2017-01-01

Describing Data And Workflow Provenance Using Design Patterns And Controlled Vocabularies

Smriti Rajkarnikar Tamrakar
University of Texas at El Paso, srajkarnikartamrakar@miners.utep.edu

DESCRIBING DATA AND WORKFLOW PROVENANCE USING DESIGN PATTERNS AND CONTROLLED VOCABULARIES

SMRITI RAJKARNIKAR TAMRAKAR

Master's Program in Computer Science

APPROVED:

Natalia Villanueva-Rosales, Ph.D., Chair

Deana Pennington, Ph.D.

Mahmud Shahriar Hossain, Ph.D.

Charles H. Ambler, Ph.D.
Dean of the Graduate School

Dedication

To Prajwol, my better half

DESCRIBING DATA AND WORKFLOW PROVENANCE USING DESIGN PATTERNS AND CONTROLLED VOCABULARIES

by

SMRITI RAJKARNIKAR TAMRAKAR

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2017

Acknowledgements

First and foremost, I would like to offer my sincere gratitude to my advisor and graduate committee chair, Professor Natalia Villanueva-Rosales, for her continuous advice, trust and encouragement during last two years of my master's degree. Professor Villanueva-Rosales supported me throughout my research with her patience and knowledge whilst allowing me a room to work in my own way. I have learnt to research in a specific field and express the knowledge that I gained through visual presentation and writing skill. She introduced me to the fields of cyber infrastructure applications and data base management, and I have increased my ability to learn and explore advancements in those fields. My appreciation is also extended to graduate committee members, Dr. Deana Pennington and Dr. Mahmud Shahriar Hossain, for providing valuable suggestions, comments, and feedback.

A special thanks to Mr. Luis Garnica Chavira for providing supportive information for my research. I would like to thank my colleagues at iLink lab for their support during my research. I appreciate the love and moral support from my father Krishna Kumar Rajkarnikar, mother Sumitra Rajkarnikar and brother Kushal Rajkarnikar for their love, care, support and encouragement for pursuing my master degree in the United States.

In addition, I would like to acknowledge the Department of Computer Science providing a Teaching Assistantship in the first year of my master's degree and iLink for providing research assistantship in the second year. This work was supported in part by the National Science Foundation under CREST Grant HRD-1242122. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF).

Abstract

In any scientific experiment, researchers are required to access, compute, and analyze data to produce useful information to the scientific community. In order to instill trust on such scientific research products, the product users need to understand the procedure applied and the assumptions incorporated. The reuse and replication of reliable scientific data need methods that help the users to understand data origin and the derivation process, i.e. provenance. Although several standards for representing provenance such as PROV model (a W3C recommendation) have been recommended, they have not been widely utilized by scientific communities due to difficulty in aligning such recommended standards to their needs. However, use of this standard has not improved, as suggested by provenance usage studies in the literature.

In this research we propose controlled vocabularies for describing provenance data using three provenance design patterns. These provenance design patterns were used in three domains, i.e., Smart Cities, Water Modeling, and Biodiversity Modeling. We evaluate the proposed vocabulary with users of the interdisciplinary, international USDA-funded Water modeling project. The results show that in general, provenance is important to understand and trust a final product. This work provides a building block to create and evaluate complex provenance design patterns that can be embedded in systems that manipulate data and executes scientific workflows.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Motivation

The Semantic Web is the foundation in which a new forms of Web content that are meaningful to the computers is provided. It is an extension to the current web that provides structural meaning to the information on the web such that the information can be consumed by both human or computers (Berners-Lee, Hendler, & Lassila, 2001). Initially, formats such as RDF and XML were used to provide some structure to the information – and are still used as a standard to this day. Furthermore, along with increasing community involvement, support, and working group standards, controlled vocabularies and data models for different domains have emerged. However, these domain experts still struggle to describe the origin, sources, and processes used to derive data products, thus imposing a barrier in the understanding, reuse, and adoption of these data products. It also makes it difficult to discover the relationships between the processing activities and data flow that span multiple institutions. This research is based upon resolving this part of the problem such that the proposed provenance design patterns help domain experts in order to use as well as reuse and express data, workflow and its associated provenance.

1.2 Problem Statement

Scientists use workflows extensively in order to perform scientific computations, they utilize workflow tools to design, validate, execute, and visualize scientific computations and their results (D. Garijo, Gil, & Corcho, 2014). Scientific workflows play a vital role in expediting complex computations and constructing new experiments. Additionally, such workflow usage encourages collaboration by the reuse of workflow fragments. However, collaborating between

different organizations and disciplines gives rise to challenges due to different domain vocabulary, heterogeneous formats of data, and approaches to solutions. The situation becomes more complicated when there is a deluge of data and metadata (data about data) is not well managed. This can result in an absence of knowledge about where a piece of data came from, who was responsible for processing raw data, what software was used for processing and so on.

