DOI: 10.1111/tgis.12555

RESEARCH ARTICLE

Transactions in GIS WILEY

A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation

Guillem Closa[1] | Joan Masó[2] | Alaitz Zabala[1] | Lluís Pesquer[2] | Xavier Pons[1]

[1]Grumets Research Group, Dep de Geografia, Edifici B, Universitat Autònoma de Barcelona, Bellaterra, 08193, Catalonia, Spain

[2]Grumets Research Group, CREAF, Edifici C, Universitat Autònoma de Barcelona, Bellaterra, 08193, Catalonia, Spain

Correspondence
Guillem Closa, Department of Geography, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona 08193, Spain.
Email: guillem.closa@uab.cat

Funding information
Catalan Government, Grant/Award Number: SGR2017 1690; European Union's Horizon 2020. ECoPotential research project, Grant/Award Number: No 641762; NEWFORLAND research project, Grant/Award Number: RTI2018-099397-B-C21/C22 MCIU/AEI/ERDF, EU

Abstract

Nowadays, there are still some gaps in the description of provenance metadata. These gaps prevent the capture of comprehensive provenance, useful for reuse and reproducibility. In addition, the lack of automated tools for capturing provenance hinders the broad generation and compilation of provenance information. This work presents a provenance engine (PE) that captures and represents provenance information using a combination of the Web Processing Service (WPS) standard and the ISO 19115 geospatial lineage model. The PE, developed within the MiraMon GIS & RS software, automatically records detailed information about sources and processes. The PE also includes a metadata editor that shows a graphical representation of the provenance and allows users to complement provenance information by adding missing processes or deleting redundant process steps or sources, thus building a consistent geospatial workflow. One use case is presented to demonstrate the usefulness and effectiveness of the PE: the generation of a radiometric pseudo-invariant areas bench for the Iberian Peninsula. This remote-sensing use case shows how provenance can be automatically captured, also in a non-sequential complex flow, and its essential role in the automation and replication tasks in work with very large amounts of geospatial data.

1 | INTRODUCTION

Provenance information, also known as lineage, is defined as the description of data origins and the processes by which a dataset is created (Buneman, Khanna, & Wang-Chiew, 2001). Provenance also includes the description of the algorithms used, their inputs and outputs, the computing environment where the process runs, the organization/person responsible for the product, and so on (Di, Yue, Ramapriyan, & King, 2013). The scientific community is interested in provenance because it provides relevant information for determining whether a product is fit for purpose and reliable. It also plays a significant role in assessing data quality and usability of the model outputs (Di, Shao, & Kang, 2013), and helps in auditing the trail of model execution, locating errors and assisting users in performing uncertainty propagation analysis (Yue et al., 2011; Zhang et al., 2017). In short, provenance allows users to determine the "what," "when," "who," "how," and "where" of the generation of geospatial data (Jiang, Kuhn, & Yue, 2017).

The tasks of preserving digital data and metadata (ISO, 2018) require contextual information (authority, process environment, software, etc.) to determine which information should be preserved to fully understand and reuse the archived data. In the case of GIS data this is a very complex task, because geospatial information is usually divided into several parts (Pons & Masó, 2016). Therefore, provenance information can be used to select the part of the information that should be preserved to ensure long-term understandability and avoid possible future geospatial data losses.

In the context of scientific models, data provenance records the workflow processing steps and the inputs or outputs that contribute to generating the final data products. Due to distributed web technology, geoprocessing tools are available as services (Di & McDonald, 1999), and a Model as a Service (MaaS) approach has recently been defined (Geller & Turner, 2007; Nativi, Mazzetti, & Geller, 2013). The task of assembling geoprocessing workflows is central to any GIS. Sharing and integrating models over the web can help organizations to save labor and computational resources by reusing methods and data (Scheider & Ballatore, 2018), thus promoting modeling research (Nativi et al., 2013).

In this paradigm, where the origin of data and algorithms has a high level of heterogeneity, several authors (Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010; Xu et al., 2010) see provenance information as even more important for inspecting and verifying quality, usability, and reliability of data. Provenance is also a central issue for dealing with remote-sensing (RS) data. RS data can be offered at different processing levels. For instance, the Copernicus (the European Union's Earth observation program) open access hub offers Sentinel-2 data at top-of-atmosphere (TOA) or bottom-of-atmosphere (BOA) reflectances. In addition, Copernicus Services offer high-level products (i.e. biophysical variables, land cover maps, etc.) obtained using specific algorithms that have been proven satisfactory for general purposes. Nevertheless, other scientific communities (e.g. regional research groups, professional associations, local developers, etc.) offer other alternative processing methods for the same outputs, which favor particular conditions (e.g. optimized for mountain areas) or allow more coherent radiometry to be obtained but require more calibration effort (e.g. determining pseudo-invariant areas), and which provide different results. Other authors go further and determine the potential of data provenance (when it is complete and points to actual data and/or metadata) for data replication (reproducibility purposes) and for workflow replication (with other inputs). Thus, provenance information can help to overcome the barrier between model providers and model users who want to reuse these models in different contexts, regions, or environments. Although the importance of provenance in the geospatial community is documented, the provenance description in geospatial products is still largely incomplete (Díaz et al., 2012). Although geospatial data usually come with some degree of provenance information, in many cases this is expressed with a simple textual description, which has a negative impact on its automated usage (Yue, Gong, & Di, 2010). According to Di, Shao, et al. (2013), there are two main obstacles that generate this situation: the lack of standards that fully describe provenance information models, thus ensuring reproducibility, and the lack of automated tools for capturing the provenance information.

An interoperable model for provenance is necessary to be able to exchange and share geospatial data provenance in a distributed information environment (He, Yue, Di, Zhang, & Hu, 2015). The geospatial community has traditionally used the ISO 19115-1 (ISO, 2014) standard to encode metadata and provenance (Masó, Closa, Gil, &

Proß, 2013). ISO 19115-2 (ISO, 2019) (initially designed as an extension for imagery) includes a model for acquisition and extends the lineage model to better capture processing metadata. Alternatively, He et al. (2015) and Jiang et al. (2018) propose extending W3C PROV (Groth & Moreau, 2013) to ISO 19115 in order to describe provenance better. Others, such as Lopez-Pellicer and Barrera (2014) and Closa, Masó, Proß, and Pons (2017) propose adapting the W3C PROV model to geospatial community requirements. However, from our point of view, there are still some issues to solve in the geospatial lineage models, such as the concrete model to capture initialization, and its basic assumptions and parameter values. These deficiencies prevent the complete description of provenance and obstruct workflow replication and data reproduction tasks. In the present article, we propose combining the Web Processing Service (WPS) standard with ISO lineage models (*LI_Lineage* and *LE_ProcessStep*) to describe provenance more precisely. The ISO models make it possible to describe provenance as a succession of processes, while the WPS schemas permit capturing the inputs and the algorithm used with a higher level of detail.

Besides the data model chosen to represent provenance, applications also need to ensure provenance capture, management, and retrieval (Miles, Groth, Branco, & Moreau, 2007). Thus, automatic tools that capture and store provenance in the metadata information are needed. Some existing workflow systems have been extended to support the capture and query of provenance, such as Kepler (Altintas, Barney, & Jaeger-Frank, 2006). Yue et al. (2011) demonstrate how geospatial services in spatial data infrastructures (SDI) can also be extended to share geospatial data provenance in the web environment. In this article we describe how we have implemented a provenance engine (PE) that automatically captures provenance information. This tool, developed in the framework of the MiraMon GIS & RS software (Pons, 2019), collects the provenance from each individual tool execution. MiraMon has the GeMM metadata editor, which is capable of graphically representing and handling the accumulated provenance information of all the tools executed in a geospatial workflow.

Most work has focused on analyzing and capturing provenance information that is created during execution, rather than on metadata generated before execution (Kim, Gil, & Ratnakar, 2006). This results in a linear description of the steps followed to generate the result. In this approach, provenance information about previous experiments, repeated iterations to obtain the correct parameters, or discarded executions are not recorded. Nevertheless, the data associated with e-science experiments have less value if other scientists are not able to access the previous tests made with these data (Greenwood et al., 2003). The current article claims that it is necessary to document the discarded executions or previous iterations as a part of current provenance information about a dataset. It is proposed to extend the potential of ISO models to capture the complete history of e-science experiments.

The remainder of this article is organized as follows: in Section 2, we identify some strengths and weaknesses of ISO and WPS models; Section 3 introduces the solution adopted to better describe provenance and the assets accomplished with this model; Section 4 describes how the system captures provenance and how it is represented. Section 5 provides a discussion based on a use case that exemplifies the usefulness of our proposal. Finally, we summarize our conclusions and identify future work in Section 6.

2 | ISO AND WPS TO DESCRIBE PROVENANCE

2.1 | WPS *describeProcess* and *Execute* documents to capture provenance

WPS is a standard protocol developed by the Open Geospatial Consortium (OGC) that makes it possible to execute remote geospatial processes on the web. The WPS interface provides a standard way to encode inputs and outputs for each of the geospatial processes offered in a service, as well as the specific input and output of each execution (OGC, 2010). WPS instances are exposed via HTTP-GET, HTTP-POST, and SOAP (Box et al., 2016) internet protocols. The potential of geoprocessing applications supported by WPS allows for application in a wide range of fields and sectors (Michaelis & Ames, 2009). In particular, it has been implemented successfully for environmental models (Castronova, Goodall, & Elag, 2013; Granell, Díaz, Schade, Ostländer, & Huerta, 2013) and in combination with other standards: WPS+OpenMI (Goodall, Robinson, & Castronova, 2011), WPS+WCS (Yu et al.,

2012), WPS+WFS (Meng, Xie, & Bian, 2010), WPS+SWE (Jirka, Nüst, & Proß, 2013), and WPS+SOS+WFS (Pesquer Mayos, Jirka, Stasch, Masó Pau, & Arctur, 2016). Its main properties are remote execution and support of multiple input and output formats. The description of the individual processes, as well as the input and output values, is made in a generic way and can be used independently of the remainder of the standard. In practice, this means that WPS process description can be applied to any processing tool (e.g. a command line application), even if it is not part of a distributed environment.

