Universitat Autònoma de Barcelona

Aportacions en el camp del Llinatge Geospacial en entorns distribuïts: de la captura a l'explotació

Guillem Closa Santos

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: http://es.creativecommons.org/blog/licencias/

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license: https://creativecommons.org/licenses/?lang=en

Tesi Doctoral
Programa de Doctorat en Geografia
Departament de Geografia

Aportacions en el camp del Llinatge Geospacial en entorns distribuïts: de la captura a l'explotació

Autor: Guillem Closa Santos
Director: Joan Masó Pau
Tutor: Xavier Pons Fernández

Bellaterra, Novembre2020

Als meus pares, l'Albert i la Pilar

Qui perd els orígens, perd la seva identitat.

Frase utilitzada i adaptada pel cantautor Raimon, tot i que el seu origen real és confús i el seu llinatge s'ha perdut. Fet que representa una paradoxa en ella mateixa.

ÍNDEX

ÍNDEX GENERAL / ÍNDICE GENERAL / TABLE OF CONTENTS

Figures

TAULES

Agraïments

Agraïments

Aquesta és una tesi sense beca que l'he anat fent en paral·lel a la meva feina diària dins el grup de recerca Grumets. Si l'he pogut tirar endavant ha estat, en part, gràcies a la solidaritat i companyonia del grup. Per tant el meu agraïment infinit per aquest col·lectiu divers i heterogeni format per grans professionals. Però sobretot format per grans persones com l'Alaitz, la Cristina Cea, la Cristina Domingo, el Cristian, l'Oscar, el Mario, la Núria Cartes, el Lluís, el Xavi, la Ivette, l'Ester, l'Alba, la Txell, el Miquel, la Núria Julià, l'Abel, el Jordi, la Caty, el Xavier i el Joan. En èpoques passades també gent com la Maria, el Juanjo, el Daniel, l'Eva, l'Edu i el José Angel. A tots vosaltres, moltes gràcies.

Evidentment les meves gratituds especials al Dr. Joan Masó per haver-me dirigit la tesi i compartit una petita part del seu coneixement amb mi.

També vull expressar el meu agraïment al Dr. Xavier Pons per col·laborar en les nostres publicacions fent sempre propostes de millora i ajudant a enriquir el resultat final.

Al Dr. Lluís Pesquer, per haver acceptat viatjar al Japó per defensar (entre d'altres) una contribució d'aquesta tesi (capítol 5) al IGARSS del 2019. Et dec unes birres. A la Núria Julià, per les hores dedicades als desenvolupaments relacionats amb el capítol 6. A la resta de coautors dels articles, la Dra. Alaitz Zabala i el Sr. Benjamin ProB.

A l'escola de doctorat en geografia, amb la Dra. Mireia Baylina i la Sra Alba Palma al capdavant, per haver-me acceptat com a estudiant de doctorat i per tot el suport i ajuda rebuda. Al DdG en general, per haver-me acollit tant bé aquests tres anys que he estat entre vosaltres. A la Dra. Anna Badia i al Dr. Pere Serra, pels consells donats en les successives comissions de seguiment que, de ben segur, han ajudat a millorar el resultat final.

També he de donar gràcies al CREAF en general, on vaig començar la tesi i on m'hi vaig trobar tant bé.

No vull oblidar-me de tots i cadascun dels companys de despatx que he tingut, el Joan, el Juanjo, la Maria, l'Oscar, el Mario i la Núria. Tots m'heu ajudat moltíssim.

Al grup de *nyamnyam*. Gent com el Roger, l'Eduard, la Sandra, el Carles, l'Íngrid, el Xavier, la Ivette, l'Ester, el Cristian, l'Abel, l'Oscar, la Cristina, la Núria i l'Alba han fet de l'hora de dinar un oasi d'esbarjo enmig de la rutina. Quant de tertulià en potència! Val a dir que l'actualitat dels darrers anys ens ho ha posat fàcil. En ocasions també ens ha superat.

Al Ricard Palanques pels anys de Sagarda. També per dir-me doctor des del principi, encara que em ruboritzés. Esperem que finalment pugui ser cert. Al Jonàs i al David, per les bones estones que també van compartir amb nosaltres a Sagarda.

Al Tòfol i l'Esperança per fer-ho tot sempre fàcil i per deixar-nos la caseta d'Es Canutells i convertir l'estiu del 2020, a pesar de les circumstàncies generals, en un dels millors estius possibles.

Quan vaig començar a fer aquesta tesi el meu pare, la meva tieta Montse i el meu tiet Ramon encara estaven entre nosaltres. Tot tres han marxat, abans d'hora i encara joves. Per a ells, el meu record constant. A pesar dels infortunis, la meva família segueix sent i estant molt i molt present. Amb la

meva mare i el seu optimisme militant al capdavant, seguida de ben a prop pels meus germans Ori i Nineta. I ara també els meus nebots la Claudia i el Marc, i també l'Elena. La meva àvia Antonyita, que enguany n'ha fet 99 . Ole tu, l'estiu vinent celebrem els 100! A més, tinc la sort de comptar amb un llaaarg equip tietes, tiets, cosins i afegits. A tots, moltes gràcies de debò.

Finalment, compartir-ho amb la Sole i la Gina que s'han carregat de paciència per aguantar el meu (bon) humor d'aquests darrers mesos. Sole, gràcies per tant! i per llegir-te la tesi😉. Espero ser capaç de compensar, ni que sigui en part, tot el que m'has donat.

Salut!

Novembre 2020

Resum

Resum (Català)

El llinatge geospacial es pot definir com aquella part de les metadades que descriu l'origen de les dades (essencialment fonts i processos emprats). Aquest té una reconeguda utilitat en la descoberta de dades, anàlisi de la qualitat i en la reproductibilitat de la informació geogràfica, entre d'altres beneficis. Malgrat l'existència de literatura científica i de models de dades *ad hoc* per representar-lo, la presència d'informació de llinatge a les metadades geospacials és en general encara escassa, i quan hi és present, no és completa.

La hipòtesi principal d'aquesta tesi doctoral es basa en que l'absència força generalitzada d'informació de llinatge dins les metadades actua com a factor limitador en la interoperabilitat i la reproductibilitat de dades, processos i models geospacials tan en entorns científics com administratius.

Davant d'aquest escenari, són necessàries investigacions que proposin nous mecanismes per potenciar una major incorporació de la informació del llinatge en les metadades geospacials. Aquesta tesi doctoral investiga, en primer lloc, les carències en les fases de representació, captura, emmagatzematge i visualització del llinatge. En segon lloc, proposa alternatives, tan a nivell teòric com aplicat, que potenciïn una millor descripció del llinatge alhora que incrementin la seva presència en les metadades. Finalment, planteja metodologies per augmentar-ne la seva utilitat tan en el context dels Sistemes d'Informació Geogràfica (SIG) com en entorns web distribuïts.

En relació a la representació del llinatge i l'ús d'estàndards i models s'ha constatat que, a dia d'avui, és encara un aspecte obert en molts sentits. Per exemple, la completesa del llinatge capturat, la interoperabilitat dels models o la granularitat de la informació que s'hi recull. Els capítols 2, 3, 4 i 6 fan propostes per millorar les capacitats dels models. En concret, el capítol 2 proposa una adaptació del model *W3C PROV* (model genèric per descriure el llinatge de tot tipus d'informació a la web) a les singularitats de la informació geogràfica i aprofitar les seves característiques per descriure el llinatge a nivell de conjunt d'informació, d'element geospacial i d'atribut. Els capítols 3 i 4 proposen l'ús combinat dels models de llinatge inclosos a la ISO 19115-1 i la ISO 19115-2 amb l'estàndard *Web Processing Service* (WPS) de *l'Open Geospatial Consortium* (OGC) per millorar-ne la completesa. Finalment, el capítol 6 emfatitza en la necessitat de representar i relacionar el llinatge de diversos conjunts de dades per maximitzar-ne els beneficis.

Les fases de captura i edició del llinatge han estat identificades com uns dels principals obstacles per una major presència del llinatge en les metadades. En aquest sentit, la manca d'eines que recullin el llinatge automàticament i la dificultat per documentar i/o editar el llinatge a posteriori de la generació de les pròpies dades en són les principals causes. Els capítols 3 i 4 presenten una eina anomenada *Provenance Engine* (PE). L'eina, implementada en el marc del programa de SIG i Teledetecció MiraMon, captura automàticament el llinatge de les execucions realitzades amb el programa. A més, permet als usuaris editar la informació de llinatge a posteriori, afegint o eliminant processos o fonts a un flux de treball geospacial.

Aspectes també importants alhora de millorar la comprensió dels processos de producció de dades són les tècniques de visualització i simbolització utilitzades. Per tant, eines que facilitin la interpretació del llinatge són necessàries i tenen un impacte directe en la seva comprensió i ús. En aquest sentit, el MiraMon permet visualitzar el llinatge com un seqüència de processos. Cada procés té una llista indentada amb tots els paràmetres utilitzats i les sortides generades. A més, el capítol 6 presenta un sistema alternatiu que proporciona i renderitza la informació de llinatge com un graf en un entorn distribuït.

En darrer lloc, s'ha treballat en generar propostes que incrementin la utilitat del llinatge i aportin un valor afegit al seu ús. El capítol 5 estableix les bases teòriques per realitzar consultes sobre la informació de llinatge de dades de teledetecció per tal de rebre només aquells fragments de dades o processos que ens poden interessar en un moment determinat. Finalment, el capítol 6 amplia i complementa el capítol 5. En concret, presenta el disseny d'un sistema de consultes inserit en un navegador de mapes. El disseny permet presentar la informació de llinatge de diverses capes incloses en el navegador en una sola vista, veure les interaccions i fer comparacions de fluxos que han donat lloc als diversos conjunts de dades.

Resumen (Castellano)

El linaje geoespacial se puede definir como aquella parte de los metadatos que describe el origen de los datos (esencialmente fuentes y procesos utilizados). Este tiene una reconocida utilidad en el descubrimiento, análisis de la calidad y en la reproducibilidad de la información geográfica, entre otros beneficios. A pesar de la existencia de literatura científica y de modelos de representación *ad hoc*, la presencia de información de linaje en los metadatos geoespaciales es en general todavía escasa, y cuando está presente, no es completa.

La hipótesis principal de esta tesis doctoral se basa en que la ausencia generalizada de información de linaje dentro de los metadatos geoespaciales actúa como factor limitador en la interoperabilidad y la reproducibilidad de datos, procesos y modelos geoespaciales tanto en entornos científicos como administrativos.

Ante este escenario, son necesarias investigaciones que propongan nuevos mecanismos para potenciar una mayor incorporación de la información del linaje en los metadatos geoespaciales. Con este fin, esta tesis doctoral investiga, en primer lugar, las carencias en las fases de representación, captura, almacenamiento y visualización del linaje. En segundo lugar, propone alternativas, tanto a nivel teórico como aplicado, que potencien una mejor descripción del linaje a su vez que incrementen su presencia en los metadatos. Finalmente, plantea metodologías para aumentar su utilidad tanto en el contexto de los Sistemas de Información Geográfica (SIG) como en entornos web distribuidos.

En relación con la representación del linaje y el uso de estándares y modelos se ha constatado que, a día de hoy, es todavía un aspecto abierto en muchos sentidos. Por ejemplo, la completitud del linaje capturado, la interoperabilidad de los modelos o la granularidad de la información que se recoge. En este sentido, los capítulos 2, 3, 4 y 6 realizan propuestas para mejorar las capacidades de los modelos. En concreto, el capítulo 2 propone una adaptación del modelo W3C PROV (modelo genérico para describir el linaje de todo tipo de información en la web) a las singularidades de la información geográfica y aprovechar sus características para describir el linaje a nivel de conjunto de datos, de elemento geoespacial y de atributo. Los capítulos 3 y 4 proponen el uso combinado de los modelos de linaje incluidos en la ISO 19115-1 y la ISO 19115-2 con el estándar *Web Processing Service* (WPS) del *Open Geospatial Consortium* (OGC) para mejorar su completitud. Finalmente, el capítulo 6 enfatiza en la necesidad de representar y relacionar el linaje de distintos conjuntos de datos con el objetivo de maximizar los beneficios que nos aporta.

Las fases de captura y edición del linaje han sido identificadas como unos de los principales obstáculos para una mayor presencia del linaje en los metadatos. En este sentido, la ausencia de herramientas que recojan el linaje automáticamente y la dificultad para documentar y/o editar el linaje a posteriori de la generación de los propios datos son las principales causas. Los capítulos 3 y 4 presentan una herramienta llamada *Provenance Engine* (PE). La herramienta, implementada en el marco del programa de SIG y Teledetección MiraMon, captura automáticamente el linaje de las ejecuciones realizadas con el programa. Además, permite a los usuarios editar la información de linaje a posteriori, añadiendo o eliminando procesos o fuentes a un flujo de trabajo geoespacial.

Aspectos importantes para mejorar la comprensión de los procesos de producción de los datos son las técnicas de visualización y simbolización utilizadas. Por tanto, herramientas que faciliten la interpretación del linaje son necesarias y tienen un impacto directo en su comprensión y uso. En este sentido, el MiraMon permite visualizar el linaje como una secuencia de procesos. Cada proceso tiene una lista indentada con todos los parámetros utilizados y las salidas generadas. Además, el capítulo 6 presenta un sistema alternativo que proporciona y renderiza la información de linaje como un grafo en un entorno distribuido.

En último lugar, se ha trabajado para generar propuestas que incrementen la utilidad del linaje y aporten valor añadido a su uso. El capítulo 5 establece las bases teóricas para realizar consultas sobre la información del linaje de datos de teledetección con el objetivo de recibir sólo aquellos fragmentos de datos o procesos que nos pueden interesar en un momento determinado. Finalmente, el capítulo 6 amplía y complementa el capítulo 5. En concreto, presenta el diseño de un sistema de consultas insertado en un navegador de mapas. El sistema permite presentar la información de linaje de distintos conjuntos de datos incluidos en el navegador en una sola vista, ver las interacciones y hacer comparaciones de los flujos que han dado lugar a los distintos conjuntos de datos.

Abstract (English)

Geospatial lineage can be defined as the part of metadata that describes the origin of the data (in essence, sources and processes used). Its usefulness has been recognized in data discovery, quality assessment, and reproducibility of geographic information. Despite the existence of scientific literature and data models to represent it, the presence of lineage information in geospatial metadata is generally still scarce, and when present, this is not comprehensive enough.

The main hypothesis of this PhD is based on the evidence the absence of lineage information in geospatial metadata acts as a barrier for the interoperability and reproducibility of data, processes and geospatial models, in both scientific and administrative environments.

In this scenario, a further research is needed in order to propose new mechanisms to promote a greater incorporation of lineage information into geospatial metadata. Firstly, this PhD investigates the deficiencies in the phases of representation, capture, storage and visualization of the lineage information. Secondly, it proposes alternatives, both at theoretical and practical level, that promote a better description of lineage to increase its presence in the metadata. Finally, it proposes methodologies to increase its usefulness in both, in the context of Geographic Information Systems (GIS) as well as in distributed web environments.

The representation of lineage and the use of standards and data models have still open issues in many aspects, such as the completeness of the lineage, the interoperability of the models, or the granularity of the information collected. Chapters 2, 3, 4, and 6 make proposals to improve the capabilities of the models. Specifically, chapter 2 proposes an adaptation of the W3C PROV model (a generic model to describe the lineage of all types of information on the web) to the particularities of geographic information in order to describe lineage at layer, feature and attribute level. Chapters 3 and 4 propose the combination of the lineage models included in ISO 19115-1 and ISO 19115-2 with Web Processing Service (WPS) standard from the Open Geospatial Consortium (OGC) to improve the completeness of the lineage data model. Finally, chapter 6 emphasizes the need to represent and relate the lineage of different datasets to maximize benefits.

The capture and editing phases of lineage were identified as one of the main obstacles for a greater presence of lineage in metadata. In this sense, the lack of tools to collect lineage automatically and the difficulty for documenting and editing lineage after the generation of data are the main causes. Chapters 3 and 4 present a tool called Provenance Engine (PE). The tool, developed in the framework of MiraMon GIS and Remote Sensing software, captures automatically the lineage of executions performed by MiraMon. In addition, permits users to edit lineage information by adding or removing processing steps or sources from a geospatial workflow after the execution.

Important aspects for improving the understanding of data production processes are the visualization and symbolization techniques used. Therefore, tools to enhance the interpretation of lineage are necessary as has a direct impact on its understanding and usefulness. In this sense, MiraMon allows to visualize lineage as an indented sequence of processes including all parameters used and the

outputs generated. In addition, chapter 6 presents a system that provides and renders lineage information as a graph in a distributed web environment.

Finally, some proposals have been formulated to increase the usefulness and provide an added value to lineage. Chapter 5 sets a theoretical basis for querying lineage information on remote sensing data in order to receive only those fragments of data or processes that we are interested on. Finally, chapter 6 expands and complements the work presented in chapter 5. Specifically, it provides the design of a query system embedded within a map browser that allows presenting lineage information of some layers included in the browser in a single view, as well as compare the workflows executed to generate the different datasets.

1. Introducció i objectius

1.1 Introducció general

Les millores tecnològiques associades a la irrupció generalitzada de la Internet van suposar una revolució en l'accés i l'intercanvi d'informació en la majoria de camps científics. En l'àmbit de les Tecnologies d'Informació Geospacial (TIG) aquest fet va possibilitar als usuaris l'accés a grans volums de dades que fins al moment no els eren accessibles, o si més no, a aquella velocitat d'accés. Per altra banda, l'observació de la terra (EO) ha contribuït molt a l'augment del volum de dades disponible degut, en part, a l'increment del nombre de satèl·lits i sensors de teledetecció promoguts per les principals agències de l'espai; com la *European Space Agency* (ESA), la *National Aeronautics and Space Administration* (NASA), la *Japan Aerospace Exploration Agency* (JAXA) o la *Chinese National Space Administration* (CNSA). A més, iniciatives per distribuir conjunts de dades[1] com a dades obertes (Open Data) han augmentat molt en els darrers anys, esdevenint, a dia d'avui, una elecció força comú per autoritats governamentals, entorns científics i, per descomptat, en iniciatives *Volunteered Geographic Information* (VGI) (Informació Geogràfica Voluntària) (Coetzee, Ivánová, Mitasova, & Brovelli, 2020).

La convergència d'aquestes i altres tendències fa que, a dia d'avui, el volum d'Informació Geospacial (en endavant IG) que qualsevol usuari mitjà té a l'abast sigui molt ampli i divers. Aquest fet contribueix a facilitar estudis que combinen dades provinents de diferents disciplines i orígens (p. ex. impactes del canvi climàtic sobre la societat (Hoegh-Guldberg, et al., 2019) o els recents estudis de correlació de la distribució de la COVID-19 amb la temperatura i altres variables atmosfèriques (Olcina, Biener, & Marti Talavera, 2020)), possibilitant la generació de nou coneixement o nous productes. Uns productes que, un cop generats, molt probablement passaran a formar part de reservoris de dades accessibles per part d'altres usuaris, incrementant així la disponibilitat de dades.

En aquest context es pot afirmar que un dels principals hàndicaps a que s'enfronten els usuaris d'IG no és la manca de dades disponibles (com podria haver succeït en el passat), sinó la dificultat per descobrir, interpretar i seleccionar aquell o aquells productes que més s'adeqüen a cada projecte, no només en relació a l'àmbit geogràfic, a la temàtica i a la resolució, sinó també en termes de confiança, autoria, qualitat, llicència, etc.

Així doncs, per interpretar les dades, necessitem disposar de la informació auxiliar sobre les pròpies dades; allò que anomenem les metadades. En el camp geospacial, les metadades ens informen sobre l'autoria, l'escala, les unitats, el sistema de projecció, les fonts, el format i la qualitat (entre d'altes propietats) de dades i/o serveis geospacials. Sense les metadades, la informació geospacial perd bona part del seu valor (Ahonen-Rainio & Kraak, 2005). A més, és necessari que el nivell de comprensió de les metadades sigui el més ampli possible; és a dir, necessitem uns estàndards que afavoreixin la interoperabilitat de les dades i metadades. Un conjunt de dades és pot considerar interoperable si pot ser accedit i utilitzat per altres sistemes sense necessitat d'un esforç tècnic i humà important (Masó, 2012).

[1] L'expressió "conjunt de dades" es fa servir com a traducció de "dataset".

En aquest sentit, organismes d'estandardització com el *World Wide Web Consortium* [W3C], l'*International Organization for Standardization / Technical Committee 211* [ISO/TC211], o l'*Open Geospatial Consortium* [OGC], treballen activament en l'establiment d'especificacions aplicables als seus respectius sectors i proporcionen les funcionalitats necessàries per millorar el rendiment dels usuaris, tot garantint la interoperabilitat dels procediments i formats emprats (Albrecht, 1999) (Percivall, 2010).

Els recursos digitals han de ser trobables (*Findable*), accessibles (*Accessible*), interoperables (*Interoperable*) i reutilitzables (*Reusable*) (FAIR Guiding Principles for scientific data management and stewardship, 2016). És evident que en el procés d'estandardització de les metadades de la IG s'han produït millores en quan a l'ús de models i llenguatges comuns (Brodeur, et al., 2020) que ens han acostat a aquests principis. Tot i això, les metadades disponibles encara estan lluny, a dia d'avui, d'oferir una visió completa del procés de producció (Spiekermann, Jolly, Herzig, Burleigh, & Medyckyj-Scott, 2019) que permeti inferir la veracitat, la qualitat i la idoneïtat de les dades i dels fluxos de treball sense un cost elevat en temps. En un context científic caracteritzat pel rigor, determinar la qualitat i la veracitat de les dades és capital. Per aconseguir tot això necessitem models de metadades rics que ens reportin la recepta completa del procés de generació de les dades, és a dir, per a que ens reportin allò que anomenem el llinatge. A més, necessitem aplicacions que despleguin completament els models i permetin explotar les dades que representen, ja que les eines dels Sistemes d'informació Geogràfica (SIG) no han considerat els aspectes relacionat amb la qualitat de les dades fins fa poc temps (Ariza López, et al., 2020).

La hipòtesi principal d'aquesta tesi doctoral es basa en que l'absència generalitzada d'informació de llinatge dins les metadades actua com a factor limitador en la interoperabilitat i la reproductibilitat de dades i models geospacials en entorns científics. Aquesta tesi aporta propostes per millorar la descripció del llinatge i així augmentar la seva presència en les metadades de la IG.

Amb l'abundància de dades obertes i de recursos de processament, els científics poden modelitzar sistemes i fer previsions sense moure's del despatx, fet que ha contribuït a l'acceleració del ritme de publicacions científiques. Aquest augment, però, no ha anat lligat necessàriament d'un increment de la seva qualitat. En alguns casos, s'han vist augmentar el nombre de publicacions amb baix rigor científic, arribant a l'extrem de resultats falsos o exagerats (Ruiz-Mallén & Gmelch, 2020). En aquest sentit, la comunitat científica ha manifestat la necessitat de la reproductibilitat científica com a mecanisme per millorar la qualitat de la recerca (Munafò, et al., 2017). En recerca sobre dades, la millor manera d'aconseguir la transparència és documentant el llinatge dels resultats, per a comunicar els elements necessaris per a reproduir la recerca feta. És necessari documentar la recerca de la manera més completa i eficient per tal d'aconseguir la reproductibilitat científica, permetent així que altres investigadors, emprant eines equivalents, puguin arribar als mateixos resultats.

Entendre el potencial del llinatge en el context de la IG, tan en entorns SIG d'escriptori com en entorns de processament distribuïts, és clau per a identificar les mancances en la representació, captura i gestió del llinatge geospacial i fer propostes de millora.

Abans d'endinsar-nos en les definicions, particularitats, models i potencial del llinatge geospacial (subapartat 0), és necessari introduir el context d'ús i aplicació d'aquesta tesi doctoral (subapartats 1.1.1 i 0).

1.1.1 La informació geospacial en entorn distribuïts

Més enllà de la cartografia oferta pels grans actors de la industria del programari i el núvol (p. ex. Google Maps o Bing Maps), que les limita a imatges ortofotogràfiques i mapes de carreteres, els organismes cartogràfics han desenvolupat les Infraestructures de Dades Espacials (IDE) per distribuir/compartir les dades multi-temàtiques que generen. Una IDE és "*marc de cooperació que té com a finalitat facilitar el coneixement, l'accés i la utilització de la informació geogràfica disponible en un àmbit territorial a través d'Internet, per mitjà de geoportals que ofereixen serveis de catàlegs de dades i de metadades; de visualització, per mitjà de serveis de mapes; de localització, per mitjà d'adreces i de nomenclàtors, i, eventualment, d'altres tipus de geoserveis específics, a més de descàrregues de dades, de documents o d'altres recursos d'informació geospacial*" (Nunes, 2012). Aquestes s'organitzen al voltant de diferents nivells administratius o temàtics, podent anar des de l'àmbit local (p. ex. GeoPortalBCN (Ajuntament de Barcelona, 2020)), fins a l'àmbit supraestatal (p. ex. INSPIRE (European Comission, 2007)). Per tal de poder acomplir els seus objectius, les IDE defensen, estimulen i coordinen l'adopció d'estàndards (Masó, 2012). A Europa, la directiva INSPIRE (el marc legal per a la implantació de IDE dins la Unió Europea) (European Comission, 2007) obliga als estats membres a proveir informació geospacial utilitzant, majoritàriament, serveis web basats en estàndards de l'OGC (European Comission, 2007).

L'OGC ha liderat la generació d'estàndards per compartir informació geospacial a la xarxa. Fins al moment s'han implementat 161 estàndards oficials (OGC, 2020) que cobreixen un ampli ventall de serveis (Bai, Di, & Wei, 2009):

- Serveis de catàleg i registre: Catalogue Service (CSW) n'és l'exemple més conegut (Nebert, Whiteside, & Vretanos, 2007).
- Serveis d'accés a dades: L'estàndard per a distribució de dades de variació contínua és el Web Coverage Service (WCS) (Bauman, 2010), l'estàndard per a distribució de conjunts d'entitats és el Web Feature Service (WFS) i l'estàndard per a observacions de sensors és el Sensor Observation Service (SOS).
- Serveis de representació i visualització: L'estàndard de mapes per tessel·les Web Map Tile Service (WMTS) (Masó, Pomakis, & Julià , 2010).
- Serveis de transformació de dades: L'estàndard per a processos genèrics sobre informació geogràfica distribuïda, Web Processing Service (WPS) (Schut, 2007), n'és l'exemple més conegut.

Tot i que les IDE han eliminat parcialment l'aïllament de dades mediambientals i geospacials, encara hi ha moltes dades que es troben, en gran mesura, aïllades respecte a d'altres dominis d'informació (Schade & Smits, 2012) (Rojas , Athanasiou, Lehmann, & Hladky , 2013) (Yue, Guo, Zhang, Jiang, & Zhai, 2016). És el que es coneix com efecte "sitja". Aquest aïllament minva la descoberta de dades i la (re)usabilitat de les mateixes.

Més enllà de les IDE i de la informació geospacial, hi ha iniciatives de dades que promouen una administració transparent i una política de dades obertes similars en esperit però emprant altres solucions informàtiques. Complementant la IG amb altres tipus de bases de dades s'evita el seu aïllament i es fa possible la integració i harmonització amb altres conjunts de dades. Un exemple de dades obertes és la DBpedia (https://wiki.dbpedia.org/) que té un gran nombre d'objectes geospacials i està basada en una estructura *Linked Data* (dades enllaçades). El *Linked Data* és un mètode per a la publicació de dades estructurades que permet que puguin ser interconnectades entre elles (Berners-Lee, 2009). És basa en els següents principis (W3C, 2016):

- Utilització d'URI (Identificador Uniforme de Recursos) com a noms per a les coses (recursos).
- Ús de HTTP per tal que la gent pugui consultar representacions d'aquestes coses.
- Inclusió d'informació útil.
- Inclusió d'enllaços a altres URI perquè és pugui descobrir més coses.

El llenguatge *Resource Description Framework* (RDF)[2] (W3C, 2014) és utilitzat per descriure els recursos en forma de tripletes (subjecte-predicat-objecte) de forma explícita en el context del *Linked Data*. L'ús del RDF permet relacionar dades mitjançant l'ús d'hipervincles i ,en el seu extrem, pot ser utilitzat per crear una única infraestructura global i compartida de dades (Bizer, 2013). A més, l'ús de llenguatges estructurats per organitzar i interconnectar dades allunyades entre sí permet la realització de consultes (queries) complexes. En aquest context, el llenguatge estandarditzat SPARQL[3] (W3C, 2013) està especialment dissenyat per expressar consultes a fonts de dades diverses o distribuïdes expressades en RDF.

En el camp geospacial, l'ús de dades enllaçades permet establir relacions entre múltiples conjunts de dades, incorporant descripcions addicionals a les dades originals (Vilches-Blázquez , Villazón-Terrazas, Corcho, & Gómez-Pérez , 2014). Això possibilita l'enriquiment dels conjunts de dades i productes finals i facilita la integració de les dades i la consulta de dades enllaçades (Harth & Gil, 2014). La combinació de la IG amb el *Linked Data* podria impulsar l'eliminació de les sitges formades per conjunts de dades geospacials aïllades, i la seva integració en una web de dades enllaçades i enriquida (W3C, 2020). Un estudi de l'EuroSDR[4] situa el Linked Data com un dels aspectes més importants a investigar en els propers anys i el situa com a factor clau per a la transició cap a les IDE de nova generació (EuroSDR, 2018). Les IDE de nova generació han de tendir cap a infraestructures de coneixement espacial (*Spatial Knowlegde Infraestructures* (SKI)) on el coneixement sigui la font

[2] El Resource Description Framework (RDF) és un model conceptual de dades[1] estandarditzat pel World Wide Web Consortium (W3C), usat per definir dades en el web semàntic i, en general, en les aplicacions que requereixen un estàndard per intercanviar dades.

[3] SPARQL és un acrònim recursiu del anglès SPARQL *Protocol and RDF Query Language*. Llenguatge estandaritzat per la consulta de grafs RDF, normalitzat per el RDF Data Access Working Group (DAWG) del World Wide Web Consortium (W3C).

[4] EuroSDR és una organització sense ànim de lucre que vincula agències nacionals de cartografia i cadastral amb instituts i universitats de recerca a Europa amb la finalitat de la investigació aplicada en subministrament, gestió i lliurament de dades espacials.

principal de valor (Duckham, Arnold, Armstrong, McMeekin, & Mottolini, 2017). Les SKI del futur hauran de possibilitar la creació, compartició, i ús de coneixements, més enllà de dades o informacions. En aquest sentit, l'ús de dades enllaçades semblen una bona oportunitat per acostar-s'hi.

En un escenari en el que podem tenir la informació relacionada i identificada inequívocament via URIs, la disponibilitat d'un llinatge geospacial pels nous conjunts de dades generats és encara més important degut a la complexitat i diversitat d'origen de les fonts i processos. És a la vegada també un escenari on el llinatge pot tenir un gran potencial, doncs amb un llinatge ben descrit i amb la disponibilitat d'identificar i enllaçar dades i processos via URIs, ens situaríem a prop de maximitzar els seu beneficis. Tot i això, la presència del llinatge és encara escassa i calen eines perquè aquesta es vegi incrementada tan en entorns IDE com en entorns *Linked Data*.

El capítol 2 d'aquesta tesi doctoral explora vies per documentar el llinatge amb una notació RDF. Els capítols 3 i 4 treuen profit dels estàndards OGC (en concret el WPS) per documentar el llinatge. Finalment el capítol 6 proposa mecanismes per documentar el llinatge de grans conjunts de dades i dissenya un sistema de consultes en un sistema servidor-navegador de mapes desenvolupat seguint estàndards OGC.

1.1.2 Models de metadades (i dades) geospacials

Un estàndard de metadades pretén establir una comprensió comuna de les característiques de les dades per assegurar-ne l'ús i la interpretació correcta per part dels usuaris. Així doncs, el coneixement subministrat per les metadades ajuda a entendre, controlar i gestionar la informació geogràfica. Les metadades són dades sobre dades de qualsevol forma (incloses aplicacions de serveis informàtics/web), sense importar en quin format (Danko, 2007). Si les metadades han de descriure les dades, per tal d'incrementar la seva efectivitat, aquestes no haurien de perdre la seva relació amb les dades. Per exemple, haurien d'enllaçar directament amb les pròpies dades (Masó, Pons, & Zabala, 2012).

Abans de presentar els diferents models de metadades, és necessari establir el marc dels principals models de dades geospacials. En el context SIG, un model de dades és una construcció matemàtica emprada per representar elements geospacials (p. ex. edificis, carreteres, muntanyes, etc) o variables de variació continua (p. ex. precipitacions, elevacions, etc) (Hoel, 2008). Aquesta definició també aplica a la IG disponible a la xarxa com a geoservei. Tradicionalment, la IG s'ha representat en dos models diferents, el model vectorial (model d'objectes o *features*) i el model ràster (model de *coverages*).Tant el model vectorial com el model ràster es guarden en suports informàtics en infinitat de variacions i formats. Al llarg d'aquesta tesi doctoral, tot i ser coneixedors d'altres models de dades (p. ex. el proposat per Goodchild (2013)), treballem fonamentalment amb informació geogràfica representada en els models vectorial i ràster disponibles tan en entorns SIG com en entorns distribuïts.

Tornant a les metadades, hi ha diverses variacions d'estàndards de metadades, com l'*Standard for Digital Geospatial Metadata (CSDGM)* (Metadata Ad Hoc Working Group, 1998) o la *Dublin Core*

Metadata Initiative (DCMI) (2020)*;* però en el camp geospacial la norma més utilitzada és la proposada per la ISO/TC211. En concret, les metadades geospacials han estat descrites principalment amb la sèrie d'estàndards de la ISO 19115. Aquesta sèrie es focalitza sobre components com la identificació, qualitat de dades (enllaçant amb la ISO 19157 (ISO/TC 211, 2013)), organització de dades espacials, referència espacial, informació de distribució, etc. Tot i que la norma defineix un extens nombre d'elements de metadades, s'estableix un "conjunt mínim" de metadades (*core*). Es tracta d'establir uns mínims per facilitar el descobriment i l'accés a les dades (Sanchez Maganto, Nogueras-Iso, & Ballari, 2008). A més, suporta la descripció en diferents nivells de granularitat de la informació: una sèrie de conjunts de dades geogràfiques, un conjunt de dades geogràfiques, un element geospacial[5] individual, un atribut[6], etc. Tot i això, la norma es centra en el nivell de conjunts de dades. Per altre banda, la sèrie ISO 19139 va ser dissenyada principalment per definir els esquemes d'implementació de la norma ISO 19115 en XML. Al llarg dels anys, aquestes dues sèries han evolucionat i algunes de les versions inicials han estat ampliades o substituïdes per normes més recents. La Taula 1 resumeix l'evolució de les diferents versions dels esquemes de les normes ISO 19115 i ISO 19139.

Taula 1: Resum dels diferents estàndards de metadades de IG. Els esquemes de totes aquestes especificacions són públics i accessibles via web (https://schemas.isotc211.org/schemas/19115/) (Font: Elaboració pròpia basat en (Brodeur, et al., 2020)).

Estàndard	Descripció	Estat	Any
ISO 19115, Informació geogràfica - Metadades	Defineix els elements de metadades i l'esquema que descriu conjunts de dades geoespacials	Substituït per la 19115-1:2014	2003
ISO 19115-1, Informació geogràfica -Metadata - Part 1: Fonaments	Defineix els elements de metadades i l'esquema que descriu conjunts de dades geoespacials	Publicat / Vigent	2014
ISO 19115-2, Informació geogràfica - Metadata - Part 2: Extensions per imatges i dades de malla	Amplia la norma ISO 19115 definint elements de metadades addicionals i esquemes per descriure imatges i dades de malla	Substituït per la 19115-2:2019	2009
ISO 19115-2, Informació geogràfica - Metadata - Part 2: Extensions per adquisició i processament	Amplia la norma ISO 19115-1 definint l'esquema necessari per a una descripció millorada de l'adquisició i el processament de informació geogràfica, incloses les imatges.	Publicat / Vigent	2019
ISO 19115-3, Informació geogràfica - Metadata - Part 3: Esquemes XML per conceptes fonamentals	Defineix una implementació XML integrada de ISO 19115-1, ISO 19115-2 i conceptes de la ISO 19139	Publicat / Vigent	2016
ISO 19139, Informació geogràfica - Metadata Esquemes XML d'implementació	Defineix els elements de metadades i l'esquema que descriu conjunts de dades geoespacials	Substituït per la 19115-3:2016 i per la 19139-1:2019	2007
ISO 19139-1, Informació geogràfica -Esquemes XML d'implementació - Part 1: Normes de codificació	Proporciona normes de codificació per implementar UML en XML.	Publicat / Vigent	2019
ISO 19139-2, Informació geogràfica - Metadata - Part 2: Esquemes XMLper imatges i dades de malla	Proporciona un esquema per implementar ISO 19115-2 en XML.	Substituït per la 19115-3:2016	2012

[5] "Element geospacial" és la traducció escollida de "feature". S'evita així l'ús d'objecte (que es reserva per referir-se a objecte en un model orientat a objecte) o de "fenomen".

[6] "Atribut" és la traducció escollida de "feature attribute"

1.1.3 Aproximació general al llinatge geospacial

Aquest apartat introdueix de manera sintètica les diferents definicions, conceptes i aspectes a tenir en compte quan es treballa amb llinatge; així com també els beneficis potencials del seu ús. En concret, els diferents subapartats introdueixen al lector les consideracions generals sobre el llinatge geospacial i formalitzen els termes que s'usaran a la resta del document.

La *Figura 1* resumeix i relaciona tots els aspectes relacionats amb el llinatge geospacial (representat en el globus central de color ocre). El globus de color verd fa referència als models de representació del llinatge (*subapartat*). Els globus de color blau fosc representen allò que, en el seu conjunt, hem anomenat *Gestió del Llinatge (subapartat 1.1.3.5).* Les diferents fases de gestió del llinatge geospacial són: els mecanismes de captura i d'emmagatzematge (*subapartat 1.1.3.5.1*), de visualització (*subapartat 1.1.3.5.2*) i de consulta (*subapartat 1.1.3.5.3*). Per altre banda, els globus representats en blanc fan referència a aspectes a tenir en compte quan s'està capturant el llinatge, com ara els elements que formen part del llinatge (*subapartat 1.1.3.2)*, els beneficis i aplicacions de l'ús del llinatge en el camp geospacial (subapartat 1.1.3.3), els models (subapartat 1.1.3.4) i aspectes clau (subapartat 1.1.3.4).

1.1.3.1 Llinatge i provinença[7] (provenance)

En el camp de la informació geospacial, els conceptes de llinatge i provinença (d'ara en endavant *provenance*) defineixen l'origen de la informació geogràfica i en molts contextos s'utilitzen com a sinònims. Tot i això, el seu significat és lleugerament diferent, si més no en el seu origen.

L'origen de la paraula llinatge prové del terme llatí antic *linea* (Grup Enciclopèdia Catalana, n.d.). El seu significat etimològic ens remet a la línia imaginaria que uneix els descendents directes d'un ancestre. El camp digital ha adaptat el terme per referir-se a l'origen de qualsevol objecte, és a dir, per referir-se a la descripció d'aquells objectes i processos implicats en la generació d'un objecte. Per objecte podem entendre qualsevol element present en el context digital (p. ex. pàgina web, conjunt de dades, document, fotografia, etc). En el camp de la IG, una de les primeres definicions de llinatge és la proposada per Lanter (1991), que defineix el llinatge geospacial com aquella part de les metadades que descriu les fonts i els processos utilitzats en l'elaboració d'un producte geogràfic determinat. En el seu sentit més ampli, la descripció dels processos inclou informació sobre les execucions dels processos, els algorismes, data d'execució i dels responsables, tan de les execucions com dels algorismes emprats. La descripció de les fonts inclou el format, l'autoria, la propietat, la data de creació i l'històric d'actualitzacions.

Mes endavant en el temps, el *Content Standard for Digital Geospatial Metadata (CSDGM)* (Metadata Ad Hoc Working Group, 1998) del *Federal Geographic Data Committee* (FGDC) va incloure un model de llinatge dins del model de qualitat de les dades geospacials i va establir que el llinatge és *"la informació dels esdeveniments, paràmetres, les fonts, i els agents participants en la construcció d'un*

[7] La traducció més directa al català del terme anglès *provenance* és procedència. No obstant això, en aquesta tesi hem preferit per traduir-ho com a provinença (acció de provenir: tenir l'origen en un lloc, alguna cosa). El terme procedència a vegades té una connotació relacionada amb el dret, i sembla suggerir un lloc origen i no un recurs d'origen.

conjunt de dades geospacials". Per la seva banda, la *International Organization for Standardization* (ISO) també va incloure un model de llinatge dins del model de qualitat de dades de la ISO 19115-1 i la ISO 19115-2 que defineix el llinatge com a "*provenance, fonts i eines emprades per produir un recurs*. El terme *provenance* utilitzat en la definició es refereix a "organització o individu que ha creat, acumulat, mantingut i utilitzat registres (ISO, 2017). Així doncs, és podria afirmar que el terme llinatge inclou la *provenance*.

Per altre banda, el terme *provenance* significa l'origen o la font d'alguna cosa (Lakshmanan, Curbera, Freire, & Sheth, 2011) (Moreau L. , 2010). Aquest concepte ha estat tradicionalment més vinculat a l'entorn web degut a l'Open Provenance Model (OPM) (Moreau, et al., 2011) inicialment, i al seu successor W3C-PROV (Moreau & Missier, 2013) (d'ara en endavant PROV). El model PROV s'ha consolidat com l'estàndard de referència per intercanviar *provenance* en entorns distribuïts (Magagna , et al., 2020). El PROV defineix la *provenance* com les entitats, activitats i persones o institucions implicades en la generació d'un objecte (Groth & Moreau, 2013).

Tot i aquestes lleugeres diferències terminològiques, són molts els autors que usen llinatge o *provenance* indistintament (Di, Yue, Ramapriyan, & King, 2013), (Yuan, Yue, Gong, & Zhang, 2013), (Yue, Zhang, Guo, & Tan, 2014), (Jiang, et al., 2018). Per referir-se al llinatge també son utilitzats termes com: historia, recepta, pedigrí, parentalitat, genealogia o filiació en funció del camp de treball.

En el context d'aquesta tesi doctoral, als capítols 2, 3, 4 i 5 s'han usat els termes llinatge i *provenance* com a sinònims. En canvi al capítol 6 s'ha introduït una lleugera diferenciació: l'ús del terme llinatge es utilitzat per descriure l'origen d'un conjunt de dades (*dataset*) i l'ús del terme *provenance* és utilitzat per descriure el llinatge de conjunts de dades ubicats en una catàleg de metadades o *ClearingHouse.*

1.1.3.2 Elements del llinatge

Per elements del llinatge entenem aquells aspectes que es capturen per tal de col·leccionar tota la informació de llinatge, com ara les fonts, les execucions, etc (Figura 2). Per tant, la completesa del llinatge queda definida per la presència o absència d'informació sobre els diversos elements, així com també per l'amplitud de la descripció de cadascun d'ells (Di, Shao , & Kang, 2013). Els elements descrits a continuació són una versió ampliada dels elements de llinatge descrits en el capítol 2 (Closa, Masó, Proß, & Pons, 2017) i emprada al capítol 6:

- Font: Informació o dades en els diferents possibles formats que han estat utilitzades en el procés de producció del resultat final. Poden ser referenciades mitjançant una citació descriptiva, un identificador, un identificador de metadades, una URI o una URI cap a les metadades. Inclou la data de creació i actualització, el sentit (input/output), així com també la vinculació cap a les execucions de les quals n'ha format part. La inclusió de la vinculació amb l'execució que l'ha generat permet ampliar el llinatge a processos ancestrals.

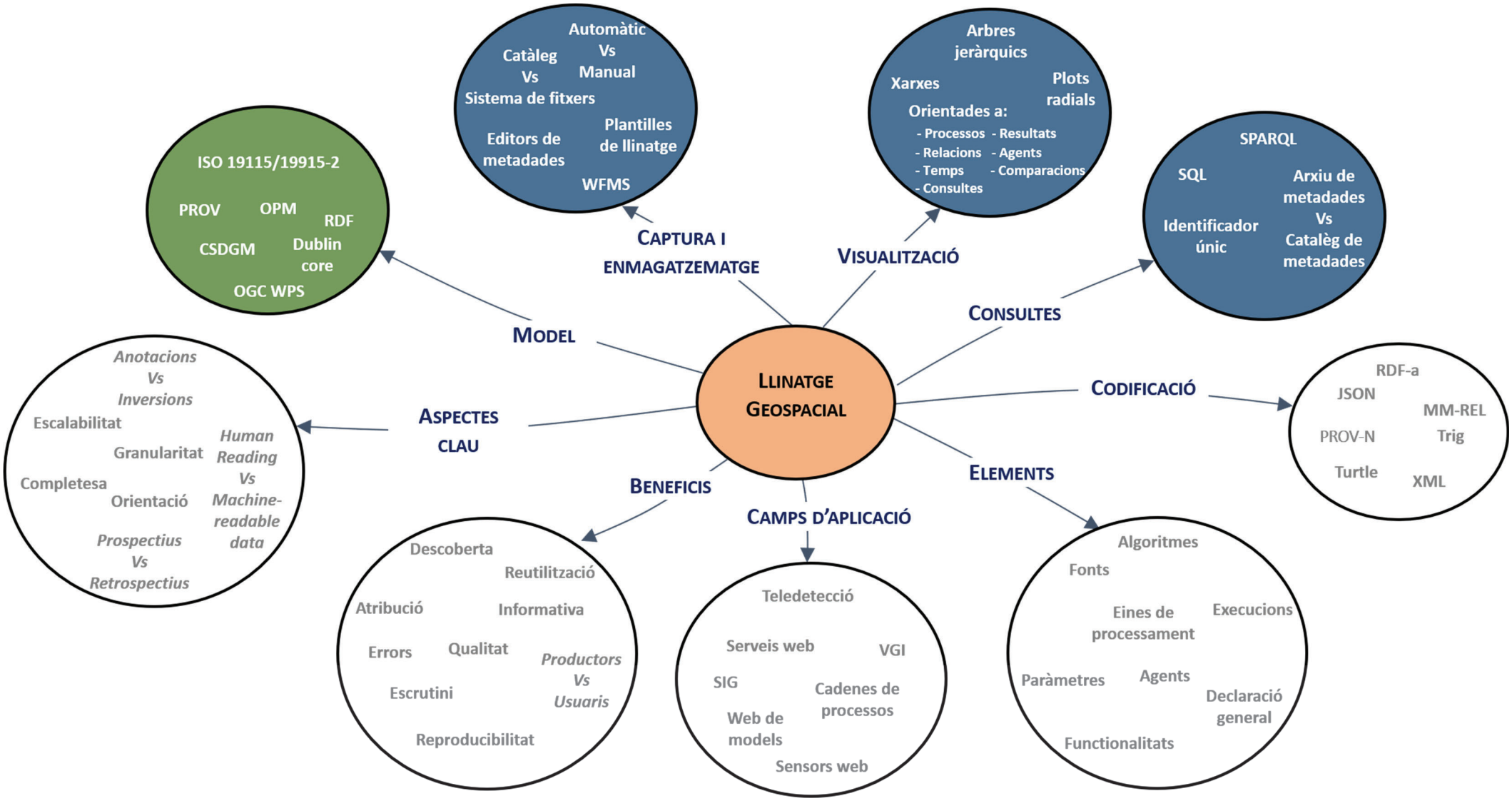

Figura 1: Aproximació general a la informació de llinatge geospacial (Font: Elaboració pròpia).

- **Execució** (procés): Operació realitzada manualment o amb una eina de processament que ha generat noves dades o informació o, si més no, ha generat noves versions. Pot ser referenciada mitjançant el nom de l'operació, una URI de l'operació o una descripció completa de l'operació. Inclou la data d'execució i la vinculació a l'eina de processament, així com també a les fonts usades (inputs) i IG generada (outputs).
- **Eina de processament**: Part de programari que implementa un algorisme i que pot ser executat moltes vegades en diferents fluxos de treball. Inclou la data de generació i/o actualització de l'eina, així com també la vinculació a l'algorisme que dirigeix l'eina.
- **Algorisme**: La descripció lògica de la implementació del procés. Inclou la data de generació i/o actualització de l'algorisme, el seu autor, així com també la vinculació a la funcionalitat que proporciona.
- **Funcionalitat**: Concepte o acció que descriu un algorisme i s'implementa en una eina de processament amb una orientació de resolució de problemes. Molts manuals de SIG descriuen llistes de funcionalitats.
- **Paràmetres**: Constant o variable que modifica el comportament de l'algorisme. Inclou el valor (en cas d'existir) i la vinculació cap a les execucions de les quals n'ha format part com a font. La inclusió de la vinculació amb l'execució que l'ha generat permet ampliar el llinatge a processos ancestrals.
- **Agents**: Persones o institucions que estan a càrrec o que han usat fonts, eines de processament i algorismes. Inclou el rol, la vinculació a fonts, eines de processament, execucions o algorismes.
- **Declaració general** (*statement*): Text lliure que inclou una explicació del procés complet de producció de les dades. Té una clara orientació *human-reading*[8].