Recent developments in provenance representation such as Open Provenance Model (OPM) (Moreau et al., 2011) leading up to the World Wide Web Consortium (W3C) PROV family of documents including PROV-Ontology (Groth & Moreau, 2013) indicate the growing recognition of the need for provenance of information. However, recent provenance studies (Groth & Beek, 2016) show that although there is an uptake of provenance usage in the Semantic Web community, its usage is less than anticipated.

This work aims to facilitate the understanding of data and data processing (i.e., workflows) to enable their reuse by providing provenance information via related design patterns. We believe that the use of provenance design patterns to represent provenance, will facilitate data and model providers in tracing of provenance and better inform users about workflow steps, the source of data, and how data were manipulated. Hence, through the improved understanding of provenance, potential users will be able to reuse and repurpose data and workflows.

1.3 Hypothesis

The hypothesis of this work is that representing provenance using controlled vocabularies and provenance design patterns will facilitate the annotation process of scientific data and workflows and improve the understanding of such products to evaluate their reuse.

1.4 Goals and Objectives

The goal of this work is to facilitate the provenance description and potential reuse of scientific data and workflows by practitioners using existing controlled vocabularies and provenance design patterns. While working towards this goal, the objectives of this research are as follows:

a. To survey the key provenance elements inspected by scientists for understanding as well as considering data and workflows for reuse.
b. To represent key provenance elements to analyze scientific data and workflows using standard controlled vocabularies organized in provenance design patterns.
c. To represent the proposed provenance design patterns in three different scenarios.
d. To evaluate how the proposed representation of key provenance elements facilitate the understanding of scientific data and workflows, and instill trust in a specific scientific product.

1.5 Contributions

The main contributions of this thesis are as follows:

a. Literature review of previous work in provenance trace representations and ontology design patters.
b. Implementation of proposed key provenance elements for scientific data and workflows using controlled vocabularies and provenance design patterns.
c. Filling the gap between the creators of standards and controlled vocabularies, and practitioners that require best practices, suggested provenance design patterns, and serialization of provenance traces for specific applications, e.g., provenance visualization.

Chapter 2: Background

In the previous chapter we provide the hypothesis, goals, and objectives of this work. In this chapter, we present the basic concepts on which this work was based.

2.1 Provenance

Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing (Moreau & Groth, 2013). Provenance represents the origin of something, where an artifact came from, how it came to be in its present state, and the steps that took place to get a resource in its current state. Hence, it offers a medium to verify results of a process, assess quality of data, and analyze the processes that produced data products. This allows the users of such data products to make an informed decision about whether they can trust them.

Provenance can broadly be categorized into: *data provenance* and *process provenance.* Data provenance mostly deals with finding associations between input and output and tracking intermediate results or steps that are responsible for the generation of a data product. Process provenance helps to find the order in which processes and their components are executed.

In the lifecycle of a scientific workflow, it is important to identify the input data, output data, and their sources. The workflow implementation is clearer when the data transformation steps and data movement are given. Data provenance represents the traces of movement of data that show how a piece of data came to be in its current state. It uses a combination of connections between what tools were used to transform data, what environment the tools were used in, and which personnel or acting personnel was responsible for the workflow process.

Data provenance is important as it provides more value to the integrity of the input and output data by allowing it to be traceable. Additionally, by providing the steps followed to manipulate data, users are able to understand, verify, and reuse the workflows. Hence, by giving the means to trace a workflow run, users are able to assess the quality of the workflow which makes the workflow and associated products more trustworthy.

As more of the scientific experimental tasks are delegated to computational methods and sensing devices, it is crucial to know the data flow path along with the computational steps. Additionally, as the amount of data resulting from scientific experiments and their analyses grow, it is even more important to capture the connection between derived data, process of derivation, and the parameters used to derive the data (Koop et al., 2010).

2.2 Semantic Web Technologies

Semantic Web technologies help to add contextual information and provide representation formats that assist in getting information rather than just documents on the web (Shadbolt, Berners-Lee, & Hall, 2006).

2.2.1 Ontologies

Ontologies represent vocabulary that formally describe a specific domain with logical axioms that express the concepts and relations between such concepts. Usually, ontology development is not an end goal in itself but rather a description of data and its structure that can later be consumed by other programs. An ontology is a formal explicit description of concepts in a domain of discourse, properties of each concept that describe features and attributes of that concept, and impose certain restrictions on properties of concepts (Noy & McGuinness, 2001). The ontology terms are depicted with `Courier New font` for rest of the document.

