Computers, Environment and Urban Systems 64 (2017) 103–117

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment

Guillem Closa [a],*, Joan Masó [a], Benjamin Proß [b], Xavier Pons [c]

[a] Grumets Research Group, CREAF, Edifici C, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain
[b] 52°North GmbH, Martin-Luther-King-Weg 24, Münster, Germany
[c] Grumets Research Group, Dep de Geografia, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain

ARTICLE INFO

Article history:
Received 28 October 2015
Received in revised form 24 January 2017
Accepted 25 January 2017
Available online 4 February 2017

Keywords:
Provenance
Distributed enviroment
Conflation
GIS
Metadata

ABSTRACT

Provenance, a metadata component referring to the origin and the processes undertaken to obtain a specific geographic digital feature or product, is crucial to evaluate the quality of spatial information and help in reproducing and replicating geospatial processes. However, the heterogeneity and complexity of the geospatial processes, which can potentially modify part or the complete content of datasets, make evident the necessity for describing geospatial provenance at dataset, feature and attribute levels. This paper presents the application of W3C PROV, which is a generic specification to express provenance records, for representing geospatial data provenance at these different levels. In particular, W3C PROV is applied to feature models, where geospatial phenomena are represented as individual features described with spatial (point, lines, polygons, etc.) and non-spatial (names, measures, etc.) attributes.

This paper first analyses the potential for representing geospatial provenance in a distributed environment at the three levels of granularity using ISO 19115 and W3C PROV models. Next, an approach for applying the generic W3C PROV provenance model to the geospatial environment is presented. As a proof of concept, we provide an application of W3C PROV to describe geospatial provenance at the feature and attribute levels. The use case presented consists of a conflation of the U.S. Geological Survey dataset with the National Geospatial-Intelligence Agency dataset. Finally, an example of how to capture the provenance resulting from workflows and chain executions with PROV is also presented. The application uses a web processing service, which enables geospatial processing in a distributed system and allows to capture the provenance information based on the W3C PROV ontology at the feature and attribute levels.

1. Introduction

According to the Union of Concerned Scientist (UCS), at the end of January 2015, there were 192 Earth observation (EO) satellites in orbit (USC_Satelite_Database, 2015) measuring different Earth parameters and generating, together with a myriad of other sensors and monitoring systems, huge volumes of geospatial data (Kogan, Powell, & Fedorov, 2011). The large and diverse Earth science data, often converted to traditional cartographic products, are consumed by scientific workflows involving multiple complex geoprocessing steps in different contexts at different times (Di, Yue, Ramapriyan, & King, 2013b). In this context, the availability of information about data provenance, which is part of the metadata that provides the description of the origin of the data and the processes involved to achieve the current status (Buneman, Khanna, & Chiew Tan, 2001), is crucial for assessing the suitable fit for purpose in each case.

The scientific community has traditionally considered geosciences data provenance as necessary. In 1991, Lanter (1991) used the word 'lineage' to define the provenance of derived products in geographic information systems (GIS) as information that describes materials and transformations applied to the derivation of data. More recently, Greenwood et al. (2003) expanded Lanter's definition of lineage, considering it as metadata recording the process of experiment workflows and annotations (notes about experiments). According to Simmhan, Plale, and Gannon (2005), provenance can be associated not only with data products but also with the processes that enabled their creation. In practice, these two concepts are difficult to separate, and in this paper, we use them as synonyms. In metadata, processes are referenced by identifiers, and this limits the information about the nature of the processes. We assume that the designated community can access to the same level of acknowledgement and that they know how the process works internally (e.g. which algorithm is involved). This can be partially solved by citing the documentation of process algorithm in the

* Corresponding author.
E-mail addresses: g.closa@creaf.uab.cat (G. Closa), joan.maso@uab.cat (J. Masó), b.pross@52north.org (B. Proß), xavier.pons@uab.cat (X. Pons).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.01.008

metadata or can be rigorously addressed by introducing spatiotemporal information generation models that express the algebra behind a process (Scheider, Gräler, Pebesma, & Stasch, 2016).

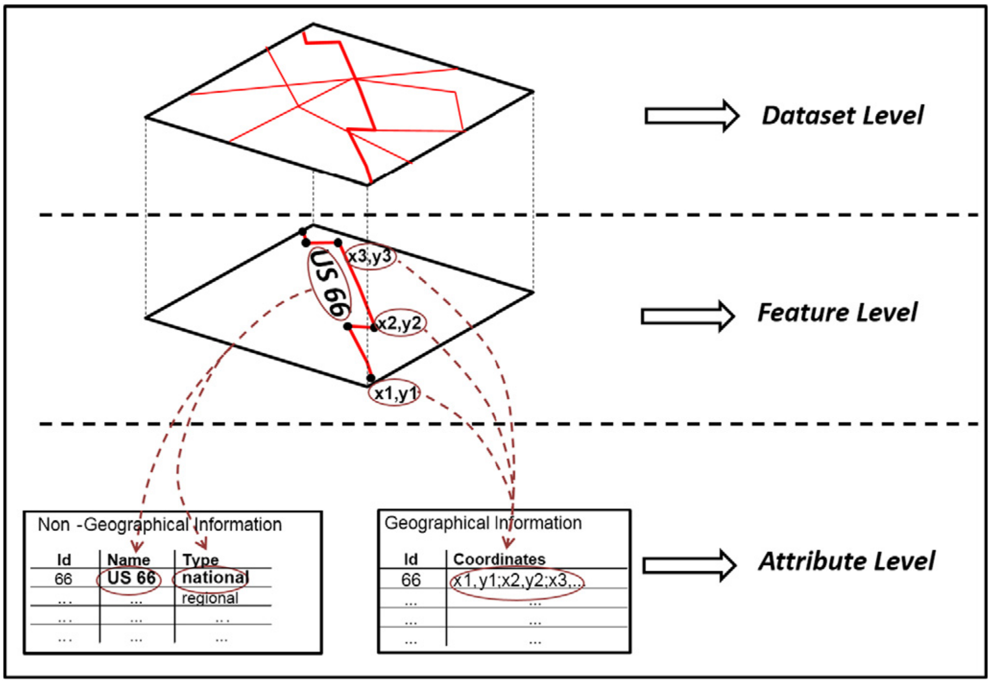
Provenance can be captured manually by editing the metadata after some process has been executed, or it can be automatically recorded though a module (Di, Shao, & Kang, 2013a). This module is called provenance engine in this document.

Despite the documented importance of provenance information, its complete description in geospatial metadata is scarce (Díaz et al., 2012). Normally, most of the geodata come with some provenance information, but in many cases, it is a simple textual form, which has a negative effect on its automated usage (Yue, Gong, & Di, 2010). Therefore, to achieve the maximum benefit of provenance information, it should be recorded according to some precise structure. Thus, before presenting the details on how to connect provenance metadata to the data, it is necessary to review the data models used in geospatial information.

Geospatial data have been traditionally represented in two different models: raster (grid coverages) and vector; this paper focusses on exemplifying the vector model. There are several works related to represent provenance derived from raster models (e.g. Yue, Zhang, Guo, & Tan, 2014). In the vector model, information is organized in features. A feature instance is an abstraction of real world phenomena [International Organization for Standardization (ISO) 19101] and can be tangible, such as a river, building or triangulation pillar, or abstract, such as a political boundary or a health district. Feature instances are grouped in collections of features that share the same feature type (what implies the same sequence of property types) and are described by a set of geometric and non-geometric properties called attributes (Fig. 1). A geometric attribute instance is the position and shape (and even topology) of a feature that can be expressed through geometries such as points, lines and polygons (as a sequence of co-ordinates). Examples of non-geometric attribute instances are the name of a river or the amount of water flowing. Attribute instances of the same kind are grouped in attribute types. In this paper, we allude to a collection of feature instances sharing the same feature type as a dataset, which in the GIS context is represented by a thematic layer (OSGeo, 2015). However, in other environments, a dataset is known as a data product that is composed of a set of feature instances of several types. Moreover, when referring to a feature level, we are talking about feature instance level, whereas when talking about attribute level, we are referring to an attribute instance level. Datasets can also be grouped in dataset collections or series.

Depending on the process type, more or less fine granularity is needed to completely describe provenance. In some cases, provenance at the dataset level would be enough as it is a re-projection of the complete dataset. Other cases may require a finer grained provenance, as in the process of conflation of two datasets using a distance threshold factor, where a part of the content (at the *feature* or *attribute* level) may be affected but the rest of the content may not. For this reason, provenance models should allow the representation of lower levels of geospatial granularity. Therefore, the common characteristics should be shared at a higher level, and just the specifics would be represented at a lower level (Di et al., 2013a). This reduces the redundancy and repetitiveness. To this end, the provenance engine is responsible for skipping the documentation of the same provenance information at more than one level simultaneously to avoid inconsistencies. Although this storage method may have its advantages, it introduces more steps in recovering the provenance of a single feature, and this can affect the service performance when resolving complex queries (Masó, Closa, Gil, & Prob, 2014).

In addition to raster and vector models, Goodchild, Yuan, and Cova (2007) proposed the concept of the geo-atom, defined as an association between a point location in space-time and a property. Geo-atom provides the foundation for discrete-object and continuous-field conceptualizations. However, no provenance-related works have been found. Another representation of data is provided by the Sensor Web and the Observation and Measurements (O&M) standard. O&M is an international standard developed by the Sensor Web Enablement (SWE) initiative of the Open Geospatial Consortium (OGC), which defines a conceptual schema encoding for observations and for features associated with the sampling process of observations (ISO 19156:2011, 2011). In applying O&M to geosciences, Cox (2015) addressed the provenance issue using an association class 'PreparationStep'. However, this approach was not fully satisfactory, particularly as the preparation step



Non -Geographical Information

Id	Name	Type
66	US 66	national
…	…	regional
…	…	…
…	…	…

Geographical Information

Id	Coordinates
66	x1,y1;x2,y2;x3,...
…	…
…	…
…	…

Fig. 1. Conceptual diagram representing the three levels of granularity of geographical information.

is not easily linked explicitly to a predecessor; there is a very wide range of specimen preparation and provenance paths. As an alternative, Cox proposes the combination of PROV with O&M to describe provenance at the attribute level. Because of the constraints of the research conducted and described in this paper, we do not further consider this approach or the geo-atom approach.

Currently, most of the geospatial metadata use ISO standards for the description of geospatial provenance information (Masó et al., 2014; Di et al., 2013a). Beyond the geospatial community, there is no single model for lineage representation across disciplines and, because of diverse needs, it is a challenge to converge all of them in a suitable single model (Myers et al., 2003). In the computer science community, the Provenance Markup Language (PML) and the Open Provenance Model (OPM) were initially proposed. Feng (2013) mapped OPM with ISO model, which allows accessing data provenance in spatial data infrastructures (SDI) by other domains that require the use of spatial data. On the basis of the OPM, the World Wide Web Consortium (W3C) led efforts to develop a more flexible and interoperable provenance ontology and data model for capturing data provenance: the W3C PROV (Moreau & Missier, 2013), hereinafter referred to as PROV.

