# A Linked Data Approach for Geospatial Data Provenance

Jie Yuan, Peng Yue, Jianya Gong, and Mingda Zhang

***Abstract*— Geospatial data provenance records sources and processing steps that are used in deriving geospatial data products. In the Web of Data environment enabled by Linked Data technologies, sources and processing steps, such as geospatial data and geoprocessing services, can be published as part of the Web of Data. To take full advantages of the machine-understandable format and linkages among heterogeneous data items in the Web of Data, this paper proposes to publish geospatial data provenance into the Web of Data. In particular, it analyzes how a catalogue for provenance, i.e., geospatial data provenance managed by a geospatial metadata catalog service, can be published into the Web of Data using a Linked Data approach. Consequently, queries over linked geospatial data provenance are analyzed and tested to demonstrate the benefits of the approach.**

***Index Terms*— Catalog service, geospatial data provenance, linked data, ontology, web of data.**

## I. Introduction

LINKED Data technologies have shown great promise for effectively sharing and interlinking Web resources [1]. They use the resource description framework (RDF) language and HTTP protocol to publish structured data on the Web [2]. Linked Data offers great opportunities in the geoscience domain, since conventionally isolated and heterogeneous geospatial data could be exposed as Linked Data on the Web, thus promoting the wide sharing and integration of geospatial information [3]. One typical example is the GeoNames-Linked Data project, in which geospatial features are interlinked with each other [4].

Data provenance, also called data lineage or data pedigree, records sources as well as a set of processing steps applied to sources [5]. Data provenance provides important information for users to determine the reliability of data products, and helps users to reproduce and validate the data products [6]. In the geoscience domain, geospatial data is heterogeneous and in diverse forms. Often, complex geoscientific workflows (or geoprocessing service chains in a service-oriented environment) are used in deriving geospatial data products [7]. The source data and processing steps (or geoprocessing services) are recorded to provide an informed understanding of the processing history. In a geospatial service environment, such provenance information could be shared through an open geospatial consortium (OGC) standard-compliant service, namely the catalog service for the Web (CSW) [6]. The catalog service could be further exposed as Linked Data [8].

The contribution of this paper is to publish geospatial data provenance in a catalog service into the Web of Data using the Linked-Data approach. While provenance management using geospatial catalog services is well suited to the existing geospatial service architecture, it is not easy for the Web crawlers to crawl the information, since the databases are hidden behind the Web interfaces. If a catalog for provenance, i.e., geospatial data provenance managed by a metadata catalog service, could be published as Linked Data (the so-called linked geospatial data provenance), it could: 1) enrich the provenance information with semantic annotations; 2) make the provenance discovery easily and efficiently by following the same data model and query language from the Semantic Web; 3) allow to merge itself into the linked-open data (LOD) cloud [9] by creating linkages to a third-party information; and 4) realize machine-accessible applications aware of the Web of data. This paper uses RDF, a fundamental standard of the semantic Web, to store the linked geospatial data provenance. Users can discover data provenance information through queries following the SPARQL query language for RDF [10] standard. The discovery takes advantages of linkages among the provenance, data/service metadata, and even third-party information.

The remainder of the paper is organized as follows. Section II provides related work. A geospatial data provenance ontology for publishing linked geospatial data provenance is described in Section III. Section IV describes how to publish geospatial data provenance as Linked Data. Section V analyzes queries on linked geospatial data provenance. Conclusion and pointers to future work are given in Section VI.

## II. Related Work

One of the earliest investigation in geospatial data provenance was conducted by Lanter *et al*. [11]. They describe a lineage system to record lineage of GIS layers. Yue *et al*. [6] describe a service-oriented approach for sharing geospatial data provenance in the Web environment. The approach extends the ebXML registry information model (ebRIM) [12] of a geospatial catalog service to store and query geospatial

Manuscript received September 29, 2012; revised January 9, 2013; accepted February 14, 2013. This work was supported by the National Basic Research Program of China under Grant 2011CB707105, the National Natural Science Foundation of China under Project 41271397 and Project 41023001, and LIESMARS (Wuhan University) Special Research Funding.

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (email: pyue@whu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2249523

data provenance. The provenance registration and discovery are compliant with OGC standards The International Organization for Standardization (ISO) 19115 Geographic Information Metadata standard defines a lineage element to document provenance [13].

Semantic Web technologies have shown promise for an informed understanding and enhanced discovery of provenance. In the $^{my}$ Grid project, the ontology was used to annotate data provenance in scientific workflows [14]. Yue *et al*. [15] propose provenance tracking in geoprocessing service chains. Semantic metadata is generated and propagated throughout the service chains [16]. They can be used to augment the provenance. In the general information domain, some well-known provenance models have been proposed, including the open provenance model, provenir ontology, and provenance vocabulary. A comparison among them has been conducted by W3C provenance incubator group [17]. The Provenir ontology focuses on the e-Science and workflow domain. It fits the context of the geoprocessing workflows in this paper, and thus could be extended in the geospatial domain. The W3C Provenance Working Group is finalizing W3C PROV Model [18]. Although the work currently adopts the Provenir ontology, it can be migrated to W3C PROV Model with less effort, since W3C PROV Model is intended to be a generic provenance model that can accommodate different application contexts of provenance.

