Computers & Geosciences 36 (2010) 270–281

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Augmenting geospatial data provenance through metadata tracking in geospatial service chaining

Peng Yue [a,*], Jianya Gong [a], Liping Di [b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[b] Center for Spatial Information Science and Systems (CSISS), George Mason University, 6301 Ivy Lane, Suite 620, Greenbelt, MD 20770, USA

ARTICLE INFO

Article history:
Received 18 April 2008
Received in revised form
6 August 2009
Accepted 20 September 2009

Keywords:
Geospatial Web Service
Service chaining
Data provenance
Metadata tracking
GIS
Semantic Web

ABSTRACT

In a service-oriented environment, heterogeneous data from distributed data archiving centers and various geo-processing services are chained together dynamically to generate on-demand data products. Creating an executable service chain requires detailed specification of metadata for data sets and service instances. Using metadata tracking, semantics-enabled metadata are generated and propagated through a service chain. This metadata can be employed to validate a service chain, e.g. whether metadata preconditions on the input data of services can be satisfied. This paper explores how this metadata can be further exploited to augment geospatial data provenance, i.e., how a geospatial data product is derived. Provenance information is automatically captured during the metadata tracking process. Semantic Web technologies, including OWL and SPARQL, are used for representation and query of this provenance information. The approach can not only contribute to the automatic recording of geospatial data provenance, but also provide a more informed understanding of provenance information using Semantic Web technologies.

1. Introduction

1.1. Metadata tracking in geospatial service chaining

Web services technologies have shown promise for providing heterogeneous data from distributed data archive centers for open worldwide use. Previously stand-alone geo-processing functions are now being wrapped as interoperable web services that can be chained to support a "Cyberinfrastructure for e-Science" (Hey and Trefethen, 2005). The Open Geospatial Consortium (OGC) is developing geospatial Web services standards by adapting or extending common-purpose Web service standards. Through the OGC Web Services (OWS) testbeds, OGC has been developing a series of interface specifications under the OGC Abstract Service Architecture (Percivall, 2002), including Web Feature Service (WFS), Web Map Service (WMS), Web Coverage Service (WCS), Catalogue Services for the Web (CSW), and Web Processing Service (WPS). To solve a complex real world problem in a service-oriented environment, multiple services must be chained together. Although the manual composition of service chains is useful, it needs considerable time and requires users to be both domain and technical experts. The wide use of accessible geospatial data and services over the Web requires a certain degree of automation for service composition.

Automatic service composition is a hot research topic in computer science (Srivastava and Koehler, 2003; Rao and Su, 2004). It consists of three phases: (1) process modeling, which involves generating an abstract composite process model consisting of the control flow and data flow among atomic processes; (2) process model instantiation, where the abstract process is instantiated into a concrete workflow or executable service chain; and (3) workflow execution, where the chaining result or workflow is executed in a workflow engine to generate on-demand data products. In the geospatial domain, the process model is a geo-processing workflow, which transforms source data into value-added data products. Each process node (i.e., atomic process) in the process model represents one type of geospatial service. All share the same functional behaviors: functionality, input and output. The descriptions of these behaviors can use service ontologies from the Semantic Web (Berners-Lee et al., 2001). Many approaches are available for generating a process model based on the service ontologies using Artificial Intelligence (AI) planning methods (Peer, 2005). A process model contains knowledge about how to generate a data product. Since this data product does not really exist in any archive, it is regarded as a *virtual data product*. This virtual data product represents a geospatial data type, not an instance (an individual data set), that the process model can produce. It can be materialized on demand for users when all required

* Corresponding author. Tel.: +86 27 68778755
E-mail address: geopyue@gmail.com (P. Yue).

0098-3004/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2009.09.002

geo-processing services and input data are available. The materialization of a virtual data product requires metadata specifications such as spatial bounding box and spatial projection. The term *metadata* in metadata tracking means descriptive information for data products such as that defined in ISO 19115 (ISO/TC 211, 2003). Through propagating these specifications to each process node of a process model, the whole process model is instantiated. Therefore, metadata tracking, the generation and propagation of geospatial metadata through the process model, is a key step towards the instantiation of the process model.

1.2. Geospatial data provenance

As the number of geospatial services grows with the wider integration of geospatial services, it becomes important to automate the recording of data provenance rather than relying on manual work (Foster, 2005). Data provenance, also referred to as lineage, contains information about the sources and production processes used in producing a data product (ISO/TC 211, 2003). With the development of multi-sensor and multi-platform technologies, the processing and transformation of multi-resolution and multi-spectral images becomes more and more frequent and complex. Therefore, data provenance is important to help users make decisions about the quality of derived data products, discover dependencies among data and services, or re-enact the process of derivation of data products.

This paper describes a synergistic effort between automatic service composition and data provenance. Most existing work on data provenance is in the domain of general information (Bose and Frew, 2005; Simmhan et al., 2005; Miles et al., 2007; da Silva et al., 2003; Zhao et al., 2004; Foster et al., 2002; Golbeck and Hendler, 2007) and does not include content specific to the geospatial domain, such as geospatial metadata standards. Although there has been some work in the geospatial domain (Lanter, 1991, 1992; Alonso and Hagen, 1997; Frew et al., 2001, 2007; Wang et al., 2008; Tilmes and Fleig, 2008), it did not consider the service-oriented environment enabled by OGC Web service standards. The emergence of Semantic Web technologies, including the Resource Description Framework (RDF) (Klyne and Carroll, 2004), the Web Ontology Language (OWL) (Dean and Schreiber, 2004), and the SPARQL Protocol and RDF Query Language (SPARQL) (Prud'hommeaux and Seaborne, 2006), provides a way to connect data for more effective discovery and integration, and thus shows considerable promise for new approaches to geospatial data provenance. In addition, the previous work focused mostly on analyses of provenance information that was created during execution, rather than on metadata generated before execution (Kim et al., 2006). The geospatial metadata generated during process model instantiation provides a context for evaluating the quality and reliability of the data product before the intensive execution of the workflow, thus contributing to the data product's provenance. Therefore, this paper presents how to interleave the Semantic Web approaches for data provenance with metadata tracking to record and query provenance information generated in instantiating a process model. The contributions of this paper are: (1) a model and semantic representation for geospatial data provenance that integrates the geospatial metadata standard and process models for geo-processing services and service chains; (2) automatic capture of geospatial data provenance through metadata tracking in the phase of process model instantiation; and (3) support to the storage and query of geospatial data provenance through Semantic Web technologies.

The remainder of the paper is organized as follows. Section 2 introduces a use case to help in understanding our work. The primary challenges for research on geospatial data provenance in a service-oriented environment are described in Section 3. In Section 4, the approach for addressing these challenges is presented, including a semantic representation of geospatial data provenance, a metadata-tracking component, and extension of the metadata-tracking component to support automatic recording and querying of geospatial data provenance. The work is compared with related work in Section 5, and conclusions and pointers to future work are given in Section 6.

2. A use case

An Earth science application serves as an example to help understand metadata tracking during service chaining and to illustrate how metadata tracking can contribute to data provenance. The application is wildfire prediction from weather and remote sensing data. The wildfire prediction process uses a variety of geospatial data items when creating the wildfire prediction product. This input data consists of the Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (FPAR), Land Cover/Use Types (LULC), daily maximum temperature, daily minimum temperature, and precipitation.

