INSTITUTE FOR GEOINFORMATICS (IFGI), UNIVERSITY OF MÜNSTER

Concepts towards a provenance-aware Web Processing Service

Bachelor Thesis

Matthias Hinz

17.07.2012

Matriculation Number: 350317

Adviser:

Benjamin Proß

Erklärung des / der Studierenden

Hiermit versichere ich, dass ich die vorliegende Arbeit mit dem Titel

....

....

selbständig verfasst habe, und dass ich keine anderen Quellen und Hilfsmittel als die angegebenen benutzt habe und dass die Stellen der Arbeit, die anderen Werken – auch elektronischen Medien – dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind.

______________________________	______________________________
Ort, Datum	Unterschrift

Abstract

Web technologies have made large contributions to geosciences and geoinformatics. A series of standardized web services has been developed to perform typical tasks of spatial data handling: Acquisition of data from distributed resources, processing of (probably massive) data sets and visualization of results [1, p. 820]. The Open Geospatial Consortium (OGC) [2]proposes the Web Processing Service (WPS) as a powerful and flexible processing unit. This service offers reusable processing components, ranging from simple spatial operations to complex models. Therefore it is a useful component in web service chains, designed to execute and automate geo-processing workflows. [1, p. 828]

Within such workflows it is important to capture data provenance, i.e. the processing history of geospatial data [3, p. 2], because it facilitates context information for comprehension or even reproduction of data analysis and its results. That helps error-tracking, evaluating process validity and data quality and supports users in further decisions.

In this context Yue et al. proposed the following approach:

> *"Provenance information can be captured by tracing the execution of the workflow engine or aggregating provenance information generated by distributed geospatial service providers" [4, p. 272].*

This bachelor thesis shall focus on the latter term, namely generation of provenance information by WPS. It shall conceptualize a WPS framework which facilitates automated recording and capturing of provenance information and thereby take care about specific properties of spatio-temporal data. Several use cases will be discussed, where provenance information is required when dealing with WPS. Conclusions will be made about which data can be derived, e.g. from process description- and execute-documents and which data shall be added to improve *provenance-awareness*.

Another issue is the choice of provenance representation. Therefore common approaches will be evaluated and involved in the conception. One such approach is the Open Provenance Model (OPM) [5], which seeks to be an interoperable provenance model. Former researches included an extension of OPM for geospatial domains [6]. Further, established metadata standards such as ISO 19115 will be taken into account.

Finally the conception shall be prototyped based on a single representative use case.

Table of Contents

1. Introduction

This bachelor thesis is about concepts. A concept is, according to the Marriam-Webster online dictionary either *"something conceived in the mind"* or *"an abstract or generic idea generalized from particular instances"* [7]. Hence, the first definition is about minds. I, the author of this paper, will describe things that I conceived in my mind, based on theoretical research and practical experience. These descriptions include introducing and analyzing concepts released by leading researchers in the field of provenance. Thus my descriptions will also reflect their minds. Chapter 2 makes clear, that provenance refers too many different understandings and manifold research fields, although many are complementary in some way and do not necessary disagree. Important is, that we can learn from them and use this derived knowledge to create our own ideas towards a specific purpose. That is how my thesis refers to the second definition of the Marriam-Webster dictionary.

Provenance-aware are applications that not only produce data, but also produce a description of their execution [8]. Hence they are explaining how outcomes of their execution were derived. This thesis envisions applying that kind of provenance-awareness to Web Processing Services (WPS). WPS is a web service standard, specified by the Open Geospatial Consortium (OGC). Distributing reusable processing components, WPS is an important component of service oriented Spatial Data Infrastructures (SDI). Typical geoprocessing workflows involve data acquisition, processing and visualization of results. In this matter, the OGC developed the OGC web service (OWS) - stack which allows different kinds services to be loosely coupled, forming up service chains [1]. Important parts of this web service family are Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Web Processing Service (WPS), Sensor Observation Service (SOS) and Catalogue Service for Web (CSW). Despite that GI Sciences are progressing in fields of Service Oriented Architectures and Cyberinfrastructures, there is still lack of concepts and solutions towards provenance-awareness [3]. Chapter 3 will discuss this topic in detail.

Main strains of provenance *provenance modeling* (section 2.2) and *data architectures for provenance management* (section 2.3) [9]. This thesis will address both, but provenance modeling will be more in focus, as questions about provenance management will gain importance when provenance-awareness is addressed in larger scale. Taverna [10] is a good example for an advancing provenance aware application for Life Sciences Workflows. The

Taverna Workflow Management System which integrates the domain-aware Janus Provenance System will be presented in section 2.4.

In summary and conclusion about both, chapters 2 and 3, a set of provenance requirements for WPS will be formulated in section 3.3. This will complete the theoretical part of work. Taking a first practical step towards a provenance-aware Web Processing Service, a prototype library for capturing provenance from WPS will be presented. Its use will be illustrated on two use cases derived from researches that were experienced during my student assistant work at Institute for Geoinformatics (Ifgi), Münster. The first use case will capture provenance of an inverse distance weighted interpolation, performed by WPS4R, the R backend of the 52°North WPS [11]. The second use case will capture the provenance of a simple service chain composing two WPS processes. It is derived from the Albatross Scenario workflow, which is part of the EU funded UncertWeb project [12].

In conclusion, this thesis aims to be a contribution to provenance research in GI Science.

2. Provenance on the Web

Among the most regarded current activities in the field of provenance are those of the W3C Provenance Working Group and its predecessor, the W3C Provenance Incubator Group [13] [14]. The Provenance Incubator Group declared

> *"a charter to provide a state-of-the art understanding and develop a roadmap in the area of provenance and possible recommendations for standardization efforts." [14]*

Their final report is a good starting point for a broad reflection about researches on capturing provenance in the web. It will be discussed below and in section 2.1. Present activities of the subsequent W3C working group on the new PROV model will be subjected in section 2.2 along with other popular approaches to model provenance.

Synonyms for the term *provenance* are *source* and *origin* [15]. Definitions and views about provenance go far apart. To get an idea about provenance in computer science, one might regard its working definition from the W3C:

> *"Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance." [14].*

Authors of the past EU Provenance Project (section 2.3) pointed out that provenance may be understood in two different ways. First, as a concept which denotes the source or derivation of an object, second, as it refers to a concrete record of such a derivation. In addition, the authors defined the *"provenance of a piece of data as the process that led to the peace of data."* [8, p. 2]

Another term closely related to provenance is lineage, which is the tracking of data transformations and processes [16]. In literature, the terms of "lineage" and "provenance" are often used synonymously, especially when discussing workflows. Different authors and communities might prefer the former or the latter terms. But a quote from Simon Miles might denote, that provenance is actually a term of broader use. Comparing two different understandings, he state that

> *"In particular, some describe the provenance of a resource in terms of attribution metadata, stating who created or modified it, and where and when, while others model provenance as a causal graph, in which occurrences trigger or influence each other, in the end leading up to the resource being as it is (the resource's lineage). [17, pp. 1-2]"*

Miles' paper proposed a mapping between the Open Provenance Model (OPM), which uses directed acyclical graphs (DAG) provenance modeling and Dublin Core (DC), which grounds on attribution metadata.

From this point, it is important to recognize that there are not just one or two provenance-communities, but manifold organizations, companies and communities spread around the globe whose doing research on provenance for individual purposes. That explains the occurrence of many competing provenance models and the lack of standardization, despite of attempts to create overarching provenance models and provenance infrastructures.

2.1 Requirements and Usage of Provenance

The incubator group collected 33 use cases of provenance and formed up three flagship scenarios (News Aggregator, Disease Outbreak and Contract Scenario). Thereafter they elaborated provenance Requirements, namely a report about "Requirements for Provenance on the Web" [18] along with a set of 140 technical requirements and "Provenance Requirements for the next Version of RDF" [19]. For the better understanding, conceptual terms mentioned in the incubator groups' final report will be formatted in italics from now on. In the report, group summed up provenance requirements in 17 dimensions and divided them into the categories *Content, Management* and *Use*.

These requirements are useful, e.g. for determining the scope of this thesis, provenance research or evaluating provenance models and applications. Scoping the thesis, the practical addresses dimensions of Content and Use, other dimensions are just theoretically considered theoretically and some were discarded. Definition, creation, provision and exchange of provenance information will be directly addressed, while provenance management is committed to provenance management systems, which are just considered in the theoretical work.

Requirements of Content

Reconsidering the W3Cs report, requirements of content address structures and attributes needed to capture provenance. They have to be conceptualized and applied. Concepts of importance are *object*, which describes the artifact that subjects statements about provenance. *Attribution* describes references to sources or entities that contributed to the artifacts creation. *Process* describes activities carried out to generate an artifact. Other content dimensions are *evolution and versioning* which addresses artifacts and *justification of decisions*.

Requirements of Management

Management requirements deal with *publication*, *access*, *dissemination control* and *scale* of provenance data. Thereby *dissemination control* refers to license information handling, access and use policies about artifacts and provenance information. In terms of management requirements, my work addresses the scale of provenance information, i.e. which data shall be captured and stored in what level of detail considering large amounts of data and complex workflows. A lot of application for provenance management, i.e. provenance stores or provenance services does already exist. Some will be presented in section 2.3 and 3.2.

Requirements of Use

Provenance information needs to be *understandable*. Understanding is an important requirement of usage and the prior condition for any research on provenance, because there is no use of any data, if meaning cannot be derived. Therefore it is important to provide appropriate visualizations and representation of provenance information. This includes different views and levels of abstraction as well as the ability to combine provenance information with domain-specific information. Section 2.4 will explicitly address domain-aware and semantic provenance models. Parts of this thesis also address provenance information within the geospatial domain. Leaving the era of monolithic applications, interoperability and ability to obtain provenance from heterogeneous systems becomes more and more important. Provenance information may be retrieved by queries across different resources. In this manner, it is useful to capture the provenance of provenance information, i.e. what sources contributed to specific provenance statements. Because provenance information needs to be exchanged and combined throughout different resources, a common provenance model or at least mappings between the models in use are important. Further dimensions stated by the W3C are comparison between artifacts regarding there provenance, accountability, trust, the possibilities to deal with imperfections of provenance records and the ability of provenance information to support debugging of processes, i.e. revealing errors. I want to point out trust and debugging. Crucial for *trust* is the ability of an application to explain processing outcomes and to proof reliability. Trust assessments are usage-dependent and subjective, but they are derived from provenance and other quality information. “Debugging” capabilities are especially important for processing workflows. Thereby provenance information can be used for searching error-prawn processes and inputs. In turn, if this information is useful to find errors, it is very useful for suspending them by doing outcome validation.