As one of the Semantic Web standards, ontologies are meant to be used to establish common understanding of concepts such that a community agrees to use the same terms in order to describe and ask questions about concepts and their relations. Additionally, ontologies are intended to be used in order to facilitate knowledge integration from different inter-related domains. According to Mao (2007), ontology mapping is finding semantic correspondences between similar elements of different ontologies. When the need to access multiple ontologies by numerous applications arises, ontology mapping provides a common layer through which information can be exchanged in a semantically sound manner (Kalfoglou & Schorlemmer, 2003). Ontology mapping is an integral step that establishes links between various knowledge bases usually through semantic relations such as subsumption and equivalence. Let us take as an example, two entities that represent similar concepts in different context. For example, if we have two ontologies that express knowledge about a university and an organization, we could assume that the concept of a person is a common occurrence in both domains. However, the same concept of a person may hold different information based on the context that the domain needs. A person could be a student that holds information for properties such as First name, Last name, Student ID, and Degree level. On the other hand, the person could hold information for properties such as Social Security Number, Marital Status, and Employee ID when the person is represented as an on-campus employee. There are logical rules that can be added to enhance such knowledge expressed in an ontology such that the on-campus employee can be represented as an extension of a university student. An example could be a concept of an undergraduate student who is a university undergraduate level student and is employed as an on-campus employee with additional descriptions such as Social Security Number, Marital Status, and Employee ID. Some of the most common standards used to represent ontologies are described in the next two sections.

According to Gangemi (2005), ontologies are broadly classified as: foundational, core, and domain ontologies. A foundational ontology contains generic domain-independent elements. A core ontology contains general terms of a certain domain. A domain ontology contains all ontology elements that are needed for that domain to be conceptualized for some task. The provenance design patterns proposed in this work fall under the core ontology for scientific data and workflows which build upon other foundational ontologies (Provenance Ontology) and core ontologies (Semantic Sensor Network Ontology, Wf-desc Ontology, ProvONE Ontology). Thus, the provenance design patterns proposed in this work, act like a glue to bind standard generic concepts to domain-specific functional concepts using concepts and relations from a core vocabulary.

2.2.1.1 Resource Description Framework (RDF)

RDF is a data model that describe concepts and relations in a subject, predicate, object format. It is a graph where nodes represent concepts or classes and edges represent the relations between them. Both nodes and edges have labels that identify them. International Resource Identifier (IRI) are used to differentiate between resources. An IRI can be used to identify a subject (resource), predicate (relations), and objects (resource), also called a triple. Figure 1 shows an example of a triple, where **user-workflow** (subject) is of **rdf:type** (predicate) **Workflow** (object). In the example, a user-workflow is a specialization of general definition of Workflow. Therefore, the subject and the object is linked by rdf:type relation. A set of such triples is called an RDF graph. An RDF vocabulary is a collection of IRIs intended for use in RDF graphs (Cyganiak, Wood, & Lanthaler, 2014). RDF is an abstract data model, as it still deals with resources on a conceptual level. Data is expressed as an edge in the graph, where a single edge represents a statement or a triple. In order for the semantic structure to be exchanged or transferred, RDF needs

to be serialized into formats such that it can be consumed by other programs. Some of the common serialization formats are N-triples, RDF/XML, OWL/XML, Turtle, JSON-LD.

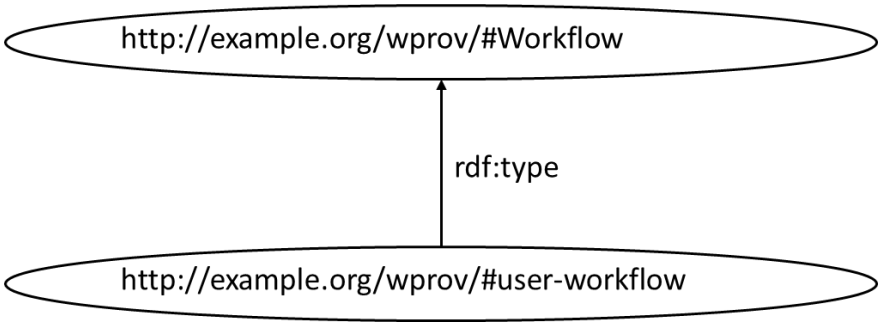

Figure 1: Example of a triple that illustrates relationship between a user workflow (subject) and workflow (object) connected via rdf:type (predicate) property

Linked data helps to interconnect data on the web and is analogous to how hypertext documents are linked on the web. However, instead of anchor tags that link Hypertext Markup Language (HTML) documents, linked data uses links as described by RDF in order to describe any concept or thing. Thus, linked data helps for humans and machines to process and explore data by following the link expressed as a URI.

2.2.1.2 Web Ontology Language (OWL)

OWL is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things ("OWL - Semantic Web Standards," n.d.). This language is used to apply computational logic to express the inter-relationships between the concepts such that it can be processed by machines. If we consider the example illustrated in Figure 1, where a **user-workflow** is a **type of Workflow**, then OWL allows us to model **Workflow** as a class of different categories of workflow. Furthermore, if we model Workflow as a type of activity, then **Activity** represents a class of concepts of **workflow**s.

In addition to adding knowledge by leveraging dependencies in relations of things, by using reasoners we can further extract implicit knowledge from the ontology and validate the consistency

of the knowledge. In the above example, if we run the reasoner then a user-workflow will be shown as an instance of Workflow as well as an Activity. As a result of the OWL documents generated, expressed knowledge can be published on the web and integrated with other OWL documents for better knowledge harvesting.