Recently, some initiatives have appeared to promote the use of PROV in the geospatial realm (e.g., Tilmes et al., 2013; Garijo, Gil, & Harth, 2014). In this sense, Ma et al. (2014) compared PROV with ISO standards, OPM and PML showing the similarities and the improvements that PROV brings to the geospatial field. Other authors such as Lopez-Pellicer and Barrera (2014) proposed to adapt and extend the PROV model to geospatial community requirements. He, Yue, Di, Zhang, and Hu (2015) combined PROV and ISO to describe provenance at the dataset and feature levels, without considering the attribute level.

Despite these examples, a comprehensive description of geospatial provenance at the attribute, feature and dataset levels, either with ISO or with PROV, remain challenging. To this aim, the present work addresses an analysis of two different alternatives available for the description of provenance at the three levels of granularity (dataset, feature and attribute levels) in distributed environments. Following this, the application of PROV is presented as a suitable one for the representation of the different provenance granularities in distributed environment contexts. As a proof of concept, a geospatial data conflation Web Processing Service (WPS) instance is presented to demonstrate the feasibility of the model. Finally, an example of how to capture the provenance resulting from workflows and chain executions by using PROV and its technological architecture is also presented. This paper is a step forward in improving the completeness of geospatial provenance at the attribute, feature and dataset levels.

2. Metadata standards for the descrption of geospatial provenance

A metadata standard intends to establish a common understanding of the semantics of data to ensure correct and proper use and interpretation of the data by their owners and users. Metadata should link directly to the data itself (Masó, Pons, & Zabala, 2012) and, when selecting a standard for describing provenance of a geospatial object, we need to ensure that the model captures the following elements (Di et al., 2013a):

- *Sources:* A geospatial object, which can be a dataset, feature or geometric/non-geometric attribute that was used to derive the resulting elements. Such elements can be referenced using a descriptive citation, an element id, a metadata id, an element URI or a metadata URI. Note that this definition encompasses the three levels of granularity.
- *Process executions* (*process steps*): These are operations applied to a dataset, feature or geometric and non-geometric attribute. They can be referenced by providing the name of the operation, a URI of the operation or a full description of the operation.
- *Process*: An engine that can execute a process step.
- *Algorithm*: The abstract logic that describes how a process engine was implemented.
- *Parameters*: Constant or variable elements that modify the behaviour of the algorithm.
- *Responsible parties*: People and institutions that are in charge of sources, algorithms and execution of geospatial operations.

Garijo et al. (2014) also found the need for these elements when elucidating on possible 38 queries on provenance metadata.

In this section, we explore the potential and the weakness of ISO 19115 and W3C PROV for representing geospatial lineage at the dataset, feature and attribute levels of granularity.

2.1. ISO 19115 family

The ISO 19115:2003 and 19115-2:2009 standards define the schema for describing geographic information and services metadata. In the ISO 19115 model, provenance information (LI_Lineage) is part of the DQ_DataQuality (ISO 19115-1:2014). The LI_Lineage is divided into three parts: Statement, which gives a textual overview of the lineage information; LI_Source, describing all the sources involved in the generation of the dataset and LI_Process, defining which processes were conducted to generate a specific data. When applied to remote sensing

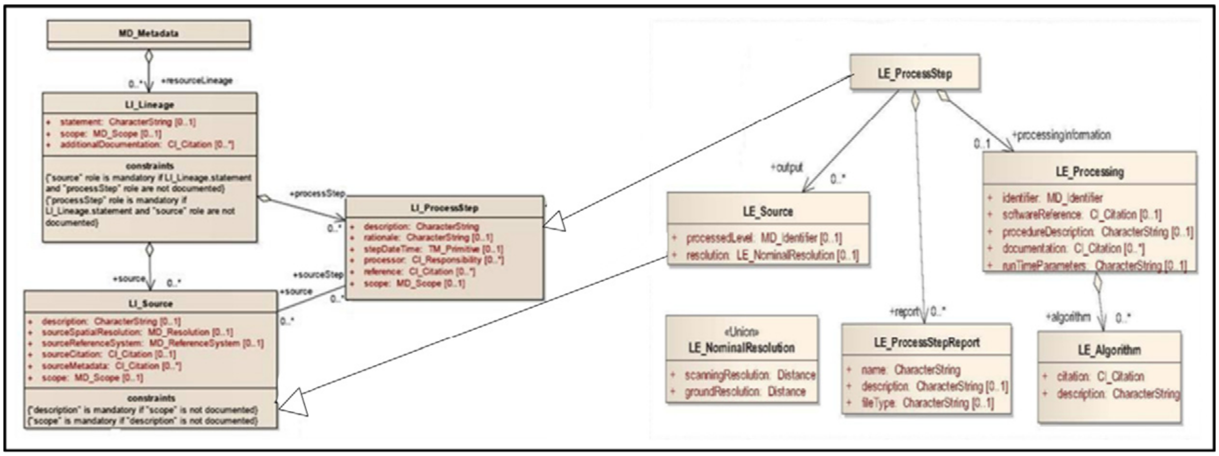

MD_Metadata
0..*
+resourceLineage
LI_Lineage
+ statement: CharacterString [0..1]
+ scope: MD_Scope [0..1]
+ additionalDocumentation: CI_Citation [0..*]
constraints
{"source" role is mandatory if LI_Lineage.statement and "processStep" role are not documented}
{"processStep" role is mandatory if LI_Lineage.statement and "source" role are not documented}
+processStep
0..*
LI_ProcessStep
+ description: CharacterString
+ rationale: CharacterString [0..1]
+ stepDateTime: TM_Primitive [0..1]
+ processor: CI_Responsibility [0..*]
+ reference: CI_Citation [0..*]
+ scope: MD_Scope [0..1]
+sourceStep
+source
0..*
+source
0..*
0..*
LI_Source
+ description: CharacterString [0..1]
+ sourceSpatialResolution: MD_Resolution [0..1]
+ sourceReferenceSystem: MD_ReferenceSystem [0..1]
+ sourceCitation: CI_Citation [0..1]
+ sourceMetadata: CI_Citation [0..*]
+ scope: MD_Scope [0..1]
constraints
{"description" is mandatory if "scope" is not documented}
{"scope" is mandatory if "description" is not documented}
LE_ProcessStep
+output
0..*
0..1
+processingInformation
LE_Processing
+ identifier: MD_Identifier
+ softwareReference: CI_Citation [0..1]
+ procedureDescription: CharacterString [0..1]
+ documentation: CI_Citation [0..*]
+ runTimeParameters: CharacterString [0..1]
LE_Source
+ processedLevel: MD_Identifier [0..1]
+ resolution: LE_NominalResolution [0..1]
+report
0..*
+algorithm
0..*
«Union»
LE_NominalResolution
+ scanningResolution: Distance
+ groundResolution: Distance
LE_ProcessStepReport
+ name: CharacterString
+ description: CharacterString [0..1]
+ fileType: CharacterString [0..1]
LE_Algorithm
+ citation: CI_Citation
+ description: CharacterString

Fig. 2. Lineage UML diagram under ISO 19115-1 and 19115-2, including Source (LI_Source and LE_Source) and Process step (LI_ProcessStep and LE_ProcessStep).

images, the LI_Lineage is insuficent. In particular, there is no place to document the processing level, the processing software, algorithm and so on. In 2009, the LI_Lineage was extended in ISO 19115-2; LE_Source and LE_Processing were added, which included the previously mentioned aspects, among others. The LE_Processing extends the process information by introducing tags about the software reference, the algorithm used and procedure description, among others, whereas the LE_Source completes the information with the process level and the resolution (Fig. 2). According to Di et al. (2013b), the combination of ISO 19115 and ISO 19115-2 serves as generic geospatial metadata models, and the lineage models defined within them can potentially document any geospatial provenance information.

The ISO model allows describing provenance information in three different ways: a list of sources and a list of processSteps, a list of sources that are used in concrete processSteps, and a list of processSteps that use sources. In a distributed environment, ISO can list the processSteps of a service-oriented architecture such as the WPS and describe the sources of the data-oriented services such as the Web Coverage Service (WCS) and the Web Feature Service (WFS). ISO 19139 provides the *eXtensible Markup Language* (XML) implementation schema for ISO 19115, specifying the metadata record format to describe, validate and exchange geospatial metadata written in XML. The benefits of this are apparent given that many of the geospatial services use XML as the primary format for message exchange (Simmhan et al., 2005).

2.1.1. Dataset-, feature- and attribute-level provenance with ISO

In the ISO model, provenance information can be specified at different levels of granularity using the role value of *scope*: '*dataset series*', '*data set*', '*feature type*', '*feature instance*', '*attribute type*' or '*attribute instance*'. Nevertheless, the hierarchical tree form of the standard generates a very deep structure that hinders comprehensibility.

We explored the possibility of combining the ISO model with Geographic Markup Language (GML) architecture to describe provenance at the attribute and feature levels. GML offers the possibility to embed an ISO document directly in a feature or a feature collection by using 'gml:metaDataProperty' to reference the provenance information. Specifically, the 'xlink:href', 'xlink:role' and 'xlink:arcrole' attributes were proposed to fully describe the relationship of features and attributes to the provenance elements in the dataset-level provenance file. However, metaDataProperty was recently deprecated in GML 3.2. Therefore, this option is not recommended. In addition, the possibility of defining a complex property type derived from 'AbstractMetadataPropertyType' was also explored, but this requires addition in the GML schema, which is not always possible. Unfortunately, there is a lack of consensus on how to implement provenance at the feature and attribute levels using the ISO 19115 Lineage model.

2.2. W3C PROV

According to Groth and Moreau (2013), provenance is information about entities, activities and people involved in producing a piece of data or a thing, which can be used to assess quality, reliability and trustworthiness. PROV defines a provenance data model (Moreau & Missier, 2013) to support the interoperable interchange of provenance in heterogeneous environments such as the web. The PROV core structure relies on the definition of the *entities*, *activities* and *agents* that are involved in producing a piece of data or a thing and on how they are related by defining the following four property types: *wasGeneratedBy*, *wasAssociatedBy*, *wasAttributedTo* and *used* (part of Fig. 3 enclosed by the dashed line).

The PROV ontology (Moreau & Missier, 2013) document expresses the PROV-DM using the W3C OWL2 Web Ontology Language (OWL2). It provides a set of classes, properties and constraints that can be used to represent and interchange provenance information. Using this ontology, provenance can be encoded in Resource Description Framework (RDF). RDF is a standard model for data interchange on the web, extending the linking structure of the web to use URIs to name the relationship between things and the two ends of the link, usually referred together as a 'triple' (W3C Semantic Web, 2015). Consequently, the RDF notation allows describing, capturing and querying provenance in a distributed environment. There are several RDF common serialization formats; in this paper, we favoured the use of Notation3 (N3). The use of RDF brings us closer to Linked Data (http://linkeddata.org), which allows the sharing of information in a way that can automatically be read by computers and enables data from different sources to be connected and queried (Bizer, Heath, & Berners-Lee, 2009). In the geospatial world, Linked Data allows the setting of relationships between multiple datasets, incorporating additional descriptions to original data (Vilches-Blázquez, Villazón-Terrazas, Corcho, & Gómez-Pérez, 2014) and enriching the final datasets and maps.

PROV can be used in heterogeneous environments and several disciplines, but its application in the geospatial domain requires a matching process between geospatial provenance concepts and PROV semantics.