Further considerations

Further, provenance may be used to determine the impact of resources on the outcome and thereby support optimization of a workflow. Provenance information can be used to reproduce and update workflow results as well. Reproducibility is not explicitly addressed in the provenance-requirements chapter of the W3C, because it’s just a summary of their detailed work, but it supplies many of the requirement dimensions stated above, e.g. understandability, accountability, trust and evolution / versioning and debugging. Therefore it is a very desirable goal of provenance research.

The Following subsections dedicate to the two main strains of provenance research. Missier et al. [9] call them *provenance modeling* (section 2.2) and *data architectures for provenance management* (section 2.3). Section 2.3 will demonstrate how to publish provenance data and gain knowledge from them by technology means, i.e. by querying and visualizing.

Section 2.4 will focus on semantic-enabled, domain-aware provenance, which is explained on a case from the bioinformatics and life science domain.

2.2 Representing Provenance: Provenance Models

Three important provenance models shall be introduced in this section. The Open Provenance Model (OPM) [5], where the main focus of this thesis is set on, the W3C Prov model [13], which is relatively new at this time and the ISO Lineage Model [16], which presents an ISO compliant way to describe provenance. The latter two might be useful alternatives for OPM and may be applied to related or further work. Notably the W3C Provenance Incubator group analyzed many other popular Provenance Models and provided vocabulary mappings between them, using OPM as the reference model for each mapping [20].

2.2.1 The Open Provenance Model

The Open Provenance Model is amongst the most popular and established provenance models at the moment. It was achieved by a unique community effort motivated by the International Provenance and Annotation Workshop (IPAW) - series [21]. The first workshop was held in Chicago 2006 with 50 participants. Recently the series lead up to the fourth IPAW, which took part in Santa Barbara, California in the week of 12-18 June 2012, during the time when this thesis had been written. It was held along *"with a number of co-located events including a tutorial on the provenance specifications from the W3C, and meetings of DataONE and the W3C Provenance Working Group* [22]*"* [13] [23]. Community activities like this show the current dynamics and brisance of provenance research.

The OPM data model, OPM v1.00, was crafted in 2007 and in 2008 revised to OPM v1.01. It is intended to be open from an interoperability viewpoint but also with respect to the community of its contributors, reviewers and users. The latest OPM version is OPM 1.1, developed after June 2009. OPM specifies an XML-Schema (OPMX), a lightweight ontology vocabulary (OPMV), an OWL ontology which extends OPMV (OPMO) and a Java Library (OPM4J). Hence OPM implementations refer to the following three namespaces:

- opmx: http://openprovenance.org/model/opmx#
- opmo: http://openprovenance.org/model/opmo#
- opmv: http://purl.org/net/opmv/ns#

OPMO and OPMX are conceptualized a way, that they complement each other and that XML serializations can be converted to RDF and vice-versa. This is one of the core feature which addresses the needs of different communities, i.e. for the use of Semantic Web technologies (i.e. RDF-triple stores, SPARQL querying) through OPMO, linked data annotations trough OPMV and XML usage, e.g. for Web Service Communication through OPMX [5]. The following introduction of PROV in subsection 2.2.2 will show that this principle has been adopted by the W3C as well.

Going more into detail, OPM represents provenance as causal directed acyclic graphs. Edges are represented as arrows between two nodes. These arrows point from effect to cause; for example, if an artifact was generated by a process, the arc points to the process, because it caused the generation of the artifact. Certain types of nodes and edges are defined. Nodes can be an Artifacts, Processes or Agents. In this order, they are represented as ellipses, rectangular or hexagons. Edges refer to causal relations between nodes. They are typed by the following terminology: a Process *used* an Artifact, an Artifact *wasGeneratedBy* a Process, a Process *wasTriggeredBy* another Process, an Artifact *wasDerivedFrom* another Artifact *wasControlledBy* an Agent. These five relations between three types of nodes are shown in Fig. 1. Understanding them is crucial for understanding OPM graphs. [24]

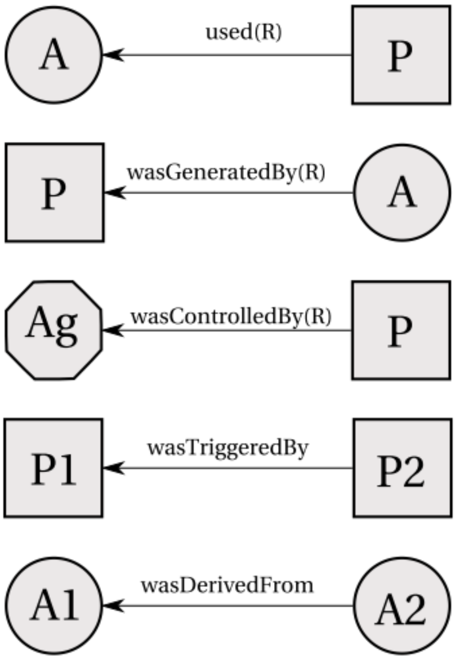


Fig. 1 Edges in the Open Provenance Model: sources are effects and destination causes. Graphic and description derived from the Open Provenance Model Core Specification (v 1.1) [24]

The editors of OPM considered weak notions to be appropriate, only to express the necessary causal dependencies between entities. Considering an open world assumption, they left interpretations open for additional, hidden factors which may have influenced an outcome. As the terminology indicates, OPM is purposed to describe what happened in the past, not what could happen in the future. OPM can be customized and extended to suite different purposes and domains. Elements of the graphs can be annotated with sets of properties. Even annotations themself can be annotated. Properties might not only declare attributes, but also subtypes of nodes, edges and other elements. OPMO and OPMV specify annotations for time and date, which can be used for process recording and workflow inferences and serve e.g. comprehension and validation of parallel activities (i.e. opmx:OTime, opmo:endTime, opmo:exactlyAt, opmo:noEarlierThan, opmo:noLaterThan, opmo:startTime, opmo:time). Roles may be assigned to certain edges (i.e. used, wasGeneratedBy, wasControlledBy) to distinguish the nature of the dependency when multiple edges are connected to a same process [24].

OPM Graphs can describe provenance in a scalable level of detail. The graph completion Rules imply the following behavior: The *wasDerivedFrom* notion between two artifacts may refer to a hidden process; the *wasTriggeredBy* notion between two processes may hide artifacts (e.g. a message or command), which were passed between them. The Karma visualization plugin for Cytoscape (Section 2.3) actually uses this completion rules for expanding and collapsing OPM graphs interactively. The notion of *accounts* in OPM enables provenance to be described within different contexts. In the practical work for example (chapter 4), this concept is applied to divide a WPS provenance records into internal and external views of WPS processing [24].

Simon Miles wrote a paper about “Mapping Attribution Metadata to the Open Provenance Modell” [17], which was already mentioned above. Its conclusions are enlightening for overthinking concepts of OPM and for relating provenance data to metadata in general. A mapping was purposed between Dublin Core (DC) and OPM. DC, in short, is a metadata model which emerged from the library and archiving community that describes resources and partially its provenance. Importantly, the concept of a DC resource is very different from that of an OPM artifact, though they might refer to similar content. This caused the work contributors several issues about the mapping:

> *"(i) DC refers to data as resources which may change over time, whereas OPM only models data artifacts at fixed instants; (ii) a single DC assertion generally does not correspond to a single causal relationship in OPM, so we have to map assertions to patterns in OPM graphs; (iii) some DC terms can refer to what occurred in the past or to the present, so do not map to provenance data in all cases (others do not describe provenance at all) [17, p. 4]".*

This section has described Open Provenance Model in detail to provide a conceptual understanding of the practical work in chapter 4 and about the decision to focus on OPM. Many concepts explained here will be directly and indirectly referred to within the following chapters.

2.2.2 The W3C PROV Model

The PROV data model represents the current work of the W3C provenance working group [13]. It is not yet shown, if it will be adopted by users. But what makes itself unique is that it is based on intensive work of the W3C Provenance Incubator Group [14], which was a team of experts spending much work on analyzing and comparing former research on provenance. The design of this model may benefit from former efforts, including those of the Open Provenance Model.

The PROV Data Model (PROV-DM), in short, contains of six components: (1) entities and activities, (2) agents bearing responsibility of entities and activities, (3) derivations of entities from entities, (4) properties that link entities referring to the same thing, (5) collections and (6) an annotation mechanism.

Some of those concepts remind of OPM, but it requires a deeper understanding of both models to conclude to which extend they map each other and to my knowledge, formal PROV-OPM mappings had not been published yet. Just to outline a comparison: instead of artifacts and processes, PROV uses the notions of entities and activities. Agents in OPM control processes, while agents in PROV bear responsibility for artifacts and activities. Both models have some annotation mechanism to increase expressiveness and to create domain-awareness. The framework presented in chapter 4 refers to OPM, but it may be considered to extend it to alternative models like PROV, where these considerations are useful.

The PROV family of specifications also addresses users of different communities. Similar to OPM, there is an XML schema specification (PROV-XML) and an ontology specification (PROV-O), other documents deal with access and query mechanisms (PROV-AQ), formal semantics (PROV-SEM) and many more.

2.2.3 ISO Lineage and ISO Metadata

The International Standardization Organization (ISO) provides a model for describing process lineage, which is proposed in the ISO specifications for geographic information - metadata. The LI_Lineage model has been specified in ISO 19115 and was extended by LE_Lineage in ISO 19115-2. Implementations are available in XML. ISO lineage conceptualizes *"sources which are either used or produced in a series of process steps. This model can be helpful in many cases despite its simplicity. Sources and process steps are linked together to describe the lineage of a resource."* Each process step is associated with processing and algorithm descriptions, while each source can be associated with any number of process steps and each process step can be associated with any number of sources [16]. These concepts are shown in Fig. 2. The ISO lineage model had been defined in UML, showing all attributions and relationships between the elements (Fig. 3).

