Computers, Environment and Urban Systems 35 (2011) 333–343

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

# Sharing geospatial provenance in a service-oriented environment

Peng Yue [a,*], Yaxing Wei [b], Liping Di [c], Lianlian He [d], Jianya Gong [a], Liangpei Zhang [a]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
[b] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6407, USA
[c] Center for Spatial Information Science and Systems (CSISS), George Mason University, 4400 University Drive, MS 6E1, Fairfax, VA 22030, USA
[d] Department of Mathematics, Hubei University of Education, Nanhuan Road 1, Wuhan, Hubei 430205, China

ARTICLE INFO

*Article history:*
Received 13 August 2010
Received in revised form 26 January 2011
Accepted 22 February 2011
Available online 16 March 2011

*Keywords:*
Geospatial Web Service
CSW
ebRIM
Service chaining
Data provenance
GIS

ABSTRACT

One of the earliest investigations of provenance was inspired by applications in GIS in the early 1990's. Provenance records the processing history of a data product. It provides an information context to help users determine the reliability of data products. Conventional provenance applications in GIS focus on provenance capture, representation, and usage in a stand-alone environment such as a desktop-based GIS software system. They cannot support wide sharing and open access of provenance in a distributed environment. The growth of service-oriented sharing and processing of geospatial data brings some new challenges in provenance-aware applications. One is how to share geospatial provenance in an interoperable way. This paper describes the development of provenance service for geospatial data products using the ebXML Registry Information Model (ebRIM) of a geospatial catalog service, which follows the interface specifications of the OGC Catalogue Services for the Web (CSW). This approach fits well the current service stack of the GIS domain and facilitates the management of geospatial data provenance in an open and distributed environment.

© 2011 Elsevier Ltd. All rights reserved.

## 1. Introduction

Recent advances in Web Service technologies have significantly promoted the wide sharing and integrated analysis of distributed geospatial data. The software architecture of Geographic Information Systems (GIS) has evolved from traditional stand-alone GIS to current Web-scale and service-oriented GIS (Peng & Tsou, 2003, chap 1; Alameh, 2003; Tu & Abdelguerfi, 2006). A number of interoperable GIS services have been available, most notably the Open Geospatial Consortium (OGC) standards-compliant services. These Web Service technologies allow science and engineering communities to set up infrastructure for collaborative sharing of such distributed resources as geospatial data and processing modules, and are widely used to support the Cyberinfrastructure (Yang, Raskin, Goodchild, & Gahegan, 2010).

Foster (2005) uses the term *Service-Oriented Science* to refer to the scientific research supported by distributed networks of interoperating services, and points out that provenance is important for quality control of data products derived in a service-oriented environment. Geoscience applications always involve diverse sources of geospatial data, which are highly multidisciplinary, complex, and heterogeneous. In geoscientific problem solving, multiple geoprocessing steps are usually needed in deriving useful data products from these data. Provenance, also called lineage, records the derivation history of data products and provides an important information context to help users determine the reliability of data products. Conventional provenance applications in GIS focus on provenance capture, representation, and usage in a stand-alone environment (Lanter, 1991; Veregin & Lanter, 1995; Alonso & Hagen, 1997; Frew & Bose, 2001). They cannot support wide sharing and open access of provenance in a distributed environment. In a service-oriented distributed environment, the data and processing utilities are becoming available as services, since Web Service technologies can significantly reduce the data volume, computing steps, and resources required by the end-user (Di & McDonald, 1999). Managing and serving provenance information using the same service-oriented paradigm, now shows great promise and consistency with the existing service-oriented architecture.

Much work in the general information domain has contributed to methods on provenance-aware applications (Simmhan, Plale, & Gannon, 2005; Miles, Groth, Branco, & Moreau, 2007; Moreau, 2010). The well known international research activities include the Provenance Challenge workshop, the International Provenance and Annotation Workshop (IPAW), and the International Workshop on the role of Semantic Web in Provenance Management (SWPM). Existing investigations suggest that research issues in a provenance-aware application include provenance representation, capture, storage, query, visualization, and applications (Yue & He, 2009). To addressing these issues in a Service-Oriented Architecture (SOA), new challenges have emerged, including the architecture design of a provenance system, and the interoperability

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

0198-9715/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2011.02.006

issue in the service development (Groth et al., 2006; Miles et al., 2007).

The paper studies the provenance support in a geospatial service environment. Thus the conventional research issues have to be investigated in a Web Service context. In such a context, distributed geospatial data and geoprocessing functions are accessible through standardized geospatial services, and can be chained as executable workflows (or called service chains in a service-oriented environment) to produce value-added products (Li, Di, Han, Zhao, & Dadi, 2010; De Longueville, 2010; Yang & Raskin, 2009). Interoperability is a crucial issue in developing geospatial services (Foerster, Lehto, Sarjakoski, Sarjakoski, & Stoter, 2010). To ensure that the provenance support can work with existing services, it is necessary to share provenance for these derived data products in an interoperable way. Provenance is a broad research area involving provenance content, provenance management, and provenance use, which can be further addressed in 17 dimensions (Gil et al., 2010). The emphasis of this paper is on the geospatial provenance publication and discovery in the Web Service context, which is subsequently referred to as geospatial provenance service in SOA.

The contribution of this paper is the proposed approach on provenance service using available geospatial standards. The proposed approach uses the ebXML Registry Information Model (ebRIM) of a geospatial catalogue service to store provenance information and discover dependencies among geospatial data, geoprocessing services, and service executions. The service interface follows the interface specifications of the OGC Catalogue Services for the Web (CSW). The strength of the approach lies in the compliance with existing standards, taking advantage of the interoperability brought by standards and fitting the architecture of service-oriented GIS.

The remainder of the paper is organized as follows. Section 2 introduces a geospatial example to help in understanding the work. The work is compared with related work in Section 3. Section 4 describes a generic architecture for provenance service, and Section 5 presents the extensions to the ebRIM for registration of provenance. A prototypical implementation is presented in Section 6. Section 7 discusses the potential and advantages of the approach. Conclusions and pointers to future work are given in Section 8.

## 2. Service scenario for wildfire prediction

A typical application in the GIS domain is to assimilate geospatial data from ground and airborne sensors such as ground weather stations and Earth observation satellites to monitor and forecast