2.2.1.3 JavaScript Object Notation for Linked Data (JSON-LD)

The Java Script Object Notation (JSON) format is considered as one of the lightweight open-standard data interchange formats ("JSON," n.d.). It has received popularity because of its ease of reading for humans and parsing for machines. Although JSON in the text format is completely an independent language, it uses several conventions as used by other languages such as C, C++, C#, Java, JavaScript, Perl, and Python. JSON uses key-value pairs in order to express triple like semantic structure. Figure 2 illustrates a JSON example where the history of a workflow and its associated steps are shown in key-value pairs.

JSON-LD is a lightweight linked data format that is convenient for humans to read and write as well as machines to parse. Since, it is based on JSON data, it provides an ideal format to programming environments, web services, and unstructured databases.

2.2.1.4 Design Patterns

Design patterns are flexible building blocks that represent common conceptual patterns that emerge in different domains when solving different tasks from various domains (Hu et al., 2012). These flexible patterns have been seen to be useful in planning for more complex ontologies. In this work, we aim to utilize these patterns in planning for provenance representation of computational tasks and associated data in a scientific workflow.

These reusable patterns act as a preliminary motivation to plan for efficient execution of services, answer core domain questions, and aide in ontology population. These design patterns are also helpful in creating initial versions of the provenance vocabulary. The design patterns

proposed in this work are focused on representing the key provenance elements. So, we refer to them as provenance design patterns for the rest of this document.

Ontology design patterns can be broadly be classified into two types:

a) **Logical Ontology Design Patterns:** These patterns are typically not dependent to any specific domain. This type of pattern deals with the issues that arise due to the restrictions depending upon the knowledge representation language.


```
{
    "user-workflow": [
        {
            "@id": "Step1: human-intervention",
            "@type": "Activity",
            "hasValue": "xsd:string",
            "hadNextStep": "climate-scenario"
        },
        {
            "@id": "Step2: climate-scenario",
            "@type": "Activity",
            "hasValue": "xsd:string",
            "hadNextStep": "customize-parameters"
        },
```


Figure 2: JSON snippet illustrating workflow execution provenance

b) Content Ontology Design Patterns: In order to solve the content design problems, these patterns can be adapted. The fundamental principle of these patterns is to propose the domain-dependent conceptual models of domain classes and properties for populating ontologies. The ontology design patterns proposed in this work mostly fall this category as they are meant to

assist in capturing the provenance of input, output, and intermediate information of scientific data and workflow from across different application domains.

2.3 Literature Review

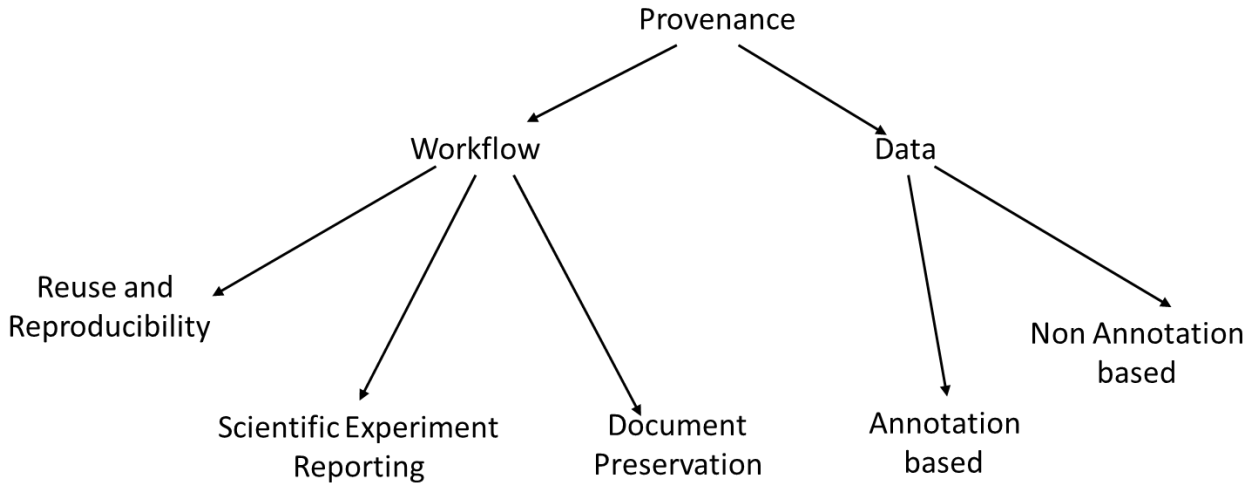

Figure 3: Branch taxonomy of literature review

The literature review for this work consists of two main branches: workflow (coarse-grained) and data (fine-grained). Here, the fine-grained provenance is termed as Data Provenance, however, it specifically is classified as tracing the origin of relational database records (unlike data provenance mentioned in Section 2.1.2). Data provenance is thus categorized into annotation-based (Eager) where the database records are annotated as the database schema is being created and non-annotation based (Lazy) approaches where methodologies perform query rewriting in order to trace the origin of a relational database triple.