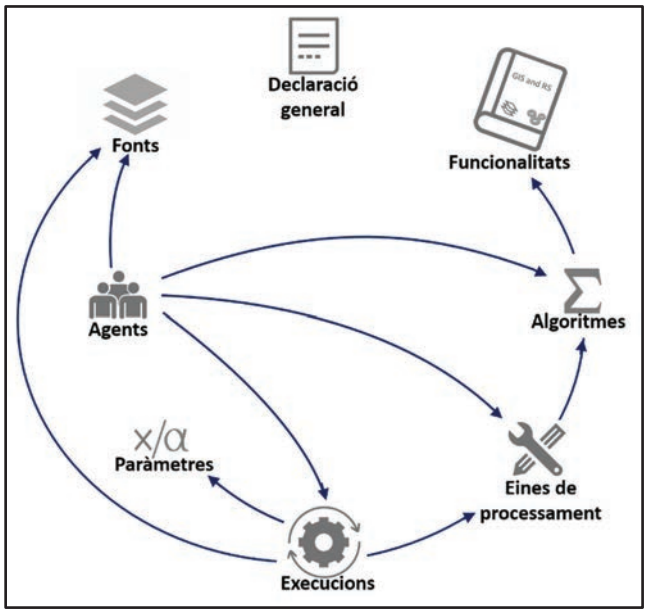

Figura 2: Elements que conformen el llinatge i les seves relacions bàsiques. (Font: Elaboració pròpia).

[8] Human-reading fa referència a dades o metadades que són interpretables per l'esser humà. En contraposició trobem el concepte de dades o metadades *machine-readable*, que es refereix a aquelles que son processables per ordinadors.

1.1.3.2.1 Exemple de llinatge

Per tal de mostrar el llinatge geospacial i les relacions entre els diferents elements que el formen, a continuació es planteja un cas d'ús en l'entorn administratiu. El llinatge representat en aquest exemple es descriu sense vincular-lo a cap model o llenguatge de codificació associat.

- *Cas d'us:*

  Infraestructures.cat, empresa pública de la Generalitat de Catalunya adscrita al Departament de la Vicepresidència i d'Economia i Hisenda, ha aprovat l'ampliació de dos a tres carrils per sentit de la carretera C-25 en el tram comprès entre Espinelves i Santa Coloma de Farners. Com a pas previ a la licitació de les obres, Infraestructures.cat ha encarregat al Negociat de Plans i Mapes del Departament de Territori i Sostenibilitat (DTES) la generació d'una cartografia amb les parcel·les cadastrals afectades per l'ampliació.

  Un cop el Negociat de Plans i Mapes ha realitzat l'encàrrec en el context d'un SIG, ha lliurat la següent IG:

  - *El conjunt de dades amb les parcel·les de cadastre afectades. Cada registre (parcel·la) conté la següent informació temàtica:*
    - *Referència cadastral*
    - *Propietari*
    - *Superfície total de la parcel·la (ha)*
    - *Àrea afectada de la parcel·la (ha)*
    - *Percentatge àrea afectada*
  - *Conjuntament amb les dades, es lliuren també les metadades associades a la IG generada. Les metadades recullen el procés de producció. La informació de llinatge derivada de la generació del conjunt de dades "parcel·les de cadastre afectades" és la presentada en les* taules següents (*Taula 2- Taula 8*):

Taula 2: Fonts que participen al procés de producció

Identificador	Descripció	Accès metadades	Data	Sentit	Execució
Carreteres_cat	Capa vectorial de carretetres	uri	20/12/2019	Input	Retalla_01012020_1335
Cadastro	Capa vectorial del Cadastre	uri	25/01/2020	Input	Retalla_01012020_1336
Carreteres_cat_ret	Capa vectorial de carretetres retallada	uri	01/01/2020	Output | Input	Retalla_01012020_1335 | Buffdist_01072020_1337
Cadastro_ret	Capa vectorial del Cadastre retallat	uri	01/01/2020	Output | Input	Retalla_01012020_1336 | Combicapa 01072020_1338
Àrea_afectació_50m	Area d'afectació	uri	01/01/2020	Output | Input	Buffdist_01072020_1337 | Combicapa 01072020_1338

Taula 3: Paràmetres que participen al procés de producció

Descripcio	Valor	Unitats	Data	Sentit	Execució
Envolupant de l'area de treball	Coord array (x,y)	UTM	na	Input	Retalla_01012020_1335 | Retalla_01012020_1336
Distància	50	m	na	Input	Buffdist_01072020_1337

Taula 4: Execucions que participen al procés de producció

Identificador	Descripció	Eina de procesament	Inputs/outputs	Sentit
Retalla_01012020_1335	Execució de l'eina Retalla del MiraMon	MM Retalla v10	Carreteres_cat | Carreteres_cat_ret	Input | Output
Retalla_01012020_1336	Execució de l'eina Retalla del MiraMon	MM Retalla v10	Cadastro | Cadastro_cat	Input | Output
Buffdist_01072020_1337	Execució de l'eina Buffdist del MiraMon	MM Buffdist v7	Carreteres_cat_ret | Àrea afectació 50m	Input | Output
Combicapa 01072020_1338	Execució de l'eina Combicapa del MiraMon	MM Combicapa v5	Cadastro_ret | Àrea afectació 50m | Parcel·les cadastre afectades	Input | Input | Output

Taula 5: Eines de processament utilitzades en el procés de producció

Identificador	Descripció	Data (versió)	Vinculació algoritme
MM Retalla v10	Eina que retalla capes a partir partir d'un objecte poligonal	12/12/2019	Extraction_algorithm
MM Buffdist v10	Eina que genera zones d'influència (buffers)	05/02/2020	Buffer_algorithm
MM Combicapa v5	Eina que combina dues capes espacialment	07/08/2020	Union_algorithm

Taula 6: Algorismes utilitzats en el procés de producció

Identificador	Descripció	Vinculació a la funcionalitat
Extraction_algorithm	Cada objecte de la capa des compara amb l'envolupant. Si aquest està dins o parcialment dins és incorporat a la capa de sortida.	Extraction
Buffer_algorithm	Algorisme que genera una capa a partir de representar la paral·leles dels segment a una distància donada.	Buffer
Union_algorithm	Algorisme que incorpora a la capa resultat cada objecte de les capes originals combinant els seus atributs temàtics.	Geometric Union

Taula 7: Funcionalitats utilitzades en el procés de producció

Identificador	Descripcio
Extraction	Selecció d'un conjunt d'elements d'una capa a partir de la seva localització.
Buffer	Area afectada per a la presencia d'un objecte
Geometric Union	Analisi de superposició espacial de dues capes

Taula 8: Agents que participen en el procés de producció

Identificador	Rol	Vinculació (font, execució, etc)
DTES	Creador/propietari	Carreteres_cat
Dirección general de Catastro	Creador/propietari	Cadastro
Negociat de Plans i Mapes	Creador	Retalla_01012020_1335 | Retalla_01012020_1336 | Buffdist_01012020_1337 | Combicapa_01012020_1338 | Carreteres_cat_ret | Cadastro_ret | Àrea_afectació_50m
Equip MiraMon	Desenvolupador	Retalla_V3 | Buffdist_v5 | Combicapa_v4
Infraestructures.cat	Propietari	**Parcel·les Cadadastre afectades**

- *Aquesta informació, si està degudament estructurada i és comprensible, pot servir a tercers per consultar el procés de producció del conjunt de dades (Figura 3) i respondre una sèrie de qüestions com per exemple:*
  - *El conjunt de dades resultat, ha estat realitzada amb fonts oficials?*
  - *Quines eines s'han emprat?*
  - *Qui ha generat (institució) el conjunt de dades resultat?*
  - *Quin ha estat el procés de producció complet per generar el conjunt de dades?*

- S'ha emprat la darrera versió del cadastre?
- De quina data és el mapa de carretes emprat?

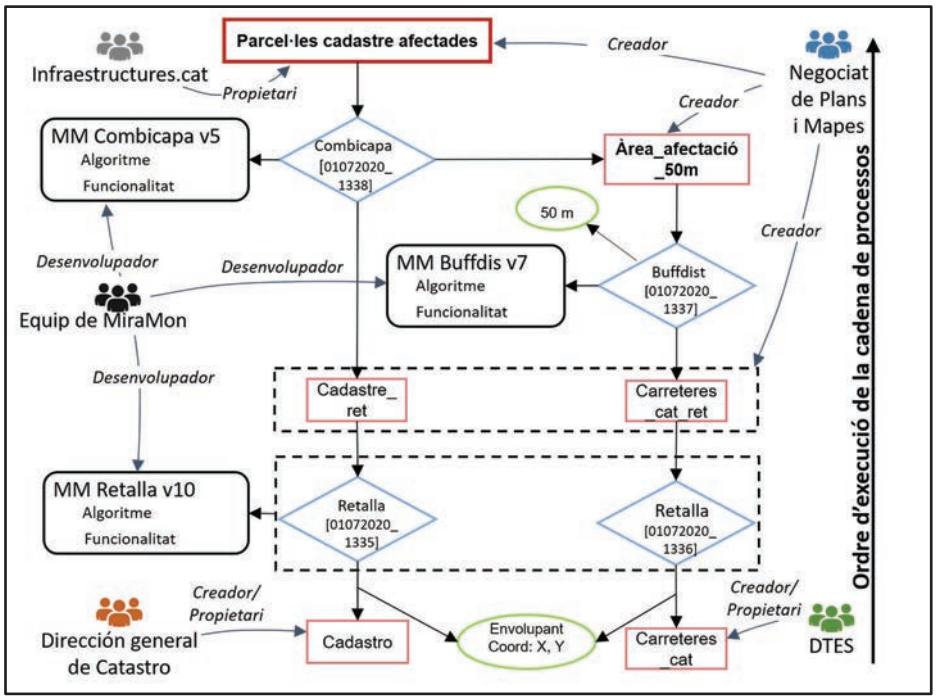

Figura 3: Llinatge del conjunt de dades "*Parcel·les cadastre afectades*". Els rectangles amb la vora de color vermell clar simbolitzen les fonts, els rombes blaus les execucions, cercles verds els paràmetres i els rectangles negres (amb cantonades arrodonides) les eines de processament que estan vinculades a l'algorisme i la funcionalitat. (Font: Elaboració pròpia)

1.1.3.3 Beneficis del llinatge

Descriure amb detall els diferents elements del llinatge de la IG en entorns científics i administratius aporta diversos beneficis. A continuació es descriuen els principals beneficis derivats de la presència d'informació de llinatge en les metadades.

- Informativa: El llinatge proveeix una visió general del procés de producció de les dades.
- La voluntat i la necessitat de transparència: Amb el ràpid increment de la compartició de dades en entorns distribuïts, les institucions necessiten mostrar l'origen de les dades per tal de mantenir la credibilitat i la reputació.
- Escrutini: En els règims democràtics, les decisions preses pels governants estan sota escrutini del poble. El llinatge és un component per l'escrutini dels processos que han conduit al coneixement que ha permès una presa de decisions informades.
- Atribució: Útil en el procés de verificació de la propietat i reconeixement dels drets d'autor de les dades. Així, els autors poden obtenir reconeixement de qui usa les dades que generen. En el camp científic, les citacions són importants per determinar l'impacte; en aquest sentit el llinatge pot actuar com a citació de conjunts de dades. També és pot emprar per determinar la responsabilitat en cas d'errors en les dades i processos.

- **Qualitat de les dades**: La qualitat de les fonts, eines i processos està relacionada amb la qualitat dels resultats de l'anàlisi de dades, alhora que ajuda a determinar, d'una forma concisa, l'adequació d'ús de les dades en els diferents projectes.
- **Descoberta**: El llinatge pot ser emprat per descobrir dades, eines i agents que estan relacionats entre sí. Per exemple, el llinatge ens permet descobrir l'existència de fonts de dades que no coneixíem.
- **Localització d'errors i avaluació de la seva propagació**: Analitzant el llinatge podem retrocedir en les cadenes de processament per localitzar els possibles errors en la selecció de fonts o processos. A més, l'anàlisi del llinatge de conjunts de dades ens pot ajudar a determinar sobre quines dades s'ha utilitzat determinada font o eina que hauria contingut errors.
- **Reproductibilitat, verificació científica i comprensió dels resultats propis o aliens**: La base per obtenir bons resultats científics està altament vinculada a una bona metodologia de treball que quedi ben documentada i permeti la reproductibilitat. La documentació de dades, eines i metodologies emprades és essencial, fins i tot per aquelles que finalment són descartades.
- **Reutilització de les dades i processos**: La disponibilitat de la recepta completa de la generació d'un producte determinat pot afavorir la reutilització de cadenes de processament com a fluxos de treball en altres contextos i àmbits geogràfics, emprant el mateix o un altre programari.

Aquests beneficis estan directament relacionats amb la completesa de la descripció del llinatge. Per tant, quan es decideixen quins elements de llinatge es capturen, s'ha de considerar també quins són els beneficis que se n'espera obtenir. En aquest sentit, la Taula 9 relaciona els beneficis (columnes) amb la completesa de la descripció del llinatge basada en la presència dels diversos elements (files). Per a cada benefici del llinatge, s'han marcat aquells elements que es consideren mínims imprescindibles, amb el benentès que la presència de més elements pot incrementar encara més el benefici.

Taula 9: Relació entre elements de llinatge (files) i els seus beneficis (columnes). Els elements marcats indiquen els mínims necessaris per a acomplir el benefici. (Font: Elaboració pròpia)

	Informativa	Transparència	Escrutini	Atribució	Qualitat	Descoberta	Errors	Reproducibilitat	Reutilització
Declaració general	✓	✓		✓					
Fonts	✓	✓	✓		✓	✓	✓	✓	✓
Paràmetres			✓		✓		✓	✓	
Execucions			✓				✓		
Eines de processament		✓			✓	✓	✓	✓	
Algoritme						✓			✓
Funcionalitat	✓								✓
Agents	✓	✓	✓	✓		✓			

Les necessitats i beneficis esperables del llinatge poden variar si es situen en el punt de vista del productor, o en el punt de vista de l'usuari de la IG (Spiekermann, Jolly, Herzig, Burleigh, & Medyckyj-Scott, 2019). En termes generals, els productors necessiten que el model sigui comprès i assegurar la

transparència del procés de producció per tal d'oferir, no només el producte en si (la IG), sinó també les bones pràctiques i la metodologia (Özkundakci, Wallace, Jones, & Hunt, 2018). Per altra banda, els usuaris estan més interessats en la determinació de la qualitat de les dades que s'usaran, assegurar la traçabilitat entre resultats obtinguts i dades utilitzades, i la reproductibilitat de dades i processos en altres contextos. La Taula 10 mostra els beneficis dels llinatge classificats en aquells que tenen una clara orientació als usuaris, i aquells que s'orienten més cap als productors.

Taula 10: Relació entre beneficis del llinatge (files) i la seva utilitat per als usuaris o per als productors de la informació geogràfica (columnes). (Font: Elaboració pròpia)

	Productors	Usuaris
Informativa		✓
Transparència	✓	
Escrutini		✓
Atribució	✓	
Qualitat		✓
Descoberta		✓
Errors	✓	
Reproducibilitat	✓	✓
Reutilització		✓

1.1.3.3.1 Camps d'aplicació

El potencial del llinatge varia dins dels diversos camps d'aplicació de la IG. A continuació es resumeixen els principals camps de la IG on el llinatge pot ser útil:

- Sistemes d'Informació Geogràfica (SIG): La incorporació d'informació de llinatge en el context d'un SIG pot ajudar a l'ús i explotació de la IG en molts dels aspectes recollits en el subapartat 1.1.3.3. El cas d'ús presentat al subapartat 1.1.3.2.1 n'és un clar exemple. Històricament, diversos treballs han valorat la captura del llinatge en el context SIG (Lanter, 1991) (Veregin & Lanter, 1995) (Alonso & Hagen, 1998) (Wang, Padmanabhan, Myers, & Tang, 2008). Encara que la majoria de SIG poden guardar les execucions generades en fitxers *log* que poden actuar parcialment com a metadades, només alguns SIG guarden algunes metadades de llinatge automàticament durant el procés de producció de les dades. En aquesta tesi presentem una implementació per formalitzar el llinatge en el SIG MiraMon.
- Teledetecció (TD): Un producte de dades de TD normalment es deriva de dades en brut adquirides per sensors mitjançant un flux de treball de geo-processament que eleva productes des de nivell 0 fins a un nivell on les dades són directament interpretables. Si la informació de llinatge es capta i es conserva adequadament, tant el flux de treball de processament com la

informació de qualitat es poden rastrejar a través de les metadades adjuntades amb el producte final. Això té utilitat en el posterior processament del producte o en la replicació del procés (Di, Shao , & Kang, 2013). La informació del llinatge és un requeriment en els productes d'*Analysis Ready Data* especificats pel CEOS (*Committee on Earth Obsebation Satellites)* (CEOS, 2020) segons s'indica en els Product Family Specifications produïts pel grup CARD4L. El capítol 5 d'aquesta tesi doctoral presenta un exemple de llinatge lligat a la teledetecció.

- Serveis web: Els serveis web basats amb dades obertes que aporten valor afegit (millors visualitzacions, densificació de productes a àmbits més locals, etc). És necessari conèixer l'origen dels productes i les incerteses generades en els processos realitzats. A més, pot ajudar a discernir com diferents conjunts de dades estan indirectament relacionats entre sí pel fet d'usar fonts comunes. Henzen (2016) indica que el llinatge pot ser emprat per identificar gaps o duplicitats.
- Web de Sensors: Els sensors són una font constant de dades que necessiten de processos de control de qualitat que haurien de ser presents en el llinatge de les dades. Les observacions realitzades pels sensors s'estandarditzen a partir de l'estàndard d'Observacions i Mesures (O&M). Cox (2017) proposa adaptar el model O&M al model PROV. Per altre banda, Jiang (2017) suggereix estendre el model PROV per incloure vocabulari propi de la web de sensors.
- Informació Geogràfica Voluntària - *Voluntered Geospatial Information* (VGI). En aquest tipus d'informació, nombrosos actors amb eines i experiència diferent contribueixen a la captura de dades que es consoliden en un sol conjunt de dades. En aquest cas, el llinatge de cada observació individual és potencialment diferent i, per exemple, pot ajudar a realitzar controls de qualitat basats en la reputació de cada observador (Celino, 2013) (Keßler & De Groot, 2013).
- Fluxos de treball – *Workflows*: En la descripció de les cadenes de processament es pot acumular molta informació de llinatge i, al mateix temps, el llinatge pot ser emprat com a font per replicar les cadenes de processament. En aquest segon cas, podem parlar de llinatge prospectiu (Lim, Lu, Chebotko, & Fotouhi, 2010).
- Web de models – *Model web:* la Web de Models és un concepte general que proposa incrementar l'accés als models i als seus resultats per augmentar la interacció entre els diferents models científics en entorns distribuïts (Nativi, Mazzetti, & Geller, 2013). Generar interdependències entre diferents models científics incrementa les capacitats de modelització i de predicció de futurs escenaris (Geller & Turner, 2007). En aquest àmbit, la incorporació del llinatge a les metadades geospacials pot ajudar al desenvolupament de la web de models.

1.1.3.4 Models de llinatge

La representació del llinatge inclou el model de dades emprat i la seva codificació i implementació en un llenguatge concret. Es podria pensar que una simple declaració general descriptiva (*statement*) pot ser suficient per documentar el llinatge, però ja s'ha comentat que encara que els humans podem llegir i extraure elements, un statement no pot ser usat per algorismes automàtics. En realitat és necessària una estructura formal reconeguda per extreure'n el màxim benefici. A més, si el model és interoperable, permetrà l'intercanvi i la compartició del llinatge en un entorn d'informació distribuït, millorant-ne la usabilitat i sobretot la utilitat (He, Yue, Di, Zhang, & Hu, 2015). Segons (Myers, Futrelle,

Gaynor, & Plutchak, 2009), els registres de llinatge han de ser digitalment visibles, accessibles i comprensibles per tal que permetin accedir al procés de producció de les dades i oferir la informació de context necessària per reproduir la informació i els resultats. Bona part d'aquestes qualitats depenen del model de metadades emprat.

No hi ha només un sol model de metadades que s'ajusti a totes les disciplines; ans al contrari, existeixen diversos models de representació i cadascun afavoreix les particularitats i necessitats del seu camp d'aplicació original. En el camp geospacial els models de llinatge es defineixen per les següents característiques:

- Completesa: Queda definida pels elements del llinatge que captura (veure subapartat 1.1.3.2) i el detall amb que els descriu.
- Codificació: Fa referència als llenguatges utilitzats per descriure el llinatge. Acordar un llenguatge de codificació incrementa la interoperabilitat (algunes alternatives són XML, RDF-a, JSON).
- Granularitat: És la capacitat de capturar el llinatge a diferents nivells de representació de la informació geospacial. Es pot representar a un nivell fi (p. ex. a nivell d'atribut), o a un nivell més general (p. ex. a nivell de *conjunt de dades*). El cost de capturar i representar el llinatge és inversament proporcional al nivell de granularitat. El Capítol 2 aprofundeix sobre les possibilitats de documentar el llinatge a diferents nivells de granularitat.
- Orientació: Indica quin és l'element del llinatge a partir del qual es pot representar. Per exemple, es pot descriure el llinatge com una successió de processos executats, com una successió de fonts utilitzades, o també pot estar orientat al temps o als agents implicats.
- Escalabilitat: És la capacitat de donar suport a la representació del llinatge de grans volums de dades sense comprometre el rendiment el sistema. La necessitat de descriure el llinatge complet pot derivar en la descripció recursiva de l'origen de les fonts intermèdies, fet que incrementa el volum i els costos d'emmagatzematge.
- Anotacions vs Inversió: Simmhan (2005) diferencia dos aproximacions per documentar el llinatge. 1) Anotacions (*Annotacions*): El model captura les fonts, processos i tota la informació rellevant del procés de creació o edició d'IG. 2) Inversió (*inversion*) es capturen les funcions o geoprocessos juntament amb dades de sortida per tal d'identificar les fonts de dades. La *inversió* té l'avantatge que la informació és més compacta que les *anotacions*, encara que no és universalment aplicable a totes les funcions. En aquesta tesi doctoral hem utilitzat només models d'anotacions.
- Models retrospectius vs Models prospectius: Els models retrospectius capturen la informació de tasques ja realitzades. Es focalitza en capturar la transformació de les dades durant i posteriorment a l'execució. Els models i les eines presentades al llarg d'aquesta tesi, es basen en models retrospectius que escriuen el llinatge a les metadades del resultat. Per altre banda, els models prospectius modelitzen tasques computacionals futures; el que en aquesta tesi anomenem fluxos de treball (*workflows*). Conté la recepta per a crear noves dades en base a requeriments específics.

En el camp geospacial s'han utilitzat diversos models de llinatge. Alguns han estat concebuts per descriure metadades geospacials (p. ex. ISO o FGDC); en canvi d'altres han estat dissenyats amb una orientació clara d'intercanvi de dades en entorns distribuïts (p. ex. OPM, PROV) i s'han aplicat posteriorment al camp geospacial. També podem afegir un tercer grup que són aquells estàndards del camp de la geoinformació que han estat ideats amb altres finalitats, però que la seva naturalesa fa que incloguin molta informació de llinatge (p. ex. WPS). A continuació és descriuen els principals models de llinatge revisats al llarg d'aquesta tesi doctoral:

- ISO 19115: Actualment, el comitè ISO/TC 211 defineix l'estàndard de metadades per a la informació geogràfica a l'ISO 19115-1 (ISO/TC 211, 2014) i a la seva extensió orientada a imatges, l'ISO 19115-2 (ISO/TC 211, 2019). El model de llinatge (*Figura 4*) (LI_Lineage) s'inclou dins el paquet de qualitat de l'ISO 19115-1. Aquest està format per una declaració o descripció general (*statement*) i per la definició de les fonts (*LI_Sources*) i els processos (*LI_ProcessStep*). Per la seva banda, l'ISO 19115-2 estén el model afegint especialitzacions de les fonts (*LE_Sources*) i dels processos (*LE_ProcesssStep*). LI_Source/*LE_Source* descriu la informació d'entrada i sortida (p. ex. extensió espacial, escala, sistema de referència, enregistrament a nivell de procés i resolució d'imatges) relacionada amb cada pas del procés de generació de les dades. Per la seva banda LI_ProcessStep/*LE_ProcesssStep* serveix com a contenidor de la informació de processament (p. ex. estat d'execució, temps d'execució, paràmetres d'execució, sortida, informació de processament i informe) de cada pas. Cal fer notar que la revisió de l'ISO 19115-2, que va donar lloc a la darrera versió del document, es va produir durant la realització de la recerca que ha donat lloc a aquesta tesi doctoral. Algunes de les propostes aquí presentades han estat adoptades per la 19115-2 com a fruit de la participació del CREAF en els processos de consens de la creació de la 19115-2.
- Open Provenance Model (OPM): És un model basat en la representació i relació entre 3 objectes primitius: agents, processos i artefactes. Fou dissenyat per satisfer els requisits següents (Moreau, et al., The Open Provenance Model core specification (v1.1), 2011): 1) permetre l'intercanvi de llinatge, 2) permetre als desenvolupadors construir i compartir eines que operin amb aquest model, 3) definir el llinatge d'una manera precisa, 4) donar suport a una representació digital del llinatge de qualsevol àmbit, ja sigui produïda per sistemes informàtics o analògics, 5) permetre la coexistència de diversos nivells de descripció i 6) definir un conjunt bàsic de regles per la representació del llinatge. En el camp geospacial s'han fet diversos esforços per definir una transformació entre el llinatge representat amb família ISO 19115 i l'OPM (Feng, 2013). L'OPM també disposa d'un esquema en XML per representar el model, *Open Provenance Model XML Schema* (Moreau & Missier, 2013).

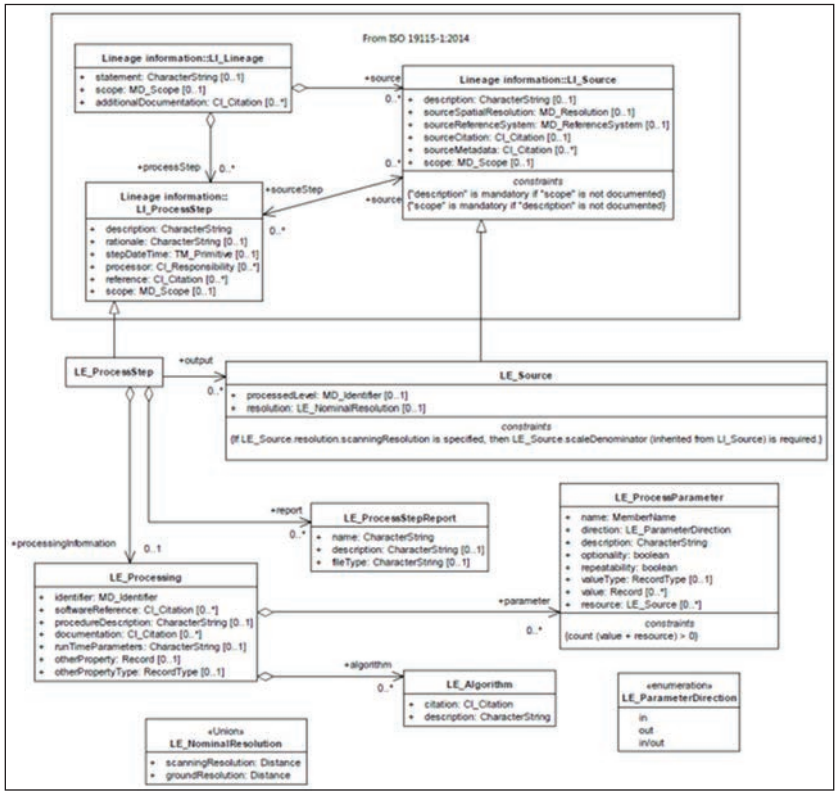

Figura 4: Diagrama UML del model de llinatge de l'ISO 19115-1 i 19115-2 (Font: ISO 19115-2)

- W3C PROV: És un model conceptual altament estructurat per a la codificació del llinatge que permet el seu intercanvi entre sistemes i l'ús automatitzat a la web (W3C, 2013b). A més, la seva estructura modular i flexible permet descriure l'origen de les dades a diversos nivells de granularitat d'una manera natural i escalable. Representa una evolució de l'OPM i descriu la provenance com la informació de les entitats (fonts i eines de geoprocessament), agents (persones o institucions) i activitats (execucions) (Figura 5) involucrades en el procés de producció de dades. Existeixen especificacions per codificar-lo en OWL, PML, XML, PROV-N i JSON. Són nombrosos els treballs que adapten l'ús d'aquest model al camp geospacial (veure capítols 2 i 6), entre elles el Prov-ES; una extensió del model desenvolupat pel grup de treball de Sistemes de dades de la Terra (ESDS) de la NASA per tal d'estandarditzar l'ús del llinatge i adaptar-lo al camp de les ciències de la terra (Hua & Tilmes, 2013).

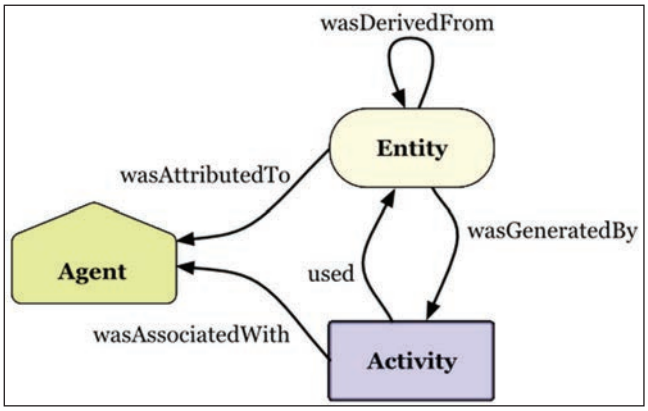

Figura 5: Elements del nucli del W3C PROV (Font: W3C PROV Data Model)

- CSDGM: És el model de metadades geospacials que fou generat i avalat pel FGDC farà ja més de 20 anys. Anàlogament al model ISO, el model de llinatge es troba dins el model de qualitat de les dades.
- OGC Services: Di (2013) afirma que els serveis OGC, tant els orientats a dades *(WMS, WFS, WCS)* com els orientats processos *(WPS),* aporten informació útil que pot complementar la informació de llinatge. El *WPS,* tot i que no ha estat dissenyat per descriure el llinatge del servei que implementa, aporta implícitament informació del llinatge. En aquest sentit els capítols 3 i 4 introdueixen l'ús del model WPS per complementar la informació del llinatge.
- Dublin Core: La *Dublin Core Metadata Initiative* (DCMI) (2020) promou les bones pràctiques en el disseny de metadades de recursos web. El nucli del model està format per quinze elements entre els quals hi ha la font, l'autor, el col·laborador, l'editor, les relacions, l'identificador del recurs, la data i la cobertura (espacial o temporal). Aspectes considerats com a elements basics del llinatge de la IG. A més dels elements principals, el model disposa d'una llarga llista de qualificadors per matisar o afinar cadascun dels elements, entre els quals trobem el *provenance statement.*

1.1.3.5 Gestió del llinatge

Per incrementar l'ús del llinatge és necessària l'existència de mecanismes que documentin i posin a disposició dels usuaris la informació en un llenguatge i format coneguts. A tal efecte, són necessaris sistemes capaços de capturar, emmagatzemar, visualitzar i consultar el llinatge. Aquestes fases conformen la gestió completa del llinatge, des del model de dades en la fase de documentació, fins a l'explotació de les dades en les fases visualització i consulta. Això és el que, en aquesta introducció, anomenem la *Gestió del llinatge*. El més habitual és trobar-nos amb diversos subsistemes que implementen alguna de les fases esmentades. Per exemple, es comú que les tasques de captura i emmagatzematge siguin realitzades pels productors de la informació, mentre que siguin els usuaris els més interessats en la visualització i la consulta del llinatge. La resta d'aquest subapartat descriu les diferents fases de la *Gestió del llinatge*.

1.1.3.5.1 Captura i emmagatzematge

Més enllà del model triat per representar el llinatge, el sistema de gestió de llinatge n'ha d'assegurar una correcta captura i emmagatzematge. Pel que fa referència a la captura (de llinatge retrospectiu), el procés pot ser manual (editant, a posteriori, les metadades associades al producte final resultant del procés de producció de la generació de la IG), o bé automàtic (mitjançant eines que registrin el llinatge concurrentment a l'execució dels processos que donaran lloc al producte final).
L'edició manual, a més de ser una tasca tediosa, tendeix a generar metadades incompletes, obre la porta a la introducció d'errors i fa que dades i metadades estiguin desconnectades entre sí (Kalantari, Olfat, & Rajabifard, 2010) (Giuliani, Ray, & Lehmann, 2013), podent trobar-se (dades i metadades) en diferents fases de producció tot i formar part d'un mateix conjunt. L'existència d'eines que guiïn la documentació de les metadades pot ajudar a minimitzar aquestes problemàtiques. En aquest sentit, podem definir tres nivells de captura manual:

- Editors de text lliure: El productor de metadades escriu els noms dels elements i el seu contingut com a text, així com també les relacions entre els elements. Generalment es comença per una plantilla. Com que el format de text emprat acostuma a ser un dialecte de l'XML requereix de molt entrenament. Més enllà de l'edició del contingut d'un element de llinatge de forma puntual o del resum general del procés de producció (*statement*), té poca utilitat per la seva baixa relació cost-benefici i la facilitat d'introduir errors de format.
- Ús d'editors de metadades i esquemes de metadades: Aquesta opció permet entrar el contingut dels diferents elements com a text lliure, però l'ús d'interfícies gràfiques, com puguin ser el CatMEdit (Advanced Information Systems Laboratory. Universidad Zaragoza, 2012), juntament amb esquemes de metadades (p. ex. el *XML Schema Definition* de l'ISO 19115-1), guien en la tipologia d'elements a definir, així com també les relacions entre els diferents elements.
- Ús d'editors de metadades i plantilles: A més dels esquemes i les interfícies gràfiques, l'editor de metadades disposa de plantilles que contenen descripcions completes de les diferents eines de geoprocessament (descripció, sintaxi, fonts, sortides, etc); com per exemple fa el Gestor de Metadades del MiraMon (GeMM). Això permet estandarditzar, en el sentit de l'ús de llistats de definicions tipificades, les descripcions del llinatge, reduint les possibilitats d'introducció d'errades i descuits.

Mes enllà de l'edició manual, són necessàries eines que capturin el llinatge automàticament per augmentar-ne la seva presència en les metadades. A més, la captura automàtica del llinatge elimina part de les problemàtiques pròpies de la captura manual (p. ex. registres incomplets, elevat cost, errors de transcripció, etc). En ocasions però, la captura automàtica pot generar registres massa detallats que continguin passos irrellevants per entendre el producte final. Llavors, les eines d'edició manual del llinatge son també necessàries i complementaries per polir el resultat inicialment generat pels sistemes de captura automàtica.

Existeixen exemples per capturar el llinatge geospacial de forma automàtica i concurrentment a l'execució dels processos, tal i com descriuen els següents treballs:

- Spiekermann (2019) presenta dos desenvolupaments per capturar el llinatge. 1) Pyluc és una eina escrita en Python que enregistra el llinatge derivat de la generació de classificacions de cobertes. Utilitza el model PROV-DM (Moreau & Missier, 2013) i emmagatzema el llinatge en la notació PROV-N (Moreau & Missier, 2013). 2) Desenvolupament realitzat sobre LUMAS (Herzig, Rutledge, Aus, & Dymond, 2019) per capturar el llinatge a diversos nivells de granularitat. Anàlogament a l'anterior exemple utilitza el model PROV-DM i emmagatzema el llinatge en la notació PROV-N.
- Di (2013) desenvolupa un mòdul sobre GeoBrian (Di, Han, Zhao, Wei, & Li, 2008) que captura el llinatge basat en els models de l'ISO 19115 i ISO 19115-2 com a arxius XML.
- Zhang (2017) demostra la captura automàtica del llinatge d'una simulació d'escolament superficial. El model es basa en OpenMI (Gregersen, Gijsbers, & Westen, 2007) per definir els elements de llinatge a capturar i escriure'l en PROV-XML i PROV-O.

També hi ha diversos exemples de Sistemes de Gestió de Fluxos de Treball - *Workflow Management Systems* (*WFMS)* que contenen funcionalitats per capturar el llinatge prospectiu:

- Kepler: Permet automatitzar i gestionar processos complexos i grans quantitats de dades. Conté un mòdul que proporciona l'enregistrament de l'historial del flux de treball.
- Pegasus: Totes les tasques capturen el llinatge automàticament La informació capturada s'emmagatzema en una base de dades i amb diverses eines o via SQL.
- Taverna: Software que permet dissenyar i executar fluxos de treball científics. El sistema captura el llinatge dels fluxos de treball incloses les iteracions individuals, les seves entrades i sortides. Aquesta informació es conserva en una base de dades interna. El llinatge es pot exportar com a PROV-O i pot ser consultat via SPARQL.

Existeixen esforços per estandarditzar i incrementar la interoperabilitat dels llenguatges que empren els WFMS, com per exemple el *Common Workflow Language* (CWL) (Amstutz, Crusoe, & Tijanić).Tot i això, els sistemes de captura del llinatge inclosos en els diversos WFMS, tenen la limitació que capturen el llinatge dins del seu context de treball; és a dir, que hauríem de recórrer a l'edició manual en cas que executéssim els fluxos de treball en un hardware diferent del que executa el WFMS (Magagna , et al., 2020). Per altra banda, exemples per capturar el llinatge, o bé són exemples de processament específics, o bé no estan pensats per executar-se en local (p. ex. GeoBrain).

Des de la introducció dels estàndards de metadades basats en XML, el llinatge de la IG s'ha emmagatzemat com una part més de la resta de metadades. Les metadades les podem trobar com a fitxers XML individuals, o formant part de catàlegs de metadades, com per exemple el GeoNetwork. En canvi, en el mon web es fa servir el W3C PROV on el llinatge és pot descriure amb RDF en les codificacions JSON, PROV-N, Trig, Turtle o RDFa.

En relació a l'emmagatzematge, els volums d'informació de llinatge poden arribar a ser considerables, en comparació als volums de la resta de les metadades. Aquest fet pot generar problemes d'espai d'emmagatzematge. En casos en els que s'emmagatzema el llinatge amb un nivell de granularitat molt fi (p. ex. nivell d'element o d'atribut), el volum d'informació pot augmentar significativament fins a ser més gran que les pròpies dades. Així doncs, són necessàries tècniques d'estalvi d'emmagatzematge. En aquest sentit, els sistemes han de ser capaços enllaçar la informació amb els nivells més alts de documentació per no malbaratar espai. Això es pot aconseguir fàcilment en els formats RDF, on podem emprar tripletes que incloguin a nivells superiors o usar els mecanismes de relacions entre documents de metadades ISO.

1.1.3.5.2 Visualització

Aspectes importants alhora de millorar la comprensió de les dades són les tècniques de visualització i simbolització utilitzades. En contextos amb un gran volum de dades, la comunitat científica ha confiat tradicionalment en les eines de visualització (Salton, Allan, Buckley, & Singhal, 1994).

En el cas del llinatge, les seves estructures complexes amb múltiples relacions i dependències entre els diferents elements que el formen, poden fer que els usuaris es sentin aclaparats quan exploren els diferents passos que han conduit a la generació d'un conjunt de dades. Així doncs, existeix la

necessitat d'eines que facilitin la interpretació del llinatge ja que aquestes tenen un impacte directe en la comprensió i ús de les pròpies dades (Yue, Gong, & Di, 2010). Les eines de visualització del llinatge han de ser escalables i tenir capacitats de generalització i d'abstracció dels diferents nivells de detall de la informació. A més, han de permetre la navegació per la informació (Chen, Plale, & Cheah, 2012). També han de tenir capacitats per visualitzar els resultats intermedis o parcials, processos derivats o qualsevol informació relativa a les fonts utilitzades (Del Rio & Da Silva, 2007). En definitiva, un major control per part dels usuaris sobre els gràfics i la quantitat d'informació que representen augmentaria la comprensió d'aquests (Wacharamanotham, Subramanian, Borchers, & Völkel, 2015).

Segons (Kunde, Bergmeyer, & Schreiber, 2008) podem trobar diversos tipus de visualitzacions de la informació del llinatge en funció dels requeriments de l'usuari:

- Orientades als processos: El centre de la visualització és la seqüència dels processos.
- Orientades als resultats: El centre de visualització són el resultats finals i intermedis.
- Orientades a les relacions: S'avaluen les interaccions entre els agents implicats, els processos i les fonts.
- Orientades als agents: En aquesta representació els agents implicats son el centre de la visualització en els que s'inclouen també elements de credibilitat dels agents.
- Orientades al temps: Es representa com una seqüència ordenada en el temps.
- Orientades a comparacions: Ressalten les diferències entre dos fluxos de treball o processos productius.
- Orientades a consultes: Es visualitzen els resultats de consultes sobre el llinatge.

En funció de la seva morfologia, les tècniques de visualització del llinatge es poden agrupar en les següents tipologies principals (Borkin, et al., 2013):

- Arbres jeràrquics: Presenten una estructura similar a un arbre genealògic, amb propietats o atributs jeràrquics. Les tècniques de visualització d'arbres jeràrquics són útils per representar el llinatge i les cadenes de processos lineals queden representades d'una forma natural. Tot i això, la representació de dependències múltiples pot resultar una mica confusa i poc natural.
- Xarxes: Basades en nodes i arcs que connecten nodes. Les dependències i relacions entre elements es representen de manera més natural que en el cas anterior. Tot i això, poden resultar una mica complexes de seguir, sobretot perquè les xarxes poden recórrer en qualsevol direcció i no necessàriament representen una seqüència temporal i es pot perdre l'ordre i el sentit amb facilitat.
- Radial plots: Les disposicions radials o circulars prioritzen un enfocament visual orientat a mostrar les relacions entre nodes, per davant de les de les ubicacions espacials relatives dels propis nodes.

Existeixen altres tècniques de visualització per a grans volum de dades (Yazici, Karabulut, & Aktas, 2018) (Gorodov & Gubarev, 2013) algunes de les quals podrien ser aplicables al llinatge geospacial. No obstant, en aquesta tesi hem destacat aquestes tres en ser les més comunes. La simbolització

aplicable dependrà d'aspectes com el model de representació utilitzat, la granularitat de les dades, l'orientació de les dades, l'ús d'identificadors únics i el volum de les dades a representar.

En el camp geospacial diversos treballs han implementat eines per a visualitzar i simbolitzar el llinatge:

- *Probe-It!* (Del Rio & Da Silva, 2007), és un navegador per representar el llinatge basat en Proof Markup Language (PML). Permet mostrar consultes, resultats, i moure focus de visualització cap a resultats intermedis o finals (Figura 6 imatge A).
- *MetaViz* (Henzen, Mäs, & Bernard, Provenance information in geodata infrastructures, 2013) és una aplicació web que il·lustra específicament dades geospacials conformes a l'ISO19115 de metadades en catàleg CSW 2.0.2 de l'OGC. El llinatge és representat en forma d'arbre. És una aplicació web basada en Java i JavaScript (Figura 6 imatge B).
- *Provis* (Spiekermann, Jolly, Herzig, Burleigh, & Medyckyj-Scott, 2019) eina per simbolitzar el llinatge basada en la llibreria de gràfics "d3" de JavaScript. Genera dos tipus de gràfics, *Standard interactive graph* (Figura 6, imatge C), que permet filtrar el elements a representar i *force directed graph,* que situa els elements finals al centre del gràfic en un estructura radial amb diferents nivells de zoom (Figura 6, imatge D).

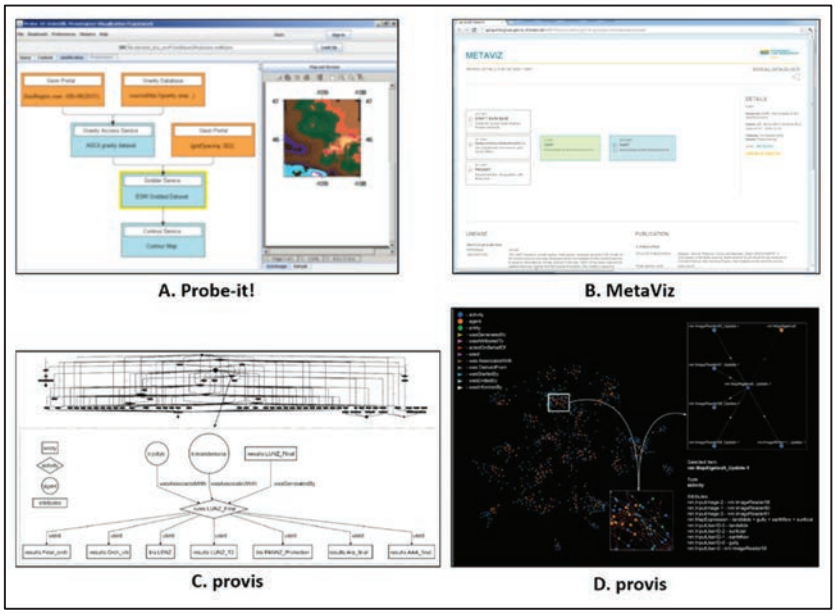

Figura 6: Exemples de visualització del Llinatge (Font: Elaboració pròpia)

1.1.3.5.3 Consultes sobre el llinatge

L'apartat *1.1.3.3 Beneficis del* ha introduït detalladament els beneficis del llinatge sobre les dades i la seva distribució a traves d'un entorn web. Aquests estan directament lligats amb les possibilitats d'explotació del llinatge, a la vegada que les possibilitats d'explotació estan relacionades, en part, amb tècniques de visualització que, com ja s'ha esmentat, poden tenir limitacions degut al volum i la complexitat de la informació a representar. En moltes ocasions, els gràfics de llinatge són d'unes dimensions que, més enllà d'una orientació sobre la quantitat de capes emprades o de processos executats, són difícilment interpretables. Per tant, disposar de mecanismes que ajudin a consultar la

informació, no només facilita la interpretació del procés de producció, sinó que ajuda a maximitzar els beneficis que el llinatge aporta. Les eines de consulta del llinatge permeten seleccionar dades (fonts) i eines de geoprocessament (processos) a partir de formular consultes sobre el "què", el "quan", el "qui" i el "com" dels elements del procés de producció de la IG. Això ajuda als usuaris a tenir un control sobre els elements mostrats en les visualitzacions i així incrementar l'explotació del llinatge en tasques com la descoberta de dades o la localització d'errors i avaluació de la seva propagació.

Autors com Kunde (2008) veuen les consultes com una forma de millorar la visualització. En part és així perquè el resultat de les consultes és típicament representat com un nou gràfic dibuixat "al vol"; tot i que la presentació dels resultats pot tenir també un format de llista. En aquesta tesi doctoral tractem les consultes com una part diferenciada de les tècniques de visualització en la gestió del llinatge geospacial.

Les consultes sobre el llinatge d'un sol arxiu de metadades serveixen per avaluar els requeriments pels quals els models de llinatge varen ser dissenyats. Consultes com les següents, són consultes bàsiques que un model de llinatge ha de ser capaç d'oferir:

- *Quines fonts s'han emprat per generar aquest producte*?
- *Quina ha estat la seqüència de processos que s'ha executat?*
- *Quin ha estat el procés que ha trigat més estona en executar-se?*
- *Qui ha generat aquest conjunt de dades?*

Ara bé, si els arxius de metadades formen part d'una col·lecció disponible en un catàleg, podem anar un pas més enllà i podem formular preguntes complexes i obtenir respostes que impliquin diversos conjunts de dades. Això ens obre la porta a formular consultes de l'estil:

- *En un catàleg de dades, quines dades s'han generat amb una versió específica d'un programa?*
- *En un catàleg de dades, quines dades s'han generat amb una font concreta?*
- *En un catàleg de dades, quines són les diferències entre la cadena de processos de dos productes similars?*
- *En un catàleg de dades, quines dades han estat generades per una persona que es va retirar el 2018?*
- *En un catàleg de dades, quina cartografia de base és la més usada i ha donat resultat a més productes derivats?*

Tot això donarà un valor afegit al llinatge i brindarà als científics i tècnics l'oportunitat d'inspeccionar no només dades geospacials específiques, sinó també els algorismes i metodologies utilitzades. A més, sempre i quan les fonts formin part del mateix catàleg, el sistema pot retornar també el llinatge de les fonts emprades i així successivament, construint un veritable arbre genealògic de dades i processos. En aquest context, el paràmetre "nivell de llinatge" és utilitzat per definir quantes "*generacions*" volem avaluar (He, Yue, Di, Zhang, & Hu, 2015).