WPS has three main operations: *getCapabilities*, *describeProcess*, and *Execute*. The three operations use the eXtensible Markup Language (XML) to encode requests and responses. Like any other OGC web service, it starts with a *getCapabilities* that includes the service metadata as well as the list of available processes.

The *describeProcess* is the operation that allows a client to request and receive a *response* with detailed information about a process that can be run on the service instance, including the inputs required and the outputs that can be produced (OGC, 2010). Inputs and outputs can be simple types expressing isolated numbers (called LiteralData), complex types (e.g. a geospatial file format) (called ComplexData), or extents (called BoundingBox). The *Execute* operation allows WPS clients to run a specified process implemented by a server, which returns the produced output values. The *Execute request* document contains the elements that identify the process that will be executed, as well as the exact data input values.

As provenance information contains the description of processes and sources, *describeProcess response* and *Execute request* documents can be used to extract information about provenance information or even store that information. Applying the same descriptions to local executions makes it possible to capture provenance automatically in a desktop GIS local framework in a standard way, and increase the completeness of the documented provenance information. Specifically, *describeProcess* documents are used to create structured documentation on how individual command-line tools work and to automatically inherit detailed descriptions of each parameter from the documentation. *Execute request* document fragments are used to capture the actual values of each execution. More details about the use of WPS to record provenance are given in Section 3.

2.2 | ISO provenance model

ISO 19115-1 and 19115-2 are commonly encoded in XML. In fact, ISO 19115-3 provides the XML implementation schema for ISO 19115-1 and 19115-2, and can be used to describe, validate, and exchange geospatial metadata. The ISO metadata standards provide a lineage model based on *sources* which are either used or produced in a series of *process steps* (*LI_Lineage* and *LE_ProcessStep*). Sources and process steps are linked together to describe the lineage of a resource. The lineage models of ISO allow the provenance information to be described in three different ways:

1. A list of process steps and a separate list of sources.
2. A list of all the sources used and then an added description of all the processes as child processes.
3. A list of all the process steps that use sources.

According to Díaz et al. (2012), describing provenance with a list of processes that use some sources is the best way to make a complete record of provenance, because it follows the workflow execution order. If it is used recursively, it can capture the complete provenance sequence. Therefore, the MiraMon metadata model uses ISO 19115 in this way because it permits the provenance of a workflow to be described fully as an ordered succession of different process steps (Figure 1).

However, as mentioned above, there are some limitations in the ISO 19115 lineage models that inhibit the reproducibility of geospatial data that use provenance information. For instance, the only way to record execution parameters that are not geospatial sources (e.g. LiteralData) is to provide them jointly as text in runTimeParameters (e.g. a sequence of key–value pairs). In ISO 19115 models there is no way to indicate them separately,

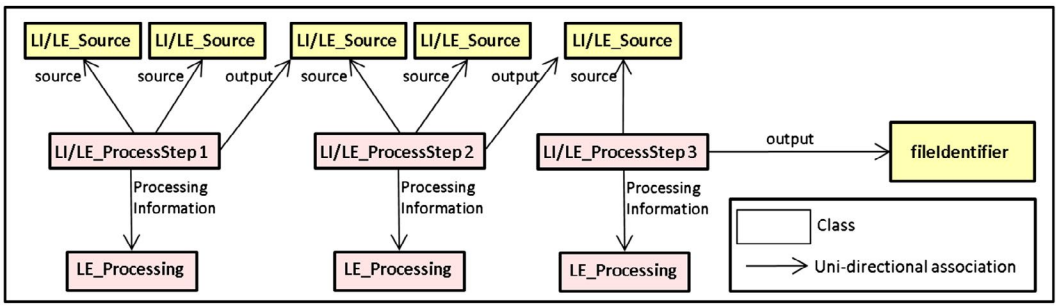

FIGURE 1 ISO 19115 and 19115-2 lineage model permits us to depict the complete sequence of the workflow

preventing the inclusion of each parameter characteristic—such as a description, the data type, and the direction (input or output).

3 | PROPOSED PROVENANCE MODEL

3.1 | Combining WPS and ISO to describe provenance

In order to overcome some of the ISO provenance model gaps, we propose the combination of the ISO provenance schemas (*LI_Lineage* and *LI_ProcessStep*) with WPS *Execute response* documents). In general, a process receives a list of inputs, some of them being geospatial datasets and others being numerical or alphanumerical parameters. In the ISO 19115 model, the *LI_ProcessStep* class has an attribute that is a composition of the *LI_Source* class that is ideal to represent input geospatial datasets but ignores other types of parameters. In Figure 2 we can see that by adding the *Input* element coming from WPS *Execute response* (in yellow) to the ISO model (in pink), we can describe provenance by a complete list of inputs of each process step. This way, for each input we capture, among other characteristics, the identifier (code) and the data type (see *literalValueChoice*) that can be a WPS *literal* or an ISO *LI_Source*. In our implementation, we link each input to its original description coming from the WPS *describeProcess response* document to add the meaning of the process inputs.

3.2 | MiraMon provenance model

In the context of the MiraMon GIS, combining ISO provenance schemas with WPS makes it possible to describe the algorithms used automatically, the processing steps, the execution dates, the data type, the units (when necessary), and data values of all parameters. Table 1 shows the correlation between the provenance elements contemplated in the MiraMon metadata model and the provenance elements in ISO 19115-1 and ISO 19115-2. The left column of the table shows the provenance elements of ISO 19115 captured by PE. The right column describes the origin of these elements: those coming from the ISO model are written in red, those coming from the WPS standard are written in blue, and the solutions natively adopted in the MiraMon metadata model are written in black. Some of the assets accomplished with this model are as follows.

3.2.1 | Source order and direction

The order of the parameters or sources might be important, but there is no place to specify this order following the actual standards. To solve this issue, we use the optional tag *ows:metadata* of WPS *describeProcess response*. Concretely, we add an incremental number to each source with this optional tag, which corresponds to the source's position in the command line (e.g. *ows:Metadata xlink:title="Param01"*).

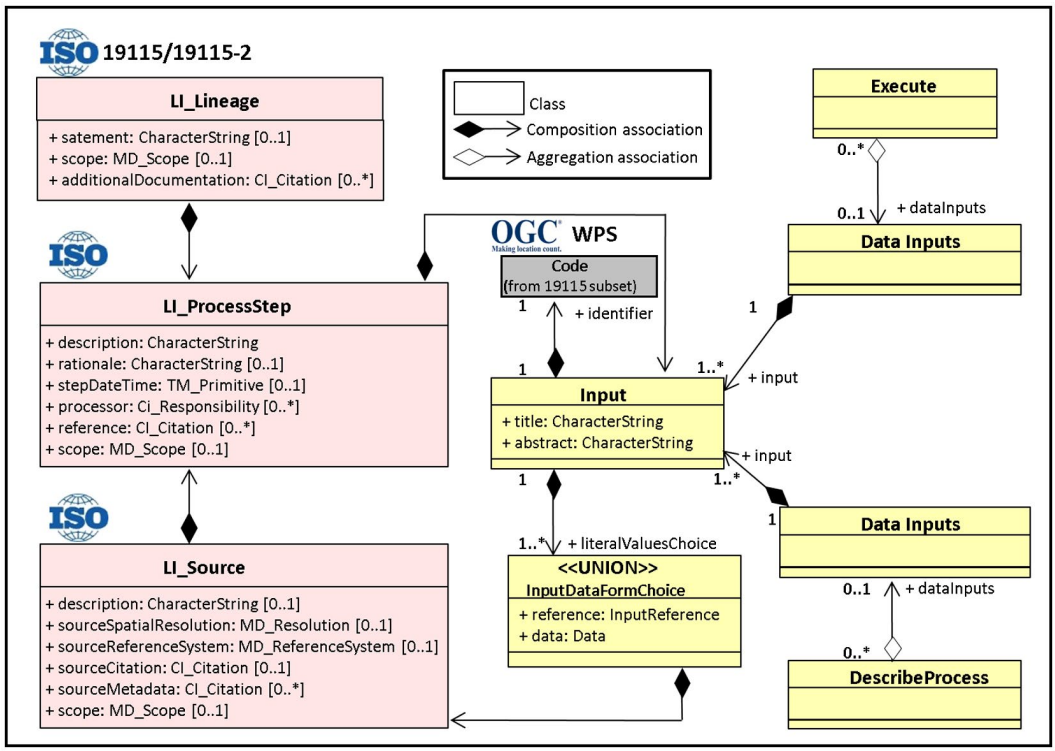

FIGURE 2 ISO 19115 *LI_Lineage* describes provenance as a sequence of *LI_ProcessStep* that uses *LI_Source*. The information contained in *LI_Source* is extended with the use of some WPS elements (UML class diagram)

To define if a source is an input or an output, we use the current WPS *describeProcess* response tags (\DataInputs\Input\LiteralData; \ProcessOutputs\Output\LiteralOutput). However, there are sources or parameters that become an output (in/out) after the execution. For this purpose we have used the tag to define the order: when a source (*Input\ows:Metadata xlink:title="ParamIdentifierX*) becomes an output, it is written again as an output (*Output\ows:Metadata xlink:title="ParamIdentifierX*) but using the same *xlink:title*.

3.2.2 | Literal data value description

As already stated, literal values of parameters are recorded using WPS *describeProcess*. Concretely, \DataInputs\Input\LiteralData in the case of data inputs; \ProcessOutputs\Output\LiteralOutput in the case of outputs. In addition, the detected gap (no placeholder to define the data type or the value used for literal data) was introduced as a change request for the ISO 19115-2 work item, and we worked in the TC211 meetings with the editors to extend the standard in this direction. The new ISO 19115-2 revisions support this request. Thus, the new ISO TC211 lineage model captures source as well as literal values through the addition of the element *LE_ProcessParameter* (Figure 3). This means that lineage information captured and represented in the MiraMon metadata manager (GeMM) is ISO compliant, thus becoming a reference implementation of ISO 19115-2:2019.