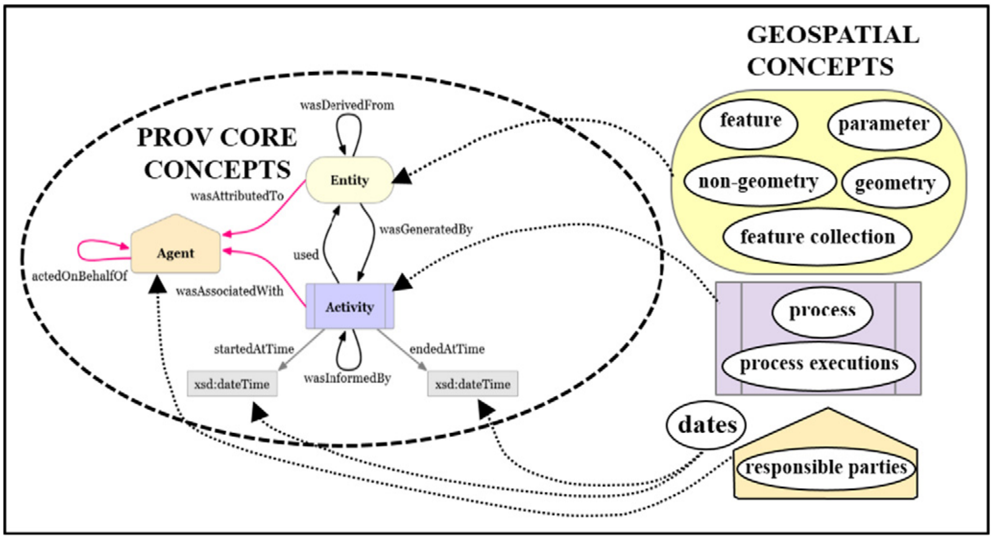


Fig. 3. Matching of geospatial data concepts with the core elements of PROV-DM.

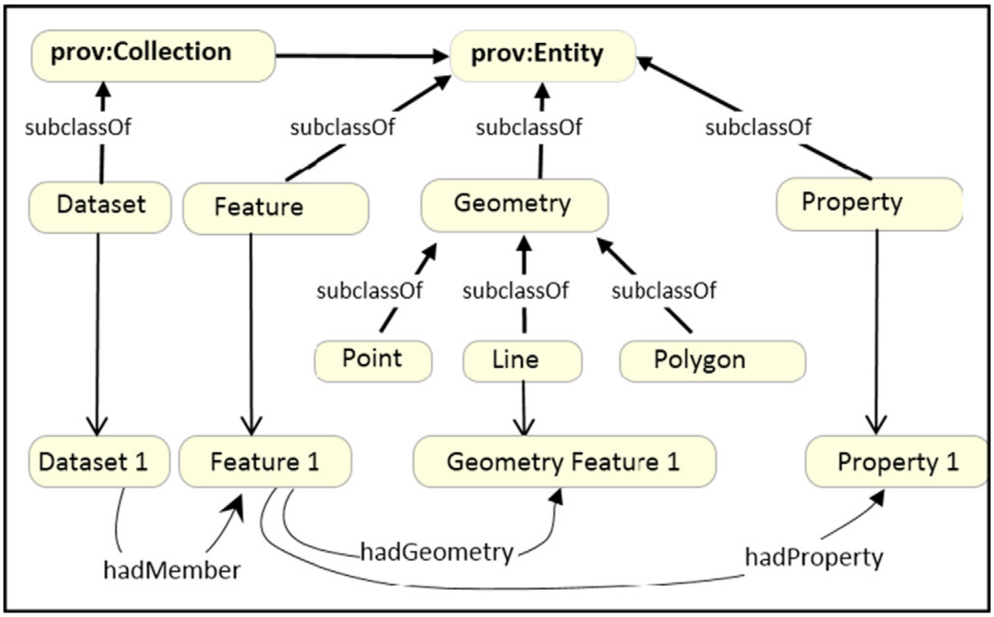


Fig. 4. Diagram representing the dataset, feature and attribute levels in PROV.

Fig. 3 shows the correspondence between the PROV core structure elements and the geospatial provenance concepts.

In addition, there is a need to define the geospatial algorithm, which does not match with any class of the PROV core structure (part of Fig. 3 enclosed by the dashed line) but matches with that of the PROV extended structures (Moreau & Missier, 2013) instead. In PROV, an algorithm is considered as a *Plan*. A plan is defined as 'an *entity* that represents a set of actions or steps intended by one or more agents to achieve some goals' (Groth & Moreau, 2013). According to this definition, the execution of an *activity* needs a *plan*.

3. Geospatial extension of PROV

Several characteristics make the PROV a suitable data model to describe geospatial provenance at the feature and attribute levels:

- It is an object-oriented data model based on the declaration of classes and objects corresponding to real-world things. This conceptual model offers a flexible solution for linking provenance information to geospatial elements with the necessary semantics and eases the description of features and attributes.
- In PROV, lineage information can be documented in RDF notation, which adapts better than XML to describe object-oriented data models and exchange data provenance in distributed environments.
- The broad definition of PROV classes such as entities and activities, which implicitly includes levels of granularity (e.g. an entity can be a dataset, a feature or an attribute), facilitates the implementation of provenance at different levels.
- PROV requires less computer storage space than that required by the combination of ISO and GML. A very simple provenance example[1] was documented with ISO and GML (https://github.com/GuillemClosa/PROV_geo_extension/tree/master/ISOGML) and with PROV (https://github.com/GuillemClosa/PROV_geo_extension/blob/master/W3CPROV/Conflation_PROV.N3). The example shows how the PROV document is much lighter (12 KB) than the same example using the combination of ISO and GML documents (23 KB), almost 100% more.

Different examples of the usage of PROV to describe provenance at the different granularities in the geospatial context already exist. We explored the possibility of embedding PROV information serialized with XML directly in the GML-encoded features for the representation of the feature and attribute levels (https://github.com/GuillemClosa/PROV_geo_extension/tree/master/W3CPROVGML). Using this method, similar to the one presented in Section 2.1.1, the same obstacles were detected. Other researchers, such as Lopez-Pellicer and Barrera (2014), suggested an expansion of the PROV-DM to adapt it to the needs of geospatial data and proposed the inclusion of ISO19115 lineage concepts such as 'primary topic' and 'scope'.

This paper contributes to this issue from a different point of view. In sub-section 3.1, we present a general provenance model (Fig. 5) for geospatial data at the three levels of granularity based on the definition of entities, agents, activities, plans and the interrelationships between these PROV classes. A use case is presented in sub-section 3.1.

3.1. Provenance model

A. **Entity**

An *entity* includes all kinds of data sources or results at all levels of granularity, even at the attribute level. Feature level is adopted as the basic level to describe the three different levels of granularity of geospatial provenance (Fig. 4). Thus, features are mapped as *entities*. Next, a dataset (considered as a collection of features) is mapped as a collection, which is also treated as an *entity*. Datasets acquire features as members by declaring *hadMember*. At the *attribute* level, both *geometric* and *non-geometric* properties are also considered as *entities*. The reason of this decision is not conceptual or practical: In PROV, things we want to describe the provenance of are called entities (Moreau & Missier, 2013), so we are forced to consider attributes as a special kind of *entity*.

Properties need to be related with features, but PROV does not have the right relation type to do this. Therefore, we propose the introduction of *had Geometry* and *hadProperty* relations, and thus, *feature* can gain geometry and property, respectively (Table 1). Geometric properties can be *points*, *lines* or *polygons*, which are sub-classes of features.

[1] This example is a simplification of the use case presented in Section 3.2.2. It is based on the conflation execution of two features of two different datasets.

Table 1
Declaration of different levels of entities and their relationships in RDF.

```
@prefix prov: <http://www.w3.org/ns/prov#> .

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix geos: < http://www.opengis.uab.cat/geos-prov#> .


geos:Dataset rdf:subClassOf prov:Collection .

geos:Feature rdf:subClassOf prov:Entity .

geos:Dataset1 prov:hadMember geos:Feature1 .

geos:Geometry rdf:subClassOf prov:Entity .

geos:Point rdf:subClassOf geos:Geometry .

geos:Line rdf:subClassOf geos:Geometry .

geos:Polygon rdf:subClassOf geos:Geometry .

geos:Property rdf:subClassOf prov:Entity .

geos:Feature1 geos:hadGeometry geos:GeometryFeature1 .

geos:Feature1 geos:hadProperty geos:Property1.
```

B. **Activity**

In a geospatial PROV implementation, a geoprocess execution is considered an *activity (*geos:Execution rdf:subClassOf prov:Activity). The definition of relationships between *entities* and *activities* implies the definition of the granularity level; an activity can act over the complete dataset or only over selected features or attributes. A more detailed explanation of the relationships between *entities* and *agents* with *activities* is given in part *E* of this sub-section.

C. **Agents**

An *agent* is something or somebody who has some responsibility over an *activity*. The definition of relationships between *agents* and *entities* implies a granularity-level definition; for example, an *agent* may have some responsibility over just some attributes or over the complete dataset. The way *agents* function is defined using the *prov:role* attribute; an *Agent* can act as the executor of a geoprocess, the developer of an algorithm and so on. A more detailed explanation of *prov:role* is provided in part *F* of this sub-section. In this example (Fig. 5), we specified that there are two *agents*: the developer of the algorithm used in the execution (Person 2) and the client or the executor of the process (Person 1). All agents act on behalf of other (*prov:actedOnBehalfOf*) *agents*; these may be, for instance, employees of a company. The delegation property extends responsibility for an *activity* and *entity* until the delegator (Groth & Moreau, 2013).

D. **Plans**

When using PROV in the geospatial context, a *plan* is used to define the provenance of the implemented algorithm. Normally, algorithms are members (*prov:hasMembers*) of a bigger service. This service, which is a sub-class of *prov:Collection*, may be composed of several geoprocesses or algorithms.

E. **Interrelationships**

The PROV model also relies on the definition of four property types that serve to relate the aforementioned class elements: *wasGeneratedBy*, *wasAssociatedBy*, *wasAttributedTo* and *used*. Fig. 5 shows how these four PROV properties are used together with the PROV classes to express geospatial provenance at the dataset, feature and attribute levels. To simplify the diagram, relationships between *activities* and *agents* with *entities* are only drawn at the feature instance level, but the same was performed at the attribute dataset levels.

The level of granularity defined in a PROV model mainly depends on two main aspects, the *entity*-level definition (dataset, feature and attribute) and the way that *activities* and *agents* are related with *entities*.

Spatial objects (datasets, features and attributes) are generated (*prov:wasGeneratedBy*) by activities. An *activity* can act over the whole dataset or just over a part of it (some attributes of features or specific features), so the definition of this relationship implies the definition of the level of granularity.

Someone runs the executions (*prov:activity*), so these are associated with (*prov:wasAssociatedWith*) an agent (e.g. the person who executes the operation). *Agents* may have responsibility over the complete dataset or just over a part of the content, dictating the level of granularity. At the same time, the *activities* use (*prov:used*) *entities* to run their operations. Finally, entities are attributed (*prov:wasAttributedTo*) to an *agent*.

A plan, which is used to capture the algorithm, is attributed to (*prov:wasAttributedTo*) an *agent* (the person who developed the algorithm). Simultaneously, *activity* used (*prov:used*) a *plan* to be executed.