In the area of workflows, provenance applications were geared towards three distinct fields: i) reuse and reproducibility of workflow, ii) scientific experiment reporting, and iii) digital preservation of workflow and related documents.

i. Reuse and reproducibility:

   Scientific workflows have been extensively used as building blocks that can be used

for common scientific tasks which can be shared, aggregated, and re-purposed to meet new requirements. However, in order to ensure their reuse and reproducibility, in addition to the workflow specifications, some additional resources are needed to understand what the workflow is meant to achieve.

To ensure that the workflow products are discovered and used as intended, users need to have tools that aide them to understand the workflow purpose such as annotations describing the process. The origin of datasets used, and resulting datasets produced by the workflow as well as the traces of a workflow execution (data provenance) helps to enrich workflow specification and thus enforces trust in the final products. As scientists are more interested in the steps taken to derive an experimental product rather than fine analysis of data itself, several studies on different areas such as provenance versioning and food safety compliance checking have been done recently.

ii. Scientific experiment validation and reporting:

Workflow execution trace is also beneficial to validate experimental data. While the scientific workflow provenance greatly facilitates workflow reuse, studies show that it does not help much with the purpose of reporting while publishing datasets (Alper, Belhajjame, Goble, and Karagoz, 2014). In a journal article by Sahoo et al., (2008)the authors propose ‘semantic provenance modules’ that supply dynamic metadata which can later be integrated into workflows on demand. This work is based on a specific domain and requires altering the workflow and execution process whereas the former study is non-intrusive and still propagates the metadata.

iii. Digital preservation of workflow and related documents:

The workflow specifications are not relevant just when they are being undertaken but for the future as well. Some studies emphasize the advantage of using collection of provenance traces from past workflow executions in order to find possible substitutes for unavailable services within workflows as time progresses (K. Belhajjame, Goble, Soiland-Reyes, & Roure, 2011). Other works like Koop et al., (2011) attempt to promote the idea of an executable paper that allows readers and reviewers to validate and explore experimental results. Thus, they reduce the barrier to reproducing and extending research.

2.4 Challenges in Provenance Representation

In the context of scientific workflows, the provenance is crucial for attribution, sharing, replication, validation, and quality analysis. W3C has recently released a set of PROV documents (Groth & Moreau, 2013) in an attempt to provide some common ground for standardization as well as the exchange of provenance amongst multiple collaborators. These recommendations are intended to be domain-agnostic and general in nature. Due to the general nature of these standard vocabularies several scientific domains and their sub-components have extended the recommendations in order to meet specific needs (Daniel Garijo & Gil, 2012; Markovic, Edwards, Kollingbaum, & Rowe, 2016; Missier, Dey, Belhajjame, Cuevas-Vicenttin, & Ludäscher, 2013; Moreau & Groth, 2013). However, they are limited in the number of representations that they can express in order to truly support interoperability between workflow process and data systems. Additionally, they are limited in their ability to provide better understanding of the research investigation.

For any kind of scientific application that deals with data from various sources it is important for the user to trust the sources from which the data has been provided, like a journal, paper, website etc. If the application uses sensors to retrieve data, deployment information, person responsible for deployment, calibration requirements and other metadata are provided, helps the end-users to trust the data from it and thus have reason to rely on it.

The importance of ontology design patterns and the scope of their usage for representing provenance has been explained in ***Section 2.2.1.3***. The modeling of workflows can also take advantage of common patterns as well as abstractions of workflows to display workflow provenance. This will help the workflow developers to reuse existing workflows and workflow users to help them understand the workflows. Both of these objectives essentially help the developers and users of the workflow to trust the end product(s) of such a workflow.

Current existing workflow provenance vocabularies tend to emphasize certain aspect of the workflow that make them specific and cannot be reused for different scenarios. Although design patterns for workflows have been studied, most of them do not provide much clarity on how the patterns should be applied (ODP-Wiki). Even though some patterns have graphical illustrations to exemplify how they are meant to be used, the majority of suggested patterns do not have good documentation. Hence, their reuse by others is limited.

In the public ontology design patterns library (ODP-Wiki), it is difficult to find patterns relevant for workflow and data provenance representation. Additionally, there is a lack of different use cases from various scenarios that display design pattern usage (Blomqvist et al., 2016). This work creates new patterns and specializes existing patterns to describe workflow processes and data using recommended provenance standards such as PROV (Groth & Moreau, 2013) and ProvONE ("The ProvONE Data Model for Scientific Workflow Provenance," n.d.). It also

provides use case scenarios where the patterns can be applied and/or extended to fit a specific need.