El disseny d'un sistema de consultes ha de tenir en compte la interfície de consulta, el llenguatge de consultes, la codificació de la informació i el model emprat. Quan més rica sigui l'estructura del

contingut d'informació de llinatge més possibilitats de consulta tindrem; però més dificultats de codificació crearem als productors de la metainformació. Llenguatges de consulta, com ara el *Structured Query Language* (SQL), proporcionen un conjunt de predicats i tipus de dades que permeten formular consultes. El capítol 2 exemplifica la generació de consultes sobre un fitxer de llinatge codificat amb la notació N3 d'RDF amb el llenguatge de consultes SPARQL (llenguatge de consultes inspirat en SQL però pensat per a consultar RDF).

L'estandardització de les metadades i la separació de conceptes en un model més estructurat facilita la creació de programari que sigui capaç de respondre les consultes complexes sobre un catàleg de metadades. En entorns web, la generació de consultes sobre el llinatge pot abraçar un gran nombre de recursos interrelacionats que pot anar més enllà d'un sol catàleg de metadades. Els capítols 5 i 6 emfatitzen sobre els aspectes relacionats amb la generació de consultes sobre el llinatge i la possibilitat de relacionar dades aparentment llunyanes a partir del llinatge gràcies a l'ús d'identificadors únics i globals en entorns distribuïts.

1.2 Motivació de la tesi

L'interès per les metadades ens remunta als principis de les discussions preparatòries dels esborranys de la directiva INSPIRE i a la iniciativa de dades obertes engegada per l'actualment desaparegut departament de medi ambient. El Grup de desenvolupament del MiraMon (Pons , 2020), que més tard va constituir-se en el Grup de Recerca Mètodes i Aplicacions en Teledetecció i Sistemes d'Informació Geogràfica (GRUMETS), va començar el desenvolupament d'un visor de metadades pels vols de l'any 2000 (GeMM) i ha participat activament en els fòrums de discussió dels estàndards OGC i les normatives ISO d'àmbit geospacials des de l'any 2006. Aquest rol ha permès al grup adquirir un coneixement i una visió holística sobre l'estat dels estàndards, permetent identificar debilitats i/o necessitats. Als inics del 2010, el grup va identificar el llinatge com a part important de les metadades i va incloure'l com objecte d'estudi dins el projecte FP7 GeoViQua (2011-2014). Durant aquest projecte es varen establir les bases aquesta tesi doctoral.

El projecte GeoViQua tenia com a objectiu afegir especificacions de qualitat a la plataforma *Global Earth Observations System of Systems* (GEOSS) (GEO Secretariat, 2005). El GEOSS és un projecte coordinat pel *Group on Earth Obser*vations (GEO, que té com a objectiu produir un sistema de sistemes públic que agrupi dades i informació d'observació de la terra, generades per les diferents administracions públiques i privades a nivell global. Durant el projecte GeoViQua es va constatar el que ja s'intuïa, que l'estat de les metadades disponibles al GEOSS tenia diverses mancances (Díaz, et al., 2012). Per exemple, hi havia relativament pocs registres amb informació de llinatge i bona part dels que en tenien, aquesta es limitava a una simple descripció.

A l'any 2013 Grumets va participar en l'experiment d'interoperabilitat *TestBed 10* de l'OGC i com a resultat d'aquesta participació es va publicar del *Testbed 10 Provenance Engineering Report* (Masó, Closa, Gil, & Proß., 2014). Durant l'experiment d'interoperabilitat ens vàrem adonar que existien alternatives més enllà de la representació del llinatge geospacial amb els models ISO vigents. Vérem

optar per un estàndard de llinatge orientat a un entorn web (com ho és el model PROV) per tal de capturar els diferents nivells de granularitat del llinatge geospacial d'una forma precisa i eficient. També vam constatar les limitacions dels models ISO per representar completament les cadenes de processos.

Així doncs, coneixedors del potencial del llinatge en la millora de la interoperabilitat, l'avaluació de la qualitat i la reproductibilitat de la informació geospacial, les motivacions d'aquesta tesi partien de totes les evidencies adquirides en diversos projectes que ens indicaven que encara existia camí per recórrer en les fases de representació, captura, emmagatzematge, visualització i explotació de la informació de llinatge.

Aquesta tesi vol constatar i situar aquestes problemàtiques, i aportar-hi discussions teòriques i crítiques per arribar a propostes de solucions concretes. La resta de la subsecció descriu mancances i oportunitats d'investigació detectades en el camp del llinatge geospacial.

1.2.1 Estat del llinatge geospacial

Tot i la reconeguda importància i rellevància del llinatge per part de la comunitat científica i de l'existència de diversos estàndards per descriure'l (p.ex. ISO, PROV), la presència d'informació de llinatge a les metadades geospacials és, en general, encara escassa i quan hi és present, aquesta no és completa (Yue, Gong, & Di, 2010). Per tal de constatar el que llegíem a la bibliografia o allò que ens trobàvem a les metadades de forma aïllada, durant la primavera del 2014 es va fer un anàlisi massiu de les metadades presents al GEOSS (Group on Earth Observations, 2020) a fi d'analitzar la quantitat i la qualitat del llinatge. Es va analitzar la completesa de 115678 registres XML de metadades en model ISO que es trobaven a la *clearinghouse (Figura 7)*.

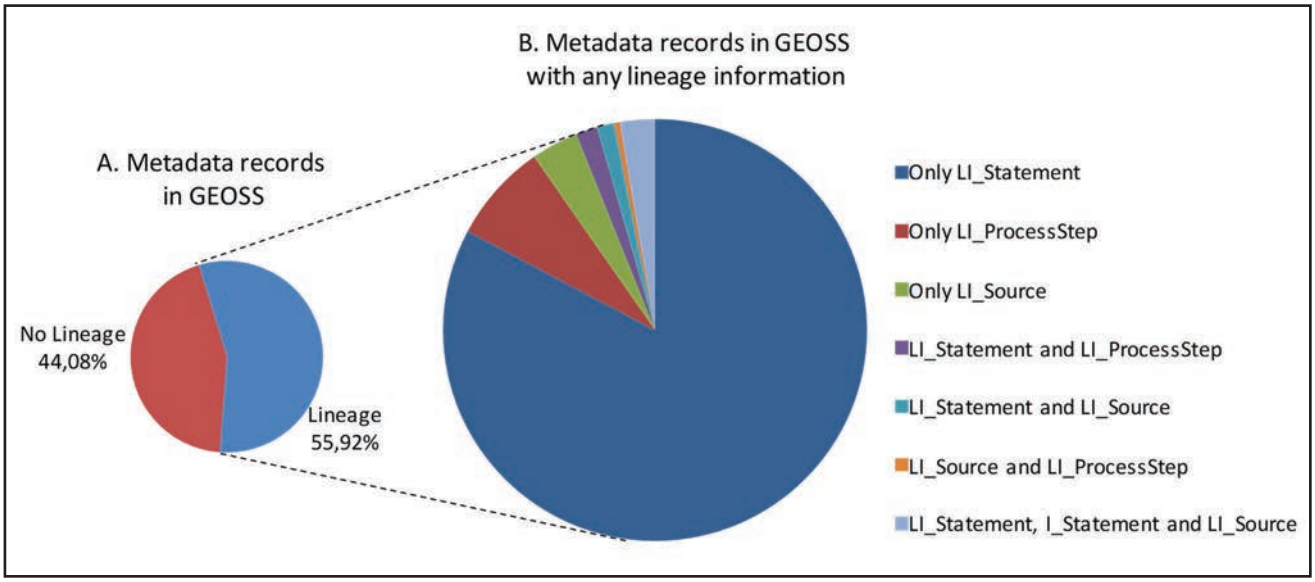

Figura 7: El gràfic circular "A" mostra el percentatge d'elements de metadades presents al GEOSS, a la primavera del 2014, que tenien informació de llinatge. El gràfic circular "B" analitza la completesa del llinatge dels arxius de metadades sobre els elements principals del model ISO. (Font: Elaboració pròpia)

Les principals xifres extretes van ser:

- Més del 44% dels elements no tenien informació de llinatge

- Del més de 55% dels elements que contenien informació de llinatge:
    - El 82.76% només contenien una descripció general (*statement*)
    - El 12.31% només contenien informació relativa als processos
    - El 3.61% només contenien informació relativa a les fonts
    - Només un 2.67% contenien una descripció, informació relativa als processos i a les fonts.

Basant-nos en les afirmacions de Di (2013), es poden citar tres causes d'aquesta situació:

- Els estàndards actuals, no descriuen completament el llinatge, limitant-ne els beneficis.
- Hi ha una manca d'eines per capturar la informació de llinatge automàticament i simultània a l'execució dels processos.
- Existeix una manca d'implementacions de sortides gràfiques eficients que ajudin a millorar la comprensió i mostrin la utilitat del llinatge més enllà de proporcionar la informació.

A més, els productors d'informació geogràfica no posen la necessària atenció en la documentació de les metadades en general i al llinatge en particular, ja que a dia d'avui és encara una tasca tediosa i costosa. Creiem que aquest fet és, en part, degut a una infravaloració dels beneficis del llinatge generada per les tres causes esmentades.

A continuació es detallen les limitacions o problemàtiques detectades en les fases de representació, captura, visualització i explotació del llinatge.

1.2.1.1 Problemàtiques en els models de llinatge

L'evolució històrica dels estàndards de metadades en el camp geospacial ha conduït a que el model ISO (ISO 19115/19115-2) sigui l'emprat de forma majoritària per la comunitat: el perfil de metadades INSPIRE (European Comission, 2007) està basat en l'ISO 19115, l'OGC avala els estàndards ISO/TC211 com a especificacions abstractes de metadades (OGC, 2016) i l'FGDC recomana una transició cap als estàndards ISO (FGDC, 2010).

Di (2013) afirma que la combinació dels models ISO 19115 i ISO 19115-2 serveix com a model genèric de metadades geospacials i, a més, que els models de llinatge definits dins d'ells són aptes per documentar qualsevol informació de llinatge d'àmbit geospacial. En canvi, altres autors com Ivánová (2017) afirmen que el model ISO ofereix una narrativa poc estructurada per descriure els recursos geospacials i impedeix un ús automatitzat del llinatge en un entorn distribuït. En sentit també es pronuncia Kalantari (2020), afirmant que el fet que la majoria d'elements de les metadades s'hagi de documentar com a text lliure és problemàtic, ja que limita la seva descoberta.

La realitat és que el model de llinatge és, a dia d'avui, un aspecte obert en molts sentits (completesa, interoperabilitat, granularitat, etc). Ho demostra el fet que són diverses les iniciatives promogudes els darrers anys per alinear el model ISO amb el model PROV, com per exemple els experiments d'interoperabilitat de l'OGC com l'OWS 9 Testbed (OGC, 2013), OWS 10 Testbed (OGC, 2014), OWS 11 Testbed (OGC, 2015) o l'OWS 12 Testbed (OGC, 2016). També ho demostra el fet que el propi model ISO s'hagi hagut d'actualitzar en diverses ocasions en els darrers anys. La constatació que hi ha altres iniciatives i treballs (moltes d'elles citades al llarg de l'apartat 1.1 d'aquesta tesi) que

assenyalen cap a diverses solucions però totes amb la voluntat d'incrementar les capacitats dels models de llinatge geospacials.

Els capítols 2, 3, 4 i 6 aporten diverses propostes per a la millora dels models de llinatge.

1.2.1.2 Mancances en la captura del llinatge

Dades i metadades encara són sovint produïdes i publicades amb programaris i en moments diferents. En el cas del llinatge, la complexitat de les seves relacions i l'elevat nombre de dades fa que la seva documentació a posteriori de la generació de les pròpies dades sigui una tasca costosa.

Existeix una manca d'eines que capturin el llinatge en el context de SIG generals. Per tant, són necessàries eines que documentin el llinatge automàticament en el context d'un SIG d'acord als principals models de llinatge.

Aquesta tesi doctoral fa propostes en el camp de la documentació automàtica del llinatge. Les propostes, ideades en un marc teòric i conceptual, s'han aplicat al camp pràctic en el context del programa de SIG i Teledetecció MiraMon. Una de les principals característiques del programa MiraMon és que les metadades es gestionen i s'integren acuradament de forma simultània a la creació de les dades. El Gestor de Metadades del MiraMon (GeMM) genera metadades posant especial atenció als aspectes de qualitat (Zabala, Masó, Bastin, & Bigali, 2013) (Zabala, Masó, & Pons, 2016), a la descripció del model de dades i a les relacions amb les bases de dades. La informació de metadades es guarda i es documenta en arxius de format REL (format de text de metadades i relacions del MiraMon, basat en el format INI de Windows) o en format XML descrit per la ISO 19139.

Els capítols 3 i 4 mostren els treballs fets en el marc del MiraMon per la generació d'una eina (*Provenance Engine*) que captura el llinatge conjuntament a l'execució dels geoprocessos. A més, s'ha treballat en la interfície gràfica del GeMM (Figura 16) per possibilitar l'edició de les fonts i dels processos a posteriori de l'execució.

1.2.1.3 Mancances en la visualització del llinatge

Més enllà dels models utilitzats per capturar i emmagatzemar el llinatge, és necessària una visualització eficaç per a una correcta comprensió i avaluació de les dades i dels processos implicats. No obstant, és estrany veure sistemes o eines que proporcionin capacitats de consulta més enllà de la clàssica visualització estàtica de metadades.

Anàlogament a les tasques fetes en l'apartat de captura del llinatge, el GeMM ens ha permès disposar d'una interfície gràfica per proporcionar també una implementació pràctica dels casos pilot.

Així doncs, els capítols 3 i 4 es mostren els treballs fets en el marc del GeMM per representar el *workflow* executat en la generació d'un producte geogràfic determinat.

El capítol 6 també mostra una proposta per visualitzar el llinatge però, a diferencia dels capítols 3 i 4, està pensada per representar el llinatge de diversos conjunts de dades en un context distribuït.

1.2.1.4 Mancances en les consultes sobre el llinatge

Són necessàries eines que complementin les tècniques de visualització i ajudin a endreçar i filtrar la informació de llinatge. Eines que permetin seleccionar dades geospacials o eines de geoprocessament, que a partir de la informació del llinatge aportin valor afegit al llinatge i ,per extensió, a les metadades en general. Tot i que l'estandardització proporciona un vocabulari interoperable que podem utilitzar per a fer consultes, no hem trobat aplicacions desenvolupades en entorns distribuïts que permetin avaluar l'origen de grans volums de dades i analitzar les interrelacions entre elles (principalment fonts i processos).

El capítol 5 es centra en l'aplicació pràctica del model de llinatge proposat per a realitzar consultes sobre la informació de llinatge de dades de Teledetecció. L'objectiu és el de rebre només aquells fragments de dades o processos que ens poden interessar en un moment determinat. A més, el capítol 6 presenta el disseny d'una eina que permet seleccionar dades geospacials o aplicacions de geoprocessament utilitzant informació de llinatge de diversos conjunts de dades.

1.3 Objectius i metodologia de la tesi

L'objectiu general d'aquesta tesi és el de proposar millores que potenciïn una major incorporació del llinatge a les metadades, per tal de millorar la interoperabilitat de dades i processos i il·lustrar el potencial del llinatge com a element de consulta en entorns distribuïts. Aquest objectiu general s'ha dividit en quatre objectius específics que alhora es concreten en diferents sub-objectius:

1. Contribució als estàndards de representació del llinatge.

   1.1 Analitzar diferents possibilitats de representació del llinatge amb diferents estàndards existents: ISO 19115-1, ISO 19115-2, W3C-PROV, etc.

   1.2 Proposar alternatives per representar el llinatge a diferents nivells de granularitat de la informació geogràfica (nivell d'element geospacials, d'atribut i conjunts de dades).

   1.3 Investigar els beneficis d'incorporar altres estàndards (p. ex. WPS) per enriquir la descripció del llinatge.

   1.4 Investigar mecanismes per utilitzar el llenguatge de serialització RDF per descriure el llinatge, i el GeoSPARQL per a poder seleccionar dades a partir de la informació del llinatge.

   1.5 Proposar mecanismes de representació del llinatge que considerin els diferents nivells d'abstracció dels geoprocessos (funcionalitat, algorisme, eina, execució).

2. Contribució a les eines de captura i reproductibilitat del llinatge.

   2.1 Introduir les millores proposades en el sistema de metadades del MiraMon.

   2.2 Analitzar les potencialitats del llinatge en el camp de la reproductibilitat de la informació geogràfica i la generació de processos encadenats en sistemes distribuïts a la xarxa d'Internet.

   2.3 Capturar i documentar les execucions descartades com a part del llinatge.

3. Contribució a les eines de visualització del llinatge.

   3.1 Proposar i implementar un nou entorn de visualització del llinatge en el context del Gestor de Metadades del MiraMon (GeMM).

   3.2 Proposar i implementar un entorn de visualització del llinatge en el context del navegador web de mapes de MiraMon.

4. Contribució a l'augment del valor afegit del llinatge geospacial com a part de les metadades.

   4.1. Generar eines que permetin elaborar consultes complexes del llinatge.

   4.2 Construir un entorn de representació del llinatge unificat de diversos conjunts de dades.

Per tal d'acomplir aquests objectius, la **metodologia** s'ha basat en l'ús i l'estudi dels estàndards i les tecnologies disponibles a fi d'identificar-ne les carències i proposar millores. Més concretament, cadascuna de les propostes realitzades en aquesta tesi doctoral ha seguit el següent procés metodològic:

- Revisió: S'han analitzat els principis bàsics del llinatge, les seves definicions i aplicacions. S'ha recopilat i estudiat la bibliografia científica, així com també dels diversos estàndards de documentació del llinatge disponibles. S'han analitzat alguns dels exemples pràctics i casos d'ús.
- Detecció de necessitats: Una vegada els principis bàsics han estat observats, s'han comparat les aproximacions actuals amb les necessitats derivades dels casos d'ús i s'han fet aparents un conjunt de mancances que calia adreçar, tant en el pla teòric com en l'aplicat.
- Discussió amb la comunitat: S'ha participat en els fòrums de discussió (p. ex. OGC, ISO) amb l'objectiu de compartir les propostes d'evolució dels estàndards o documents de referència per tal de validar, amb la comunitat, la possible integració d'algunes d'aquestes propostes en els nous estàndards internacionals.
- Pilots: S'ha validat el conjunt de propostes a nivell pràctic en un entorn científic rellevant. En aquest punt s'han desenvolupat exemples d'aplicació a nivell de programari que han estat provades en els casos d'ús exposats. Els desenvolupaments s'han inclòs en el programari de SIG i de teledetecció MiraMon pel seu ús en entorns de producció. Això ha permès millorar les contribucions científiques que formen els diferents capítols del compendi.

La Figura 8 relaciona els objectius de la tesi i la seva implementació en els diferents capítols d'aquest document.

1.4 Organització de la tesi

Aquesta introducció (**capítol 1**) té com objectiu situar el lector en el context general necessari que li serveixi de base pels capítols següents. Cal considerar que en ser aquesta una tesi per compendi de publicacions, el nucli central del document el formen les 5 publicacions que es troben incloses del capítol 2 al capítol 6. Naturalment, cadascuna d'aquestes publicacions també conté un resum (*abstract*) i la seva pròpia introducció individualitzada, el que permet aprofundir una mica més en els

aspectes tractats en cada capítol. Cal mencionar que el capítol 6 reprodueix un text en format d'article que, en el moment de redactar aquests línies, encara no havia estat enviat per a la seva consideració com a publicació.

Malgrat que les publicacions presenten i desenvolupen el tema del llinatge des de diferents angles, el conjunt presenta una sola unitat temàtica amb una estructura coherent que primer contribueix a l'estudi i la millora conceptual dels models de llinatge, per anar passant a tractar problemàtiques progressivament més aplicades en les diverses fases de gestió i explotació. Des del principi, les publicacions realitzades durant l'elaboració de la recerca es varen dissenyar per a formar part d'una tesi temàticament unitària i coherent. Tot i això, aquestes també han de ser comprensibles individualment. Per tant, existeix una repetició de certs conceptes bàsics en les diferents publicacions. A més, aquesta introducció recull i agrupa conceptes i vocabulari general per tal de facilitar la comprensió dels 5 articles.

El capítol 2 analitza la representació del llinatge geospacial en un entorn distribuït quan s'aplica a diferents nivells de granularitat: nivells d'atribut, element i conjunt de dades dels models ISO 19115 i PROV. A continuació, es presenta una proposta per aplicar el model PROV a tots els nivells de granularitat d'un entorn geospacial.

El capítol 3 presenta una eina que captura i representa el llinatge basada en l'ús combinat de l'estàndard *Web Processing Service* (WPS) i el model de llinatge ISO 19115. L'eina, desenvolupada en el marc del programa MiraMon, mostra una visualització gràfica de la provenance.

L'article del capítol 4 és en part una continuació del treball discutit en el capítol 3 on es presenta de nou l'eina implementada al GeMM per captar i representar automàticament mitjançant una combinació de WPS i l'ISO 19115. L'eina permet als usuaris editar informació de *provenance* afegint o suprimint passos o fonts a un flux de treball geospacial. Aquest treball inclou també la descripció de com capturar els paràmetres involucrats en les execucions i com això ha quedat reflectit a la darrera versió de l'ISO 19115-2. A més es defensa la necessitat de capturar les execucions descartades com a part del procés de producció científic i es fa una proposta al respecte. Parts d'aquestes propostes són ara incloses en l'ISO19115-2 com a fruit de processos de consens.

El capítol 5 il·lustra l'ús del llinatge en les tasques d'avaluació de dades de Teledetecció. La contribució es focalitza en com consultar la informació de llinatge per tal d'obtenir les fonts, els agents implicats o els geoprocessos, no només d'un conjunt de dades, sinó de diversos conjunts de dades que formen un catàleg. L'objectiu és potenciar els beneficis del llinatge per inferir la qualitat de les dades, rastrejar fonts d'errors i determinar la confiança de la informació geospacial, entre d'altres.

El treball del capítol 6 és, en part, una continuació de les idees presentades al capítol 5. L'article parteix de la idea de poder fer consultes sobre la totalitat del llinatge contingut en un catàleg per tal d'afegir-hi valor a la informació. A l'article s'introdueixen les bases per poder generar un model que documenti els diferents nivells d'abstracció dels processos, el disseny d'una eina fonamental per tal de connectar el llinatge provinent de diferents programaris. A més, es presenta una implementació en un entorn web que permet generar gràfics de llinatge i fer-hi consultes.

El capítol 7 inclou un resum dels resultats extrets dels capítols 2 a 6. Finalment, el capítol 8 inclou les conclusions que serveixen per donar una visió general de les fites assolides.

Per altra banda, s'ha afegit un annex que conté una llista dels acrònims que apareixen en la introducció, el resum de resultats i les conclusions de la tesi.

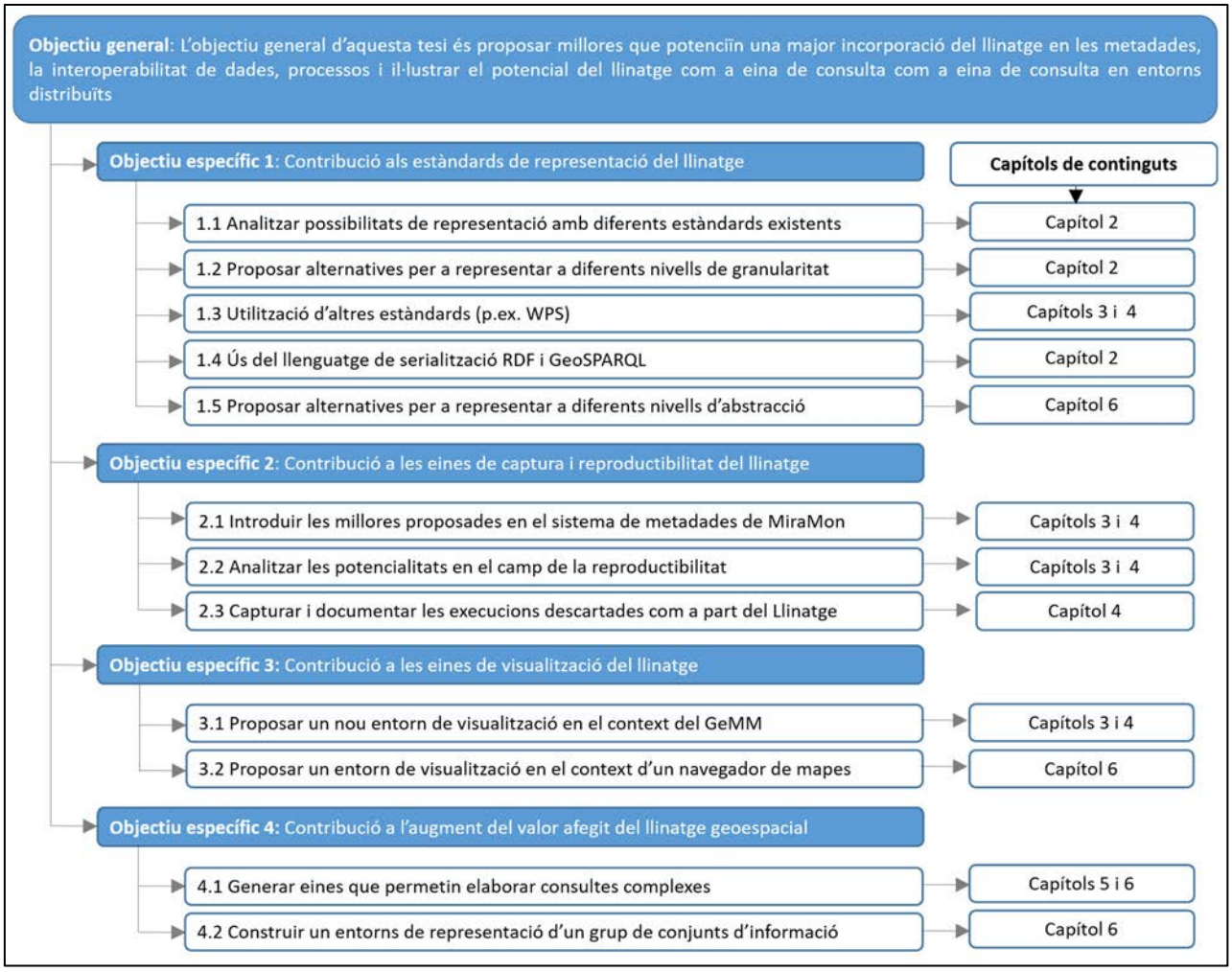

Figura 8: Resum gràfic que relaciona els objectius de la tesi i la seva implementació en els diferents capítols. (Font: Elaboració pròpia)

2. Article 1: W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment

Aquest capítol és una reproducció de: G Closa, J Masó, B Proß, X Pons (2017). *W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment. Computers, Environment and Urban Systems. Volume 64, July 2017, Pages 103-117. DOI: 10.1016/j.compenvurbsys.2017.01.008* https://doi.org/10.1016/j.compenvurbsys.2017.01.008

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.

© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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
0198-9715/© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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

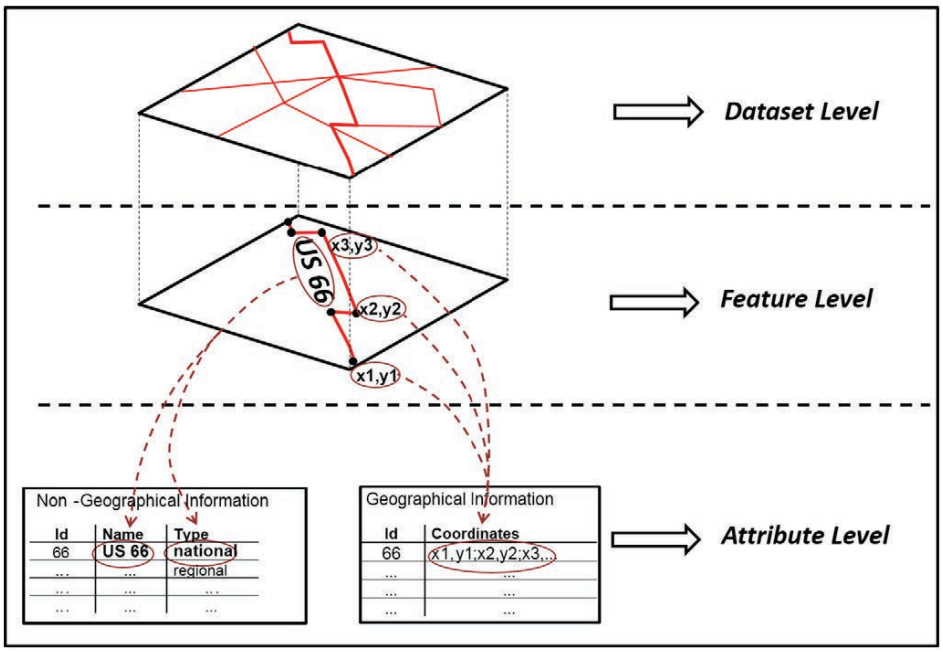

Fig. 1. Conceptual diagram representing the three levels of granularity of geographical information.

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

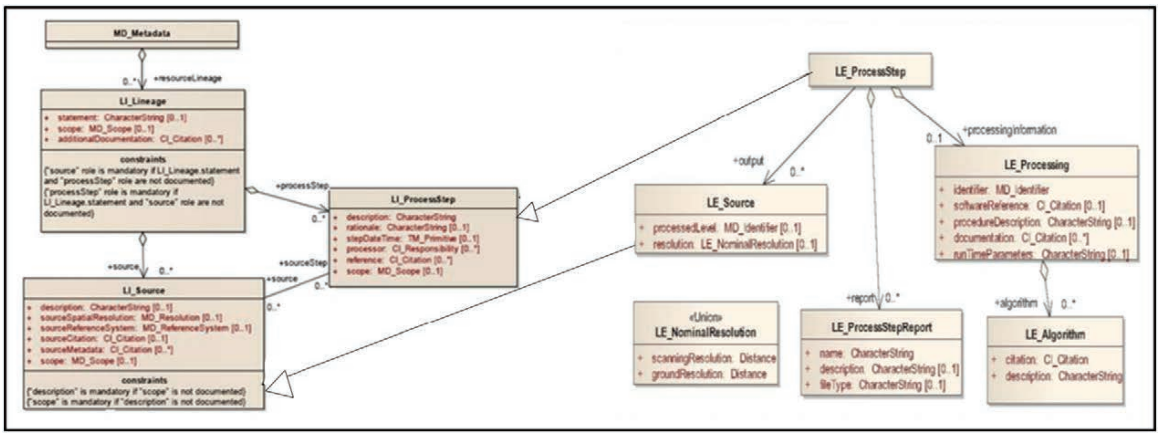

Fig. 2. Lineage UML diagram under ISO 19115-1 and 19115-2, including Source (LI_Source and LE_Source) and Process step (LI_ProcessStep and LE_ProcessStep).

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.

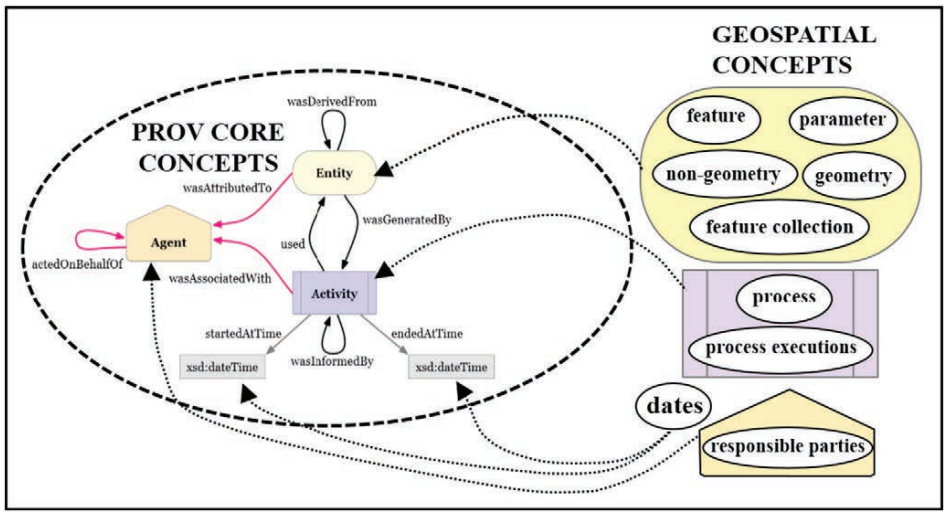

Fig. 3. Matching of geospatial data concepts with the core elements of PROV-DM.

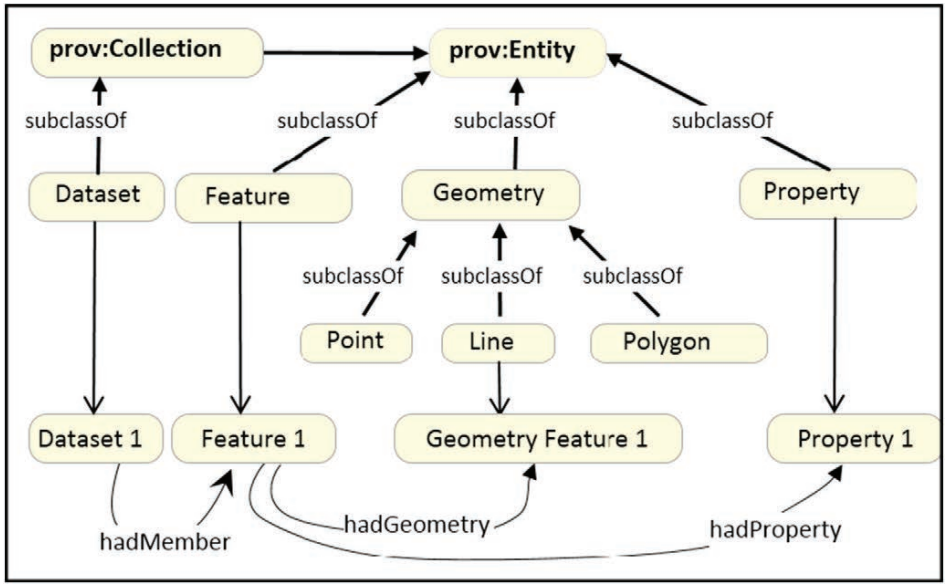

Fig. 4. Diagram representing the dataset, feature and attribute levels in PROV.

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.

[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.

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.

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*

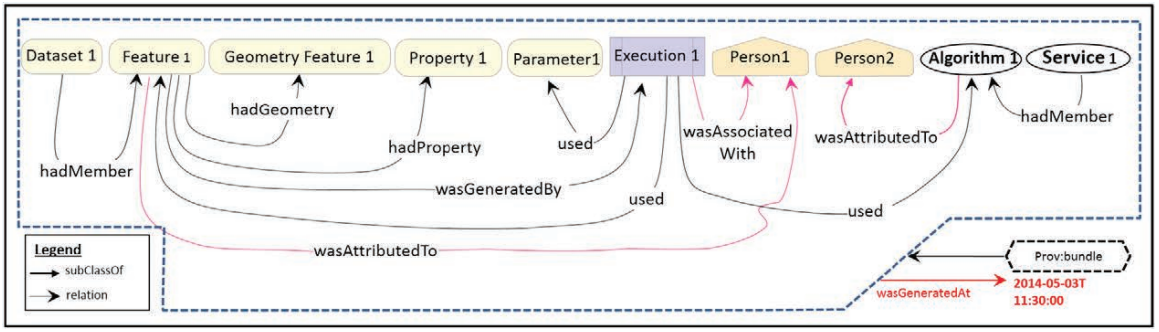

Fig. 5. PROV model for geospatial provenance representing feature, attribute and dataset levels.

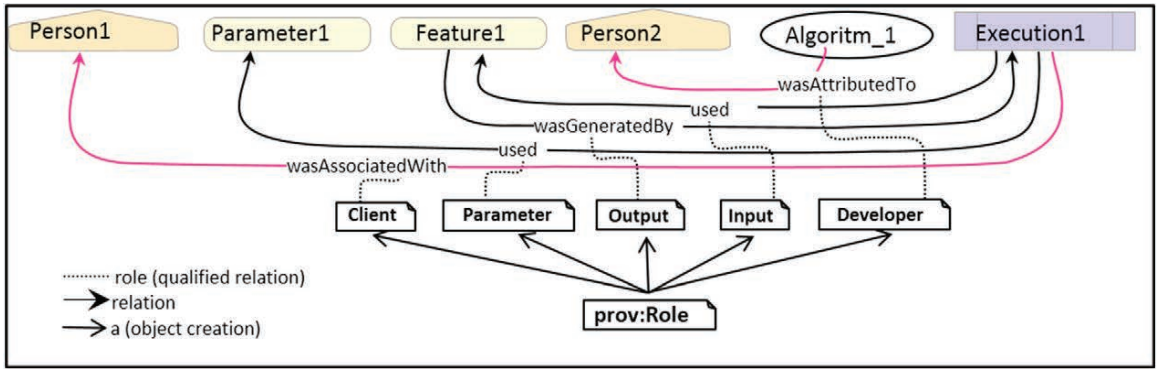

Fig. 6. Diagram representing the PROV roles corresponding to Fig. 5.

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:

- 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.