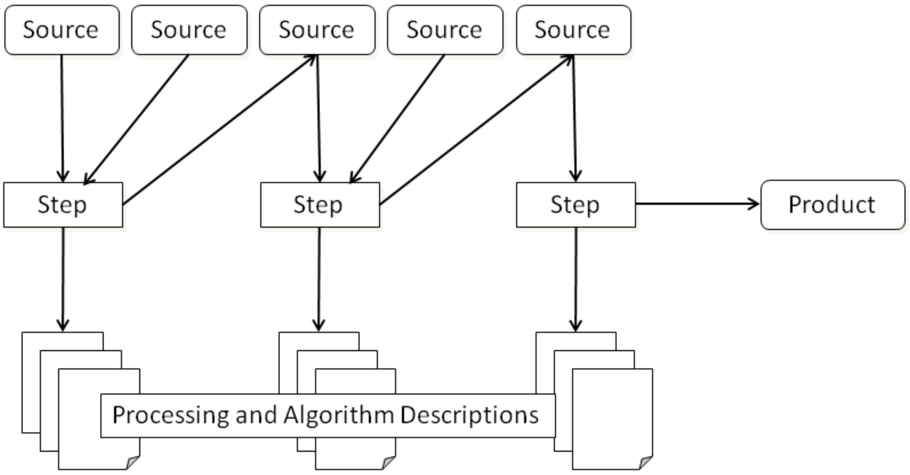


Fig. 2 Overview of the ISO Lineage model [Online] Available: https://geo-ide.noaa.gov/wiki/index.php?title=File:ISOLineageModel.png [Accessed 10 July 2012]

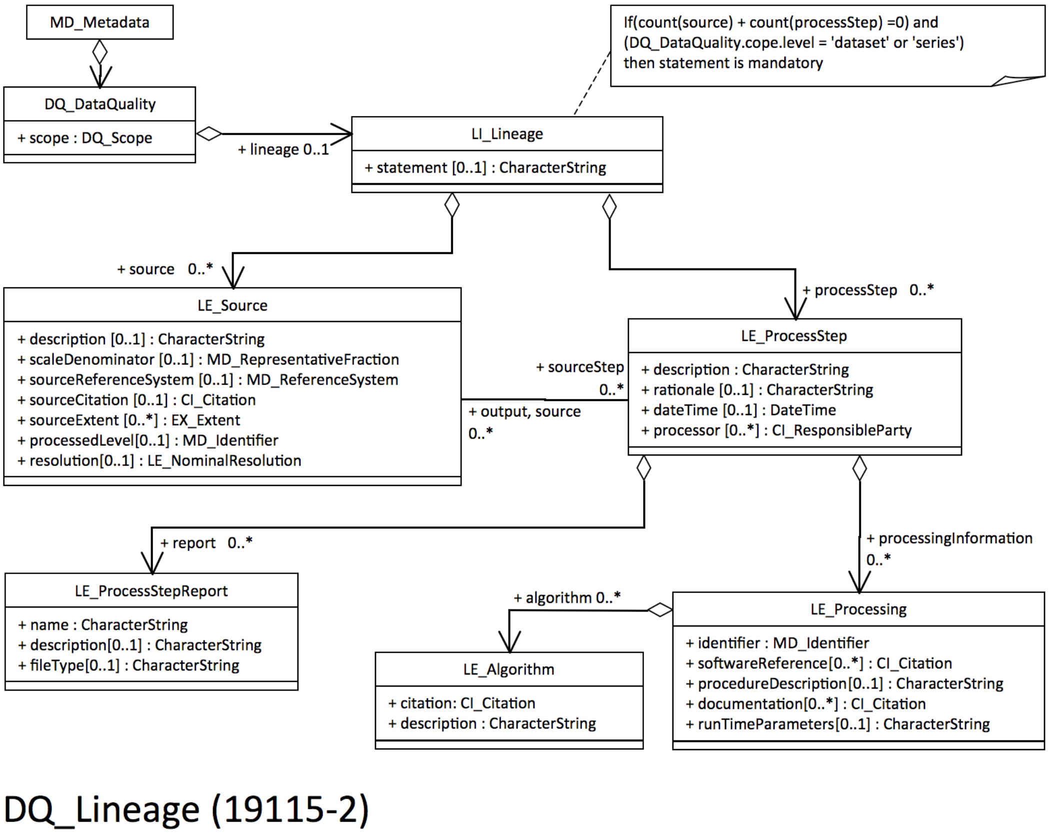

MD_Metadata
DQ_DataQuality
+ scope : DQ_Scope
+ lineage 0..1
LI_Lineage
+ statement [0..1] : CharacterString
If(count(source) + count(processStep) =0) and
(DQ_DataQuality.cope.level = 'dataset' or 'series')
then statement is mandatory
+ source 0..*
LE_Source
+ description [0..1] : CharacterString
+ scaleDenominator [0..1] : MD_RepresentativeFraction
+ sourceReferenceSystem [0..1] : MD_ReferenceSystem
+ sourceCitation [0..1] : CI_Citation
+ sourceExtent [0..*] : EX_Extent
+ processedLevel[0..1] : MD_Identifier
+ resolution[0..1] : LE_NominalResolution
+ processStep 0..*
LE_ProcessStep
+ description : CharacterString
+ rationale [0..1] : CharacterString
+ dateTime [0..1] : DateTime
+ processor [0..*] : CI_ResponsibleParty
+ sourceStep
0..*
+ output, source
0..*
+ report 0..*
LE_ProcessStepReport
+ name : CharacterString
+ description[0..1] : CharacterString
+ fileType[0..1] : CharacterString
+ processingInformation
0..*
+ algorithm 0..*
LE_Algorithm
+ citation: CI_Citation
+ description : CharacterString
LE_Processing
+ identifier : MD_Identifier
+ softwareReference[0..*] : CI_Citation
+ procedureDescription[0..1] : CharacterString
+ documentation[0..*] : CI_Citation
+ runTimeParameters[0..1] : CharacterString
DQ_Lineage (19115-2)

Fig. 3 ISO 19115-2 Lineage UML [Online] Available: https://geo-ide.noaa.gov/wiki/index.php?title=File:LI_Lineage-2.png [Accessed 10 July 2012]

Although this model has not been in use during this thesis' practical work, it has been inspired the concepts and software architecture. Like it was mentioned for W3C PROV, extensions and mappings for this model have to be considered. ISO lineage is certainly the first choice for implementations compliant to international geographic metadata standards, but even though other criteria might dominate the choice of data model, referring to ISO standards is crucial for interoperability. An approach to bridge this gap may be to annotate provenance graphs with ISO conform terminology and to use semantic annotations which refer to ISO metadata standards. Related work on semantic annotations has been explained by Maué [25].

2.3 Managing and Understanding Provenance: Related Technologies

The final public report of the EU Provenance Project was published in December 2006 with under the headline "Enabling and Supporting Provenance in Grids for Complex Problems [26]". The project team developed a provenance architecture of industrial strength that facilitates so called provenance-aware application within grid environments. The idea behind is, that applications in general produce data, but they do not explain sufficiently, how derived it. In contrast, provenance-aware applications produce descriptions of their executions during their run. These process documentations have to be managed and stored. Therefore the project members developed a *provenance store*, which is central to their architecture. This store basically contains of a recording interface for collecting data from applications and a query interface for retrieval of process documentations as well as a particular data item's provenance. The recording interface allows stateless and asynchronous out-of-order recording by actors. Provenance stores preserve provenance data is preserved in its original form, which means that data cannot be modified or deleted. The provenance architecture is grounded on the self-introduced p-assertions data model but since it does not relate to this paper's considerations it won't be further discussed. Important about this work is the purpose of provenance architecture influenced by service oriented architecture-style, which enables services or actors to exchange provenance information on the web. Project members considered issues and requirements about *scalability*, *security* and *management.* Conclusions about these fundamental insights may be to integrate a provenance-aware Web Processing Service or a provenance-aware WPS client application in such an infrastructure. [8]

Fig. 4 illustrates the scope of the EU Provenance Project in a simplified form. Achievements of the EU Provenance Project are open to the public, including implementations of the corresponding Provenance Store Service, the Provenance Tools suite and the Provenance Client Side Library. [27]

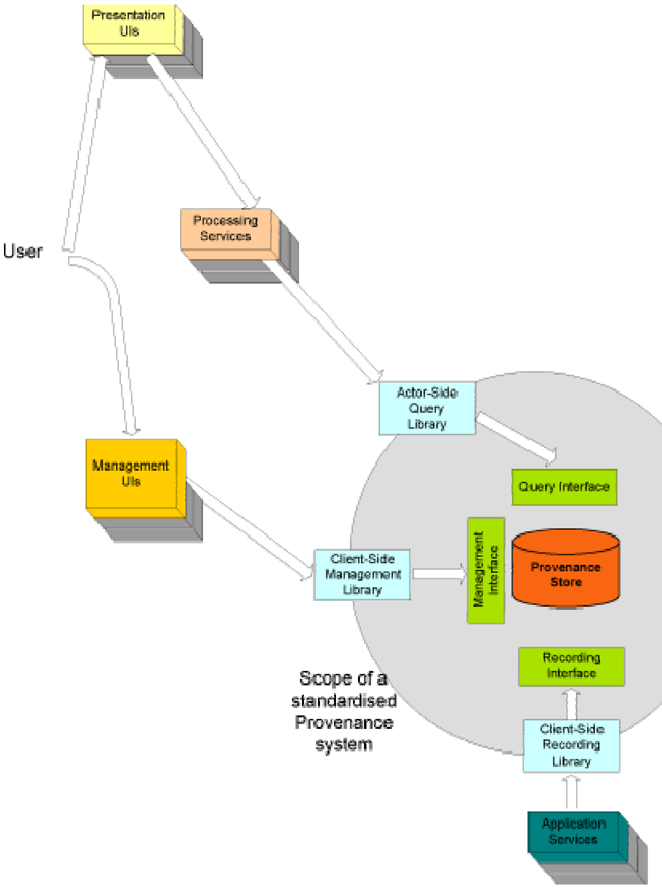


Fig. 4 Scope of the specification effort from the EU Provenance Project. [Online] Available: http://www.gridprovenance.org/openSpecification/ [Accessed 11 July 2012]

Another application which refers to provenance in this context is the Karma provenance collection tool from the Indiana University, USA [28]. It relates closer to the framework presented by this thesis, because it supports the Open Provenance Model. It contains of the Karma provenance service, which is provides functionalities similar to the mentioned provenance store, namely to collect and query over provenance data and to store it in XML format.

The Karma Toolkit further contains a visualization plugin for Cytoscape [29]. Cytoscape is a desktop open source platform for complex network visualization and analysis, originally developed for research in biological science. It is a very powerful tool for visualizing large graphs, including various mechanism and algorithm to browse, to organize and to edit them. This Karma plugin enables Cytoscape to retrieve provenance data from Karma web services and to visualize OPM graphs. The latter functionality had been in use during the practical work of this thesis. As mentioned before (section 2.2.1), the visualization plugin uses OPM completion rules to expand and collapse provenance graphs interactively. Thereby it provides views in a customized level of detail [30].