The majority of geospatial operations require the use of special parameters that modify the behaviour of the execution, e.g. map projection, geographic datum, resolution, distance threshold, etc. In PROV, because an *entity* is any kind of thing (part A of this sub-section), parameters are also described as entities. Parameters are used (*prov:used*) by *activities*.

F. **Roles**

Entities and agents may have different functions inside the model: geospatial features (*prov:entities*) can be an input or an output of a process (*prov:activity*), and a responsible party (*prov:agent*) can be the developer of an algorithm (*prov:plan*) or the executor of a geoprocess (*prov:activity*). The *prov:hadRole* property and *prov:Role*

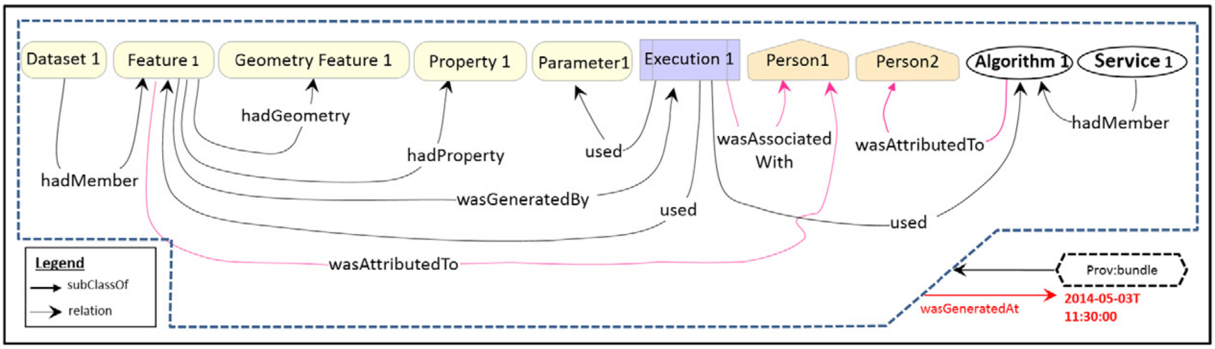


Fig. 5. PROV model for geospatial provenance representing feature, attribute and dataset levels.

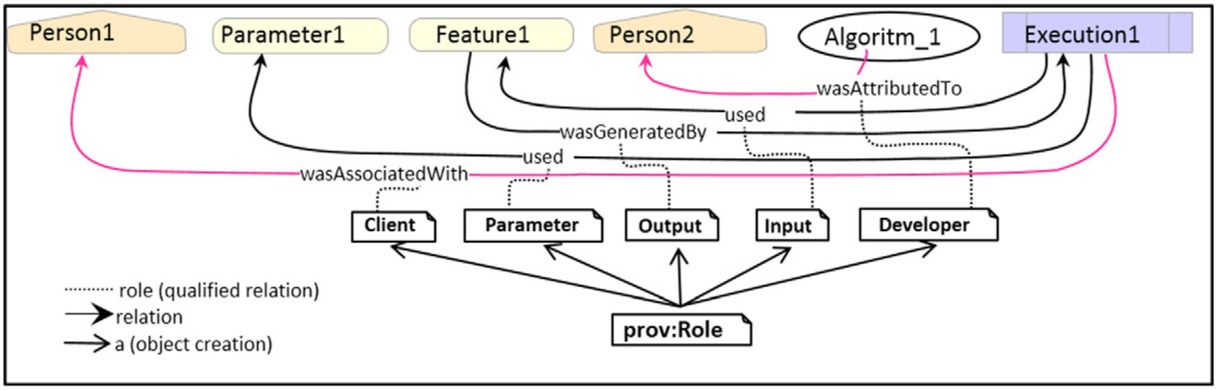


Fig. 6. Diagram representing the PROV roles corresponding to Fig. 5.

class are used to describe the functions that an *agent* and an *entity* have with respect to an *activity*. The role is defined in the context of a usage, generation, invalidation, association, initialization and finalization of a qualified property. A role is defined by connecting it to a qualified relation property in which the influencer of that relation property receives the role defined in the prov:Role class.

Fig. 6 shows the roles of the example presented in Fig. 5, and one role is illustrated in N3 notation language in Table 2.

3.2. Example of use: web processing conflation service

A conflation process between different datasets is the selected geoprocess to demonstrate the implementation previously presented, i.e., a model to describe geospatial provenance at the feature and attribute levels with PROV in a distributed environment. The aim of this process is to enhance the Base Map [U.S. Geological Survey (USGS)] using a Target Dataset [National Geospatial-Intelligence Agency (NGA)]. As our example is designed to be executed remotely, a WPS, which enables geospatial processing in a distributed system, fits with our needs. In Section 3.2.2, a PROV model ontology for a conflation example is described. Then, in Section 3.2.3, the provenance captured in the previous example is described.

3.2.1. Geospatial conflation process

Geospatial data conflation is the compilation or reconciliation of two different geospatial datasets covering overlapping regions (Saalfeld, 1988). The purpose of conflation is to combine the elements of highest quality of different datasets created at different times or based on different levels of accuracy and/or precision, with the final objective of improving the quality of the resulting dataset (Fig. 7). Depending on the types of geospatial datasets, the conflation process can be classified as vector to vector, vector to image, or image to image (Chen, Knoblock, & Shahabi, 2008); moreover, Ruiz, Ariza, Ureña, and Blázquez (2011) added the raster to DEM and DEM to DEM types. In this paper, a vector to vector case is developed and tested. The extensions to other conflation types is left as future work. A conflation use case can merge both attribute and geometric information or just one of these:

Table 2
The prov:Role defined as *geos:Input* is associated with the qualified relation *prov:Usage* between Execution1 and Feature1 receiving this *geos:Input* role.

```
Geos:Input a prov:Role .

geos:Execution1 prov:used geos:Feature1 .

geos:Execution1 prov:qualifiedUsage [

                    a prov:Usage ;

                    prov:entity  geos:Feature1;

                    prov:hadRole geos:Input  ] .
```

- Attribute: The process of adding new attribute information to a dataset based on feature matching because the information is missing or the data are outdated.
- Geometric: The process of adding a new feature or correcting the position and shape of a feature based on an algorithm.

Our conflation example consists of adding new features and updating the geometry or other attributes, which are based on two conflation rules, the *Id matching rule* and the *Distance matching rule*. The *Id matching rule* adds features from the source dataset if they do not exist in the target dataset. A *Distance matching rule* acts as a threshold, where NGA features closer to a USGS feature than the distance threshold can be considered the same. Deriving from these specific rules, different situations can emerge as follows:

- Some completely new features can be added to the USGS dataset, and in these cases, feature-level provenance should be provided.
- Other features are conflated at the attribute level: the geometrical property (location) is modified or non-geometrical properties (attributes) are added from the NGA dataset. In both cases, an attribute-level provenance is needed.

3.2.2. Provenance model for a conflation process

To describe the presented conflation example process executed in WPS, we implemented a conceptual model divided into six levels (which we call layers) of abstraction, from the most general and abstract concepts to more specific executions. This structure facilitates the model comprehensibility and the correspondence between PROV and geospatial concepts.

Fig. 8 shows the complete provenance conflation diagram. The different colours represent the six layers of abstraction of the approach. The central part of the figure enclosed by a dashed line represents the *bundle*. In PROV, a *bundle* is a named set of provenance descriptions and is itself an entity, thus allowing provenance of the provenance to be expressed. Thus, in our example, which represents a single execution, the bundle includes the provenance information that emanates from that specific execution. The provenance ontology used in this conflation example can be found in Notation (N3) serialization in https://github.com/GuillemClosa/PROV_geo_extension/blob/master/model.N3.

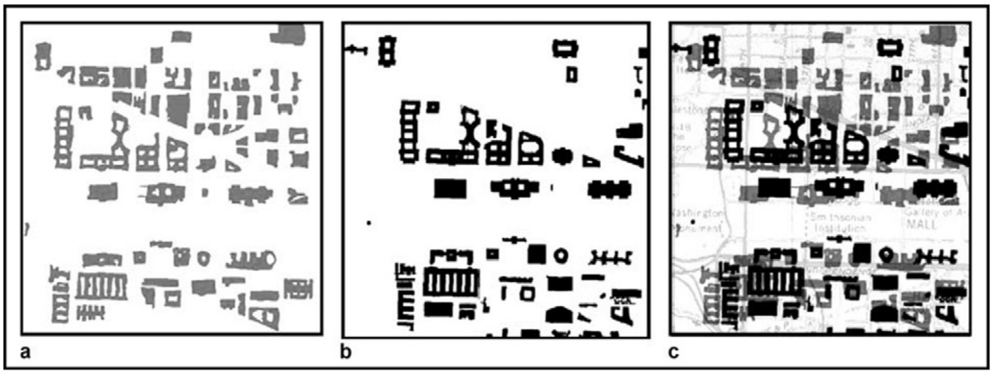


Fig. 7. Graphical example of conflation of two different sources (Seth & Samal, 2007).

3.2.2.1. Levels of abstraction.

- The **Geospatial Concepts** (layer 0) describes how the general geospatial concepts (explained in Section 2) are related to the PROV semantics. This includes the abstract of WPSService and WPSProcess, the WPS Execution and the Responsible Party. This level also defines all *entities*: parameter, feature collection (dataset), feature and both properties of features (geometric and non-geometric). These concepts are shown in the topmost layer and derive directly from the PROV OWL definition. All the elements have a defined role in the model. This level also defines the generic roles, which in this example are Developer, Client, Process Input and Process Output.
- The **WPS Conflation Profile** (layer 1) defines a generic conflation WPS, which is a sub-class of WPSExecution, and a generic conflation algorithm, which is a sub-class of WPSProcess. The first one is considered an activity, whereas the second one is a plan. This level also specifies the kind of input and output roles declared in a conflation process. Specifically, these roles are *Conflation distance threshold*, which filters the executions depending on the distance (beyond this distance, the algorithm will not look for new matches), a *Reference Map* role, which is the map that is being updated, and a *Crowds Map* role, which nourishes the reference map. In addition, as a result of the conflation process, there is a conflation output map role.
- The **Conflation WPS** (layer 2) describes the conflation example process, in our case developed by 52North. In this example, the '52NWPSService' has a member '52N Conflation Algorithm version 1', which is used during the '52N ConflationWPSExecution' and is attributed to a 'Benjamin', who acts as a developer.
- The **User** level (layer 3) defines the agent who executed the process. In this example, *David*, who has the role of the executor, *actedOnBehalf* of the NGA.
- The **Conflated Map Definition** (layer 4) defines the conflation example inputs and outputs at the three levels of granularity. This defines datasets (the source maps and the conflated map), feature types and attribute types and describes the generic concept of distance. In this level, features and attributes (both geometric and non-geometric) are sub-classes of entities, but datasets are sub-classes of collections of entities.
- In the **Conflation layer** (layer 5), the user supervises the conflation steps that involve a set of a few features, and, for each step, we document the specific features and specific attributes participating in it (Fig. 9). In this level, the value of the distance parameter is defined. All the elements of this level are objects themselves. The relationship between the features and attributes and their execution with the specific input and output parameters are defined.