The process model is derived in the process modeling phase. Its representation is formalized through an ontology approach. In information sciences, an ontology is a formal, explicit specification of a conceptualization that provides a common vocabulary for a knowledge domain and defines the meaning of the terms and the relations between them (Gruber, 1993). Ontologies are crucial to making the semantics of the exchanged content machine-understandable. OWL, recommended by W3C as the standard Web ontology language, is designed to enable the creation of ontologies and the instantiation of these ontologies in the description of resources. Therefore, process models for geo-processing workflows are addressed through the introduction and design of OWL-based ontologies conveying semantic information on geospatial services and data. The following ontology entities are linked to the process model for wildfire prediction in the upper part of Fig. 1: "WildFirePrediction" for the semantics of service functions, "FPAR", "LAI", "IGBP_CLASS[1]", "Maximum_Temperature", "Minimum_Temperature", and "Precipitation_Amount" for the semantics of input data, and "Wildfire_Danger_Index" for the semantics of output data.

We can refer to this process model as a virtual data product for wildfire prediction. Therefore, an instance of this virtual data product can be generated with metadata specifications through the materialization process. For example, a user provides the spatial (e.g. Bakersfield, CA, United States) and temporal (e.g. August 26, 2006) information. A semantically augmented geospatial catalogue service (Yue et al., 2006) can be used to automatically determine that the National Oceanic & Atmospheric Administration (NOAA) National Digital Forecast Database[2] (NDFD) can provide the weather data (MAXT, MINT and QPF) and National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Moderate Resolution Imaging Spectroradiometer (MODIS)[3] products can provide FPAR, LAI, and LULC.

[1] Land cover classes defined by the International Geosphere–Biosphere Program (IGBP).

[2] The operational NDFD data provided by the NOAA National Weather Service (NWS) are stored in the GRIB2 data format with a Lambert conformal coordinate reference system and a spatial resolution of 5-km.

[3] The operationally available NASA data in the Land Processes Distributed Active Archive Center (LPDAAC) are stored in HDF-EOS data format, and in a sinusoidal grid coordinate reference system at a spatial resolution of 1-km. The MODIS grids are stored as tiles, each covering approximately 1200×1200 square kilometers.

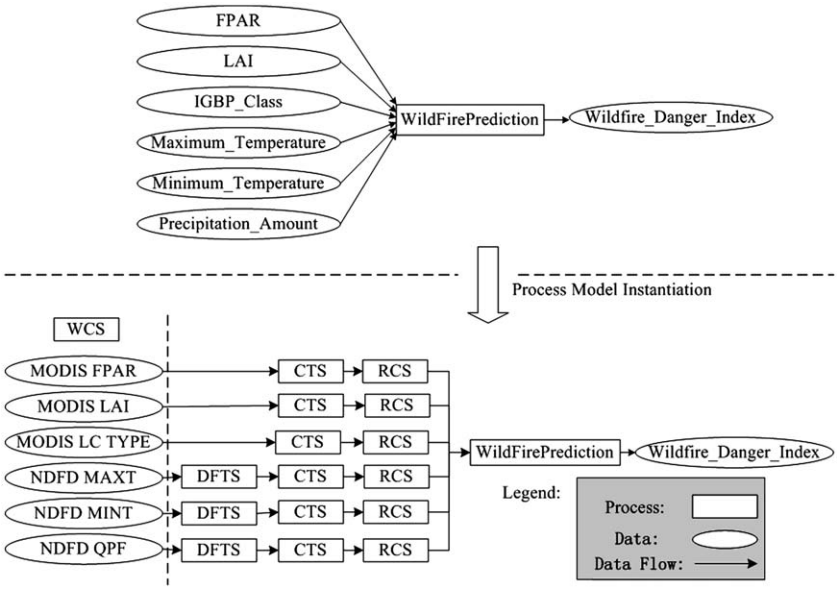

Fig. 1. Wildfire prediction case.

These data are accessible through a standards-compliant service, WCS. As illustrated in the lower part of Fig. 1, the following data reduction and transformation services, Coordinate Transformation Service (CTS), Data Format Transformation Service (DFTS) and Resolution Conversion Service (RCS), are needed to transform the NDFD and MODIS data into a form that can be readily accepted by the wildfire prediction service.

3. Challenges in geospatial data provenance

Determining geospatial data provenance in a service-oriented environment faces the following challenges. Although the following list is not intended to be exhaustive, it includes the primary issues that need to be considered.

- Representation of geospatial data provenance: the way in which provenance information is represented has direct impacts on its understanding and usage. Examples of possible content for geospatial data provenance are metadata descriptions of source data (e.g. NDFD or MODIS data from data archive centers), transformation functionalities (e.g. geo-processing services), geo-processing workflow (e.g. geospatial service chaining), parameters used, intermediate geospatial data product(s), and date and time. The representation may have a level of granularity. The common characteristics could be shared at a high level, while the differences would be represented at a low level. For example, the common provenance information for instances of a virtual data product can be represented at a high level. Extensible Markup Language (XML), a primary format for message exchange in a service-oriented environment, is suitable for representation of provenance. The ontology language based on XML, i.e., OWL, can be used to formalize the semantics conveyed in the provenance information, thus supporting ontology-based query, navigation and reasoning capabilities.
- Automatic capturing of geospatial data provenance: provenance information can be captured by tracing the execution of the workflow engine or aggregating provenance information generated by distributed geospatial service providers. Automating provenance capture requires extending legacy geospatial applications with provenance-capturing functions, from either the workflow engine or geospatial services, to enable provenance-aware geospatial applications. The provenance information captured can be converted to a semantic-enabled representation and propagated from source data to derived data.
- Management of geospatial data provenance: provenance information can be tightly coupled with geospatial data, as by using the lineage tag in the ISO 19115 metadata standard. However, in the general information domain, provenance metadata tends to be separate from other metadata and data provenance is maintained through a provenance management store. This approach integrates the provenance information of dependable data products, thus facilitating storing and querying of provenance.

4. Augmenting geospatial data provenance through metadata tracking

Fig. 2 shows a high-level modular view of the overall architecture designed to address the challenges mentioned in Section 3. This figure shows that an ontology-based representation for geospatial data provenance is proposed for semantic description of geospatial data provenance (Section 4.1). The metadata-tracking component (Section 4.2) from semantics-based automatic service composition is extended to support the automatic recording of geospatial data provenance (Section 4.3). Provenance information is managed in provenance stores, using Semantic Web technologies (Section 4.4). Finally, Section 4.5 describes the implementation.