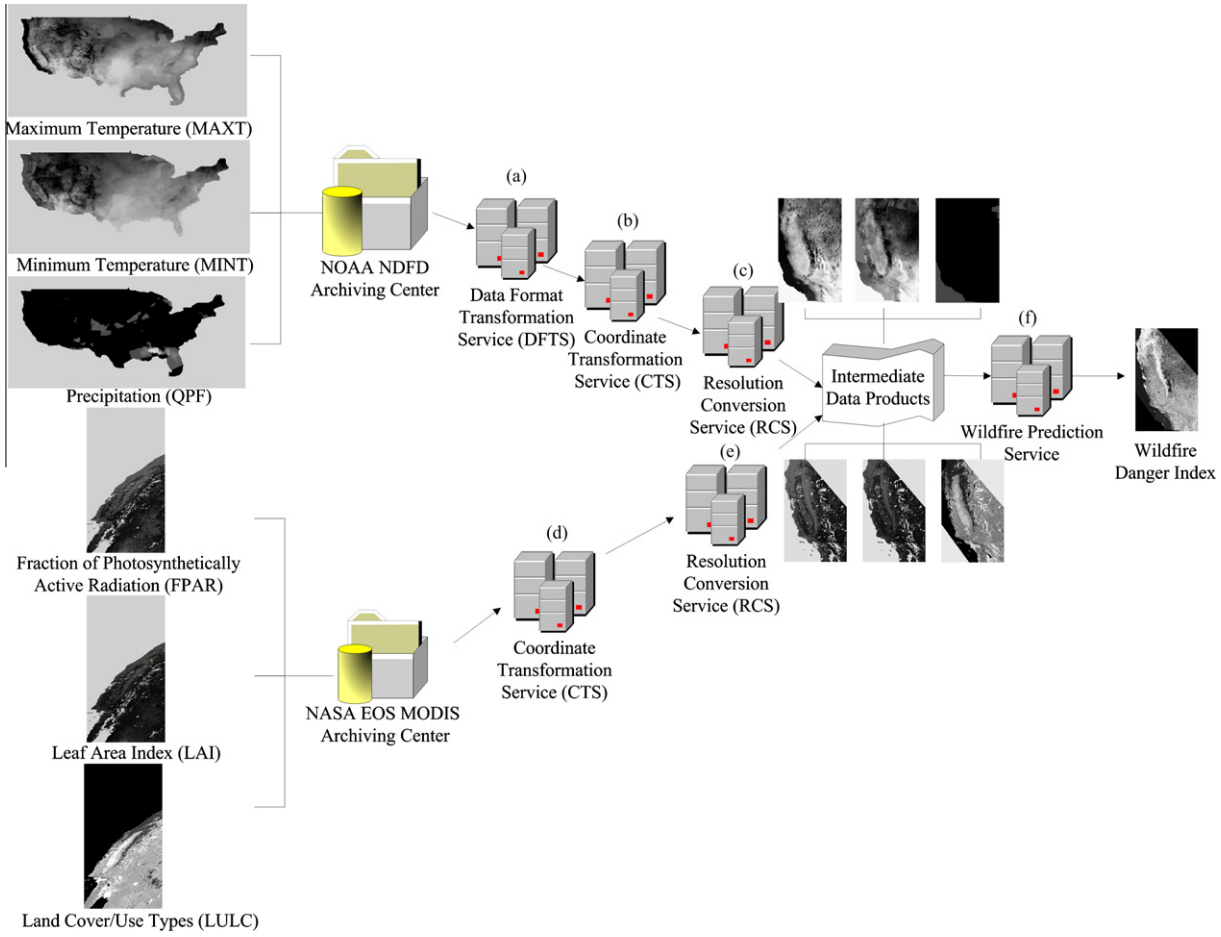


**Fig. 1.** Service chain for wildfire prediction using weather data from NOAA NDFD and MODIS data from NASA EOS. (a) A Data Format Transformation Service (DFTS) transforms the NDFD data from the GRIB2 format to the HDF-EOS format. (b) A Coordinate Transformation Service (CTS) transforms the data from the Lambert conformal coordinate reference system to the Lambert Azimuth Equal Area projection. (c) A Resolution Conversion Service (RCS) coregisters the weather data through operations of resampling/interpolation/regridding. (d) A CTS transforms the MODIS data from the sinusoidal grid coordinate reference system to the Lambert Azimuth Equal Area projection. (e) A RCS coregisters the MODIS data. (f) The wildfire prediction service that generates the wildfire danger index takes input data in HDF-EOS format, using the Lambert Azimuth Equal Area projection and 1-km spatial resolution. The wildfire prediction data product, thus, is derived from heterogeneous data and various geoprocessing services distributed on the Web.

environmental changes. Fig. 1 illustrates a service scenario involving weather data from the National Oceanic and Atmospheric Administration (NOAA) National Digital Forecast Database (NDFD) and Moderate Resolution Imaging Spectroradiometer (MODIS) data from National Aeronautics and Space Administration (NASA) Earth Observing System (EOS). The temporal and spatial coverage, resolution, format, and map projections of the data obtained from these data centers are incompatible. Several geoprocessing services, including data format conversion, coordinate system transformation, and resampling/interpolation/regridding, are chained together to transform these data into a form that can be readily accepted by the geoprocessing service that operates the prediction model. Such a scenario raises some interesting questions for users:

- How was the dataset derived?
- What are the source data and their spatial and temporal range?
- Is there an error in the source data and geoprocessing functions involved?

All these questions raise the important requirement for a provenance service to support scientific analysis – specifically, how to discover dependencies, such as ancestry or input/output relationships, among geospatial data, geoprocessing services/chains, and service/chain executions to the users' demand.

## 3. Related work

Provenance investigation in GIS can be traced back to Lanter's work in the early 1990's (Lanter, 1991). Lineage information was recorded when performing spatial analyses on vector data using commands in GIS software, and can be used to support analysis on error propagation (Veregin & Lanter, 1995). Geo-Opera, a geospatial extension to the Open Process Engine for Reliable Activities (OPERA), provided lineage support for geospatial workflows (Alonso & Hagen, 1997). Frew and Bose (2001) added lineage-tracking support for remote sensing data processing in a script-based environment. Wang, Padmanabhan, Myers, Tang, and Liu (2008) proposed a provenance-aware architecture to record the lineage of spatial data. Tilmes and Fleig (2008) discussed some general concerns of provenance tracking for Earth science data processing systems. Plale, Cao, Herath, and Sun (2010) described architectural considerations to support provenance collection and management in geosciences. The ISO 19115 Geographic Information – Metadata standard has addressed data provenance in the data quality part of the metadata. 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, e.g., discovery of dependencies among data, services, and executions using unstructured texts. Stuiver and Crompvoets (2009) propose a structured way to represent process steps and sources in the metadata standard on lineage. The OGC Web Processing Service specifies the lineage element in the request message of the Execute operation. In the OGC Sensor Web Enablement standards, Sensor Model Language (SensorML) can provide an explicit description of the process by which an observation has been obtained (i.e., observation lineage). However, how geospatial provenance information can be stored and accessed in a scalable and loosely coupled service-oriented GIS environment remains unresolved.

In the general information domain, the emergence of new information infrastructures such as e-Science or Cyberinfrastructure has caused intensive investigations of the provenance problem in the past several years (Moreau, 2010). The Provenance Challenge (Moreau, Ludäscher, Altintas, et al., 2008a) series have resulted in the proposal of the Open Provenance Model (OPM) (Moreau et al., 2008b), which aims to provide an interoperable model of provenance representation in different provenance systems. Provenance can be captured by tracing the execution of the workflow engine (Zhao, Goble, Greenwood, Wroe, & Stevens, 2003), aggregating provenance information generated by distributed service providers as a workflow executes (Foster, Vockler, Wilde, & Zhao, 2002), or a combination of the previous two methods (Miles et al., 2007). There are already some systems that are actively used to capture the provenance. Kepler, a scientific workflow system, provides a generic provenance framework for use with scientific workflows and has been applied in the geological and biological domains (Altintas, Barney, & Jaeger-Frank, 2006; Bowers, McPhillips, Riddle, Anand, & Ludäscher, 2008). In the MyGrid project, provenance in the Taverna workflow system is captured with semantic annotations to answer provenance-related questions in the bioinformatic domain (Zhao, Goble, Stevens, & Turi, 2008; Missier, Sahoo, Zhao, Goble, & Sheth, 2010). Karma system collects provenance from a BPEL (Web Services Business Process Execution Language, shortly known as BPEL) engine and provides a provenance framework for managing provenance. The implementation is evaluated using an example in the meteorology domain (Simmhan, Plale, & Gannon, 2008). The virtual data system (VDS) (Clifford, Foster, Voeckler, Wilde, & Zhao, 2008) can collect provenance from either the shell program or Pegasus workflow system. Pegasus itself can also be extended to capture provenance (Kim, Deelman, Gil, Mehta, & Ratnakar, 2008). PASS system collects provenance when executing scripts in the UNIX shell environment (Holland, Seltzer, Braun, & Muniswamy-Reddy, 2008), and Redux system collects provenance from the Windows Workflow Foundation (WinWF) engine (Barga & Digiampietri, 2008). Our previous work proposes an approach on capturing the provenance of geospatial data before execution of service chains (Yue, Gong, & Di, 2010). The focus in this paper is on sharing provenance instead of capturing provenance, thus it assumes that provenance can be collected, either using existing systems or adding support on provenance capture to other workflow systems. The encodings of provenance collected from provenance capturing processes, using either XML or Resource Description Framework (RDF), can be transformed into the standard form that is acceptable to the CSW based on the shared understanding of the provenance model.

Provenance information can be tightly coupled with a metadata store, using an existing metadata catalogue for storage and management, such as the virtual data catalog in VDS (Clifford et al., 2008). It could also be managed using a separated storage system, or called the provenance store in the European Union funded project on provenance-aware SOA (PASOA) (PASOA, 2006). The provenance store can be implemented using a database as the PASS and Redux do (Holland et al., 2008; Barga & Digiampietri, 2008). In a Semantic Web environment, where semantics of provenance are represented as RDF triples, a RDF triple store can be employed to store provenance and support queries using the SPARQL query language (Sahoo, Sheth, & Henson, 2008; Chebotko, Lu, Fei, & Fotouhi, 2010). In the PASOA project, it is suggested that the development of a new provenance architecture should be aware of existing standards and ensure the software interoperates with that which already exists (Miles et al., 2007). Our work addresses this requirement by using existing geospatial services to store and query provenance, thus ensuring compliance with existing standards and architecture in the geospatial domain.