The LOD Cloud includes 295 data sets, which consist of over 31 billion RDF triples and cover different domains like media, biology, and geography [9]. In the geospatial domain, GeoNames is the first case to be published as Linked Data, which provides geographical entities as Linked Data for users and can be linked to other Linked Data sets. Linking GeoNames, Revyu and Flickr together, a DBpedia Mobile could provide user locations for linked information [19]. In the Linked GeoData project, OpenStreetMap data has been published according to Linked Data principles [1], and is linked with DBpedia, GeoNames, and geospatial data sets from Food and Agriculture Organization of the United Nations [20]. Another linked data set, named GeoLinkedData publishes Spanish hydrographical data as Linked Data [21]. CSW2LD toolkit is a tool developed to expose metadata from OGC CSW as Linked Data [8]. Once provenance is registered into CSW as a special kind of metadata, it is possible, as proposed in this paper, to publish it as Linked Data to facilitate linkages among source data, processing services, and third-party information. Carsten *et al*. [22] propose a method to construct links between linked sensor data and the LOD Cloud. There is also some work on publishing geography markup language (GML) data as Linked Data [23]. In addition, Lynn *et al*. [24] propose an approach on how to publish raster data as Linked Data. They extract geospatial information from raster data, and then transform it into RDF. Harshal *et al*. [25] investigate provenance in linked sensor data by publishing sensors and observations in the Sensor Web environment. Our work focuses on the provenance in geospatial domain and its publishing from the metadata catalog.

Some work in the general information domain has tried to link provenance with the LOD Cloud [26]. In this paper, data provenance is enriched with domain-specific annotations from the LOD Cloud, thus enhancing the discovery capability. This paper proposes a semantic, domain-aware provenance framework, supported by the Janus Ontology, for answering typical user questions. Based on that work, Zhao *et al*. [27] discuss how to consume linked provenance data by taking advantages of linkages between provenance and the LOD Cloud. These approaches, although applied in bioinformatic applications, are intended to be general, and could be used in the geospatial domain by enriching it with geospatial-specific semantics.

## III. Geospatial Data Provenance Ontology

There are already some provenance ontologies in the general information domain. As mentioned in Section II, the Provenir ontology [28] is extended in geospatial domain to illustrate the approach.

### A. Provenir Ontology

Provenir ontology serves as an upper_level ontology in the geospatial data provenance ontology. The Provenir ontology defines three fundamental classes to represent the major content of data provenance, i.e., data, agent, and process. The data class has five subclasses: data_collection, parameter, temporal_parameter, spatial parameter, and domain_parameter. Data represent material, product, and so on. Process represents activities_changing data. Agent affects and controls process There are eleven basic relations developed in the Provenir ontology; for example, derives_from links data and represents ancestry relationships between data has_participant links data and process. It records data participating in a process. Icons with white colors in Fig. 1 describe the structure of Provenir ontology. The extensions to Provenir ontology are introduced in Section B.

### B. Extensions to Provenir Ontology

Extensions to Provenir ontology are designed to allow domain specific queries and support the Linked Data-enriched discovery. In particular, the work in this paper focuses on the provenance in geoprocessing service chains. The following extensions are made (Fig. 1). It is noted that these extensions are for demonstration only and not intended to be complete. More comprehensive domain-specific elements can be added once they are needed by applications.

1) Service class describes metadata information of a service, which follows the Dublin Core Metadata Initiative (DCMI) Metadata model [29]. It has the following properties: dc:title that labels the service name; dc:identifier that records a service's ID; dc:abstract that describes the function of a service; rdf:type that records whether a service is an atomic service or a service chain; dc:creator that links a service to the organization that offers the service; dc:version that records the version of a service; and dc:relate that provides the URL of a service. The property has_type is used to annotate services and datasets using ontological concepts

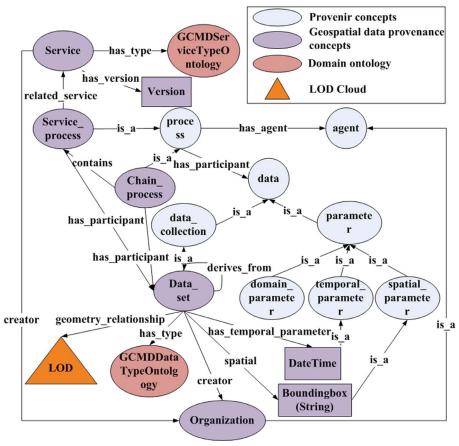


Fig. 1. Geospatial data provenance ontology.