2.5 Related work

Previous works on modeling provenance have been done in various domains, such as semantic trajectory description (Hu et al., 2012) where ontology design patterns were used to describe the trajectory. However, the work did not describe provenance or use any recommended standards during that time such as are incorporated in this research. Other works like streamflow forecasting (Shu, Taylor, Hapuarachchi, & Peters, 2012) describe provenance by adapting ontologies based upon standards such as Open Provenance Model (OPM). The OPM was basically conceived to characterize what caused "things" to be and how "things" depended on others and resulted in the current state (Moreau et al., 2011). Artifacts are immutable pieces of state which may be a physical object or a digital representation in a computer system. Processes are actions or series of actions which in turn produce other artifacts. Agents act as catalysts of a process that are responsible for enabling, facilitating, controlling, or affecting the process execution. The W3C standard model that we adopt in this work has been influenced by the OPM model. Hence, we see a lot of similarity in the context of the basic concept representation with slight variation in terminologies. In PROV, artifacts and processes are called entities and activities whereas derived-from, generated-by, used, and informed-by edges are used as past facts such as `prov:wasDerivedFrom`, `prov:wasGeneratedBy`, `prov:used`, and `prov:wasInformedBy` respectively (Kwasnikowska, Moreau, & Bussche, 2015). We do not incorporate OPM representations in our provenance design patterns, as provenance expressed in one system using OPM can be easily mapped into PROV. This work reuses PROV by reusing and

extending PROV concepts and relations by adapting them to represent process and data provenance in different scenarios.

Similarly, Salayandia (2012) created a framework that provides a workflow-driven perspective as observed by scientists in order to describe the data collection process, the collection conditions, and experiment provenance can be found in. However, this work uses Proof Markup Language (PML) instead of W3C recommended OPM, which is more aligned to the recommended PROV model. In D-PROV (Missier et al., 2013), the PROV model is extended to adapt to a workflow structure for data and workflow preservation of large projects. The D-PROV team tries to emphasize the data flow part of the workflow by using ports and channels as source and sinks. This work is similar to D-PROV in that it tries to blend both prospective and retrospective provenance and provides relations to express them, however, we do not express provenance relations in such fine granularity in order to make the provenance design patterns extensible and reusable.

Another similar work on research objects (Khalid Belhajjame et al., 2015), goes beyond workflow preservation and focuses on describing aggregated resources by implementing a suite of ontologies. This work uses domain-agnostic annotations similar to our work in order to increase the understandability, reusability, and reproducibility. This work uses a suite of ontologies and allows expression of a research object as an aggregated form of resources. Although the provided provenance representations extend well-known ontologies like our work, we do not deal with workflow evolution in detail.

Chapter 3. Application Scenarios

The provenance design patterns created in this work are applied to systems from three different scenario domains: the Water Modeling Platform, Life Species Distribution Modeling, and the Smart City Sensor Platform. As the provenance design patterns are reusable and extensible according to their application domain, the usage of these patterns have been exemplified based on these three distinct domains.

3.1 Integrated Water Modeling Platform

The Integrated Water Modeling Platform is a part of the Sustainable Water Resources Project funded by the United States Department of Agriculture under Grant No. 2015-68007-23130. The project aims to provide tools that allow individual stakeholders to have a common understanding of future scenarios of water availability of the Middle Rio Grande River valley. The Modeling Platform exposes complex scientific models to the web by providing a middleware application which acts as a broker between the offline modeling software and the modeling interface. Through the web-based interface, users can select a model, customize parameter values (constrained automatically to valid values from a knowledge base), and visualize the results in different formats.

The water model simulates all major sources, sinks, uses, and losses of water for Middle Rio Grande between inflow at Elephant Butte Reservoir and Fort Quitman on the Rio Grande. This model is designed to be a tool in understanding hydrology, institutions, and economy in order to guide analysis of policy and management questions of importance to stakeholders (Ward, 2016). The model (see Figure 4) requires inputs such as hydrologic data, crop requirements, surface treaty delivery requirements from U.S to Mexico, evaporation rates, and reservoir capacity. Water use

demand is informed by crop yields, cost of production, crop price, price elasticity of demand, and urban population. The current version gives the model outputs for water users of irrigated agriculture, urban demands, and environmental / recreational demands. The water usage pattern is identified such that the discounted net current value of water is maximized by adjusting water usage patterns for past and future time periods.