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  ] .
```

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.

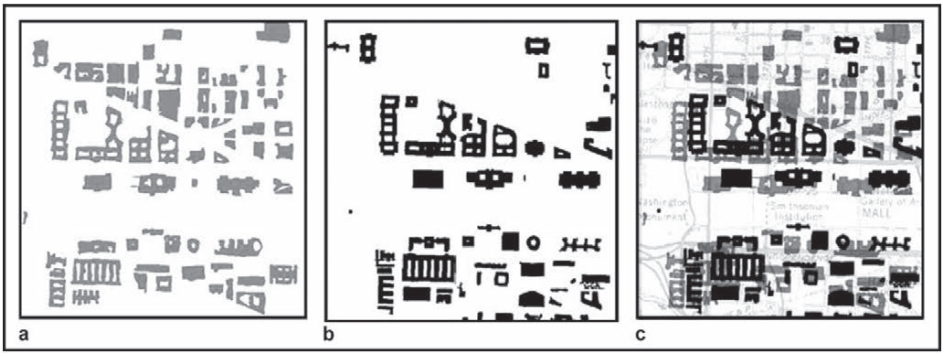

Fig. 7. Graphical example of conflation of two different sources (Seth & Samal, 2007).

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

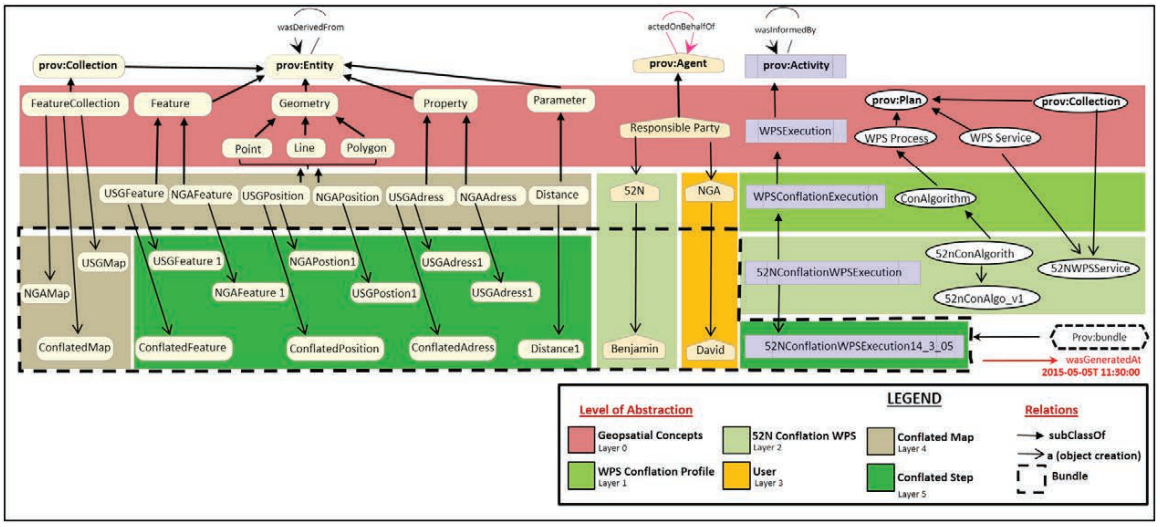

Fig. 8. Diagram showing the main PROV elements involved in the conflation process example. The complete diagram of the PROV conflation example with all the relations between different elements can be found at https://github.com/GuillemClosa/PROV_geo_extension/blob/master/conflation_PROV_model_legend.png,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

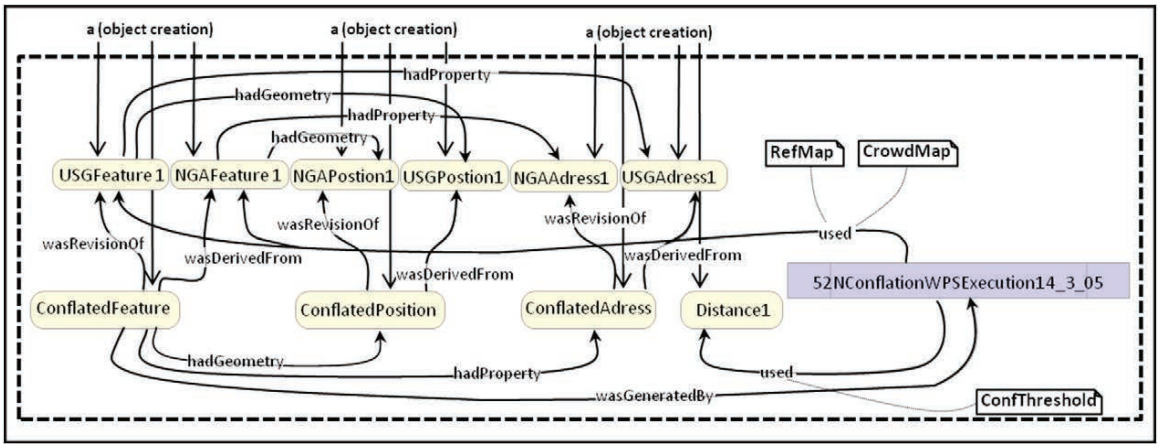

Fig. 9. Diagram showing the conflation use case definition.

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 .
#   Possition 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

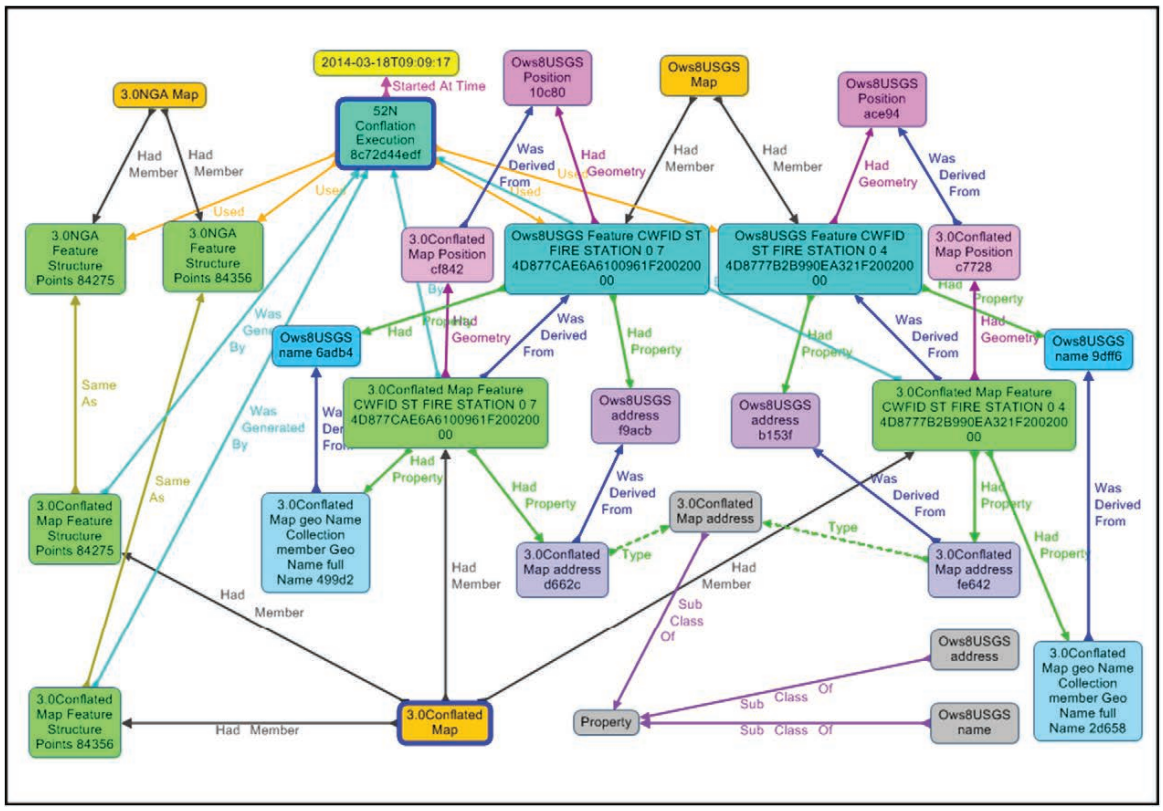

Fig. 10. Provenance graph of a reduced number of entities derived from the conflation example presented above.

- 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/

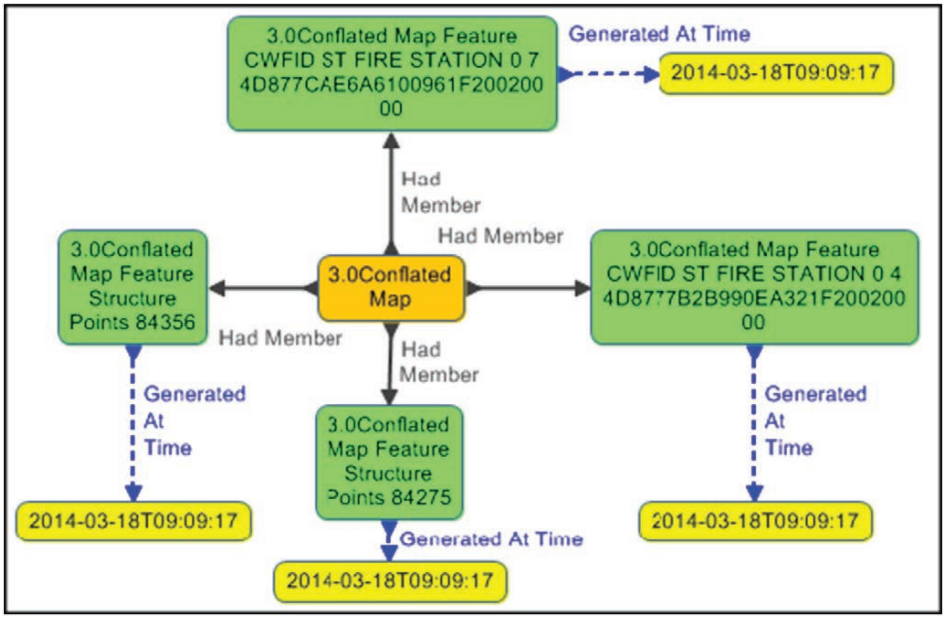

Fig. 11. Graphical representation of the results of Table 13 SPARQL query. Elements that were generated at 2014-03-18T09:09:17 are represented.

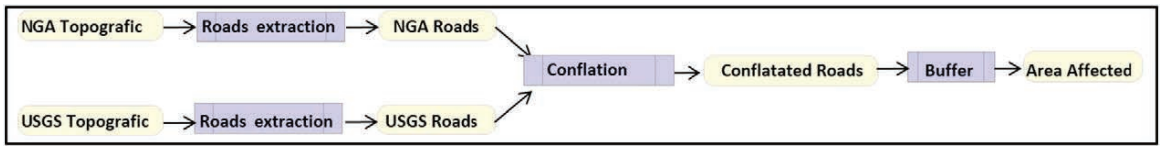

Fig. 12. Example of Geospatial Workflow determining the area affected in conflated roads.

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.

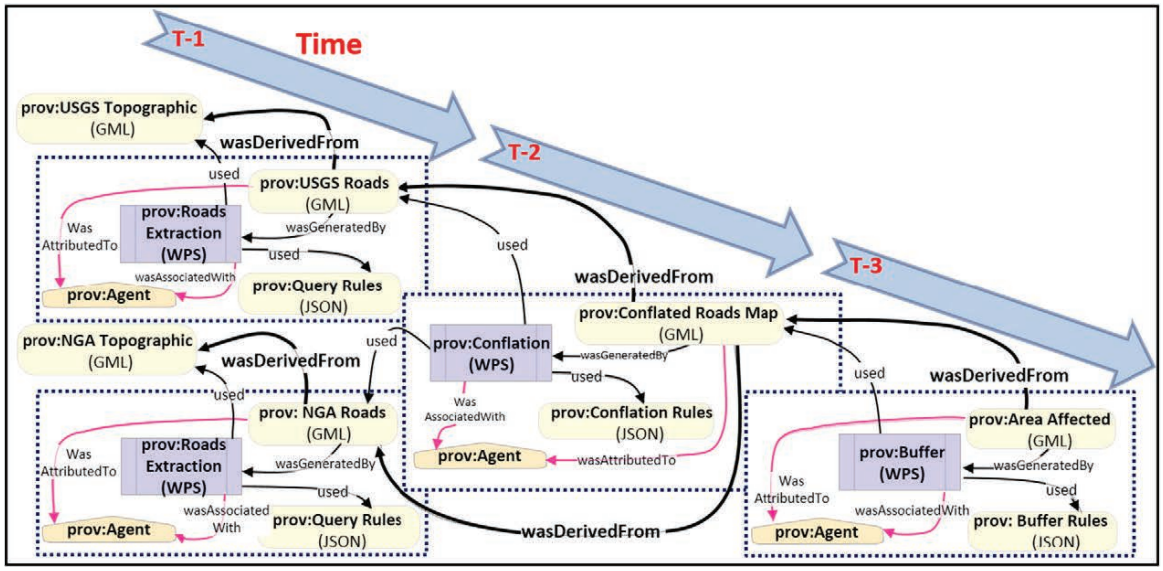

Fig. 13. Diagram illustrating that prov:*wasDerivedFrom* allows the generation of chains using entities as threads.

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.

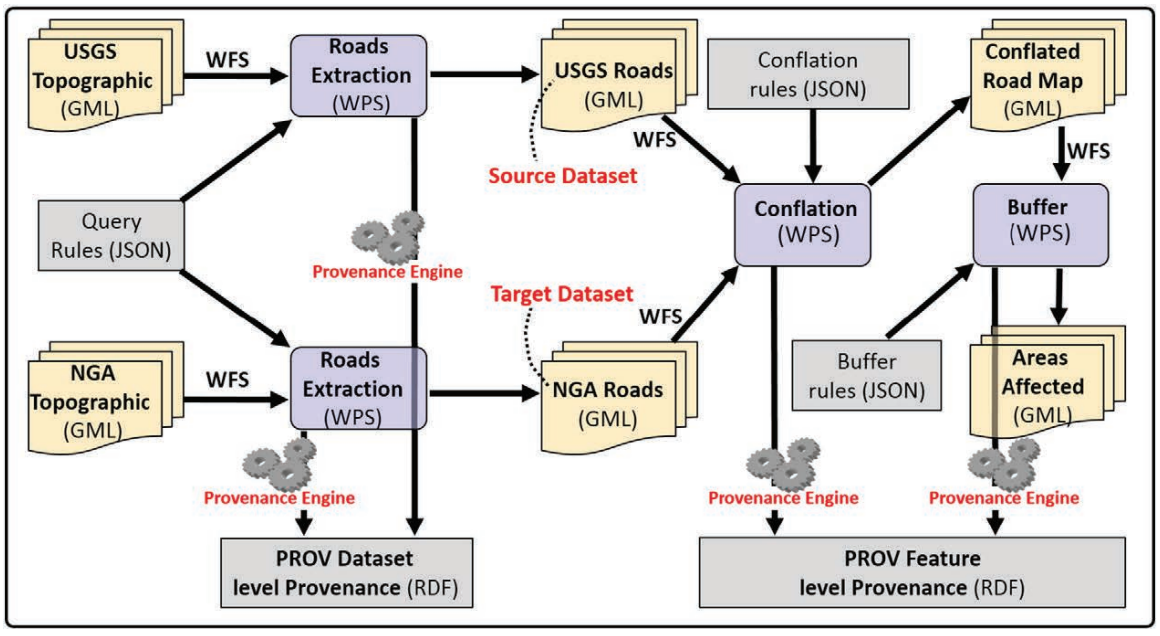

Fig. 14. Diagram of the workflow execution of the conflation process and provenance engine using WPS.

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).

3. Article 2: Web Processing Services to Describe Provenance and Geospatial Modelling

Aquest capítol és una reproducció de: **G Closa, J Masó, N Juliá, L Pesquer, A Zabala. (2017)**
Web Processing Services to Describe Provenance and Geospatial Modelling. *GEOProcessing 2017: The Ninth International*
Conference on Advanced Geographic Information Systems, Applications, and Services. ISBN: 978-1-61208-539-5

GEOProcessing 2017 : The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services

Web Processing Services to Describe Provenance and Geospatial Modelling

Guillem Closa, Joan Masó, Núria Julià, Lluís Pesquer
Grumets Research Group, CREAF
g.closa@creaf.uab.cat, joan.maso@uab.cat,
n.julia@creaf.uab.cat, l.pesquer@creaf.uab.cat
Edifici C, Universitat Autònoma de Barcelona
Bellaterra, Catalonia, Spain

Alaitz Zabala
Grumets Research Group, Dep de Geografia
alaitz.zabala@uab.cat
Edifici B, Universitat Autònoma de Barcelona
Bellaterra, Catalonia, Spain

***Abstract*— There are still some gaps regarding the complete geospatial provenance description. These gaps prevent the use of provenance information for replication and reproducibility task. In addition, the lack of automated tools for capturing the provenance is an obstacle to a widely generation of provenance information. In this sense, we present a tool that captures and represents provenance information based on the combined use of Web Processing Service (WPS) standard and the ISO 19115 lineage model. The tool, developed in the frame of the MiraMon GIS & RS software, shows a graphical visualization of provenance and allow users to edit provenance information by adding or deleting process steps or sources to a geospatial workflow. The automatic capture of lineage information is a step forward in the development of a model constructor tool. It will allow reproducing previous process workflows and applying them to other similar situations.**

Keywords- Provenance; WPS; Modelling

I. INTRODUCTION

Buneman [1] defines provenance information as the description of data origins and the processes by which a dataset is created. This includes also the description of the algorithms used, the processing steps, the inputs and outputs, the computing environment where the process runs, the organization/person responsible for the product, etc [2][3]. In the context of scientific models, data provenance records the workflow processing steps and their inputs/outputs that contribute to the production of the final data products [4].

The scientific community is interested in provenance information because it provides important information to determine the fit for purpose and the reliability of a product. In the geospatial domain data provenance plays a significant role in data quality and usability assessment [5], among others qualities. Moreover, if data provenance information is complete and points to real data and metadata, it can be used as a source for a workflow replication (with other inputs) or for data replication (reproducibility purposes) [2].

As a result of web-technology improvements that have reduced the data volume, computing steps, and resources required by the end-user, geospatial data and geoprocessing tools are available as services [6]. More recently, *Model as a Service* (MaaS) approach has been defined [7][8]. In this paradigm, where the origin of data has a high level of heterogeneity, several authors [9][10] see that provenance information is even more important for inspecting and verifying quality, usability and reliability of data.

Although that the importance of provenance in the geospatial community is documented, its complete description in geospatial metadata is still scarce [11]. Usually, most of the geodata come with some provenance information but in many cases only as a simple textual description, thus having a negative impact on its automated usage [12]. According to *Di* et al. [5], there are two main obstacles that generate this situation: the lack of standards that fully describe provenance information models ensuring reproducibility, and the lack of automated tools for capturing the provenance information.

To exchange and share geospatial data provenance in a distributed information environment, an interoperable model for provenance is needed [13]. The geospatial community has traditionally used the ISO 19115 [14] and 19115-2 standards to encode metadata and provenance [15]. However, there are still some gaps in the ISO models, such as the concrete model initialization, its basic assumptions and parameters values. These deficiencies prevent the complete description of provenance and blocks it use in workflow replication and data reproduction tasks.

Besides representation, provenance applications also need to ensure provenance capture, management and retrieval [16]. In addition, automatic tools that capture and store provenance as a part of metadata information are needed. Most of the work has focused on analysing and capturing provenance information that was created during execution, rather than on metadata generated before execution [17]. However, tools that document provenance before and after the execution are needed too.

In this regard, we have implemented a provenance engine tool that automatically captures and represents provenance information based on the combined use of Web Processing Service (WPS) standard and ISO 19115 lineage model. The tool, developed in the framework of the MiraMon Geographic Information System (GIS) and Remote Sensing (RS) [18], presents a graphical visualization of provenance and allow users to edit provenance information of a geospatial workflow before and after the execution. This automatic acquisition of geospatial provenance represents a step forward in the development of

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-539-5

GEOProcessing 2017 : The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services

a model constructor tool in the context of MiraMon software.

This paper is structured as follows. Section II introduces related work, then sections II and III present the use of WPS to capture provenance and the developed tool. Following section IV introduces the efforts done in generating geospatial models from the captured provenance information. Finally, the conclusions are presented in the last section.

II. RELATED WORK

When selecting a standard for describing provenance in the geospatial domain, some requirements should be taken into account [3]. For Di et al. [5] ISO 19115 and ISO 19115-2 templates are enough to record the complete geospatial lineage. Alternatively, He et al. [14] combines ISO 19115 with W3C PROV [19] to better describe provenance. Others, such as Lopez-Pellicer et al. [20] propose to adapt and extend the W3C PROV model to geospatial community requirements.

Beyond the models used to capture and store provenance, an effective visualization of provenance is also necessary to understand and evaluate data [21]. There are different types of visualization proposals [22], namely:

- Provenance as node-links: data is represented as points and processes as lines. [23][24].
- Provenance as a radial plots: Brings a visual focus to the relationships rather than the relative spatial locations [25].
- Tree diagrams: This technique displays a tree-form diagram starting from the data that is being analysed. Most provenance data have hierarchical properties or attributes [26][27]. Thus we found this type a suitable one to describe provenance.

When generating a geospatial model from concrete executions, a generalization process have to be carried out to standardise and reference the common processing functions. Yue et al. [28] use three levels of encapsulation to reduce the difficulty of sharing and use geo-analysis models in the web. Otherwise, Müller [29] proposes a hierarchical approach to process definitions with different abstraction levels. WPS process profiles [30] are also useful to determine which information from the concrete execution needs to be added to the model to ensure it reusability. An Application Profile is essentially the same as the ProcessDescription document obtained in response to a *DescribeProcess* request [31] (Fig. 1). This approach is in line with our approach of using the WPS standard to capture provenance, consequently we will use *DescribeProcess* documents to generalise models.

III. WPS TO CAPTURE GEOSPATIAL PROVENANCE

A. DescribeProcess documents to capture Provenance

The Web Processing Service (WPS) Interface provides rules for standardizing inputs and outputs (requests and responses) for geospatial processing services [32]. WPS instances are exposed via HTTP-GET, HTTP-POST and SOAP [33] Internet protocols. The potential of geoprocessing applications supported by the WPS allows to apply it in a wide range of fields [34]. Its main properties are: remotely execution, chain of several processes and standardized encodings for data and metadata. WPS is applied in many different fields and sectors that need geoprocessing applications; in particular it is successfully implemented for environmental models [35][36] and in combination to other standards: WPS+OpenMI [37], WPS+WCS [38], WPS+WFS [39]. WPS has three main operations: *getCapabilities, describeProcess and Execute.*

The *describeProcess* is the operation that allow a client to request and receive back detailed information about the processes that can be run on the service instance, including the inputs required, the allowable formats, and the outputs that can be produced [32]. The *describeProcess response* documents use the eXtensible Markup Language (XML). The information described in the WPS describeProcess documents(Fig 1) is the following:

- Process Description: A description of the process and an Identifier.
- Inputs: The input description, the dataType (*ComplexData*, *BoundingBox*, *LiteralData)*, the MIME type, an identifier and the name.
- Outputs: The output description, the dataType (*ComplexData*, *BoundingBox*, *LiteralData)*, the MIME type, an identifier and the name.

Considering that provenance information is the description of processes and sources, *describeProcess* documents could also be used to document provenance information. In addition, *describeProcess* operation can be requested in a local environment. This provides a magnificent opportunity to capture provenance automatically in a GIS local instance. In our case, we have used the *describeProcess* documents to describe all the MiraMon Applications (App), and capture its provenance information when executed. This permits the system to reference sources as a complex data, bounding box or capture the values of the *LiteralData* type.

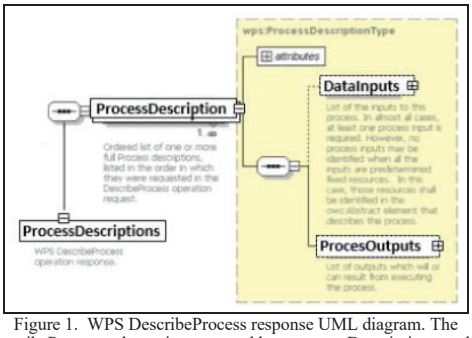

Figure 1. WPS DescribeProcess response UML diagram. The DescribeProcess schema is composed by a processDescription tag that includes a list of DataInputs and ProcessOutputs.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-539-5

WPS is enough well known for the geospatial community, and this allows to jump the interoperability wall. More detail about the use WPS *describeProcess* documents in the context of MiraMon GIS & RS software are provided in Section III.

B. *Combining WPS and ISO to describe provenance*

As mentioned, we have detected some limitations in the ISO 19115 models that prevents the reproducibility of geospatial data using provenance information. In order to overcome this issue we propose the combination of the ISO provenance schemas (LI_Lineage and LE_ProcesStep) with *WPS describeProcess* documents (Fig. 2) presented in the previous section. Combining these two models allows to describe provenance as an ordered list of processes with ISO, including a WPS description of sources and outputs of each process step.

The ISO 19115 and 19115-2 can be described using the *eXtensible Markup Language (*XML). In fact, the ISO 19115-3 provides the XML implementation schema for ISO 19115 and 19115-2 and may be used to describe, validate, and exchange geospatial metadata. The lineage models of ISO (LI_Lineage and LE_ProcessStep) allow to describe the provenance information in three different ways:

- A list of process steps and a list of sources separately.
- A list of all the sources used and then add the description of all the processes as a child.
- A list of all the process steps that use some sources.

Describing provenance with a list of processes that use some sources provides the better way to report a complete record of provenance [12], because it follows the workflow execution. Thus, we use ISO in this way because permits the full description of provenance of a workflow as an ordered succession of different process steps**.** ISO model describes for each intermediate step the sources used and the outputs generated.

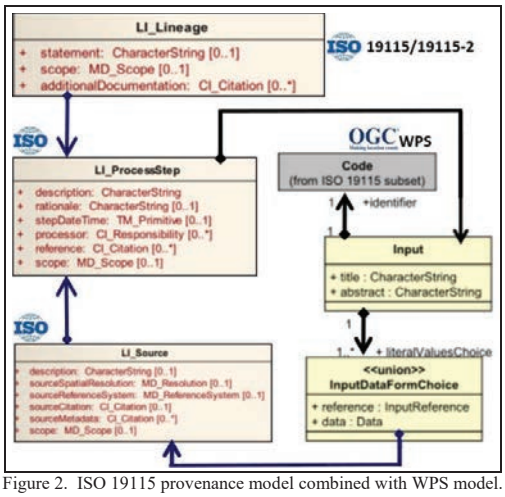

Figure 2. ISO 19115 provenance model combined with WPS model.

However, when describing sources, there is no place to indicate the data type or the value used (for literal data). In this context, to improve the description of the sources and the outputs of each step we introduce the use the *WPS DescribeProcess* to capture, among others characteristics, the data type and the literal data values. The sources and outputs used in each process step of the WPS are connected via identifiers to the ISO schemas.

The combination of ISO provenance schemas with WPS permits the automatic description of the algorithms used, the processing steps, the execution dates, the data type, the units (when necessary) and data values or data location.

The detected gap (no place to define the data type or the value used for literal data) has been introduced as a request for the revision of ISO 19115-2 and we are working with the editors to extend the standard to include this information.

IV. PROVENANCE ENGINE TOOL

A. *Provenance capturation in the context of GEM*M

MiraMon is a Geographic Information System (GIS) and Remote Sensing (RS) software [18]. The main characteristic of MiraMon software is that metadata are carefully managed and completely integrated with the dataset, which allows, at every processing step, to program automatic decisions based on quality information from the previous steps in the process chain [40]. MiraMon incorporates a Metadata Manager (GeMM) to ensure maximum documentation of layers. GeMM allows generating, editing and saving metadata, including the description of the data model and the relations with databases for several hierarchical levels (dataset to several dataset series). The metadata information is stored in REL format documents, which are the native format of MiraMon to document and store metadata information. These files store metadata about identification, extent, related databases, responsible party, technical specification and quality information [41][42]. In addition, as a part of quality information, there is also place for documenting provenance information. REL documents conform to INSPIRE ISO 19115 and FGDC standards and. moreover, metadata can be exported to HTML or XML (ISO 19139) files. Unlike others purely documentary applications, GeMM maintains the dependencies and consistency by checking coherence between metadata and datasets.

MiraMon software has more than 90 applications. In order to capture provenance information automatically, the main task has been the generation, for each App, of a *DescribeProcess response* template that describes the process and its allowed input and output data types. In addition, we use the optional tag *ows:metatada* to define the exact syntax and order of the parameters.

The provenance engine, using the WPS *DescribeProcess* templates, captures provenance of each process carried out and stores it in the metadata files as a part of the quality information of the dataset.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-539-5

The provenance engine is a piece of code that is shared by the visual interface of the GeMM and the MiraMon Apps. It is encoded as a library of C functions that can be linked to each module. Each App uses these functions to read metadata of the source datasets, load it, integrate it, and add the current App process step in the provenance information of the resulting dataset.

The provenance engine writing function can select between two alternatives: a) include all lineage details: complete sequence and description of process steps and previous data sources; or b) write only the last process step and link to the metadata sources. To save space, the generic purpose of each process step and its parameters is not stored. Instead, only identifiers are recorded. The reading function supports the two alternatives described before, being able to read the provenance information by following the links to previous sources recursively if needed. The graphical interface of GeMM requires a more elaborated set of functions to enrich the presentation of provenance information extracted from a *DescribeProcess* response template.

This allows the GeMM to capture, concurrently to an App execution, provenance information using the *DescribeProcess response* templates of each App (Fig. 3).

The system captures the exact parameters and values involved in an execution (that can be numbers, text strings, or bounding box data) and references to datasets or to data services. The system updates metadata information at every intermediate step maintaining the dependencies between the datasets and metadata files during all the workflow execution. The tool keeps track of the dependencies to source datasets and can browse to their metadata too.

B. *Provenance editing and visualization*

In complex environments, scientists rely on visualization tools to help them understand large amounts of data that are generated from experiments [24].

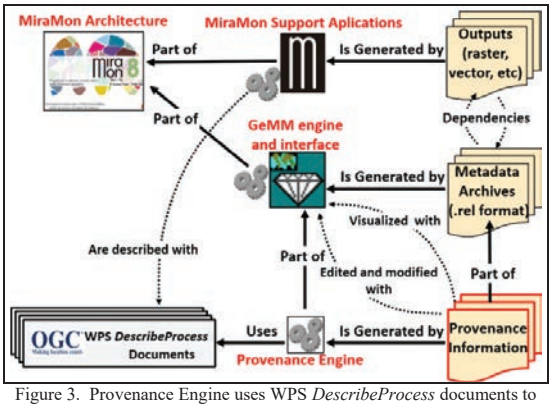

Figure 3. Provenance Engine uses WPS *DescribeProcess* documents to extract provenance information and then the GeMM interface allows users to edit and modify provenance.

According to Steele et al. [43], there are two categories of data visualization *Exploratory*, designed to support researcher who has not certain what is in it; and *Explanatory*, when a researcher is trying to explain the data to someone else. This differentiation remits also to the contraposition of the "*data user needs*" in front of "*data producer needs*", where the user needs more exploratory visualization ways, while producers more explanatory. The graphical interface of our provenance engine fits for both, exploratory and explanatory data visualization approaches.

The provenance engine presented in this paper helps data users to navigate and interpret provenance. The tool represents provenance information as a succession of processes. Each process has an indented list of all parameters used and outputs generated. At the same time, some parameters of the workflow are derived by previous processes (child process), which have, in a deeper level, its own indented list of parameters used, and so on. Thereby, the structure of the provenance schema is progressively increasing its profundity reminding a hierarchical indented form (Fig. 4). From our point of view, this tree-like provenance structure is a suitable way to visualise the provenance information because can easily represent the flow of a specific chain of processes.

The graphical interface of GeMM allows also editing provenance information by adding or deleting child processes or child sources to a geospatial workflow. Moreover, the algorithm description, the processing steps carried out, the execution dates, the responsibility of the product and the processes order can be edited and adapted to each scenario if necessary. This allows data producers to complete the provenance description automatically captured during the process or workflow execution.

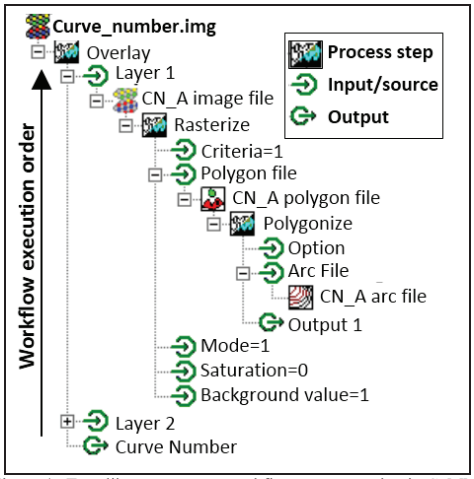

Figure 4. Tree-like provenance workflow representation in GeMM. The example shows processes and sources used in the layer (Curve_Number.img) generation.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-539-5

V. GEOSPATIAL MODELLING

The automatic acquisition of geospatial provenance provides the complete recipe of the geospatial data generated. This supposes an opportunity to develop a model constructor tool in the MiraMon architecture. A model constructor allows the reproduction of previous chains of processes in different scenarios and applying them to similar situations using the provenance captured from previous executions.

Models, as a general representation of a system, are used to understand and simulate a geospatial phenomenon. Thus, a model have to provide enough information to enable the model users to apply it in different scenarios. As pointed in section II there are different approaches in order to generalize specific workflows. In our case, to document models we use the same WPS *DescribeProcess* templates generated to capture provenance. The WPS templates provide the necessary information of each App (process description, process syntax, algorithm location and parameters data type) to allow users to understand each individual process that conforms the model.

The provenance tool (presented in Section III) provides the specific order of the process chain and allows browsing the data inputs of each intermediate step, if necessary.

Finally, all captured information can be automatically exported as a batch file. The generated batch files points to processes and sources used to run workflows. Thus, this allows users to easily reproduce a workflow, or replicate it with different conditions (scope, data, parameters, algorithm options, etc). In addition, the collection of MS-DOS command lines permits automatize executions and ease the use of loops to process large volumes of data.

VI. CONCLUSIONS

Geospatial provenance facilitates geospatial data evaluation for reuse, and brings us closer to the replication of process chains and geospatial modelling. We have detected that there still some gaps regarding to the complete geospatial provenance description, affecting the provenance usefulness. Some gaps detected in the ISO 19115 lineage model has been introduced as a request for the revision of ISO 19115-2.

In this paper, we have shown that the combination of WPS *DescribeProcess* documents with ISO model provides a more complete provenance description. As a proof of concept, we have presented a provenance engine in the framework of MiraMon GIS and Remote Sensing software. The tool allows automatically capturing provenance information and its manually edition if needed. In addition, the automatic description of provenance information is a step forward in the development of a model constructor tool in the context of MiraMon software.

The near future efforts should point to enhance the process chaining and model generation in a distributed environment using provenance information.

ACKNOWLEDGMENT

This work has been conducted within the framework of the Geography PhD program of the Universitat Autònoma de Barcelona, and was supported by the European Commission [grant agreements H2020-641538: ConnectinGEO, H2020-641762: ECOPotential and H2020-689744: Ground Truth 2.0], Spanish Ministry of Economy and Competitiveness [ACAPI (CGL2015-69888-P MINECO/FEDER)] and Catalan Government [SGR2014-1491]).

REFERENCES

[1] P. Buneman, S. Khanna,W and Chiew Tan. Why and Where: A Characterization of Data Provenance. In Database Theory—ICDT. *Springer Berlin Heidelberg.* pp. 316-330, 2001.

[2] L. Di, P. Yue, H. Ramapriyan and R. King. Geoscience Data Provenance: An Overview. Geoscience and Remote Sensing, IEEE Transactions on, 51(11). pp. 5065-5072, 2013.

[3] D. Garijo, Y. Gil and A. Harth. Challenges in Modelling Geosaptial Provenance. Proceedings of the Fifth 24 International Provenance and Annotation Workshop (IPAW), Cologne, Germany, June 9-13, 2014.

[4] A. Chebotko, S. Chang, S. Lu, F. Fotouhi, and Yang, P. Scientific workflow provenance querying with security views. In Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management (Washington, DC, USA), WAIM '08, IEEE Computer Society, pp. 349–356, 2008.

[5] L. Di,Y. Shao and L. Kang.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). pp 5082-5089, 2013.

[6] L. Di and McDonald, K. Next generation data and information systems for earth sciences research, in: 10 Proceedings of the First International Symposium on Digital Earth, vol. I. , Science Press, Beijing, 11 China, pp. 92–101, 1999.

[7] G. Geller and W. Turner. The model web: a concept for ecological forecasting. In IEEE International Geoscience and Remote Sensing Symposium. Pp. 2469-2472, 2007.

[8] S. Nativi, P. Mazzetti, and G. Geller. Environmental model access and interoperability: The GEO Model Web initiative. Environmental Modelling & Software, 39. pp. 214-228, 2013.

[9] S. Bechhofer, D. De Roure, M. Gamble, C. Goble and I. Buchan. Research objects: Towards exchange and reuse of digital knowledge. In, *The Future of the Web for Collaborative Science,* Raleigh, NC, USA. *2010.*

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

[11] P. Díaz, et al. Analysis of Quality 19 Metadata in the GEOSS Clearinghouse. International Journal of Spatial Data Infrastructures Research, 20 7. pp. 352-377, 2012.

[12] P. Yue, J. Gong and L. Di. Augmenting geospatial data provenance through metadata tracking in 28 geospatial

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-539-5

service chaining. Computers & Geosciences, 36(3). pp. 270-281, 2010.

[13] L. He, P. Yue, L. Di, M. Zhang and L. Hu. Adding Geospatial Data Provenance into SDI—A Service-Oriented Approach. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 8(2).pp. 926-936, 2015.

[14] ISO 19115-1:2014 (2014). "Geographic Information-Metadata- Part 1: Fundamentals".

[15] J. Masó, G. Closa, Y. Gil and B. Prob. OGC® Testbed 10 Provenance Engineering Report OGC Public Engineering Report (pp. 1-87): Open Geospatial Consortium. 2013.

[16] S. Miles, et al. The requirements of using provenance in e-science experiments. Journal of Grid Computing, 5 (1).pp. 1–25, 2007.

[17] J. Kim, Y. Gil and V. Ratnakar. Semantic metadata generation for large scientific workflows, Proceedings of the 5th International Semantic Web Conference, Athens, Georgia, USA, Lecture Notes in Computer Science (LNCS) 4273. Springer, Berlin, Germany, pp. 357–370, 2006.

[18] X. Pons. (2004). MiraMon. Geographical information system and remote sensing software. Centre de Recerca Ecològica i Aplicacions Forestals (CREAF).

[19] P. Groth, and L. Moreau. PROV-Overview: An Overview of the PROV Family of Documents. Working group note, W3C. 2013.

[20] F. Lopez-Pellicer and J. Barrera. D16.1 Call 2: Linked Map VGI Provenance Schema. In Linked Map subproject of Planet Data. Seventh Framewok Programe, 2014.

[21] M. Kunde, H. Bergmeyer and A. Schreiber. Provenance and annotation of data and processes. In J. Freire, D. Koop, and L. Moreau, editors, IPAW '08, chapter Requirements for a Provenance Visualization Component. p.p. 241–252, 2008

[22] M.Borkin et al. Evaluation of filesystem provenance visualization tools. IEEE Transactions on Visualization and Computer Graphics, 19(12). pp. 2476-2485, 2013.

[23] N. Del Rio and P. Da Silva. Probe-it! visualization support for provenance. In International Symposium on Visual Computing. Springer Berlin Heidelberg .pp. 732-741,2007.

[24] G. Salton, J. Allan, C. Buckley, and A. Singhai. Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 264. pp. 1421–1426, 1994.

[25] C. Scheidegger, et al. Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, *20*(5). Pp. 473-483, 2008.

[26] G.Closa and J. Masó. A provenance visualization tool for global earth observation system of systems. In EGU General Assembly Conference Abstracts (Vol. 15, p. 8266). April, 2013.

[27] L. Gou and X. Zhang. Treenetviz: Revealing patterns of networks over tree structures. IEEE TVCG, 17(12), December 2011.

[28] S. Yue, M. Chen, Y. Wen and G. Lu. Service-oriented model-encapsulation strategy for sharing and integrating heterogeneous geo-analysis models in an open web environment. ISPRS Journal of Photogrammetry and Remote Sensing, 114. pp. 258-273, 2016.

[29] M. Müller. Hierarchical process profiles for interoperable geoprocessing functions. In Proceedings of the 16th AGILE Conference on Geographic Information Science, Leuven, Belgium. 2013.

[30] OGC® WPS 2.0 Interface Standard. OGC 10-59r2, 2010 14-065

[31] WPS concepts (November, 2016) Retrieved from:http://geoprocessing.info/wpsdoc/Concepts

[32] OGC® WPS 2.0 Interface Standard. OGC 10-59r2, 2010 14-065

[33] D. Box et al. Winer, Simple Object Access Protocol (SOAP) 1.1, W3C Note. Retrived: November, 2016. http://www.w3.org/TR/SOAP

[34] C. Michaelis and D. Ames. Evaluation and implementation of the OGC web processing service for use in client-side GIS. Geoinformatica, 13(1). pp. 109-120, 2009.

[35] A. Castronova, J. Goodall and M. Elag. Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard Environmental Modelling and Software Volume 41, pp. 72-83, 2013.

[36] L. Granell, S. Díaz, N. Schade,J. Ostländer and J. Huerta. Enhancing Integrated Environmental Modelling by Designing Resource-Oriented Interfaces. Environmental Modelling & Software, 39.pp. 229-246, 2013.

[37] J. Goodall, B. Robinson and A. Castronova. Modeling water resource systems using a service-oriented computing paradigm. Environmental Modelling & Software, 26(5). pp. 573-582, 2011.

[38] G. Yu, P. Zhao, L. Di, A. Chen, M. Deng and Y. Bai. BPELPower-A BPEL execution engine for geospatial web services Computers and Geosciences Volume 47. pp. 87-101, 2012.

[39] X. Meng, Y. Xie and F. Bian. Distributed geospatial analysis through web processing service: A case study of earthquake disaster assessment Journal of Software Volume 5, Issue 6, pp. 671-67, 2010.

[40] L. Pesquer, J. Masó, G. Moré, X. Pons, J. Peces and E. Doménech. Servicio interoperable (WPS) de procesado de imágenes Landsat. Teledetección, 37. pp. 51-56, 2012.

[41] A. Zabala, J. Masó, L. Bastin and L. Bigali. Increasing dataset quality metadata presence: Quality focused metadata editor and catalogue queriables. . Inspire Conference. Florence, Italy, June 23-27, 2013.

[42] A. Zabala, J. Masó and X. Pons. Quality and user feedback metadata: theoretical aspects and a practical implementation in the MiraMon metadata editor. Inspire Conference. Barcelona, Spain, Setember 26-30, 2016.

[43] J. Steele and N. Iliinsky. Beautiful visualization: looking at data through the eyes of experts. "O'Reilly Media, Inc.", 2010.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-539-5

View publication stats

4. Article 3: A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation

Aquest capítol és una reproducció de: G Closa, J Masó, A Zabala, L Pesquer, X Pons. (2019). *A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation*. *Transactions in GIS, Volume23, Issue5 October 2019 Pages 1102-1124*. DOI: 10.1111/tgis.12555. https://doi.org/10.1111/tgis.12555

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.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. *Transactions in GIS* published by John Wiley & Sons Ltd

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,

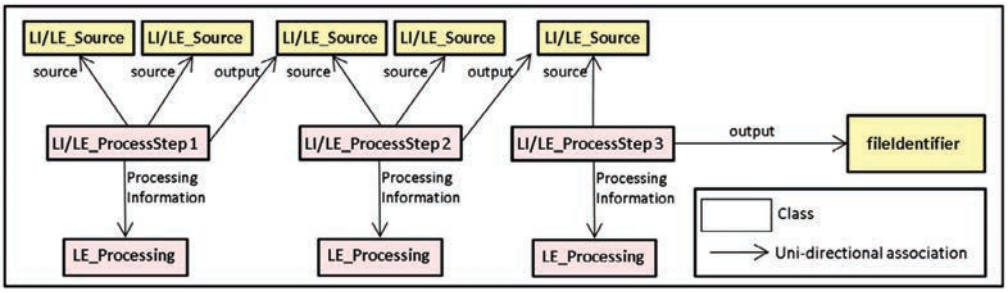

FIGURE 1 ISO 19115 and 19115-2 lineage model permits us to depict the complete sequence of the workflow

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"*).

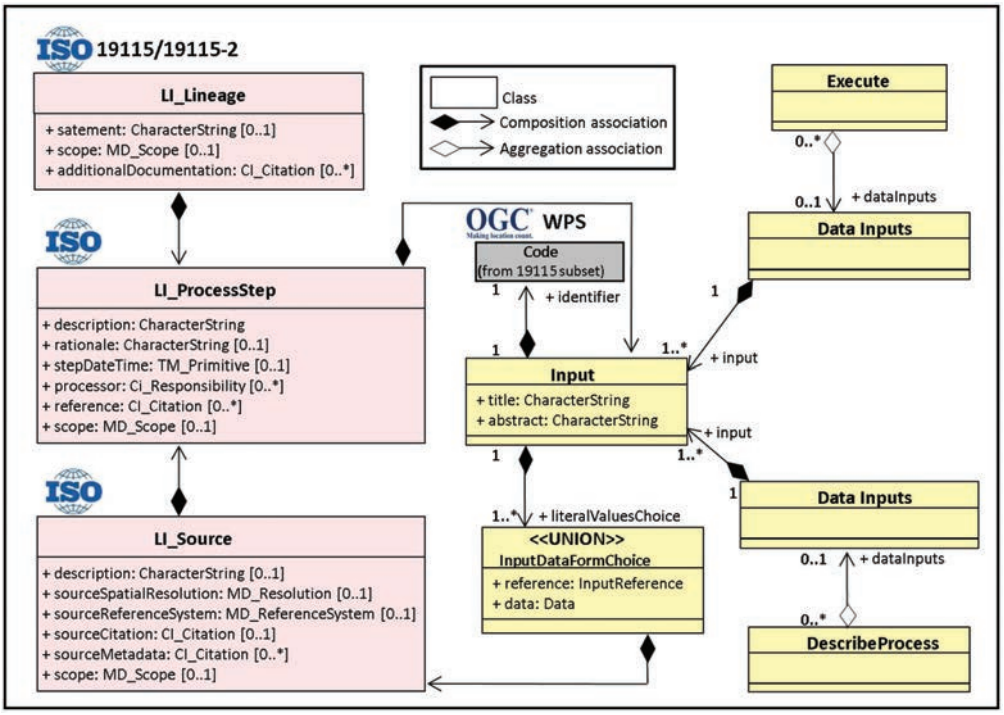

FIGURE 2 ISO 19115 *LI_Lineage* describes provenance as a sequence of *LI_ProcessStep* that uses *LI_Source*. The information contained in *LI_Source* is extended with the use of some WPS elements (UML class diagram)

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).

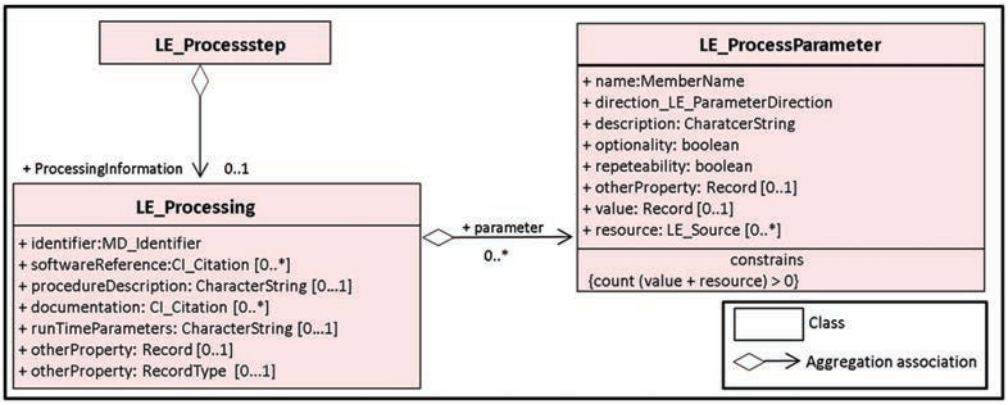

FIGURE 3 *LE_Processing* has an aggregation type: *LE_ProcessParameter*

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

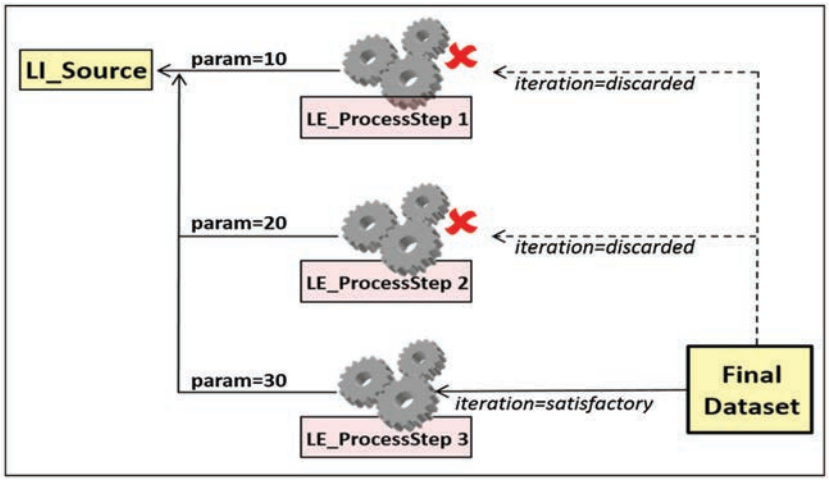

FIGURE 4 The attribute of *LE_ProcessStep*, *OtherProperty*, is used in a Boolean way to document whether the execution is satisfactory or not. In this figure the final output is generated after three iterations of the same process (different executions) with a different parameter value: the first two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

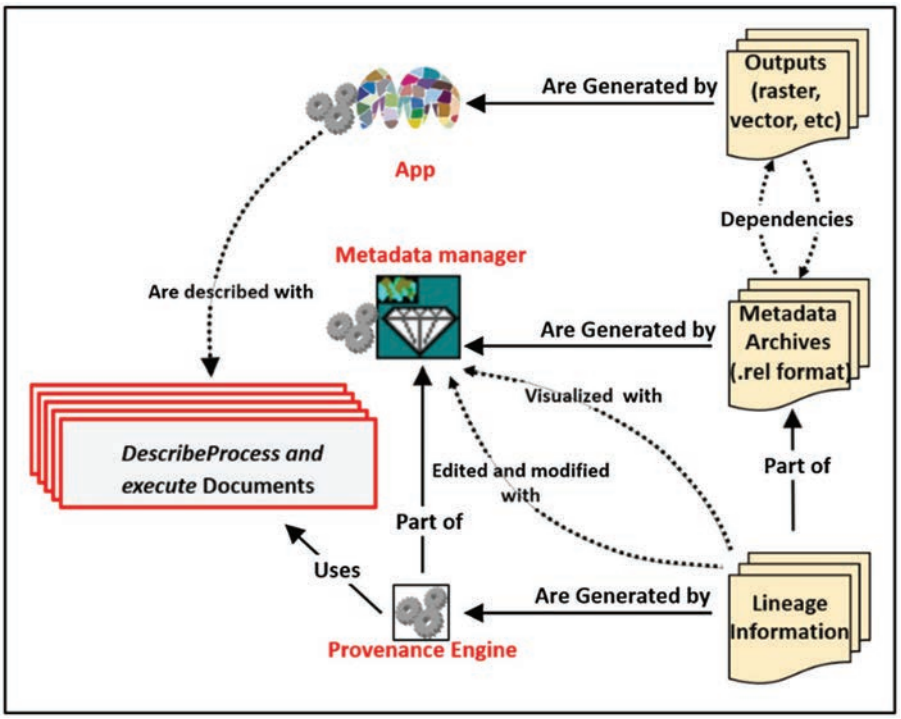

FIGURE 5 The PE uses WPS *DescribeProcess* documents to extract provenance information and then the GeMM (metadata manager) interface allows users to edit and modify the provenance of the geospatial data generated by MiraMon apps

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

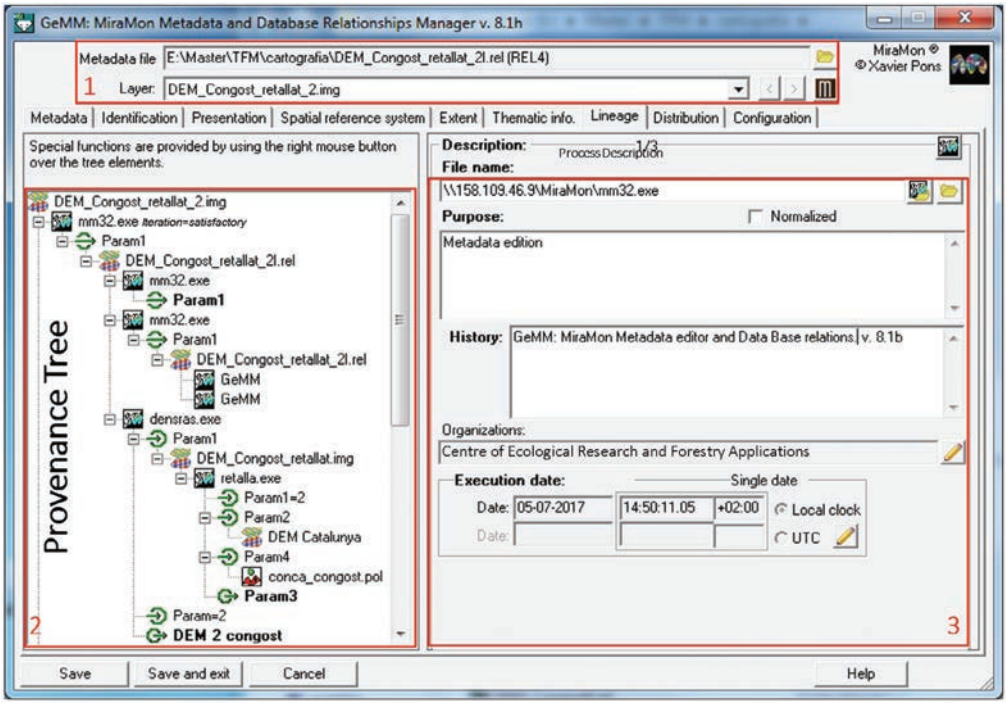

FIGURE 6 The GeMM graphical interface: (1) the path of the metadata file and the name of the geospatial file; (2) the exploratory mode with a tree including all processes and sources used in the history of the creation of the dataset; (3) the explanatory mode to view or edit the attributes of each source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

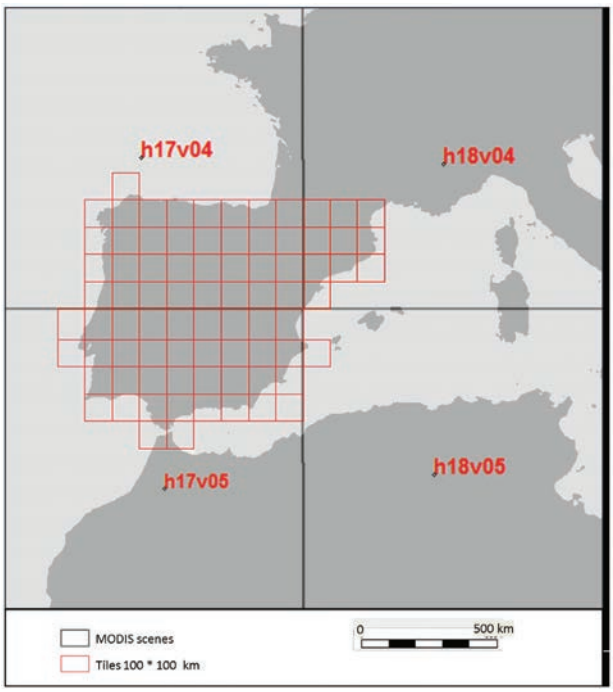

FIGURE 7 Main MODIS tiles and 100 × 100 km sub-tiles over the Iberian Peninsula

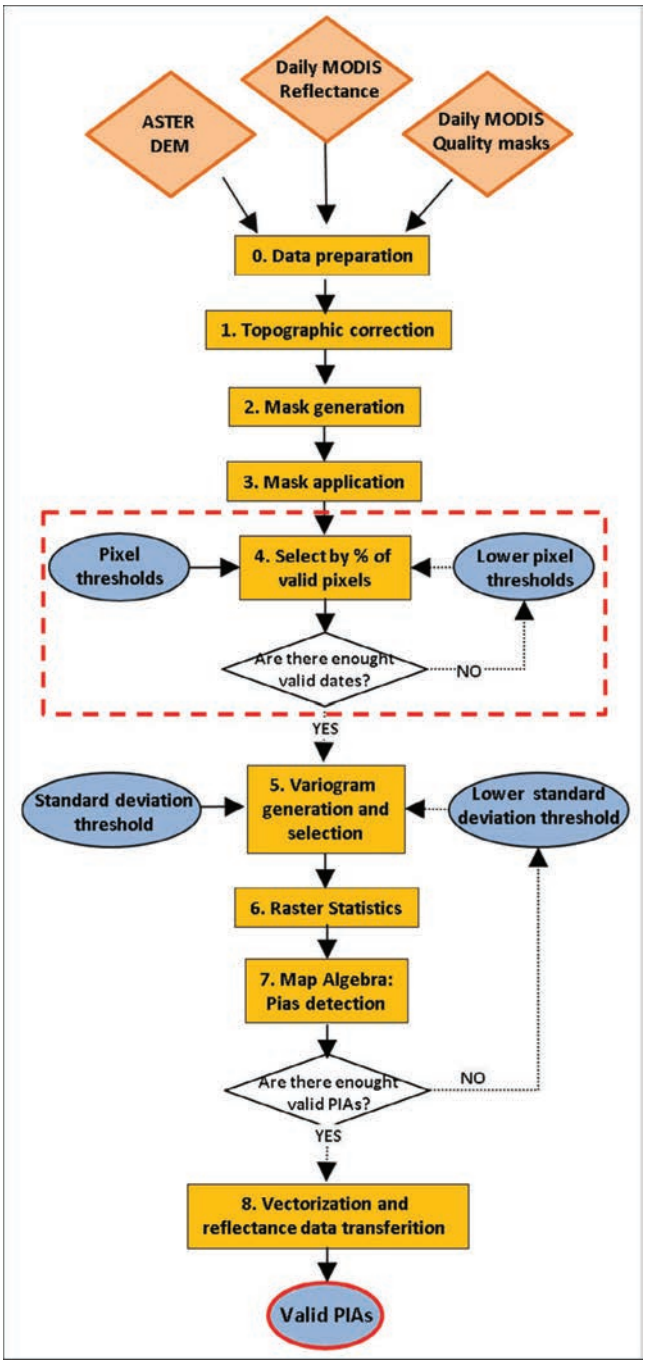

FIGURE 8 Complete workflow of PIAS generation. The dashed red lines mark the part of the workflow used to exemplify how provenance is captured in the Section "Replication with different pixel thresholds"

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

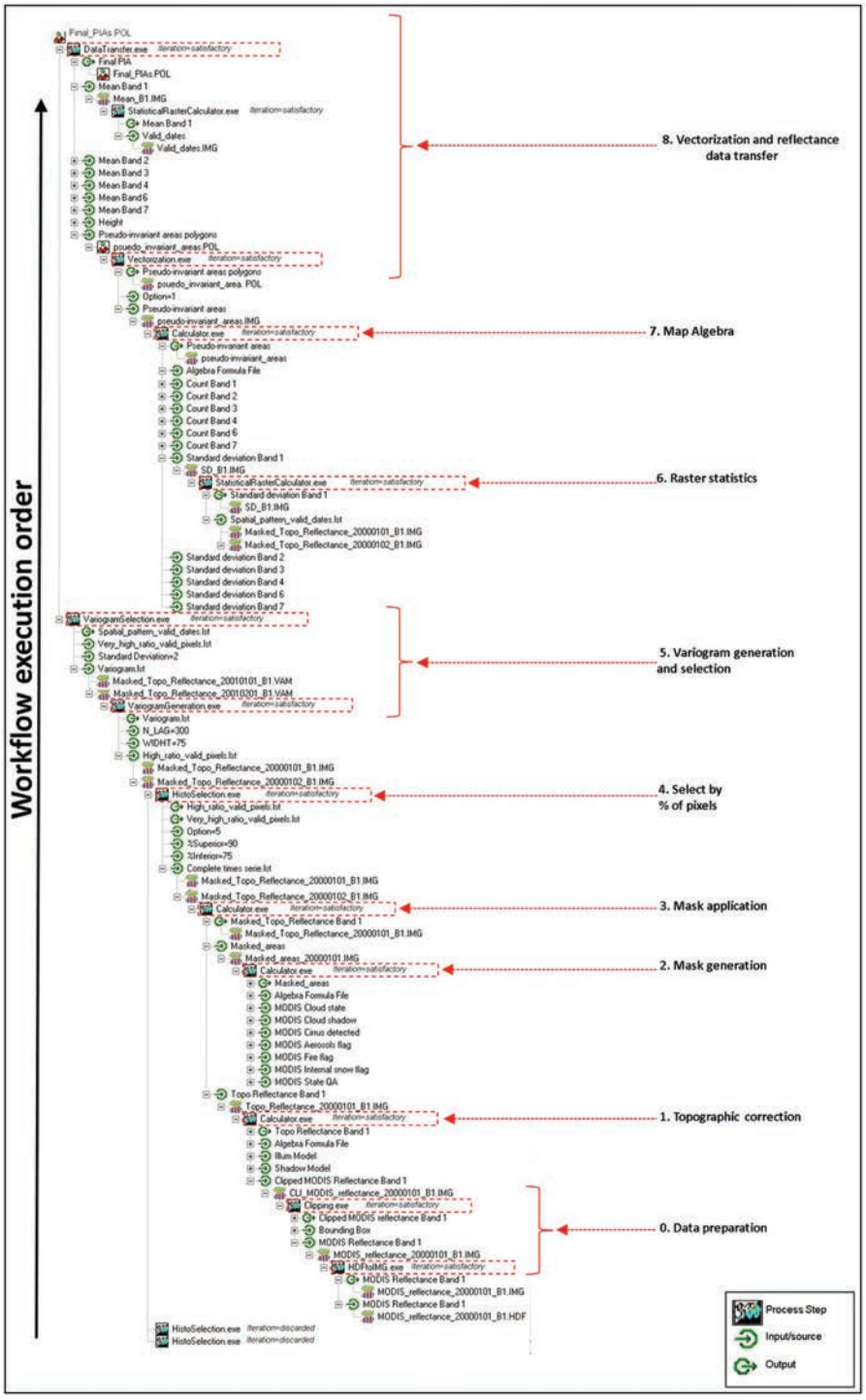

FIGURE 9 Example of provenance tree including all processes and sources used in PIA file generation. To show the provenance tree more clearly, only one branch is shown completely, some parts of the workflow are minimized, and step number and name labels have been added

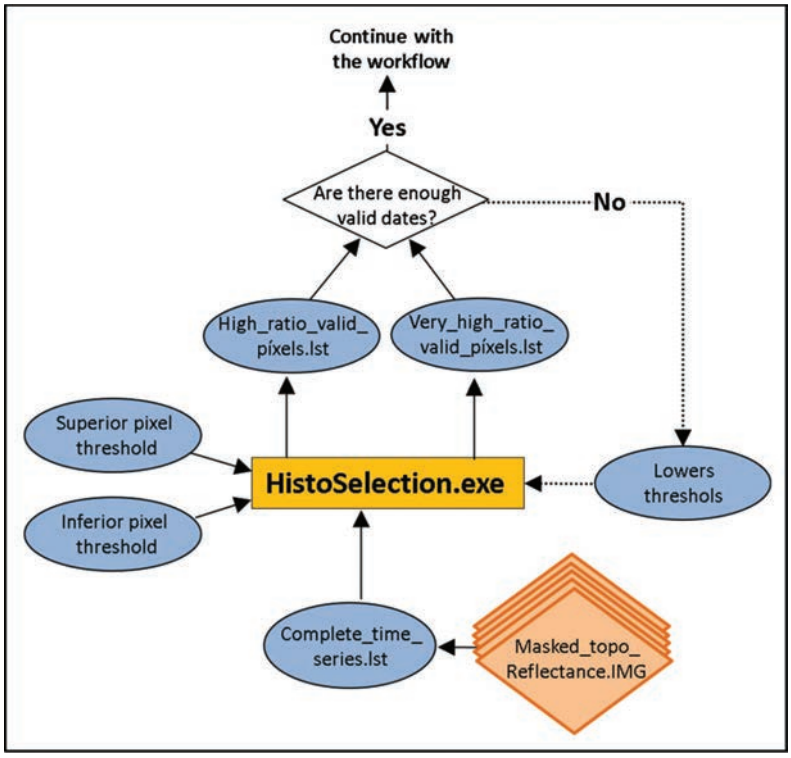

FIGURE 10 Detailed graph of step 4. The execution is repeated until the number of valid dates is sufficient

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
```