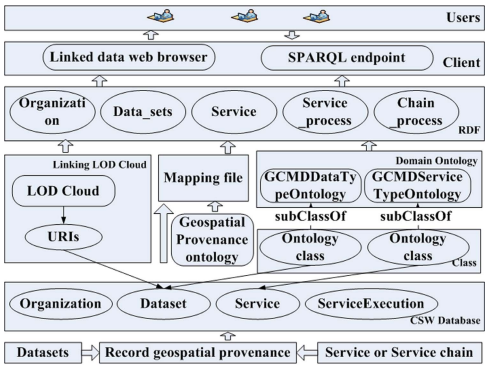


Fig. 2. Architecture of publishing linked geospatial data provenance.

from geospatial domain ontologies such as geospatial data types (GCMDDataTypeOntology) and service types (e.g., GCMDServiceTypeOntology)[1] [30].

2) Service_process class extends the class process. It records the execution information of a Service. The property-related_service links Service and Service_process. The Service_process is linked to Data_set using the has_participant property.
3) Chain_process class extends the class process and describes an execution of a service chain. The property contains links a Chain_process to its component Service_process. The Chain_process is also linked to Data_set using the has_participant property.
4) Organization class inherits from agent. It is defined for describing groups or persons who publish or create entities such as Data_set, Service, Chain_process, and Service_process. If the Service and Organization are linked by dc:creator, this means that the organization publishes the service. If the Service_process and Organization are linked by has_agent, this means that the organization executes the service.
5) Data_set class is defined to represent geospatial datasets. It inherits from data_collection, and includes some properties from DCMI. In particular, the following two properties are defined: dc:spatial and geometry_relationship. The property dc:spatial records the boundingbox of a dataset. The property geometry_relationship is a top property for recording topological relations between the data set and data in the LOD. Specific relations, such as within, contains, intersects, and overlaps, are defined as sub properties of the geometry_relationship. More details on the role of this link are presented in Section IV. The property derives_from links different datasets that have ancestry relationships. A Data_set can be annotated using ontological concepts from the GCMDDataTypeOntology through the property has_type.
6) DateTime class extends the class temporal_parameter. When it is linked by the property has_temporal_parameter, it records the production time of a Data_set.

[1]It is noted that Global Change Master Directory (GCMD) keywords sets are also available in the SKOS (Simple Knowledge Organization System)/RDF format. Here we use the OWL format, which by its nature supports ontology representation.

## IV. Publishing Geospatial Data Provenance As Linked Data

Provenance could be registered in a geospatial catalog service as a special kind of metadata [6]. The Linked Data approach explored in this paper tries to publish geospatial data provenance in a metadata catalog as Linked Data, so that metadata and provenance could be part of the Web of Data and enjoy linkages in the LOD Cloud. The following will introduce how provenance in a CSW-ebRIM profile could be published and linked to the LOD Cloud. We first introduce an architecture and describe a general process in Section A. Then some key issues in the process are highlighted in Sections B, C, and D.

### *A. Architecture to Publish Linked Geospatial Data Provenance*

Fig. 2 shows an architecture to publish linked geospatial data provenance. When geospatial data provenance is collected by third parties, it is registered in CSW and stored in CSW databases. Domain semantics are registered in CSW by adding extensions to the ebRIM model. A mapping file will be defined to convert a database to RDF, which then is published as Linked Data by D2R-server [31]. The D2R-server converts records in the CSW database into RDF using the mapping file, and then can publish RDF as Linked Data. The triples of linkage between linked geospatial data provenance and datasets from LOD are added to RDF stores. In the front end, two access modes are provided for users: Linked Data Web browser and SPARQL endpoint.

### *B. Mapping Provenance Records in ebRIM to RDF*

An existing ebRIM based provenance registration model [6] is adopted here, which is then extended and mapped to the geospatial data provenance ontology presented in Section III. Such a mapping is implemented using the D2R-server.

The existing ebRIMbased provenance registration model is extended by adding slots to record semantic annotations. classification slots are added to Dataset and Service to record semantic annotations using ontological entities from GCMDDataTypeOntology and GCMDServiceTypeOntology.

TABLE I
SNIPPET OF THE MAPPING FILE

```
map:Data_set a d2rq:ClassMap;
        d2rq:dataStorage map:Database1;
        d2rq:uriPattern "data/@@extrinsicobject.id@@";
        d2rq:condition
"extrinsicobject.objecttype='urn:x-ogc:specification:csw-ebrim:ObjectType:D
ata'";
        d2rq:condition "extrinsicobject.id=name_.parent";
        d2rq:class pro:Data_set;    .
map:DataType a d2rq:PropertyBridge;
        d2rq:property pro:has_type;
        d2rq:uriPattern
"http://geobrain.laits.gmu.edu/ontology/2004/11/gcmd-science.owl#@@slot.v
alue@@";
        d2rq:condition "slot.name_='type'";
        d2rq:condition "slot.parent=extrinsicobject.id";
        d2rq:belongsToClassMap map:Data_set;.
map:datamodel a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:Data_set;
        d2rq:property dc:format;
        d2rq:condition "slot.name_='Format'";
        d2rq:condition "slot.parent=extrinsicobject.id";
        d2rq:column "slot.value";    .
map:DataTitle a d2rq:PropertyBridge;
         d2rq:belongsToClassMap map:Data_set;
         d2rq:property dc:title;
         d2rq:condition "name_.parent=extrinsicobject.id";
         d2rq:column "name_.value";    .
…
```