4.1. Semantic description of geospatial data provenance

An ontology is defined for representing provenance. In the Web ontology context, the basic elements for ontology are classes, properties, and individuals. Classes group individuals into categories; properties stand for binary relations between those individuals. The organization of these elements follows the RDF triple form: subject–predicate–object, because RDF, the basis of

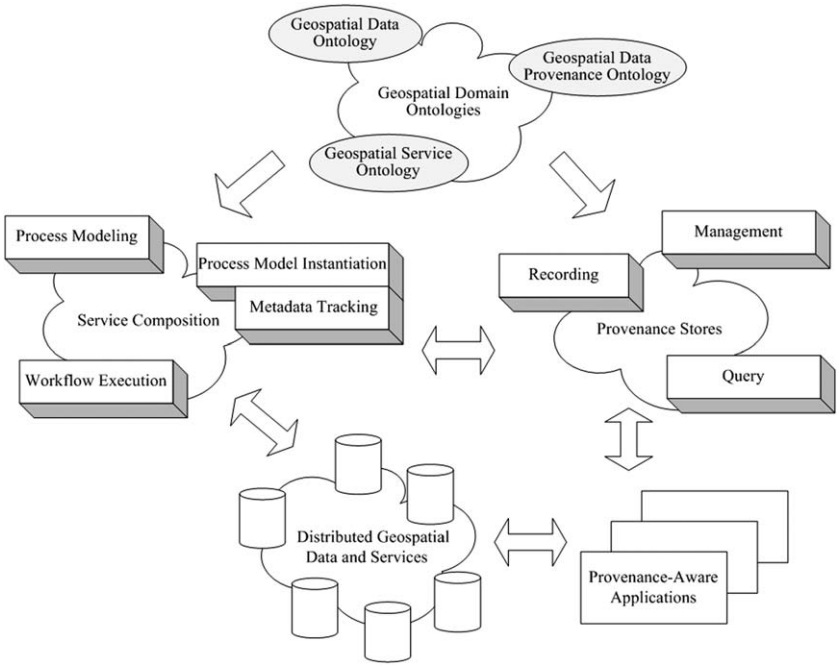

Fig. 2. A high-level modular view of overall architecture.

OWL, provides a flexible model for describing Web resources and relations among these resources. An OWL file can be defined based on RDF. Using these technologies, the implicit provenance information such as indirect ancestor data products can be discovered using OWL reasoners.

The ontology design of geospatial data provenance is within the context of a Semantic Web-enabled geospatial Web services environment. In this environment, geospatial Web services that are either standards-compliant or not compliant with standards are semantically described using OWL-S (Martin et al., 2004). In the Web service domain, semantics for Web services can be classified into four types (Cardoso and Sheth, 2005): (1) data/information semantics, (2) functional/operational semantics, (3) execution semantics, and (4) Quality of Service (QoS) semantics. The geospatial "DataType" and "ServiceType" ontologies address the data and functional semantics of geospatial Web services, respectively (Yue et al., 2007a). Since metadata standards have been defined after many years of work by geospatial experts and community consensus, the geospatial "DataType" ontology can be enriched with metadata ontologies (Bermudez and Piasecki, 2006) to allow more precise description of geospatial data, and support cross-metadata-standards discovery (Bermudez, 2004). The execution semantics of a geospatial service can be specified using the metadata statement in the preconditions and effects (Yue et al., 2007b). Furthermore, the "DataType" ontology, enriched with such metadata ontologies in the OWL-S service grounding, can be used to define a set of bidirectional mappings between the schemas of the OGC-compliant services and the ontologies, thus addressing the structural interoperability required by geospatial Web services. In the wildfire prediction case, OWL-S descriptions for WPS-based geospatial services such as wildfire prediction service, CTS, and RCS have been developed.

Semantic description for geospatial data provenance in the Semantic Web environment should integrate existing semantic descriptions for geospatial Web services with provenance-specific information. The Semantic Web approach of Golbeck and Hendler (2007) is employed, enhanced with geospatial content. The ontology for geospatial data provenance is defined using OWL classes and properties for four types of entities: geospatial data products, geospatial Web services, atomic service executions, and service chain executions:

- Geospatial data products: the execution of OWL-S files for geospatial Web services in the Semantic Web environment creates OWL individuals for classes in the geospatial "DataType" ontology, each representing a geospatial data instance. We propose the "ProvenanceGeoDataType" entity class in the context of the geospatial "DataType" ontology, defining it to be a subclass of "GeoDataType". The individuals of this class represent provenance information for individual geospatial data products, including intermediate geospatial data products. "GeoDataType" is a top-level class in the geospatial "DataType" ontology. Ontology entities such as "FPAR" and "LAI" in Section 2 are all sub-classes of "GeoDataType". For each individual of "ProvenanceGeoDataType", an inherited property "hasMD_Metadata" links it to an ISO 19115-based metadata description. The property "hasGeoDataTypeAncestor" connects a geospatial data product to its ancestor geospatial data product. It is defined as a transitive property so that a reasoner can infer an indirect ancestry relation based on the transitive relation of existing data products. The property "hasGeoDataTypeParent" can be defined to link a geospatial data product to its direct ancestor geospatial data product. Another property "producedBy" describes the service execution that produces the geospatial data.
- Geospatial Web services: ontologies for geospatial Web services can use OWL-S directly since OWL-S has provided an ontology framework for Web services. OWL-S specifies the semantics of a geospatial Web service, including the geospatial data exchanged (i.e., input/output), the geo-processing functionality, and pre-/post-conditions. The class "Service" from ontologies can refer in OWL-S to either an atomic geospatial service or a geo-processing service chain.
- Atomic service executions: the class "ServiceExecution" is the superclass of "AtomicServiceExecution" and "CompositeServiceExecution". It describes instances of a geospatial service execution. The properties "hasInput" and "hasOutput" give the

input and output geospatial data products of the execution of a specific geospatial service. Service execution is connected to OWL-S descriptions through the property "hasService". Additional properties, such as annotating date and time for the service execution can also be added to the class "ServiceExecution". For an atomic service, the property "isContainedBy" links it to the geo-processing service chain execution that invoked it.

- Service chain executions: the class "CompositeServiceExecution" describes the execution of a geo-processing service chain. Therefore, the property "hasService" inherited from the class "ServiceExecution" links it to a geo-processing service chain described in the composite process of OWL-S.

Fig. 3 illustrates the relations among these entities using OntoViz (OntoViz, 2006). This example is a lightweight ontology, for purpose of demonstration. More properties and entities can be added to provide richer provenance information.

4.2. Metadata-tracking component

The role of metadata tracking for geospatial service composition was introduced in Section 1.1. According to Yue et al. (2007b), the main functions of this component are

- Semantic metadata generation: data repository services, such as OGC CSW allow the input data of the service chain, i.e., those that already physically exist in a data archive, to be queried with the user-specified metadata to obtain detailed metadata information. If such information is not available, a metadata generation component can be used to extract metadata from data encoded in self-describing file formats such as HDF-EOS and GeoTIFF. This metadata is represented using Semantic Web technologies. To ensure interoperability, this metadata uses an existing OWL ontology (Drexel, 2004) for the ISO 19115 metadata standards in order to follow international standards.
- Metadata validation: when all the metadata of the input data have been generated, the preconditions of the service can be validated with the generated metadata. Preconditions identify the metadata constraints that the input data of an individual service must follow, e.g. a particular file format or a specific coordinate reference system. These preconditions are specified through the expressions in OWL-S.
- Metadata precondition satisfaction: we enhance the capabilities of the metadata-tracking component by formulating built-in rules when preconditions are not satisfied. These rules ensure automatic insertion of data reduction and transformation services to modify the data, so that the preconditions can be satisfied. For example, according to the preconditions of the wildfire prediction service, the data reduction and transformation services shown in the lower part of Fig. 1 are all inserted automatically using these rules.
- Metadata propagation: the output metadata can be derived from the input metadata by modifying, deleting, or inserting metadata elements in the input affected by the service operation, using the service capabilities of OWL-S. Table 1 shows an example where resolution and coordinate reference system information are updated while the file format information is still the same as in the original MODIS data.