Some efforts have been devoted on provenance visualization, such as the Kepler provenance browser (Anand, Bowers, & Ludäscher, 2010) or VisTrails (Callahan et al., 2006). Other work focuses on the applications of provenance. Simmhan et al. (2005) have summarized several categories of applications on provenance: data quality, audit trail, replication recipes, attribution, and informational. We intend to leave the visualization and use of provenance as future work.

## 4. Geospatial provenance service

Fig. 2 shows an architecture to support the geospatial provenance service. To integrate provenance support into the geospatial service architecture and make the geospatial provenance service compliant with existing geospatial standards, we propose to use the geospatial catalogue service to publish and discover geospatial provenance information. A geospatial data server provides geospatial data, for example point observation, feature, coverage, and map. The NOAA NDFD and NASA MODIS data in the scenario are examples of coverage data. A geoprocessing server deals with both atomic processes and service chains. OGC Web Services provided by servers are self-describing and support a GetCapabilities operation that returns a capabilities document describing the service's metadata such as data collections (e.g., the dataset from NOAA NDFD and NASA MODIS) or geoprocessing processes (e.g., the resolution conversion service or wildfire prediction service in the scenario). Geography Markup Language (GML), SensorML, Metadata standard, and provenance in addition, provide information models that facilitate the information exchange between geospatial services and their clients. For example, in the wildfire prediction scenario, metadata for geospatial data follows the ISO 19115 metadata standard, and metadata for geospatial data services and geoprocessing services follows the ISO 19119 service standard. Geospatial provenance can be collected from either individual geospatial data and processing services or geoscientific workflow engines. This paper does not investigate approaches to automatic recording of geospatial provenance. It focuses on how recorded provenance can be managed and queried in the current geospatial service architecture. The provenance information is represented as a special type of metadata entity and registered in a catalogue server by extending its registration information model. Thus, geospatial provenance information can be provided by using a geospatial catalogue service.

### 4.1. Service-oriented approach

A geospatial provenance service can be thought of as a specialized database of information about provenance information available to a group or community of users. Such provenance information includes metadata descriptions of source data (e.g. NDFD or MODIS data from data archive centers), transformation functionalities (e.g. geoprocessing services), geoprocessing workflows (e.g. service chains for wildfire prediction), parameters used, intermediate geospatial data products, and date and time of execution. When integrated with catalogue services, provenance information provides clues for locating and retrieving related geospatial resources including data and services.

The publication and discovery of geospatial data provenance using Web Service technologies follows the component-based software engineering principle and can be integrated into the architec-

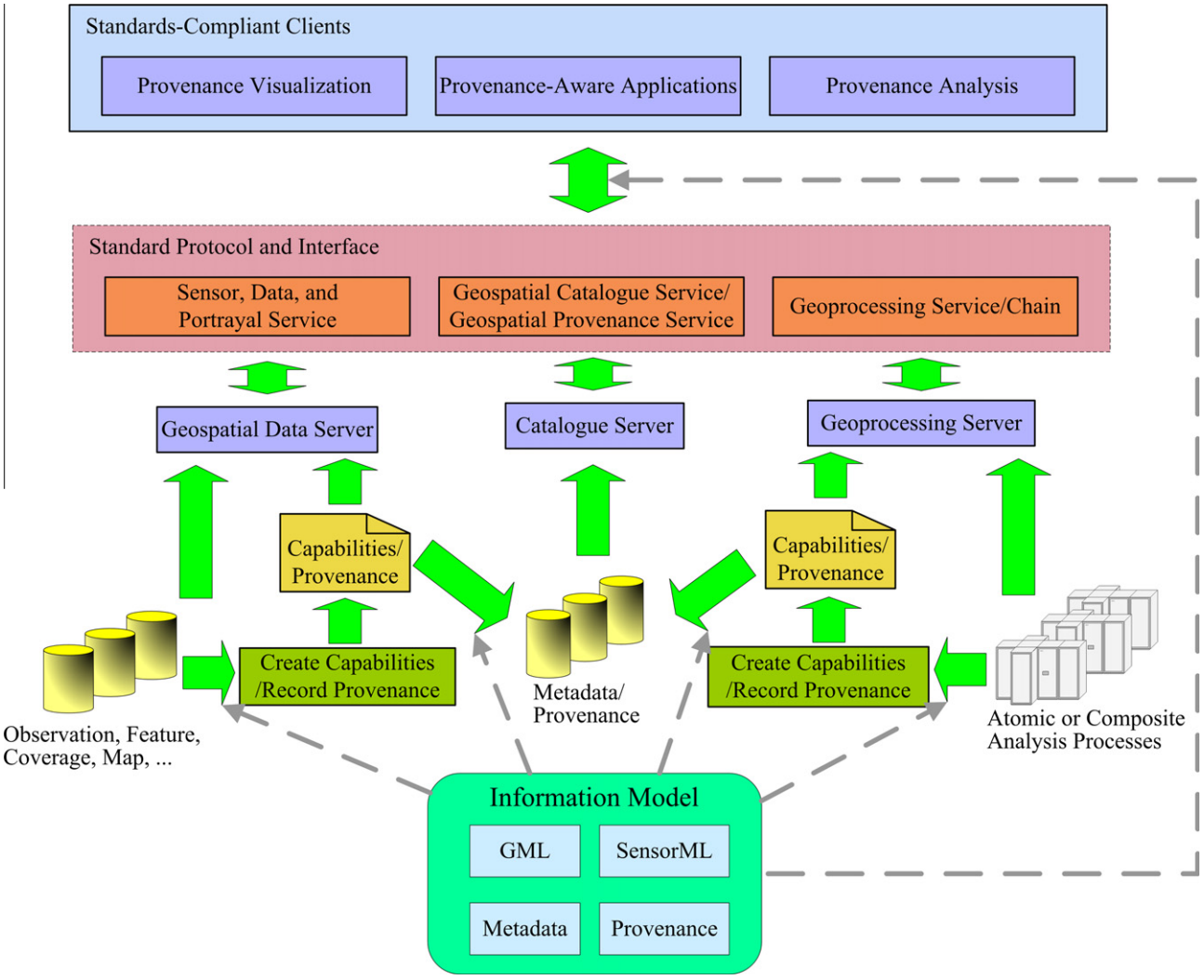


**Fig. 2.** The architecture to support geospatial provenance service. The figure shows how geospatial provenance service can be provided in the current service-oriented GIS.

ture of service-oriented GIS. Such a geospatial provenance service provides three essential functions:

- Presenting a common model for provenance information. The model defines the content, structure, and semantics of provenance information that would allow a shared understanding and exchange of provenance information.
- Arranging provenance descriptions to facilitate easy access. The fragmented provenance information for distributed data products, originally known by diverse organizations or individuals, is organized and managed in a single, searchable location and indexed by the service.
- Defining an interface for publishing and discovering provenance information. Clients interact with provenance services using this interface.

### 4.2. Compliant with existing standards

The use of geospatial catalogue services to support geospatial provenance services takes advantages of scalable Web Service and registry technologies, while at the same time is compliant with existing standards. OGC CSW is an industry consensus that defines an open, standard interface to online metadata catalogs for geospatial resources (Nebert, Whiteside, & Vretanos, 2007). According to the CSW specification, any implementation of CSW consists of two components.

- Catalogue information model: The ebRIM standard has been defined by the Organization for the Advancement of Structured Information Standards (OASIS) and selected by OGC as the information model for specifying how catalogue content is structured and interrelated. An ebRIM profile of CSW (Martell, 2008) has been developed and recommended for CSW implementation.
- HTTP protocol binding: CSW is a specification focusing on catalogue operations in the Web environment. It follows the HTTP protocol binding and can support XML encoding of the OGC Filter query language (Vretanos, 2005). Seven operations are defined in the CSW interface, among them GetCapabilities, DescribeRecord, GetRecords, and GetRecordById (Nebert et al., 2007).

To provide provenance service using CSW, provenance information is registered into the ebRIM-based catalogue information model. The ebRIM model is a general information model. It provides standard mechanisms to define and associate registered objects. Geospatial provenance information, therefore, can be registered using these standard mechanisms.