Table I shows a snippet of the mapping file. For example, instances of Dataset in the ebRIM model are published as instances of the class Data_set in the geospatial data provenance ontology. The ebRIM association hasGeoDataTypeAncestor is mapped to the property derives_from in the ontology. Instances of SeviceExecution in the ebRIM model are mapped to instances of Service_process or Chain_process in the geospatial data provenance ontology. More relations about instances of Service, Service_process, Organization and Data_set are constructed from the provenance registration model in ebRIM using the mapping file. There are also a lot of attributes of Service, Data_set, Service_process, and Organization that can be extracted from the ebRIM model using the mapping file.

### C. Linking Domain Semantics

The ebRIM model could be extended to annotate datasets and services using ontological entities from domain semantics including geospatial data types and service types [32]. An ontology of geospatial data types such as GCMDDataTypeOntology conceptualizes scientific meanings of geospatial datasets, and an ontology of geospatial service types such as GCMDServiceTypeOntology is defined according to scientific problems that geospatial services focus on solving. After transforming from such an extended ebRIM model to the geospatial data provenance ontology, the Data_set and Service in the ontology can then be enriched with domain semantics from GCMDDataTypeOntology and GCMDServiceTypeOntology respectively. The domain semantics provide possibilities for enhancing provenance discovery by taking advantages of taxonomical relations in ontologies. For example, if users try to locate the provenance of NDVI (Normalized Difference Vegetation Index) products, the provenance of Landsat-Enhanced Thematic Mapper (ETM) NDVI could be retrieved, since the ontological entity ETM_NDVI is a subClassof NDVI.

### D. Constructing Links to the LOD Cloud

Using the third-party information to enrich query requests and results is a significant benefit of the Linked Data. For example, administrative divisions, once linked in the provenance, can add social information of divisions to sources. There are already many Linked Data sets relating geospatial information in the LOD Cloud, such as DBpedia, GeoNames, LinkedGeoData, and GADM-RDF [33]. The Linked Data approach allows users to link geospatial data provenance to a bunch of information in such a cloud.

The Web of Data has two basic ideas. First, to employ the RDF data model to publish structured data on the Web, and second, to set explicit RDF links to interlink data items from different data sources [34]. The first one can be implemented using the D2R-server. For the second one, both spatial and nonspatial LOD content could be linked to provenance information. In this paper, we are interested in the spatial information. Therefore, we suggest that the location information could be retrieved from the existing LOD Cloud, and thus can be linked to source data in the provenance. The links are topological relations defined in the provenance ontology, or named spatial links in the context of this paper. Although automatic generation of RDF links among large datasets is still an open issue [34], in this paper we focus on the automatic generation of spatial links between provenance and LOD data sets. The boundingbox of data items from the linked geospatial data provenance is compared with the spatial region of data items from LOD data sets. The topological relation between two geometries is determined by traditional primitive spatial operators in GIS. Once the relation is determined, the dataset will be linked to the data item using one type of geometric relations. For example, the boundingbox of a dataset "DonghuLake.tif" is within the boundary of the city "Wuhan" from GADM-RDF, then the linked provenance of the dataset will include a link to GADM-RDF using the property pro:within with the URI value http://gadm.geovocab.org/id/2_9906_geometry, which is the geometry of the city "Wuhan" in GADM-RDF in the LOD cloud. Thus third-party information like the administrative information can be linked to linked geospatial data provenance.

Some well-known linked data sets such as DBpedia and GeoNames are core parts of the current LOD cloud; however, they offer only coordinates of a central point in the spatial region of a geographic feature. GeoLinkedData publishes European hydrographical data as Linked Data, yet it does not include spatial data of China, the same as the LinkedGeodata. The linked data set from the Food and Agriculture Organization of the United Nations only provides the boundingbox information at the country level in the world. Since the computation of spatial links here focuses on topological relations between polygons, we choose the GADM-RDF. GADM is a spatial database of locations of the world's administrative areas. It provides administrative boundaries and hierarchical

TABLE II
EXAMPLE FOR SETTING A SPATIAL LINK BETWEEN TWO DATASETS (SRCDS REPRESENTS THE DATASET FROM LINKED DATA PROVENANCE, AND TGTDS REPRESENTS THE DATASET TO BE LINKED TO THE SRCDS)