3.2.3 | Capturing scientific experiments, previous iterations, or discarded executions

As part of the scientific process, it is important for researchers to be able to verify the correctness of their own experiments, or to review the correctness of their peers' work (Miles et al., 2007). Validation ensures that results generated from experiments are meaningful. This is also necessary in the geospatial domain, especially when we

TABLE 1 This table shows the equivalences between the lineage elements of the ISO model and the lineage elements of the GeMM metadata model. The ISO model elements are written in red, those coming from the WPS standard are written in blue, and the solutions natively adopted in the MiraMon metadata model are written in black

LI_Lineage (ISO 19115-2:2017)		Lineage Data Model of GeMM
Lineage information::LI_Lineage		
statement		LI_Lineage\statement
LI_ProcessStep/LE_ProcesStep		**wps:ProcessDescriptions\wps:ProcessDescription**
description		\ows:Abstract
rationale		LI_ProcessStep\rationale
stepDateTime		LI_ProcessStep\stepDateTime
reference		LI_ProcessStep\reference
scope		LI_ProcessStep\scope
LE_Processing		**wps:ProcessDescriptions\wps:ProcessDescription**
identifier		\ows:Title
softwareReference		LI_ProcessStep\LE_Processing\softwareReference
procedureDescription		\ows:Abstract
documentation		\ows:Abstract
runTimeParameters		N/A
otherProperty: iteration=	satisfactory	LE_Processing\otherProperty :iteration=satisfactory
	discarded	LE_Processing\otherProperty :iteration=discarded
source		DataInputs\Inputs\ComplexData
output		ProcessOutputs\Output\ComplexOutput
source/output		input\ows:Metadata xlink:title="ParamX" | Output\ows:Metadata xlink:title="ParamX" *
source/output position		@xlink:title="ParamX" *
LE_Algortihm		\ows:Identifier
CI_Citation\		<ows:Metadata xlink:title="Title:
	Onlineresource	LE_Algortihm\Ci_Citation\Onlineresource
Description		<ows:Metadata xlink:title="Abstract:
LE_ProcessParameter		**\DataInputs\Input\LiteralData | \ProcessOutputs\Output\LiteralOutput**
name		..\Title
resource		..\Identifier
description		..\Abstract
optionally		..\@minoccurs
repeatibility		..\@maxoccurs
value Type		\ows:DataType
value		\wps:Execute_request\...\ows:Value
direction:LE_ParameterDirection=	in	DataInputs\Inputs\LiteralData
	out	ProcessOutputs\Output\LiteralOutput
	in/out	input\ows:Metadata xlink:title="ParamX" | Output\ows:Metadata xlink:title="ParamX"
LE_Parameterposition		@xlink:title="ParamX"
LI_Source/LE_Source		**DataInputs\Input\ComplexData | DataInputs\Output\ComplexOutput**
name		..\Title
\@id		..\Identifier
description		..\Abstract
sourceCitation		LI_Source/sourceCitation
sourceMetadata		LI_Source/sourceMetadata

** the root of this tag is \DataInputs\Input\LiteralData | \ProcessOutputs\Output\LiteralOutput*

are working in Big Data environments and in scientific contexts where we should repeat and replicate processes and results frequently.

As pointed out, ISO 19115 captures provenance as a succession of process steps, *LI_ProcessStep*. In ISO19115-2, *LI_ProcessStep* was extended into *LE_ProcessStep*, adding details of the algorithm and software used for processing (*LE_Processing* and *LE_Algortihm*). However, occasionally there are executions that are not purely sequential and require some iterative flow that progressively adjusts the final result (e.g. the generation of training areas in a supervised classification: several versions of this file are usually produced in order to improve the final classification). Iterative loops (that might overwrite a dataset) are not commonly recorded in the provenance of the result. Our proposal is that these previous executions are part of the workflow itself and therefore should be recorded as additional process steps. With this purpose, the added steps can use *otherProperty* of *LE_Processing* as a flag to document that the output was not the intended result. Therefore, we can still document the discarded executions as well as the satisfactory (final) iteration. *otherPropertyType* is mapped to a *recordtype* with a single field called *"iteration"* and *otherProperty* states that *"iteration=discarded"* (default value is *"satisfactory"*) (Figure 4).

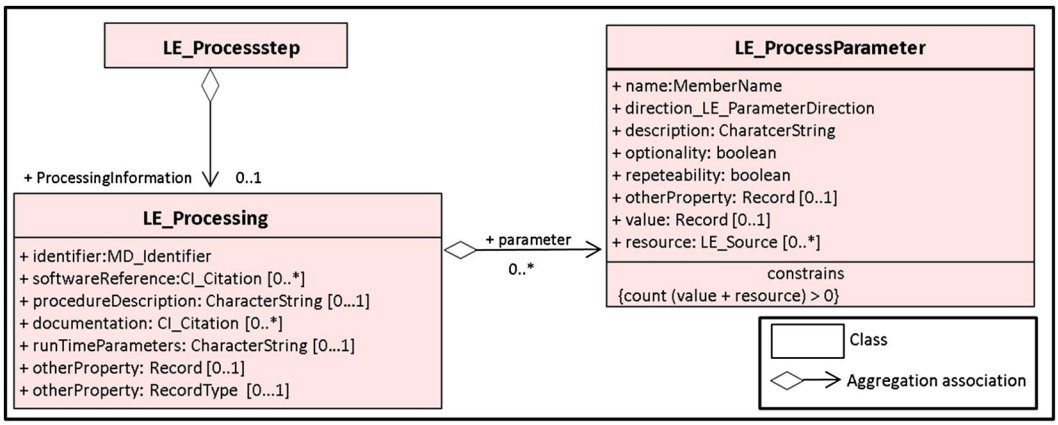

FIGURE 3 *LE_Processing* has an aggregation type: *LE_ProcessParameter*

This extension of the provenance model is useful for documenting decisions and conditions encountered during the execution of a workflow. Even if this brings provenance a bit closer to workflows, constructs such as conditional branches and loops will not be recorded because provenance only reflects the actual path followed in an execution and not all the possible alternatives. For example, in a condition presenting two options, only the selected option is recorded. However, the "iteration" extension provides a way to document branches in conditional clauses that have been tested and, despite being wrong, are needed to know the final correct path. Including typical constructs of programming languages, like conditional branches or loops, is the mission of a workflow but is beyond the scope of the current provenance models.

3.2.4 | Semantic algorithm enrichment

Examining previously recorded provenance information has potential in choosing the best-suited algorithm. However, not much research has been done to solve this issue. The current ISO 19115-2 permits us to point directly to the algorithm used and to its description (*LE_Algortihm*). Furthermore, the possibility of linking the algorithm to online resources via *CI_Citation* is provided. One possible implementation of this element can make use of citing a GIS generic vocabulary (such as https://gisgeography.com/gis-dictionary-definition-glossary), where a semantic description of the standard GIS operations is provided. Potentially, users in need of a particular algorithm can discover a variety of implementations of it in previously recorded usages, in their respective contexts, and learn which tool fits their specific case.

3.3 | Provenance exchange and interoperability

The completeness and interoperability of the model are two key aspects that should be considered when a standard is selected for describing provenance. In the MiraMon metadata model, the completeness of the provenance information has been increased by combining the original version of the ISO 19115-1 model (*LI_lineage*) and the original version of the ISO 19115-2 model (*LE_lineage*) with WPS (*describeProcess response* and *Execute request* documents). Thereby, the provenance captured automatically by the PE describes what occurred during the workflow execution more precisely than the ISO lineage standards.

The completeness achieved by combining three different models could be a handicap in terms of interoperability. The PE is able to export the provenance as an extended ISO 19139 XML document using the first version of the ISO model and the WPS elements as extensions. Other applications most probably will not use the ISO–WPS

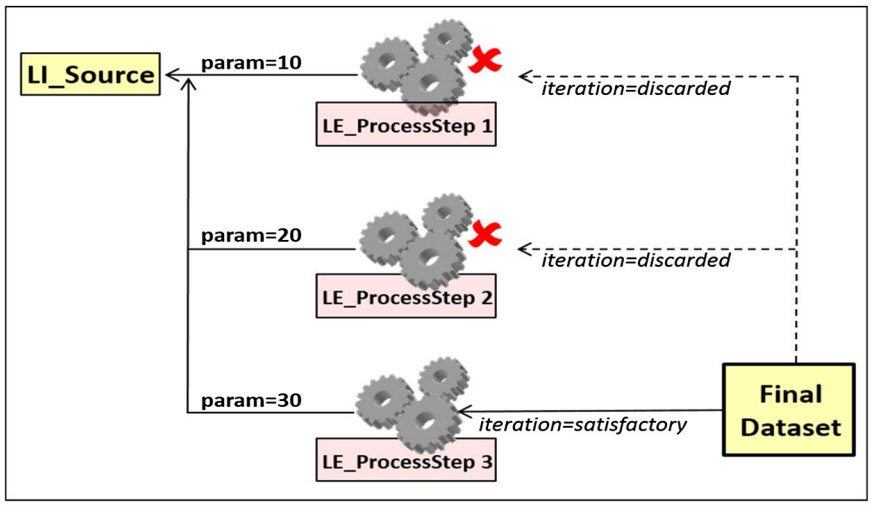

FIGURE 4 The attribute of *LE_ProcessStep*, *OtherProperty*, is used in a Boolean way to document whether the execution is satisfactory or not. In this figure the final output is generated after three iterations of the same process (different executions) with a different parameter value: the first two iterations were discarded and the third one was considered satisfactory

combinations and will not be able to fully interoperate with the generated XML. As explained in Section 3.2.2, to solve the issue of combining standards, the authors of the present work collaborated in the revision of ISO 19115-2 to include the necessary elements of WPS, mainly the description of the non-geospatial parameters that were finally mapped to the new *LE_ProcessParameter*. Later on, the conceptual encoding was included in the ISO 19115-3 XML encoding. The metadata editor is also able to generate this new XML document, which only uses the new ISO TC211 XML elements and schemas and will be more portable when other metadata tools adopt it.