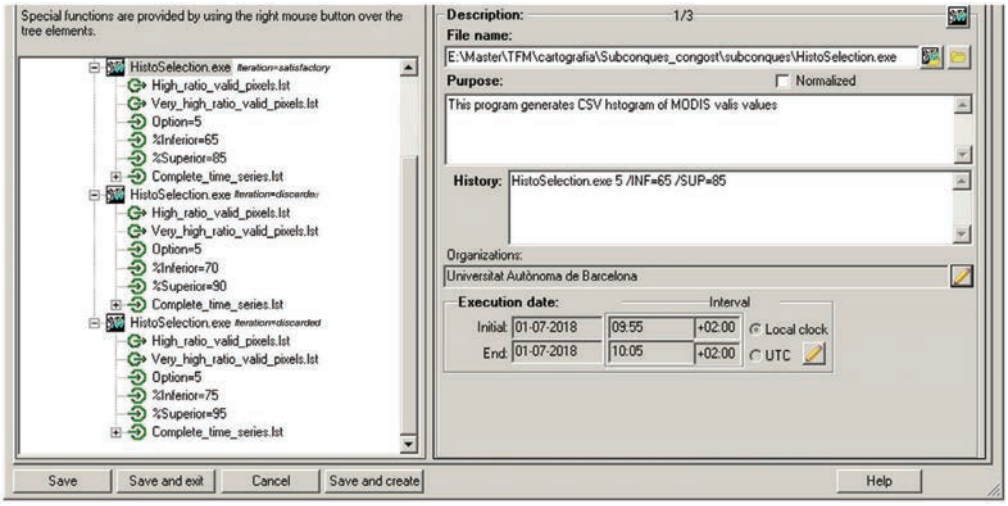

FIGURE 11 Step 4 tree-like provenance workflow representation in the GeMM. The graph records the satisfactory and discarded executions with the parameters used in each loop. The right-hand panel shows the properties of each parameter and the processes executed, such as the file name, execution date,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

5. Article 4: Auditing Remote Sensing Data Using Geospatial Provenance

Aquest capítol és una reproducció de: **G Closa, J Masó, L Pesquer, X Pons (2019).** Auditing Remote Sensing Data Using Geospatial Provenance. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. https://doi.org/10.1109/IGARSS.2019.8898350

AUDITING REMOTE SENSING DATA USING GEOSPATIAL PROVENANCE

Guillem Closa[1], Joan Masó[2], Lluís Pesquer[2], Xavier Pons[1]
[1]Grumets Research Group, Dep de Geografia, Edifici B, Universitat Autònoma de Barcelona
08193 Bellaterra, Catalonia, Spain
guillem.closa@uab.cat; xavier.pons@uab.cat
[2]Grumets Research Group, CREAF, Edifici C, Universitat Autònoma de Barcelona
08193 Bellaterra, Catalonia, Spain
joan.maso@uab.cat; l.pesquer@creaf.uab.cat

ABSTRACT

Auditing Remote Sensing (RS) data can be a tough task due to the huge heterogeneity of RS data and geoprocessing tools available. In this sense, provenance information, part of the metadata that describes "what", "when", "who", "how" and "where" have generated a particular geospatial data, can be very useful. This contribution focuses on how to query provenance information in order to receive back data sources and geoprocessing tools that have participated in the generation of datasets in a catalog. This way, queries over data provenance can be used to infer data quality, trace errors sources, trust and authority of a geospatial information.

Keywords- Provenance; Data queries; Metadata

1. INTRODUCTION

Geospatial provenance, information about the origins of geospatial data products, has been pointed as a fundamental issue in distributing spatial information over the web [1]. Provenance includes information about sources, process executions, algorithms used, algorithms definition, moment of execution and responsible parties [2]. This way, tracking back the production process, scientists can assess the usability, reliability, data quality, errors location and propagation [3] [4] among other qualities and uncertainties.

In Remote Sensing (RS) applications, data can be offered at different processing levels. Moreover, the origin of the algorithms and data used can have a high level of heterogeneity and be crucial for the proper understanding and usage of this data. Provenance can be useful in this field as presented in [5]. For instance, Copernicus (the European Union's Earth Observation Programme) open access hub offers Sentinel-2 data at Top-Of-Atmosphere [TOA] or at Bottom-Of-Atmosphere [BOA] reflectance. In addition, Copernicus Services offer high-level products (i.e., biophysical variables, land cover maps, etc.) that were 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.) might need other alternative processing methods, which favor particular conditions (e.g., optimized for mountain areas) or allow more coherent radiometry to be obtained, even if they might require more precise calibration (e.g., determining pseudo-invariant areas), and which can provide a slightly different results. In this context, the question is how scientists and technicians can easily choose the more convenient algorithms or sources for their specific purposes based on trust, data quality, authority, property type, etc.

Some GIS and RS tools provide users with some functionalities to store provenance information. In addition, spatial agencies have sometimes been particularly careful on saving provenance information

as part of their data preservation strategies. Despite the enormous potential of the stored provenance information, it is rare to see a system that provides query capabilities beyond basic metadata visualization. In this regard, this contribution presents the initial efforts done in order to generate a provenance queries information system. The possibility to formulate queries over datasets or clearinghouses (e.g. given a clearinghouse, which data have been generated with a specific version of an algorithm?), will give an added value to data provenance, and will provide scientists and technicians the opportunity to inspect not only specific geospatial data, but also the algorithms and methodologies used.

2. BACKGROUND

One could assume that a simple statement could be enough for documenting provenance, but to be able to query provenance, a well-known internal structure is needed. Indeed, the usability of provenance increases if an interoperable metadata model is used, permitting the exchange and share of the geospatial data provenance in a distributed information environment [6].

The geospatial community has traditionally used the standards ISO 19115-1 [7] and 19115-2 [8] (the later initially designed for imagery). In the work (Closa et al, 2017) [9] a provenance system based partly on ISO standards is presented. The system, developed in the MiraMon GIS and RS software [10], is composed by a *provenance engine* that captures complete provenance contemporaneously to a process execution, and provides a graphical interface that shows provenance information in a tree form. We believe that is necessary to go further and extract more profit from provenance. To this goal, users should have the possibility to search (provenance queries) among geospatial data and processing tools.

3. PROVENANCE QUERIES

In order to illustrate the potential benefits that provenance can provide to users, a list of 28 general queries is formulated (*Table 1)*. The table relates the benefits of provenance (columns), with standard elements of provenance that users are querying (rows). The queries are grouped into the following 4 benefit areas:

- *Trust and authority*: Based on the sources and tools used. The authority can help in determining liability.
- *Data quality*: Related to quality of the sources and processes used, and can help to estimate error propagation (both uncertainty and blunder propagation). E.g. which sources have been produced with a tool that we know now that have a bug?
- *Documentation and reproducibility*: The documentation of the complete workflow can help in reproducibility task, especially if provenance points to real data, metadata or used tools.
- *License and accessibility*: Is related to the author right of a source and to enhancing or not the accessibility.

The table 1 is now applied to a hypothetical and plausible scenario in table 2. The scenario is:

"A European research group has been generating land cover maps by automatic classification of satellite images every five years during the last 30 years. Depending on the map version, Landsat-5, Landsat-7, Landsat-8 or Sentinel-2 images have been used. Orthophotos and regional topographic maps provided by the national mapping agency and road maps provided by the land policy department are also used as ancillary data.

Provenance elements	Provenance Benefits			
	Trust and authority	**Data Quality**	**Documentation and reproducibility**	**License and accesibility**
Sources	1. Which data have been created from untrusted sources? 2. Which are the most used sources?	10. Which data have been created using images with the highest positional accuracy? 11. Which sources have quality metadata?	16. Which are the differences between the list of sources used to produce two versions (or revisions) 17. Which sources points to real data or metadata?	22. Which data have been created using an open access data sources? 23. Which sources used are accessible in a distributed system? (list of url)
Process executions and methodology	3. In a family of methodologies, which products have used a specific member of the family? 4. Which methodologies are based on peer reviewed articles?	12. Is the methodology using the tool that produces the best results?	18. List the diferent workflows (process chained) used.	24. Which items have a citation pointing to a specific methodology?
Process algortihm	5. Which data have been created with untrusted software vendors 6. Which are the most used tools?	13. Which data have been created with the most recent vesion of the software?	19. Which are the differences between the list ofalgorithms used to produce two versions (or revisions) of the same product?	25. Which data have been created using an open source tools? 26. Which used tools are accessible in a distributed system? (list of url)
Time	7. Is one of the used sources too old for specific pursposes?	14. Are all the sources from the same temporal interval as the output?	20. Is this methodology following the state of the art?	27. Which results have been created based on sources that were not openly accessible in a period of time?
Responsible parties	8. Which institution generated the maps? 9. Who was the technician in charge to generate each map?	15. How much data was elaborated by parties with a quality certification?	21. Who (scientist/institution) is responsible of this methodology?	28. Which institution is distributing the used sources?

Table 1: This table shows the 28 generic queries. Columns are the benefits or applications from provenance, while rows represent the different standard elements of provenance.

Provenance elements	Provenance Benefits			
	Trust and authority	**Data Quality**	**Documentation and reproducibility**	**License and accesibility**
Sources	1. Which versions of land cover maps have been created using untrusted sources? 2. Who (scientist/institution) is the owner of the Ortophotos used? 3. Are all the sources using ETRS89?	10. Which automatic classifications have been generated based on more satellite images? 11. Which automatic classifications were done with less test areas? 12. In a specific version of land covers map, which land cover class have less test areas? 13. Which is the positional accuracy for the maps generated with Landsat-5 images? Landsat-8? 14. Which is the thematic accuracy for maps generated with Landsat-5 images? Landsat-7?	23. Which land cover maps were created with Landsat ? And which ones with Sentinel? 24. Which are the differences between the lists of sources used to produce two versions of the land cover maps? 25. Which are the differences between the list of sources used to produce a revision of a land cover map?	34. Are the Ortophotos used an open access data? 35. Are the ropads maps used accessible in a distributed system? (list url)
Process executions and methodology	4. Is the land cover maps generation methodology based on any peer reviewed article?	15. Which is the complete workflow to generate land cover maps? 16. Is the land cover maps methodology following the state of the art? 17. Is the land cover methodology consistent in time?	26. Which versions of land cover maps are produced with same complete workflow? 27. Does the workflow suffered any change in the different version of the land cover map? 28. Does the workflow suffered any change in a revision of a land cover map?	36. Is there any stament or citation explaining the land cover maps methodology? 37. Which is the url of the citation pointing to the land cover maps methodology?
Process algortihm	5. Who (scientist/institution) is the developer of the algorithm used to generate automatic classifcations? 6. Which land cover maps have been created using an algorithm developed by trusted software vendors?	18. Which is the thematic accuracy for the land cover maps generated with specific version of software? 19. Which is the positional accuracy for the land cover maps generated with specific version of software? 20. A bug in an algorithm is discovered, which land cover maps have been created using that version of the algorithm?	29. Which are the differences between the lists of algoritms used to produce two diferent versions of the land cover maps? 30. Which are the differences between the list of algoritms used to produce a revision of a land cover map?	38. Are the algorithms used an open access tools? 39. Is this algorithm accessible in a distributed system? 40. Which is the algorithm of this source?
Time	7. Is the ancillary data used from the same temporal interval of the land cover map?	21. Which temporal resolution have the satellite images used to generate automatic classifications?	31. Does the methology reduced the time consuming? 32. Is this methodology following the state of the art of the sector?	41. Which results have been created based on sources that were not openly accessible in a period of time?
Responsible parties	8. Which institution has generated the maps? 9. Which technician did the manual editing task in the second temporal version of the land cover maps?	22. Who (scientist/institution) generated the land cover maps.	33. Who (scientist/institution) is responsible of the land cover maps generation methodology?	42. Which institutions are distributing the land cover maps?

Table 2: This table shows 42 queries corresponding to the presented use case. Columns are the benefits or applications of provenance, while rows represent the different standard elements of provenance.

Since the last 30 years the available software have been evolving, appearing new methodologies, geoprocessing possibilities and new versions of the same applications. Finally, manual editing tasks to fix the unavoidable errors have been done by the technicians, as well. Several revisions of the same version of the land cover maps are generated until the product is validated as the final one, and all of them are stored in a database."

4. Future work and conclusions

It is recognized that geospatial provenance has benefits in distributing spatial information over the web. Thus, near future efforts should point to enhance the exploitation of the provenance information captured and tools for facilitating the queries about the "what", "when", "who", "how" and "where" of the generation of geospatial data will give an added value to the provenance information captured. In this sense, the implementation of a tool that allows to select geospatial data or geoprocessing tools on the basis of the information from provenance queries will be the main goal. Then, based on a real RS scenario, the list of the generated queries will be validated.

Important questions are not possible to solve if provenance is not provided. The standardization provides an interoperable vocabulary that we can use to make queries. The separation of concepts in the data model facilitates building software that is able to answer complex the queries to a metadata catalogue.

Acknowledgment

This work has been conducted within the framework of the Geography PhD program of the Universitat Autònoma de Barcelona. This work has been supported by the Catalan Government [SGR2017 1690]; by the European Union through the projects ECOPOTENTIAL (641762-2 EC) and ERA-PLANET (689443 EC); by the Spanish MCIU Ministry through the NEWFORLAND research project (RTI2018-099397-B-C21/C22 MCIU/AEI/ERDF, EU). Xavier Pons is the recipient of an ICREA Academia Excellence in Research Grant (2016–2020).

References

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

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

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

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

[5] Yue, P., Zhang, M., Guo, X., & Tan, Z. *Granularity of geospatial data provenance*. In Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International (pp. 4492-4495), 2014.

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

[7] ISO 19115-1."*Geographic Information- Metadata-Part 1: Fundamentals*". 2009.

[8] ISO 19115-2 . "*Geographic Information- Metadata-Part 2: Extensions for acquisition and processing*. 2014.

[9] Closa, G., Masó, J., Julià, N., Pesquer, L., & Zabala, *A. Web processing service to describe provenance and geospatial modeling*, GEOProcessing, Nice, 2017.

[10] X. Pons MiraMon. *Geographical information system and remote sensing software*. Centre de Recerca Ecològica i Aplicacions Forestals (CREAF), 2019.

6. Article 5: Geospatial Queries on a Data Collection using a common Provenance Data Model

Aquest capítol és una reproducció de: G Closa, J Masó, N Julia, X Pons. Geospatial Queries on a Data Collection using a common Provenance Data Model (article no publicat encara)

GEOSPATIAL QUERIES ON A DATA COLLECTION USING A COMMON PROVENANCE DATA MODEL

Guillem Closa [a], Joan Masó [a], Núria Juliá [a], Xavier Pons [b]

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

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

ABSTRACT

Lineage information is the part of the metadata that describes "what", "when", "who", "how" and "where" a piece of geospatial data was generated. If it is well-presented and queryable, lineage becomes very useful information to infer data quality, trace errors sources and increase trust in geospatial information. In addition, if lineage from a collection of datasets can be related and presented together, it allows for comparing datasets, workflows and methodologies.

To do that, this paper proposes the extension of process step lineage descriptions into four explicit levels of abstraction (process run, tool, algorithm and functionality). The inclusion of functionalities and algorithms descriptions as a part of lineage provides a high-level information that is independent from the details of the software used. This permits to transform lineage metadata that is initially documenting concrete processing tools into an abstract workflow that focus on the aim of the processing chain.

This paper presents a system that provides lineage information as a service in a distributed environment. The system is complemented by an integrated provenance window that is capable of visualizing and querying a provenance graph composed by the lineage of a collection of datasets.

In order to integrate provenance of a collection of datasets, we have combined ISO 19115 standards family with W3C PROV. To represent lineage elements, we chose the ISO 19115-2 lineage class names (LI_Lineage) because they express the names of the geospatial objects involved in a more geographic and precise way. The relation naming conventions of W3C PROV are used to represent relations among these elements. Elements and relations are presented as a queryable graph.

Keywords: Provenance; Data queries; Metadata

1. INTRODUCTION

The volume of Earth Observation (EO) data available is increasing progressively due to, in part, the rise of satellites and sensors promoted mainly by national and international spatial agencies such as ESA, NASA or JAXA, which are capturing data with an unprecedented level of velocity (near Real Time and Real Time). In addition to this big volume, EO data is being produced in a variety of formats with a wide range of veracity. Moreover, the continuous technological improvements allow the treatment of bigger volumes of data giving an added value to geospatial scientific research. Thus, Big Data concept (IBM 2019) also applies to the geospatial field. The convergence of all these trends together has generated that geospatial data users have difficulties to choose the more convenient EO data product for their specific purposes, based on terms of trust, data quality, authority, license and accessibility, etc. This also applies when trying to compare and select geoprocessing services (Muller 2015). Geoprocessing tools can have technical limitations, work under some licence types or have certain access restrictions. Knowing those characteristics *a priori* would help geospatial data users to choose the most suitable products for their specific purpose.

According to Baker (2016) over two-thirds of the Earth and Environment works are not reproducible because of 1) the lack of methodology or code, 2) access limitations to raw data and 3) incomplete metadata documentation. This is what Lemos (2012) has called a "usability gap". Thus, tools that improve the comprehension of data production process are needed (Spiekermann et al., 2019). Brinckman (2019) presented some recommendations to increase the level of transparency and to capture the "*Whole Tale*" of the computational environments. In the geospatial world there are some initiatives such as "Analysis Ready Data" (ARD), data prepared with minimum processing requirements and metadata (Giuliani et al., 2017), that facilitates the use and interoperability of remote sensing (RS) products and attempts to reduce the usability gap. However, ARD data may not be applicable under certain circumstances because for apparently identical implementations of common algorithms, occasional case studies generate slightly and sometimes clearly different products (Fisher, 2006) (Lutz et al., 2003). For instance, processing methods which favours particular conditions (e.g. mountain areas or Mediterranean climate). That is why it is necessary to have precise information on how an ARD product was created and, this way, to determine its a priori limitations. In this sense, information at the different abstraction levels of geoprocessing services will help to distinguish and discriminate geoprocessing tools (CEOS, 2020).

In this paradigm, geospatial lineage, information about the origins of geospatial data products, has been pointed out as a fundamental issue in spatial information (Jiang et al., 2018). Tracking back the production workflow, scientists can assess the usability, in terms of data quality, conditioned by steps more sensible to uncertainties and error propagation (Yue et al., 2011) (Zhang et al., 2017). Moreover, when lineage information is complete and points to actual data sources and code, it can help to data replication (reproducibility purposes) and to workflow reuse (with other inputs). Summarizing, lineage information helps to overcome the knowledge gap between data providers and data consumers who want to reuse these models, data or algorithms in different contexts, regions or purposes.

One could assume that a simple statement could be enough for documenting lineage, but a well-known internal structure is needed to extract the maximum benefit of it. Indeed, the usability and the utility of lineage increases if an interoperable metadata model is used, permitting the exchange and share of the geospatial data lineage in a distributed information environment (He et al., 2015). The Provenance Working Group of the World Wide Web Consortium (W3C) defined a model to represent provenance over the web. The W3C-PROV (PROV from now on) defines provenance as information of sources, process executions, algorithms used, conceptual frameworks behind algorithms, moment of execution and responsible parties involved in producing a piece of data or thing (Groth and Moreau 2013). In the geospatial domain the

term lineage was used to define the provenance of geographic information systems (GIS) products (Lanter 1991). The Spatial Data Transfer Standard (SDTS) (SDTS 1998) of the Federal Geographic Data Committee (FGDC) defined a lineage model, and the International Organization for Standardization (ISO) included a lineage model first in the ISO 19115:2003 an later in ISO 19115-1:2014 (ISO 2014) and ISO 19115-2 (ISO 2019). Although the term "lineage" is more associated to geospatial standards, several works use both terms as a synonymous (Di et al., 2013) (Di et al., 2013). In this paper we introduce a slight differentiation: we consider the term lineage as the history of a single dataset, while provenance refers to the integrated history of more than one dataset.

According to Ivanová (2017), the initial version of the ISO model offered an unstructured narrative to the history of the spatial resource and thereby it is unsuitable for automation purposes in the web. To cover this gap, López-Pellicer et al. (2014) and Closa et al. (2017) propose adapting the PROV model to the geospatial community requirements. Other authors such as He et al. (2015), and Jiang (2018) went further and semantically enriched the PROV structure with ISO concepts in order to permit the representation of geospatial particularities with PROV. However, the recently edited version of ISO (ISO 19115-2:2019) has improved substantially in structure and now is able to better represent the workflow of a production line (Closa et al., 2019).

Another key factor to enhance the comprehension of the data production description is the selected provenance visualization approach. In complex environments, scientists rely on visualization tools to help them understand large amounts of data that are generated from experiments (Salton et al., 1994). Visualization tools are essential in the steps of *discovery* and *inspection* of data and workflows (Konkol & Krai 2019). Provenance can have a complex structure with multiple relations and dependencies. This can make users feel overwhelmed when exploring the different steps that lead to a dataset. Given the linked nature of the provenance information, one possible approach to simplify visualization is the use of graphs that summarize the process effectively (Yazici., et al 2018).

Some GIS and RS tools provide users with some functionality to store lineage information. In addition, spatial agencies have sometimes been particularly careful on saving lineage information as part of their data preservation strategies. However, despite the potential of the recorded lineage information, it is rare to see a system that provides query capabilities that goes beyond basic metadata visualization. In this regard, there is a need for interactive systems and tools able to visualize, query or mine provenance information (Cohen-Boulakia et al., 2017). Nevertheless, given the multiple relations and dependencies between different datasets that provenance information can describe, designing such tools is a challenging task.

This contribution tackles this issue presenting a system that provides lineage information as a service in a distributed environment that can be queried. The system is developed in order to be able to query lineage not only of specific data, but the provenance of a collection of datasets or federated metadata services. Our graphs go beyond to the typical lineage graphs which are sources or process workflow oriented; instead, the system can show the used tools, the executions done, the outputs generated, or the agents involved in the collection of datasets. In order to be able to include all the heterogeneity and variety of provenance information in a single graph, four different levels of geoprocessing abstraction are proposed, concretely at execution, tool, algorithm and functionality levels. The preliminary results are tested in a map server.

The rest of this paper is organized as follows: in Section 2, the chosen provenance representation system, paying special attention to the different level of granularity of the geoprocessing tools, is presented; in Section 3, the potential for provenance query is described; in Section 4, a query provenance system is presented; in Section 5, a use case to show the implementation of the system and the visualization tool is described. Section 5 also provides a discussion based on a use

case that exemplifies the usefulness of our proposal. The identification of future work is covered in Section 6 and finally the summary the conclusions obtained are presented in section 7.

2. PROVENANCE MODEL

This paper takes advantage of the legacy of the lineage data model in the ISO 19115 family standard and the W3C PROV provenance data model to propose an evolved data model that relates collections of datasets in a network using provenance as a basis. On top of this legacy, the proposed model presents a set of levels of abstraction for process steps.

2.1. Levels of Abstraction of process steps

The definition and capture of the different levels of granularity of geospatial provenance data has motivated several works such as the one of Yue et al. (2014). In the work of Closa et al. (2017) ISO and PROV were compared to describe the provenance of different levels of abstraction of geospatial data (feature types, features, attribute types, attributes) and a proposal to describe the different levels of granularity with PROV was stated. In this present work we see the necessity to define the different levels of abstraction of the processing steps too: process run, processing tool, algorithm and functionality (see Figure 1). The meaning of these concepts is (going from more concrete to more abstract):

- Process run (process step): An individual execution of a processing tool where a concrete set of parameters is used. It is a single GIS operation.
- Processing tool (executable): A concrete version of an implementation of an algorithm in a piece of software that obviously can be executed several times with different sources. It is what we can find in the GitHub, buy from a software vendor, etc.
- Algorithm (model): A set of mathematical and logical steps that allows to transform some inputs into some outputs. It can be implemented in software in different ways and programming languages. It is what usually a scientific paper describes.
- Functionality (operation): an operation that transforms data into other data or information with spatial problem-solving orientation. It is a black box that can be implemented with different algorithms giving slightly different results. It is what a GIS and RS textbook describes.

These levels of abstraction are related as follows (see Figure 1): the process runs, as a single execution, *executes* a processing tool. The processing tool *implements* an algorithm. Finally, the algorithm *gives* a functionality.

The inclusion of functionalities and algorithms descriptions as a part of provenance provides a high-level provenance information that is independent from the software used. This permits to take a provenance graph that is initially documenting concrete processing tools and abstract it into a higher-level diagram that describes the aim of the processing chain. This idea goes beyond pure reproducibility by providing reasoning and the intentions that are behind each process step. Exploiting this approach allows to:

- Represent together in a single provenance representation the origin of different datasets.
- Formalise provenance queries at different levels of abstraction and over a complete dataset. For instance (from more abstract to more concrete):
    - What functionalities are more frequently used in my organization?
    - Given a quality test in my final products, what is the best algorithm that I can use?

- How are my results affected by a specific processing tool version that has a bug?

- Translate a lineage description done in one software brand into another software product and to reproduce the results.

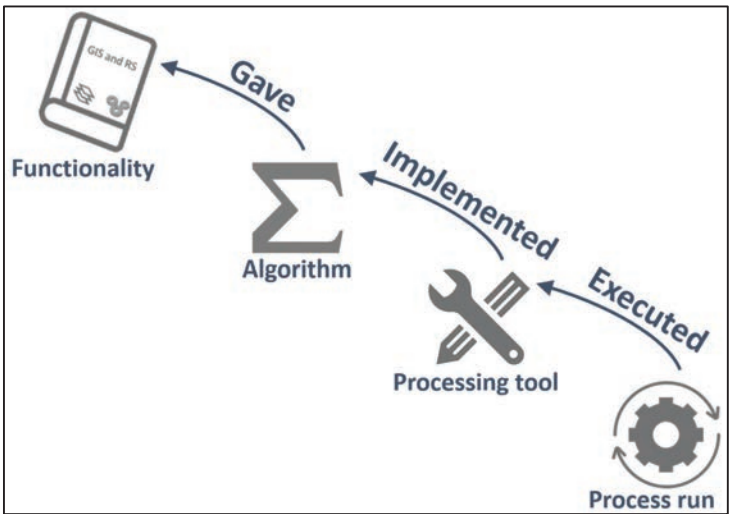

Figure 1: The representation of the four levels of process steps abstraction

The description of all functionalities used in a process chain depicts the task that the workflow was designed for. In the geospatial domain a task describes all actions that require human input or the knowledge about context and is usually composed by functions (Albrecht, 1998) (Sun et al., 2012) (e.g. watershed delineation or a polluting industry buffer zone delimitation). In any case, tasks are not bound with specific tools. Back in 1998, Albrecht (1998) demonstrated that is feasible to translate flow charts based in a universal GIS functionality into specific GIS software flow charts when the link between functionalities and GIS operations (tools used) is clear. To this purpose the different GIS software tools should be clearly linked to common GIS functionalities. In this sense, even though several classifications of the principles of GIS functionalities have been formulated (Goodchild 1991) (Kuhn 2015), semantic descriptions are still ambiguous or incomplete (Yue et al. 2015). This is possible the reason why the authors of this work fail to find a common, universally accepted, classification of the GIS functions. However, in spite of this situation, it is generally acknowledged that the main GIS software perform a common core of functionalities (tools with the same problem-solving intentionality (in concordance with the exposed in section 1, we are aware that each software tool favours specific conditions and, then, results will have differences)). The same can be applied to RS functionalities. In this regard, Table 1 shows a subset of this common GIS and RS functionalities and the name of the implementation in ArcGIS (ESRI, 2020), MiraMon (Pons, 2020), GRASS (GRASS Development Team, 2017) and SNAP (SNAP – ESA, 2020).

Table 1: A relation of some of the functionalities in GIS and RS with the different implementation names in ArcGIS, MiraMon, GRASS and SNAP software

GIS functionality	ArcGIS tool	MiraMon tool	GRASS tool
Geometric union	Union	CombiCapa	v.overlay(or)
Extraction	Clip	Retalla	v.overlay(and)
Proximity	Buffer	BufDist	v.buffer
Distance	Distance	BufDist	r.distance
Surface interpolation	Interpolation	InterPNT	r.resamp.interp
Slope	Slope	Pendent	r.slope.aspect

Aspect	Aspect	Pendent	r.slope.aspect
Shade	Hillshade	Illum	r.relief
Visibility	Viewshed	Visible	r.viewshed
Cell statistics	Cell statistics	EstRas	r.univar
Field statistics	Field statistics	EstCamp	v.vect.stats
Joining	Join	GestBD	v.db.join
Merging	Merge	GestBD	v.patch
Aggregation	Dissolve	Ciclar	v.disolve
Feature selection	Select by features	VecSelect	v.extract
RS functionality	**SNAP**	**MiraMon tool**	**GRASS tool**
Georeferencing	Orthorectification	CorrGeom	i.ortho.photo
Radiometric correction	Sen2Cor	CorRad	i.atcorr

2.2. Linking geospatial datasets collections through the history of processes

In Closa et al (2019) a lineage system was presented partly based on ISO 19115 family standard. The interactive metadata visualization tool (GeMM), part of the MiraMon GIS and RS software (Pons 2020), provides a graphical interface that shows lineage information in a hierarchical tree form. A tree represents the lineage of an isolated geospatial dataset, but it is not possible to represent the provenance of a collection of datasets as a forest.

While the lineage model of the ISO metadata standards focuses on the final product instances and their history, PROV focuses on the relation of different pieces of information in terms of knowledge creation flow and is commonly represented as a graph. This does not mean that the ISO model cannot be represented as a graph, but the nature of their relations makes more indirect to express it and sometimes implies the use of multiple connections for representing a single relation. These difficulties increase when representing provenance of a dataset collection. In this sense, several authors (Jiang et al. 2018) (Lopez-Pellicer et al. 2015) have proved that mapping both models is possible (Table 2) and the two models (PROV and ISO) can be used to represent provenance of equivalent data. Thus, the combination of both models is possible.

Table 2: This table show the PROV-ISO equivalences

PROV relations	ISO elements and properties relating them
Used	LI_ProcessStep/source →LI_Source (or LE_Source)
WasAssociatedwith	LI_ProcessStep/processor→CI_Responsability
WasGeneratedBy	LI_ProcessStep/output→LE_Source
WasAttributedTo	LE_ProcessStep/processingInformation/LE_Processing/softwareReference/CI_Citation/citedResponsibleParty→CI_Responsability
WasDerivedFrom: Gave	LE_ProcessStep/processingInformation/LE_Processing/procedureDescription →CharacterString

wasDerivedFrom: Implemented	LE_ProcessStep/processingInformation/LE_Processing/algorithm → LE_Algorithm
Use: Executed	LE_ProcessStep/processingInformation →LE_Processing

2.3. W3C PROV to represent the abstraction levels of processes

In PROV, provenance is the representation of entities (e.g. source, output, executable, etc), agents (person, institution, etc) and activities (e.g. execution). Entities, Activities and Agents form the PROV Data Model (PROV-DM) core types are related by the PROV-DM core relations:

- Used→ Relates activities (PROV:Activity) with (PROV:Used) entities (PROV:Entity).
- WasAssociatedwith→ Relates activities (PROV:Activity) with (PROV:wasAssociatedwith) agents (PROV:Agent).
- WasGeneratedBy→ Relates entities (PROV:Entity) with (PROV:WasGeneratedBy) activities (PROV:Activity).
- WasAttributedTo→ Relates entities (PROV:Entity) with (PROV:wasAttributedTo) agents (PROV:Agent).
- WasInformedBy[1] → Relates activities (PROV:Activity) with (PROV:WasInformedBy) activities (PROV:Activity).
- WasDerivedFrom[1] → Relates entities (PROV:Entity) with (PROV:WasDerivedFrom) entities (PROV:Entity).
- ActedOnBehalfOf[1] → Relates agents (PROV:Agent) with (PROV: ActedOnBehalfOf) agents (PROV:Agent).

In addition to the PROV core types relations, which are high-level descriptions, there are mechanisms to 'open up' those descriptions to a lower level specification. In this sense, we have introduced three subtyping of PROV-DM core relations to relate the four level of processing abstraction described in section 2:

- Executed → The subtype *Used:Executed*, is introduced to relate the Process runs (PROV:Activity) with the Processing tool (PROV:Entity). A Process run *executed* a Processing tool once.
- Implemented → The subtype *WasDerivedFrom:Implemented* is used to relate the Processing tool (PROV:Entity) with an Algorithm (PROV:Entity). A Processing tool implemented an Algorithm.
- Gave → The subtype WasDerivedFrom:Gave is used to relate an Algorithm (PROV:Entity) with a Functionality (PROV:Entity). An algorithm gave Functionality.

2.3.1. The combination of W3C PROV and ISO19115 to represent provenance

In this paper, to encode provenance of a collection of datasets, we have chosen a composed solution: the combination of ISO with PROV. To represent each element, we chose the ISO 19115-2 lineage class names (LI_Lineage) because it expresses the names of the geospatial objects involved in a more geographic and precise way. The relation naming conventions of PROV are used to represent relations among agents (CI_Responsability), actions (LE_ProcessingSteps) and entities (anything else). The result of this combination can be seen in Figure 2.

[1] These PROV core relations are not used in this proposal.

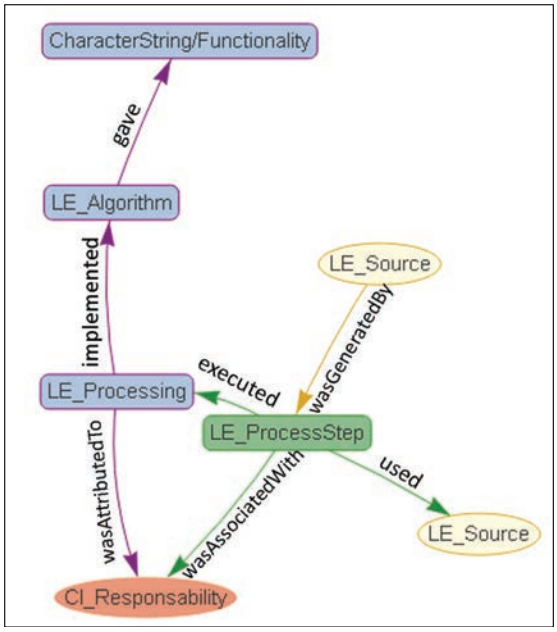

Figure 2: This graph shows how the use of PROV is used to relate the different lineage elements. The ISO naming is restricted to the description of the lineage elements

Figure 2 can be read as follows:

> *A Source (LI_Lineage:LE_Source) is used (PROV:use) by a Process step (LI_Lineage:LE_ProcessStep). An output (LI_Lineage:LE_Source) was generated by (PROV:WasGeneratedBy) a Process step. A Process step was associated with (PROV:wasAssociatedWith) an Agent (LI_Lineage:CI_Responsability). A Processing tool is executed (PROV:used:executed) by a Process step. A processing tool implements (PROV:wasDerivedFrom:implemented) an algorithm (ISO:LE_Algorithm). The algorithm gave (PROV:wasDerivedFrom:gave) the Functionality (LE_ProcessStep/processingInformation/LE_Processing/ procedureDescription) of a geoprocess.*

Relating the four levels of abstraction with PROV we can describe the provenance of many products in a single relational graph. We propose the use the global identifiers to sources and processing tool (ISO MD_Identifier) (Masó et al. 2012) , to coalesce repeated objects and to integrate remaining objects in a provenance graph connected by PROV relations (see Figure 3). By no longer separating the individual product lineage instances, PROV opens the door to a new set of possibilities in terms of representation and query over the data of a collection of datasets.

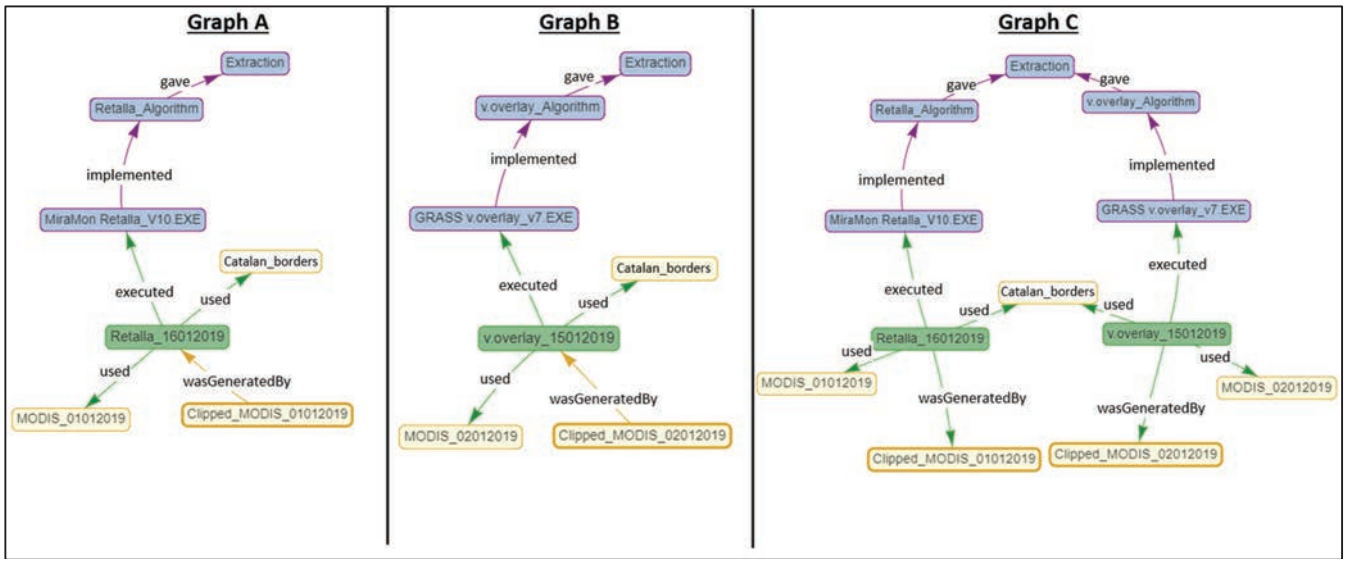

Figure 3: Graph A and graph B represent individual executions with the same functionality done with MiraMon and GRASS respectively. Graph C represents provenance of both executions merged into one.

3. QUERIES FACILITATED BY THE PROVENANCE DATA MODEL

The exploitation of provenance is really enhanced when formulating queries. There are several examples in the literature where a consolidated and standardized data model and the associated interoperable vocabularies is the base for a query language that exploits the data expressed in the data model. Erwig et al (1999) see this relation for spatio-temporal data; Koubarakis et al (2012) show this relation in the linked data; and Amman et al (1993) recognizes the link between a model and query language for graphs. The same happens for the presented provenance information data model. The separation of concepts and the introduction of the different levels of abstraction in the data model facilitates formulating formal queries that involve this concepts and relations. In addition, allows for querying not only on a specific dataset, but also in a dataset collection.

Queries can be formulated over the different lineage elements. In *Table 3* forty-four general queries over lineage elements are provided. The queries presented are only an example of the potential of what we can get if we have the possibility to query provenance from a collection of datasets. The table is a dual entry table which relates the different lineage elements between them. Only the first row restricts queries to the ones that affect only to a one lineage element.

Table 3: Forty-four examples of queries to provenance**.**

	Process run	**Processing tool**	**Algorithm**	**Functionality**	**Agent**	**Source**	**Time**	**Output**
	Was any execution covering Africa?	Was version 3 of *InterPNT* used?	Was *kriging* algorithm used?	Was a *reprojection* functionality used?	Had the user Bob a role in the creation of this dataset?	Did we used a dataset called Rivers?	Was something executed in 2013?	Did we create a rain dataset?
Process run	**What was executed after the Process step 5?**	Did the Process step 5 use the version 3 of *InterPNT*?	Was the Process step 5 a *krigring* interpolation?	What was the purpose of Process *step 5* ?	Was the Process step 5 executed by Bob?	Did the Process step 5 use a DEM of 2m?	How long did the Process step 5 last?	Which data was generated with Process step 5?
Processing tool		**Which tool is often used right after *InterPNT*?**	Does the version 3 of *InterPNT* support an *IDW* interpolation?	Did *InterPNT* and *r.resamp.intep* implement equivalent functionalities?	Which interpolation tools were developed by a trusted software vendor?	Did the version 3 of *InterPNT* used any GeoJSON format?	Is the version 3 of *InterPNT* the last version available?	Which outputs were created with the version 3 of *InterPNT*?

Algorithm			**Which are the different versions of the *buffer* algorithm used?**	Did GRASS and MiraMon *buffer* tools use the same algorithm?	Was Bob the author of any of the used algorithms?	Is this algorithm suitable for categorical data?	When was this algorithm developed?	Which outputs were created using this algorithm?
Functionality				**Have all the corrections been done with the same software?**	Who did the radiometric corrections?	Which used datasets were reprojected?	Was something reprojected in 2015?	Which outputs were reprojected?
Agent					**Who used tools developed by Bob?**	Which of the sources used were produced by a public institution?	When did Bob make his first execution in this collection?	Which institution generated the resulting maps?
Source						**Which two sources were used together**?	Were all the sources from the same temporal interval as the output?	Was a rain intensity dataset needed to create rivers flow dataset?
Time							**How long it took to complete production?**	When was this output generated?
Output								**Was this output a revision of another output?**

Answers allow users to inspect directly the datasets generation history. Depending on which aspect of provenance is queried, different benefits can emerge:

- Information and transparency: Description of the lineage allows us to learn how datasets are done
- Trust and authority: of the sources and tools used: The authority can help in determining liability.
- Data quality: Sources and processes used can be used to estimate uncertainty and blunder propagation. E.g. which sources have been produced with a tool that we know now that has a bug?
- Documentation and reproducibility: The documentation of the complete workflow can help in the reproducibility, especially if provenance points to actual and exact datasets, parameters and tools used.
- License and accessibility: Related to the author rights of a source.

4. REPRESENTATION AND QUERY PROVENANCE TOOL

4.1 MiraMon Map Brower

The MiraMon Map Browser is a visualization, analysis and download web application that runs in web browsers (Masó et al., 2020). The browser is coded with HTML5 and JavaScript, it uses Open Geospatial Consortium web standards service protocols and APIs to communicate with web services to get the minimum subsets of the information necessary to create a fast and dynamic user interaction. The MiraMon Map Browser can be configured to provide an integrated view of several datasets having something in common (geographic or thematic or both). These datasets might come from a single service or by several services from different institutions. In this paper we present the design of a new characteristic of the MiraMon Map Browser to provide an integrated provenance visualization and demonstrate the potentialities of querying provenance.

Lineage is communicated from servers hosting the metadata to the client that is capable of merging and presenting it in a provenance graph (see Figure 4). In our implementation, the lineage information of each dataset is communicated from service to client independently. The client relates the lineage of each dataset in an integrated provenance graph. The integrated view takes advantage of the connections created by common processes or sources at the different abstraction levels that the individual dataset lineage may have. To achieve that, for each dataset, the client software needs to identify the common elements processes and sources and merged them before presenting the graph. On top of that, the client offers different ways to filter and query the resulting graph. This permits the user to control the amount of content of the graph and progressively increase understanding of the graph itself and, with that, of the provenance information that represents.