| Algorithm for setting implementing "adjacent to some X" | Software and Standards |
|---|---|
| Select all data items from SrcDS<br>*e.g.*:SELECT ?x ?z WHERE { ?x <http://purl.org/dc/terms/spatial> ?z. } | SPARQL, RDF |
| Convert the format of coordinates into the "WKTLiteral". | OGC Well-known Text (WKT) |
| Select data items from TgtDS<br>*e.g.*:SELECT ?xM ?z WHERE { ?xM <http://www.opengis.net/rdf#asWKT> ?z. } | SPARQL, RDF |
| Dump WKT into geospatial database | PostGIS |
| For each data item *a* in SrcDS do<br>For each data item *b* in TgtDS do<br>If *a* has a geometric relation (*within*, *overlaps*, *contains*, and *intersects*) with *b* then<br>Add the spatial link between *a* and *b*<br>*e.g.*:select source.name, 'within', target.name from source, target where st_within (source.geometry, target.geometry). | Spatial SQL |
| Output the computation results | N-triples |

relationships among administrative divisions. GADM is now published as linked data, namely GADM-RDF. Although the work in this paper chooses the GADM-RDF for the demonstration, the algorithm in creating links can still apply to other linked geospatial data sets. All datasets, ontologies, computation software, and online demo are available at http://geopw.whu.edu.cn:8099/provb/about.htm.

Table II shows an informal representation of an algorithm containing the spatial analysis steps in creating spatial links. The right column shows the software and standards used. SPARQL queries are used to get identifiers of data items and their coordinates in both data sets. They are dumped into the spatial database of PostGIS. This allows us to use the spatial index of the spatial database when invoking primitive spatial operators such as st_within in PostGIS. Since the computation focuses on topological relations between polygons, the following relations are considered: within, overlaps, contains, and intersects. The cross relation is not applicable. The equal and touch are not practical since the source geometry is an envelope, while the target geometry is a polygon with an irregular shape. The disjoint relation is the inverse of intersects. If the links on disjoint were added, the number of such links could be very large, which have little values in practices.

## V. QUERYING LINKED GEOSPATIAL DATA PROVENANCE

### A. Example of the Water Coverage Analysis

The paper uses a geoscientific workflow analyzing the coverage area of water as a running example for provenance queries. The workflow is implemented by a geoprocessing service chain in a service-oriented environment. The service chain includes three geoprocessing services.

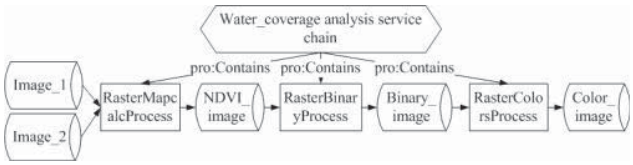


Fig. 3. Service chain for the water_coverage analysis.

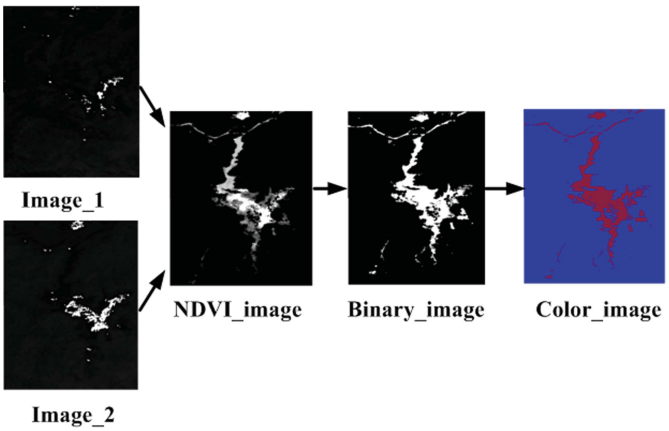


Fig. 4. Example results generated by the water_coverage analysis.

1) *RasterMapcalcProcess:* It calculates NDVI using the near-infrared (NIR) and red bands of MODIS data.
2) *RasterBinaryProcess:* It extracts the water coverage through binarization of the output in the first service.
3) *RasterColorsProcess:* It assigns colors to the output of NDVI binarization.

Through analysis of the same lake using input data at different times, the changes of water coverage areas can be detected. Fig. 3 illustrates the provenance information of data products generated by the service chain of the water_coverage analysis. Fig. 4 shows example images that are processed by the service chain. Users can ask various questions related to the provenance, which can be answered by the linked geospatial data provenance. The following three examples are provided to show how linked geospatial data provenance can answer users' queries.

1) Find the provenance of Hydrosphere data that is within Hubei, China.
2) List all final data that is processed by the water_coverage analysis service chain and published between 2010 and 2011.
3) List all final data that are processed by the RasterColorsProcess.

### B. Analysis of the Query on the First Question

Provenance query can be regarded as a transitive closure operation. Such a transitive closure operation can be implemented using a recursive traversal of the RDF graph based on (Data_set, derives_from) class-property pair of instances. The SPARQL1.1 defines transitive closure operations using the symbol "+" [10]. In the query, the symbol