4 | IMPLEMENTING PROVENANCE ENGINES

4.1 | Provenance capture in the context of MiraMon

MiraMon is GIS & RS software (Pons, 2019), free for students, universities, and so on. One of the main characteristics of the MiraMon software is that metadata are carefully managed and integrated in the dataset, which makes it possible at every processing step to program automatic decisions based on metadata information from the previous steps in the process chain (Pesquer et al., 2012). Its metadata manager, GeMM, generates metadata paying special attention to quality aspects (Zabala, Masó, Bastin, & Bigali, 2013; Zabala, Masó, & Pons, 2016), the description of the data model, and the relationships with databases. If necessary, the MiraMon metadata model can be structured in hierarchical levels (dataset item to dataset series) (Zabala & Masó, 2005). The metadata information is stored and documented in REL format documents (open native text MiraMon metadata format) or in ISO 19139 XML. In addition, as part of the quality information, there is also a place for documenting provenance information. Unlike other metadata tools, the PE maintains the dependencies with previous source datasets and ensures consistency between metadata and datasets.

The MiraMon software has more than 100 independent processing command-line applications handling different data models, mainly vector and raster layers; most of these applications can work with both data models in the same process. Some of them have already been migrated to WPS and the remainder will be migrated in the near future. For each app a *describeProcess response* document is generated, describing the process and the allowed input and output data types. *DescribeProcess response* syntax fits with the purpose of describing

command-line syntax with one exception, the order of the parameters in the command line; as already stated, this issue is solved using the optional tag *ows:metadata*.

The PE uses the generated WPS *describeProcess* template to capture, concurrently with an app execution, provenance information and store it in the metadata files (Figure 5). All captured information can be exported automatically as a batch file which collects the MS-DOS command line. This permits users to easily reproduce workflows, replicate them with different conditions (scope, data, parameters, algorithm options, etc.), and automatize executions.

The PE is a library that is shared by the visual interface of the GeMM and the MiraMon apps. It is encoded as a C library that can be linked to all GIS and RS apps. Each app uses these functions to read the metadata of the source datasets, load them, integrate them, and add the current app process step to the provenance information of the resulting dataset. In addition, GeMM's graphical interface requires a more elaborate set of functions to enrich the presentation of provenance information extracted from a *DescribeProcess* response template.

The PE writing function has two alternatives: (a) to include all lineage details—the complete sequence and description of process steps and previous data sources; or (b) to write only the last process step and link to the data sources. The generic purpose of each process step and the description of its parameters is not stored. Instead, only identifiers linking to the *describeProcess* documents are recorded. The reading function supports the two alternatives described above, and is able to read the provenance information by following the links to previous sources recursively if necessary.

The MiraMon system captures the exact parameters and values involved in an execution (which can be numbers, text strings, or bounding box data) and references them to datasets or data services. The system updates

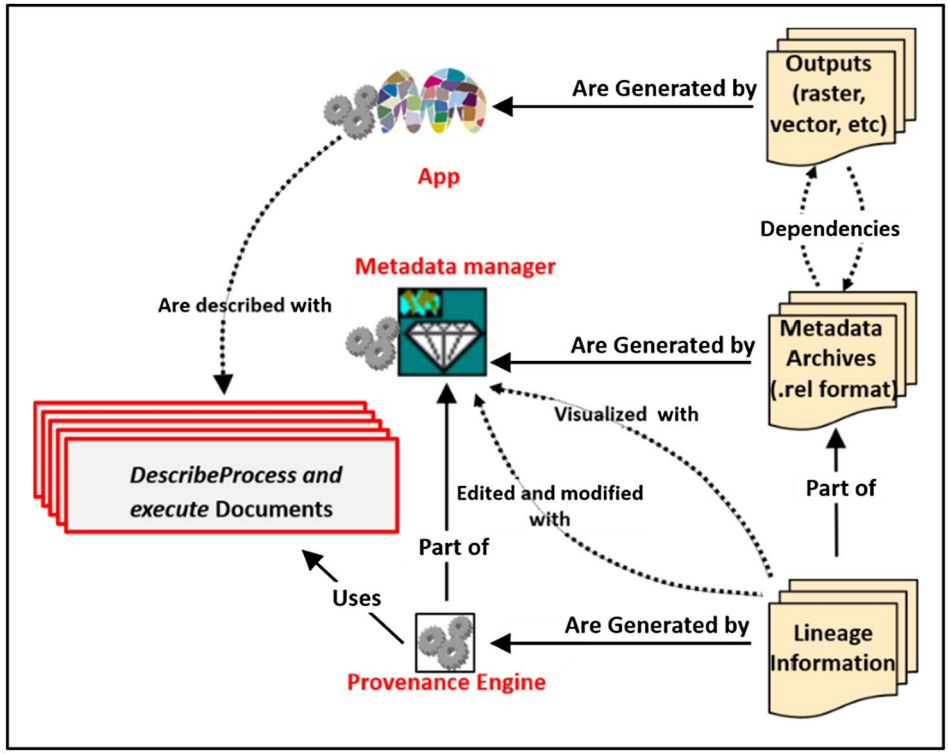

FIGURE 5 The PE uses WPS *DescribeProcess* documents to extract provenance information and then the GeMM (metadata manager) interface allows users to edit and modify the provenance of the geospatial data generated by MiraMon apps

metadata information at every intermediate step, maintaining the dependencies between the datasets and metadata files during the entire workflow execution.

As mentioned before, all the executions are recorded as satisfactory (*iteration=satisfactory*) by default. However, in a non-sequential flow where the system detects that the output of an execution already exists and is generated by the same algorithm that is being re-executed, the PE will ask the data producer to also overwrite the provenance of the last step (removing the history of the previous loops), or to keep the last execution documented as a discarded execution (*iteration=discarded*).

4.2 | Provenance editing and visualization in GeMM

In complex environments, scientists rely on visualization tools to help them understand large amounts of data that are generated from experiments (Salton, Allan, Buckley, & Singhal, 1994). Beyond the models used to capture and store provenance, an effective visualization of provenance is also necessary to understand and evaluate data and the processes involved (Kunde, Bergmeyer, & Schreiber, 2008). According to Steele and Iliinsky (2010), there are two categories of data visualization: *Exploratory*, designed to support researchers who are not certain about what is in the data; and *Explanatory*, when a researcher is trying to explain the data to someone else. This differentiation also refers to the contraposition of the "*data user needs*" compared to the "*data producer needs*," where the user requires more exploratory visualization tools, while the producer requires more explanatory information. In addition to these two viewing approaches, a review and edit functionality can help the producer to supplement the information captured automatically with extra details. Thus, a graphical interface in a provenance representation tool should fit the three purposes: exploratory, explanatory, and editing.

The GeMM graphical interface (Figure 6) presented in this article helps data users to navigate and interpret provenance. The tool represents the provenance information of a top-level dataset as a list of processes. Each process has an indented list of all the parameters used and all the outputs generated. At the same time, some parameters of the workflow (mainly the data sources) are derived by previous processes (child process), which are represented at a deeper level with their own indented (set in from the margin) list of parameters used, and so on. Thereby, the structure of the provenance schema increases progressively in profundity. From our point of view, this tree-like provenance structure is a suitable way to visualize the provenance information because it graphically represents the flow and dependencies of a specific chain of processes. The exploratory mode is facilitated by the left-hand-side tree view, while the explanatory mode is provided with the extra information of each node of the tree in the right window.

The GeMM graphical interface also allows provenance information to be edited by adding or deleting child processes or child parameters in a geospatial workflow. Moreover, the algorithm description, the processing steps carried out, the execution dates, the responsibility of the product, and the order of the processes can be edited and adapted to each scenario if necessary. This allows data producers to complete or adjust the provenance description that was automatically captured during the processes or workflow execution. By default, the provenance of a data source is linked to the provenance of a previous source; however, this has the disadvantage that if the source is removed, the provenance tree is broken and some part of the provenance is also lost. In editing the provenance, the producer can decide to embed the source provenance in the dataset description instead of linking it, thus ensuring its preservation.

5 | DISCUSSION

In order to discuss the capabilities of the presented solution, the system was tested against a real use case. Concretely, we tested with a workflow to detect pseudo-invariant areas in remote sensing. Even if the use case is mainly focused on raster data, the presented implementation has also been tested for vector data and for raster

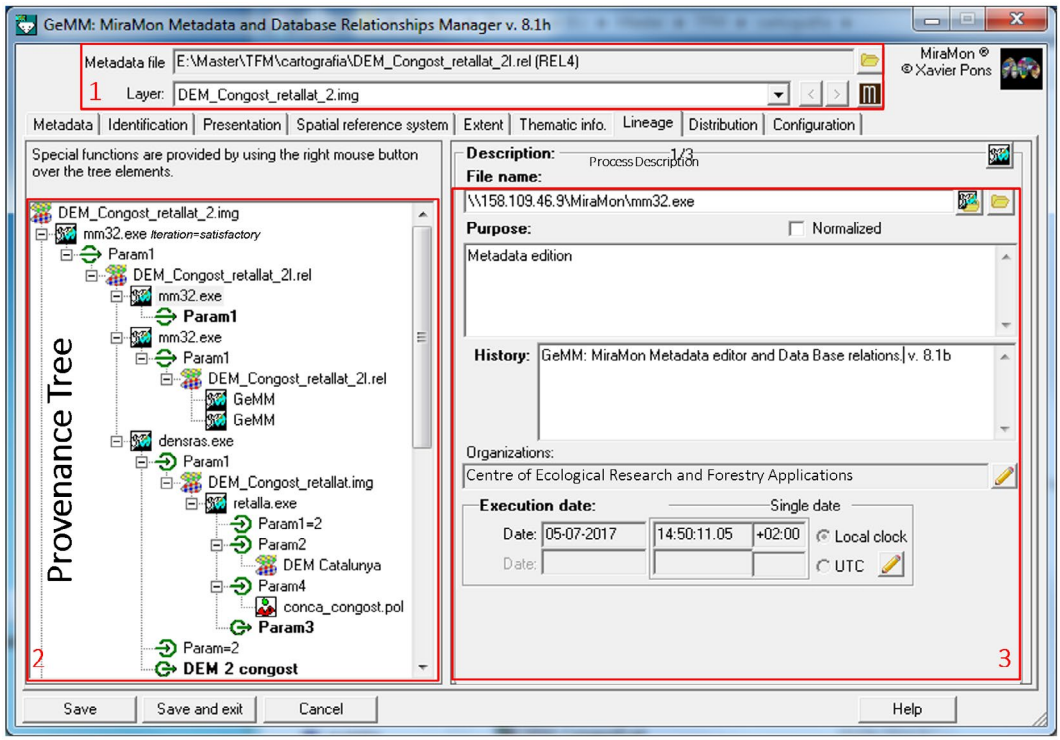

FIGURE 6 The GeMM graphical interface: (1) the path of the metadata file and the name of the geospatial file; (2) the exploratory mode with a tree including all processes and sources used in the history of the creation of the dataset; (3) the explanatory mode to view or edit the attributes of each source or process—attribution, execution date, process description, execution description, and so on

and vector data in the same process. We consider this an example of how the management of metadata information, and specifically provenance information, can be used to efficiently manage Big GeoData projects.