Another way to visualize OPM graphs, which had been in use during the practical work too, comes from the OPM core framework. The java library contained in OPM toolbox, called OPM4J, couples with the Graphviz graph visualization software [31]. Graphiz contains of a command line tool called dot, which can visualize directed graphs based on the dot-language. OPM4J can translate OPM terms into dot-language and serializes OPM graphs as dot-files. Afterwards, based on command line requests, OPM4J can retrieve graph visualizations in many graphics formats like png, jpg, pdf. This concept is very useful because it is a low-level programming approach which can be integrated within various applications (including WPS) to visualize provenance without direct user interaction [5].

Since many provenance models contain of OWL and RDF based implementations, it is evident that provenance data can be managed and handled by semantic web technologies as well. Therefore a vast range of RDF triple stores can be used, as some were compiled by the W3C [32]. Yue et al [4] used in an approach to retrieve provenance data from geoprocessing workflows, the Joseki SPARQL server [33], which builds on the popular Apache Jena Framework to publish provenance data on the web. This approach will be further discussed in chapter 3.2, but for now, this concludes that SPARQL servers as Joseki can be used as provenance services too, though they were not necessary designed for it.

Semantic web technologies are also capable for querying and visualizing provenance data. Many desktop applications, such as Protégé [34] or TopBraid Composer [35] are capable to query and visualize RDF graphs but also on the web there are some related tools, e.g. the Sindice Web Data Inspector [36]. All these applications mentioned ground on W3C specifications of the Recourse Description Framework (RDF), Web Ontologie Language (OWL) and the SPARQL query language.

2.4 Domain-aware and Semantic Provenance Models

Many approaches for making scientific research reproducible and recording provenance information lack of broadness and interoperability. For example, paper laboratory notebooks served this purpose for hundreds of years, but they finally have been replaced by Laboratory Information Management Systems (LIMS), databases and electronic notebooks. Despite that, according to the W3C provenance incubator group [14], these systems didn’t get adopted easily and they have significant limitations, particularly as community-scale infrastructures:

> *“Tracking provenance across laboratories, sharing data with provenance, and integrating data from multiple sources remains very labor intensive today. Provenance records for reference information and community data sets are either produced by hand […], produced by ad hoc applications developed for a given kind of data, or not produced at all.” [14]*

On the other hand, provenance models which are proposed to be broad and interoperable are not necessarily ready to get adopted by a particular community, because they are *domain-agnostic* in its core specification. They describe causal relationships amongst data products, but they are not applicable for capturing domain-specific information and answering domain-specific questions. Luckily it is possible to extend certain provenance models and create *domain-awareness* thereby. This is often done by subtyping, annotating and adding vocabulary terms, so that the original semantics stay unchanged on higher level. Hence, it is not sufficient to include only domain-specific information in provenance reports, but also domain-specific semantics. [9]

A showcase example is Janus, which is a domain-aware extension of the domain-agnostic Provenir Ontology. It was made to capture data provenance of Life Sciences workflows and is implemented as a Provenance Model for Taverna, which is an open source and domain-independent Workflow Management System [10]. Missier et al explained their strategy and concepts in detail within the paper “Janus: from Workflows to Semantic Provenance and Linked Open Data” [9]. Technically, they modeled provenance records as RDF graphs. Considering provenance graphs in general, they stated that domain domain-aware graphs are useful to answer a broader class of questions than their domain-agnostic counterparts. Regarding previous work on providing semantic annotations to provenance logs by post-processing, they argued that this approaches lacked of a clear data model. These were the primary intentions for designing a domain-aware semantic provenance model.

Important is, that the Janus-developers assigned domain-specific terminology elements of the provenance graphs. Basically this is done by subtyping through annotations in RDF. This enables researchers to uses the same well-known terminologies within queries, which means, e.g. they can ask direct questions about chromosomes, because it was made explicit, that a certain resource has the type *so:chromosome*.

Taking it a step further, they wanted to ground the model within the semantic web and integrate provenance graphs in the web of data, to derive further knowledge from external data sources, specifically those which are compliant with the Linked Open Data (LOD) project.

That means, resources, which are present in provenance records, whether they are inputs, outputs, intermediate or intermediate results, might be present within other knowledge-bases as well. They demonstrated it on an example workflow, dealing with the analysis of mouse genomes. The key point is that the genes occurring in the workflow are registered Bio2RDF dataset as well. They were able to do a mapping between them and by having defined semantics, they managed to query over both, the provenance graph and the BIO2RDF dataset. Therefore they combined information from both data resources to answer complex questions about genes.

The implementation strategy also brings some illuminating insights. The Janus developers reused four publicly available ontologies from Life Sciences and the OWL Time ontology from the W3C. This enables Interoperability of their model for large public datasets using the same terminology. The annotation framework relies on *manual* annotations within Taverna workflows. By default, Taverna produces domain-agnostic content for each workflow, but as users annotate input and output ports as well as processors with domain-aware terms, the annotations get automatically propagated to the graphs. In RDF, this is done by some additional statements, which carry over to each of relevant values through some inference rules. Chapter 3.2 of the source paper explains this in detail [9].

That Janus semantics relies on manual annotations is a disadvantage. However, according to the authors it is realistically to annotate Taverna workflows this way, because they typically contain only a hand full of services. In the long run, they plan to retrieve annotations from web service registries such as the BioCatalogue registry for Life Science Web Services [37] [38], which will further contribute to automation of provenance recording.

The Taverna website [10] informs about their current activities in the field of provenance. These include the development of a whole provenance management infrastructure [39] (see section 2.3), involving database backends (MySQL and Derby), a provenance client API in Java and a provenance query language.

In conclusion, Taverna and Janus provide a truly reusable concept of provenance, which is capable to be extended within the GI Science domain as well. Still, works on Taverna provenance are in progress and still some features are in experimental status, but meanwhile Taverna exports OPM graphs as well. Web Processing Services had been extended to support WSDL and REST, which makes them compatible with the Taverna workbench. Some Taverna workflows involving WPS and other geospatial Web services were already published on the myExperiement online platform, so enthusiastic researchers in this field may be very welcome.

3. Provenance in GI Science

In 2009, Yue et al. [2] published a review about "Geospatial Data Provenance in Cyberinfrastrucure". They stated that *"although GIS is one of the earliest applications domain to conduct the data provenance research, little work has been done to investigate geospatial data provenance in the emerging Cyberinfrastructure* [3, p. 1]*"*. Still there are large technology gaps within the GI Science domain, especially in web environments, but current dynamics shall not be underestimated. Therefore, section 3.1 will explore the state of the art of provenance in GI Science, focusing on web-enabled approaches. Section 3.2 will focus on provenance in geoprocessing workflows. Section 3.3 leads up from theoretical insights to direct conclusions about requirements for a provenance-aware Web Processing Service.

3.1 Introduction and State of the Art

Defining geospatial data provenance as the processing history of a geospatial data product, Yue et al. [3] stated that

> *"examples of possible content in the processing history includes metadata descriptions of source data, transformation functionalities (e.g. geo-processing services), geo-processing workflow (e.g. geospatial service chaining), parameters used, intermediate geospatial data product, and date and time. [3, p. 2]"*

going through recent achievements of provenance research within the geospatial domain, he mentioned the ISO 19115 specification for geographic information – metadata (Chapter 2.2.3), OGC standards like the Sensor Model Language (SensorML), which can provide descriptions of the process by which an observation is obtained, and the "lineage"-element within a WPS execute-request, which demands for a complete copy of input parameter values as received in the execute request. To add a comment on the latter statement, the lineage component might be specified within the WPS standard, but in practice it still lacks of proper implementations based on a sufficient data model. Amongst other provenance-aware application and researches, Yue et al. mentioned the work of Wang et al. [6], published in the paper "Towards Provenance-Aware Geographic Information Systems." To add another comment, special about this work in context of this thesis is, that Wang et al.

proposed a three-tier service-oriented architecture for the provenance management component of GIS applications and that it represents geospatial data provenance in OPM. Further, Yue et al. mentioned the work of on a proposal of metadata tacking in geospatial service composition to instantiate a geo-processing model into an executable service chain [40], where another related publication appeared later [4].

Another relevant publication made by Yue et al. was about "Sharing geospatial provenance in a service-oriented environment" [41]. Stash et al. [42] presented work about Provenance in Observation Data and proposed a mapping between Linked Sensor Data and OPM. They also mentioned related work and amongst others an introduction of a provenance aware virtual sensor system based upon the OPM, done by Liu et al. [43]

Finally I want to mention the present NASA funded InstantKarma project [44], which builds upon the Karma toolkit discussed in chapter 2.3. The purpose of this project is to improve the usage of provenance information within the NASA Earth Science community. They adhere to OPM as well. Karma will be customized and integrated in the NASA data production by collecting and disseminating provenance of Advanced Microwave Scanning Radiometer - Earth Observing (AMSR-E) standard data products, while initially focusing on Sea Ice.

3.2 Provenance in Geoprocessing Workflows

Yue et al. stated that "Constructing distributed and complex geoprocessing workflows is also a special attention in Cyberinfrastructure, since it can help to enhance a data-rich geoscientific research environment to an analysis-rich environment" [3].

Spatial Data Infrastructures (SDI) on the web provide distributed geospatial data, either in original or value-added form. Value-added data products may be the outcome of single process steps performed by a particular service or the outcome of complex analysis archived by complex scientific models and extensive service chaining. More precisely said:

> *"The value-adding component in spatial data handling is the acquisition of information through processing and concatenation of data. This procedure involves the acquisition of problem-specific data, the application of specific computations (e.g. spatial intersection, spatial buffering, etc.), and the visualization of results (usually as maps or map-like presentations). It turns data into information [1, p. 1]"*

The most successful standardized approach to archive such architecture has been archived by the Open Geospatial Consortium (OGC) throughout the OGC Web Service (OWS) – stack.