4.3. Automatic recording of geospatial data provenance

The discussion in this paper concentrates on augmenting provenance information through metadata tracking before the intensive execution of service chains. The metadata-tracking component is developed to deal specifically with the geospatial domain and focuses on tracking geospatial metadata. This component interacts with the OGC CSW to generate geospatial metadata and manages validation of geospatial metadata, precondition satisfaction, and propagation. Provenance capturing through metadata tracking is implemented by extending the metadata-tracking component. In addition to those functionalities mentioned in Section 4.2, this component automatically

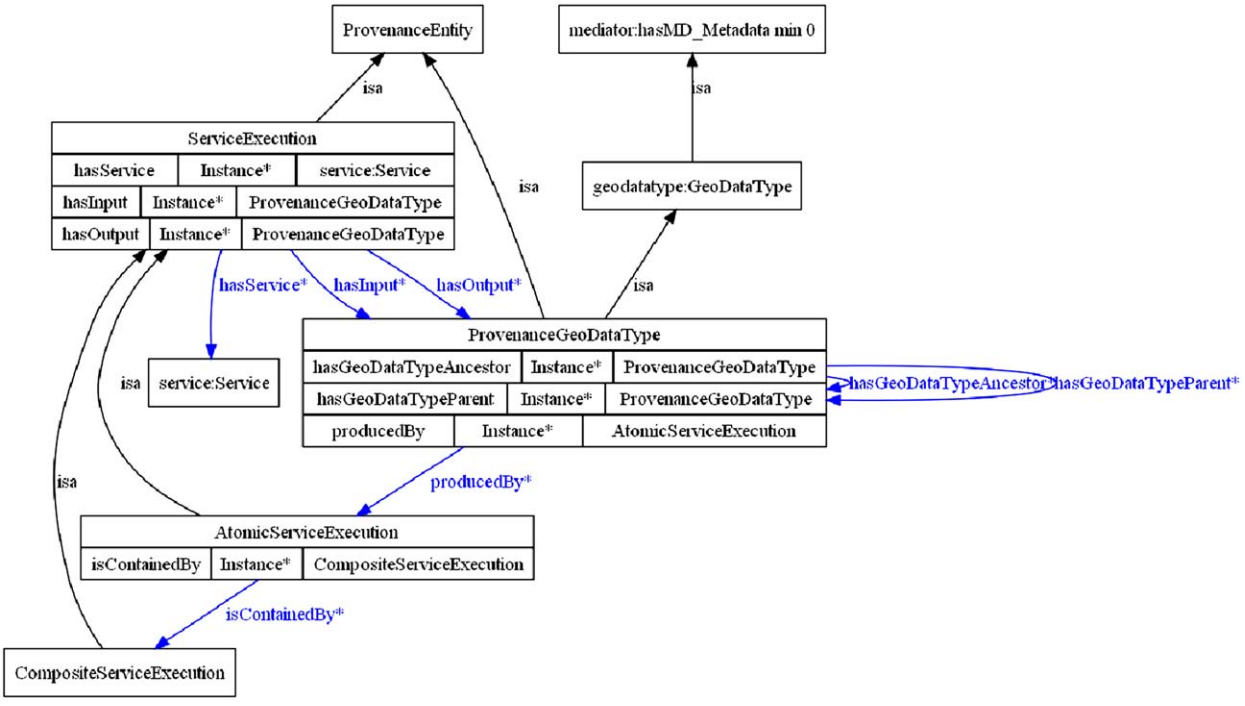

Fig. 3. An ontology for geospatial data provenance.

Table 1
Metadata tracking results.

Before	After
hasMD_Metadata: MD_Metadata: distributionInfo: MD_Distribution: distributionFormat: MD_Format: name_MD_Format: application/HDF-EOS identificationInfo: MD_DataIdentification: dataExtent: … upperCorner: DirectPosition: coordinates: **-10007554.677000,5559752.598333** lowerCorner: DirectPosition: coordinates: **-11119505.196667,3335851.559000** referenceSystemInfo: MD_ReferenceSystem: referenceSystemIdentifier: RS_Identifier: code: **AUTO2:42016,1,0,0** spatialRepresentationInfo: MD_GridSpatialRepresentation: axisDimensionProperties: MD_Dimension: dimensionSize: **2400** dimensionName: row axisDimensionProperties: MD_Dimension: dimensionSize: **1200** dimensionName: column …	hasMD_Metadata: MD_Metadata: distributionInfo: MD_Distribution: distributionFormat: MD_Format: name_MD_Format: application/HDF-EOS identificationInfo: MD_DataIdentification: dataExtent: … upperCorner: DirectPosition: coordinates: **-1300330.654000,-38780.983000** lowerCorner: DirectPosition: coordinates: **-2038330.654000,-1241780.983000** referenceSystemInfo: MD_ReferenceSystem: referenceSystemIdentifier: RS_Identifier: code: **AUTO2:42004,1,-100,45** spatialRepresentationInfo: MD_GridSpatialRepresentation: axisDimensionProperties: MD_Dimension: dimensionSize: **1203** dimensionName: row axisDimensionProperties: MD_Dimension: dimensionSize: **738** dimensionName: column …

instantiates classes in provenance ontologies and assigns properties with values generated from the metadata tracking process. It parses OWL-S and traverses each geospatial service in the geo-processing service chain. OWL-S can represent a geo-processing service chain using a composite process.

Each metadata-tracking process run uses two ontology files. The first is the geospatial data provenance ontology, which provides a scheme for provenance recording. The second is an OWL-S composite service for a geo-processing service chain. It contains OWL-S descriptions for atomic geospatial services, links between input/output data types and services, and relationships between atomic services and the composite service. This information can be part of the provenance information. The metadata-tracking component relies on it to interrogate the OGC CSW for semantically matched input data.

At the beginning of the metadata tracking process, an instance of "CompositeServiceExecution" is generated. The instance is represented as an OWL individual. During the metadata tracking process, the metadata-tracking component assigns semantic geospatial metadata that has been generated to instances of "ProvenanceGeoDataType" using the property "hasMD_Metadata". Each traverse through a geospatial service in the geo-processing service chain generates an instance of "AtomicServiceExecution", recoding its input data, output data, and related OWL-S description, and linking it to the instance of "CompositeServiceExecution" using the property "isContainedBy". At the end of the metadata tracking process, all this geospatial data provenance information is available and can be aggregated into a RDF triple store.

4.4. Querying geospatial data provenance

The storage management of a RDF triple store can use existing Semantic Web systems such as Jena (Jena, 2006). We use SPARQL to query the triple store. The query example in Table 2 finds what geospatial services were executed to generate geospatial data products. Additional queries can be formulated by following the links enabled by ontologies.