TABLE III
SPARQL QUERY OF THE FIRST QUESTION

```
prefix dc: <http://purl.org/dc/terms/>
prefix gcmdsc:
<http://geobrain.laits.gmu.edu/ontology/2004/11/gcmd-science.owl#>
prefix pro: <http://geopw.whu.edu.cn:8099/provb/datas/provenance.owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix gadm: <http://gadm.geovocab.org/ontology#>
SELECT  distinct  ?value ?t ?x ?e  ?y ?st ?sp
WHERE { ?ty rdfs:subClassOf+ gcmdsc:Hydrosphere.
        ?x pro:has_type ?ty.
        ?x  pro:within ?z.
        ?p <http://geovocab.org/geometrygeometry> ?z.
        ?p  gadm:name "Hubei".
        ?x dc:title ?e.
        ?x dc:identifier ?m.
        ?x dc:abstract ?de.
        ?x pro:derives_from+  ?value.
        ?value dc:title ?t.
        ?value dc:identifier ?i.
        ?value ^pro:derives_from ?output.
        ?y pro:has_participant ?value.
        ?y pro:related_service ?res.
        ?y dc:title ?st.
        ?value dc:spatial ?sp. }
```

TABLE IV
SPARQL QUERY OF THE SECOND QUESTION

```
prefix dc: <http://purl.org/dc/terms/>
prefix pro: <http://geopw.whu.edu.cn:8099/provb/datas/provenance.owl#>
SELECT  distinct ?z ?sp
WHERE {?z dc:date ?Date.
        FILTER(?Date >= "2010-01-01")
        FILTER(?Date <= "2012-01-01")
        ?z pro:derives_from{3} ?d.
        ?x pro:has_participant ?d.
        ?x dc:title "water_coverage_analysis_Process".
        ?z dc:title ?m .
        ?z dc:spatial ?sp. }
```

Data_set is used together with the property derives_from+ (Table III), which can locate all instances of Data_set linked by the recursive traversal of the property derives_from. Then all instances of Service_process and agent classes linked to those Data_set instance are included in the query result.

The concept "Hydrosphere" in the first question is a domain concept, which can be located in the GCMDDataTypeOntology. Through querying GCMDDataTypeOntology, we can obtain all subclasses of Hydrosphere. Those instances of the Data_set, whose property has_type has a value from either Hydrosphere or its subclasses, can be returned. Thus, domain semantics can support the provenance query. Since linked geospatial data provenance is connected to GADM-RDF, queries that span across multiple linked data sets could be formulated. Table III lists an example SPARQL query to answer the first question in our SPARQL endpoint. It uses query filters pro:within to return provenance of the data that is within Hubei.

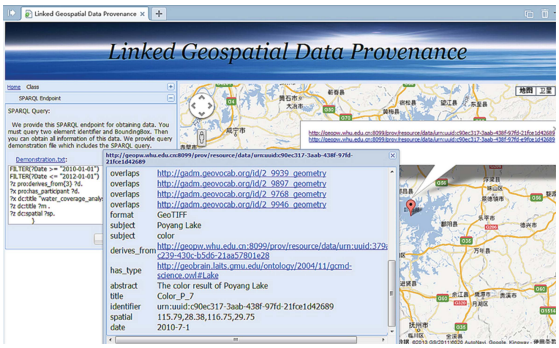


Fig. 5. Result of the second query.

TABLE V
SPARQL QUERY OF THE THIRD QUESTION

```
prefix dc: <http://purl.org/dc/terms/>
prefix pro: <http://geopw.whu.edu.cn:8099/provb/datas/provenance.owl#>
SELECT  ?z ?sp
WHERE {
        ?x pro:has_participant ?d.
        ?x dc:title "RasterColorsProcessStep".
        ?z pro:derives_from+ ?d.
        FILTER EXISTS{
        ?y  dc:title "water_coverage_analysis_Process".
        ?y pro:has_participant ?e.
        ?z pro:derives_from{3} ?e.          }
        ?z dc:spatial ?sp.          }
```

TABLE VI
SPARQL QUERY FOR EVALUATION

```
prefix dc: <http://purl.org/dc/terms/>
prefix gcmdsc:
<http://geobrain.laits.gmu.edu/ontology/2004/11/gcmd-science.owl#>
prefix pro: <http://geopw.whu.edu.cn:8099/provb/datas/provenance.owl#>
prefix gadm: <http://gadm.geovocab.org/ontology#>
SELECT ?x ?m ?de ?value ?y ?st ?sp ?date ?typ ?su ?for ?inter ?over
WHERE {     ?x dc:title "Lake_D" .
            ?x dc:identifier ?m.
            ?x dc:abstract ?de.
            ?x pro:derives_from+  ?value.
            ?y pro:has_participant ?x.
            ?y pro:related_service ?res.
            ?y dc:title ?st.
            ?x dc:spatial ?sp.
            ?x dc:date ?date.
            ?x pro:has_type ?typ.
            ?x dc:subject ?su.
            ?x dc:format ?for.
            OPTIONAL {?x pro:within ?inter.}
            OPTIONAL {?x pro:overlaps ?over.}
            OPTIONAL {?x pro:contains ?over.}       }
```

### C. Analysis of the Queries on the Second Question and Third Question

While the first query is used to obtain the provenance information of a specific dataset, the second and third queries (Tables IV and V) are used to acquire datasets that have specific provenance. The query in Table IV specifies the processing chain that derives the data. In Table IV, $z$ is the final data that is processed by the water_coverage analysis service chain, $x$ represents the instance of Chain_process, and

*d* is the input of Chain_process. Using the query, we could locate data sets that were derived by the service chain. As shown in Fig. 5, two are located in Poyang Lake, China, in July, 2010 and August, 2010 respectively. The dialog shows provenance information of one dataset located in Poyang Lake, China, in July, 2010.