3.2.3. Provenance captured

The following is a sequence of tables (Tables 3–12) illustrating fragments of an example of provenance conflation output encoded in RDF. The reader can find explanations by reading the lines starting with a '#' symbol. The complete provenance data derived from the WPS

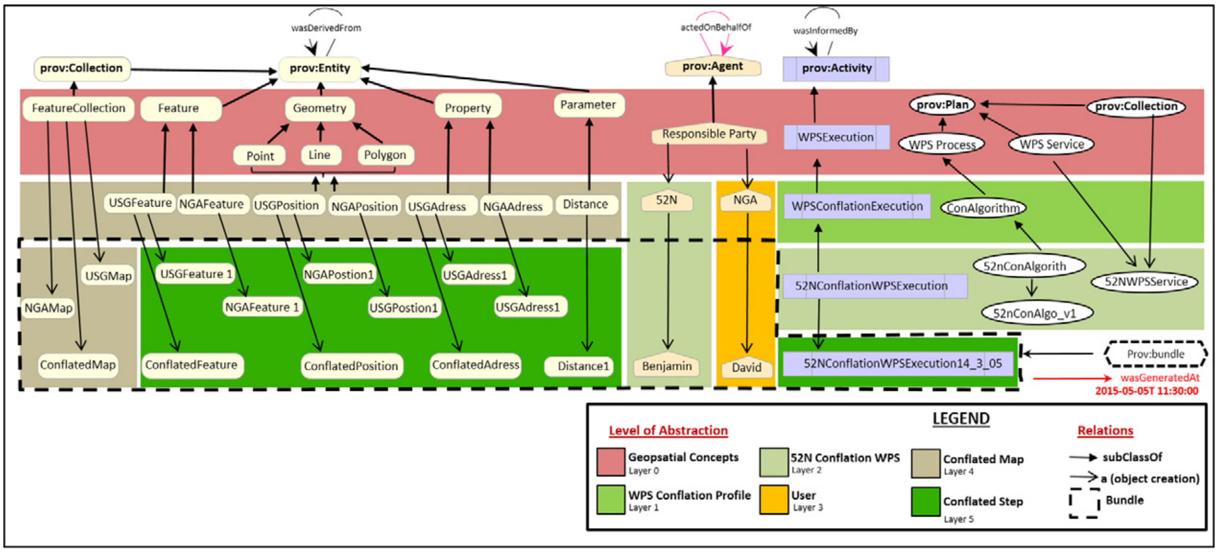


Fig. 8. Diagram showing the main PROV elements involved in the conflation process example. The complete diagram of the PROV conflation example with all the relations between different elements can be found at https://github.com/GuillemClosa/PROV_geo_extension/blob/master/conflation_PROV_model_legend.png, and the complete N3 notation can be found at https://github.com/GuillemClosa/PROV_geo_extension/blob/master/model.N3.

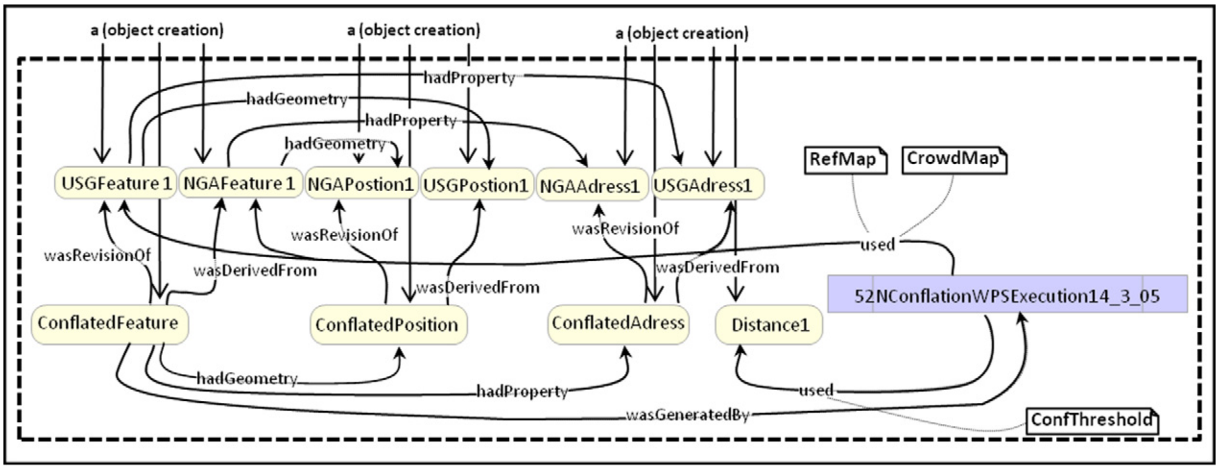
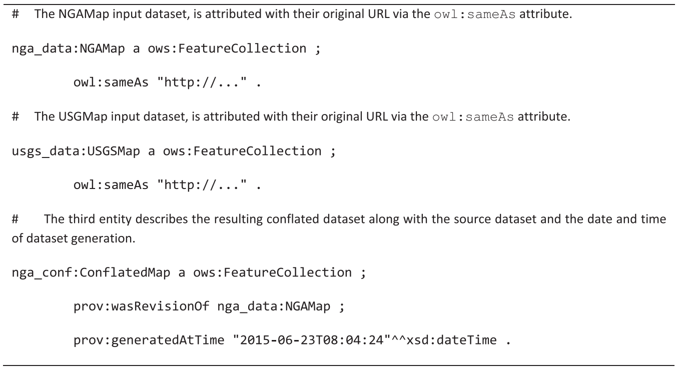


Fig. 9. Diagram showing the conflation use case definition.

execution in N3 notation can be found at https://github.com/GuillemClosa/PROV_geo_extension/blob/master/prov.N3.

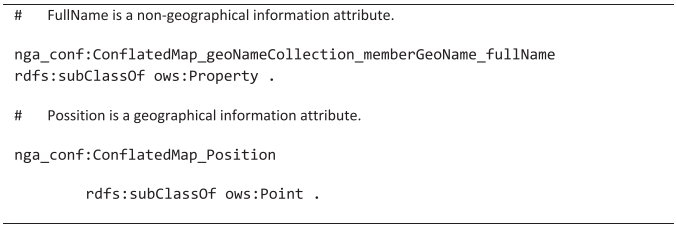
- **Datasets involved in the conflation process**

Table 3
Datasets are described as ows:FeatureCollection.


```
#   The NGAMap input dataset, is attributed with their original URL via the owl:sameAs attribute.

nga_data:NGAMap a ows:FeatureCollection ;

        owl:sameAs "http://..." .

#   The USGMap input dataset, is attributed with their original URL via the owl:sameAs attribute.

usgs_data:USGSMap a ows:FeatureCollection ;

        owl:sameAs "http://..." .

#     The third entity describes the resulting conflated dataset along with the source dataset and the date and time
of dataset generation.

nga_conf:ConflatedMap a ows:FeatureCollection ;

        prov:wasRevisionOf nga_data:NGAMap ;

        prov:generatedAtTime "2015-06-23T08:04:24"^^xsd:dateTime .
```

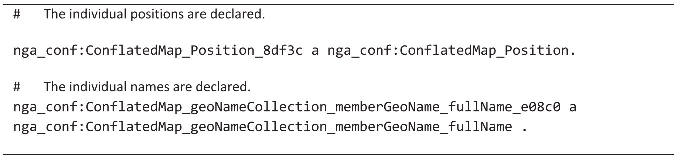
- **Attribute types involved in the conflation process**

Table 4
Attribute types involved in the process are described by the rdfs:subClassOf attribute.


```
#     FullName is a non-geographical information attribute.

nga_conf:ConflatedMap_geoNameCollection_memberGeoName_fullName
rdfs:subClassOf ows:Property .

#     Position is a geographical information attribute.

nga_conf:ConflatedMap_Position

         rdfs:subClassOf ows:Point .
```

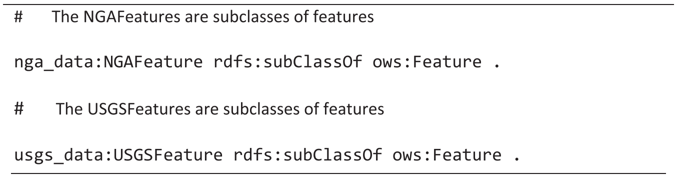
In this example, the positions and the name properties of the features are taken into account. The non-positional properties are specified by the conflation rules. Fixed attribute values are not taken into account here.

Table 5
RDF properties are used to describe individual feature properties.


```
#     The individual positions are declared.

nga_conf:ConflatedMap_Position_8df3c a nga_conf:ConflatedMap_Position.

#     The individual names are declared.
nga_conf:ConflatedMap_geoNameCollection_memberGeoName_fullName_e08c0 a
nga_conf:ConflatedMap_geoNameCollection_memberGeoName_fullName .
```

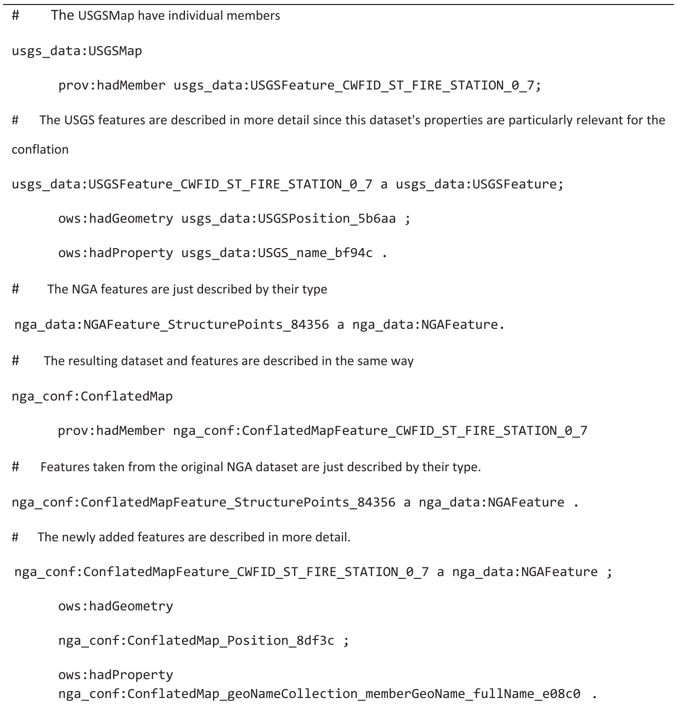
- **Feature types involved in the conflation**

Table 6
Feature types of the involved datasets are also described by the rdfs:subClassOf attribute.


```
#     The NGAFeatures are subclasses of features

nga_data:NGAFeature rdfs:subClassOf ows:Feature .

#      The USGSFeatures are subclasses of features

usgs_data:USGSFeature rdfs:subClassOf ows:Feature .
```