The PREFIX lines simply define several prefixes for selected namespaces, so that they need not be entered every time there is a reference to them. The rest of the query has a SQL-like style. The SELECT clause specifies what the query should return, in this example, the service desired. The WHERE clause consists of a set of triple patterns that will be matched in a RDF triple store.

The execution result of the above query example is shown in the SPARQL query panel (Fig. 4) of Protégé (Protégé, 2006). Each row in the results panel shows the execution of an atomic service and its corresponding output data.

The results of a SPARQL query can be returned in SPARQL Query Results XML Format (Beckett and Broekstra, 2008). This format bridges RDF and existing XML tools so that the results of a SPARQL query can be transformed into various formats using XSLT (Clark, 1999), XPATH (Clark and DeRose, 1999), etc.

SPARQL can be used to query various aspects of geospatial data provenance, including spatial characteristics. The query example in Table 2 shows geospatial services that created geospatial data products. The query in Table 3 finds the file format of an intermediate geospatial data product, and Table 4 shows the results of this query in XML format. Table 5 shows the query for

Table 2
Provenance query example (1).

```
PREFIX geoprov: <http://www.laits.gmu.edu/geo/ontology/domain/geoprovenance.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?pgdt ?execution ?service
WHERE {
  ?pgdt rdf:type geoprov:ProvenanceGeoDataType .
  ?pgdt geoprov:producedBy ?execution .
  ?execution geoprov:hasService ?service
}
```

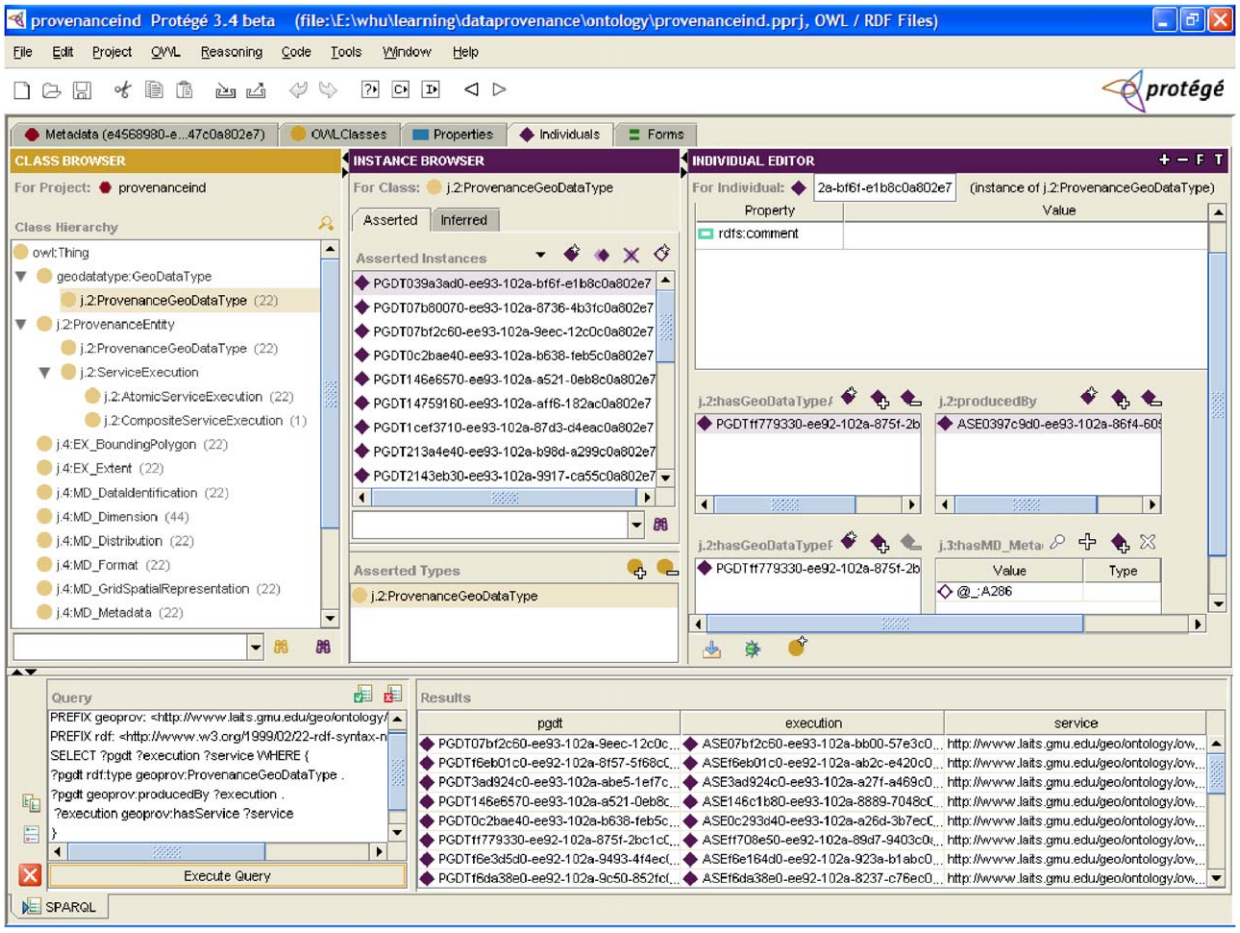

Fig. 4. Provenance information for wildfire prediction case.

the spatial bounding box and grid dimension size of a geospatial data product in the wildfire prediction case, and the result of this query is shown in Table 6.

4.5. Implementation

We have developed a metadata-tracking component[4] in our prototype system for geospatial service composition to enable the instantiation of process models (Yue et al., 2007b). The process modeling is implemented on top of an OWL-S Manager (OWLS-Manager) (Yue et al., 2007a), a component for OWL-S Files Management, which can deploy and undeploy OWL-S files for geospatial services into the knowledge base. The instantiation process interacts with a semantically augmented CSW (Yue et al., 2006). OWL-S API (OWL-S API, 2004) is used for parsing and traversing each service in the service chain.

To run the wildfire prediction scenario in this system, we have implemented related Web services and created OWL-S descriptions for these geospatial services. Fig. 4 shows the provenance information generated from the extended metadata-tracking component for the wildfire prediction case. Totally 22 individuals

[4] The metadata-tracking component was successfully demonstrated in July 2007 at Summer ESIP Federation meeting in University of Wisconsin, Madison, Wisconsin, USA.

Table 3
Provenance query example (2).

```
PREFIX iso19115: <http://loki.cae.drexel.edu/~wbs/ontology/2004/09/iso-19115#>
PREFIX mediator: <http://www.laits.gmu.edu/geo/ontology/domain/v3/mediator_v3.owl#>
PREFIX fileformat: <http://www.laits.gmu.edu/geo/ontology/domain/v2/fileformat.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geoprov: <http://www.laits.gmu.edu/geo/ontology/domain/geoprovenance.owl#>
PREFIX geoprovind: <http://www.laits.gmu.edu/temp/e4568980-ee92-102a-a26e-8647c0a802e7#>
SELECT ?file_format_name
WHERE {
  geoprovind:PGDTf6da38e0-ee92-102a-9c50-852fc0a802e7 mediator:hasMD_Metadata ?md_metadata .
  ?md_metadata rdf:type iso19115:MD_Metadata .
  ?md_metadata iso19115:distributionInfo ?md_disinfo .
  ?md_disinfo rdf:type iso19115:MD_Distribution .
  ?md_disinfo iso19115:distributionFormat ?file_format .
  ?file_format iso19115:name_MD_Format ?file_format_name
}
```

Table 4
Results of the query (2) in XML format.