## 5. Provenance registration in the CSW-ebRIM profile

The publication and discovery of geospatial provenance uses the existing interface and protocol of geospatial catalogue service. Therefore, the key issues to develop a geospatial provenance

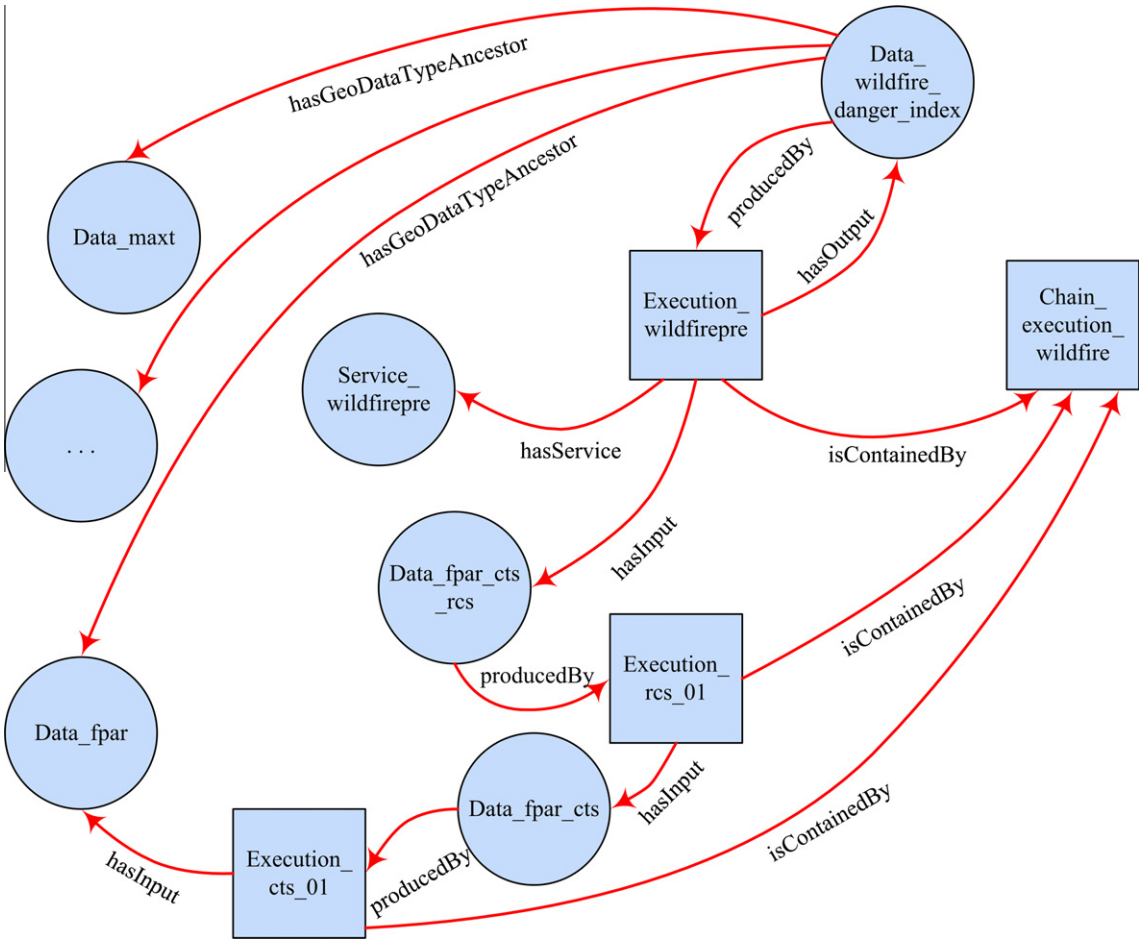


**Fig. 3.** Provenance information for the wildfire danger index. The figure follows the styles of OPM provenance graph (Moreau et al., 2008b) and illustrates dependency relations among these provenance entities. Complete provenance information is extensive. For the brevity of figure size, only part of the information is presented as an example.

service are the representation of geospatial data provenance and its registration in the catalogue service to support provenance queries.

### 5.1. Representing geospatial data provenance

When concerning about the derivation history of the data product for the wildfire prediction, geospatial users are interested in datasets used or derived, geoprocessing functions imposed on those datasets, and execution information including argument values and runtime environment. This information can be modeled as entities – geospatial data, services, and executions, and relations among them. From this perspective, we define four types of provenance entities when modeling geospatial data provenance: geospatial data products, geospatial Web Services, atomic service executions, and service chain executions. A service chain as a whole can be conceived of as a service. For example, a service chain defined using BPEL runs as a Web Service. Thus service chains are regarded as a special kind of services. Space and time are intrinsic characteristics of geospatial data products. Spatial and temporal elements in the ISO 19115 metadata standard are included as attributes for geospatial data products entity in the provenance definition. Four types of relations among these entities exist.

- Relations between geospatial data products: Such relations include connections between a geospatial data product and its ancestor geospatial data product. For example, the relation *hasGeoDataTypeAncestor* is defined to link the wildfire danger index (Data_wildfire_danger_index) to its ancestor NOAA MAXT (Data_maxt) (Fig. 3).
- Relations between geospatial data products and service executions: The relations define that the execution of a service requires or results in a geospatial data product. For example, the relations *hasInput* and *hasOutput* in Fig. 3 give the input (Data_fpar_cts_rcs) and output (Data_wildfire_danger_index) geospatial data products of an execution of the wildfire prediction service (Execution_wildfirepre). Another relation *producedBy* specifies that a geospatial data item (e.g. Data_wildfire_danger_index) was produced by executing a service (e.g. Execution_wildfirepre).
- Relations between atomic service executions and service chain executions: In a service-oriented geoprocessing workflow, an atomic service execution is triggered by a service chain execution. As Fig. 3 shows, atomic service execution (Execution_wildfirepre) is linked to the execution of the geoprocessing service chain (Chain_execution_wildfire) by the relation *isContainedBy*.
- Relations between service executions and services: A service execution must specify the service it executes. The relation *hasService* is used to link a service execution (Execution_wildfirepre) to the service descriptions (Service_wildfirepre).

These relations can be identified as causal dependencies in the OPM (Moreau et al., 2008b). Although ISO 19115 defines a lineage metadata tag, it does not cover the full content of the proposed definition.

The provenance representation is encoded in XML and registered into the CSW-ebRIM profile. Furthermore, if Semantic Web technologies are adopted, this provenance representation can also be implemented in RDF, stored in a RDF store, and queried using the SPARQL query language (Yue et al., 2010).