5.1 | Use case: Pseudo-invariant detection areas

Pseudo-invariant areas (PIAs) are used to deduce atmospheric effects in images captured by passive sensors in the solar spectrum (Pons, Pesquer, Cristóbal, & González-Guerrero, 2014; Padró et al., 2017). The idea is that radiance captured by satellite sensors varies due to changes in the Earth's surface, such as land cover phenology dynamics, land cover changes, and so on, but also due to other conditions (illumination angle, atmospheric conditions, etc.). To be able to separate land cover response (the most common interest) from other factors, it is useful to find areas where reflectance is almost invariant. These PIAs can be used in algorithms to remove atmospheric effects (Hadjimitsis, Clayton, & Retalis, 2009), which allows us to obtain not only better radiometric corrections for improved land cover classifications, but also images that are physically comparable. In addition, highly coherent time series can be generated from remote-sensing data (Vidal-Macua, Zabala, Ninyerola, & Pons, 2017).

Pesquer et al. (2012) proposed a methodology to generate an extensive bench of PIAs using the Terra-MODIS (MODerate resolution Imaging Spectroradiometer) MOD09GA daily surface reflectance product (Vermote & Kotchenova, 2008). This cited methodology has now been applied to the four MODIS tiles (h17v05, h17v04, h18v05, h18v04) that cover the Iberian Peninsula (IP) (Figure 7). The bench of PIAs is generated using 17 years (2000–2016) of daily MODIS products, specifically the bands numbered 1, 2, 3, 4, 6, and 7 of the solar spectrum

(visible, near-infrared, and short-wave infrared). There are more than 850,000 MODIS images for this period and area in the NASA archives. In addition, in order to ease the workflow execution and to better regionalize the spatial pattern analysis, each MODIS tile has been divided into smaller sub-tiles of 100 × 100 km. Thus, there are 81 scenes corresponding to the IP.

5.1.1 | Data and workflow description

The methodology is based on selecting a subset of high-quality images and defining a threshold of low deviation values (Pesquer, Domingo, & Pons, 2013). The selection of the highest-quality MODIS images combines the quality assessment of USGS (Roy et al., 2002) with a geostatistical spatial pattern analysis (Pesquer, Domingo, & Pons, 2019). Throughout the workflow execution and depending on the results obtained, a block of steps might need to be re-executed more than once. This loop of steps iterates parts of the workflow until proper results are generated. In addition, the entire workflow is replicated for each MODIS sub-tile generated. The complete workflow (Figure 8) has the following steps:

0. Data preparation: import and clipping the images to the 100 × 100 km sub-tiles.
1. Accurate topographic correction of the MOD09GA product.
2. Total mask generation from quality assessment (QA) of the MOD09GA: *topographic mask* and *geometric quality mask*.
3. Application of the total mask to the result of step 1 for bands 1, 2, 3, 4, 6, and 7 of the solar spectrum (Wang, Zeng, Li, & Shen, 2011).
4. Image classification depending on the number of non-valid pixels: a first subset of images that contains a very high ratio of valid pixels and a high ratio of valid pixels. If there is a low number of daily images containing a very

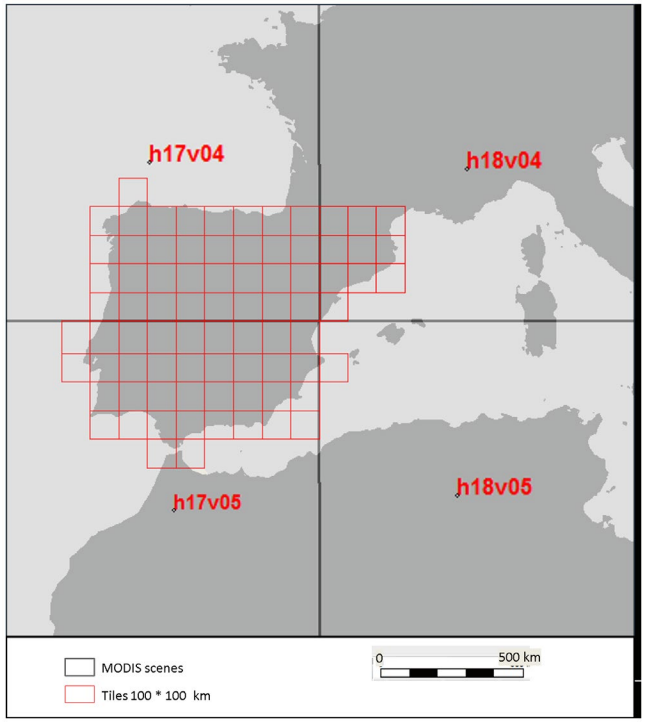

FIGURE 7 Main MODIS tiles and 100 × 100 km sub-tiles over the Iberian Peninsula

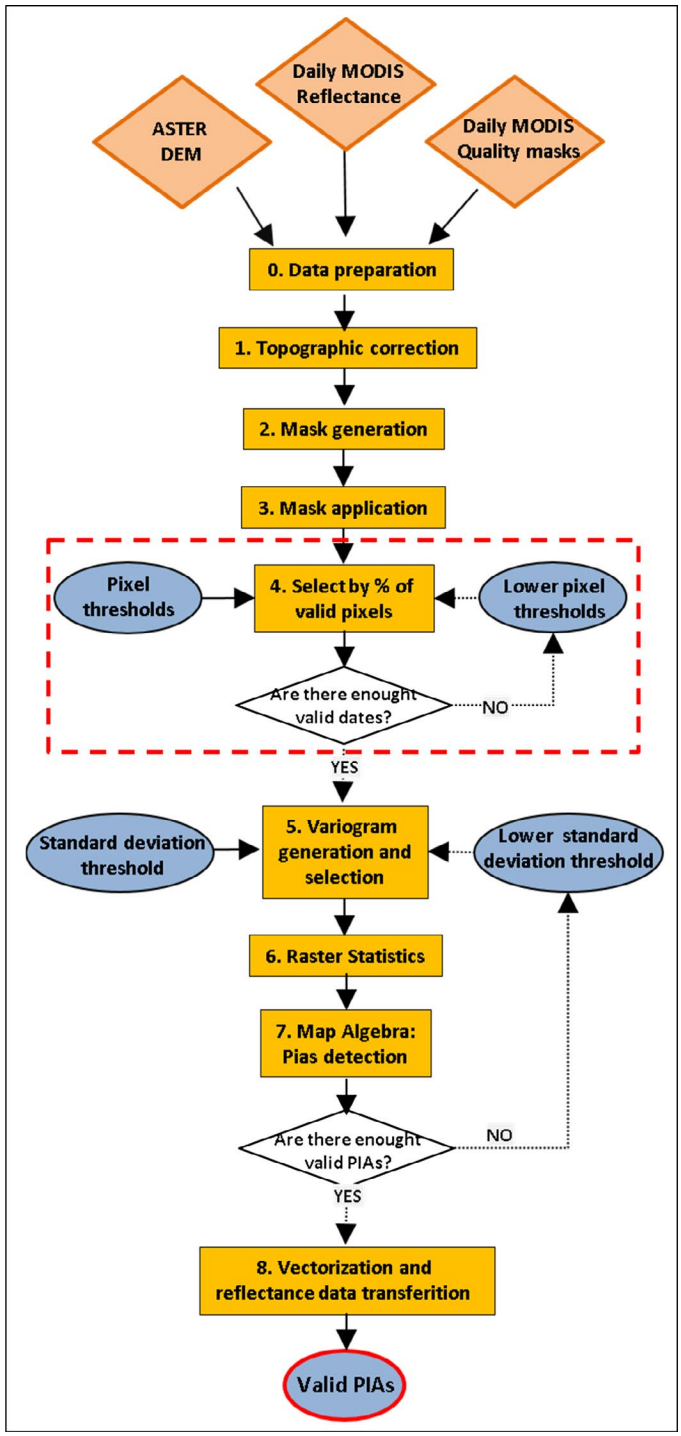

FIGURE 8 Complete workflow of PIAS generation. The dashed red lines mark the part of the workflow used to exemplify how provenance is captured in the Section "Replication with different pixel thresholds"

high ratio and/or a high ratio of valid pixels, progressively lower the threshold until sufficient images have been selected.

5. Selection of high-quality images by comparing the spatial pattern model and each daily image. Select the subset that presents a variogram structure with parameters within predefined thresholds (Kitanidis, 1997).
6. Calculation, for the highest-quality subset of image series (2000–2016) of standard deviation, the number of valid images and the average reflectance value for each pixel and band.
7. Application of map algebra to all bands in order to select the near-invariant pixels that are considered as PIAs. If the number of PIAs is very low, increase the threshold deviation.
8. Vectorization and transfer of the average reflectance value from the high-quality subset. Each vectorized contiguous group of pixels results in a PIA entity.

5.1.2 | Provenance retrival and management

Figure 9 illustrates the provenance information captured by the PE during the PIA workflow execution. In order to shorten the explanation, we will concentrate on a single step of the PIA workflow execution (inside red-dashed rectangle of Figure 8). Specifically, we focus on step number 4 (select by % of pixels using a Histoselection.exe app) to demonstrate how provenance information is captured, stored, and then how it can be used. The fragment has been selected because it includes a re-execution loop with different parameters until proper results are generated.