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
    <head>
        <variable name="file_format_name"/>
    </head>
    <results>
        <result>
            <binding name="file_format_name">
                <literal>application/HDF-EOS</literal>
            </binding>
        </result>
    </results>
</sparql>
```

of "AtomicServiceExecution" and "ProvenanceGeoDataType" are generated, respectively, linked with each other through OWL properties defined in the provenance ontology.

We use the Web interface provided by Joseki (Joseki, 2008), an application for publishing RDF data on the Web that was built on Jena, to answer the provenance query (Fig. 5). To find hidden relations, we add the reasoner "OWLMicroFBRuleReasoner" into the Jena assembler description (Assembler, 2008) when configuring the provenance RDF data set for the Joseki. As shown in Fig. 5, the execution of the query on the property "hasGeoDataTypeAncestor" returns three results on an html page after using XSLT. If no reasoner is configured, the result of this query will be the only one, because the reasoner can infer additional relations from the transitive property "hasGeoDataTypeAncestor". The demonstration using the Joseki and related OWL files can be downloaded and tested with different SPARQL queries.[5]

5. Related work and discussion

In a service-oriented environment, service chaining is a major approach for service integration and bears workflow characteristics (Percivall, 2002). Much work in the general information domain has contributed to determining the provenance of

[5] Related OWL-S files, Joseki demonstration and other resources are available at www.laits.gmu.edu/geo/nga/provenance.html.

Table 5
Provenance query example (3).

```
PREFIX iso19115: <http://loki.cae.drexel.edu/~wbs/ontology/2004/09/iso-19115#>
PREFIX iso19107: <http://loki.cae.drexel.edu/~wbs/ontology/2004/09/iso-19107#>
PREFIX mediator: <http://www.laits.gmu.edu/geo/ontology/domain/v3/mediator_v3.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geoprov: <http://www.laits.gmu.edu/geo/ontology/domain/geoprovenance.owl#>
PREFIX geoprovind: <http://www.laits.gmu.edu/temp/e4568980-ee92-102a-a26e-8647c0a802e7#>
SELECT ?dimRow ?dimCol ?lowerCorner ?upperCorner
WHERE {
  geoprovind:PGDT2e3f6f80-ee93-102a-8da0-6f7fc0a802e7 mediator:hasMD_Metadata ?md_metadata .
  ?md_metadata rdf:type iso19115:MD_Metadata .
  ?md_metadata iso19115:spatialRepresentationInfo ?md_gridrep .
  ?md_gridrep rdf:type iso19115:MD_GridSpatialRepresentation .
  ?md_gridrep iso19115:axisDimensionProperties ?md_dimrow .
  ?md_dimrow rdf:type iso19115:MD_Dimension .
  ?md_dimrow iso19115:dimensionName iso19115:row .
  ?md_dimrow iso19115:dimensionSize ?dimRow .
  ?md_gridrep iso19115:axisDimensionProperties ?md_dimcol .
  ?md_dimcol rdf:type iso19115:MD_Dimension .
  ?md_dimcol iso19115:dimensionName iso19115:column .
  ?md_dimcol iso19115:dimensionSize ?dimCol .
  ?md_metadata iso19115:identificationInfo ?md_dataid .
  ?md_dataid rdf:type iso19115:MD_DataIdentification .
  ?md_dataid iso19115:dataExtent ?md_dataex .
  ?md_dataex rdf:type iso19115:EX_Extent .
  ?md_dataex iso19115:geographicElement ?ex_boundingpoly .
  ?ex_boundingpoly rdf:type iso19115:EX_BoundingPolygon .
  ?ex_boundingpoly iso19115:polygon ?gm_env .
  ?gm_env rdf:type iso19107:GM_Envelope .
  ?gm_env iso19107:lowerCorner ?lcpos .
  ?lcpos iso19107:coordinates ?lowerCorner .
  ?gm_env iso19107:upperCorner ?ucpos .
  ?ucpos iso19107:coordinates ?upperCorner
}
```

workflow-oriented data (Bose and Frew, 2005; Simmhan et al., 2005; Miles et al., 2007). The provenance of a derived data product is determined by tracing the execution of the workflow and input/output data of each processing step. Several methods are available for provenance acquisition: generating provenance through the workflow engine (Zhao et al., 2003), aggregating provenance information generated by each service provider (Foster et al., 2002), or a combination of the previous two methods (Miles et al., 2007). Some efforts have been devoted to the use of Semantic Web technologies, including RDF, OWL, and SPARQL SPARQL, for representing and querying data provenance information (da Silva et al., 2003; Zhao et al., 2004; Golbeck and Hendler, 2007). We use a provenance representation similar to that of Golbeck and Hendler (2007). An important characteristic of the geospatial domain is that an application often includes multiple modeling or processing steps involving large and heterogeneous data volumes, such as the wildfire prediction example presented in this paper. Geospatial data provenance therefore should include complex metadata for source or intermediate data products, such as data format, map projection, and spatial and temporal resolution. We have developed geospatial "DataType" and "ServiceType" ontologies for semantic description

Table 6
Results of the query (3) in XML format.

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
    <head>
        <variable name="dimRow"/>
        <variable name="dimCol"/>
        <variable name="lowerCorner"/>
        <variable name="upperCorner"/>
    </head>
    <results>
        <result>
            <binding name="dimRow">
                <literal>2400</literal>
            </binding>
            <binding name="dimCol">
                <literal>1200</literal>
            </binding>
            <binding name="lowerCorner">
                <literal>-155.572392,30.000001</literal>
            </binding>
            <binding name="upperCorner">
                <literal>-103.923052,50.000001</literal>
            </binding>
        </result>
    </results>
</sparql>
```

of geospatial Web services. The geospatial data provenance ontology can then be linked to these geospatial ontologies by using the extensible capability of OWL and adding new properties and classes. We have shown how to incorporate ISO 19115-based metadata description into geospatial data provenance in Section 4.1. Additional geospatial metadata can be added through this approach.

In the geospatial domain, geospatial metadata standards such as ISO 19115 have addressed data provenance in the data quality part of metadata. A lineage metadata tag is defined in ISO 19115. It allows description of process steps or sources used in creating data. However, this description uses free text and does not readily support the automatic processing of provenance information. Lanter (1991) carried out the pioneering research on data provenance in Geographic Information Systems (GISs). Lineage information is collected from the commands performing spatial analyses. Lineage software, called Geolineus (Lanter, 1992), was developed using early command-line based GIS software, namely ARC/INFO. Another example has been Geo-Opera (Alonso and Hagen, 1997), which supports lineage tracking for geospatial applications. It is actually a geospatial extension to the OPERA workflow management system. Frew et al. (2001; 2007) provide lineage support for remote sensing data processing in a script-based environment. Wang et al. (2008) propose a provenance-aware GIS architecture to record the spatial data lineage and related analysis workflows. Tilmes and Fleig (2008) discuss some general concerns of provenance tracking for Earth science data processing systems, although the development of possible solutions is an ongoing work. To the best of our knowledge, how to capture provenance within the context of geospatial Web services and geo-processing service chains has not been addressed in the literature. Our work proposes the use of metadata tracking to automatically derive data provenance before the intensive execution of service chains. In addition, the use of Semantic Web technologies for linking provenance information and semantic descriptions for geospatial Web services and discovering dependencies provides an informed understanding of geospatial data provenance.