The third question is to obtain datasets that are processed by a specific service component in a service chain. Table V shows the SPARQL query to answer the question.

The query performance is not the concern of this paper, since it takes advantage of the third-party developed tools, i.e., TDB and Fruseki [35], for queries over a large number of RDF triples. For example, the size of GADM-RDF is 1.7 GB. All RDF triples from the D2R-server, GADM-RDF, and spatial links are stored in the TDB. There are totally eight hundred thousand triples in the TDB. The running times of four queries (Tables III, IV, V, and VI) are compared. The fourth query combines multiple spatial links (within, contains, and overlaps) to find linked data items (Table VI). The experiments ran on a Linux server with 1600 MHz Intel Xeon processor, 512 MB of RAM, and a Redhat operating system. Each query was executed twenty times. The average running times of four queries are 113, 91, 88.5, and 107 m respectively. Since the first and fourth queries have more filters, they took more time. The results show that the performance is acceptable.

## VI. Conclusion

This paper defined a geospatial data provenance ontology based on the Provenir ontology, published geospatial data provenance as Linked Data, and analyzed queries of linked geospatial data provenance. The geospatial data provenance ontology facilitated the interoperable sharing and enhanced discovery of provenance information. By publishing geospatial data provenance as Linked Data, users can browse geospatial data provenance on the Web, and traverse different parts of geospatial data provenance through URI links. By augmenting provenance with domain semantics using GCMDDataTypeOntology and GCMDServiceTypeOntology and constructing links to LOD Cloud using GADM-RDF, this paper provided capabilities to support complex queries based on domain knowledge and enrich results of provenance queries with third-party information. The results demonstrated how geospatial data provenance can be migrated from geospatial services into Linked Data and become a part of Web of Data.

Although the work in this paper demonstrated the applicability of Linked Data approach to the geospatial data provenance, there are still some issues that need further investigations. The first issue is how to achieve geospatial reasoning based on the linked geospatial data provenance. Some challenges on achieving the qualitative spatial reasoning are identified [36]. OGC proposes GeoSPARQL as a spatial extension to SPARQL to support the spatial reasoning and query. Previous research can provide references on integrating geospatial reasoning into the linked geospatial data provenance.

Another issue is to construct more links with the LOD Cloud. The LOD Cloud includes numerous data sets about geospatial information, which can be used as third-party information to enrich geospatial data provenance and support complex queries. In future work, we will investigate how to provide more links with the LOD Cloud.

## References


[1] T. Berners-Lee. (2006, Jul.). *Linked Data* [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[2] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked Data on the Web(LDOW2008)," in *Proc. WWW* 2008, pp. 1265–1266.

[3] S. Cox and S. Schade, "Linked data: What does it offer earth sciences?" in *Proc. EGU General Assembly*, Vienna, Austria, 2010, pp. 1–2079.

[4] *GeoNames*. (2009) [Online]. Available: http://www.geonames.org/about.html

[5] A. Woodruff and M. Stonebraker, "Supporting fine-grained data lineage in a database visualization environment," in *Proc. 13th Int. Conf. Data Eng.*, 1997, pp. 7–11.

[6] P. Yue, Y. Wei, L. Di, L. He, J. Gong, and L. Zhang, "Sharing geospatial provenance in a service-oriented environment," *Comput., Environ. Urban Syst.*, vol. 35, no. 2, pp. 333–343, 2011.

[7] P. Yue, J. Gong, L. Di, J. Yuan, L. Sun, Z. Sun, and Q. Wang, "GeoPW: Laying blocks for the geospatial processing web," *Trans. GIS*, vol. 14, no. 6, pp. 755–772, 2010.

[8] F. López-Pellicer, A. J. Florczyk, J. Nogueras-Iso, P. R. Muro-Medrano, and F. J. Zarazaga-Soria, "Exposing CSW catalogues as linked data," in *Proc. 13th AGILE Conf.*, 2010, pp. 183–200.

[9] *Linking Open Data*. (2006) [Online]. Available: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[10] S. Harris and A. Seaborne. (2012). *SPARQL 1.1 Query Language. W3C Recommendation* [Online]. Available: http://www.w3.org/TR/sparql11-query/

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

[12] R. Martell, "CSW-ebRIM registry service, part 1: ebRIM profile of CSW," Open Geospatial Consortium, Inc., Redlands, CA, USA, Tech. Rep. OGC 07-110r2, 2008.

[13] *Geographic Information—Metadata*, BS ISO Standard 19115:2003, 2003.

[14] J. Zhao, C. Wroe, C. Goble, R. Stevens, D. Quan, and M. Greenwood, "Using semantic web technologies for representing e-science provenance," in *Proc. 3rd Int. Semantic Web Conf.*, 2004, pp. 92–106.