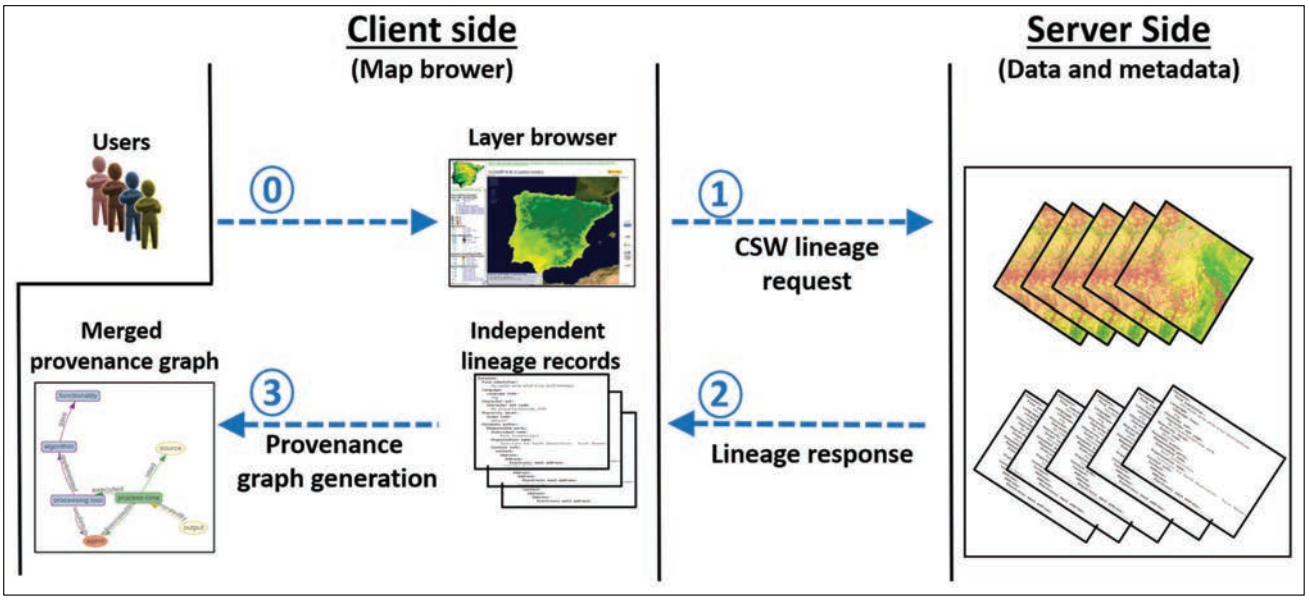

Figure 4: Steps that make possible that users retrieve information to build a integrated provenance graph. A CSW lineage request (1) to the server allows the retrieval of the independent lineage of each requested dataset (2). Finally, the client generates the provenance graph (3)

4.2 Lineage server

Lineage is part of the metadata, so one possible solution to retrieve lineage in a standard way is to use the OGC Catalogue Service for the Web (CSW). The *GetRecordById* request retrieves the default representation of metadata records characterized by this identifier. However, instead of getting the whole metadata record, we wanted a way to retrieve only the lineage information from the servers. Thus, we came up with a small extension of the CSW protocol that includes the *ELEMENTSETNAME* key that has "*lineage"* as a value. In addition, to facilitate the reading process in the JavaScript client we also included the *OUTPUTFORMAT* key and the possibility to request the lineage in a JSON encoding of the ISO 19115 data model. Currently there is no official JSON encoding for the ISO19115, so we defined one using the draft rules proposed in the OGC Architecture DWG JSON best practice (https://github.com/opengeospatial/architecture-dwg/tree/master/json-best-practice). All these modifications were implemented in the MiraMon Map Server. The MiraMon Server is a stand-alone CGI application that runs on Windows operating systems and can be used in combination with a general-purpose web server (e.g. Internet Information Service, Apache, etc).

The example in Figure 5: A CSW *GetRecordById request* operation

```
www.ogc3.uab.cat/cgi-bin/mcsc/MiraMon.cgi?SERVICE=CSW&REQUEST=GetRecordById&OUTPUTSCHEMA=
http://www.isotc211.org/2005/gmd&ELEMENTSETNAME=lineage&id=MCSCv2Nivell2:EPSG:4326&OUTPUTFORMAT=application/js
on
```

shows a CSW *GetRecordById* operation which returns the lineage information in a JSON encoding. Part of the JSON response is in Figure 6: Fragment of the CSW *GetRecordById* operation response in JSON

```
      "lineage":{
                  "statement": {"cat":" The present raster is the result of the change of cartographic
projection of the CLCM2 level 2 layer from SR UTM-31N-ED50 to SR UTM-31N-ETRS89."},
                  "processes":[
               {"processor":[
                     {
                     "role":"processor",
                     "party":{
                        "organisation":
                           {
                           "name":"CREAF"
                           }
                        }
                     }
                     ],
                  "purpose":"Metadata edition of the CLCM2 level 2",
                  "timeDate":"2020-06-19T19:29:38.069+02:00",
                  "executable":
                     {
                     "reference": "c:/miramon/GeMM.exe",
                     "compilationDate":"2020-06-19T19:29:38.069+02:00",
                     "functionality": "Metadata management and edition "
                     },
                  "parameters":
                  [
                     {
                     "id":"Param1",
                     "name":"MCSCv2Nivell2_1",
                     "direction":"in",
                     "purpose": null,
                     "valueType":"source",
                     …
```

.

Figure 5: A CSW *GetRecordById request* operation

```
www.ogc3.uab.cat/cgi-bin/mcsc/MiraMon.cgi?SERVICE=CSW&REQUEST=GetRecordById&OUTPUTSCHEMA=
http://www.isotc211.org/2005/gmd&ELEMENTSETNAME=lineage&id=MCSCv2Nivell2:EPSG:4326&OUTPUTFORMAT=application/js
on
```

Figure 6: Fragment of the CSW *GetRecordById* operation response in JSON

```
      "lineage":{
                  "statement": {"cat":" The present raster is the result of the change of cartographic
projection of the CLCM2 level 2 layer from SR UTM-31N-ED50 to SR UTM-31N-ETRS89."},
                  "processes":[
               {"processor":[
                     {
                     "role":"processor",
                     "party":{
                        "organisation":
                           {
                           "name":"CREAF"
                           }
                        }
                     }
                     ],
                  "purpose":"Metadata edition of the CLCM2 level 2",
                  "timeDate":"2020-06-19T19:29:38.069+02:00",
                  "executable":
                     {
                     "reference": "c:/miramon/GeMM.exe",
                     "compilationDate":"2020-06-19T19:29:38.069+02:00",
                     "functionality": "Metadata management and edition "
```

```
    },
"parameters":
[
    {
    "id":"Param1",
    "name":"MCSCv2Nivell2_1",
    "direction":"in",
    "purpose": null,
    "valueType":"source",
    …
```

4.3 Provenance representation

The use of a JSON encoding is particularly convenient for a JavaScript client. A JSON file can be converted into a JavaScript data structure with only one sentence of code. Internally, the MiraMon Map Browser also relies on a JSON file to define the configuration of the client. In particular, contains an array of data layers (representations of datasets) that describes their metadata and schemas. Among those layers metadata and schemas, lineage is described exactly in the same way as the JSON response of our lineage server. This provides the capability to have lineage specified directly in the configuration file when the layer has no CSW server to respond. The configuration file can be validated with the JSON file following the schema language (https://json-schema.org/) that includes the data model description and the validation information for the lineage part. The MiraMon Map Browser JSON schema file can be found in GitHub (https://github.com/grumets/MiraMonMapBrowser/blob/master/src/config-schema.json).

4.4 Provenace interface

To provide a more flexible provenance visualization, we departed from the tree visualization implemented in the GeMM and mentioned in section 2.2. In the map browser we wanted to represent lineage of one dataset or to combine lineage from more than one dataset in a single provenance view. Thus, we have opted for using a network representation by using vis.js library, which has an implementation of graph diagram visualizations. A graph is defined as a set of nodes that have identifiers and a set of edges that connect nodes. In the vis.js library, nodes and edges are described as two independent arrays of JavaScript objects in an encoding that is very different from the one used by the ISO19115 (and our JSON transcription of it), that are based on the concept of objects (e.g. process steps) that have other objects (e.g. sources) as properties, recurrently. A JavaScript code converts the JSON encoding based on the ISO 19115 into the JSON arrays required by vis.js. In this conversion, a process step is represented as a blue box with purple border, a tool as dark green box, an algorithm as purple box and a functionality as green box. A source is represented as yellow ellipsis and an agent as oranges circle. Edges use the color of their origin and have the PROV relations as labels (see section 2.3). Finally, the bright yellow ellipse is reserved for the resulting dataset itself (see Figure 7, panel 2).

Users start the process of visualizing integrated provenance by checking for the presence of lineage information of a single dataset in the legend (see Figure 7). Then all the processes steps, processing tools, algorithms, functionalities, agents and sources documented in the production process of that dataset are displayed in a provenance window. The vis.js library calculates the optimal position of nodes in provenance window to avoid overlaps. This window can be resized, moved and there are a zoom-in and zoom-out options provided by the vis.js library. Users can still move nodes around as required. Additionally, JavaScript code handles the *onclick* events and shows more information about the node in a text area.

Once users have the provenance graph with a single dataset represented, there are have several options to continue exploring provenance (see Figure 7, panel 1):

- Users can select some lineage element types to be hidden in order to simplify the visualization (see Figure 8, panel 2):
    - Agents: they can be hidden with no consequences to simplify visualization.
    - Leaf sources (sources that existed independently of the executed process step): to make the workflow simple, they can be removed from the view,
    - Internal and many times temporary sources (sources which were produced during the workflow execution): In order to enhance the comprehension of the workflow execution they can be removed from the view.
    - Processes steps: When steps are removed, tool names take their place.
    - Tools: When tools are removed, they are replaced by the algorism names.
    - Algorithms: When algorithms are removed, they are replaced by the functionality provided.

    In the last 3 described sequence of events, the represented provenance becomes more abstract and less dependent from the details of the software used, making the lineage of two products created with different software comparable.

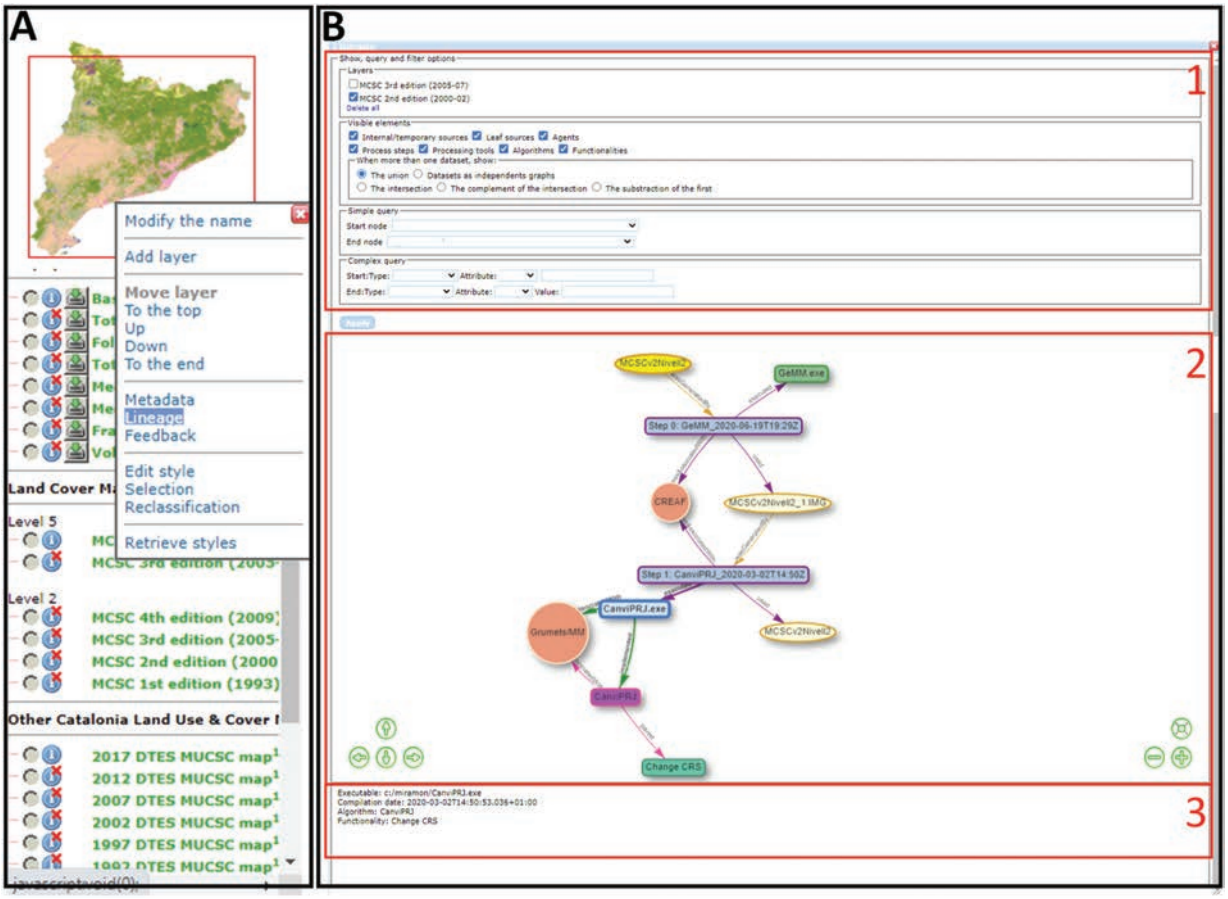

Figure 7: **A.** Legend panel: users can navigate to a specific layer present in panel, click with the mouse button and select *Lineage* selection option in the layer. **B.** Provenance window: **1**- Visibility and query options panel (detailed in Figure 8). **2**- Provenance graph panel. **3**- Element attributes panel

- Users can select another dataset. The "incoming" lineage elements are accumulated in the provenance window. This combined graph can be represented in different ways (see Figure 8, panel 3):
    - A new independent graph that is presented next to the previous one in the same window.
    - The union of all lineage element in a provenance graph: The common elements are represented only once, allowing users to see the full picture of the provenance, including provenance connections between two production processes, such as shared sources, tools, agents, etc.
    - The intersection between the two graphs: The common elements are represented, creating a view that only presents the nodes that connect and are shared by both lineage graphs. Common elements are the most useful.
    - The subtraction of the first graph: Only elements of the first graph that are not present on the other graph are represented. This puts the emphasis on what is different in the first graph form the second one.
    - The complement of the intersection: The elements that are not common in the graphs are represented, putting the emphasis in the elements used only once.
- Users can click on the box of the process steps and request to group processes with the previous step or with the next step. This creates a "virtual" process that is the sequence of the previous two; in the same way as we create batch processes.
- User can check the lineage statement by clicking with the right button on the resulting dataset (bright yellow ellipsis).

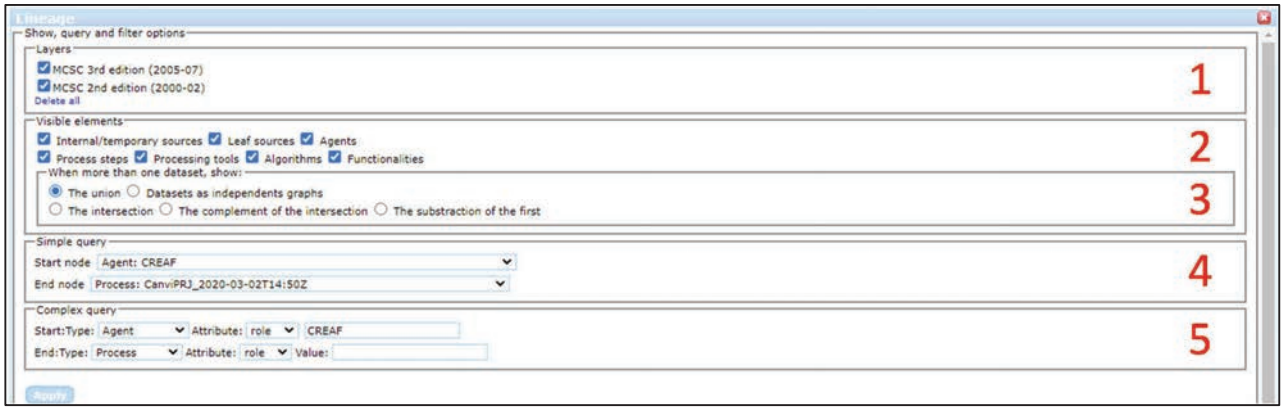

Figure 8: Visibility query and filter options panel: **1**- Layers panel. **2**- Lineage elements filtered/visible panel. **3**- Lineage fusion options. **4**- Simple queries panel. **5**- Complex queries panel.

4.5 Provenance query tool

Users can generate queries to the provenance graph resulting from the dataset activated in the Layers panel (see Figure 8, panel 1). The selection of lineage elements (see Figure 8, panel 2) described in section 4.4 also applies on the query result. There are two types of queries:

- **The simple query** interface offers two lists of all objects that are present in the graph (classified per type). An algorithm determines if the start object is connected with the end object and masks them as well as all intermediate nodes that connect them. For instance, we would like to check the activity of the Agent "CREAF" regarding to a concrete CanviPrj execution (see Figure 8, panel 4).

- **The complex query** allows selecting two objects types and their respective attribute values. In this case, the user does not know the start and the end nodes names, several start and end nodes might be marked as selected. Not filling the attribute value will result in selecting all the nodes with the same attribute type as starting or ending point. For instance, we would like to check the activities of the Agent "CREAF" regarding to all the executions (see Figure 8, panel 5) whatever they are. As in the previous case, the objects that match the query are selected as well as all objects that connect them.

Once the provenance queries are solved, the provenance window can present the resulted provenance information in two different forms:

- A graph representing only the elements that were selected by the query. The result is simpler but some relations to other objects that are essential to understand the graph might not be visible.
- A full graph with all element but with the selected elements emphasized. This option is more useful for graph that contains a limited number of elements.

5. USE CASE: CATALONIA LAND COVER MAP

The Catalonia Land Cover Maps server (http://www.opengis.uab.cat/mcsc/) provides the first (1993), the second (2000), the third (2005) and the fourth (2009) edition of Catalonia Land Cover Map (MCSC) (https://www.creaf.uab.es/mcsc/), and 1987, 1992, 1997, 2002, 2007, 2012, 2017 editions of Land Use and Cover Maps of Catalonia (DTES MUCSC) (DTES web site[2]) (see Figure 8). Even though the usability of both products is similar, its workflows generation and what are representing is quite different.

On the one hand, the MCSC were done by photointerpretation and digitizing on computer screen, which permits the use of other digital cartography elements as a direct support for the process. The base materials for the photointerpretation are a set of orthophotos in natural colour from the Cartographic and Geologic Institute of Catalonia (ICGC). The 2005 and 2009 editions have hierarchical levels of complexity of the legend (the simplest is the level 1 and the more complex is the level 5) (Ibàñez et al., 2010). The server contains the level 2 of all editions and the level 5 of 3rd and 4th edition. On the other hand, MUCSC were generated using automatic classification of satellite imagery and auxiliary cartography. While the 1987, 1992, 1997 and 2002 MUCSC maps editions where generated by the ICGC, the 2007, 2012 and 2017 editions were generated by the UAB. In addition, 1987, 1992, 1997, 2002 and 2012 MUCSC were done using Landsat images (Landsat 5, Landsat 7 or Landsat 8 depending on the edition), the 2017 was done with Sentinel 2 images (Gonzalez-Guerrero et al., 2019). The software used has evolved, with new methodologies and new versions of the same applications. Finally, manual editing tasks to fix the unavoidable errors of the automatic classification have been done. This scenario represents a good example to validate the visualization and queries techniques developed within the framework of MiraMon Map Browser.

5.1 Provenance visualization examples

Based on the MCSC scenario, some examples of provenance visualization capabilities can be showed:

- **Visualization example *1*** (see Figure 9, left image): A provenance graph shows the agents involved in the generation of the MCSC version 1. The attributes text panel shows the attribute of the lineage selected element

[2] https://territori.gencat.cat/ca/01_departament/12_cartografia_i_toponimia/bases_cartografiques/medi_ambient_i_sostenibilitat/bases_miramon/territori/mapa-dusos-i-cobertes-del-sol/

(in this case, *processor:CREAF*). The visibility, query and filter options panel has only the layer MCSCv1Nivell2 selected and the agents are visible.

- **Visualization example *2*** (see Figure 10, right image): The provenance graph panel shows processing tools and sources involved in the MCSC version 2 generation. In this example process steps have been abstracted into used tools. The attributes text panel shows the attribute of the lineage selected element (in this case, *Resulting dataset*). The visibility query and filter options panel has only the layer MCSCv1Nivell2 selected and the internal sources, leaf sources and processing tools are visible.
- **Visualization example *3*** (see Figure 11): The provenance graph panel shows a representation of the combination of the lineage of the MCSC version 1 and 2. In this example the shared lineage elements are detected and represented only once. The attributes panel shows the attribute of the lineage selected element in this case, *processor:Grumets/MM*). The visibility, query and filter options panel have both layers selected, MCSCv1Nivell2 and MCSCv2Nivell2, and all lineage elements are selected.

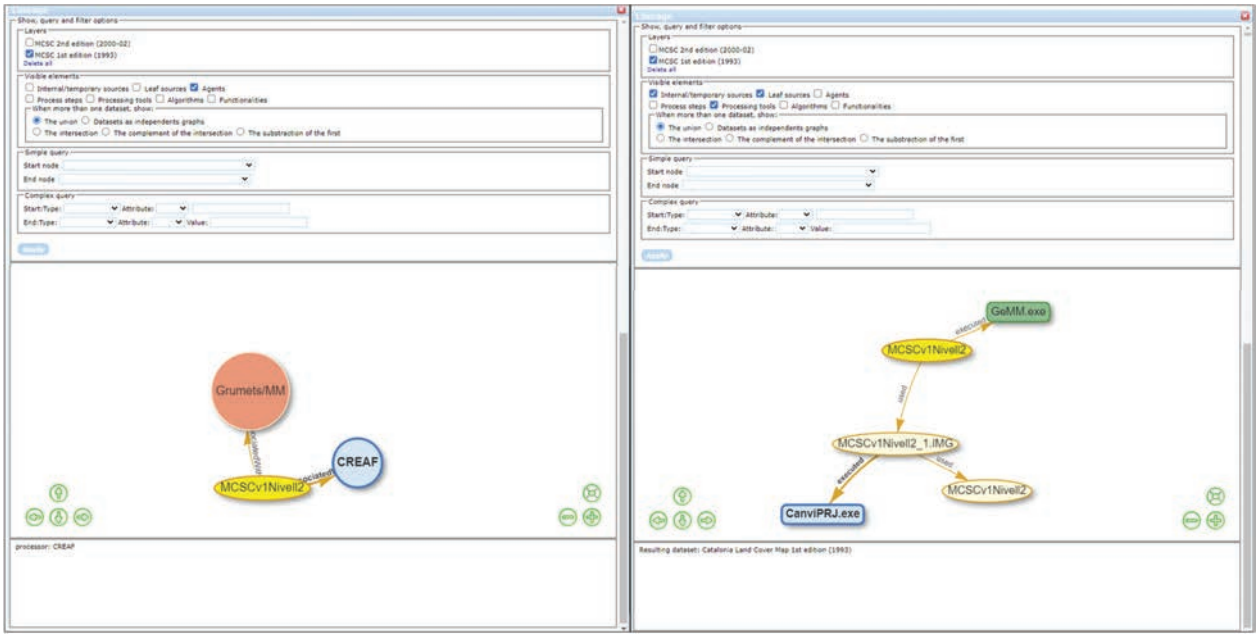

Figure 9: Visualization example 1 (left image) and Visualization example *2* (right image)

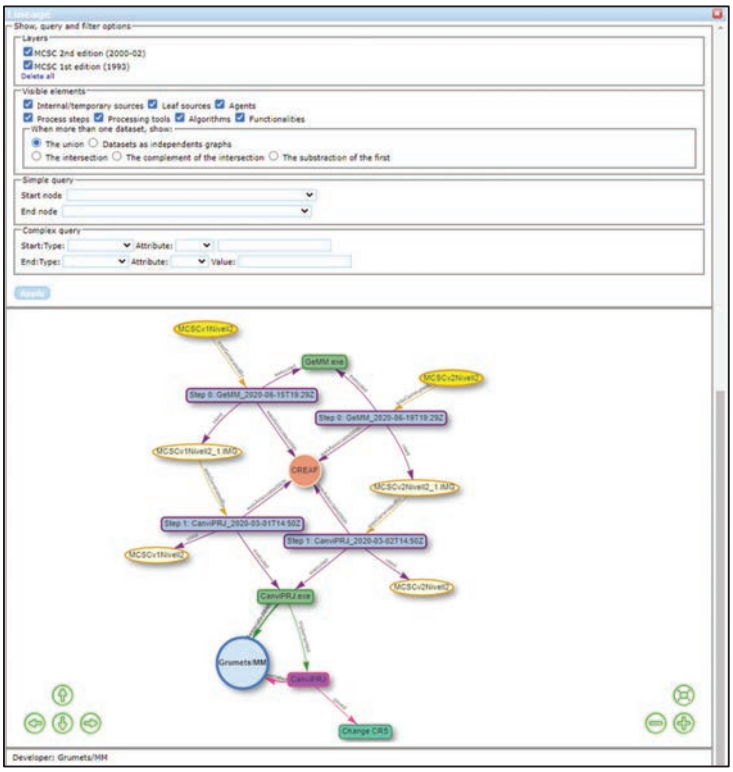

Figure 10: Visualization example *3*

5.2 Provenance Query examples

Table 4 shows forty-four possible queries that we can apply on the provenance of our land cover map use case. The presented queries are only examples of the potential of querying provenance from a collection of datasets. In the dual entry table, we present queries that relate two lineage elements. Only the first row restricts queries to the ones asking only about one lineage element.

Table 4: Forty-four examples of queries that can be formulated on the Catalonia Land Cover Maps

	Process run	**Processing tool**	**Algorithm**	**Functionality**	**Agent**	**Source**	**Time**	**Output**
(over a complete dataset)	Q1. Was any execution not using MiraMon software?	Q2. Was version 3 of Class*Knn* used?	Q3. Was the *Knn* algorithm used?	Q4. Was *supervised classification* functionality used?	Q5. Which roles had CREAF done?	Q6. Was a dataset called Orto25m used?	Q7. Was something executed in before 2013?	Q8. Was any output not owned by CREAF?
Process run (over MUCSC 2017 generation)	Q9. What was executed after version 5 of ClassKnn?	Q10. Did the process Step 5 use the version 3 of Class*Knn?*	Q11. Was the Process step 5 a *Knn classification*?	Q12. What is the functionality provided by Process step 5?	Q13. Which process steps where executed by Grumets?	Q14. Did the Process step 5 use a DEM of 2m?	Q15. Which was the last process step?	Q16. Which outputs where generated with Step 5 execution?
Processing tool (over MUCSC series generation)		Q17. Which tool was more often used right after *ClassKnn* tool*?*	Q18. What algorithm implemented version 3 of *ClassKnn*?	Q19. Provided the tools *ClassKnn* and *IsoMM* equivalent functionalities ?	Q20. Were the used tools developed by trusted software vendors?	Q21. Did the version 3 of *ClasskNN tool* need PIA[3] sources?	Q22. Was the version 3 of *ClassKnn tool* the last version available?	Q23. Which versions of MUCSC used version 2 of Class*Knn tool*?

[3] Pseudoinvariant areas

Algorithm (over a complete dataset)			Q24. Which were the different versions of *Knn* algorithm used?	Q25. Did all the *Sup.classification* tools used the same algorithm?	Q26. Was CREAF the author and owner of any of the used algorithms?	Q27. Was the reclassification algorithm suitable to work with categorical data?	Q28. When was the current version of the *Sup.classification* developed?	Q29. Which versions of MUCSC were created using Knn algorithm?
Functionality (over a complete dataset)				Q30. Were all radiometric corrections done with the same software?	Q31. Which institution performed the radiometric corrections?	Q32. Which of the used datasets were reclassified?	Q33. Was some dataset classified before 2015?	Q34. Which Land Cover Maps were photointerpreted?
Agent (over a complete dataset)					Q35. Who used tools developed by CREAF?	Q36. Which of the sources used in this collection have open access licences?	Q37. When did CREAF made his first execution in this collection?	Q38. Did CREAF create a MUCSC 2007dataset?
Source (over 2012 and 2017 MUCSC series generation)						Q39. Which sources were used in all workflows?	Q40. Had the orthos used the same temporal interval as the MUCSC outputs?	Q41 Which orthos were used in the generation MUCSC of 2012 and 2017?
Time (over 2017 MUCSC generation)							Q42. How long took the complete workflow to be completed?	Q43. When was the MUCSC map finalized?
Output (over 20017 MUCSC generation)								Q44. Was this LULC a revision of another MUCSC map?

From the forty-four examples we are showing how the results look like in two examples:

- **Query example 1** (see Figure 11): Over the complete MCSC dataset series, which processes steps were executed by Grumets research group? (Q13 in Table 4)

  The provenance graph shows all the processes steps involved in the generation of any of the MCSC version executed by Grumets and related to the generated dataset. The isolated datasets (1987, 1992, 1997 and 2002) mean that there are no processes steps executed by Grumets its lineage.

  The *layers* panel contains all MCSC versions; in the *visible elements* panel only agents and process steps are selected, and finally a complex query is filled in order to obtain only the end elements. The *attributes* panel shows the attributes of the selected element (in this case, *processor: Grumets research group*).

- **Query example 2** (see Figure 12): Over the complete MCSC dataset series, which versions have been generated using kNN (Classification by number of nearest neighbours) algorithms? (Q29 in Table 4)

  The provenance graph shows the kNN algorithm, including the functionality, related to the different versions of tools that implemented the kNN algorithm: ClassKnn_v2.exe and ClassKnn_v3.exe. These tools are related to the generated datasets. The isolated datasets (1987, 1992, 1997 and 2002) mean that the kNN algorithm have not been used in its lineage.

  The *layers* panel contains all MCSC versions; in the *visible elements* panel functionalities, algorithms and processing tools are selected and finally a *complex query* is filled in order to obtain only the target elements (those one related to knn algorithm). The attributes panel shows the attributes of the selected element (in this case, *ClaskNN_v2.exe*).

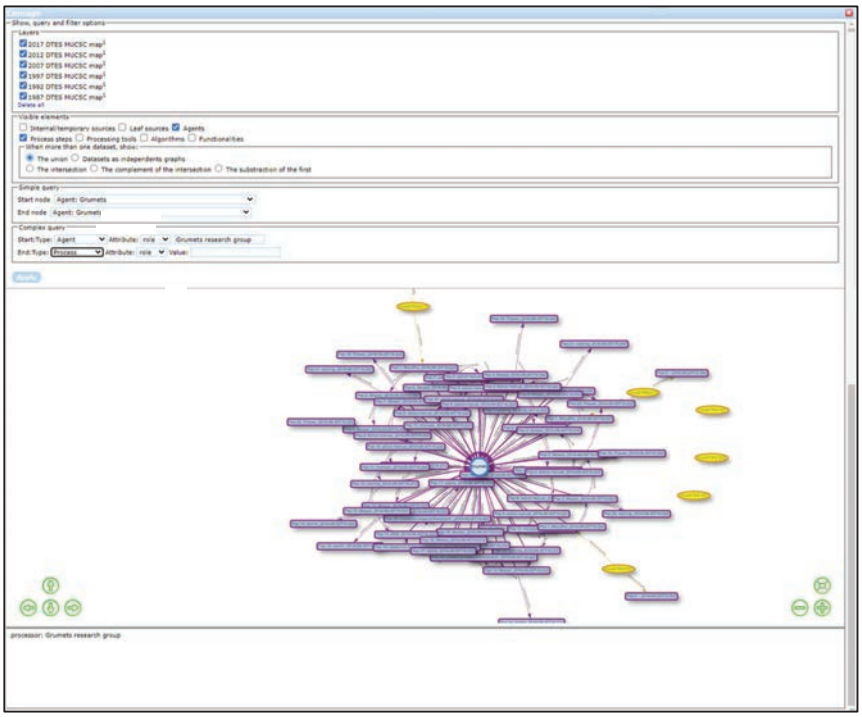

Figure 11: Query example 1

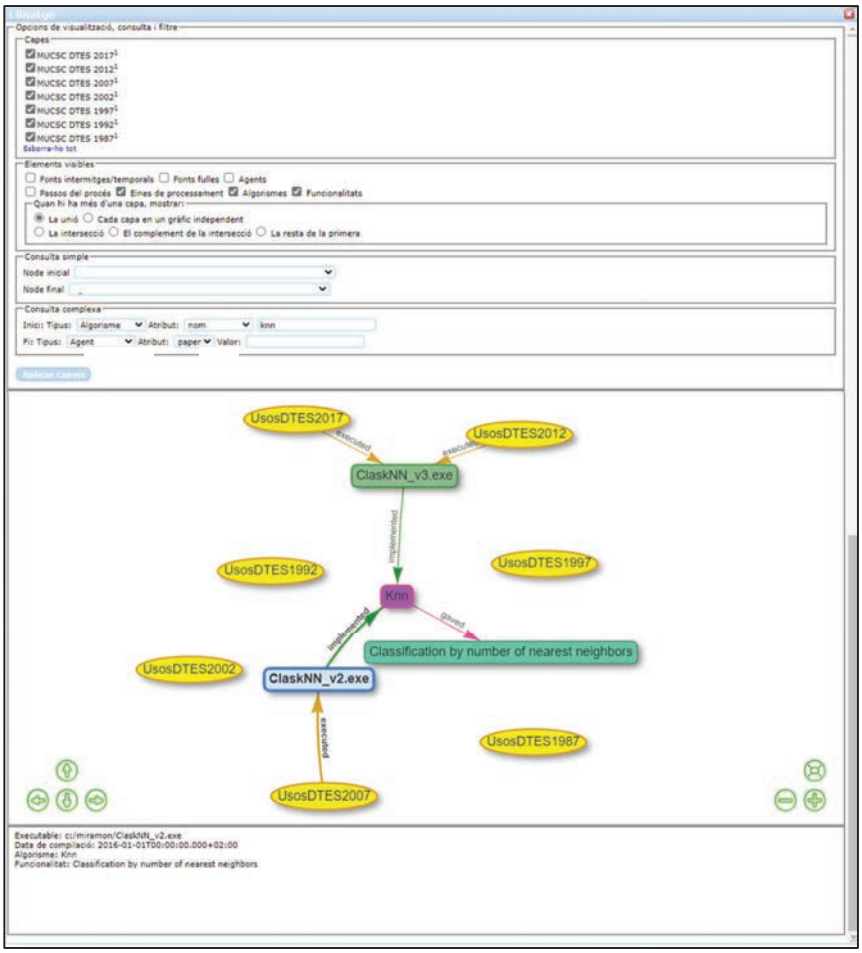

Figure 12: Query example

5.3 Discusion

In a network graph, dependencies and relationships between elements are represented in a more natural way than in a hierarchical tree form (see Figure 10). Even though it can be more difficult to follow than in a linear way, particularly in complex scenarios, it allows for presenting collections of datasets in as single view. Although a provenance graph with full detail is more informative, filtering the graph results in an easier to interpret (Figure 9 and 12) diagram. In addition, the possibility to formulate queries over provenance helps to take advantage of it. Answers to provenance questions allow users to inspect directly dataset generation history and, depending on which aspect of provenance is queried, different benefits can emerge[4]:

- Informative and transparency. They provide a better understand and compare methodologies in Q1, Q3, Q4, Q10, Q11, Q12, Q15, Q16, Q18, Q21, Q24 and Q25, Q27, Q28. Q32, Q33, Q39, Q42 and Q43
- Trust and authority: agents and its responsibility can be inferred based on the sources and tools used in Q2, Q5, Q13, Q20, Q31, Q37 and Q38
- Data quality can be deduced from the quality of the sources and precisions of the processing tools used in Q6, Q7, Q10, Q14, Q19, Q21 Q22, Q25, Q30, Q34, Q40 and Q44.
- Documentation and reproducibility can be achieved if all necessary details about the actual dataset, metadata or tools are present such as in Q9, Q16, Q17, Q23, Q24 Q29 and Q41.
- License and accessibility: Informs about the needed resources that were accessed and licences needed such as in Q8, Q26, Q35 and Q36

6. FUTURE WORK

In the conceptual side, the levels of abstraction introduced for processes allow for transforming a precise lineage graph in more abstract workflow diagram or even in a list of functionalities that inform a GIS operator or student on how to create datasets by chaining GIS tools. By doing that, we realize that it could be also possible to abstract the sources into generic ones by pointing to the schema of the product or in the extreme, by providing only the topic category they belong. This information can be extracted from source metadata by looking at the ISO 19115 metadata fields not directly related to lineage.

This paper is presenting a sketch of a provenance window that has been co-designed in collaboration with the MiraMon maps browser implementers. The development was completed only the first loops of an agile methodology that has been prototyped and only been tested with the data presented in this paper. Further loops into the development can reveal the need for extra functionalities or the need to repurpose the user interface of the provenance query and filter panel.

Extending the queries presented in this paper to bigger collections is a challenge for the visualization but combined with the right queries could be applied to an organization to determine the most useful datasets and tools. This will be a new benefit for the organizations they are facing, for the first time: the challenge to have to decide which datasets should be preserved from the too big organizational digital legacy, and what information can be forgotten and erase from the archives. A comprehensive provenance study can help to this purpose.

[4] Some queries could provide more than one benefit. The most relevant benefit is presented in this classification

7. CONCLUSIONS

The geospatial lineage is a necessary component in the metadata of spatial information distributed over the web. However, it is recognized that these benefits cannot be materialized if there are no proper tools to help users in visualizing and interpreting it. This paper has made two main contributions to overcome this situation,

On one hand, the introduction of four levels of abstraction of the process step description (process run, processing tool, algorithm and functionality) has proven as a valuable way to better describe lineage. The inclusion of functionalities and algorithm descriptions as a part of lineage provides a high-level information and representation that is independent from the software used or the moment in time the step was executed. This solution has provided benefits: allows for interconnecting datasets coming different workflows and opens the possibility to compare workflows at the methodology level. In this sense, the combination of the lineage model of ISO19115 family (to express the object types of the geospatial objects involved the production processes) with W3C PROV (to convey the relation naming conventions) was demonstrated. In this paper, a provenance symbolization as a provenance graph instead of hierarchical tree was explored as more flexible alternative.

The web tool presented in this paper helps users to interpret lineage by making connections among processes and source more visible and allowing for filtering and querying lineage elements. The tool facilitates the formulation of queries to interrogate the origins of geospatial data of a collection of datasets. The tool generates on-demand visualizations that provides answer to queries that emphasize the benefits of lineage information: Informative and transparency; trust and authority, data quality; documentation and reproducibility; and license and accessibility.

The possibility to formulate queries over a collection of datasets will give an added value to provenance and will provide scientists and technicians the opportunity to inspect dataset interrelations and workflow performance. Provenance graph could help on the difficult task of determining the most useful datasets and eventually deciding what information should be preserved for future generations.

ACKNOWLEDGMENT

This work was supported by the Catalan Government [SGR2017 1690]. This work was done under ECOPOTENTIAL, e-shape, and ERA-PLANET projects. These projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641762-2, 776740 and 689443 respectively This work has been also supported by the Spanish Ministry of Economy and Competitiveness through the NEWFORLAND project (RTI2018-099397-B-C21 (MINECO/FEDER)). Xavier Pons is the recipient of an ICREA Academia Excellence in Research Grant (2016–2020).

REFERENCES

1. Albrecht, J. (1998). Universal analytical GIS operations: A task-oriented systematization of data structure-independent GIS functionality. *Geographic information research: Transatlantic perspectives*, 577-591
2. Amann, B., & Scholl, M. (1993). Gram: a graph data model and query languages. In *Proceedings of the ACM conference on Hypertext* (pp. 201-211).
3. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452.
4. Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., ... & Stodden, V. (2019). Computing environments for reproducibility: Capturing the "Whole Tale". *Future Generation Computer Systems*, *94*, 854-867.

5. CEOS Interoperability Terminology, Version 1.0. (2020) CEOS – WGISS Interoperability and Use Interest Group.
6. 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 and Urban Systems, *64*, 103-117, M.
7. Closa G, Masó J, Zabala A, Pesquer L, Pons X. (2019*). A provenance metadata model integrating ISO geospatial lineage and the OGC WPS: Conceptual model and implementation*. Transactions in GIS 23(4): 683-705. DOI: 10.1111/tgis.12555.
8. Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., ... & Mareuil, F. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284-298.
9. 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 and remote sensing*, *51*(11), 5082-5089.
10. Di, L., Yue, P., Ramapriyan, H. K., & King, R. L. (2013). Geoscience data provenance: An overview. *IEEE transactions on geoscience and remote sensing*, *51*(11), 5065-5072.
11. Erwig, M., & Schneider, M. (1999). Developments in spatio-temporal query languages. In *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99* (pp. 441-449). IEEE.
12. ESRI. (2020). ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.
13. Fisher, P. F. (2006). Algorithm and Implementation Uncertainty: Any Advances? *Classics from IJGIS: Twenty years of the International Journal of Geographical Information Science and Systems*, 225-228.
14. Giuliani, G., Chatenoux, B., Bono, A. D., Rodila, D., Richard, J. P., Allenbach, K., & Peduzzi, P. (2017). Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). Big Earth Data, 0 (0), 1–18.
15. González-Guerrero, Ò., i Fernández, X. P., Bassols-Morey, R., & Fernández, F. X. C. (2019). Dinàmica de les superfícies de conreu a Catalunya mitjançant teledetecció en el període 1987-2012. *Quaderns agraris*, 59-91.
16. Goodchild, M. F. (1991). Geographic information systems. *Progress in Human geography*, *15*(2), 194-200.
17. GRASS Development Team. (2017). Geographic Resources Analysis Support System (GRASS) Software, Version 7.2. Open Source Geospatial Foundation. Electronic document: http://grass.osgeo.org
18. Groth, P., and Moreau, L. (2013). PROV-Overview: An Overview of the PROV Family of Documents. Working group note, W3C. 2013
19. 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 and Remote Sensing*, *8*(2), 926-936
20. Ibàñez, J.J., and Burriel, J.A. (2010). Mapa de cubiertas del suelo de Cataluña: características de la tercera edición y relación con SIOSE. En: Ojeda, J., Pita, M.F. y Vallejo, I. (Eds.), Tecnologías de la Información Geográfica: La Información Geográfica al servicio de los ciudadanos. Secretariado de Publicaciones de la Universidad de Sevilla. Sevilla. Pp. 179-198. ISBN: 978-84-472-1294-1
21. IBM Big Data & Analitics Hub. (2020). The Four V's of Big Data. Retrieved from https://www.ibmbigdatahub.com/infographic/four-vs-big-data
22. ISO 19115."*Geographic Information- Metadata*". 2003.
23. ISO 19115-1."*Geographic Information- Metadata- Part 1: Fundamentals*". 2014.
24. ISO 19115-2. "*Geographic Information- Metadata- Part 2: Extensions for acquisition and processing*. 2019.

25. Ivánová, I., Armstrong, K., & McMeekin, D. (2017). Provenance in the next-generation spatial knowledge infrastructure. In *22nd International Congress on Modelling and simulation (MODSIM 2017)* (pp. 410-416).
26. 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.
27. Koubarakis, M., Karpathiotakis, M., Kyzirakos, K., Nikolaou, C., & Sioutis, M. (2012). Data models and query languages for linked geospatial data. In *Reasoning Web International Summer School* (pp. 290-328). Springer, Berlin, Heidelberg.
28. Kuhn, W., & Ballatore, A. (2015). Designing a language for spatial computing. In *AGILE 2015* (pp. 309-326). Springer
29. Lanter, D. P. (1991). Design of a lineage-based meta-data base for GIS. *Cartography and Geographic Information Systems*, *18*(4), 255-261
30. Lemos, M. C., Kirchhoff, C. J., & Ramprasad, V. (2012). Narrowing the climate information usability gap. *Nature climate change*, *2*(11), 789.
31. Lopez-Pellicer, F. J., & Barrera, J. (2014). D16. 1 Call 2: Linked map VGI provenance schema. In *In Linked Map subproject of planet data. Seventh framewok programe*.
32. Lopez-Pellicer, F. J., Lacasta, J., Espejo, B. A., Barrera, J., & Agudo, J. M. (2015). The standards bodies soup recipe: an experience of interoperability among ISO-OGC-W3C-IETF standards. In *INSPIRE-GEOSPATIAL WORLD FORUM. Proceedings... Lisboa, Portugal*.
33. Lutz, M., Riedemann, C., & Probst, F. (2003). A classification framework for approaches to achieving semantic interoperability between GI web services. In *International Conference on Spatial Information Theory* (pp. 186-203). Springer, Berlin, Heidelberg.
34. Masó, J., X. Pons; A. Zabala. (2012). Building the World Wide Hypermap (WWH) with a RESTful architecture. International Journal of Digital Earth. Vol.7(3), pp.175-193. ISSN: 1753-8947. DOI: 10.1080/17538947.2012.669414
35. Masó J., Zabala A., Pons X. (2020) Protected Areas from Space Map Browser with Fast Visualization and Analytical Operations on the Fly. Characterizing Statistical Uncertainties and Balancing Them with Visual Perception . ISPRS International Journal of Geo-Information. Vol.9 (5), pp.30-Article Id: 300. DOI: 10.3390/ijgi9050300. In Internet: https://www.mdpi.com/2220-9964/9/5/300/pdf
36. Markus Konkol & Christian Kray. (2019). In-depth examination of spatiotemporal figures in open reproducible research, Cartography and Geographic Information Science, 46:5, 412-427, DOI: 10.1080/15230406.2018.1512421
37. Müller, M. (2015). Hierarchical profiling of geoprocessing services. *Computers & Geosciences*, *82*, 68-77.
38. Pons, X. (2020). MiraMon: Geographical information system and remote sensing software. Barcelona, Spain: Centre de Recerca Ecològica i Aplicacions Forestals.
39. Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, *264* (5164), 1421-1426.
40. SNAP – ESA. (2020). Sentinel Application Platform v8.0.0, http://step.esa.int
41. Spatial Data Transfer Standard (SDTS). (1998). American National Standards Institute's (ANSI). ANSI/NCITS320.1998
42. Spiekermann, R., Jolly, B., Herzig, A., Burleigh, T., & Medyckyj-Scott, D. (2019). Implementations of fine-grained automated data provenance to support transparent environmental modelling. *Environmental Modelling & Software*.

43. Sun, Z., Yue, P., & Di, L. (2012). GeoPWTManager: a task-oriented web geoprocessing system. *Computers & Geosciences*, *47*, 34-45
44. Yazici, I. M., Karabulut, E., & Aktas, M. S. (2018). A Data Provenance Visualization Approach. In *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)* (pp. 84-91). IEEE.
45. Yue, C., Baumann, P., Bugbee, P., & Jiang, L. (2015). Towards intelligent giservices. *Earth Science Informatics*, *8*(3), 463-481.
46. Yue, P., Zhang, M., Guo, X., & Tan, Z. (2014). Granularity of geospatial data provenance. In *2014 IEEE Geoscience and Remote Sensing Symposium* (pp. 4492-4495). IEEE.
47. Yue, P., Wei, Y., Di, L., He, L., Gong, J., & Zhang, L. (2011). Sharing geospatial provenance in a service-oriented environment. *Computers, Environment and Urban Systems*, *35*(4), 333-343.
48. 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.

7. Resum de resultats

El resum de resultats d'aquesta tesi es realitza presentant de manera unificada[9] les principals fites assolides al llarg dels capítols 2, 3, 4, 5 i 6. En primer lloc, es presenten els resultats relacionats amb els *Aspectes derivats de les propostes en els models de representació (*subapartat *7.1).* A continuació, es presenten els resultats relacionats amb els *Aspectes derivats de les propostes en la captura i visualització (*subapartat *7.2)*. Finalment, es presenten aquells resultats relacionats amb els *Aspectes derivats de les propostes de consultes sobre el llinatge (*subapartat *7.3).*

7.1 Aspectes derivats de les propostes en els models de representació

En aquesta secció es detallen les diverses propostes realitzades sobre els aspectes relacionats amb la representació del llinatge, l'estudi dels models disponibles i les seves codificacions.

7.1.1 PROV i RDF per descriure llinatge a nivell d'atribut, element geospacial i conjunt de dades

En determinats processos geospacials és fonamental capturar el llinatge a nivell d'atribut, d'element geospacial i de conjunt de dades. En són un exemple els processos recurrents de fusió de dades, on els conjunts de dades resultants poden contenir objectes amb diversos orígens. Per aquest motiu, els models de llinatge han de permetre representar els nivells més baixos de granularitat (nivell d'element i/o atribut), a la vegada que compartir les característiques comunes dels nivells més alts (nivell de capa) i especificar les característiques concretes només quan sigui necessari. Així es redueix la redundància i s'aconsegueix una notació més compacte.

A tal efecte, es proposa l'ús de l'estàndard PROV per capturar els diferents nivell de granularitat en base a les següents evidències:

- PROV és un model de dades orientat a objectes, basat en la declaració d'objectes corresponents a aspectes del món real agrupats en classes.
- Amb PROV es pot documentar informació de llinatge en notació RDF. Aquesta notació s'adapta millor que l'XML de la ISO per a descriure models de dades orientats a objectes i intercanviar el llinatge en entorns distribuïts.
- La definició àmplia de les classes PROV, com ara entitats i activitats, inclou implícitament els diferents nivells de granularitat.
- PROV requereix menys espai d'emmagatzematge que una alternativa que usi la combinació d'ISO i GML.