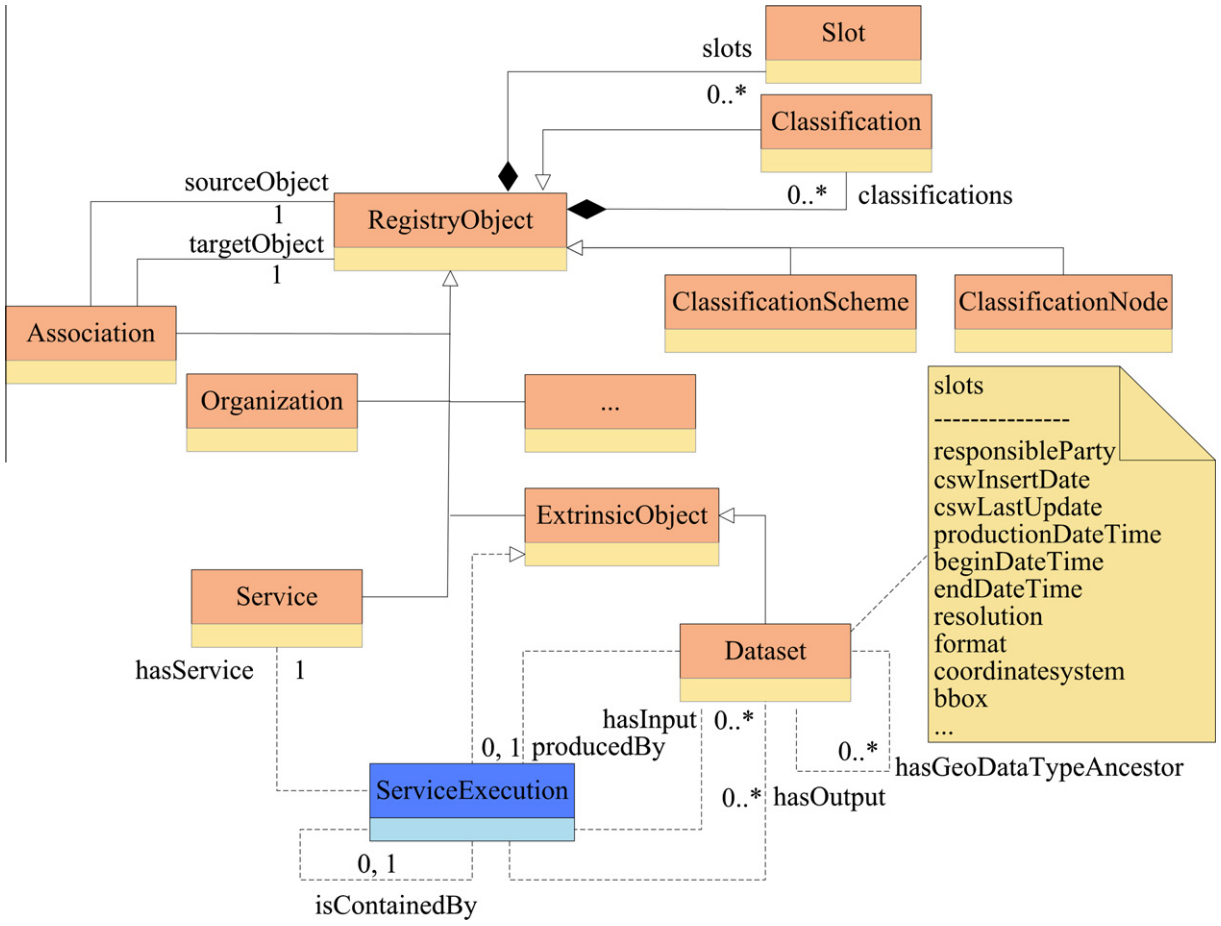


**Fig. 4.** The extended ebRIM information model. The UML style graph of this figure shows the relationships of the metadata classes defined by the model. The dashed lines show extensions for registration of provenance.

### 5.2. Extension of the ebRIM information model for registration of provenance

Extensions for registering provenance information are created in the CSW-ebRIM profile. The extensions use the extensibility features provided by ebRIM such as new classes inherited from ExtrinsicObject, new types of associations, classifications, and additional slots (Martell, 2008). The ebRIM model depicts the metadata for information resources by using a set of classes and relationships among these classes (Fig. 4). The core metadata class is the RegistryObject. Most of other metadata classes in the information model are derived from this class. An instance of RegistryObject may have a set of zero or multiple Slot instances that serve as extended attributes for this RegistryObject instance. An Association instance represents an association between a source RegistryObject and a target RegistryObject. Each association instance has an associationtype attribute that specifies the type of that association. A Classification instance classifies a RegistryObject instance by referring to a ClassificationNode instance defined within a classification scheme. A classification scheme, defined by a ClassificationScheme instance in the ebRIM model, is a tree structure made up of ClassificationNode instances that can be used to describe a taxonomy.

The following extensions are made to enable the registration of geospatial data provenance:

- Creating a new class in ebRIM for representing provenance entities. The ebRIM model has provided the Service class that supports the registration of service descriptions. A service chain as a whole can be registered as a service using this class. ExtrinsicObject provides metadata that describes submitted content whose type is not intrinsically known to the registry and therefore must be described by means of additional attributes. In the CSW-ebRIM profile, metadata for geospatial data is registered as a subclass of ExtrinsicObject – Dataset. Therefore, provenance entities including geospatial data products and services can utilize existing classes. To register service executions, one new subclass of ExtrinsicObject, ServiceExecution, is created. It can be used for registration of both atomic and service chain execution.

**Table 1**
The declaration of service execution in XML.

```
<ClassificationScheme ...>
  <ClassificationNode ...>
    ...
    <ClassificationNode xmlns = "urn:oasis:names:tc:ebxml-
  regrep:xsd:rim:3.0" xmlns:dsig = "http://www.w3.org/2000/09/
  xmldsig#" xmlns:xsi = "http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation = "urn:oasis:names:tc:ebxml-regrep:xsd:rim:3.0
  http://laits.gmu.edu:8099/csw/schema/rim-3.0.xsd"
  id = "urn:uuid:0275cd39-8ce0-45cb-b3b5-c52cde536568"
  home = "http://laits.gmu.edu:8099/csw/"
  objectType = "urn:uuid:555c406c-2850-4b34-b75f-fe936f670960"
  status = "Approved » parent = »urn:uuid:6902675f-2f18-44b8-888b-
  c91db8b96b4d" code = "ServiceExecution" path = "/ObjectType/
  RegistryObject/ExtrinsicObject/ServiceExecution">
      <Name>
        <LocalizedString xml:lang = "en-US" charset = "UTF-8"
  value = "ServiceExecution"/>
      </Name>
      <Description>
        <LocalizedString xml:lang = "en-US" charset = "UTF-8"
  value = "execution of the service"/>
      </Description>
    </ClassificationNode>
  </ClassificationNode>
  ...
</ClassificationScheme>
```

- Adding slots to declare attributes of provenance entities. Attributes such as spatial and temporal properties can be added to Dataset by defining additional slots. The execution begin-time and end-time slots are added to both ServiceExecution and Dataset classes to support the use of timestamps in validating data derivation and understanding causality. More attributes such as execution environment descriptions (operating system, software library, hardware configuration, etc.) and values of service parameters can be added to ServiceExecution using additional slots.

**Table 2**
An example of slots representation in XML.

```
< ServiceExecution id = "urn:uuid:1af249d3-2328-46a0-a1ea-ffb165867014"
   objectType = "urn:uuid:0275cd39-8ce0-45cb-b3b5-c52cde536568"
   expiration = "2012-12-17T09:30:47" ...>
  <Name>
    <LocalizedString lang = "en-US" charset = "UTF-8"
   value = "WildfirePredictionExecution "/>
  </Name>
  <Description>
    <LocalizedString lang = "en-US" charset = "UTF-8" value = "Execution of
   Wildfire Prediction Service"/>
  </Description>
   < !–execution timestamps –>
  <Slot name = "beginTime" slotType = "ServiceExecution">
     < ValueList>
       <Value>2009-12-17T09:30:47</Value>
     </ValueList>
  </Slot>...
   < !–execution environment information –>
  <Slot name = "operatingSystem" slotType = "ServiceExecution">
    <ValueList>
      <Value>Name = Linux</Value>
      <Value>Release = 2.6.23.1-42.fc8</Value>
    </ValueList>
  </Slot>
  <Slot name = "processor" slotType = "ServiceExecution">
    <ValueList>
      <Value>ClockSpeed = 3000</Value>
      <Value>InstructionSet = x86</Value>
    </ValueList>
  </Slot>
  <Slot name = "mainMemory" slotType = "ServiceExecution">
    <ValueList>
      <Value>RAMSize = 2026</Value>
      <Value>RAMAvailable = 463</Value>
    </ValueList>
  </Slot>
  <Slot name = "softwareLibrary" slotType = "ServiceExecution">
     < ValueList>
       <Value>HDF41r2</Value>
     </ValueList>
  </Slot>...
   < !–values of service parameters –>
  <Slot slotType = "inputValues" name = "ServiceExecution">
    <ValueList>

      <Value>MAXT = urn:uuid:375c8d90-d0b6-4e3b-a9da-
   e68e2bea4e7d</Value>
      <Value>MINT = urn:uuid:be236647-b48b-4ffb-bc04-5f4fef97ca39</
   Value>
      <Value>QPF = urn:uuid:cc9c7fc2-80f6-4beb-bc8d-bc555860555c</
   Value>
      <Value>FPAR = urn:uuid:16c5c071-e0f1-4213-9373-da141308e12a</
   Value>
      <Value>LAI = urn:uuid:e942b31a-dddf-492b-9661-53516c9f36a8</
   Value>
      <Value>LULC = urn:uuid:a091c3ce-6471-445c-b15a-67c7631815ee</
   Value>
      <Value>OutputFormat = application/HDF-EOS</Value>
    </ValueList>
  </Slot>...
</ ServiceExecution>
```

- Building new associations based on relations among provenance entities. Relations among provenance entities are registered using associations. As illustrated in Fig. 4, six association types, i.e. hasGeoDataTypeAncestor, hasInput, hasOutput, producedBy, isContainedBy, and hasService, are defined to associate Service, Dataset, and ServiceExecution objects. The ebRIM model provides several standard classification schemes, such as ObjectType and AssociationType as a mechanism to provide extensible types. These classification schemes are called canonical classification schemes and can be extended by adding additional classification nodes. The AssociationType classification scheme defines the types of associations between RegistryObjects. The association types are defined as classification nodes in the AssociationType classification scheme.

## 6. Implementation

The CSW-ebRIM implementation (Wei et al., 2005), developed and maintained by the Laboratory for Advanced Information Technology and Standards (LAITS) of George Mason University (GMU), is used. The GMU CSW implementation has extended ebRIM using international geographic standards: ISO 19115 (including part 2: Extensions for imagery and gridded data) and ISO 19119 Geographic Information – Services. The new Dataset class is used to describe geographic datasets. Many new attributes are added to the Dataset class based on ISO 19115 and its part 2. The Service class included in ebRIM can be used to describe geographic services, but the available attributes in the class Service are not sufficient to describe geospatial Web Services. New attributes derived from ISO 19119 are added to the Service class through Slots.