- **Individual features of the datasets**

Table 7
Individual members of the input datasets are described as members of a dataset.


```
#       The USGSMap have individual members

usgs_data:USGSMap

        prov:hadMember usgs_data:USGSFeature_CWFID_ST_FIRE_STATION_0_7;

#    The USGS features are described in more detail since this dataset's properties are particularly relevant for the
conflation

usgs_data:USGSFeature_CWFID_ST_FIRE_STATION_0_7 a usgs_data:USGSFeature;

        ows:hadGeometry usgs_data:USGSPosition_5b6aa ;

        ows:hadProperty usgs_data:USGS_name_bf94c .

#     The NGA features are just described by their type

nga_data:NGAFeature_StructurePoints_84356 a nga_data:NGAFeature.

#     The resulting dataset and features are described in the same way

nga_conf:ConflatedMap

        prov:hadMember nga_conf:ConflatedMapFeature_CWFID_ST_FIRE_STATION_0_7

#     Features taken from the original NGA dataset are just described by their type.

nga_conf:ConflatedMapFeature_StructurePoints_84356 a nga_data:NGAFeature .

#     The newly added features are described in more detail.

nga_conf:ConflatedMapFeature_CWFID_ST_FIRE_STATION_0_7 a nga_data:NGAFeature ;

        ows:hadGeometry

        nga_conf:ConflatedMap_Position_8df3c ;

        ows:hadProperty
        nga_conf:ConflatedMap_geoNameCollection_memberGeoName_fullName_e08c0 .
```

- **How individual conflated features relate to sources**

Table 8
Relationship between result and input features is described by the prov:wasDerivedFrom predicate for newly created features and the owl:sameAs predicate for unchanged features.

```
#    The newly added features in NGA map is related to source.

nga_conf:ConflatedMapFeature_CWFID_ST_FIRE_STATION_0_7

      prov:wasDerivedFrom usgs_data:USGSFeature_CWFID_ST_FIRE_STATION_0_7 .

#    There features that have not suffer any changes.

nga_conf:ConflatedMapFeature_StructurePoints_84356

      owl:sameAs nga_data:NGAFeature_StructurePoints_84356 .
```

- **How individual conflated feature properties relate to sources**

Table 9
Relationship between properties is captured by the prov:wasDerivedFrom and wasRevisionOf predicate.

```
#    The newly added feature property (position) in NGA map is related to source.

nga_conf:ConflatedMap_Position_8df3c

      prov:wasDerivedFrom usgs_data:USGSPosition_5b6aa ;

      prov:wasRevisionOf nga_data:NGAPosition_8df3c .

#    The newly added feature property (fullname) in NGA map is related to source.

nga_conf:ConflatedMap_geoNameCollection_memberGeoName_fullName_e08c0

      prov:wasDerivedFrom usgs_data:USGS_name_bf94c ;

      prov:wasRevisionOf    nga_data:NGA_geoNameCollection_memberGeoName_fullName_e08c0.
```

- **Relations between individual features and individual executions**

Table 10
Relationships between the execution and the used input features are established.

```
#    The conflation execution uses NGA and USGS features as a sources.

f2n:52N_ConflationExecution_9a6c2

      prov:used usgs_data:USGSFeature_CWFID_ST_FIRE_STATION_0_7 ;

      prov:used nga_data:NGAFeature_StructurePoints_84356 ;
```

Table 11
New entities generated are related to individual executions.

```
#    The conflated features were generated by a concrete execution.

nga_conf:ConflatedMapFeature_StructurePoints_84356

      prov:wasGeneratedBy f2n:52N_ConflationExecution_9a6c2 .

#    The conflated features were generated specific time.

nga_conf:ConflatedMapFeature_StructurePoints_84356

      prov:generatedAtTime "2015-06-23T08:04:47"^^xsd:dateTime .
```

- **Roles for individual executions and features**

Table 12
Roles of executions and features are defined by connecting them it to a qualified relation property.

```
#    Here a ReferencedMapSource and ConflatedMapOutput roles are defined.

f2n:52N_ConflationExecution_9a6c2

      prov:qualifiedUsage [

      a prov:Usage ;

      prov:entity usgs_data:USGSFeature_CWFID_ST_FIRE_STATION_0_7 ;

      prov:hadRole ows10:ReferencedMapSource ; ] .

nga_conf:ConflatedMapFeature_StructurePoints_84356

      prov:qualifiedGeneration [

      a prov:Generation ;

      prov:activity f2n:52N_ConflationExecution_9a6c2 ;

      prov:hadRole ows10:ConflatedMapOutput ;] .

#    Other attributes of the execution are also described, including the algorithm used.

f2n:52N_ConflationExecution_9a6c2 a f2n:F2N_WPSConflationExecution .

f2n:52N_ConflationExecution_9a6c2

      prov:used f2n:Kinda_Generic_ConflationProcess_v120 .

f2n:52N_ConflationExecution_9a6c2

      prov:startedAtTime "2015-06-23T08:04:24"^^xsd:dateTime ;

      prov:endedAtTime "2015-06-23T08:04:24"^^xsd:dateTime
```

3.2.4. The usefulness of provenance

Once provenance information is captured and serialized with N3 notation, it can be exploited and used to audit the origin of elements of the geospatial dataset. Graphical representations of provenance aid the comprehension of geographical products. The Gruff 5.8.0 software (http://franz.com/agraph/gruff/) is used to interpret the triples and to generate automatically a graph of provenance (Fig. 10).

The RDF N3 language also allows the generation of SPARQL queries over provenance. SPARQL (http://www.w3.org/TR/rdf-sparql-query/) is a RDF query language that can retrieve and manipulate data stored in RDF format. Thus, provenance data can be used to select specific geospatial data. This is very useful in scenarios where datasets are updated periodically and new versions of the same dataset are generated. Several queries over provenance can be done, such as:

- Show data derived from a particular geospatial process.
- Show features conflated on a specific date (Table 13 and Fig. 11).
- Show attributes derived from a specific dataset.
- Show attributes that are new in the dataset.

4. Workflows and chain executions with PROV

The scientific community requires complex models that normally process data in a chain of executions that configure a complete workflow. Thus, there is a necessity to track and represent the provenance of all these intermediate processes, intermediate results, parameters, inputs, agents, and dates and times. For instance, we can imagine a situation such as that in Fig. 12 where there is a need to establish a safety buffer area of 50 m beside conflated roads' maps.

4.1. Chain of executions with PROV

PROV has two ways of capturing provenance of workflows: (1) by generating a chain of activities and (2) by generating a chain of entities that forms the workflow execution. Thus, following the Fig. 12 example

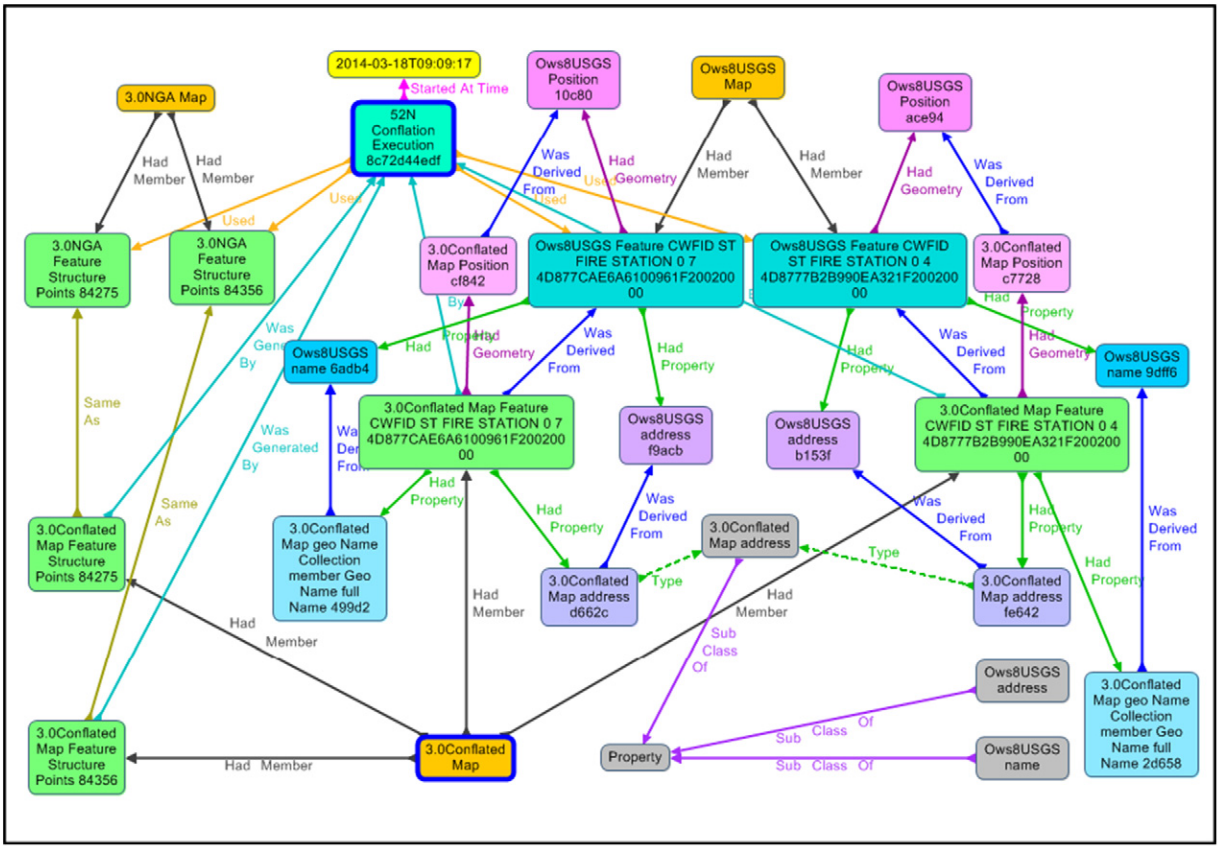

2014-03-18T09:09:17
Started At Time
3.0NGA Map
52N Conflation Execution 8c72d44edf
Ows8USGS Position 10c80
Ows8USGS Map
Ows8USGS Position ace94
Had Member
Used
Was Derived From
Had Geometry
3.0NGA Feature Structure Points 84275
3.0NGA Feature Structure Points 84356
3.0Conflated Map Position cf842
Ows8USGS Feature CWFID ST FIRE STATION 0 7 4D877CAE6A6100961F200200 00
Ows8USGS Feature CWFID ST FIRE STATION 0 4 4D8777B2B990EA321F200200 00
3.0Conflated Map Position c7728
Was Generated By
Ows8USGS name 6adb4
Had Property
Ows8USGS name 9dff6
Same As
3.0Conflated Map Feature CWFID ST FIRE STATION 0 7 4D877CAE6A6100961F200200 00
Ows8USGS address f9acb
Ows8USGS address b153f
3.0Conflated Map Feature CWFID ST FIRE STATION 0 4 4D8777B2B990EA321F200200 00
3.0Conflated Map Feature Structure Points 84275
3.0Conflated Map geo Name Collection member Geo Name full Name 499d2
3.0Conflated Map address
Type
3.0Conflated Map address d662c
3.0Conflated Map address fe642
Sub Class Of
Ows8USGS address
3.0Conflated Map geo Name Collection member Geo Name full Name 2d658
3.0Conflated Map Feature Structure Points 84356
3.0Conflated Map
Property
Ows8USGS name

Fig. 10. Provenance graph of a reduced number of entities derived from the conflation example presented above.

- The *wasDerivedFrom* relation connects (new) entities derived from previous entities generated in previous executions. A derivation is a transformation of one entity into another. This property allows the generation of chains using entities as threads (Table 14 and Fig. 13).

- The *wasInformedBy* property connects activities. This permits the description of provenance of a workflow consisting of several process steps. Activities are informed by previous activities, and this connectivity provides information on dependency without explicitly providing the activity start and end times (Groth & Moreau, 2013).