La proposta es basa en la definició d'atributs, elements geospacials i conjunt de dades com a entitats (en PROV qualsevol cosa a descriure és una entitat) i la introduccció de *hadGeometry* i *hadProperty* per relacionar els elements geospacials amb els atributs i les seves geometries (*Figura 9*). La relació *hadMember* (relació definida pel model PROV) vincula els conjunts de dades amb els elements geospacials.

[9] Per tal de donar una visió unitària, s'ha preferit no mencionar en quin capítol s'arriba a cada resultat.

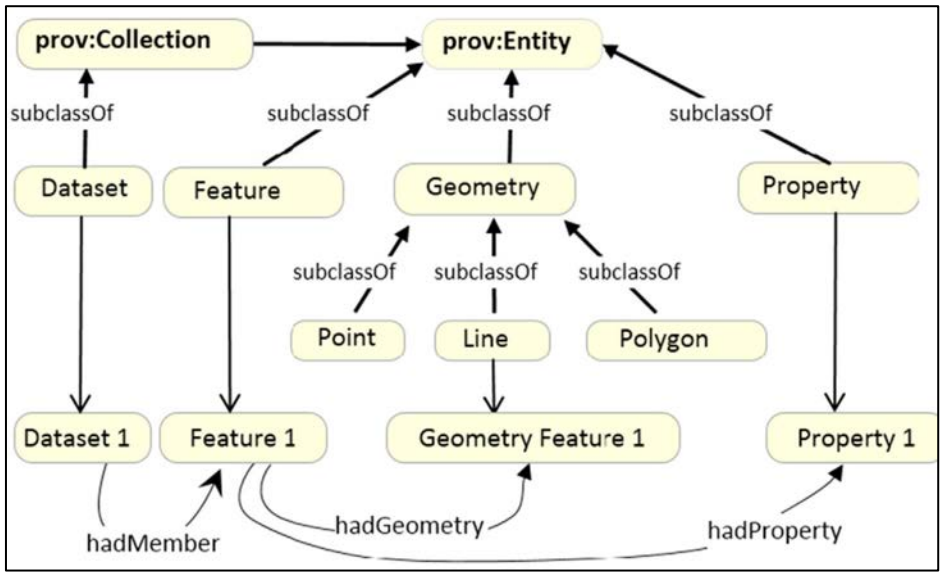

Figura 9: Diagrama que representa la relació dels nivells de conjunt de dades, element geospacial i atribut amb PROV (Font: Figura 4 capítol 2)

Per altra banda, l'ús de l'RDF ens apropa al concepte de dades enllaçades (*Linked Data*), permetent relacionar-les amb altres dades, de manera que es fan més útils mitjançant consultes semàntiques. La Taula 11 mostra les declaracions dels diferents nivells d'entitats del diagrama de la Figura 9.

Taula 11: Declaració dels diferents nivells d'entitats i les seves relacions amb RDF en codificació N3 (Font: Taula 1 capítol 2)

```
@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.
```

7.1.2 Combinació del WPS i el model ISO per descriure el llinatge

La constatació que el model de llinatge ISO té certes limitacions ens ha portat ha experimentar amb la combinació d'aquest model amb el model WPS de l'OGC per tal de descriure el llinatge més completament. Tal i com es pot observar a la Figura 10, presentem un model de llinatge on el LI_Lineage (ISO) és ampliat amb l'estàndard WPS. En concret, aprofundeix en la descripció de les

fonts i les sortides. D'aquesta manera, el llinatge és pot llegir com una seqüència de processos descrits amb el model ISO, que usen una sèrie de paràmetres (inputs) descrits amb el WPS, alguns dels quals poden ser fonts.

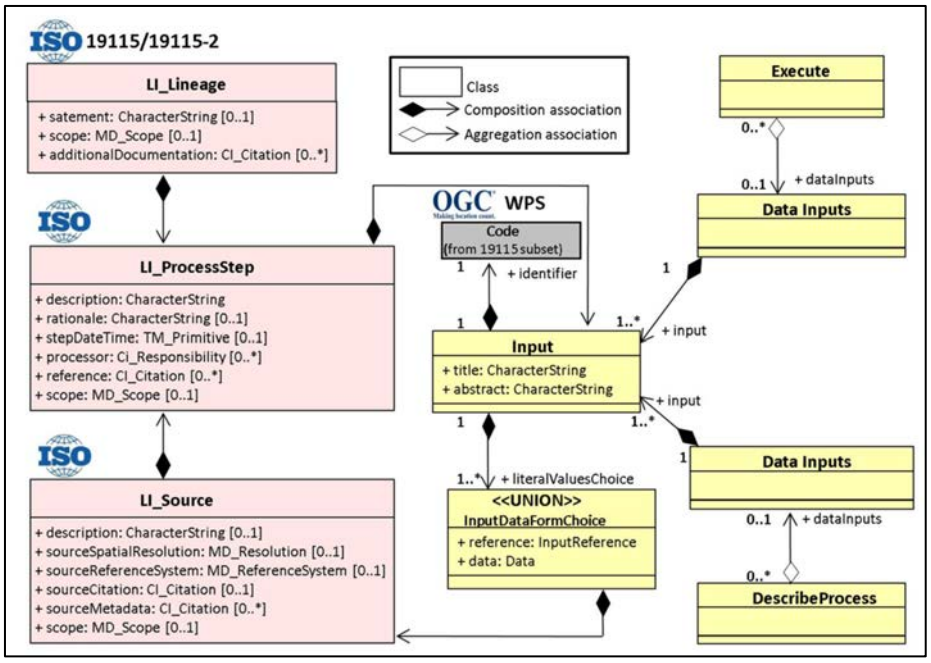

Figura 10: L'ISO 19115 LI_Lineage descriu la provenance com una seqüència de LI_ProcessStep que usa LI_Source. La informació continguda a LI_Source és ampliada amb l'ús d'alguns elements del WPS (diagrama de classes UML) (Font: Figura 2, capítol 4)

La incorporació del WPS ens ha permès presentar els resultats següents:

- Establir l'ordre de les fonts i la direcció

L'etiqueta opcional *ows:Metadata* del *WPS describeProcess response* ens permet associar cada paràmetre i/o font amb un paràmetre de la sintaxi i establir l'ordre (p. ex. *ows:Metadata xlink:title="Param01*" pel paràmetre 1 d'una línia de comanda) .

Per definir si una font és una entrada o una sortida, utilitzem les etiquetes de la resposta a un *WPS describeProcess* actuals: *DataInputs**Input* i *ProcessOutputs**Output* respectivament. Per les fonts o paràmetres que es converteixen en una sortida (entrada/sortida) després de l'execució, hem utilitzat l'etiqueta per definir l'ordre: quan una font (*Input**ows:Metadata xlink:title="ParamIdentifierX"*) es converteix en una sortida, es torna a escriure com una sortida (*Output**ows:Metadata xlink:title="ParamIdentifierX"*) utilitzant el mateix *xlink:title*.

- Descripció dels valors literals

El model ISO tenia una mancança en quan a la descripció dels paràmetres de tipus literal (numèric o alfanumèric), doncs no preveia la captura del valor individualitzat de cada paràmetre. En el model proposat, els valors literals de paràmetres es capturen fent servir el document de resposta d'un *WPS describeProcess*. Concretament, *DataInputs**Inputs**LiteralData* en el cas d'entrades de dades; \ProcessOutputs\Outputs\LiteralData en el cas de les sortides.

Aquesta proposta es va introduir com a sol·licitud de canvi de l'ítem de treball ISO 19115-2 i es va treballar en les reunions del TC211 amb els editors per ampliar l'estàndard en aquesta direcció. Les noves revisions de l'ISO 19115-2 (*Figura 11*) ja incorporen aquesta sol·licitud.

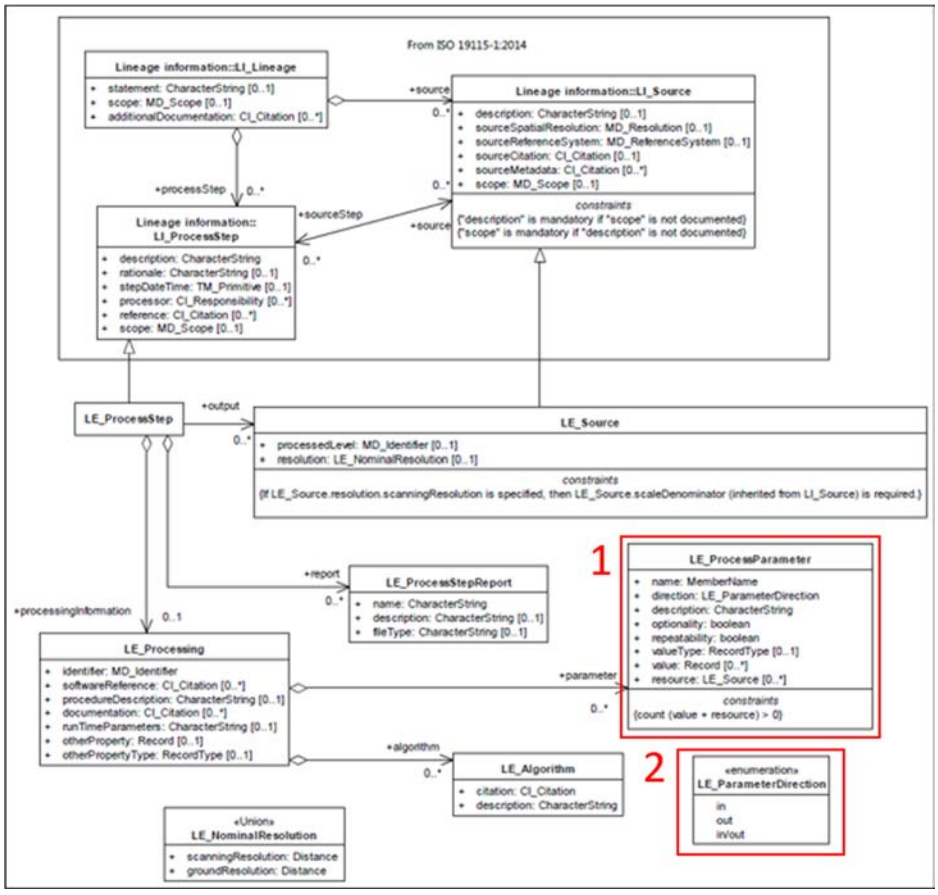

Figura 11: Diagrama UML del model de llinatge de l'ISO 19115-1 i l'ISO 19115-2. Els rectangles 1 i 2 formen part dels canvis introduïts a la revisió de l'ISO. El rectangle vermell 1 mostra la nova classe LE_ProccessParametre i els seus atributs. El rectangle vermell 2 mostra com es codifica si el paràmetre és entrada o sortida (Font: Elaboració pròpia sobre il·lustració extreta de la ISO 19115-2:2019)

- **Captura dels experiments científics, iteracions prèvies o execucions descartades**

Era necessària una proposta per capturar les iteracions prèvies o les execucions descartades com a part de les anotacions dels experiment científics. Amb aquest objectiu, proposem l'ús del marcador *otherProperty* de *LE_Processing* del model ISO per documentar si el resultat obtingut és el resultat previst o definitiu. En aquest sentit vam assignar el *otherProperty* com *recordtype* amb un camp anomenat *"iteration"* que pot tenir dos valors: "*iteration=discarded" o "iteration= satisfactory".*

7.1.3 Abstracció dels diferents nivells de processament

La definició de diferents nivells d'abstracció de processament: execució, eina de processament, algorisme i funcionalitat(*Figura 12*) s'introdueix amb un objectiu doble:

- Enriquir semànticament les descripcions dels processos.
- Relacionar les descripcions de llinatge de conjunts de dades per tal de possibilitar la seva representació gràfica conjunta, comparació i generació de consultes sobre llinatges de diversos conjunts de dades.

- Abstracció dels diferents nivells de processament

S'han ampliat els diferents nivells d'abstracció que descriuen un procés:

- Execució (*process run o process step*): Execució individual d'una eina de geoprocessament. És una operació SIG individualitzada.
- Eina de processament (executable): Versió concreta d'una implementació d'un algorisme en un programari que, òbviament, es pot executar diverses vegades amb fonts diferents.
- Algorisme (model): conjunt de passos matemàtics i lògics que permeten transformar algunes entrades en sortides. Es pot implementar en programari de diferents maneres i amb diferents llenguatges de programació. Es sol descriure en un document científic o tècnic.
- Funcionalitat (operació): Nom de l'operació que transforma dades en altres dades o informació amb una orientació clara a la resolució de problemes espacials. És un dels elements del llistat de capacitats que descriu un llibre de text de SIG i/o de teledetecció.

Aquests nivells d'abstracció es relacionen de la manera *següent Figura 12):*

L'execució executa (*Executed)* una eina de processament. L'eina de processament implementa (*Implemented*) un algorisme. Finalment, l'algorisme proporciona (*Gave*) una funcionalitat[10]

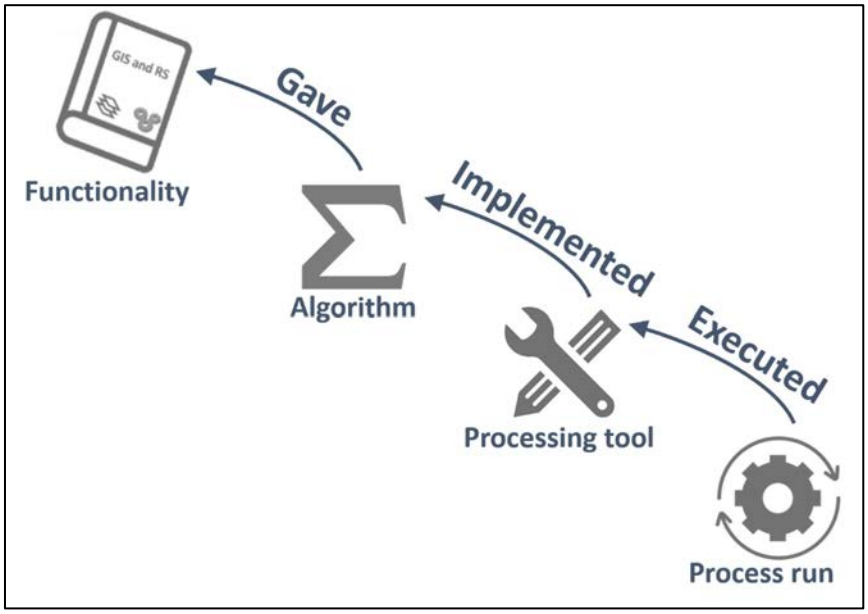

Figura 12. Nivell d'abstracció de les eines de processament (Font: Figura 1, Capítol 6).

La inclusió de funcionalitats i descripcions d'algorismes com a part de la *provenance* proporciona una informació d'alt nivell que és independent del programari utilitzat. Això permet que un diagrama de *provenance* que inicialment està documentant eines de processament concretes (per exemple part del propi programari; com fa el MiraMon), es pugui abstraure a un nivell superior per descriure l'objectiu o la metodologia general de la cadena de processament.

[10] Per seguir amb les regles proposades per W3C PROV aquestes relacions entre objectes s'expressen en passat.

- Combinació del model PROV i el model ISO per representar els diferents nivells d'abstracció

Per codificar el llinatge d'un catàleg complet de dades, hem escollit una solució composta basada en la combinació del model PROV i el model ISO (Figura 13):

- Les convencions de PROV s'han utilitzat per representar relacions entre agents (CI_Responsability), accions (LE_ProcessingSteps) i entitats (qualsevol altra cosa).
- Els elements de llinatge s'han representat amb els noms de classe de l'ISO 19115-2 (LI_Lineage). Aquests proveeixen noms més concrets per referir-se al que PROV anomenaria entitats, agents i accions.

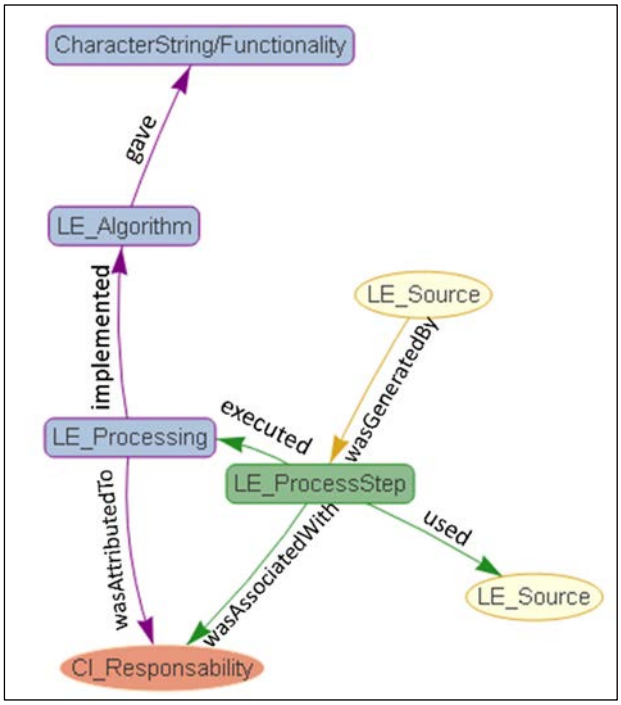

Figura 13 Aquest gràfic mostra com s'usa PROV per descriure les relacions, mentre que l'ISO està restringida a la descripció dels elements (Font: Figura 2, Capítol 6).

Presentant els quatre nivells d'abstracció amb PROV, podem descriure l'origen de diversos conjunts de dades en una sola visualització que mostra les relacions entre elles (Figura 14). S'ha proposat l'ús d'identificadors globals per les fonts i les eines de processament (ISO MD_Identifier), per a co-escriure objectes repetits i integrar els objectes restants en un únic graf de *provenance* connectat per relacions PROV.

En integrar diversos conjunts de dades de llinatge en una representació, PROV obre les portes a un nou conjunt de possibilitats en termes de representació i consulta sobre les dades d'una col·lecció de conjunts de dades o catàlegs de metadades sencers.

El subapartat 7.2.2 descriu els resultats obtinguts en la representació conjunta del llinatge de conjunts de dades i el subapartat 7.3.3 presenta els resultats per l'eina de generació de consultes en entorns distribuïts.

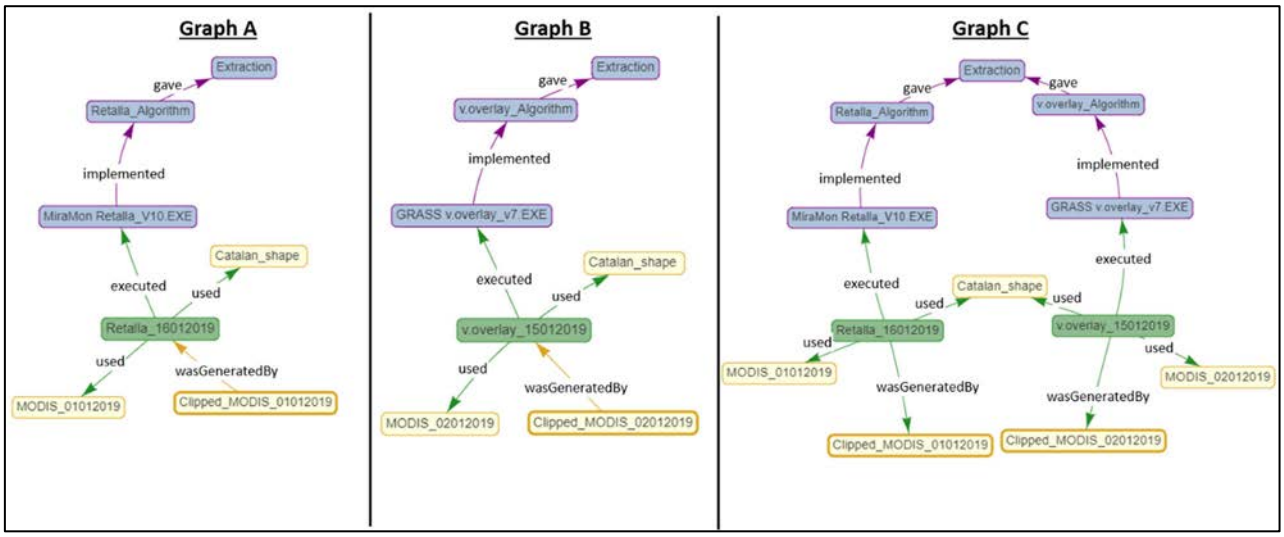

Figura 14: Els gràfics A i B representen execucions individuals amb la mateixa funcionalitat realitzada amb dos programaris diferents: MiraMon i GRASS, respectivament. El gràfic C representa la provenance d'ambdues execucions fusionades en una sola instància(Font: Figura 3, Capítol 6).

7.2 Aspectes derivats de les propostes en la captura i visualització

Aquesta secció detalla els resultats més destacats dels aspectes relacionats amb la captura i la visualització del llinatge.

7.2.1 Provenance Engine: eina integral de gestió del llinatge

El *Provenance Engine* (PE) és l'eina integral de gestió del llinatge desenvolupada en el marc del GeMM i el MiraMon per capturar, editar, visualitzar i exportar el llinatge geospacial.

- Captura automàtica del llinatge

El PE és una biblioteca de funcions, escrita en llenguatge C, compartida entre la interfície visual del GeMM i les aplicacions de SIG i teledetecció del MiraMon. Així, cada aplicació utilitza aquestes funcions per llegir les metadades dels conjunts de dades font, i integrar aquesta informació al llinatge del conjunt de dades resultant.

Per tal de millorar la gestió del llinatge del PE, per a cada aplicació del MiraMon es va escriure un document plantilla que fa servir l'esquema dels document de resposta del *describeProcess* de l'estàndard WPS. El PE utilitza la plantilla generada per capturar, durant l'execució de l'aplicació, tota aquella informació de llinatge comuna per totes les execucions (p. ex. la descripció de cada paràmetre de la sintaxi, el propòsit de l'aplicació, etc). Per altra banda, el PE empra els documents de resposta *execute* del WPS per emmagatzemar en el llinatge informació especifica de cada aplicació (p. ex. fitxers fonts, valors literals, sortides, etc)(Figura 15).

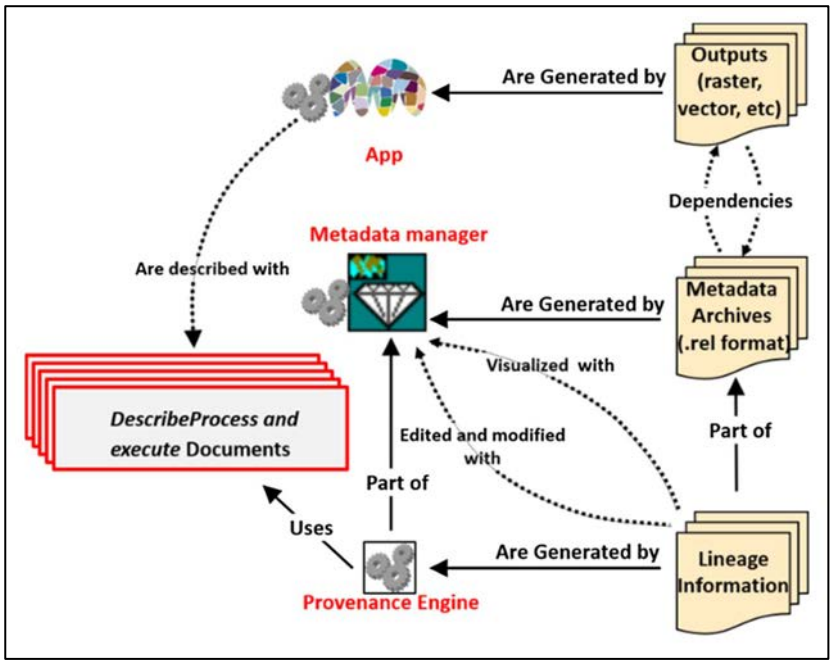

Figura 15: La PE utilitza documents resposta de WPS DescribeProcess per extreure informació i incorporar-la al llinatge. La interfície del GeMM (gestor de metadades) permet als usuaris editar i modificar la *provenance* de les dades geospacials generades per les aplicacions del MiraMon (Font: Figura 6, capítol 4).

- Edició del llinatge

Tal i com mostra la Figura 15 un cop la informació de llinatge està capturada, aquesta pot ser editada a posteriori amb el GeMM. La interfície gràfica del GeMM (Figura 16) permet editar la informació de llinatge afegint o suprimint processos o paràmetres en un flux de treball. A més, la descripció de l'algorisme, els passos de processament, les dates d'execució, la responsabilitat del producte i l'ordre dels processos es poden editar i adaptar a cada escenari. Això permet als productors de dades completar o ajustar la descripció de llinatge que es captura automàticament durant l'execució del flux de treball.

- Visualització del llinatge com un arbre

La interfície gràfica del GeMM (Figura 16) ajuda als usuaris d'IG a navegar i interpretar el llinatge. L'eina representa l'origen d'un conjunt de dades com a llista de processos. Cadascun d'aquests processos té una llista indentada de tots els paràmetres utilitzats i de totes les sortides generades. Al mateix temps, alguns dels paràmetres del flux de treball (les fonts de dades) es deriven de processos anteriors (procés pare), que són representats a un nivell més profund amb la seva pròpia llista indentada de paràmetres utilitzats, etc. Per tant, l'estructura de l'esquema de *provenance* augmenta progressivament en profunditat.

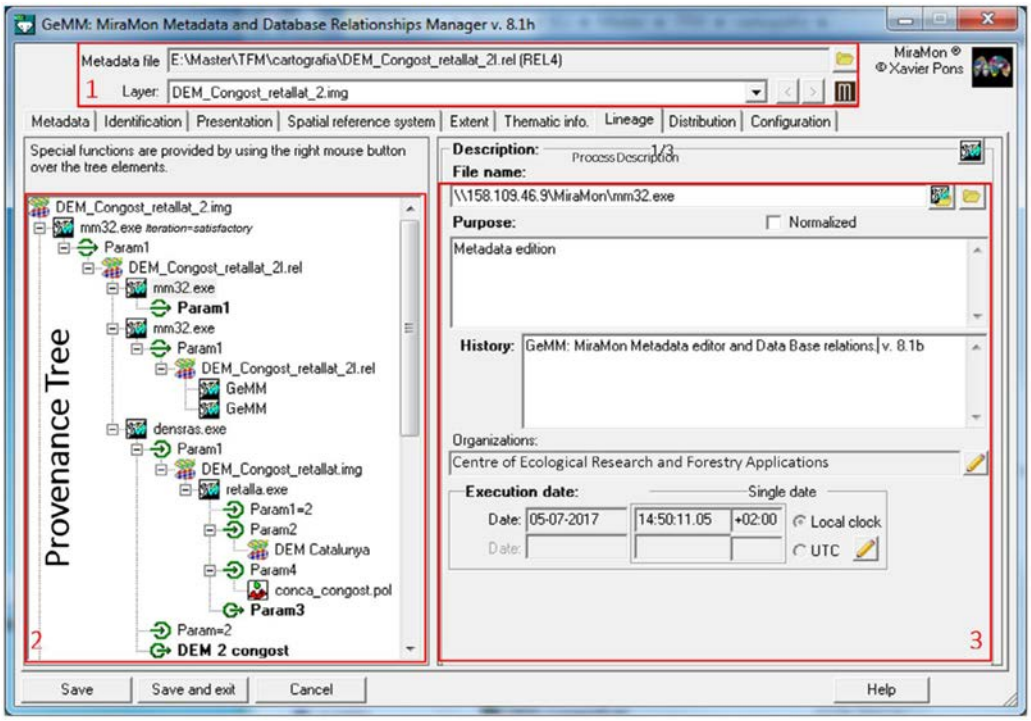

Figura 16: Interfície gràfica del GeMM: (1) ubicació de l'arxiu de metadades i nom del conjunt de dades; (2) caixetí amb l'arbre de llinatge que inclou tots els processos i les fonts utilitzades en la història de la creació del conjunt de dades; (3) caixetí per visualitzar o editar els atributs de cada font o procés: atribució, data d'execució, descripció del procés, descripció d'execució, etc (Font: Figura 6, capítol 4).

7.2.2 Visualització del llinatge com un graf

En un entorn distribuït, on es necessita mostrar el llinatge de més d'un conjunt de dades, la visualització en forma xarxa és preferible. Aquest tipus de visualització s'ha desenvolupat dins el navegador de mapes del MiraMon. Aquest permet proporcionar una visió integrada del llinatge de diversos conjunts de dades que tenen alguna cosa en comú (geogràfic o temàtic) i que poden ser produïts per més d'una organització.

En el sistema servidor-navegador del MiraMon el llinatge es recupera del servidor però la representació gràfica del llinatge es fa al costat client. S'ha utilitzat la biblioteca visjs (2020), que té una implementació per la generació de grafs, per a simbolitzar el llinatge.

S'ha creat un codi JavaScript que converteix la codificació JSON basada en la norma ISO 19115 a les matrius JSON requerides per la biblioteca de funcions del vis.js. El graf es defineix com un conjunt de nodes que tenen identificadors i com un conjunt de vores que es connecten als nodes que identifica. La Figura 17 mostra un exemple de representació del llinatge d'un sol conjunt de dades.

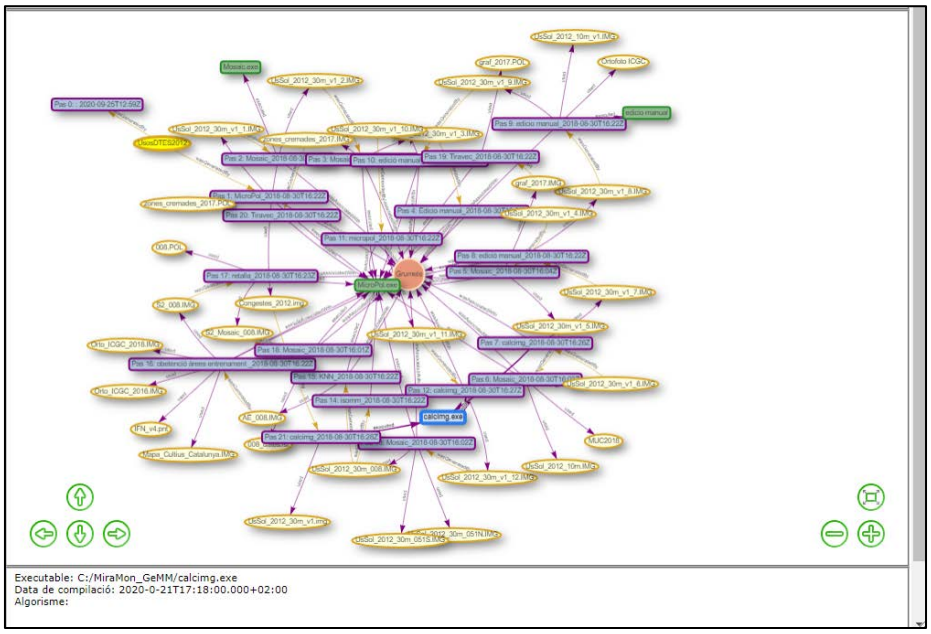

Figura 17: Vista del llinatge d'un conjunt de dades del navegador de mapes del MiraMon. Les fonts emprades estan representades amb el·lipses grogues. Les execucions amb rectangles porpra i les eines emprades amb rectangles de color verd. En aquest cas només hi ha un agent (cercle de color taronja). L'el·lipse de color groc brillant és el conjunt de dades resultant del llinatge descrit. *(*Font: Elaboració pròpia*).*

7.3 Aspectes derivats de les propostes de consultes sobre el llinatge

A continuació es detallen els resultats més destacats quant als aspectes relacionats amb les consultes i l'explotació del llinatge.

7.3.1 Demostració d'ús SPARQL per consultar el llinatge geospacial

Un cop la informació de llinatge és capturada i serialitzada amb la notació N3 d'RDF (tal i com s'apunta al subapartat 7.1.1), es pot emprar el llenguatge de consultes SPARQL per explotar les dades i seleccionar fragments específics del llinatge (Taula 12).

Taula 12: Consulta amb SPARQL per seleccionar elements generats d'una data específica (Font: Taula 13 capítol 2)

```
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 }
```

Les representacions gràfiques de la totalitat, o de fragments derivats de les consultes sobre el llinatge, ajuden a la comprensió dels productes geogràfics. El programari Gruff (http://franz.com/agraph/gruff/) es pot utilitzar per interpretar les tripletes i generar automàticament un graf. La Figura 18 mostra la representació de les tripletes derivades de la consulta de la Taula 12.

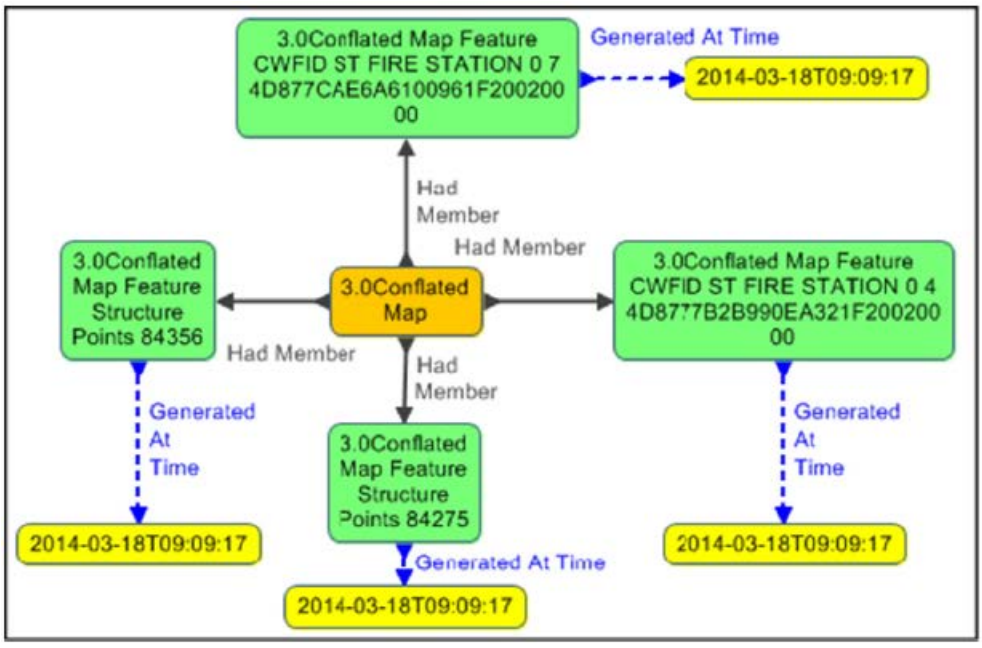

Figura 18: Representació gràfica dels resultats de la consulta de la Taula 12 (Font: Figura 11, capítol 2)

7.3.2 Consultes sobre el llinatge de catàlegs de dades en entorns distribuïts

S'ha constatat que és rar trobar un sistema de llinatge que ofereixi funcionalitats de consulta que vagin més enllà de la visualització bàsica de les metadades. En aquest sentit, s'ha plantejat la necessitat de fer cerques sobre conjunts de dades geospacials i eines de processament. Amb tal propòsit, els usuaris haurien de disposar d'un entorn (interfícies gràfiques) que facilités la generació de consultes així com la clara representació de resultats. A més, es planteja la necessitat de poder consultar tot el llinatge present en un catàleg de metadades, augmentant-ne el valor afegit, tot proporcionant als científics i tècnics l'oportunitat d'inspeccionar no només dades geospacials específiques, sinó el conjunt de dades d'un catàleg , algorismes i metodologies utilitzades.

Disposar d'un model de dades de llinatge sòlid i complet permet plantejar consultes complexes i obtenir resultats satisfactoris. La *Taula 13* mostra un conjunt de consultes que es poden formular no només sobre un conjunt de dades, sinó sobre la totalitat del catàleg de metadades. Els resultats d'aquestes consultes ens poden aportar informació sobre la credibilitat, la qualitat de les dades, la reproductibilitat de la informació i l'accessibilitat a les dades.

Taula 13: 28 consultes genèriques sobre el llinatge. Les columnes representen els avantatges o les aplicacions, mentre que les files representen els diferents elements que es consulten (Font: Taula 1 capítol 5).

Provenance elements	Provenance Benefits			
	Trust and authority	**Data Quality**	**Documentation and reproducibility**	**License and accesibility**
Sources	1. Which data have been created from untrusted sources? 2. Which are the most used sources?	10. Which data have been created using images with the highest positional accuracy? 11. Which sources have quality metadata?	16. Which are the differences between the list of sources used to produce two versions (or revisions) 17. Which sources points to real data or metadata?	22. Which data have been created using an open access data sources? 23. Which sources used are accessible in a distributed system? (list of url)
Process executions and methodology	3. In a family of methodologies, which products have used a specific member of the family? 4. Which methodologies are based on peer reviewed articles?	12. Is the methodology using the tool that produces the best results?	18. List the diferent workflows (process chained) used.	24. Which items have a citation pointing to a specific methodology?
Process algortihm	5. Which data have been created with untrusted software vendors 6. Which are the most used tools?	13. Which data have been created with the most recent vesion of the software?	19. Which are the differences between the list ofalgorithms used to produce two versions (or revisions) of the same product?	25. Which data have been created using an open source tools? 26. Which used tools are accessible in a distributed system? (list of url)
Time	7. Is one of the used sources too old for specific purspose?	14. Are all the sources from the same temporal interval as the output?	20. Is this methodology following the state of the art?	27. Which results have been created based on sources that were not openly accessible in a period of time?
Responsible parties	8. Which institution generated the maps? 9. Who was the technician in charge to generate each map?	15. How much data was elaborated by parties with a quality certification?	21. Who (scientist/institution) is responsible of this methodology?	28. Which institution is distributing the used sources?

7.3.3 Disseny d'una eina de consultes sobre el llinatge de catàlegs de dades en entorns distribuïts

S'ha dissenyat una eina dins el sistema servidor-navegador del MiraMon que permet filtrar i consultar la informació de *provenance* i representar el resultat d'acord al subapartat 7.2.2. L'eina (*Figura 19*) permet:

- La generació de consultes complexes sobre diversos conjunts de dades.
- La comparació entre diferents fluxos de processament, mostrant només aquells elements de llinatge compartits, tots, els que no estan compartits, etc.
- Els filtratge d'elements de llinatge (processos, agents, fonts, etc) que han participat en un flux de processament.
- La abstracció a nivell de funcionalitat de fluxos de processament concrets (p.ex. per mostrar el nom de la funcionalitat en lloc del nom del programa executat)

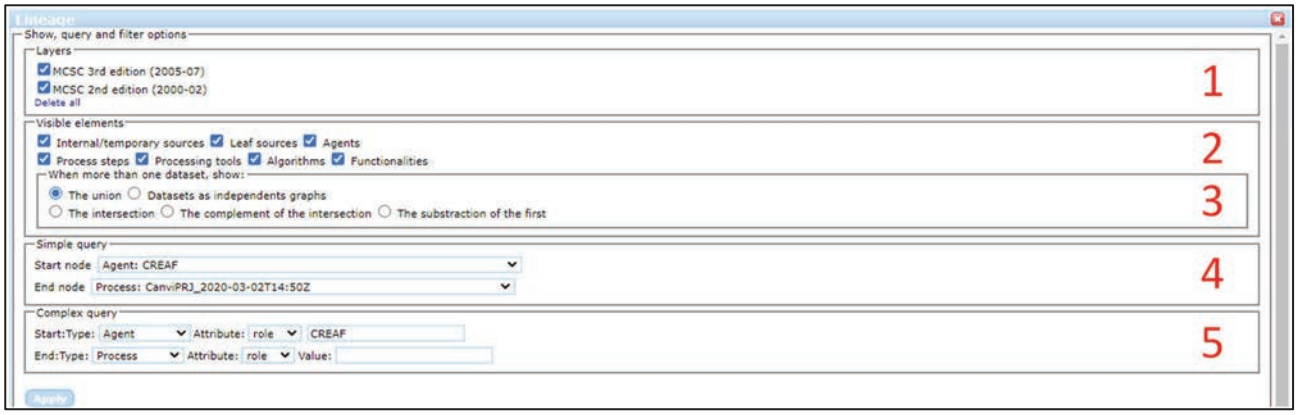

Figura 19: Detall del disseny del caixetí de l'eina de generació de consultes sobre la provenance del navegador de mapes. 1- Panell de conjunts de dades. 2- Panell visibilitat dels elements del llinatge. 3- Panell de fusió de dades. 4- Panell de consultes simples. 5- Panell de consultes complexes (Font: Figura 8, capítol 6).

8. Conclusions i reflexions

8.1 Conclusions (versió en català)

Aquesta tesi doctoral aprofundeix en les possibilitats i beneficis que es deriven d'una major i millor presència d'informació de llinatge en les metadades geospacials. Es realitzen aportacions, tant teòriques com aplicades, que milloren les fases de representació, captura, emmagatzematge, visualització i explotació del llinatge geospacial. Durant la realització de la tesi ha quedat palès que l'ús d'estàndards de representació i codificació del llinatge detallat és la millor via per facilitar la interpretació de la informació, assegurar-nos una documentació el més completa possible, incrementar-ne l'accés per part dels usuaris i fer possible consultes complexes. També s'han constatat les dificultats existents en l'aplicació dels estàndards i l'explotació de la informació que se n'extreu.

Tot i les millores que s'han produït els darrers anys en quan a la completesa de la descripció del llinatge (p. ex. millores en el model LI_Lineage o l'aparició del W3C PROV), les metadades disponibles a dia d'avui encara estan lluny treure el màxim profit de les possibilitats dels models i oferir una visió completa del procés de producció que ens acosti als beneficis que té la documentació del llinatge. Aquest fet és atribuïble a:

- L'escassetat d'eines que ajudin a la documentació automàtica del llinatge fa que l'esforç d'adquisició sigui elevat.
- L'escassetat d'eines i interfícies que en facilitin la interpretació i explotació posterior provoca una falta de motivació per a recollir el llinatge en detall.

Al llarg dels diversos capítols d'aquesta tesi s'han analitzat en profunditat el model de llinatge inclòs en l'ISO 19115 i el model W3C PROV. A més, tot i no proveir explícitament un model de llinatge, també s'ha estudiat el potencial de l'estàndard WPS per millorar la documentació de la informació de llinatge. En aquest sentit podem concloure que:

- El model de llinatge de la família d'estàndards ISO 19115 expressa els noms dels objectes geoespacials implicats d'una manera precisa i amb una nomenclatura geogràfica. Això, juntament amb el fet que l' ISO 19115 és el model de metadades emprat per les principals aplicacions geospacials, el situa com un model adient per representar el llinatge geoespacial des del punt de vista de la interoperabilitat i la semàntica. Ara bé, la seva estructura en forma d'arbre jeràrquic focalitzat en el producte final dificulta la representació i, conseqüentment, la interpretació del llinatge en casos de cadenes de processos complexos o que no siguin seqüencials.
- El model W3C PROV, desenvolupat per representar el llinatge de qualsevol objecte a Internet, és adaptable a les necessitats especifiques de cada disciplina i, per tant, es pot aplicar també a la geoinformació. La seva estructura modular, centrada més en la relació dels diferents elements que en el producte final, el fa adient per representar fluxos complexos no lineals i relacionar el llinatge de diversos conjunts de dades. A més, el fet que la seva notació es basi en identificadors universals (URIs), el fa òptim en entorns distribuïts en general i en el *linked data* en particular. Per contra, la manca d'una semàntica geospacial que defineixi els conceptes concrets de la disciplina (p. ex. col·lecció, atribut o la geometria) fa necessari un

procés d'especialització dels diferents elements del model PROV per reflectir els conceptes geogràfics.

- Els models de dades inclosos a l'estàndard WPS, tot i que no han estat dissenyats per descriure llinatge, ens aporten informació molt valuosa. Els models, orientats a la descripció de fonts i processos implicats en una execució, s'han mostrat molt útils per complementar els models explícits de llinatge existents.

En relació als models de representació del llinatge s'han extret les següents conclusions més especifiques:

- El nivell de detall (granularitat) necessari per descriure el llinatge varia en funció de la complexitat i naturalesa dels geoprocessos realitzats sobre un conjunt de dades. En alguns casos, el llinatge a nivell de conjunt de dades és suficient (p. ex. en un buffer de distàncies d'un sol conjunt de dades) mentre que en altres casos és necessari que els models permetin la representació del llinatge als nivells d'element geospacial i d'atribut (p. ex. en la fusió de conjunts de dades en un de sol). S'ha evidenciat que malgrat que el model ISO 19115 podria permetre diferents nivells de granularitat, el seu disseny resulta en una notació indirecta i complicada. A més, la combinació del model ISO amb GML (*Geographic Markup Language*) proposada com una solució per documentar el llinatge a nivell d'atribut, resulta en una codificació massa enrevessada que requereix capacitats d'emmagatzematge elevades. Per altra banda, el model W3C PROV, convenientment adaptat, resulta apropiat per documentar el llinatge als diferents nivells de granularitat degut a la seva estructura de relacions i la seva flexibilitat en la semàntica.
- El model de llinatge ISO (en la seva versió ISO19115:2014) contenia llacunes respecte a la descripció dels paràmetres literals. Aquestes llacunes, exposades als capítols 3 i 4, són claus per dotar al model ISO d'una narrativa precisa que l'acosti a la reproductibilitat dels processos que documenta el llinatge. L'aplicabilitat d'aquestes idees conceptuals ha quedat demostrada pel fet que les propostes aquí detallades varen entrar en els processos formals de revisió de l'estàndard ISO i algunes, com és el cas de la descripció dels valors literals del paràmetres de les execucions, van ser acceptades i es troben incloses en la versió vigent ISO 19115-2 aprovada el 2019.
- La combinació del model ISO amb els documents retornats per a la operació *describeProcess* del WPS és una bona solució per documentar el llinatge d'una manera més completa. La generació de plantilles WPS *describeProcess response* permet estandarditzar les descripcions del llinatge de cadascuna de les eines del MiraMon. Aquesta solució ens aproxima molt a la literalitat de les execucions, fet que permet reproduir processos amb molta fidelitat. L'exemple d'ús presentat al capítol 4 n'és la constatació.
- Es necessari que el llinatge documenti també les execucions d'aquells experiments científics que no són purament seqüencials, com ara bucles o les execucions descartades. La proposta d'usar *LE_Processing:otherPropertyType* del model *LE_Lineage* (ISO 19115-2) és una solució pràctica per a la documentació de les execucions descartades.

- La inclusió de diferents nivells d'abstracció dels processos (execució, eina, algorisme i funcionalitat) és una solució efectiva per tal de dotar a la informació de llinatge de capacitats per comparar cadenes de processament i metodologies amb independència del software utilitzat. La combinació del model ISO, per documentar els elements geospacials, amb el model W3C PROV, per representar-ne les relacions, ha resultat ser una bona combinació per documentar els diferents nivells d'abstracció.

Donat l'elevat nombre de relacions i la complexitat de les dependències que en ocasions arriba a incloure el llinatge, s'ha constatat que la documentació del llinatge a posteriori de la generació de la pròpia informació és un procés molt costós. Un procés que, en cas de fer-se manual, obre la porta a la introducció d'errors. Per tant, són necessàries eines que capturin el llinatge de forma concurrent a l'execució dels processos. A més, també són necessàries eines que permetin l'edició posterior del llinatge capturat. En aquest sentit, el *Provenance Engine* (PE) desenvolupat dins del MiraMon s'ha demostrat útil per capturar el llinatge. L'exemple d'ús presentat al capítol 4 n'és la constatació del seu funcionament i capacitats:

- El seu funcionament en paral·lel a l'execució del procés possibilita la documentació del llinatge automàticament.
- La interfície del GeMM possibilita l'edició del llinatge a posteriori. Aquest fet és molt important per:
  - Documentar el llinatge d'aquells elements que no formen part del MiraMon i no participen del PE.
  - Completar el llinatge o afegir descripcions del context de generació de les dades en aquells casos que sigui necessari.
  - Simplificar el propi llinatge per fer-lo més entenedor eliminant passos irrellevants.
- El fet que les descripcions del llinatge estiguin sustentades en plantilles *describeProcess* WPS que contenen tant les descripcions (en estil *human reading*) de l'eina i de la resta de paràmetres, com la sintaxi usada per l'eina, ens possibilita principalment dues coses:
  - S'eviten els errors derivats de l'entrada manual del llinatge.
  - El procés de producció queda completament documentat possibilitant la replicació de conjunts de dades i reutilització de cadenes de processos.