Reconsidering the quote above, the value adding component contains of three functionalities: data acquisition, processing and visualization. Different OWS may be the primary choice for each of these tasks, for example sensor observations may be acquired by Sensor Observation Services (SOS), processing tasks may be done by Web Processing Services (WPS) and map visualizations may be done by Web Map Services. Since many analyses require multiple tasks in each category, e.g. acquiring different kinds of data from distributed sources, it gets clear, that most geoprocessing workflows require chaining of multiple services of different kinds, which was made possible by standardizing these interoperable, self-describing web services. [1]

Provenance tracking can help to understand these service chains and evaluate their results. Furthermore it can even facilitate automated service composition, as Yue et al. [4] demonstrated in a metadata-tracking approach, using semantic web technologies. Further benefits may be derived from the provenance requirements discussed in chapter 2, e.g. considering optimization and debugging facilities and reproducibility. Though this thesis focusses on Web Processing Service, it becomes clear that provenance capturing functionalities may cover the whole infrastructure in the long run. This case is not only about provenance-aware applications anymore, but about provenance-aware data infrastructures.

Yue et al. [41] took a concrete step towards that idea. They proposed provenance data management in an OGC and ISO compliant way. The suggested infrastructure is displayed in Fig. 5. Essentially, it propagates a geospatial provenance service as an extension of a geospatial catalogue service, which refers to the OGC Web Catalogue Service (CSW) standard. The presented catalogue service shares the same capabilities as WCS to store metadata, but is customized to tread provenance data as a special kind of metadata. Besides, the general information within the infrastructure, based on GML, SensorML, and other metadata, shall be extended with provenance data. That way, provenance information can be wide shared with open access within a Spatial Data infrastructure.

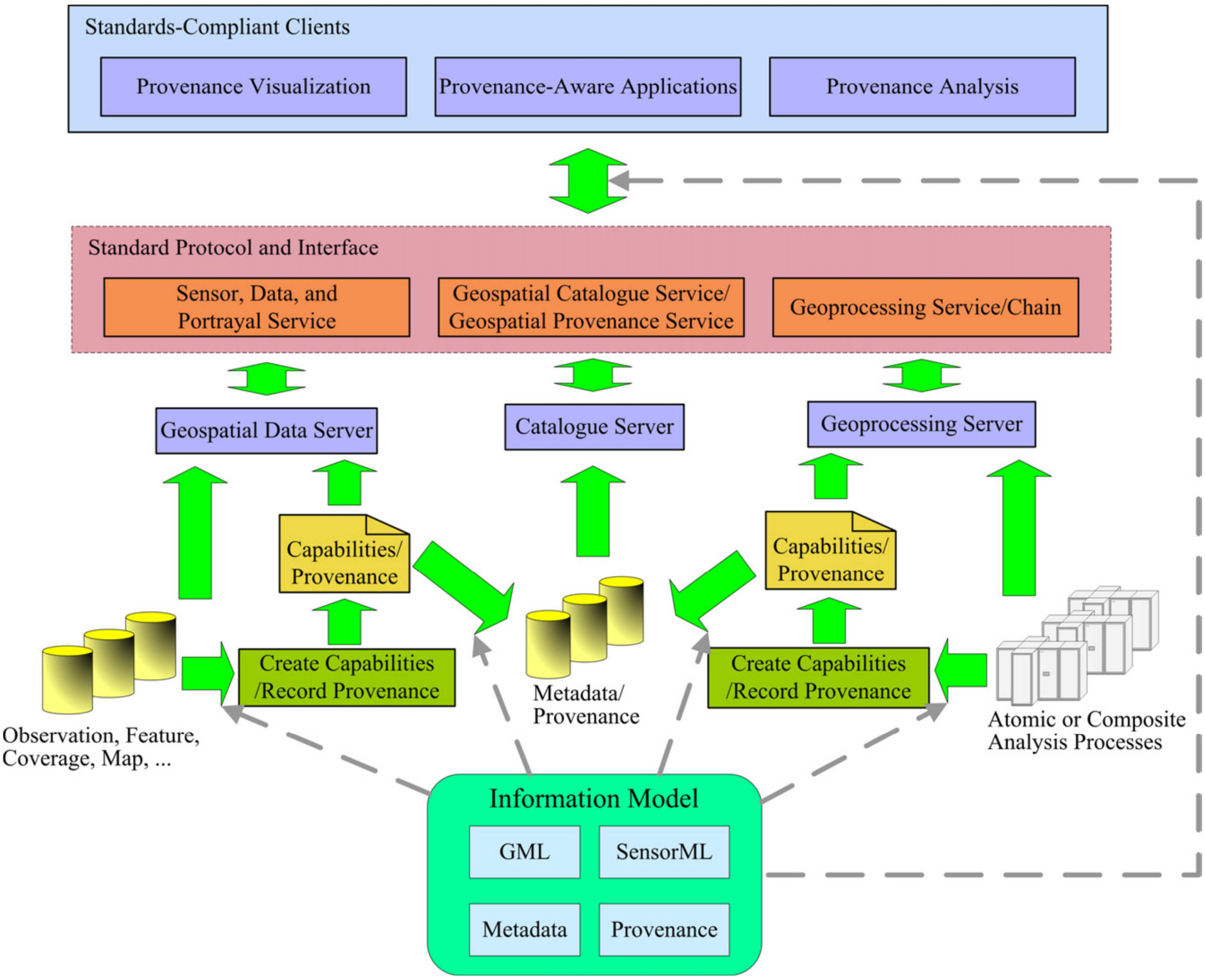

Standards-Compliant Clients
Provenance Visualization
Provenance-Aware Applications
Provenance Analysis
Standard Protocol and Interface
Sensor, Data, and Portrayal Service
Geospatial Catalogue Service/ Geospatial Provenance Service
Geoprocessing Service/Chain
Geospatial Data Server
Catalogue Server
Geoprocessing Server
Capabilities/ Provenance
Capabilities/ Provenance
Create Capabilities /Record Provenance
Create Capabilities /Record Provenance
Observation, Feature, Coverage, Map, ...
Metadata/ Provenance
Atomic or Composite Analysis Processes
Information Model
GML
SensorML
Metadata
Provenance

Fig. 5 The architecture to support geospatial provenance service. The figure shows how geospatial provenance service can be provided in the current service-oriented GIS. Graphic and description derived from Yue et al. [41]

These environmental considerations are important for conceptualizing a provenance-aware Web Processing Service. Provenance information can be either generated from workflow engines or from distributed Service providers [4]. Hence it is useful to enable sharing, exchanging and combining of provenance information. Regarding the typical way how OGC web services are chained, called web service orchestration, Kiehle et al. [1] (2007) listed up three different approaches:

> *"The orchestration of web services can be archived manually, by feeding the output of a web service as input into another web service, semi-automatically by defining the sequence of web service interactions in a configuration file, or fully automatically by providing capabilities to establish a self-organizing net. In contrast to the choreography of web services, where each service knows his successor and ancestor, web service orchestration leaves the services loosely coupled* [1]*"*

This principle of loosely coupled services may bear additional challenges for provenance tracing, i.e. the input of a Web Processing Service may be concrete references to (spatial) data, but also references to ancestor services providing the data, e.g. WMS, WFS, WCS or WPS providing value-added data. A WPS process may accept many supported input formats and -types for the same process and determines necessary data handling, i.e. parsers and generators at runtime. Capturing the provenance-dependent nature of inputs, outputs and intermediate results within a may require different complementary approaches.

3.3 Requirements for a provenance-aware Web Processing Service

This section concludes about the discussions within former chapters and sums up requirements for a WPS provenance framework. The listing is complementary to the general provenance requirements discussed in chapter 2.1. It shall not be exhaustive, nor may each requirement be mandatory, but it shall bring up key considerations towards a provenance-aware Web Processing Service.

Domain awareness requirements

Domain awareness has been discussed in chapter 2.4. Many principles described in the example about *Janus* [9] and *Taverna* [39] can be applied to WPS provenance too. Crucial to this approach is a proper annotation framework, using domain-specific terminology and domain-specific semantics to capture the true nature of provenance. The strategy of vocabulary and ontology reuse enabled the combination provenance data and data from external knowledge bases within SPARQL queries, but it may have more benefits as well. Subtyping of nodes and edges within provenance graphs may enable expressive visualizations and queries.

To outline this idea, a Semantic Web approach may be to annotate inputs and outputs of provenance records with datatypes, mimetypes, and maybe case-specific types. By referring these types to proper ontologies, reasoning may enable queries over interconnected domain-specific terms like *Literal Data Type*, *Number*, *Complex Data Type, Rasterdata*, *Vectordata*, *Points*, *Polygons*, *Geospatial Data Type*, *Spatio-Temporal Data Type* and more.

Interoperability requirements

Interoperability is a broad term with many aspects, but I want to point out that the choice of data model and its compatibility with related technology is important. Three selected models were discussed in chapter 2.2. It turned out that OPM and PROV a broad, flexible and extendable models seeking interoperability with many domains and communities. They might emerge to standards within the general eScience domain. An alternative way to represent provenance is the ISO lineage model, which addresses standard compliance within the Geoscience and GI Science domain. However, technical mappings between the models may compromise advantages and disadvantages committed by the choice of data model.

Standard compliance requirements

Standard compliance is an aspect very related to interoperability, because achieving interoperability and compatibility between technologies is easier when developers agree on certain standards, as shown in the case of Spatial Data Infrastructures of chapter 3.2. Same applies exchange and combination of data. Important standards within the GI Science domain were established by OGC and ISO. Seeking compliance with ISO metadata standards is an argument which pledges for adoption of the ISO lineage model [16], specified within 19115 and ISO 19115-2, but even though different provenance model may be in use, they should refer to metadata standards by using common terminology, using semantic annotations or adopting XML patterns (chapter 2.2). WPS may handle provenance data in an OGC compliant way by declaring provenance process input and process output or by using the lineage-element specified for execute request and –response to generate provenance information. Declaring provenance as and input makes sense, because if the provenance of input data is acquired by WPS, the WPS may internally complete and combine existing provenance graphs to model a more informative complete graph about a workflow execution.

Data management requirements

It shall be considered integrate the WPS provenance framework into a proper system architecture where provenance data is managed and handled. Those architectures were presented in chapter 2.3 for the general eScience domain when considering the EU Provenance project and the Karma toolkit. An approach to create such architecture in an OGC and ISO compliant way has been presented in chapter 3.2. Recording, sharing, exchange and combination of provenance data shall be considered. This may involve a provenance store, which might be realized as provenance service.

Managing provenance data requires decisions about what to store and for how long. As stated by Yue et al. [3], given the storage of intermediate results and many workflows reruns, provenance data might become very large and outsize the data products. Hence time period management of storage may be necessary.

Storing intermediate results in general might be not an option for larger projects, as scientific analysis becomes data-intensive. Indeed, so called dataflow applications fed by sensor networks may be composed of hundreds of services and reach gigabytes to terabytes in a single run [45]. Hence provenance data management shall address granularity and the level of detail in which data is stored.