Since no actual service is executed during the metadata tracking process, the provenance information collected is partial and does not include the information available only from service executions, for example, execution date and time or physically existing data products. However, the provenance information captured during the metadata tracking process is common to all executions of a service chain. As stated in Section 3, the representation of provenance information may be at a level of

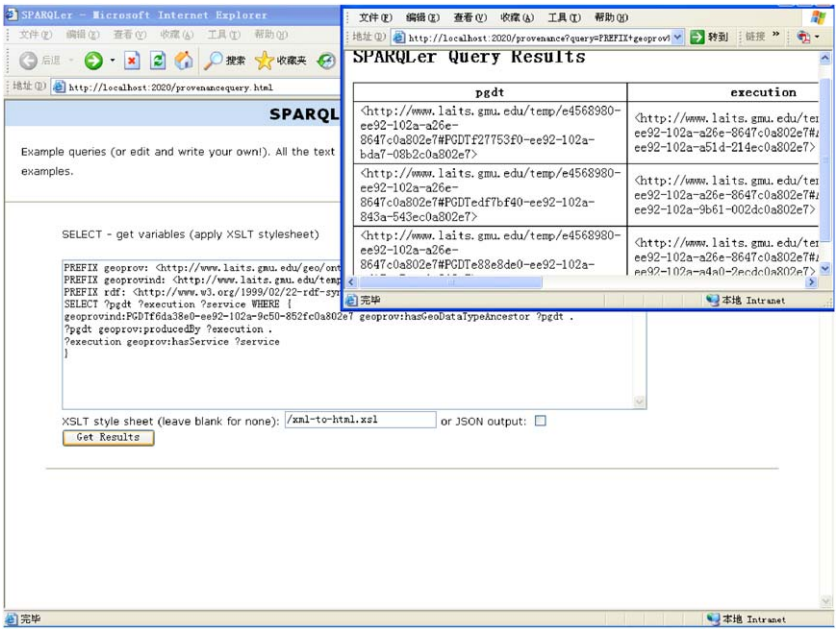

Fig. 5. Provenance query using Joseki.

granularity. From this perspective, we believe that provenance information represents a high level of granularity for geospatial data provenance.

6. Conclusions and future work

Data provenance is an important research issue in Cyberinfrastructure. This paper addresses geospatial data provenance. It describes how to use metadata tracking to augment the provenance of geospatial data. Semantic Web technologies are used to represent, store and query geospatial data provenance. A geospatial data provenance ontology is developed to describe the dependencies among geospatial data products, geospatial Web services, atomic service executions, and service chain executions. The metadata description for geospatial data products follows the ISO 19115 metadata standard and is represented using OWL. The semantics-enabled metadata generated during the metadata tracking is incorporated into geospatial data provenance. The provenance information is stored in a RDF store and is queried using SPARQL. The work demonstrates that metadata tracking can contribute to geospatial data provenance. Ontologies are useful for conveying semantics for geospatial data, services, service chains, and data provenance. Through the underlying connections among entities in geospatial data provenance, Semantic Web technologies provide an effective way towards more informed understanding and query of geospatial data provenance.

A number of observations point to important future directions. Although geospatial metadata standards already include provenance information in the data quality report, aggregating provenance information is more flexible. As the number of accessible data products increases over the Web, the amount of lineage information and the depth of the lineage increase. Therefore, systematic provenance management across different organizations and scientific activities will allow more effective discovery and integration of provenance information. Future work will consider the development of a full-fledged geospatial data provenance system to support the applications of geospatial data provenance. Such a system would provide a user-friendly interface and allow users to browse and navigate provenance information of geospatial data easily. Our work provides an initial framework for geospatial data provenance. Additional application-specific features can be added later to support such applications as quality assessment, data reproducibility, result interpretation, and problem diagnosis.

Although no consensus method is yet available for geospatial domain ontologies, the ontology approach has shown promise along with its emerging applications in research on data provenance. Combined with the underlying RDF model, the ontology approach from the Semantic Web allows the connections among data products to be more explicit and discoverable. Adding extensions to the SPARQL query language to support spatial semantics has been proposed (Kolas, 2008). Combining these extensions with geospatial data provenance query is a worthwhile technique for supporting spatial reasoning for discovering implicit provenance information. An approach towards an interoperable model for data provenance is emerging, namely the Open Provenance Model (OPM) (Moreau et al., 2008). Wang et al. (2008) propose to extend OPM for the development of a spatial data provenance model. Although OPM does not yet discuss the implementation syntax and query language, mapping our provenance model established through the ontology approach onto OPM is a worthwhile technique for improving the interoperability of our implementation. Future work will consider adding richer semantics in the representation of geospatial data provenance and mapping our provenance model onto OPM.

Acknowledgements

We are grateful to the anonymous reviewers, to Prof. Boyan Brodaric and to Dr. Barry Schlesinger for their detailed comments that helped improve the quality of the paper. This work was funded fully or partially by U.S. NGA NURI program (HM1582-04-1-2021), Project 40801153 supported by NSFC, 863 Program of China (2007AA120501, 2007AA12Z214), LIESMARS and SKLSE (Wuhan University) Special Research Funding.

References

Alonso, G., Hagen, C., 1997. Geo-Opera: workflow concepts for spatial processes. In: Proceedings of the 5th International Symposium on Spatial Databases (SSD' 97), Berlin, Germany, 238–258.