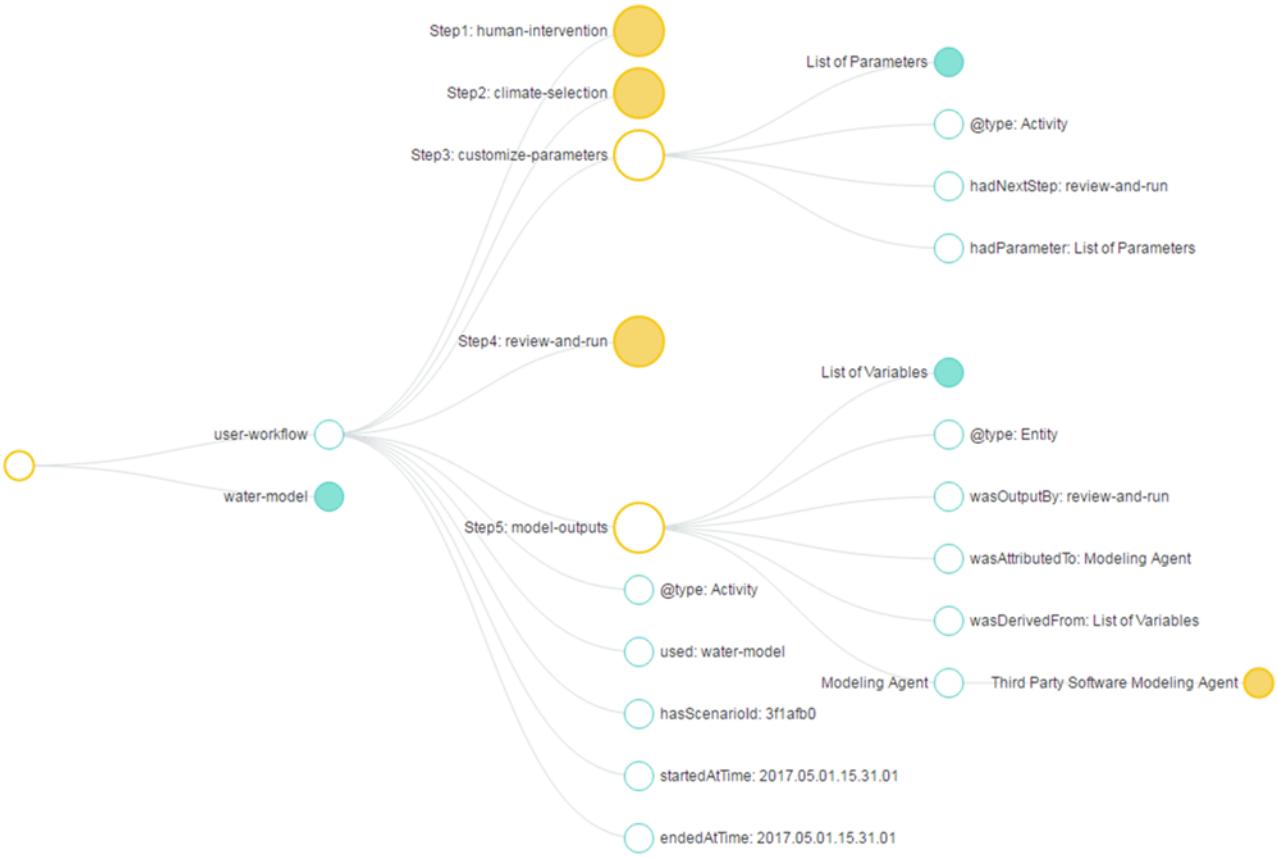

Figure 4: Workflow execution provenance design pattern applied to Integrated Water Modeling Platform (Ward, 2016) as a JSON visualization that illustrates input, processing steps, and model outputs

With such a large number of parameters and sources of data the workflow execution provenance design pattern helps users of the water modeling platform to understand the history of the created model. If a user X wants to view what a parameter A's default value was after a user customized it, the pattern visualization helps user do so. Likewise, the provenance design patterns

categorize the conceptual terms in the water modeling workflow as entity, activity, or agent in the PROV vocabulary or as an extension of it.

It also uses other controlled vocabularies that are commonly used to express provenance terms and relations in respective scenarios such as **wfdesc** (Workflow Description) and **dcterms** (Dublin Core Metadata Initiative).

3.2 Biodiversity Modeling

The Earth, Life and Semantic Web (ELSEWeb) system uses heterogeneous data sources and Species Distribution Modeling (SDM) in order to study the effect of climate and human activities on biodiversity scenario (Villanueva-Rosales, Rio, Pennington, & Chavira, 2015). SDM helps to determine the chances of a species' survival based on current conditions where it is found, human activities, and environmental data. It automates the process of collecting and pre-processing the relevant data so that it can be ingested by an SDM thus facilitating data-to-model integration. The ELSEWeb framework implements data integration between datasets from the University of New Mexico Earth Data Analysis Center (EDAC) with the SDM service provider called Lifemapper.

The SDM is a statistical model that uses the location of known occurrences of species (dependent variables) and various environmental conditions at those locations (independent predictor variables)(Cavner, Stewart, Grady, & Beach, 2012). These variables are either measured using instruments or observed by humans. The ELSEWeb allows the data retrieval and transformation into appropriate formats as required by the modeling service, which is often the most time-consuming task. The provenance design patterns can be used to illustrate the procedure of data collection and processing activities.