[15] P. Yue, J. Gong, and L. Di, "Augmenting geospatial data provenance through metadata tracking in geospatial service chaining," *Comput. Geosci.*, vol. 36, no. 3, pp. 270–281, 2010.

[16] P. Yue, J. Gong, L. Di, and L. He, "Automatic geospatial metadata generation for Earth science virtual data products," *GeoInformatica*, vol. 16, no. 1, pp. 1–29, 2012.

[17] S. Sahoo, P. Groth, and O. Hartig. (2010, Aug.). *Provenance Vocabulary Mappings* [Online]. Available: http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

[18] L. Moreau and P. Missier. (2012, Dec.). *PROV-DM: The PROV Data Model, W3C Working Draft* [Online]. Available: http://www.w3.org/TR/2012/CR-prov-dm-20121211/

[19] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *Proc. 6th Int. Semantic Web Conf.*, 2007, pp. 11–15.

[20] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "LinkedGeoData: A core for a web of spatial open data," *Semantic Web*, vol. 3, no. 4, pp. 333–354, 2012.

[21] L. M. Vilches-Blázquez, B. Villazón-Terrazas, A. De León, F. Priyatna, and Ó. Corcho, "An approach to publish spatial data on the web: The geolinked data case," in *Proc. Workshop Linked SpatioTemporal Data 2010 Conjunct. 6th Int. Conf. Geograph. Inf. Sci., GISci.* 2010, pp. 1–9.

[22] C. Keßler and K. Janowicz, "Linking sensor data—Why, to what, and how?" in *Proc. 3rd Int. Workshop Semantic Sensor Netw. Conjunct. 9th Int. Semantic Web Conf.*, 2010, pp. 7–11.

[23] S. Tschirner, A. Scherp, and S. Staab, "Semantic access to INSPIRE: How to publish and query advanced GML data," in *Proc. Workshop Conjunct. 10th Int. Semantic Web Conf.*, Bonn, Germany, Oct. 2011, pp. 1–13.

[24] E. L. Usery and D. Varanka, "Design and development of linked data from the national map," *Semantic Web*, vol. 3, no. 4, pp. 371–384, 2012.

[25] P. Patni, S. Sahoo, C. Henson, and A. Sheth, "Provenance aware linked sensor data," in *Proc. 2nd Workshop Trust Privacy Social Semantic Web*, 2010, pp. 1–12.

[26] P. Missier, S. S. Sahoo, J. Zhao, C. Goble, and A. Sheth, "Janus: From workflows to semantic provenance and linked open data," in *Proc. 3rd Int. Provenance Annotat. Workshop*, Jun. 2010, pp. 1–13.
[27] J. Zhao, S. S. Sahoo, P. Missier, A. Sheth, and C. Goble, "Extending semantic provenance into the web of data," *IEEE Internet Comput.*, vol. 15, no. 1, pp. 40–48, Jan.–Feb. 2011.
[28] S. S. Sahoo and A. Sheth, "Provenir ontology: Toward a framework for escience provenance management," in *Proc. Microsoft eSci. Workshop*, 2009, pp. 15–17.
[29] A. Powell, M. Nilsson, A. Naeve, P. Johnston, and T. Baker. (2007). *DCMI Abstract Model* [Online]. Available: http://dublincore.org/documents/abstract-model/
[30] P. Yue, L. Di, W. Yang, G. Yu, and P. Zhao, "Semantics-based automatic composition of geospatial Web services chains," *Comput. Geosci.*, vol. 33, no. 5, pp. 649–665, 2007.
[31] *D2RQ Platform*. (2006) [Online]. Available: http://d2rq.org/
[32] 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*, vol. 15, no. 2, pp. 273–303, 2011.
[33] *GADM-RDF*. (2007) [Online]. Available: http://gadm.geovocab.org/
[34] T. Heath and C. Bizer, "Linked Data: Evolving the web into a global data space," in *Synthesis Lectures on the Semantic Web: Theory and Technology*. San Rafael, CA, USA: Morgan & Claypool, 2011.
[35] A. Owens, A. Seaborne, N. Gibbins, and M. Schraefel, "Clustered TDB: A clustered triple store for Jena," Dept. Electron. Comput. Sci., Univ. Southampton, Southampton, BJ, USA, Tech. Rep., 2008.
[36] M. Koubarakis, K. Kyzirakos, M. Karpathiotakis, C. Nikolaou, M. Sioutis, S. Vassos, D. Michail, T. Herekakis, C. Kontoes, and I. Papoutsis, "Challenges of qualitative spatial reasoning in linked geospatial data," in *Proc. Workshop Benchmark Appl. Spatial Reason.*, 2011, pp. 33–38.

**Jie Yuan** is currently pursuing the Ph.D. degree with Wuhan University, Wuhan, China.

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

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

**Mingda Zhang** is pursuing the M.S. degree with Wuhan University, Wuhan, China.