El potencial d'ús del llinatge va estretament lligat a la capacitat d'interpretació del llinatge per part dels usuaris. Més enllà de la seva representació, les tècniques de visualització i simbolització són claus. Aquesta tesi ha contribuït a la visualització del llinatge en dos vessants diferents:

- L'estil de representació en forma d'arbre jeràrquic emprat dins la PE ha resultat un model adequat per representar el llinatge d'un conjunt de dades. Aquest estil resulta especialment adequat per representar el llinatge com una successió de processos vinculant d'una manera directa les eines, els paràmetres i els resultats.
- La representació en forma de xarxa s'ha demostrat útil per tal de representar el llinatge de diversos conjunts de dades. A més, la possibilitat de representar els diversos nivells d'abstracció dels processos (execució, eina, algorisme i funcionalitat) en un sol gràfic permet visualitzar el llinatge des d'una perspectiva més didàctica i descriptiva.

La possibilitat de generar consultes sobre el llinatge (implementacions incloses als capítols 5 i 6) és un mecanisme útil per treure partit de la documentació del procés de generació de les dades. A més, el filtratge de la informació de llinatge potencia l'explotació dels seus beneficis (p. ex. detecció d'errors, escrutini, reproducció de metodologies), motivant als usuaris a la seva captura.

Finalment, del conjunt de propostes presentades al llarg de la tesi en podem fer algunes reflexions més generals:

Hom podria pensar que seria possible representar, en una sola instància, el llinatge del catàleg de col·leccions de dades utilitzant els mecanismes de representació dels diferents nivells de granularitat. En part és així, tal i com hem vist al capítol 2. El model ISO disposa d'un vocabulari concret per representar les diferents jerarquies de la IG (sèrie de conjunts de dades, conjunt de dades, elements geospacials individuals, atributs, etc). Per altre banda el model PROV és prou flexible per representar els diferents nivell de granularitat de la IG d'una manera directa, eficient i poc redundant. En el capítol 2, aquests mecanismes serveixen per a vincular dades sota un mateix paraigües a partir del seu llinatge. Aquest paraigües està delimitat pels diferents nivells d'abstracció (granularitat) de la IG i restringeix connexions amb dades allunyades més enllà que el que les pròpies jerarquies permeten. En canvi, l'abstracció de les eines de processament presentada al capítol 6 ens permet vincular sota un sol paraigües processos que comparteixen el mateix algorisme o funcionalitat (amb el benentès que la majoria d'IG és generada per un conjunt limitat i ben descrit de funcionalitats). Aquest fet, a la vegada que ens permet comparar metodologies i eines, ajuda a evitar l'aïllament temàtic de la informació tot connectant-la a partir d'una xarxa de llinatge de la mateixa manera que ho fa el *Linked Data*.

Cal destacar també l'amplitud de les propostes presentades dins el del camp de la gestió del llinatge. Aquesta tesi exposa aproximacions i millores tant en les fases de captura i emmagatzematge del llinatge, com en les fases de visualització, ús i explotació. Aquest fet fa que les propostes tinguin interès tant pels productors d'informació geogràfica (les fases de captura i emmagatzematge), com pels usuaris o consumidors (visualització, ús i explotació). El conjunt de propostes s'ha basat sempre en l'ús i la combinació de models de dades i estàndards de metadades existents per determinar-ne la seva complementarietat i detectar i cobrir possibles mancances. Aquest fet afavoreix l'estandardització i la interoperabilitat, i evita duplicitats. A més, el fet que les propostes hagin estat implementades i testejades dins un SIG i un navegador de mapes aporta validesa al conjunt de la tesi.

Línies de futur:

La recerca plantejada obre noves vies de treball en les aplicacions de consulta del llinatge geospacial, tant en les aplicacions d'escriptori, com en els navegadors de mapes. Cada vegada és més freqüent que les aplicacions del món dels SIG associades als processos tinguin actualitzacions freqüents que solucionin defectes (bugs) o aportin noves funcionalitats. Cal investigar de quina manera es pot usar el llinatge per determinar els impactes dels defectes o errors descoberts en aplicacions i/o en productes ja creats amb les versions antigues d'aquestes mòduls.

Cal aprofundir en la utilització del llinatge retrospectiu de conjunts de dades existents i com convertir-lo en llinatge prospectiu de noves cadenes de processament de cara a afavorir la idea de la reproductibilitat de cadenes de processos. En un context en el que dades (fonts) i eines (processos) són accessibles com a serveis web, la reproductibilitat de les dades hauria de poder ser més directe i eficient. A més, els diferents nivells d'abstracció dels processos presentats al capítol 6 (documentació de l'eina-algorisme-funcionalitat de processament) pot ajudar a exportar, no només cadenes de processos sinó les metodologies i utilitzar-les en altres programaris. Aquestes abstraccions en els processos es podrien complementar amb abstraccions en els conjunts de dades (les dades també es poden abstreure a nivell dels conceptes i temes que representen).

També veiem la necessitat d'estudiar com el llinatge pot contribuir a la formalització de llibreries de models i simulacions a la web que contribuirien a les infraestructures científiques emergents com la *European Open Data Cloud*. En aquest sentit, una cadena de processos es pot abstraure a un nivell superior que els agrupés sota un model: un model que externament no es mostra com una cadena de processos, sinó com una eina concreta que necessita unes fonts per generar unes sortides determinades.

8.2. Conclusiones (versión en catellano)

Esta tesis doctoral profundiza en las posibilidades y beneficios que se derivan de una mayor y mejor presencia de la información de linaje en los metadatos geoespaciales. Se realizan aportaciones, tanto teóricas como aplicadas, que mejoran las fases de representación, captura, almacenamiento, visualización y explotación del linaje geoespacial. Durante la realización de la tesis ha quedado patente que el uso de estándares de representación y codificación del linaje detallado es la mejor vía para facilitar la interpretación de la información, asegurarnos una documentación lo más completa posible, incrementar el acceso por parte de los usuarios y posibilita la realización de consultas complejas. A su vez, también se han constatado las dificultades existentes en la aplicación de los estándares y su explotación de la información que se extrae.

A pesar de las mejoras que se han producido en los últimos años en cuanto a la completitud de la descripción del linaje (p. ej: mejoras en el modelo LI_Lineage o aparición W3C PROV), los metadatos disponibles hoy en día todavía están lejos de expresar las posibilidades de los modelos y ofrecer una visión completa del proceso de producción que nos acerque a los beneficios que se obtienen de la documentación del linaje. Este hecho es atribuible a:

- La escasez de herramientas que ayuden a la documentación automática del linaje hace que el esfuerzo para su adquisición sea elevado.
- Escasez de herramientas e interfaces que faciliten su interpretación y explotación posterior provoca una falta de motivación para recoger el linaje en detalle.

A lo largo de los distintos capítulos se han analizado en profundidad el modelo de linaje incluido en la ISO 19115 y el modelo W3C PROV. Además, a pesar de no ser un modelo de linaje explícito, también se ha estudiado el potencial del estándar WPS para documentar información de linaje. En este sentido podemos extraer las siguientes conclusiones:

- El modelo de linaje de la familia de estándares de la ISO 19115 expresa los nombres de los objetos geoespaciales implicados de un modo preciso y con una nomenclatura geográfica. Esto, juntamente con el hecho que es el modelo de metadatos utilizado por los principales geoservicios, lo sitúa como un modelo adecuado para representar el linaje geoespacial desde el punto de vista de la interoperabilidad y la semántica. Ahora bien, su estructura en forma de árbol jerárquico focalizado en el producto final dificulta la representación y, consecuentemente, la interpretación del linaje en casos de cadenas de procesos complejos o que no sean lineales o secuenciales.
- El modelo PROV, desarrollado para representar el linaje de cualquier objeto en Internet es adaptable a las necesidades específicas de cada disciplina y, por lo tanto, se puede aplicar también a la geoinformación. Su estructura modular, centrada más en la relación de los distintos elementos que en el producto final, lo hace adecuado para representar flujos complejos no lineales y relacionar el linaje de distintos conjuntos de datos. Además, el hecho que su notación se base en identificadores universales (URIs) lo hace óptimo para entornos distribuidos en general y el *linked data* en particular. Por el contrario, la falta de una semántica geoespacial que defina los conceptos concretos de la disciplina (p.ej: resolución, nivel de

procesamiento o geometría) hace necesario un proceso de especialización de los distintos elementos del modelo PROV para reflejar los elementos geográficos.

- Los modelos de datos incluidos en el estándar WPS, aunque no han sido diseñados para describir el linaje, aportan información muy valiosa. Los modelos, orientados a la descripción de fuentes y procesos implicados en una ejecución, se han mostrado muy útiles para complementar los modelos explícitos de linaje existentes.

En relación con la representación del linaje se han extraído las siguientes conclusiones más concretas:

- El nivel de detalle (granularidad) necesario para describir el linaje varía en función de la complejidad y naturaleza de los geoprocesos realizados sobre un conjunto de datos. En algunos casos, el linaje a nivel de conjunto de datos es suficiente (p. ej: *buffer* de distancias de un solo conjunto de datos), mientras que en otros es necesario que los modelos permitan la representación del linaje a los niveles de elemento y de atributo (p.ej: en la fusión de conjuntos de datos). Se ha evidenciado que, pese a que el modelo ISO podría permitir distintos niveles de granularidad, su diseño resulta en una notación indirecta y complicada. Además, la combinación del modelo ISO con GML (Geographic Markup Language) propuesta como una solución para documentar el linaje a nivel de atributo, resulta en una codificación demasiado enrevesada que requiere capacidades de almacenamiento elevadas. Por otra parte, el modelo W3C PROV, convenientemente adaptado, resulta apropiado para documentar el linaje a los distintos niveles de granularidad debido a su estructura modular y su flexibilidad en la semántica.
- El modelo de linaje ISO (en su versión ISO: 2014) contenía lagunas respecto a la descripción de los parámetros literales. Estas lagunas, convenientemente presentadas en los capítulos 3 y 4, son claves para dotar al modelo ISO de una narrativa precisa que le acerque a la reproducibilidad de los procesos que documenta el linaje. La aplicabilidad de estas ideas conceptuales ha quedado demostrada por el hecho de que las propuestas formaron parte de los procesos formales de revisión del estándar ISO y algunas, como es el caso de la inclusión de los valores literales de las ejecuciones, fueron aceptadas y se encuentran incluidas en la versión actual ISO 19115-2 aprobada en 2019.
- La combinación del modelo ISO con los documentos devueltos por la operación *describeProcess* del WPS es una buena solución para documentar el linaje de una manera más completa. La generación de plantillas WPS *describeProcess response* permite estandardizar las descripciones del linaje de cada una de las herramientas del MiraMon. Esta solución nos aproxima mucho a la literalidad de las ejecuciones, lo que nos permite reproducir procesos con mucha fidelidad. El ejemplo de uso presentado en el capítulo 4 es la constatación.
- Es necesario que el linaje documente también las ejecuciones de aquellos experimentos científicos que no son puramente secuenciales, tales como bucles o ejecuciones descartadas. La propuesta de usar *LE_Processing: otherPropertyType* del modelo LE_Lineage (ISO 19115-2) es una solución práctica para la documentación de las ejecuciones descartadas.

- La inclusión de distintos niveles de abstracción de los procesos (ejecución, herramienta, algoritmo y funcionalidad) se ha mostrado como una solución efectiva para dotar a la información de linaje de capacidades para comparar cadenas de procesamiento y metodologías con independencia del software utilizado. La combinación del modelo ISO, para documentar los elementos geoesopaciales, con el modelo PROV, para representar las relaciones, ha resultado ser una buena combinación para documentar los distintos niveles de abstracción.

Dado el elevado número de relaciones y la complejidad de las dependencias que en ocasiones llega a documentar el linaje, se ha constatado que la documentación del linaje a posteriori de la generación de la propia información es un proceso muy costoso. Un proceso que, en caso de hacerse manual, abre la puerta a la introducción de errores. Por tanto, son necesarias herramientas que capturen el linaje de forma concurrente a la ejecución de los procesos. Además, herramientas que permitan la edición y la documentación posterior del linaje también son necesarias. En este sentido la *Provenance Engine* (PE) desarrollada dentro del MiraMon se ha demostrado útil para capturar el linaje. El ejemplo de uso presentado en el capítulo 4 es la constatación de su funcionamiento y capacidades:

- Su funcionamiento en paralelo a la ejecución del geoproceso posibilita la documentación del linaje automáticamente.
- La interfaz del GeMM posibilita la edición del linaje *a posterior*i. Este hecho es muy importante para:
    - Documentar el linaje de aquellos elementos que no forman parte de MiraMon y no participan del PE.
    - Completar el linaje o añadir descripciones del contexto de generación de los datos en aquellos casos que sea necesario.
    - Simplificar el propio linaje para hacerlo más comprensible eliminando pasos irrelevantes.
- El hecho que las descripciones del linaje estén sustentadas en plantillas *describeProcess* del WPS que contienen tanto las descripciones (en estilo *human reading*) de la herramienta y del resto de parámetros, como la sintaxis usada para la herramienta en el programa MiraMon, nos posibilita dos cosas principalmente:
    - Se evitan los errores derivados de la entrada manual del linaje.
    - El proceso de producción queda completamente documentado posibilitando la replicación de conjuntos de datos y la reutilización de cadenas de procesos.

El uso potencial del linaje va estrechamente ligado a la capacidad de interpretación de este por parte de los usuarios. En este sentido, más allá de su representación, las técnicas de visualización y simbolización son claves. Esta tesis ha contribuido a la visualización del linaje en dos maneras distintas:

- El estilo de representación en forma de árbol jerárquico empleado en la PE ha resultado un modelo adecuado para representar el linaje de un conjunto de datos. Este estilo resulta

especialmente adecuado para representar el linaje como una sucesión de procesos vinculando directamente las herramientas y los parámetros con los resultados.

- La representación en forma de red se ha demostrado útil para representar el linaje de varios conjuntos de datos. Además, la posibilidad de representar los distintos niveles de abstracción de los procesos (ejecución, herramienta, algoritmo y funcionalidad) en un solo gráfico permite visualizar el linaje con orígenes de procesamiento muy dispares.

La posibilidad de generar consultas sobre el linaje (implementaciones incluidas en los capítulos 5 y 6) se ha demostrado como un mecanismo útil para sacar partido de la documentación del proceso de generación de los datos. Además, el filtrado de la información del linaje potencia la explotación de sus beneficios (p. ej. detección de errores, escrutinio, reproducción de metodologías), motivando a los usuarios a su captura.

Finalmente, del conjunto de propuestas presentadas a lo largo de la tesis podemos hacer algunas reflexiones más generales:

Se podría pensar que sería posible representar, en una sola instancia, el linaje de catálogos de colecciones de datos utilizando los mecanismos de representación de los distintos niveles de granularidad. En parte es así, tal y como vimos en el capítulo 2. El modelo ISO dispone de un vocabulario concreto para representar las diferentes jerarquías de la IG (series de conjuntos de datos, conjuntos de datos geográficos, fenómenos geográficos individuales, propiedades de los fenómenos, etc). Por otro lado, el modelo PROV es suficientemente flexible para representar los distintos niveles de granularidad de la IG de una manera directa, eficiente y poco redundante. En el capítulo 2, estos mecanismos sirven para vincular datos bajo un mismo paraguas a partir de su linaje. Este paraguas está delimitado por los distintos niveles de abstracción (granularidad) de la IG y restringe las conexiones con datos alejados más allá de las propias jerarquías. En cambio, la abstracción de las herramientas de procesamiento presentada en el capítulo 6 nos permite vincular bajo un solo paraguas procesos que comparten el mismo algoritmo o funcionalidad (entendiendo que la mayoría de IG es generada por un conjunto limitado y bien descrito de funcionalidades). Este hecho, a la vez que nos permite comparar metodologías y herramientas, ayuda a evitar el aislamiento temático de la información conectándola a partir de una red de linaje del mismo modo que lo hace el *Linked Data*.

Cabe destacar también la amplitud de las propuestas presentadas en el del campo de la gestión del linaje. La presente tesis ha expuesto aproximaciones y mejoras tanto en las fases de captura y almacenamiento del linaje, como en las fases visualización, uso y explotación. Este hecho hace que las propuestas tengan interés tanto para los productores de información geográfica (las fases de captura y almacenamiento), como para los de usuarios o consumidores (visualización, uso y explotación). El conjunto de las propuestas se ha basado siempre en el uso y/o la combinación de modelos de datos y metadatos existentes con el objetivo de favorecer la estandarización, la interoperabilidad y evitar en la medida de lo posible las duplicidades. Además, el hecho de que las propuestas hayan sido implementadas y testeadas en un SIG y un navegador de mapas aporta validez al conjunto de la tesis.

Líneas de futuro:

La investigación planteada abre nuevas vías de trabajo en las aplicaciones de consulta del linaje geoespacial, tanto en las aplicaciones de escritorio, como en los navegadores de mapas. Cada vez es más frecuente que los procesos y las aplicaciones del mundo de los SIG tengan asociadas actualizaciones frecuentes que solucionen defectos (bugs) o aporten nuevas funcionalidades. Hay que investigar de qué manera se puede usar el linaje para determinar los impactos de los defectos o errores descubiertos en aplicaciones y/o en productos ya creados con versiones antiguas de estas herramientas.

Hay que profundizar en la utilización del linaje retrospectivo de conjuntos de datos existentes y convertirlo en linaje prospectivo de nuevas cadenas de procesamiento de cara a favorecer la reproducibilidad de los procesos. En un contexto en el que datos (fuentes) y herramientas (procesos) son accesibles como geoservicios, la reproducibilidad de los datos debería poder ser más directa y eficiente. Además, los diferentes niveles de abstracción de los procesos presentados en el capítulo 6 (documentación de la herramienta-algoritmo-funcionalidad de procesamiento) puede ayudar a exportar no sólo cadenas de procesos sino las metodologías y utilizarlas en otros contextos o softwares.

Un paso más en el proceso de abstracción de los procesos y la utilización del prospectivo sería la generación de modelos. En este sentido una cadena de procesos se podría abstraer a un nivel superior que los agrupara bajo un solo modelo. Un modelo que externamente no se comportara como una cadena de procesos, sino como una herramienta concreta que necesita unas fuentes concretas para generar unas salidas determinadas. En definitiva, al final de todos los desarrollos el linaje se podría convertirse en la principal fuente de una herramienta de generación de modelos.

8.3. Conclusions (English version)

This PhD study the possibilities and benefits derived from a greater and better presence of lineage information into geospatial metadata. At the same time, makes contributions, both theoretical and applied, to improve the phases of representation, capture, storage, visualization and exploitation of the geospatial lineage. During this PhD it has become clear that the use of standards to represent and encode a detailed lineage is the best way to enhance the interpretation of information, ensure documentation as much comprehensive as possible, increase access and facilitate the generation of complex queries. In addition, have been verified the difficulties in the application of the standards and the exploitation of the information extracted from them.

Despite the progress achieved in recent years in the completeness of the lineage description (e.g. improvements in the LI_Lineage model or the release of the W3C PROV), the metadata available nowadays does not take full advantage of the possibilities that models offers and does not present a complete view of the production process that brings users closer to the benefits of lineage documentation. This fact can be attributed to:

- The scarcity of tools to help in the automatically lineage documentation generates a high effort to capture lineage.
- The scarcity of tools and interfaces to enhance the interpretation and the subsequent exploitation results in lack of motivation to collect the lineage in a detailed way.

Throughout the various chapters of this document, the lineage model included in ISO 19115 and the W3C PROV model have been deeply analysed. In addition, even though does not provide an explicit lineage model, the potential of the WPS standard to improve the documentation of lineage information has been also studied. In this sense we can conclude that:

- The lineage model of the ISO 19115 standard family expresses the names of the geospatial objects involved in a precise manner and with a geographical nomenclature. This, together with the fact that ISO 19115 is the metadata model used for the major geospatial applications, places it as a suitable model to represent the geospatial lineage from the point of view of interoperability and semantics. However, its hierarchical tree structure focused on the final product complicate the representation and, consequently, the interpretation of lineage in complex chains or in non-sequential processes.
- The W3C PROV model, developed to represent the lineage of any resource on the Internet, is adaptable to the specific needs of any discipline and, therefore, can also be applied to geoinformation. Its modular structure, focused more on relations of different resources than on the final product, makes it suitable for representing complex nonlinear workflows and relate the lineage of several datasets. In addition, the fact that its notation is based on universal identifiers (URIs) makes it optimal in distributed web environments, in general, and in linked data particularly. In contrast, the lack of a geospatial semantics that defines specific concepts of the discipline (e.g. resolution, processing level, or geometry) creates the need for specialization of the different elements of the PROV model to reflect geographic concepts.

- The data models included in WPS, although are not designed to describe lineage, provide very valuable information. The models, oriented to the description of sources and processes involved in an execution, have been very useful to complement the explicit lineage models.

In addition, more specific conclusions about the models of the lineage have been extracted:

- The level of detail (granularity) needed to describe lineage varies depending on the complexity and nature of the geoprocesses performed on a data set. In some cases, the lineage at the data set level is enough (e.g. in a distance buffer of a single layer) while, in other cases, it is necessary to allow the representation of the lineage at the feature and attribute levels (e.g. merging datasets into one). Although the ISO 19115 model may allow for representing the different levels of granularity, its design results less direct and complex notation. In addition, the combination of the ISO model with Geographic Markup Language (GML) proposed as a solution for documenting the lineage at the attribute level, results very verbose encoding and requires high storage capacities. On the other hand, the W3C PROV model, conveniently adapted, is appropriate to document the lineage at different levels of granularity due to its structure based on relations and its flexibility in semantics.
- ISO lineage model (in ISO19115:2014 version) had some gaps concerning to the description of the literal parameters. These gaps, pointed in chapters 3 and 4, are key to provide ISO model with a precise narrative that brings it closer to the reproducibility of the processes documented by the lineage. The applicability of these conceptual ideas has been demonstrated: the proposals detailed here entered in the formal processes of revision of the ISO standard and some, such as the description of the literal values of the parameters of the executions, were accepted and are included in the current version ISO 19115-2 approved in 2019.
- Combining the ISO model with the response documents of the WPS *describeProcess* operation has proved as a good solution for documenting the lineage in a more comprehensive way. The generation of *describeProcess response* templates allowed to standardize the lineage descriptions for each MiraMon application. This solution brings us very close to the literal nature of executions, which allows us to reproduce processes with great fidelity. The use case presented in chapter 4 is a demonstration of this.
- Lineage has to include also the documentation of the executions of scientific experiments that are not purely sequential, such as loops or discarded executions. The proposal of using the *LE_Processing: otherPropertyType* of the *LE_Lineage* model (ISO 19115-2) is a practical solution for documenting discarded executions.
- The inclusion of different levels of process abstraction (execution, tool, algorithm and functionality) is an effective solution in order to provide lineage information with capabilities to compare processing chains and methodologies regardless the software used. The combination of the ISO model, to document the geospatial elements, with the PROV model, to represent relationships, is a good combination to document the different levels of abstraction.

Given the high number of relations and the complexity of the dependencies that sometimes lineage information has, the documentation of lineage after the generation of the information itself is a tough process. A process that, if it is done manually, opens the door to the introduction of errors. Therefore, tools that automatically capture the lineage concurrently with the execution of the processes are needed. In addition, tools to edit and prune the captured lineage are needed too. In this sense, the Provenance Engine (PE) developed within MiraMon has proven as useful for capturing the lineage. The use case presented in chapter 4 is the verification of its performance and capabilities:

- The tool runs in parallel with the execution of the geoprocessing tools allowing the documentation of lineage automatically.
- The GeMM interface allows editing the lineage after the dataset generation. This is very important for:
    - Documenting the lineage of those elements that are not part of MiraMon and do not participate in the PE.
    - Complete the lineage or adding descriptions of the data generation context when necessary.
    - Simplify the lineage itself to make it more understandable by removing irrelevant steps.
- The descriptions of the lineage are based on *DescribeProcess* WPS templates that contain both the descriptions (in human reading style) of the tool and the other parameters, as well as the tool syntax used. This results mainly in two things:
    - Errors derived from manual lineage entry are avoided.
    - The production process is fully documented enabling dataset replication and reuse of process chains.

The potential of lineage is closely linked to the interpretation done by the users. Beyond their pure representation, visualization and symbolization techniques are key factors. This PhD has contributed to the visualization of the lineage in two ways:

- The hierarchical tree representation style used within PE has proved to be a suitable model for representing the lineage of a dataset. This style is especially suitable for representing lineage as a succession of processes that directly link tools, parameters, and results.
- The representation of lineage as a network has proved useful for representing the lineage of several datasets. In addition, the ability to represent several levels of abstraction of processes (execution, tool, algorithm and functionality) in a single graph allows users to view the lineage with a more didactic and informative perspective.

The ability to generate lineage queries (implementations included in chapters 5 and 6) has been proved to be a useful mechanism for leveraging information from the data generation process. In addition, filtering lineage information enhances the exploitation of its benefits (e.g. error detection, scrutiny, reproduction of methodologies), motivating users to capture it.

Finally, based on the on the set of proposals presented throughout this thesis some more general reflections can be added:

One might think that it would be possible to represent, in a single instance, the lineage of data collection catalogue using the mechanisms of representation of the different levels of granularity. This is partly true as we saw in chapter 2.The ISO model has a specific vocabulary to represent the different hierarchies of the geographical information (dataset series, geospatial datasets, individual geographic features, properties of a feature, etc). On the other hand, the PROV model is flexible enough to represent the different levels of granularity in a direct, efficient and non-redundant way. In chapter 2, these mechanisms serve to link data under the same umbrella using lineage. However, this umbrella is bounded by the different levels of abstraction (granularity) of the geographical information and does not permit connections beyond the possibilities provided by these hierarchies. Instead, the abstraction of the processing tools presented in chapter 6 allows to link under a single umbrella processes that share algorithms or functionalities (recognizing that most geographical information is generated by a limited and well-described set of GIS functionalities). This allows to compare methodologies and tools. In addition, it helps to avoid the thematic isolation of information by connecting it from a lineage network in the same way that *Linked Data* does.

It is important to underline the range of the presented proposals in the field of lineage management. This PhD sets out approaches and improvements from the phases of capture and storage of lineage, to the phases of visualization, usage and exploitation. This wide scope makes the findings interesting for both, producers (the phases of capture and storage) and users or consumers (visualization, use and exploitation) of geographical information. The set of proposals has always taken into consideration the use and combination of existing data models and metadata standards to determine their complementarity and to detect and cover possible gaps. This promotes standardization and interoperability and avoids duplication. In addition, the proposals have been implemented and tested in a GIS and in a map browser which adds validity to the thesis findings.

Future work:

The proposed research introduces new avenues of work in geospatial provenance query applications, both in desktop applications and in map browsers. It is increasingly common that GIS applications that are executing processes has frequent updates to fix bugs. It is necessary to investigate how the lineage can be used to determine the impacts of defects or errors discovered in applications and / or products already created with older versions of these modules.

It is necessary to investigate the use of the retrospective lineage of existing data sets and how to turn it into a prospective lineage of new processing chains in order to favour the idea of the reproducibility of process chains. In a context where data (sources) and tools (processes) are accessible as web services, data reproducibility should be able to be more direct and efficient. In addition, the different levels of process abstraction presented in chapter 6 (documentation of the tool, algorithm, processing and functionality) can help to export not only process chain but methodologies and use them in other software. This abstraction in processes can be complemented by abstraction in dataset (the data can also be abstracted to the level of the concepts and topics they represent).

We also see the need to study how lineage can contribute to the formalization of libraries of models and simulations on the web that would contribute to emerging scientific infrastructures such as the

European Open Data Cloud. In this sense, a processes chain can be abstracted to a higher level by grouping them under a model: a model that externally is not shown as a chain of processes, but as a specific tool that needs sources to generate certain outputs.

Bibliografía

Advanced Information Systems Laboratory. Universidad Zaragoza. (2012). *CatMDEdit*. Zaragoza.

Ahonen-Rainio, P., & Kraak, M.-J. (2005). Deciding on fitness for use: evaluating the utility of sample maps as an element of geospatial metadata. *Cartography and geographic information science*, 101-112.

Ajuntament de Barcelona. (2020). *GeoPortalBCN*. Retrieved 2020, from http://www.bcn.cat/geoportal/ca/estandards.html

Albrecht, J. (1999). Geospatial information standards. A comparative study of approaches in the. *Computers and Geosciences*, 9–24.

Alonso, G., & Hagen, C. (1998). Geo-Opera: Workflow concepts for spatial processes. *International Symposium on Spatial Databases* (p. 238-258). Berlin : Springer.

Amstutz, P., Crusoe, M., & Tijanić, N. (sense data). *Common Workflow Language, v1.0. Specification, Common Workflow Language working group.* doi:10.6084/m9.figshare.3115156.v2

ANSI. (1998). Spatial Data Transfer Standard (SDTS).

Ariza López, F., Barreira González, P., Masó Pau, J., Zabala Torres, A., Rodríguez Pascual, A., Moreno Vergara, G., & García Balboa, J. (2020). Geospatial data quality (ISO 19157-1): evolve or perish. *Revista Cartográfica 100*, 129-154.

Bai, Y., Di, L., & Wei, Y. (2009). A taxonomy of geospatial services for global service discovery and interoperability. *Computers & Geosciences*, 783-790.

Bauman, P. (2010). *OGC Web Coverage Service (WCS) Standard – Core, Ver. 2.0, OGC 09-110r3.* Retrieved 2020, from OGC: http://portal.opengeospatial.org/files/?artifact_id=41437

Berners-Lee, T. (2009). The next Web of open, linked data. *TED*.

Bizer, C. (2013). Interlinking scientific data on a global scale. *Data Science Journal*.

Borkin, M., Yeh, C., Boyd, M., Gajos, P., Seltzer, M., & Pfister, H. (2013). Evaluation of filesystem provenance visualization tools. *IEEE Transactions on Visualization and Computer Graphics*, (pp. 2476-2485).

Brackin, R., Gonçalvez, P., Masó, J., & David, R. (2014). OGC OWS Context Conceptual Model. *Open Geospatial Consortium*.

Brodeur, J., Coetzee, S., Danko, D, Garcia, S., & Hjelmager, J. (2020). Geographic Information Metadata—An Outlook from the International Standardization Perspective. *SPRS International Journal of Geo-Information*.

Celino, L. (2013). Human computation VGI provenance: semantic web-based representation and publishing. *Transactions on Geoscience and Remote Sensing*, 5137-5144.

CEOS. (2020). *Analysis Ready Data for Land (CARD4L)*. Consultat el 05 / 08 / 2020, a Analysis Ready Data for Land (CARD4L): http://ceos.org/ard/

Chen, P., Plale, B., & Cheah, Y. (2012). Visualization of Network Data Provenance. *19th International Conference on High Performance Computing* (pp. 1-9). IEEE.

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 and Urban Systems*, 103-117.

Coetzee, S., Ivánová, I., Mitasova, H., & Brovelli, M. (2020). Open Geospatial Software and Data: A Review of the Current State and A Perspective into the Future. *ISPRS Int. J. Geo-Inf.*

Cox, S. (2017). Ontology for observations and sampling features, with alignments to existing models. *Semantic Web*, 453-470.

Danko, D. (2007). Metadata and Interoperability, Geospatial. A S. Shekar, & H. Xiong, *Enciclopedia of GIS* (p. 1395). New York: Springer.

Del Rio, N., & Da Silva, P. (2007). Probe-it! visualization support for provenance. (p. In International Symposium on Visual Computing). Berlin,: Springer,.

Di, L., & McDonald, K. (1999). Next generation data and information systems for earth sciences research. 92–101.

Di, L., Han, W., Zhao, P., Wei, Y., & Li, X. (2008). Design and implementation of GeoBrain online analysis system (GeOnAS). *International Symposium on Web and Wireless Geographical Information Systems* (p. 27-36). Heidelberg.: Springer.

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. *Geoscience and Remote Sensing, IEEE Transactions*, 5082-5089.

Di, L., Yue, R., Ramapriyan, H., & King, R. (2013). Geoscience data provenance: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 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. *Int. J. Spat. Data Infrastruct. Res*, 352-377.

Dublin Core Metadata Initiative. (2020). *DCMI Metadata Terms.* Retrieved 08 18, 2020, from DCMI Metadata Terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms

Duckham, M., Arnold, L., Armstrong, K., McMeekin, D., & Mottolini, D. (2017). Towards a spatial knowledge infrastructure. *White paper*.

European Comission. (2007). *Directive 2007/2/EC of the European Parliament and the council of 14 March 2007 establishing an Infraestructure for spatial Information in the European Community (INSPIre).* Brusels.

European Comission. (2007). *Infraestructure for spatial information in Europe*. Retrieved from https://inspire.ec.europa.eu/

EuroSDR. (2018). *EuroSDR Annual Report 2018.* Consultat el 2020, a EuroSDR Annual Report 2018: http://www.eurosdr.net/sites/default/files/images/inline/eurosdr_annual_report_2018.pdf

FAIR Guiding Principles for scientific data management and stewardship. (2016). *FAIR Principles*. Consultat el 12 / 08 / 2020, a FAIR Principles: https://www.go-fair.org/fair-principles/

Feng, C. (2013). Mapping geospatial metadata to open provenance model. *IEEE transactions on geoscience and remote sensing, 51 (11)*, 5073-5081.

FGDC. (2010). *OMB Circular A-16*. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-016.pdf

Geller, G., & Turner, W. (2007). The model web: a concept for ecological forecasting. *In Geoscience and Remote Sensing Symposium, IGARSS* (pp. 2469-2472). IEEE International.

GEO Secretariat. (2005). *The Global Earth Obserbation System of Systems (GEOSS) 10-Year Implementation Plan: 11.*

Giuliani, G., Ray, N., & Lehmann, A. (2013). Building Regional Capacities for GEOSS and INSPIRE: a journey in the Black Sea Catchment. *International Journal of Advanced Computer Science and Applications*, 19-27.

Goodchild, M., Yuan, M., & Cova , T. (2013). Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 239-260.

Gorodov, E., & Gubarev, V. (2013). Analytical review of data visualization methods in application to big data. *Journal of Electrical and Computer Engineering.*

Gregersen, J., Gijsbers, P., & Westen, S. (2007). OpenMI: open modelling interface. *Journal of hydroinformatics*, 175-191.

Groth, P., & Moreau, L. (2013, 04 30). *PROV-Overview.* Retrieved 2020, from PROV-Overview: https://www.w3.org/TR/prov-overview/

Group on Earth Observations. (2020). *Global Earth Observation System of Systems (GEOSS)*. Recollit de Global Earth Observation System of Systems (GEOSS): https://www.earthobservations.org/geoss.php

Grup Enciclopèdia Catalana. (n.d.). *diccionari.cat*. Retrieved 07 09, 2020, from diccionari.cat: http://www.diccionari.cat/

Harth, A., & Gil, Y. (2014). Geospatial data integration with linked data and provenance tracking. *In W3C/OGC Linking Geospatial Data Workshop*, (pp. 1-5).

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

Henzen, C., Mäs, S., & Bernard, L. (2013). Provenance information in geodata infrastructures. *In Geographic Information Science at the Heart of Europe* (p. 133-151). Springer.

Henzen, C., Mäs, S., Zander, F., Schroeder, M., & Bernard, L. (2016). Representing research collaborations and linking scientific project results in spatial data infrastructures by provenance information. *AGILE.* Helsinki (Findland).

Herzig, A., Rutledge, D., Aus, A., & Dymond, J. (2019). LUMASS The Land Use Management Support System. Dubln.

Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bolaños, T., Bindi, M., Brown , S., . . . Engelbrecht, F. (2019). The human imperative of stabilizing global climate change at 1.5 C. *Science*.

Hoel, E. (2008). Data Models in Commercial GIS Systems. A S. Shekhar, & H. Xiong, *Enciclopedia of GIS* (p. 1370). New York: Springer.

Hua, H., & Tilmes, C. (2013). Earth Science Provenance Ontology (PROV-ES). *Summer Meeting 2013. ESIP Commons April 2013*.

Huynh, T., & Moreau, L. (2014). ProvStore: a public provenance repository. *At 5th International Provenance and Annotation Workshop (IPAW'14)*.

ISO. (2017). *ISO 5127 Information and documentation — Foundation and vocabulary.* Geneve, Switzerland: ISO.

ISO/TC 211. (2007). *ISO 19139:2007: Geographic information—Metadata.* Geneva, Switzerland: ISO.

ISO/TC 211. (2012). *ISO 19139-2: Geographic information—Metadata—Part 2: XML schema for imagery and gridded data.* Geneva, Switzerland: ISO.

ISO/TC 211. (2013). *ISO 19157 - Geagraphic Information - Data Quality.* Geneva, Switzerland: ISO.

ISO/TC 211. (2014). *ISO 19115-1:2014: Geographic Information – Metadata – Part 1: Fundamentals.* Geneva, Switzerland: ISO.

ISO/TC 211. (2016). *ISO 19115-3:2016: Geographic information - Metadata - Part 3: XML schema implementation for fundamental concepts.*

ISO/TC 211. (2019). *ISO 19115-2:2019: Geographic information – Metadata – Part 2: Extensions for acquisition and processing.* Geneva, Switzerland: ISO.

Ivánová, I., Armstrong, K., & McMeekin, D. (2017). Provenance in the next-generation spatial knowledge infrastructure. *In 22nd International Congress on Modelling and simulation* (pp. 410-416). Hobart, Tasmania: Modelling and Simulation Society of Australia and New Zealand.

Jiang, L., Kuhn , W., & Yue, P. (2017). An interoperable approach for Sensor Web provenance. *6th International Conference on Agro-Geoinformatics*, (p. 1-6).

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*, 21-31.

Kalantari, M., Olfat, H., & Rajabifard, A. (2010). Automatic spatial metadata enrichment: reducing metadata creation burden through spatial folksonomies. *In GSDI 12 World Conference: Realising Spatially Enabled Societies,.* Singapore.

Kalantari, M., Syahrudin, S., Rajabifard, A., & Subagyo, H. (2020). Spatial Metadata Usability Evaluation. *ISPRS International Journal of Geo-Information*.

Keßler, C., & De Groot, R. (2013). Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. *Geographic information science at the heart of Europe*, 21-37.

Kunde, M., Bergmeyer, H., & Schreiber, A. (2008). Requirements for a provenance visualization component. In D. K. J. Freire (Ed.), *IPAW* (pp. 241–252). Salt Lake City: Springer.

Lakshmanan, G., Curbera, F., Freire, J., & Sheth, A. (2011). Guest editors introduction: Provenance in Web applications. *IEEE Internet Computing*, 17–21.

Lanter, D. (1991). Design of a Lineage-Based Meta-Data Base for GIS. *Cartography and Geographic Information Systems*, 255-261.

Lim, S., Lu, S., Chebotko, A., & Fotouhi, F. (2010). Prospective and retrospective provenance collection in scientific workflow environments. *International Conference on Services Computing* (p. 449-456). IEEE .

Magagna , B., Goldfarb, D., Martin, P., Atkinson, M., Koulouzis , S., & Zhao , Z. (2020). Data Provenance. In Z. Zhao, & M. Hellström , *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences.* Springer.

Masó, J. (2012). *Models de dades dels SIG a Internet. Aspectes teòrics i aplicats.* Barcelona: Universitat Autònoma de Barcelona.

Masó, J., Closa, G., Gil, Y., & Proß., B. (2014). *OGC Public Engineering Report: OGC®Testbed 10 Provenance Engineering Report.* Wayland, MA, USA: Open Geospatial Consortium Inc.

Masó, J., Pomakis, K., & Julià , N. (2010). *Web Map Tile Service Implementation Standard (WMTS). OGC 07-057r7.* Consultat el 01 / 08 / 2020, a OGC: http://www.opengeospatial.org/standards/wmts

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

Metadata Ad Hoc Working Group, F. G. (1998). Content Standard for Digital Geospatial Metadata.

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

Moreau, L., & Missier, P. (2013, 04 30). *PROV-DM: The PROV Data Model.* Retrieved 2020, from https://www.w3.org/TR/2013/REC-prov-dm-20130430/

Moreau, L., & Missier, P. (2013, 04 30). *PROV-N: The Provenance Notation.* Retrieved from PROV-N: The Provenance Notation: https://www.w3.org/TR/prov-n/

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., . . . Van den Bussche, J. (2011). The open provenance model core specification (v1. 1). *Future generation computer systems*, 743-756.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., . . . Van den Busshce, J. (2011). The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems, 27 (6)*, 743-756. doi:10.1016/j.future.2010.07.005

Munafò, M., Nosek, B., Bishop, D., Chambers, C., Du Sert, N., Simonsohn , U., . . . Ioannidis , J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*. Retrieved from A manifesto for reproducible science.

Myers, L., Futrelle, J., Gaynor, J., & Plutchak, J. (2009). Embedding Data within Knowledge Spaces. *arXiv preprint arXiv:0902.0744.*

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

Nebert, D., Whiteside, A., & Vretanos, P. (2007). *OGC Catalogue Service Implementation.* Consultat el 08 / 01 / 2020, a OGC: http://portal.opengeospatial.org/files/?artifact_id=20555

Nunes, J. (2012). *Diccionari terminològic de sistemes d'informació geogràfica (Diccionaris Terminològics).* Barcelona: Institut Cartogràfic de Catalunya: Enciclopèdia Catalana.

OGC. (2013). *OWS-9 Cross Community Interoperability (CCI) Conflation with Provenance.* Retrieved 17 07, 2020, from Engineering Report: www.opengeospatial.net/def/doc-type/per/cciconflation-provenance

OGC. (2014). *Testbed 10 Provenance.* Retrieved 07 17, 2020, from Engineering Report: http://www.opengis.net/doc/ER/testbed10/provenance

OGC. (2015). *Testbed 11 Data broker specifications.* Retrieved 07 017, 2020, from Engineering Report: http://docs.opengeospatial.org/per/16-045r2.html

OGC. (2016). In The OpenGIS Abstract Specification, version 4;. *Topic 11: Metadata (OGC 11-111r1).*

OGC. (2016). *Testbed 12 Semantic portrayal, registry and mediation.* Retrieved 07 17, 2020, from Engineering Report: http://docs.opengeospatial.org/per/16-059.html

OGC. (07 / 06 / 2020). *Community Standards*. Recollit de https://www.ogc.org/resource/products

Olcina, J., Biener, S., & Marti Talavera, J. (2020). Aspectos atmosféricos y climáticos en la expansión de la pandemia (COVID-19) en la provincia de Alicante. *Investigaciones Geográficas*, 275-297.

Oliveira, W., Ambrósio, L., Braga, R., Ströele, V., David, J., & Campos, F. (2017). A framework for provenance analysis and visualization. *Procedia Computer Science*, 1592-160.

Özkundakci, D., Wallace, P., Jones, H., & Hunt, S. (2018). Building a reliable evidence base: Legal challenges in environmental decision-making call for a more rigorous adoption of best practices in environmental modelling. *Environmental science & policy*, 52-62.

Percivall, G. (2010). The application of open standards to enhance the interoperability of. *International Journal of Digital Earth,*, 14–30.

Pons , X. (2020). *MiraMon: Geographical information system and remote sensing software.* Barcelona: Centre de Recerca Ecològica i Aplicacions Forestals.

Rojas , A., Athanasiou, S., Lehmann, J., & Hladky , D. (2013). Garcia-Rojas, A., Athanasiou, S., Lehmann, J., & Hladky, D. (2013). GeoKnow: leveraging geospatial data in the Web of data. Open Data on the Web. *Open Data Workshop, W3C*.

Ruiz-Mallén, I., & Gmelch, N. (2020, 07 21). *Responsible Research & Innovation for researchers. An introduction.* Retrieved from Responsible Research & Innovation for researchers. An introduction: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/76886/1/Responsible%20Research%20%26%20Innovation%20%28RRI%29%20for%20researchers.%20An%20introduction.pdf

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

Sanchez Maganto, A., Nogueras-Iso, J., & Ballari, D. (2008). *Normas sobre metadatos (IS019115, ISO19115-2, ISO19139, ISO 15836).*

Schade, S., & Smits, P. (2012). Why linked data should not lead to next generation SDI. *IEEE International Geoscience and Remote Sensing Symposium* , 2894-2897.

Schut, P. (2007). *OGC Web Processing Service (WPS), Version 1.0.0, OGC 05-007r7.* Consultat el 01 / 08 / 2020, a OGC: http://portal.opengeospatial.org/files/?artifact_id=24151

Simmhan, Y., Plale, B., & Gannon, D. (2005). *A survey of data provenance techniques.* Bloomington: Computer Science Department, Indiana University.

Spiekermann, R., Jolly, B., Herzig, A., Burleigh, T., & Medyckyj-Scott, D. (2019). Implementations of fine-grained automated data provenance to support transparent environmental modelling. *Environmental Modelling & Software*, 134-145.

Van den Brink, L., Janssen, P., Quak, W., & Stoter, J. (2017). Towards a high level of semantic harmonisation in the geospatial domain. *Computers, Environment and Urban Systems*, 233-242.

Veregin, H., & Lanter, D. (1995). Data-quality enhancement techniques in layer-based geographic information systems. *Computers, Environment and Urban Systems*, 23-26.

Vilches-Blázquez , L., Villazón-Terrazas, B., Corcho, C., & Gómez-Pérez , A. (2014). Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth*, 554-575.

vis.js. (2020). *GitHub vis.js*. Retrieved 2020, from https://github.com/visjs

W3C. (2013, 03 26). *SPARQL Query Language for RDF*. Retrieved 07 08, 2020, from SPARQL Query Language for RDF: https://www.w3.org/TR/rdf-sparql-query/

W3C. (2014, 02 05). *Resource Description Framework (RDF)*. Retrieved 07 08, 2020, from Resource Description Framework (RDF): https://www.w3.org/RDF/

W3C. (01 / 08 / 2016). *LinkedData wiki*. Consultat el 08 / 07 / 2020, a LinkedData wiki: https://www.w3.org/wiki/LinkedData

W3C. (07 / 07 / 2020). *Geospatial Semantic Web (GeoSemWeb) Wiki*. Consultat el 07 / 07 / 2020, a https://www.w3.org/community/geosemweb/wiki/Main_Page

Wacharamanotham, C., Subramanian, K., Borchers, j., & Völkel, S. (2015). Statsplorer: Guiding Novices in Statistical Analysis. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2693-2702). ACM.

Wang, S., Padmanabhan, A., Myers, J., & Tang, W. (2008). Towards provenance-aware geographic information systems. *SIGSPATIAL international conference on Advances in geographic information systems* (p. 1-4). ACM.

Wikipedia. (2020, 01 25). *e-Science*. Retrieved 07 20, 2020, from e-Science: https://en.wikipedia.org/wiki/E-Science

Yazici, I., Karabulut, E., & Aktas, M. (2018). A Data Provenance Visualization Approach. *14th International Conference on Semantics, Knowledge and Grids (SKG)* (p. 84-91). IEEE.

Yuan, J., Yue, P., Gong, J., & Zhang, M. (2013). A linked data approach for geospatial data provenance. *IEEE transactions on geoscience and remote sensing,*, 5105-5112.

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., Guo, X., Zhang, M., Jiang, L., & Zhai, X. (2016). Linked Data and SDI: The case on Web geoprocessing workflows. *ISPRS Journal of Photogrammetry and Remote Sensing*, 245-257.

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

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

Zabala, A., Masó, J., Bastin, L., & Bigali, L. (2013). Increasing dataset quality metadata presence: Quality focused metadata. *Inspire Conference.* Florence, Italy.

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*, 95-105.

Annexos

Annex I: Acrònims

Annex I: Acrònims

Aquesta és la llista d'acrònims que apareixen en la introducció, el resum de resultats i les conclusions d'aquesta tesi.

CNSA	Chinese National Space Administration
CSDGM	Content Standard for Digital Geospatial Metadata
CSW	Catalogue Service Web
DCMI	Dublin Core Metadata Initiative
ESA	European Space Agency
ESDS	Earth Science Data Systems Program (NASA)
EO	Eath Observation
FGDC	Federal Geographic Data Commite
GEO	Group on Earth Observations
GEOSS	Global Earth Observations System of Systems
GIS	Geographic Information System
GML	Geographic Markup Language
IDE	Infraestructura de Dades Espacials
ISO	International Organization for Standardization
ISO/TC 211	ISO Tecnnical Comitte - Geographic information/Geomatics
JAXA	Japan Aerospace Exploration Agency
NASA	National Aeronautics and Space Administration
OGC	Open Geospatial Consortium
OPM	Open Provenance Model
RDF	Resource Description Framework
SIG	Sistema d'Informacio Geografica
SQL	Structured Query Language
SOS	Sensor Observation Service
SPARQL	Protocol and RDF Query Language
TD	Teledetecció
TIG	Tecnologia de la Informació Geogràfica

URI	Uniform Resource Identifier
VGI	Volunteered geographic information
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service
WMTS	Web Map Tile Service
WPS	Web Processing Service
W3C PROV	W3C Provenance Data Model
XML	eXtensible Markup Language