Transparency and understanding requirements

WPS is essentially a web service which distributes reusable processing components, ranging from simple computational operations (e.g. buffering or intersection of polygons) to highly complex scientific models [1]. Thus WPS can be seen as a black box which might bear complex processes which need to be explained to the user. Therefore process metadata such as the process description and website documentations are helpful, but provenance data can provide supply this understanding. The 52°North WPS [46] for example uses different processing backends, i.e. the Grass Backend, ArcGIS Server Backend, Sextante Backend and WPS4R. These backends may refer to different software versions and point to different external services. WPS instances may have different behavior dependent on how they were configured and customized. Different IO handlers, i.e. parser and generators may be in use for different inputs, outputs and processes. Understanding of WPS may benefit from modeling those internal components and operation in provenance data, referring to corresponding online documentations and metadata. More data to include are responsible parties and originators for each component, license information and restrictions of use, as far as available.

Because users may be interested in different aspects of provenance information and different levels of detail, I suggest dividing the data into different views. OPM provides the notion of different accounts. Therefore WPS internals may be recorded in a separate account that delimits these occurrences from all others, even though they can occur in the same graph.

4. Prototype Development

Achievement of the prototype development shall be a core WPS provenance library written in Java. This library shall be extendable for different usages within WPS provenance and provides basic software architecture. Two use cases will demonstrate how the library can be applied. The first use case deals with the 52°North WPS, which performs an inverse distance weighted interpolation that grounds on the WPS4R backend [11]. The use case is aimed to capture provenance related metadata and WPS internals. The second use case deals with a simple service chain of two processes. It is derived from the Albatross workflow, which is part of a scenario from the EU funded UncertWeb project [12].

4.1 Software Architecture

The architecture conceptualized in a way that it can integrate in a WPS as well as in a WPS client or workflow engine to trace and record WPS provenance. The architecture is conceptualized in a way that provenance related metadata is traced from execute-requests and process-descriptions, other data may be recorded from processing components or manually assigned during the integration. This data is mapped to OPM, but the mapping components and OPM helper classes were separated from the provenance tracing and recording components. Both components aim to be loosely coupled to allow mappings to other provenance models if the code will be extended.

The library builds upon the OPMToolbox [5], wich is used to produce serialize and visualize OPM graphs. Java libraries related to the 52°North WPS has been integrated as well, for the purpose of handling WPS related XML documents and mapping metadata. Eclipse Indigo was used as development environment. Development tools were Apache Maven [47], used for Java dependency management and UML Lab [48], used for roundtrip-engineering and for creating class diagrams. Provenance graphs have been visualized by Graphviz [29] and Cytoscape [29]. Cytoscape was enhanced with the Karma visualization plugin therefore [28].

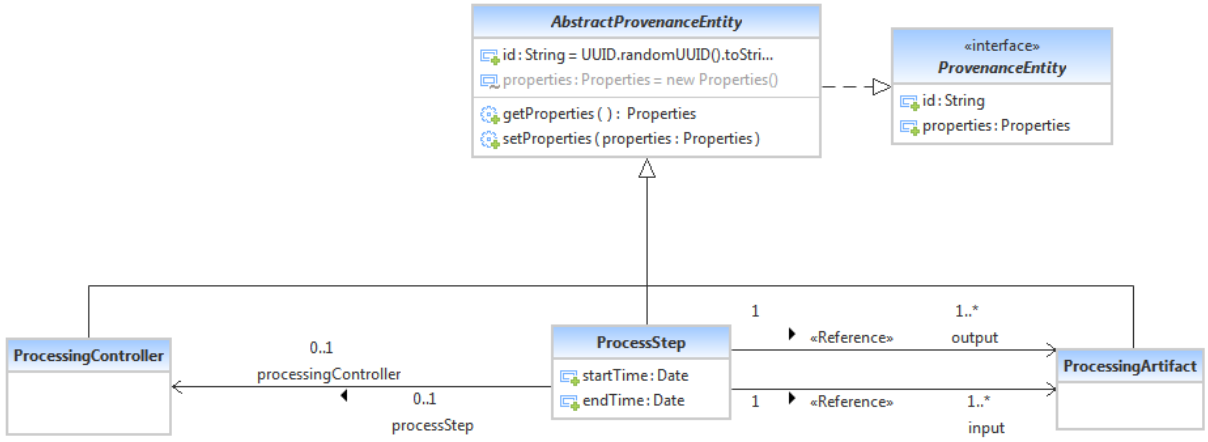


Fig. 6 Top-level class architecture of the wps provenance library, created by UML Lab

Fig. 6 displays the top-level architecture of the library. Different classes are meant to encapsulate provenance data, each one of type *ProvenanceEntity,* containing the attribute fields *id* and *properties*. The Properties-Object contains of hash maps where key and values are both strings. Keys are either property namespaces or string-representations that refer to *ProvPropertyType*-instances. *ProvPropertyType* is an enumeration type that defines semantic namespaces and labels assigned to properties. *ProcessingController*, *ProcessingStep* and *ProcesingArtifact* are classes which correspond with the OPM graph nodes *Agent*, *Process* and *Artifact*. *ProcessingStep* specifies the attributes *startTime* and *endTime*. They are internally converted to semantic properties that refer to the OPM ontology namespaces *opmo:startTime* end *opmo:endTime*. Attributes of other classes are handled similarly, either being converted directly or being disjoint for extracting properties, which is case for the process description in example.

The three concrete top-level-classes are suptyped according their meaning, i.e. *ProcessingController* is extended by *Parser*, *Generator*, *Service*, and *Backend*. *ProcessingStep* is extended by *ProcessExecution*, referring to the WPS-process execution as whole. But a *ProcessExecution* might refer to additional *ProcessingSteps* which may be internal *processingSteps* triggered by the process execution. They might be controlled by specific parsers, backends, etc. The framework is completed by several helper classes: *OPMMapper,* performs the final mapping from *ProcessExecution*-Objects to *OPMGraph*-Objects and *OPMPublisher* which serializes and visualizes the graph. The *ProvenanceFactory* class eases up the library handling by assembling important methods to generate provenance data, e.g.

a ProcessExecution – Object is generated by attributing an execute-request and a service URL.

4.2 Use Case1: Interpolation with WPS4R

In our WPS4R executes a simple inverse distance weighted interpolation, using an R-script. Special about WPS4R is that WPS is actually performing processing tasks by triggering a backend-process, which is performed by Rserve [49]. Rserve is a standalone and thread-save TCP/IP server that distributes R [50] functionality for remotely connected applications. The Use Case is special about modeling WPS internals, but important because it involves spatial in- and outputs and literal input parameters.

Essentially the capture was triggered by a small main-method in Java, which passes and execute request and a wps-host URL to the *ProvenanceFactory*, which gererates a *ProcessExecution* instance. This instance is mapped to an *OPMGraph* by a call to *OPMMapper* and serialized in xml, dot, pdf and png formats by the *OPMPublisher* helper class. Serializing rdf is also possible, but not yet supported. The pdf and png files are Graphviz outputs, generated from a dot file. They contain of graphical visualizations as shown in Fig. 7. The xml serialization is compatible with Cytoscape (Karma plugin), so that provenance graphs can by visualized and explored interactively, as shown in Fig. 8.

The figures show artifacts as ovals, which are either process inputs or outputs. WPS-processes are rectangles. Agents, which are WPS instances in each of the figures, are hexagons.

The active execution, i.e. values of process outcomes and process duration have not been captured yet, because the prototype is not implemented so far. Essentially the provenance data was captured from the execute request and enriched by the process description, which was directly retrieved from the service.

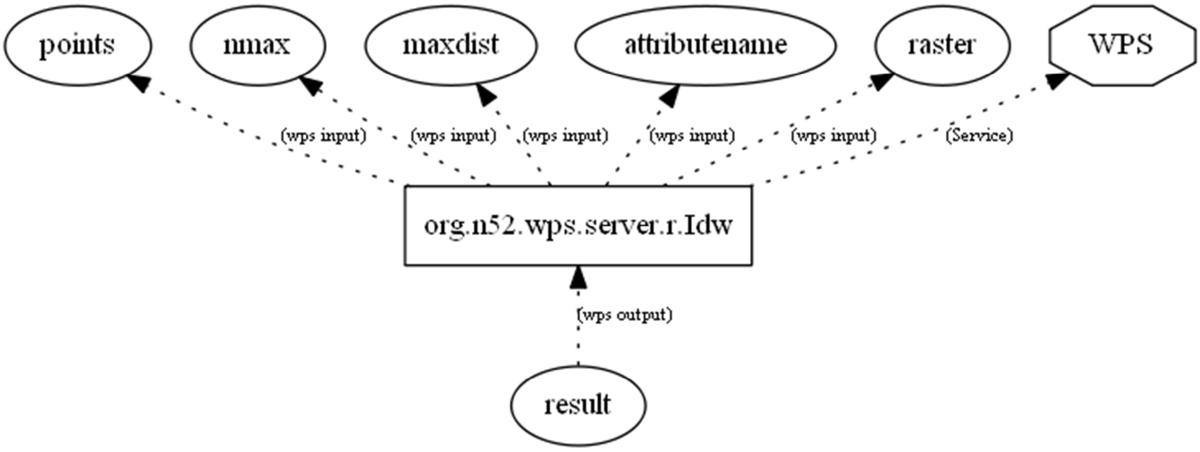
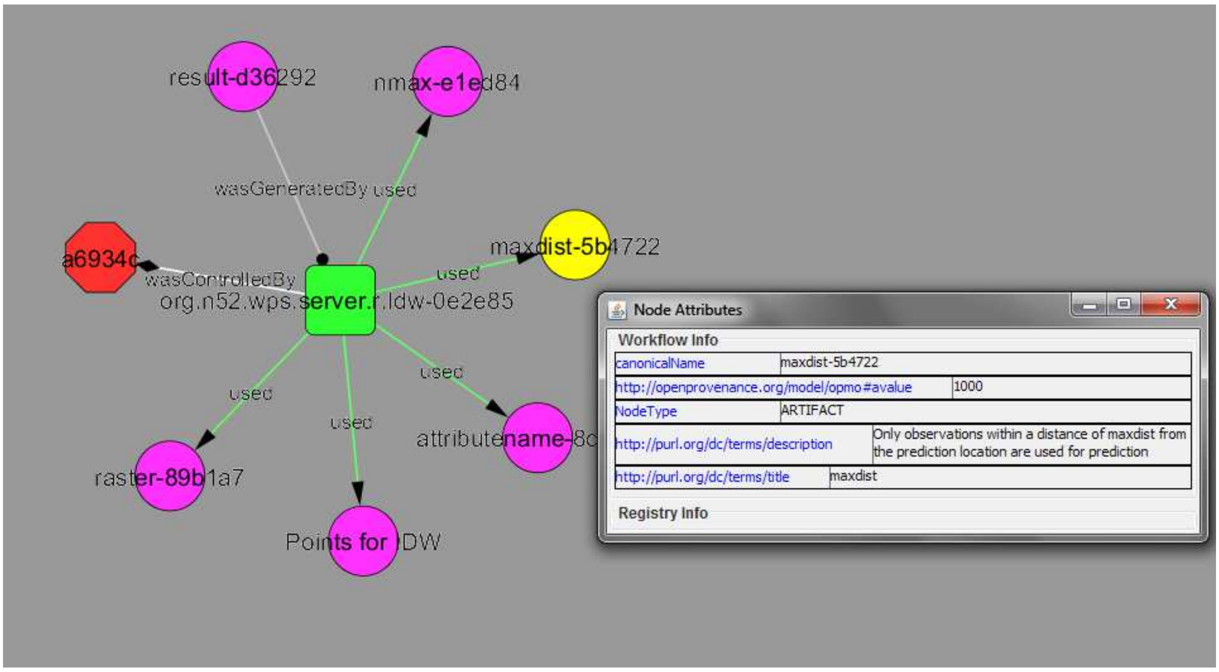