Assembler, 2008. Jena Assembler. Hewlett-Packard Labs Semantic Web Programme, www.jena.sourceforge.net/assembler/, (accessed 19.11.2009).
Beckett, D., Broekstra, J., 2008. In: SPARQL query results XML format. World-Wide Web Consortium (W3C) www.w3.org/TR/rdf-sparql-XMLres/, (accessed 19.11.2009).
Bermudez, L., 2004. Ontomet: ontology metadata framework, Ph.D. Dissertation. Drexel University, Philadelphia, USA 177 pp.
Bermudez, L., Piasecki, M., 2006. Metadata community profiles for the semantic web. Geoinformatica 10 (2), 159–176.
Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. Scientific American 284 (5), 34–43.
Bose, R., Frew, J., 2005. Lineage retrieval for scientific data processing: a survey. ACM Computing Surveys 37 (1), 1–28.
Cardoso, J., Sheth, A., 2005. Introduction to semantic web services and web process composition. In: Cardoso, J., Sheth, A. (Eds.), Proceedings First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004). Springer, Berlin, Germany, pp. 1–13 Lecture Notes in Computer Science (LNCS) 3387.
Clark, J., 1999. In: XSL transformations (XSLT). World Wide Web Consortium (W3C) www.w3.org/TR/xslt, (accessed 19.11.2009).
Clark, J., DeRose, S., 1999. In: XML path language. World Wide Web Consortium (W3C) www.w3.org/TR/xpath, (accessed 19.11.2009).
da Silva, P.P., McGuinness, D.L., McCool, R., 2003. Knowledge provenance infrastructure. IEEE Data Engineering Bulletin 26 (4), 26–32.
Dean, M., Schreiber, G., 2004. In: OWL Web ontology language reference. World Wide Web Consortium (W3C) www.w3.org/TR/owl-ref, (accessed 19.11.2009).
Drexel, 2004. In: ISO 19115 Ontology for Geographic Information Metadata. Drexel University, USA www.loki.cae.drexel.edu/~wbs/ontology/, (accessed 17.10. 2005).
Foster, I., 2005. Service-oriented science. Science 308 (5723), 814–817.
Foster, I., Vockler, J., Wilde, M., Zhao, Y., 2002. Chimera: a virtual data system for representing, querying, and automating data derivation. In: Kennedy, J. (Ed.), Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM'02). IEEE Computer Society, Edinburgh, Scotland, pp. 37–46.
Frew, J., Bose, R., 2001. Earth system science workbench: a data management infrastructure for earth science products. In: Kerschberg, L., Kafatos, M. (Eds.), Proceedings of the 13th International Conference on Scientific and Statistical Database Management (SSDBM'01). IEEE Computer Society, Fairfax, Virginia, USA, pp. 180–189.
Frew, J., Metzger, D., Slaughter, P., 2007. Automatic capture and reconstruction of computational provenance. The First Provenance Challenge (Special Issue), Concurrency and Computation: Practice and Experience 20 (5), 485–496.
Golbeck, J., Hendler, J., 2007. A semantic web approach to the provenance challenge. The First Provenance Challenge (Special Issue), Concurrency and Computation: Practice and Experience 20 (5), 431–439.
Gruber, T.R., 1993. A translation approach to portable ontology specification. Knowledge Acquisition 5 (2), 199–220.
Hey, T., Trefethen, A.E., 2005. Cyberinfrastructure for e-Science. Science 308 (5723), 817–821.
ISO/TC 211, 2003. Geographic information—metadata (ISO/DIS 19115). ISO (International Organization for Standardization) TC (Technical Committee) 211/WG 3, Switzerland, 149.
Jena, 2006. Jena. Hewlett-Packard Labs Semantic Web Programme, www.jena. sourceforge.net, (accessed 19.11.2009).
Joseki, 2008. Joseki. Hewlett-Packard Labs Semantic Web Programme, www.joseki. org/, (accessed 19.11.2009).
Kim, J., Gil, Y., Ratnakar, V., 2006. Semantic metadata generation for large scientific workflows, Proceedings of the 5th International Semantic Web Conference, Athens, Georgia, USA, Lecture Notes in Computer Science (LNCS) 4273. Springer, Berlin, Germany, pp. 357–370.
Klyne, G., Carroll, J.J., 2004. Resource Description Framework (RDF): concepts and abstract syntax. World Wide Web Consortium (W3C) www.w3.org/TR/2004/ REC-rdf-concepts-20040210/, (accessed 19.11.2009)..
Kolas, D., 2008. Supporting spatial semantics with SPARQL. Transactions in GIS 12 (s1), 5–18.
Lanter, D.P., 1991. Design of a lineage-based meta-data base for GIS. Cartography and Geographic Information Systems 18 (4), 255–261.
Lanter, D.P., 1992. GEOLINEUS: data management and flowcharting for ARC/INFO. Technical Software Series S-92-2, National Center for Geographic Information and Analysis, Santa Barbara, California, USA.
Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K., 2004. OWL-based Web service ontology (OWL-S). www.daml.org/services/ owl-s/1.1/overview/, (accessed 19.11.2009).
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., Plale, B., Miles, S., Goble, C., Missier, P., Barga, R., Simmhan, Y., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P., Bowers, S., Ludaescher, B., Kwasnikowska, N., den Bussche, J.V., Ellkvist, T., Freire, J., Groth, P., 2008. In: The Open Provenance Model (v1.01). University of Southampton, United Kingdom 35 pp. www.eprints.ecs.soton.ac.uk/16148/1/opm-v1.01.pdf, (accessed 19.11.2009).
OntoViz, 2006. OntoViz Tab. Protégé, www.protegewiki.stanford.edu/index.php/ OntoViz, (accessed 19.11.2009).
OWL-S API, 2004. OWL-S API. Maryland Information and Network Dynamics Lab Semantic Web Agents Project (MINDSWAP), www.mindswap.org/2004/owl-s/ api/, (accessed 19.11.2009).
Peer, J., 2005. Web service composition as AI planning—a survey, University of St. Gallen 63 pp.
Percivall, G., 2002. In: The OpenGIS Abstract Specification Topic 12: OpenGIS Service Architecture (OGC 02-112, Version 4.3). Open Geospatial Consortium Inc., USA 78 pp.
Protégé, 2006. In: Protégé. Stanford University www.protege.stanford.edu/, (accessed 19.11.2009)..
Prud'hommeaux, E., Seaborne, A, 2006. In: SPARQL Query Language for RDF. World Wide Web Consortium (W3C) www.w3.org/TR/rdf-sparql-query/, (accessed 19.11.2009).
Rao, J., Su, X., 2004. A survey of automated web service composition methods. In: Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004), San Diego, California, USA. 43–54.
Simmhan, Y., Plale, B., Gannon, D., 2005. A survey of data provenance in e-Science. ACM (Association for Computing Machinery) SIGMOD (Special Interest Group on Management of Data) Record 34 (3), 31–36.
Srivastava, B., Koehler, J., 2003. Web service composition — current solutions and open problems. In: Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS) 2003 Workshop on Planning for Web Services, Trento, Italy, 28–35.
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.
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, 4.
Yue, P., Di, L., Yang, W., Yu, G., Zhao, P., 2007a. Semantics-based automatic composition of geospatial Web services chains. Computers & Geosciences 33 (5), 649–665.
Yue, P., Di, L., Yang, W., Yu, G., Zhao, P., Gong, J., 2007b. Semantics-enabled metadata generation, tracking and validation in the geospatial web service composition for distributed image mining. In: Proceedings of the 2007 IEEE International Geoscience and Remote Sensing Symposium (IGARSS07), Barcelona, Spain. 334–337.
Yue, P., Di, L., Zhao, P., Yang, W., Yu, G., Wei, Y., 2006. Semantic augmentations for geospatial catalogue service. In: Proceedings of the 2006 IEEE International Geoscience and Remote Sensing Symposium (IGARSS06), Denver, USA. 3486–3489.
Zhao, J., Goble, C., Greenwood, M., Wroe, C., Stevens, R., 2003. Annotating, linking and browsing provenance logs for e-Science. In: Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Florida, USA, 6.
Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., Greenwood, M., 2004. Using semantic web technologies for representing e-Science provenance, In: Proceedings of the Third International Semantic Web Conference (ISWC2004), Hiroshima, Japan, Lecture Notes in Computer Science (LNCS) 3298. Springer, Berlin, Germany, pp. 92–106.