Replication with different pixel thresholds

After applying the total quality mask over the MOD0GA daily surface reflectance product to the six abovementioned bands, the next step is the generation of a list with higher-quality daily images. Therefore, images are selected depending on the percentage of valid pixels compared to the total number of pixels of the image:

- Images with at least the inferior threshold (by default 75% of pixels) of valid pixels. These images, written in the *High ratio valid pixels list*, are used to generate the PIA bench.
- Images with more than the superior threshold (by default 90% of pixels) of valid pixels. These images, written in the *Very high ratio valid pixel list*, are also used to generate the PIA bench, and to obtain a representative variogram of the area, the *variogram model*.

The invariant (in time) property of the PIA is not guaranteed with a poor subset of quality images. In some regions, depending on the particular regional climatic and atmospheric conditions, there are not sufficient representative dates to generate PIAs (at least 10% of the time series), or to create the variogram (at least five or six images are needed), using the default threshold values. In these cases, it is necessary to lower the thresholds and repeat the process in order to increase the statistically representative set of selected daily images (Figure 10).

Provenance capture and description

The metadata editor represents provenance as a process source-oriented tree. In each process the list of all parameters used and the outputs generated are shown. Processes can be tagged as discarded (*iteration=discarded*) by data producers and no output is generated, or can be considered satisfactory (*iteration=satisfactory*) and an output is generated. However, when producers consider it appropriate, the PE also saves all provenance related to discarded iterations.

Table 2 shows the provenance documented during the generation of the *High_ratio_pixels.lst* and *Very_high_ratio_pixels.lst*, and Figure 11 shows the provenance tree generated by the PE. It can be seen that the HistoSelection.exe app is executed three times until the results are considered satisfactory according to the intrinsic scientific

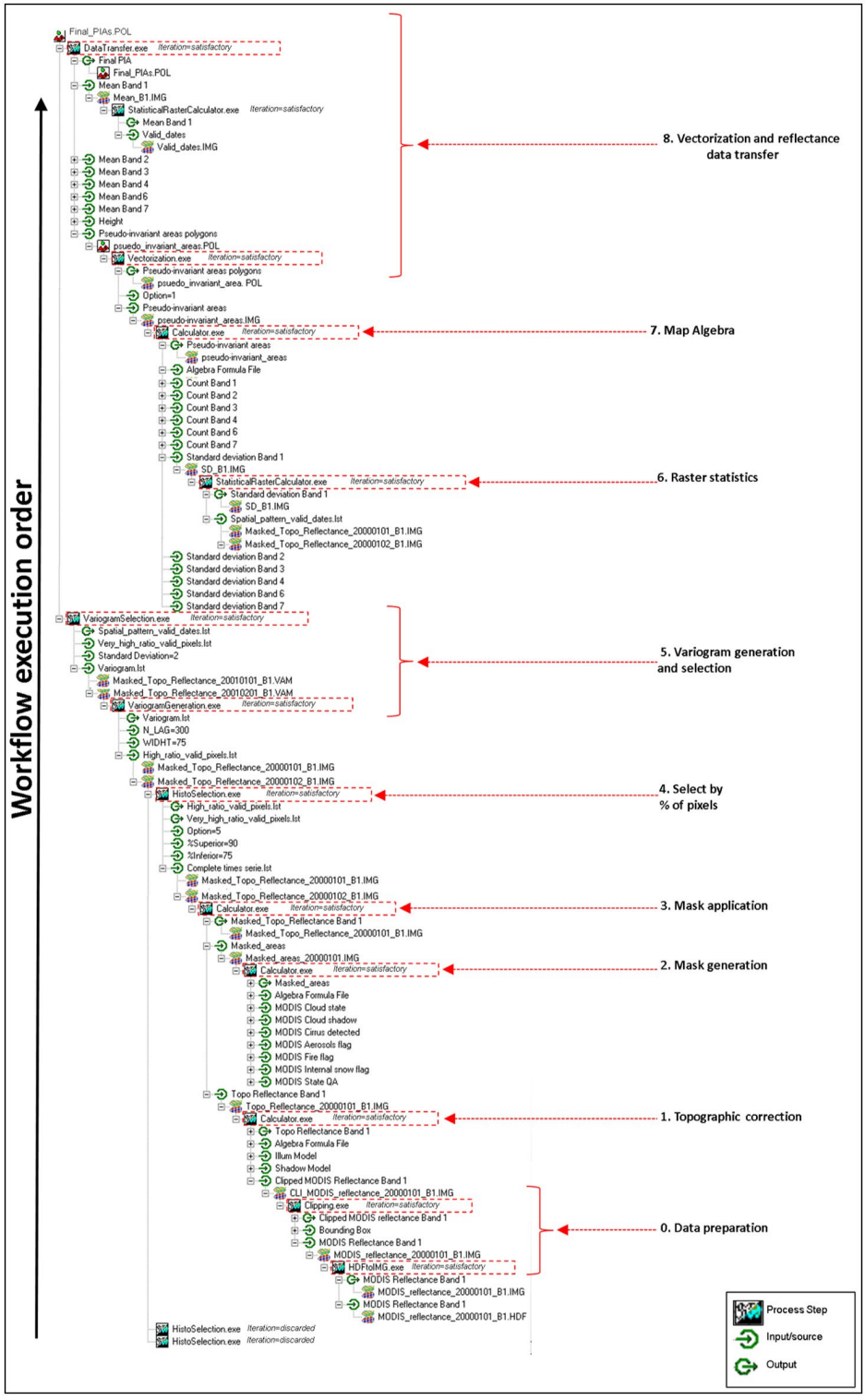

FIGURE 9 Example of provenance tree including all processes and sources used in PIA file generation. To show the provenance tree more clearly, only one branch is shown completely, some parts of the workflow are minimized, and step number and name labels have been added

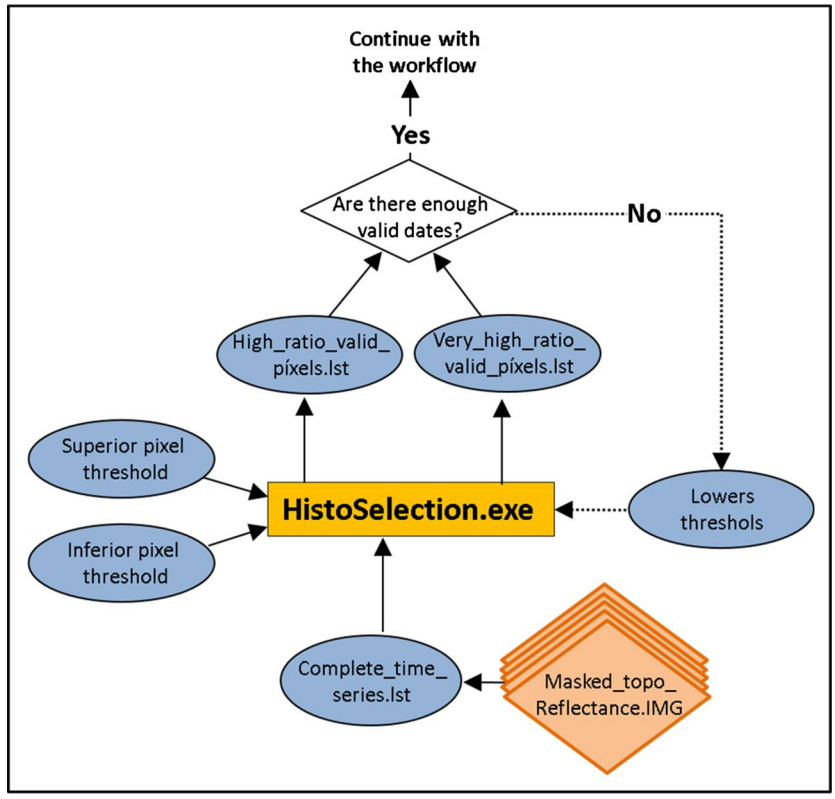

FIGURE 10 Detailed graph of step 4. The execution is repeated until the number of valid dates is sufficient

requirements (subsection 5.B.i). For each iteration of the HistoSelection.exe app, the PE recorded the sources used and, in the case of the parameters, also recorded the real values. Data users can observe how the data producers have lowered the superior and inferior thresholds to generate a proper result.

In order to test the implementation against ISO 19115-1 and 19115-2, this use case has been exported to ISO 19139 XML and imported in GeoNetwork (GeoNetwork, 2019). We consider GeoNetwork a reference implementation of the standards. The non-native ISO elements and the parameters are lost, due to the fact that the new version of ISO is not published yet, but the remainder of the elements are exported correctly.

5.2 | Analysis of the captured provenance

The use case presented has shown that combining WPS with the ISO lineage model provides a more complete provenance description and allows us to overcome some of the identified gaps (such as the documentation of parameters or sources order). In addition, the PE included in the MiraMon GIS & RS software captures provenance automatically. This more complete capture of provenance information can be used to infer quality, attribution, and trust about the generated PIAS, or to help in reproduction tasks. For instance:

- In the context of step 4 (Section "Replication with different pixel thresholds"), data users can reuse the *percentage of pixels threshold* in similar cloud regime areas or can replicate the task with a larger *percentage of pixels threshold* in more favorable cloud regimes (the number of valid dates is inversely correlated to the cloud regime, among other factors).