Fig. 7 Visualization of IDW interpolation process, done by Graphviz. Attributes has been suspressed to simplify the graph.



canonicalName	maxdist-5b4722
http://openprovenance.org/model/opmo#avalue	1000
NodeType	ARTIFACT
http://purl.org/dc/terms/description	Only observations within a distance of maxdist from the prediction location are used for prediction
http://purl.org/dc/terms/title	maxdist

Fig. 8 Visualization of IDW interpolation process, done by Cytoscape (screen capture). On the right side is an attribute table of the maxdist - process input, which is marked in yellow.

4.3 Use Case2: Simple Service Chain from Albatross Workflow

The second use case is a simple service chain which combines two WPS processes. It is derived from the Albatross Scenario Workflow, which is part of the EU funded UncertWeb project [12]. The first process generates a so called synthetic-population, which is fed to the second process, which performs an Albatross model run. The workflow can be executed in three different ways, but this case will be the simplest, performed with minimal input parameters that are randomly chosen. Albatross is a rule-based system of activity-travel behavior. Reports and publications are available on the UncertWeb website referenced above. The case of Albatross has been explained by Rasouli et al. [51].

Essentially the provenance capture was done in the same way as described in the precious use case. In difference, two service URLs and two process identifiers had to be specified, along with relations between both processes. These relations are the following: Albatross process *was triggered by* Synthetic population service, export-file input *was derived from* export-file output and export-file-bin output *was derived from* export-file-bin input. The resulting visualization with all parameters is shown in Fig. 9.

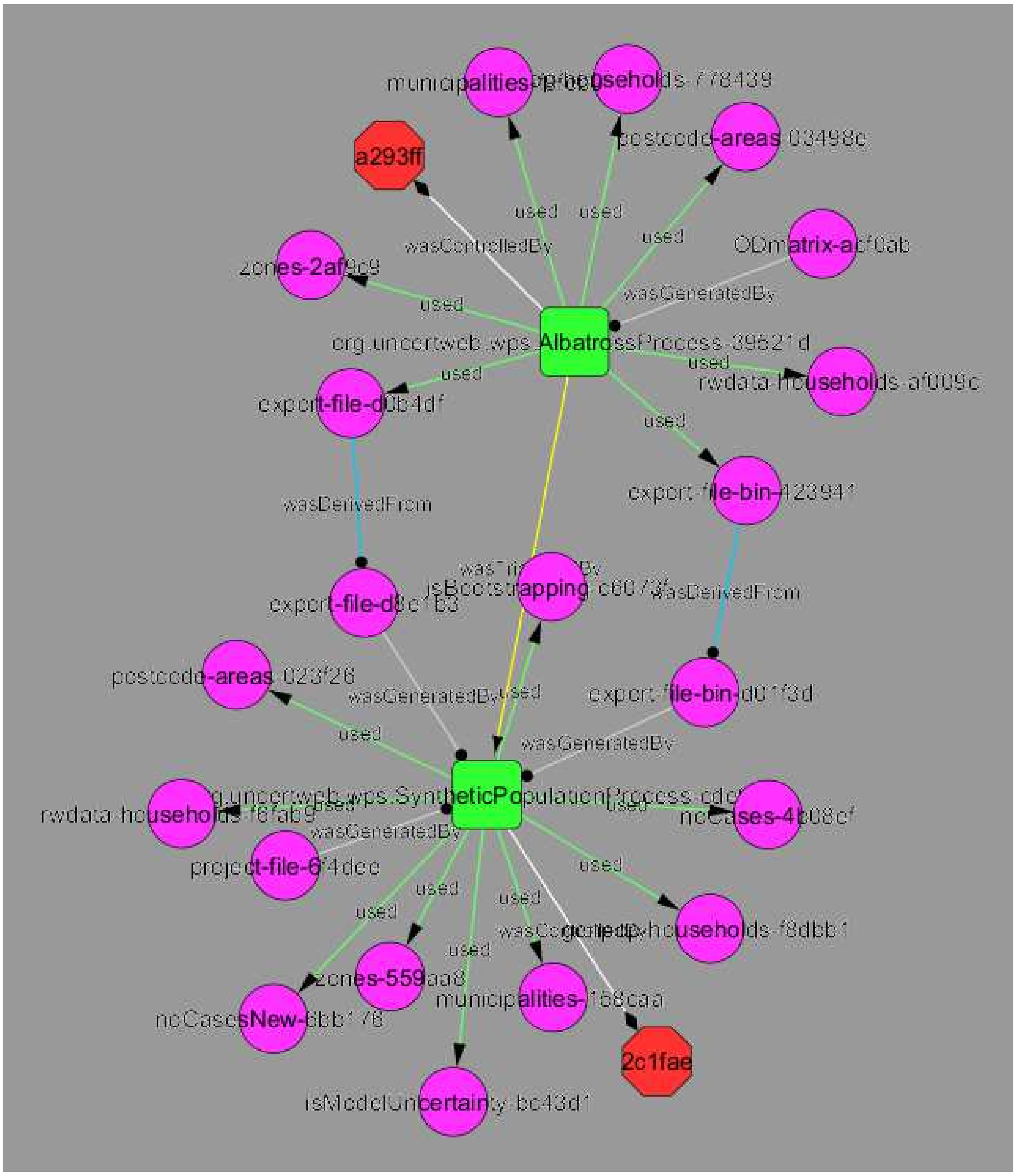


Fig. 9 Visualization of a simple Albatross workflow execution, done by Cytoscape (screen capture). The Albatross process (upper green node) was triggered by the synthetic population process (lower green node). Two inputs of the Albatross process were derived from outputs of the synthetic population service (blue edges).

5. Discussion

This thesis explained concepts and researches about provenance across multiple scientific disciplines. Chapter 2 started from the general eScience domain, pointing over to exemplary researches within the domain of Life Sciences and geoinformatics in section 2.4. Parts of Chapter 2 were also addressing the GI Science domain, either directly, when explaining the ISO lineage model in section 2.2.3, or indirectly when concluding about GI Science and Web Processing Service from researches that do not explicitly address this field. Chapter 3 was explicitly addressing the field of GI Science and the subject of this thesis, but it cannot stand without the fundamental and supplemental considerations of the previous chapter.

Considering issues of provenance modeling, choices have to be done within a vast range of provenance models. This thesis introduced the Open Procenance Model (OPM), the W3C PROV model which is very new at the present and the ISO lineage model. Broadness and interoperability plays a role in this choice, as well as domain awareness and standard compliance. Addressing geographic information standards from ISO and OGC is very important within modern Spatial Data Infrastructures. Further work might dedicate to extension and support of selected provenance models. In case of OPM this may involve specifying a publically available OPM profile for geospatial provenance.

Technical infrastructures that facilitate provenance-aware applications needs to be addressed, as done by the EU Provenance Project [8] (section 2.3) and proposed by Yue et al. [41] in an ISO and OGC compliant way (section 3.2). Selection criteria may be scalability, security and management capabilities, abilities to record, store, exchange and distribute provenance information as well as standard compliance. Much work will be required to integrate these architectures within existing Spatial Data Infrastructures. Handling geospatial provenance also requires sufficient interfaces for querying and visualization, which need to be integrated in the described environment.

My thesis presented a set of requirements for a provenance-aware Web Processing Service. Although the list is not exhaustive and rather informal, it may provide initial considerations for related projects in GI Science. Concrete project requirements depend on individual outsets, thus may exclude some of the mentioned points and further elaborate others.

Outcome of my practical work is a WPS provenance library written in Java. Its use has been demonstrated on two practical use cases, but still it is an experimental prototype. If it is found to be useful and gets acquired by research, further work may include functional enhancements, error handling, systematic testing and detailed documentation. The library may be applied to provenance-aware applications such as WPS instances, WPS clients and workflow engines that support geospatial provenance. Stable releases may become publically available so that the geoprocessing community can benefit from this achievement.

6. Conclusion

Ideas and concepts are often conceived after specific questions. Such a question can be: *Do applications sufficiently explain, how they produce there outcomes?* The answer may be, that users cannot get necessary information that they need e.g. for evaluating or reproducing these outcomes. Similar insights led this thesis to considerations about WPS and related technologies. Thereby it discovered that there is still a lot of work to do within the GI Science domain. It still lacks of concepts and solutions about provenance. This work took a first step towards a provenance-aware Web Processing Service and may be a contribution to provenance research in GI Science.

7. References

[1] C. Kiehle, C. Heier and K. Greve, "Requirements for Next Generation Spatial Data Infrastructures-Standardized Web Based Geoprocessing and Web Service Orchestration," *Transactions in GIS,* no. 11(6), pp. 819-834, 2007.

[2] Open Geospatial Consortium, "Open Geospatial Consortium | OGC(R)," [Online]. Available: http://www.opengeospatial.org/. [Accessed 15 July 2012].

[3] P. Yue and L. He, "Geospatial Data Provenance in Cyberinfrastructure," *Proceedings of the 17th international conference on geoinformatics (Geoinformatics 2009),* 2009.