The use of GMU CSW for implementing provenance service is demonstrated as follows. The registration and query of provenance in CSW use the XML-based messages. The creation of the new class ServiceExecution is shown in Table 1. The ServiceExecution is defined as a classification node in the ObjectType classification scheme, which defines the different types of RegistryObjects a registry may support. The parent of the ClassificationNode instance representing ServiceExecution object type is a unique identifier referring to the ClassificationNode instance representing ExtrinsicObject object type. The code of the ServiceExecution contains a code that can be used in constructing the path. The path of the ServiceExecution contains the canonical path from the root ClassificationScheme. The use of slots for declaring attributes of provenance entities such as execution timestamps, execution environment information, and values of service parameters is demonstrated in Table 2. More attributes can be added according to the specific application requirements when applying provenance information to support scientific analysis. The associations among provenance entities are represented using predefined association types. Fig. 5 shows an example of the association using the association type hasService. The bottom part of Fig. 5 is an XML encoding example to illustrate this association.

The extended catalogue contents are used to formulate queries. Those extended catalogue contents are queried through the standard CSW interface. The query in Table 3 is an example. It uses the spatial and temporal filters to locate the wildfire prediction data of interest and find the service execution that produces this data using the OGC filter specification (Vretanos, 2005). The association type producedBy, as a classification node in the AssociationType classification scheme, is used. Queries that are more sophisticated, such as queries based on various relations among

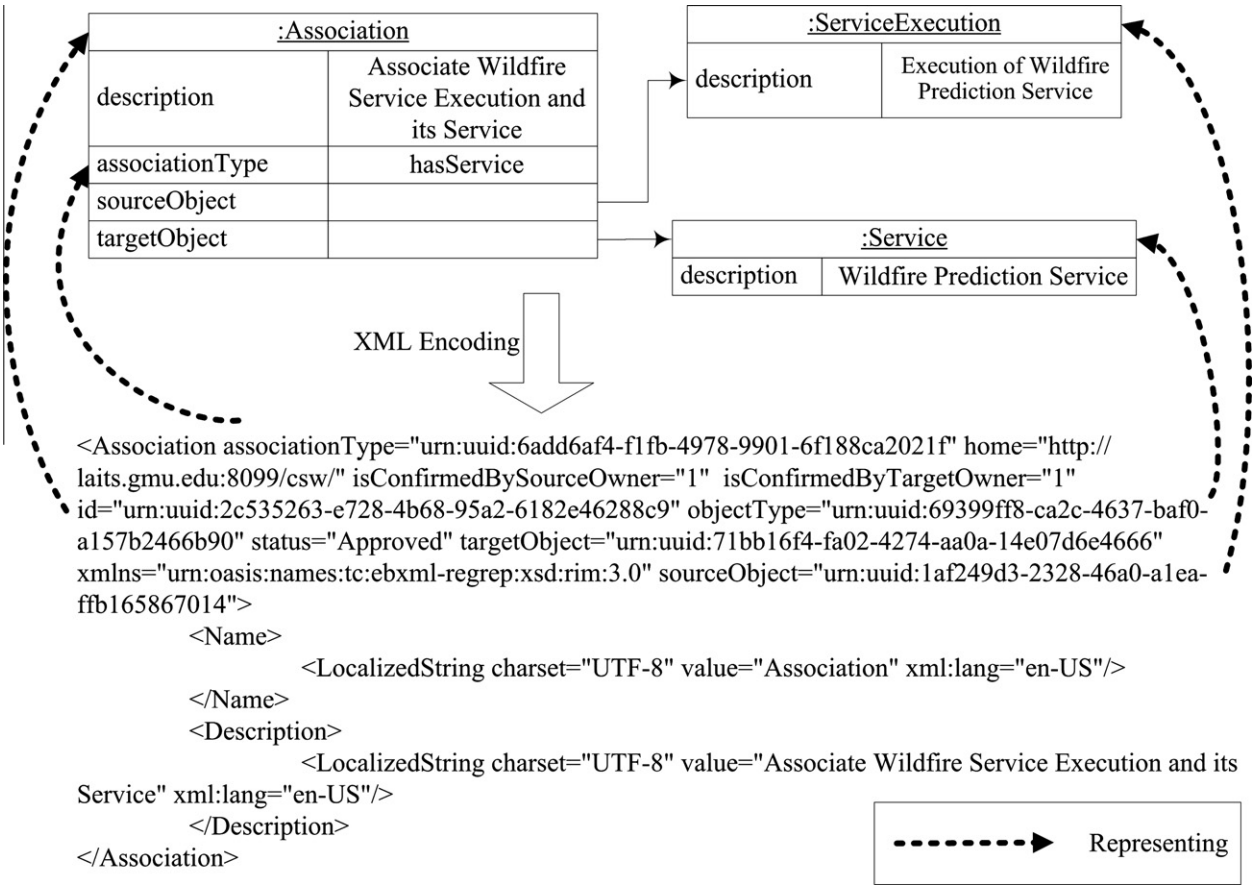


**Fig. 5.** An association between a service execution and its service.

**Table 3**
Provenance information query using the standard GetRecords operation of CSW. The query finds the service execution that creates the wildfire prediction data.

```
<?xml version = "1.0" encoding = "UTF-8"?>
<csw:GetRecords xmlns = "http://www.opengis.net/cat/csw"
  xmlns:csw = "http://www.opengis.net/cat/csw" xmlns:ogc = "http://
  www.opengis.net/ogc" xmlns:gml = "http://www.opengis.net/gml"
  version = "2.0" outputFormat = "text/xml" charset = "UTF-8"
  outputSchema = "http://www.opengis.net/cat/csw" startPosition = "1"
  maxRecords = "50">
 <csw:Query typeNames = "ServiceExecution Association Dataset
  ClassificationNode">
   < csw:ElementSetName > full</csw:ElementSetName>
  <csw:ElementName>/ServiceExecution/</csw:ElementName>
  <csw:Constraint version = "1.0.0"><ogc:Filter><ogc:And>
    < !–temporal condition –>
    < ogc:PropertyIsGreaterThanOrEqualTo>
      < ogc:PropertyName>/Dataset/beginDateTime</ogc:PropertyName>
      < ogc:Literal > 2009-08-26T12:00:00Z</ogc:Literal></
  ogc:PropertyIsGreaterThanOrEqualTo>
    < ogc:PropertyIsLessThanOrEqualTo><ogc:PropertyName>/Dataset/
  endDateTime</ogc:PropertyName>
      <ogc:Literal>2009-12-26T23:59:59Z</ogc:Literal></
  ogc:PropertyIsLessThanOrEqualTo>
    < !–spatial condition –>
   <ogc:BBOX><ogc:PropertyName>/Dataset/BBOX</ogc:PropertyName>
     <gml:Box srsName = "EPSG:4326">
       <gml:coordinates > -125.068871,31.473307-
  116.248549,43.478614</gml:coordinates>
     </gml:Box></ogc:BBOX>
    < !–keyword –>
    < ogc:PropertyIsEqualTo>
      < ogc:PropertyName>/Dataset/Name/LocalizedString/@value</
  ogc:PropertyName>
      < ogc:literal > wildfire_danger_index</ogc:literal></
  ogc:PropertyIsEqualTo>
    < !–producedBy association –>
    < ogc:PropertyIsEqualTo><ogc:PropertyName>/Dataset/@id</
  ogc:PropertyName>
      < ogc:PropertyName>/Association/@sourceObject</
  ogc:PropertyName></ogc:PropertyIsEqualTo>
    < ogc:PropertyIsEqualTo><ogc:PropertyName>/ServiceExecution/
  @id</ogc:PropertyName>
      < ogc:PropertyName>/Association/@targetObject</
  ogc:PropertyName></ogc:PropertyIsEqualTo>
    < ogc:PropertyIsEqualTo><ogc:PropertyName>/Association/
  @associationType</ogc:PropertyName>
      <ogc:PropertyName>/ClassificationNode/@id</ogc:PropertyName></
  ogc:PropertyIsEqualTo>
   <ogc:PropertyIsEqualTo><ogc:PropertyName>/ClassificationNode/
  @code</ogc:PropertyName>
      <ogc:Literal>producedBy</ogc:Literal></ogc:PropertyIsEqualTo>
 </ogc:And></ogc:Filter></csw:Constraint></csw:Query></
  csw:GetRecords>
```

provenance entities, can be established using the OGC filter specification. Fig. 6 shows the Web query interface to access the CSW service. The query defined in Table 3 can be applied to query provenance information registered in the CSW-ebRIM implementation. The query result is return in XML (Fig. 6).

