI: 10.1007/s10707-009-0096-1.

[14] A. Woodruff and M. Stonebraker, "Supporting fine-grained data lineage in a database visualization environment," In: Proceedings of International Conference on Data Engineering (ICDE), 1997, pp. 91-102.

[15] Y. Cui, J. Widom, and J. L. Wiener, "Tracing the lineage of view data in a warehousing environment," ACM Transactions on Database Systems, 25(2), pp. 179-227, 2000.

[16] P. Buneman, S. Khanna, and W. C. Tan, "Why and where: a characterization of data provenance," In: Proceedings of International Conference on Database Theory (ICDT), 2001, pp. 316-330.

[17] Y. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," SIGMOD Record, vol. 34, pp. 31-36, 2005

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

[19] J. Zhao, C. Goble, M. Greenwood, C. Wroe, and R. Stevens, "Annotating, linking and browsing provenance logs for e-Science," In: Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, FL., USA, 6 pp, 2003.

[20] J. Golbeck and J. Hendler, "A semantic web approach to the provenance challenge," The First Provenance Challenge (this issue), Concurrency and Computation: Practice and Experience, 20 (5), pp. 431-439, 2007.

[21] S.S. Sahoo, A. Sheth, and C. Henson, "Semantic provenance for eScience: managing the deluge of scientific data," IEEE Internet Computing, 12 (4), pp. 46-54, 2008.

[22] S. Zednik, P. Fox, D. L. McGuinness, P. P. da Silva, and C. Chang, "Semantic provenance for science data products: application to image data processing," in: Proceedings of the First International Workshop on the Role of Semantic Web in Provenance Management (SWPM 2009), Washington DC, USA, CEUR-WS, vol. 526, October 25 2009, 7 pp.

[23] D. P. Lanter, "Design of a lineage-based meta-data base for GIS," Cartography and Geographic Information Systems, vol. 18, No. 4, pp. 255-261, 1991.

[24] J. Frew, D. Metzger, and P. Slaughter, "Automatic capture and reconstruction of computational provenance," Concurrency and Computation: Practice and Experience. 20(5). John Wiley & Sons, Ltd. pp. 485-496, 2007.

[25] S. Wang, A. Padmanabhan, D. J. Myers, W. Tang, and Y. Liu, "Towards provenance-aware geographic information systems," In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (ACM GIS 2008). 4pp, 2008.

[26] C. Tilmes and J. A. Fleig, "Provenance tracking in an earth science data processing system," In: Proceedings of Second International Provenance and Annotation Workshop (IPAW), LNCS 5272, pp. 221-228, 2008.

[27] B. Plale, B. Cao, C. Herath, and Y. Sun, "Data provenance for preservation of digital geoscience data," Geological Society of America (GSA), Memoir Volume 12, 14pp, 2010.

[28] P. Yue, J. Gong, and L. Di, "Augmenting Geospatial Data Provenance through Metadata Tracking in Geospatial Service Chaining," Computers & Geosciences, vol. 36, no. 3, pp. 270-281, 2010