4.2. Web architecture of chain execution

4.2.1. Chain of execution architecture using WPS

A practical implementation of this workflow was written in Java using the Jena API[2] and creates provenance information in RDF. Fig. 14 shows the internal architecture of the workflow execution using WPS.

The workflow starts with the extraction of all roads (*features*) from the topographic maps. This process is done by exporting the result of a query that selects all the roads into a new road map. The operation is executed twice: for the NGA topographic map and for the USGS topographic map. At this stage, all features in each of the two new roads maps have the same origin; therefore, the dataset provenance level is enough.

The second step is a conflation process between the two road maps. During execution, the process iterates over all target features (NGA). To determine which features from the target dataset should be added, the IDs of all target features are checked against each ID of the current source feature (more detailed explanation of the conflation rules can be found in sub-section 4.2.2.3). If the target feature does not exist in the source dataset (NGA), a new empty feature is created according to the schema of the source dataset. Attributes of new features subject to a conflation rule are accordingly mapped against an attribute of the original feature or set to a fixed value. All other attribute values are set to their respective default value. The ID of the target feature is used as the ID of the new feature. The relationship between the newly created feature and the target feature is preserved by annotating this relationship in the provenance information. At this step of the workflow, feature-level provenance is needed because of the different origin of some features: Some features were extracted from the NGA dataset and others from USGS dataset. Thus, the system provides feature-level provenance in RDF.

Finally, a buffer of 50 m over all the entities of the road map is generated and exported into a new affected area map. This newly generated map inherits the need for feature-level provenance.

4.2.2. Web conflation process service

Table 13
SPARQL query to select features at a specific date.

```
select ?Feature where {

        ?Feature http://www.w3.org/ns/prov#generatedAtTime "2014-03-

        18T09:09:17"^^http://www.w3.org/2001/XMLSchema#dateTime .

        http://metadata.dod.mil/mdr/ns/GSIP/3.0/tds/3.0ConflatedMap
        http://www.w3.org/ns/prov#hadMember ?Feature }
```

[2] https://jena.apache.org/

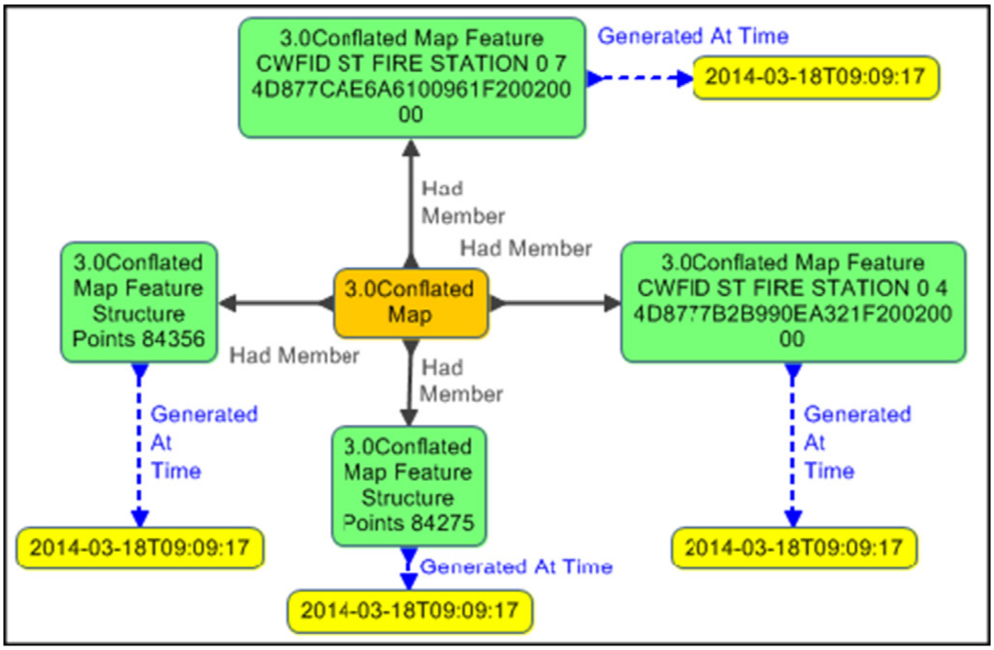


Fig. 11. Graphical representation of the results of Table 13 SPARQL query. Elements that were generated at 2014-03-18T09:09:17 are represented.

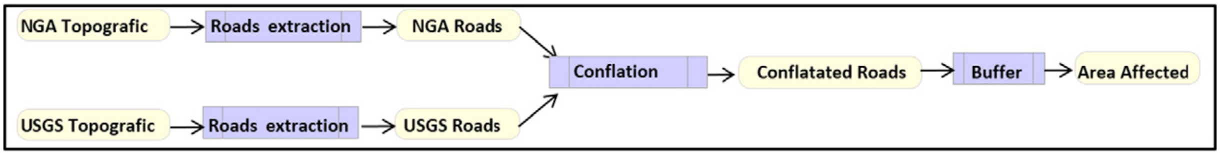


Fig. 12. Example of Geospatial Workflow determining the area affected in conflated roads.

The geospatial process is implemented as a WPS process and is internally divided into a conflation process and a provenance engine. The inputs, outputs and internals of the implemented process are described in the following sections.

4.2.2.1. User inputs. The user must enter the following information to start the process:

- Source: All features of the source dataset will be included in the conflated dataset. The schema of the data will be applied to the features derived from the target dataset. Features are expressed in GML and are passed to the process as a reference to a WFS.
- Target: Non-existing features in the source dataset will be taken from the target dataset and added to the resulting conflated dataset. As is the case with the source features, these are expressed in GML and are passed to the service as a reference to WFS.
- Rules: Section 4.2.2.3 illustrates how the *Id matching* rule was used in the implementation of the conflation processes.

Table 14
Declaration of derivation between entities in RDF.

```
#   The Area Affected entity wasDerivedFrom Conflated Roads Map entity.

            prov:Area_Affected prov:wasDerivedFrom prov:Conflated_Road_Map .

#   Conflated Road Map  wasDerivedFrom the conflation of USGS Roads map and NGA Roads map entities.

            prov:Conflated_Road_Map prov:wasDerivedFrom prov:USGS_Roads ;

                                                         prov:NGA_Roads .

#   Both maps had been derived (wasDerivedFrom) from their respective topographic maps.

            prov: prov:NGA_Roads prov:wasDerivedFrom prov:NGA_Tographic .

            prov: prov:USGS_Roads prov:wasDerivedFrom prov:USGS_Tographic .
```

4.2.2.2. Process outputs. The following outputs can be requested:

- Conflated result: The resulting conflated dataset including all features of the source dataset in addition to the new ones extracted from the target dataset is generated by the geospatial process. Features are expressed in GML.
- Provenance: The provenance information about the process and involved features and attributes is generated by the provenance engine. Provenance is expressed in RDF.

4.2.2.3. Conflation rules. Conflation rules for the conflation WPS process were encoded in JSON. With this encoding, some rules can be specified (e.g. which attribute values are to be taken over from the target features and which attribute values shall be set to fixed values). The structure of the JSON code is shown in Table 15.

An example taken from the scenario is given in Table 16:

This example only covers simple rules for conflation scenarios in which features from the target dataset not existing in the source dataset are added to the result. Extensions of this technique to more complex conflation scenarios, e.g. updating the source features/attributes on the basis of distance rules, would also be possible.

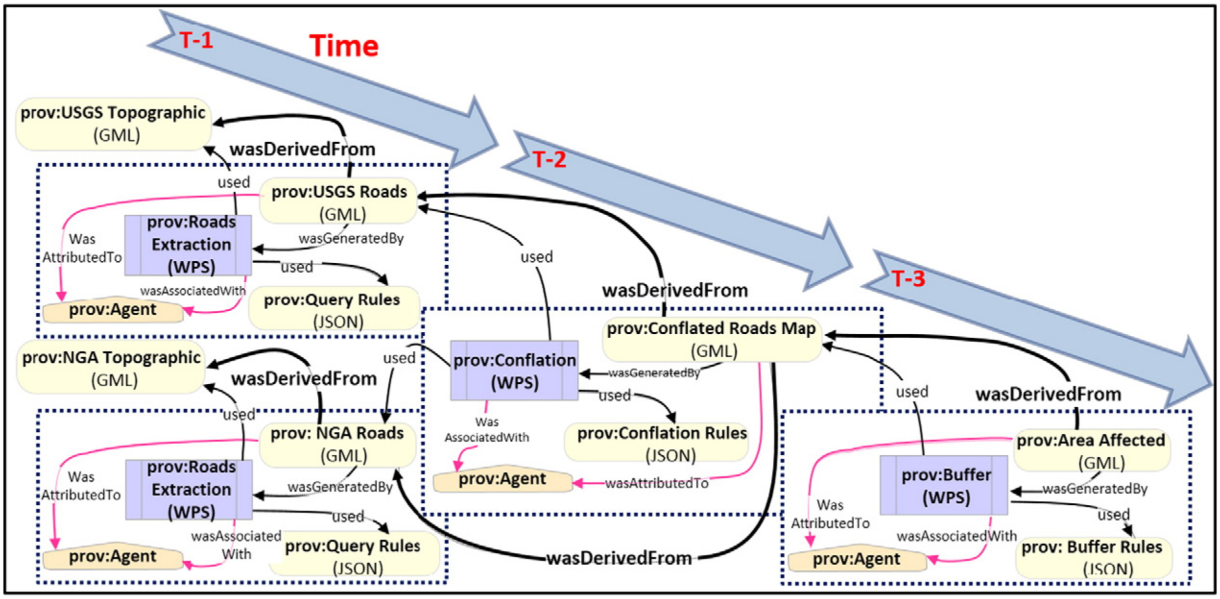


Fig. 13. Diagram illustrating that prov:*wasDerivedFrom* allows the generation of chains using entities as threads.

5. Conclusions

Because of the heterogeneity and complexity of the geospatial data derived from diverse geospatial processes, the required fineness of provenance granularity can change depending on the geospatial process requirements (e.g. conflation process and buffer execution). Thus, the provenance models should allow for the representation of lower levels of geospatial granularity and the generation of the dependencies between different levels. This paper highlights that the common mechanisms for describing provenance at the dataset, feature and attribute levels using ISO 19115 and W3C-PROV are not satisfactory.

In case of ISO 19115, its combination with GML documents has been explored, but this quickly becomes a very verbose solution that demands large amounts of computer storage space and does not entirely satisfy the requirements of attribute-level provenance. Moreover, the ISO 19115 solution requires modification in the GML 3.2 or newer application schemas, and this may require extensive community dialogue to permit a change. Thus, it is still not clear how to write provenance at the feature and attribute levels using the ISO 19115 model.

Regarding the PROV model, although this model was not originally designed for describing geospatial provenance, in this paper, we have shown a way to apply PROV for use in the geospatial domain: its modular structure, the flexibility of its semantics and the definition of relationships between different elements make it ideal to describe geospatial provenance at the dataset, feature and attribute levels. This paper has presented an application of the PROV ontology to describe provenance without introducing major changes in the PROV model; just by adding two entity property types (*hadProperty* and *hadGeometry*), the PROV model was used to connect the feature level with the attribute level.