## 7. Discussion

Provenance is crucial in determining the reliability of data products and can be regarded as part of metadata for data quality. In geospatial domain, lineage is defined as a type of data quality information in ISO 19115. The argument values, transformations, and base data included in the provenance information can assist users in evaluating the quality of the data based on specific quality metrics. Provenance can serve as a means to audit the trail of execution and help locate errors or exceptions. The transformation steps included in the provenance information can act as a recipe to reproduce a particular data product. The intellectual property of contributors or copyright can also be identified through provenance information, e.g. the NASA and NOAA in providing the data, and GMU LAITS in providing geoprocessing services. Often, this kind of information is already included in the metadata for geospatial data and services, and can be used in the provenance context. Interleaving provenance information and data products together, discovery and interpretation of data products can be more informational.

The use of the ebRIM information model allows provenance to be organized in a structured way and enjoys the flexibility of the ebRIM. It overcomes the weakness of unstructured data such as irregularities and ambiguities caused by using free text based provenance descriptions, thus provides support to automated processing by software. The identification of specific extensions to ebRIM permits the queries on the essential aspects of how geospatial data products were consumed or produced in various geoprocessing processes. The integration of three types of provenance information – geospatial data, services, and executions – into a unified registration information model in the catalogue, allows these provenance information entities to be connected in a coherent fashion, and support the navigation of these entities in a consistent and seamless way by exploring the linkages among the underlying ebRIM elements. Using registered associations, users can find the derivation history including the source data products and their spatial and temporal metadata, geoprocessing services, and chains. The assessment of data quality for derived dataset can use provenance by locating source data for error source identification and checking geoprocessing functions and parameter values for error propagation.

The CSW specification has been recommended by OGC and supported by many geospatial software vendors including Galdos Systems Inc., ERDAS Inc., and Environmental Systems Research Institute (ESRI) (OGC, 2011). The existing CSW implementations already support the registration of metadata for geospatial data and services and are being deployed to allow standards-compliant discovery. For example, the geospatial data products such as NASA MODIS data or geoprocessing services like coordinate transformation services could be already registered in the CSW to facilitate the discovery. A practical approach to sharing provenance is to reuse these existing implementations so that our approach can be widely employed. Existing registrations of geospatial data and services in CSW can be reused, and included as part of provenance information when taking a provenance view of catalogue entities. Work on publishing provenance focuses more on adding execution entities and making linkages among data, services, and executions. The use of CSW for sharing provenance, therefore, can work with existing implementations, follow the service-oriented architecture, and be compliant with existing standards.

It is noted that the ebRIM model itself specifies a provenance information model for RegistryObjects by defining a set of classes such as Organization and PostalAddress, which can describe parties responsible for creating, publishing, or maintaining a RegistryObject (Fuger, Najmi, & Stojanovic, 2005)[1]. Our registration extensions for data provenance focus on the registration of derivation history for data products instead of the history of RegistryObjects. Thus, the approach is tailored to the applications in the geospatial domain.

## 8. Conclusions and future work

By providing geospatial provenance service in the current OGC service architecture, the publication and discovery of geospatial provenance interoperate with Web-scale GIS software and follow the service-oriented paradigm. The adoption of OGC CSW interface

[1] It is also defined in the latest draft for ebRIM Version 4.0.

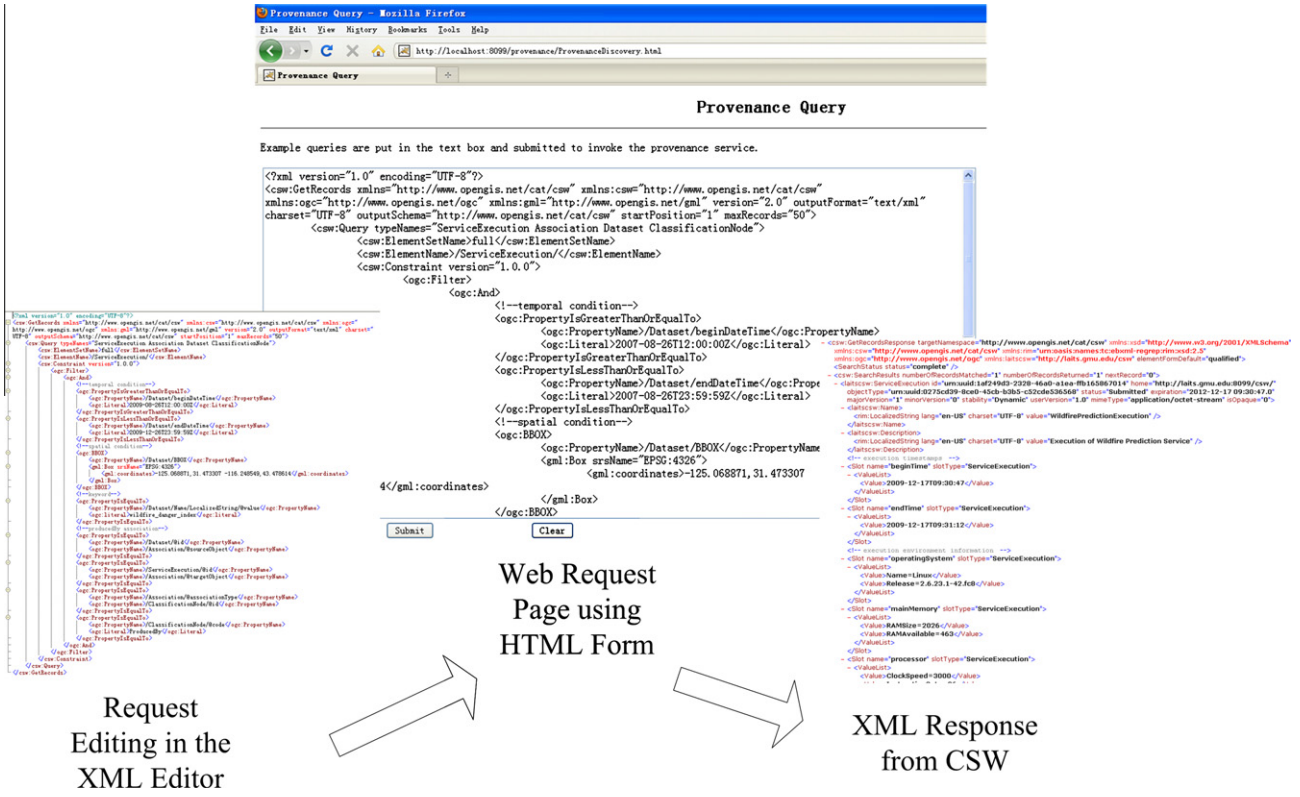


**Fig. 6.** Invocation of the CSW service through the HTML Form.

for provenance service provides promise to integrate easily with existing GIS applications. The registration of provenance using the ebRIM information model not only provides a common model to facilitate the exchange of provenance information, but also arranges provenance information in a consistent way with other catalogue information entities such as metadata for data and services. Such an arrangement facilitates the easy access to the provenance and allows free navigation among provenance, data, services, and chains.

The provenance extensions to the ebRIM increases the complexity of query formulation, since more search conditions are required in query filters when discovering dependencies among geospatial data, geoprocessing services/chains, and service/chain executions. We are implementing a minimal set of canonical parameterized query functions that can provide a simple means of search using the causal dependencies among provenance entities. In turn, we can create provenance-aware GIS software systems using interoperable geospatial provenance services, to demonstrate the benefits of provenance in supporting scientific analysis such as data quality analysis.

## Acknowledgements

We are grateful to the anonymous reviewers, and to Dr. Barry Schlesinger for their detailed comments that helped improve the quality of the paper. This work was funded jointly by National Basic Research Program of China (2011CB707105), Project 40801153 and 41023001 supported by NSFC, LIESMARS and SKLSE (Wuhan University) Special Research Fundings.

## References

Alameh, N. (2003). Chaining geographic information web services. *IEEE Internet Computing, 7*(5), 22–29.

Alonso, G., & Hagen, C. (1997). Geo-Opera: Workflow concepts for spatial processes. In *Proceedings of the fifth international symposium on spatial databases (SSD' 97)* (pp. 238–258). Berlin, Germany.