[4] P. Yue, J. Gong and L. Di, "Augmenting geospatial data provenance through metadata tracking in geospatial service chaining," *Computers & Geosciences,* no. 36, pp. 270-281, 2010.

[5] "The OPM Provenance Model (OPM)," [Online]. Available: http://openprovenance.org/. [Accessed 10 July 2012].

[6] S. Wang, A. Padmanabhan, J. D. Myers, W. Tang and Y. Liu, "Towards Provenance-Aware Geographic Information Systems," *Proceedings of the 16th ACM SIGSPATIAL*

international conference on advances in geographic informationsystems (ACM GIS 2008), 2008.

[7] Mariam-Webster Online Dictionary, [Online]. Available: http://www.merriam-webster.com/dictionary/concept?show=0&t=1342391015. [Accessed 2012 July 15].

[8] P. Groth, S. Jiang, S. Miles, S. Munroe, V. Tan, S. Tsasakou and L. Moreau, "An Architecture for Provenance Systems. Technical Report," 29 November 2006. [Online]. Available: http://eprints.ecs.soton.ac.uk/13216/. [Accessed 10 July 2012].

[9] P. Missier, S. S. Shahoo, J. Zhao, C. Goble and A. Sheth, "Janus: from Workflows to Semantic Provenance and Linked Open Data," in *Proceedings of the third international provenance and annotation workshop (IPAW 2010)*, Troy, NY, USA, 2010.

[10] "Taverna - open source and domain independent Workflow Management System," [Online]. Available: http://www.taverna.org.uk/. [Accessed 4 July 2012].

[11] 52° North, "WPS4R - the R Backend of the 52°North WPS," [Online]. Available: http://52north.org/wps4r. [Accessed 15 July 2012].

[12] "UncertWeb - the uncertain model web," [Online]. Available: http://www.uncertweb.org/. [Accessed 15 July 2012].

[13] W3C Provenance Working Group, "Provenance WG Wiki," [Online]. Available: http://www.w3.org/2011/prov/wiki/Main_Page. [Accessed 7 July 2012].

[14] Y. Gil, J. Cheney, P. Groth, O. Hartig, S. Miles, L. Moreau and P. Pinheiro da Silva, "Provenance XG Final Report. W3C Provenance Incubator Group," 8 December 2010. [Online]. Available: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/. [Accessed 5 July 2012].

[15] Merriam-Webster Online Dictionary, [Online]. Available: http://www.merriam-webster.com/dictionary/provenance. [Accessed 14 May 2012].

[16] NOAA Environmental Data Management Wiki, "ISO Lineage," 2012. [Online]. Available: https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Lineage. [Accessed 15 May 2012].

[17] S. Miles, "Mapping Attibution Metadata to the Open Provenance Model," *Future Generation Computer Systems,* 2010.

[18] J. Cheney, Y. Gil, P. (. Groth and S. Miles, "Requirements for Provenance on the Web," 2010. [Online]. Available:

http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements. [Accessed 2012 July 13].

[19] J. Zhao, C. Bizer, Y. Gil, P. Missier and S. Sahoo, "Provenance Requirements for the Next Version of RDF," [Online]. Available: http://www.w3.org/2005/Incubator/prov/wiki/images/3/3f/RDFNextStep_ProvXG-submitted.pdf. [Accessed 13 July 2012].

[20] S. Sahoo, G. Paul, O. Hartig, S. Miles, S. Coppens, J. Myers, Y. Gil, L. Moreau, J. Zhao, M. Panzer and D. Garijo, "Provenance Vocabulary Mappings," 6 August 2010. [Online]. Available: http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings. [Accessed 7 July 2012].

[21] "International Provenance and Annotation Workshop Series," [Online]. Available: http://www.ipaw.info/. [Accessed 7 July 2012].

[22] "IPAW 2012 - 4th International Provenance and Annotation Workshop," [Online]. Available: http://ipaw2012.bren.ucsb.edu/index.php/IPAW_2012_-_4th_International_Provenance_and_Annotation_Workshop. [Accessed 7 July 2012].

[23] "DataOne. Scientific Workflows and Provenance Working Group," [Online]. Available: http://www.dataone.org/working_groups/scientific-workflows-and-provenance-working-group. [Accessed 7 July 2012].

[24] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan and J. Van den Bussche, "The Open Provenance Model Core Specification (v1.1)," *Future Generation Computer Systems,* 2010.

[25] P. Maué, "Semantic annotations in OGC standards," 16 July 2009. [Online]. Available: http://www.opengeospatial.org/standards/dp/. [Accessed 12 July 2012].

[26] L. Moreau and J. Ibbotson, "The EU Provenance Project: Enabling and Supporting Provenance in Grids for Complex Problems," 6 December 2006. [Online]. Available: http://www.gridprovenance.org/publications/public-final-report.pdf. [Accessed 7 July 2011].

[27] "The EU Provenance Project: Enabling and Supporting Provenance in Grids for Complex

] Problems," [Online]. Available: http://www.gridprovenance.org/. [Accessed 7 July 2012].

[28] I. U. Data Insight Center, "Karma Provenance Collection Tool," [Online]. Available: http://d2i.indiana.edu/provenance_karma. [Accessed 2012 July 11].

[29] "Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization," [Online]. Available: http://www.cytoscape.org/. [Accessed 11 July 2012].

[30] "Karma Provenance Retrieval and Visualization Plugin for Cytoscape. User Manual v1.1.0," 5 January 2011. [Online]. Available: http://heanet.dl.sourceforge.net/project/karmatool/v3.2.1/KarmaVisualizationUserManual.pdf. [Accessed 11 July 2012].

[31] "Graphviz - Graph Visualization Software," [Online]. Available: http://www.graphviz.org.

[32] W. Wiki, "LargeTripleStores," [Online]. Available: http://www.w3.org/wiki/LargeTripleStores. [Accessed 11 July 2012].

[33] "Joseki - A SPARQL Server for Jena," [Online]. Available: http://joseki.sourceforge.net/. [Accessed 11 July 2012].

[34] "The Protégé Ontology Editor and Knowledge Acquisition System," [Online]. Available: http://protege.stanford.edu/. [Accessed 11 July 2012].

[35] "Topquadrant TopBraid Composer," [Online]. Available: http://www.topquadrant.com/products/TB_Composer.html. [Accessed 11 July 2012].

[36] "Sindice Web Data Inspector," [Online]. Available: http://inspector.sindice.com. [Accessed 11 July 2012].

[37] "The BioCatalogue: providing a curated catalogue of Life Science Web Services," [Online]. Available: http://www.biocatalogue.org/. [Accessed 2012 July 7].

[38] "Annotating Web Services in the BioCatalogue," [Online]. Available: http://www.biocatalogue.org/wiki/doku.php?id=public:help:biocatalogue:annotating_web_services. [Accessed 7 July 2012].

[39] "Taverna Provenance Management," [Online]. Available: http://www.taverna.org.uk/documentation/taverna-2-x/provenance/. [Accessed 7 July 2012].

[40 P. Yue, L. Di, W. Yang, G. Yu, P. Zhao and J. Gong, "Semantics-enabled metadata

] generation, tracking and validation in geospatial web service composition for mining distributed images," *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International,* pp. 334-337, 2007.

[41] P. Yue, L. Di, L. He, J. Gong and L. Zhang, "Sharing geospatial provenance in a service-oriented environment," *Computers, Environmet and Urban Systems 35,* pp. 333-343, 2011.

[42] C. Stash, S. Schade, A. Llaves, K. Janowicz and A. Bröring, "Aggregating Linked Sensor Data," *4th International Workshop on Semantic Sensor Networks 2011 (SSN 2011), Workshop of the 10th International Semantic Web Conference (ISWC 2011),* 2011.

[43] Y. Liu, J. Futrelle, J. Myers, A. Rodriguez and R. Kooper, "A provenance-aware virtual sensor system using the open provenance model," *2010 International Symposium on Collaborative Technologies and Systems (CTS),* pp. 330-339.

[44] I. U. Data to Insight Center, "InstantKarma," [Online]. Available: http://d2i.indiana.edu/provenance_instantkarma. [Accessed 13 July 2012].

[45] Y. L. Simmhan, B. Plale and D. Gannon, "A Framework for Collecting Provenance in Data-Centric Scientific Workflows," *ICWS '06 Proceedings of the IEEE International Conference on Web Services,* pp. 427 - 436, 2006.

[46] 52°North, "52°North WPS," [Online]. Available: http://52north.org/communities/geoprocessing/wps/index.html. [Accessed 14 July 2012].

[47] Apache Software Foundation, "Maven - Welcome to Apache Maven," [Online]. Available: http://maven.apache.org/. [Accessed 2012 July 16].

[48] Yatta Solutions GmbH, "UML Lab - Yatta: Software, Tools, Consulting," [Online]. Available: http://www.uml-lab.com/de/uml-lab/. [Accessed 2012 July 16].

[49] RForge.net, "Rserve - Binary R server - RForge.net," [Online]. Available: http://www.rforge.net/Rserve/index.html. [Accessed 16 July 2012].

[50] R Foundation, "The R Project for Statistical Computing," [Online]. Available: http://www.r-project.org/. [Accessed 16 July 2012].

[51] S. Rasouli, T. Arentze and H. Timmermans, "Analysis of Uncertainty in Performance Indicators of a Complex Activity-Based Model: The Case of the Albatross Model System,"

in *Innovations in Travel Modeling*, 2012.

[52] L. Moreau, J. Freire, J. Futrelle, J. Myers and P. Paulson, "Governance of the Open Provenance Model," 15 June 2009. [Online]. Available: http://twiki.ipaw.info/pub/OPM/WebHome/governance.pdf. [Accessed 10 July 2012].

Appendix

Figures

Acronyms

CSW	Catalogue Service for Web
DC	Dublin Core
EU	European Union
GI Science	Geographic Information Science
HTTP	Hypertext Transfer Protocol
IPAW	International Provenance and Annotation Workshop
ISO	International Organization for Standardization
MIME	Multipurpose Internet Mail Extensions
OPM	Open Provenance Model
SOS	Sensor Observation Service
OGC	Open Geospatial Consortium, also referred to as OpenGIS
OWL	Web Ontology Language
OWS	OGC Web Service
RDF	Recourse Description Framework
SDI	Spatial Data Infrastructure
SensorML	Sensor Model Language
UML	Unified Modeling Language
URL	Uniform Recourse Locator
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service
WPS	Web Processing Service
W3C	World Wide Web Consortium
XML	Extensible Markup Language