TABLE 2 Snippet of provenance captured in the example of replication with different pixel thresholds. The dashed line indicates the end of each loop

```
[QUALITY:LINEAGE:PROCESS1]
nOrganismes=1
history=HistoSelection.exe 5 /INF=75 /SUP=95
purpose=This program generates CSV hstogram of MODIS valis values
date=20180629 09559900+0200,20180629 10059900+0200
NomFitxer=HistoSelection.exe
[QUALITY:LINEAGE:PROCESS1:ORGANISME_1]
OrganisationName=Universitat Autònoma de Barcelona
[QUALITY:LINEAGE:PROCESS1:INOUT1]
identifier=High_ratio_valid_pixels.lst
iteration=discarded
[QUALITY:LINEAGE:PROCESS1:INOUT2]
identifier=Very_high_ratio_valid_pixels.lst
sentit=1
[QUALITY:LINEAGE:PROCESS1:INOUT3]
identifier=Option
TypeValues=C
ResultValue=5
[QUALITY:LINEAGE:PROCESS1:INOUT4]
identifier=%Inferior
TypeValues=C
ResultValue=75
[QUALITY:LINEAGE:PROCESS1:INOUT5]
identifier=%Superior
TypeValues=C
ResultValue=95
[QUALITY:LINEAGE:PROCESS1:INOUT6]
identifier=Complete_time_series.lst
TypeValues=S
-----------------------------------------------------------------
[QUALITY:LINEAGE:PROCESS2]
nOrganismes=1
history=HistoSelection.exe 5 /INF=70 /SUP=90
purpose=This program generates CSV hstogram of MODIS valis values
date=20180730 09559900+0200,20180730 10059900+0200
NomFitxer=HistoSelection.exe
iteration=discarded
[QUALITY:LINEAGE:PROCESS2:ORGANISME_1]
OrganisationName=Universitat Autònoma de Barcelona
[QUALITY:LINEAGE:PROCESS2:INOUT1]
identifier=High_ratio_valid_pixels.lst
sentit=1
[QUALITY:LINEAGE:PROCESS2:INOUT2]
identifier=Very_high_ratio_valid_pixels.lst
sentit=1
[QUALITY:LINEAGE:PROCESS2:INOUT3]
identifier=Option
TypeValues=C
ResultValue=5
[QUALITY:LINEAGE:PROCESS2:INOUT4]
identifier=%Inferior
TypeValues=C
ResultValue=70
[QUALITY:LINEAGE:PROCESS2:INOUT5]
identifier=%Superior
TypeValues=C
ResultValue=90
[QUALITY:LINEAGE:PROCESS2:INOUT6]
identifier=Complete_time_series.lst
TypeValues=S
-----------------------------------------------------------------
[QUALITY:LINEAGE:PROCESS3]
nOrganismes=1
history=HistoSelection.exe 5 /INF=65 /SUP=85
purpose=This program generates CSV hstogram of MODIS valis values
date=20180701 09559900+0200,20180701 10059900+0200
NomFitxer=HistoSelection.exe
iteration=satisfactory
[QUALITY:LINEAGE:PROCESS1:ORGANISME_1]
OrganisationName=Universitat Autònoma de Barcelona
[QUALITY:LINEAGE:PROCESS3:INOUT1]
identifier=High_ratio_valid_pixels.lst
sentit=1
[QUALITY:LINEAGE:PROCESS3:INOUT2]
identifier=Very_high_ratio_valid_pixels.lst
sentit=1
[QUALITY:LINEAGE:PROCESS3:INOUT3]
identifier=Option
TypeValues=C
ResultValue=5
[QUALITY:LINEAGE:PROCESS3:INOUT4]
identifier=%Inferior
TypeValues=C
ResultValue=65
[QUALITY:LINEAGE:PROCESS3:INOUT5]
identifier=%Superior
TypeValues=C
ResultValue=85
[QUALITY:LINEAGE:PROCESS3:INOUT6]
identifier=Complete_time_series.lst
TypeValues=S
```

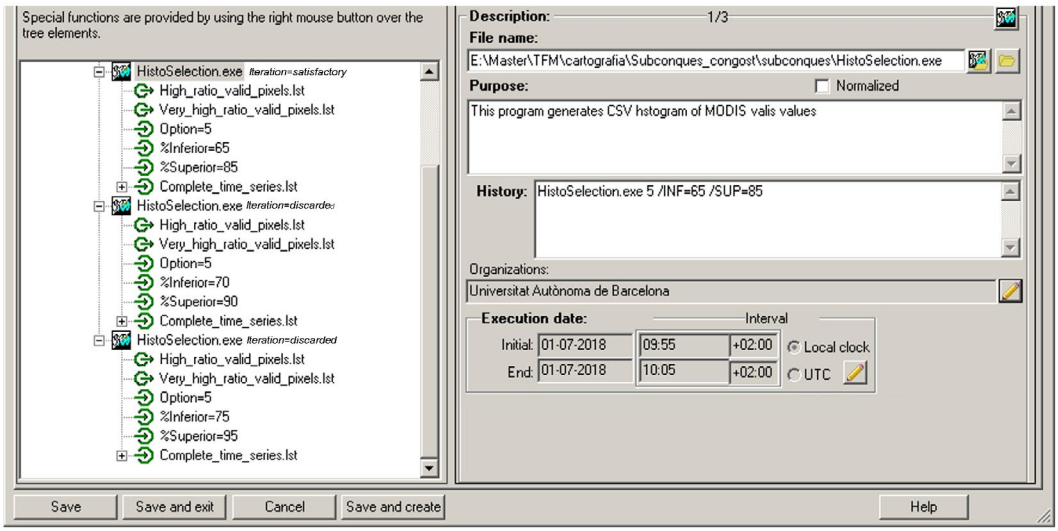

FIGURE 11 Step 4 tree-like provenance workflow representation in the GeMM. The graph records the satisfactory and discarded executions with the parameters used in each loop. The right-hand panel shows the properties of each parameter and the processes executed, such as the file name, execution date, command line (history), and purpose

- In the context of the whole PIA *generation* process, users can use provenance to check if there are sufficient images to ensure PIAs with high-quality data (the quality and consistency of the PIAs generated are directly related to the final number of images used). Or, to check if the final selection of images is representative enough of the whole dataset series (to define an area as pseudo-invariant it is necessary to have a homogeneous distribution of image dates).
- In the context of the whole PIA generation process, users can use provenance to check which of the algorithms used are open source.
- Users can replicate the entire workflow that was used to generate PIAs using provenance information with the same or different parameters.

Beyond the applicability of the provenance model improvements in the presented use case, most of them can also be applied to many other workflows where provenance has been captured. As a proof of fact, some of them were included as a change request in the revision of ISO 19115-2 (the documentation of parameters), and the new revision recognizes the usefulness of this improvement and includes this request. Therefore, the PE captures provenance in compliance with the current version of ISO 19115-2.

The use case also points out the utility of documenting scientific experiments that are not purely sequential, such as loops of discarded executions performed during data-generation processes. Step 4 (Sections "Replication with different pixel thresholds" and "Provenance capture and description") is an evident example. To solve the issue, this article proposes a practical solution using the *LE_Processing:otherPropertyType* of the *LE_Lineage* model (ISO 19115-2). The *otherPropertyType* tag is mapped to a *recordtype* with a single field called *"iteration"* and *otherProperty* states that *"iteration=discarded"* (default value will be *"satisfactory"*). In the future, it might be useful to have in the ISO model a new attribute to contain the *"iteration"* information.

Concerning visualization, the MiraMon metadata editor (GeMM) has successfully represented the provenance information captured during PIAS generation (Figures 9 and 11), allowing users to interpret it. Moreover, the GeMM graphical interface (Figure 6) permits us to edit and complete the information captured. Nevertheless,

representing a graph of provenance is a difficult task due to its complexity (multiple relations and different hierarchical levels). Thus, more effort to enhance the comprehension of provenance should be made.

6 | CONCLUSIONS

This article claims that data provenance is useful in the phases of quality, reliability, the fitness-for-use assessment, and workflow replication and data reproduction, when provenance information is complete.

However, we have detected that there are still some gaps in the full geospatial provenance description, which affect the provenance usefulness. In this sense the article has proposed some improvements to the ISO 19115 lineage model, to provide more complete and accurate provenance information. In addition, the article presents the PE to capture complex workflows like the one presented as a use case for generating a PIA bench. This relevant amendment and the automatic acquisition of geospatial provenance provides a complete recipe for generating geospatial data for data users.

The automatic acquisition of geospatial provenance represents a step forward in the development of a model constructor tool in the context of the MiraMon software. A model constructor would allow scientific modelers to reproduce previous chains of processes in different scenarios, using the provenance captured from previous executions. Future efforts should also aim to enhance the exploitation of catalogues of provenance information previously captured. Therefore, tools for facilitating queries about the "what," "when," "who," "how," and "where" of the generated geospatial data will give added value to the provenance information captured. These queries should provide information about geoprocessing tools implementing generic algorithms (e.g. in a given dataset where data have been generated with a specific algorithm). This will help users to more precisely choose not only the appropriate geospatial data, but also the correct algorithm and geoprocessing tool.

ACKNOWLEDGMENTS

This work has been conducted within the framework of the Geography PhD program of the Universitat Autònoma de Barcelona. This work has been partially funded by the Spanish MCIU Ministry through the NEWFORLAND research project (RTI2018-099397-B-C21/C22 MCIU/AEI/ERDF, EU), by the Catalan Government (SGR2017 1690), and by the ECoPotential and ERA-PLANET research projects. ECoPotential received funding from the European Union's Horizon 2020 research and innovation program under grant agreement (GA) No. 641762; ERA-PLANET under GA No. 689443. Xavier Pons is the recipient of an ICREA Academia Excellence in Research Grant (2016–2020).

ORCID


Guillem Closa https://orcid.org/0000-0002-1333-171X

REFERENCES

Altintas, I., Barney, O., & Jaeger-Frank, E. (2006). Provenance collection support in the Kepler scientific workflow system. In L. Moreau & I. Foster (Eds.), *Provenance and annotation of data: IPAW 2006* (Lecture Notes in Computer Science, Vol. *4145*, pp. 118–132). Berlin, Germany: Springer.

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). *Research objects: Towards exchange and reuse of digital knowledge*. Retrieved from http://precedings.nature.com/documents/4626/version/1

Box, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H. F., Thatte, S., & Winer, D. (2016). Simple *Object Access Protocol (SOAP) 1.1*. Retrieved from http://www.w3.org/TR/SOAP

Buneman, P., Khanna, S., & Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. In J. Van den Bussche & V. Vianu (Eds.), *Database theory: ICDT 2001* (Lecture Notes in Computer Science, Vol. *1973*, pp. 316–330). Berlin, Germany: Springer.

Castronova, A. M., Goodall, J. L., & Elag, M. M. (2013). Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard. *Environmental Modelling & Software*, *41*, 72–83.

Closa, G., Masó, J., Proß, B., & Pons, X. (2017). W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment. *Computers, Environment & Urban Systems*, *64*, 103–117.