Altintas, I., Barney, O., & Jaeger-Frank, E. (2006). Provenance collection support in the Kepler scientific workflow system. In *Proceedings of the international provenance, annotation workshop (IPAW'06). Lecture notes in computer science (LNCS) 4145* (pp. 118–132). Berlin, Germany: Springer.

Anand, M., Bowers, S., & Ludäscher, B. (2010). Provenance browser: Displaying and querying scientific workflow provenance graphs. In *Proceedings of the 26th IEEE international conference on data engineering (ICDE 2010)* (pp. 1201–1204). Long Island, CA, USA: IEEE Publication.

Barga, R. S., & Digiampietri, L. A. (2008). Automatic capture and efficient storage of e-Science experiment provenance. *Concurrency and Computation: Practice and Experience, 20*(5), 419–429.

Bowers, S., McPhillips, T., Riddle, S., Anand, M., & Ludäscher, B. (2008). Kepler/pPOD: Scientific workflow and provenance support for assembling the tree of life. In *Proceedings of the second international provenance and annotation workshop (IPAW 2008), Salt Lake City, UT, USA. Lecture Notes in Computer Science (LNCS) 5272* (pp. 70–77). Berlin, Germany: Springer.

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. (2006). Managing the evolution of dataflows with VisTrails. In *Proceedings of the IEEE workshop on workflow and data flow for scientific applications (SciFlow 2006)*, 22nd international conference on data engineering workshops (ICDEW'06) (5pp).

Chebotko, A., Lu, S., Fei, X., & Fotouhi, F. (2010). RDFProv: A relational RDF store for querying and managing scientific workflow provenance. *Data & Knowledge Engineering, 69*(8), 836–865.

Clifford, B., Foster, I., Voeckler, J.-S., Wilde, M., & Zhao, Y. (2008). Tracking provenance in a virtual data grid. *Concurrency and Computation: Practice and Experience, 20*(5), 565–575.

De Longueville, B. (2010). Community-based geoportals: The next generation? Concepts and methods for the geospatial Web 2.0.. *Computers, Environment and Urban Systems, 34*(4), 299–308.

Di, L., & McDonald, K. (1999). Next generation data and information systems for earth sciences research. *Proceedings of the 1st international symposium on digital earth* (Vol. I, pp. 92–101). Science Press.

Foerster, T., Lehto, L., Sarjakoski, T., Sarjakoski, L. T., & Stoter, J. (2010). Map generalization and schema transformation of geospatial data combined in a Web Service context. *Computers, Environment and Urban Systems, 34*(1), 79–88.

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 Y. Zhao (Ed.), *Proceedings of 14th international conference on scientific and statistical database management (SSDBM'02)* (pp. 37–46). Edinburgh, Scotland: IEEE Computer Society.

Frew, J., & Bose, R. (2001). Earth system science workbench: A data management infrastructure for earth science products. In *Proceedings of the 13th international conference on scientific and statistical database management (SSDBM'01), Fairfax* (pp. 180–189). Virginia, USA: IEEE Computer Society.

Fuger, S., Najmi, F., & Stojanovic, N. (Eds.) (2005). *ebXML Registry Information Model Version 3.0*. OASIS Standard, regrep-rim-3.0-os (78pp).

Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., et al. (Eds.) (2010). Provenance XG final report, W3C provenance incubator group. <http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/> Accessed 12.01.11.

Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S., et al. (2006). *An architecture for provenance systems, Technical Report*. ECS, University of Southampton. 162pp.

Holland, D. A., Seltzer, M. I., Braun, U., & Muniswamy-Reddy, K.-K. (2008). PASSing the provenance challenge. *Concurrency and Computation: Practice and Experience, 20*(5), 531–540.

Kim, J., Deelman, E., Gil, Y., Mehta, G., & Ratnakar, V. (2008). Provenance trails in the Wings/Pegasus workflow system. *Concurrency and Computation: Practice and Experience, 20*(5), 587–597.

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

Li, X., Di, L., Han, W., Zhao, P., & Dadi, U. (2010). Sharing geoscience algorithms in a Web service-oriented environment (GRASS GIS example). *Computers and Geosciences, 36*(8), 1060–1068.

Martell, R. (Ed.) 2008. *CSW-ebRIM Registry Service – Part 1: ebRIM profile of CSW*. Version 1.0.0, OGC 07-110r2 (57pp). Open Geospatial Consortium, Inc.

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.

Missier, P., Sahoo, S. S., Zhao, J., Goble, C., & Sheth, A. (2010) Janus: From workflows to semantic provenance and linked open data. In *Proceedings of the third international provenance and annotation workshop (IPAW 2010)* (13pp). Troy, NY, USA.

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

Moreau, L., Ludäscher, B., Altintas, I., et al. (2008). The first provenance challenge. *Concurrency and Computation: Practice and Experience, 20*(5), 409–418.

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. (2008b). The open provenance model (v1.01) (35pp). United Kingdom: University of Southampton. <http://eprints.ecs.soton.ac.uk/16148/1/opm-v1.01.pdf> Accessed 19.11.09.

Nebert, D., Whiteside, A., & Vretanos, P. (Eds.) (2007). *OpenGIS® catalog services specification*. Version 2.0.2, OGC 07-006r1 (218pp). Open Geospatial Consortium Inc.

OGC (2011). Implementations – Compliant Products. Open Geospatial Consortium, Inc. <http://www.opengeospatial.org/resource/products/byspec> Accessed 22.01.11.

PASOA (2006). Provenance Aware Service Oriented Architecture (PASOA). PASOA consortium. <http://twiki.pasoa.ecs.soton.ac.uk/bin/view/PASOA/WebHome> Accessed 12.01.11.

Peng, Z. R., & Tsou, M. H. (2003). *Internet GIS: Distributed geographic information services for the internet and wireless networks*. New Jersey: John Wiley & Sons.

Plale, B., Cao, B., Herath, C., & Sun, Y. (2010). Data provenance for preservation of digital geoscience data. Geological Society of America (GSA), Memoir Volume, 12/2010, 14pp. <http://www.cs.indiana.edu/~plale/papers/PlaleDataProvenancePreservationPreprint.pdf> Accessed 13.07.10.

Sahoo, S. S., Sheth, A., & Henson, C. (2008). Semantic provenance for eScience: Managing the deluge of scientific data. *IEEE Internet Computing, 12*(4), 46–54.

Simmhan, Y. L., Plale, B., & Gannon, D. (2008). Karma2: Provenance management for data driven workflows. *International Journal of Web Services Research, 5*(2), 1–22.

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

Stuiver, J., & Crompvoets, J. (2009). Data Lineage, an essential step in data processing. *GIM International, 23*(9), 33–39.

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 (pp. 221–228). Berlin, Germany: Springer.

Tu, S., & Abdelguerfi, M. (2006). Web services for geographic information systems. *IEEE Internet Computing, 10*(5), 13–15.

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

Vretanos, P. A. (Ed.) (2005). *OpenGIS® filter encoding implementation specification*. Version 1.1.0, OGC 04-095 (40pp). Open Geospatial Consortium, Inc.

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)* (4pp). Irvine, California, USA.

Wei, Y., Di, L., Zhao, B., Liao, G., Chen, A., Bai, Y., et al. (2005). The design and implementation of a grid-enabled catalogue service. In *Proceedings of the 25th anniversary of IEEE international geoscience and remote sensing symposium (IGARSS 2005)*, July 25–29 (pp. 4224–422). COEX, Seoul, Korea.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems, 34*(4), 264–277.

Yang, C., & Raskin, R. (2009). Introduction to distributed geographic information processing research. *International Journal of Geographical Information Science, 23*(5), 553–560.

Yue, P., Gong, J., & Di, L. (2010). Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Computers & Geosciences, 36*(3), 270–281.

Yue, P., & He, L. (2009). Geospatial data provenance in cyberinfrastructure. In *Proceedings of the 17th international conference on geoinformatics (Geoinformatics 2009)*, 12–14 August 2009 (4pp). Fairfax, USA: IEEE Publication.

Zhao, J., Goble, C., Stevens, R., & Turi, D. (2008). Mining Taverna's semantic web of provenance. *Concurrency and Computation: Practice and Experience, 20*(5), 463–472.

Zhao, J., Goble, C., Greenwood, M., Wroe, C., & Stevens, R. (2003). Annotating, linking and browsing provenance logs for e-Science. In Proceedings of workshop on semantic web technologies for searching and retrieving scientific data (6pp). Sanibel Island, Florida, USA.