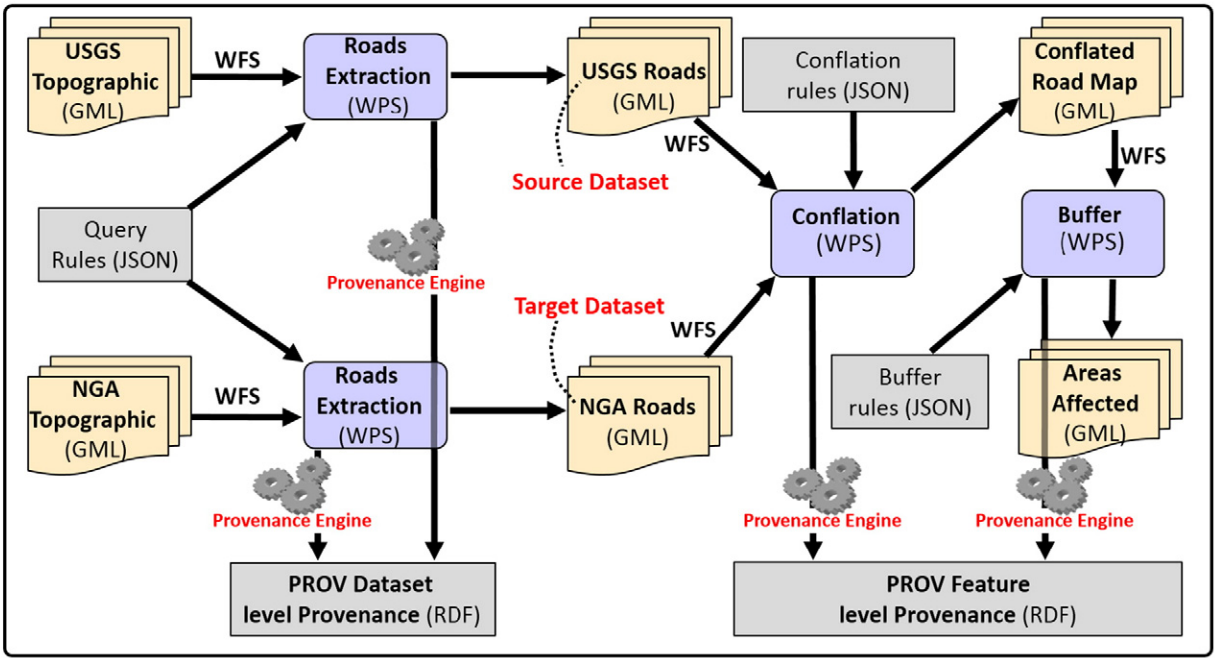


Fig. 14. Diagram of the workflow execution of the conflation process and provenance engine using WPS.

Table 15
Structure of the JSON-based encoding for conflation rules.

```
{
    "mappings": {
        "attribute-name-target":"attribute-name-source"
    },
    "fixedAttributeValues": {
        "attribute-name-target":"fixed-value"
    }
}
```

Table 16
Example of JSON-based encoding for conflation rules.

```
{
    "mappings": {
        "Road_id":"Road_id",
        "name":"geoNameCollection.memberGeoName.fullName"
    },
    "fixedAttributeValues": {
        "featureFunction-1":"roads_network",
        "restriction.NetworkAttributesGroup_resClassification" :"U"
     }
}
```

The feasibility to serialize PROV with RDF notation triples makes PROV an optimum model for the description of provenance in a distributed environment and in the linked data sphere. The combination of the presented provenance model and the WPS allows connection between the results of an analysis and the original sources. This is very beneficial when assessing the quality of results or when reproducing the workflows. In addition, the use of SPARQL enables powerful queries that involve data, metadata and provenance.

The presented example demonstrates that it is possible to use PROV to describe geospatial provenance at the three levels of granularity in a distributed environment. In addition, an example of the architecture of a workflow and the chain implementations written in Java using the Jena API shows how provenance information serialized in N3 notation language can be retrieved satisfactorily in a distributed environment.

Acknowledgements

This paper was written with the support of the European Commission through the grants H2020-641538 - ConnectinGEO, the H2020-641762 - ECOPotential and the H2020-689744 - Ground Truth 2.0; to the Consolidated Research Groups given by the Catalan Government (2014 SGR 1491); and Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under Grant GCL-2015-69888-P (ACAPI). Xavier Pons is recipient of an ICREA Academia Excellence in Research grant (2016–2020). The paper also incorporates the previous efforts done in the elaboration of the Provenance Engineering Report 14-001.

References

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—The story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22.

Buneman, P., Khanna, S., & Chiew Tan, W. (2001). *Why and where: A characterization of data provenance. In database theory—ICDT 2001 (pp. 316–330)*Berlin Heidelberg: Springer. http://dx.doi.org/10.1007/3-540-44503-X_20.

Chen, C. C., Knoblock, C. A., & Shahabi, C. (2008). Automatically and accurately conflating raster maps with orthoimagery. *GeoInformatica*, *12*(3), 377–410. http://dx.doi.org/10.1007/s10707-007-0033-0.

Di, L., Shao, Y., & Kang, L. (2013a). Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 19115-2 lineage model. *Geoscience and Remote Sensing, IEEE Transactions on*, *51*(11), 5082–5089. http://dx.doi.org/10.1109/TGRS.2013.2248740.

Di, L., Yue, P., Ramapriyan, H., & King, R. (2013b). Geoscience data provenance: An overview. *Geoscience and Remote Sensing, IEEE Transactions on*, *51*(11), 5065–5072. http://dx.doi.org/10.1109/TGRS.2013.2242478.

Díaz, P., Masó, J., Sevillano, E., Ninyerola, M., Zabala, A., Serral, I., & Pons, X. (2012). Analysis of quality metadata in the GEOSS clearinghouse. *International Journal of Spatial Data Infrastructures Research*, *7*, 352–377.

Feng, C. C. (2013). Mapping geospatial metadata to open provenance model. *Geoscience and Remote Sensing, IEEE Transactions on*, *51*(11), 5073–5081. http://dx.doi.org/10.1109/TGRS.2013.2252181.

Garijo, D., Gil, Y., & Harth, A. (2014). Challenges in modelling geosaptial provenance. *Proceedings of the fifth international provenance and annotation workshop (IPAW), Cologne, Germany* (June 9–13, 2014).

Greenwood, M., Goble, C. A., Stevens, R. D., Zhao, J., Addis, M., Marvin, D., ... Oinn, T. (2003). Provenance of e-science experiments-experience from bioinformatics. *Proceedings of UK e-science all hands meeting 2003* (pp. 223–226).

Groth, P., & Moreau, L. (2013). *PROV-overview: An overview of the PROV family of documents. Working group note, W3C.*

Goodchild, M. F., Yuan, M., & Cova, T. J. (2007). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, *21*(3), 239–260. http://dx.doi.org/10.1080/13658810600965271.

He, L., Yue, P., Di, L., Zhang, M., & Hu, L. (2015). Adding geospatial data provenance into SDI—A service-oriented approach. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, *8*(2), 926–936. http://dx.doi.org/10.1109/JSTARS.2014.2340737.

ISO 19115-1:2014 (2014). “Geographic information- metadata- part 1: Fundamentals”.

ISO 19156:2011. (2011). “Geographic information – Observations and measurements”.

Kogan, F., Powell, A., & Fedorov, O. (2011). *Use of satellite and in-situ data to improve sustainability.* Springer. http://dx.doi.org/10.1007/978-90-481-9618-0.

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

Lopez-Pellicer, F., & Barrera, J. (2014). D16.1 Call 2: Linked map VGI provenance schema. *In Linked Map subproject of planet data. Seventh framewok programe.*

Ma, X., Zheng, J. G., Goldstein, J. C., Zednik, S., Fu, L., Duggan, B., ... Fox, P. (2014). Ontology engineering in provenance enablement for the National Climate Assessment. *Environmental Modelling & Software*, *61*, 191–205. http://dx.doi.org/10.1016/j.envsoft.2014.08.002.

Masó, J., Pons, X., & Zabala, A. (2012). Tuning the second-generation SDI: Theoretical aspects and real use cases. *International Journal of Geographical Information Science*, *26*(6), 983–1014.

Masó, J., Closa, G., Gil, Y., & Prob, B. (2014). *OGC® Testbed 10 provenance engineering report OGC public engineering report.* , 1–87 Open Geospatial Consortium.

Moreau, L., & Missier, P. (2013). *PROV-DM: The prov data model. W3C recommendation.*

Myers, J., Pancerella, C., Lansing, C., Schuchardt, K., Didier, B., Ashish, N., & Goble, C. A. (2003). Multi-scale science, supporting emerging practice with semantically derived provenance. *ISWC workshop on Semantic Web Technologies for searching and retrieving scientific data* (Florida, October 2003).

Resource Description Framework (RDF): https://www.w3.org/RDF/. Accessed 2017-01-15.

Ruiz, J. J., Ariza, F. J., Ureña, M. A., & Blázquez, E. B. (2011). Digital map conflation: A review of the process and a proposal for classification. *International Journal of Geographical Information Science*, *25*(9), 1439–1466. http://dx.doi.org/10.1080/13658816.2010.519707.

Saalfeld, A. (1988). Conflation automated map compilation. *International Journal of Geographical Information System*, *2*(3), 217–228. http://dx.doi.org/10.1080/02693798808927897.

Seth, S., & Samal, A. (2007). Conflation of features. (Coord) In S. Shekhar, & H. Xiong (Eds.), *Encyclopedia of GIS* (pp. 129–132). Springer US.

Scheider, S., Gräler, B., Pebesma, E., & Stasch, C. (2016). Modeling spatiotemporal information generation. *International Journal of Geographical Information Science*, 1–29. http://dx.doi.org/10.1080/13658816.2016.1151520.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record*, *34*(3), 31–36. http://dx.doi.org/10.1145/1084805.1084812.

Cox, S. (2015). Ontology for observations and sampling features, with alignments to existing models. *Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal*, 1138–2350. http://dx.doi.org/10.3233/SW-16021.

Tilmes, C., Fox, P., Ma, X. L., McGuinness, D. L., Privette, A. P., Smith, A., & Zheng, J. G. (2013). Provenance representation for the national climate assessment in the global change information system. *Geoscience and Remote Sensing, IEEE Transactions on*, *51*(11), 5160–5168. http://dx.doi.org/10.1109/TGRS.2013.2262179.

The Open Source Geospatial Foundation (OsGeo) (2015). Starter_Dictionary. http://wiki.osgeo.org/wiki/Starter_Dictionary#Feature_Schema (Accessed 2015-09-30)

Union of Concerned Scientist (2015). USC_Satelite_Database Downloads. Retrieved from. https://s3.amazonaws.com/ucs-documents/nuclear-weapons/sat-database/3-1115+update/UCS_Satellite_Database_officialname_2-1_15.xls (Accessed 15-09-30)

Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O., & Gómez-Pérez, A. (2014). Integrating geographical information in the linked digital earth. *International Journal of Digital Earth*, *7*(7), 554–575. http://dx.doi.org/10.1080/17538947.2013.783127.

Yue, P., Gong, J., & Di, L. (2010). Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Computers & Geosciences*, *36*(3), 270–281. http://dx.doi.org/10.1016/j.cageo.2009.09.002.

Yue, P., Zhang, M., Guo, X., & Tan, Z. (2014). Granularity of geospatial data provenance. *2014 IEEE Geoscience and Remote Sensing Symposium* (pp. 4492–4495) (IEEE).