Di, L., & McDonald, K. (1999). Next generation data and information systems for Earth sciences research. In *Proceedings of the First International Symposium on Digital Earth* (Vol. *1*, pp. 92–101). Beijing, China.

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

Di, L., Yue, P., Ramapriyan, H. K., & King, R. L. (2013). Geoscience data provenance: An overview. *IEEE Transactions on Geoscience & Remote Sensing*, *51*(11), 5065–5072.

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 Infrastructure Research*, *7*, 352–377.

Geller, G. N., & Turner, W. (2007). The model web: A concept for ecological forecasting. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium* (pp. 2469–2472). Barcelona, Spain: IEEE.

GeoNetwork. (2019). *GeoNetwork open source community website*. Retrieved from http://geonetwork-opensource.org/

Goodall, J. L., Robinson, B. F., & Castronova, A. M. (2011). Modeling water resource systems using a service-oriented computing paradigm. *Environmental Modelling & Software*, *26*(5), 573–582.

Granell, C., Díaz, L., Schade, S., Ostländer, N., & Huerta, J. (2013). Enhancing integrated environmental modelling by designing resource-oriented interfaces. *Environmental Modelling & Software*, *39*, 229–246.

Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., … Watson, P. (2003). Provenance of e-science experiments: Experience from bioinformatics. In *Proceedings of the UK e-Science Programme All Hands Conference* (pp. 223–226). Nottingham, UK.

Groth, P., & Moreau, L. (2013). *PROV-Overview: An overview of the PROV family of documents*. Retrieved from http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Hadjimitsis, D. G., Clayton, C. R. I., & Retalis, A. (2009). The use of selected pseudo-invariant targets for the application of atmospheric correction in multi-temporal studies using satellite remotely sensed imagery. *International Journal of Applied Earth Observation & Geoinformation*, *11*(3), 192–200.

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

ISO. (2014). *ISO 19115-1:2014: Geographic Information – Metadata – Part 1: Fundamentals*. Retrieved from https://www.iso.org/standard/53798.html

ISO. (2018). *ISO 19165-1:2018: Geographic information – Preservation of digital data and metadata – Part 1: Fundamentals*. Retrieved from https://www.iso.org/standard/67325.html

ISO. (2019). *ISO 19115-2:2019: Geographic information – Metadata – Part 2: Extensions for acquisition and processing*. Retrieved from https://www.iso.org/standard/67039.html

Jiang, L., Kuhn, W., & Yue, P. (2017). An interoperable approach for Sensor Web provenance. In *Proceedings of the 6th International Conference on Agro-Geoinformatics*. Fairfax, VA: IEEE.

Jiang, L., Yue, P., Kuhn, W., Zhang, C., Yu, C., & Guo, X. (2018). Advancing interoperability of geospatial data provenance on the web: Gap analysis and strategies. *Computers & Geosciences*, *117*, 21–31.

Jirka, S., Nüst, D., & Proß, B. (2013). Sensor web and web processing standards for crisis management. In *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management*. Baden-Baden, Germany.

Kim, J., Gil, Y., & Ratnakar, V. (2006). Semantic metadata generation for large scientific workflows. In *Proceedings of the 5th International Semantic Web Conference* (pp. 357–370). Athens, GA: ACM.

Kitanidis, P. K. (1997). *Introduction to geostatistics: Applications in hydrogeology*. Cambridge, UK: Cambridge University Press.

Kunde, M., Bergmeyer, H., & Schreiber, A. (2008). Requirements for a provenance visualization component. In J. Freire, D. Koop, & L. Moreau (Eds.), *Second International Provenance and Annotation Workshop, IPAW 2008, Salt Lake City, UT, USA, June 17–18, 2008, Revised Selected Papers* (pp. 241–252). Berlin, Germany: Springer.

Lopez-Pellicer, F. J., & Barrera, J. (2014). *D16 1 Call 2: Linked map VGI provenance schema* (Linked Map Subproject of Planet Data, Seventh Framework Programme). Brussels, Belgium: European Commission.

Masó, J., Closa, G., Gil, Y., & Proß, B. (2013). *OGC® Testbed 10 Provenance Engineering Report*. Wayland, MA: Open Geospatial Consortium.

Meng, X., Xie, Y., & Bian, F. (2010). Distributed geospatial analysis through web processing service: A case study of earthquake disaster assessment. *Journal of Software*, *5*(6), 671–679.

Michaelis, C. D., & Ames, D. P. (2009). Evaluation and implementation of the OGC web processing service for use in client-side GIS. *Geoinformatica*, *13*(1), 109–120.

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.

Miles, S., Wong, S. C., Fang, W., Groth, P., Zauner, K. P., & Moreau, L. (2007). Provenance-based validation of e-science experiments. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(1), 28–38.

Nativi, S., Mazzetti, P., & Geller, G. N. (2013). Environmental model access and interoperability: The GEO Model Web initiative. *Environmental Modelling & Software*, *39*, 214–228.

OGC. (2010). *OGC® WPS 2.0 Interface Standard: OGC 10-59r2, 14-065*. Wayland, MA: Open Geospatial Consortium.

Padró, J. C., Pons, X., Aragonés, D., Díaz-Delgado, R., García, D., Bustamante, J., … Lange, M. (2017). Radiometric correction of simultaneously acquired Landsat-7/Landsat-8 and Sentinel 2A imagery using pseudoinvariant areas (PIA): Contributing to the Landsat time series legacy. *Remote Sensing*, *9*(12), 1319.

Pesquer, L., Domingo, C., & Pons, X. (2013). A geostatistical approach for selecting the highest quality MODIS daily images. In J. M. Sanches, L. Micó, & J. Cardoso (Eds.), *Pattern recognition and image analysis* (Lecture Notes in Computer Science, Vol. *7887*, pp. 608–615). Berlin, Germany: Springer.

Pesquer, L., Domingo, C., & Pons, X. (2019). Spatial and spectral pattern identification for the automatic selection of high quality MODIS images. *Journal of Applied Remote Sensing*, *13*(1), 014510.

Pesquer, L., Masó, J., Moré, G., Pons, X., Peces, J., & Doménech, E. (2012). Servicio interoperable (WPS) de procesado de imágenes Landsat. *Teledetección*, *37*, 51–56.

Pesquer Mayos, L., Jirka, S., Stasch, C., Masó Pau, J., & Arctur, D. (2016). RiBaSE: A pilot for testing the OGC web services integration of water-related information and models. In *Proceedings of the 2016 Geospatial Sensor Webs Conference*. Münster, Germany.

Pons, X. (2019). *MiraMon: Geographical information system and remote sensing software*. Barcelona, Spain: Centre de Recerca Ecològica i Aplicacions Forestals.

Pons, X., & Masó, J. (2016). A comprehensive open package format for preservation and distribution of geospatial data and metadata. *Computers & Geosciences*, *97*, 89–97.

Pons, X., Pesquer, L., Cristóbal, J., & González-Guerrero, O. (2014). Automatic and improved radiometric correction of Landsat imagery using reference values from MODIS surface reflectance images. *International Journal of Applied Earth Observation & Geoinformation*, *33*, 243–254.

Roy, D. P., Borak, J. S., Devadiga, S., Wolfe, R. E., Zheng, M., & Descloitres, J. (2002). The MODIS land product quality assessment approach. *Remote Sensing of Environment*, *83*(1–2), 62–76.

Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, *264*(5164), 1421–1426.

Scheider, S., & Ballatore, A. (2018). Semantic typing of linked geoprocessing workflows. *International Journal of Digital Earth*, *11*(1), 113–138.

Steele, J., & Iliinsky, N. (2010). *Beautiful visualization: Looking at data through the eyes of experts*. Sebastopol, CA: O'Reilly Media.

Vermote, E. F., & Kotchenova, S. Y. (2008). *MOD09 (surface reflectance) user's guide, version 1.1*. Retrieved from http://modis-sr.ITdri.org

Vidal-Macua, J. J., Zabala, A., Ninyerola, M., & Pons, X. (2017). Developing spatially and thematically detailed backdated maps for land cover studies. *International Journal of Digital Earth*, *10*(2), 175–206.

Wang, R., Zeng, C., Li, P., & Shen, H. (2011). Terra MODIS band 5 stripe noise detection and correction using MAP-based algorithm. In *International Conference on Remote Sensing, Environment and Transportation Engineering* (pp. 8612–8615). Nanjing, China: IEEE.

Xu, Z. W., Wang, Y. P., Li, Y., Ma, F., Zhang, F., & Ye, C. J. (2010). Sediment transport patterns in the eastern Beibu Gulf based on grain-size multivariate statistics and provenance analysis. *Acta Oceanologica Sinica*, *32*(3), 67–78.

Yu, G. E., Zhao, P., Di, L., Chen, A., Deng, M., & Bai, Y. (2012). BPELPower: A BPEL execution engine for geospatial web services. *Computers & Geosciences*, *47*, 87–101.

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., Wei, Y., Di, L., He, L., Gong, J., & Zhang, L. (2011). Sharing geospatial provenance in a service-oriented environment. *Computers, Environment & Urban Systems*, *35*(4), 333–343.

Zabala, A., & Masó, J. (2005). Integrated hierarchical metadata proposal: Series, layer, entities and attributes. In *Proceedings of the International Cartographic Conference on Mapping Approaches in a Changing World*. A Coruña, Spain: ICA.

Zabala, A., Masó, J., Bastin, L., & Bigali, L. (2013). Increasing dataset quality metadata presence: Quality focused metadata editor and catalogue queriables. In *Proceedings of the Inspire Conference*. Florence, Italy.

Zabala, A., Masó, J., & Pons, X. (2016). Quality and user feedback metadata: Theoretical aspects and a practical implementation in the MiraMon metadata editor. In *Proceedings of the Inspire Conference*. Barcelona, Spain.

Zhang, M., Yue, P., Wu, Z., Ziebelin, D., Wu, H., & Zhang, C. (2017). Model provenance tracking and inference for integrated environmental modelling. *Environmental Modelling & Software*, *96*, 95–105.

How to cite this article: Closa G, Masó J, Zabala A, Pesquer L, Pons X. A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation. *Transactions in GIS*. 2019;23:1102–1124. https://doi.org/10.1111/tgis.12555