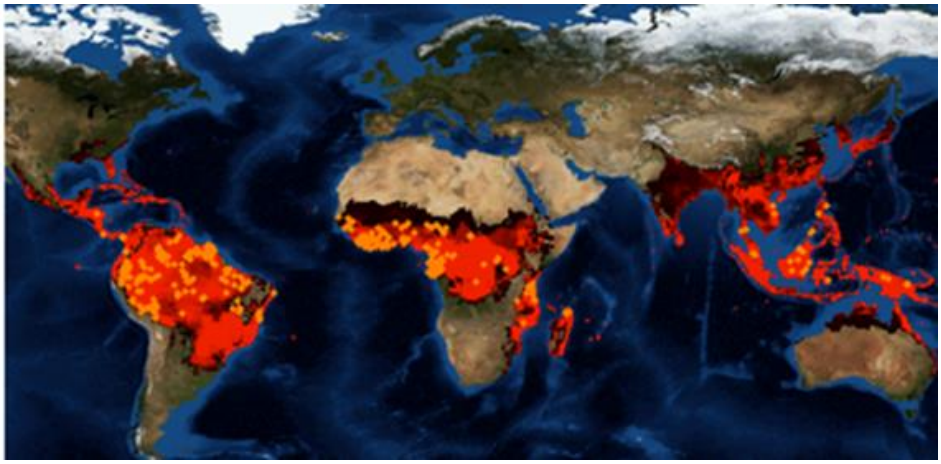

Figure 5: Illustrating the Biodiversity model outputs (species projections) ("Lifemapper," 2017) with integrated environmental data and historical species occurrence data (input)

The environmental data are obtained from many different satellite imagery sources using remote sensing instruments such as the Moderate Resolution Imaging Spectroradiometer (MODIS). The provenance design pattern helps an ELSEWeb modeling user to understand the details of data processing steps performed by service agents. Analyzing the provenance trace of a particular run of the model can help another user to verify that experiment. Also, if a user to wants to perform a similar experiment with different parameters, he can compare his own results with others' results. Furthermore, if someone needs to perform a similar experiment but with different types of data, he can have a reference of the different data processing steps and the software agents associated with corresponding process.

The proposed data collection and processing provenance design pattern represents the history of the observed data such as what type of instrument was used to get the land cover data. Similarly, the data processing provenance design pattern helps to understand all processing activities that took place such that land cover data is processed into a format that generates a dataset that can be used for species distribution modeling. Figure 5 shows the resulting Biodiversity Model

outputs as species projections taking into account the historical species occurrence datasets and environmental data obtained by satellite imagery.

3.3 Smart Cities

IEEE Smart Cities defines a smart city as a city that brings together technology, government and society to enable characteristics such as: smart economy, smart mobility, smart environment, smart people, smart living, and smart governance. The improvement of quality of city life through the above mentioned sectors requires data management from multiple domains such as public transportation, road maintenance, waste collection and urban fault management (Consoli et al., 2015). Such data management benefits by leveraging the semantic tools in order to describe, analyze, and annotate city data in the form of linked open data. Publically available web services can further present the city model and data in way that people can observe and interact with them.

As the urban population grows and rural settlement are abandoned, along with the global growth caused by urbanization, other issues such as air pollution and citizen mobility arise. Sensory devices are a trivial part of data collection, aggregation and analysis. Although there are recommended standards for describing sensing device information such as W3C Semantic Sensor Network (SSN) (Taylor, Janowicz, Phuoc, & Haller, 2017), they do not offer a means to express sensor data provenance. As the sensor data may be used in multiple domains for different purposes, sensor data provenance is crucial in order to understand and reuse the data, quality (Corsar et al., 2013) and trustworthiness assessment (Umuhoza & Braun, 2012) .

As a provenance design pattern, we illustrate the sensor observation provenance design pattern (a `prov:Activity`) that shows what property (`prov:Entity`) the sensing device is

measuring, what type of sensor (`prov:Entity`) was used for measurement, and what other entities (`prov:Entity`) affect the property being measured. As an outcome of the observation we get the value of the measured property.

```
"@id": "http://ontology.cybershare.utep.edu/smart-cities/scllv",
"@type": "iot:Model",
"label" : "Smart Cities Living Lab Vocabulary",
"iot:attribute": [
{
 "@type": "xsd:float",
 "@id": "#light",
 "iot:purpose": "iot-purpose:light",
 "iot:unit": "iot-unit:light.si.lumen",
 "iot:type": "iot:type.number",
 "iot:read": true,
 "iot:write": false,
 "iot:sensor": true,
 "iot:actuator": false,
 "label" : "light",
 "comment" : "numeric light value measured in lumens by a light sensor"
},
```

Figure 6: JSON-LD representation of Smart City Vocabulary that describes light as a measurable attribute using a sensor

Chapter 4. Methodology and Results

In the previous chapter, we presented the background information of scenarios where the proposed provenance design patterns of this work have been applied. In this chapter we discuss the methodology that we use in order to extend, align, and evaluate design patterns using controlled vocabularies.

4.1 Provenance Design Patterns

Provenance design patterns are flexible and self-contained building blocks of semantic annotations. They are intended to solve commonly recurring modeling issues such as how to model a scientific process, inputs, and outputs in alignment with provenance standards. It is quite challenging for domain experts from different backgrounds to agree upon a stable domain ontology. Some foundational ontologies have been able to provide basic, agreed upon, common ground in order to build a domain ontology. However, in the sector of linked data, these foundational ontologies are found to be too generic and have ontological commitments and relations that are difficult to understand for a layman. In order to deal with this, the provenance design patterns are proposed. Such patterns were found to be useful as they could be combined, re-purposed, and aligned with other foundational ontologies to serve as a base for creating complex ontologies.

The following section covers three types of scientific data and workflow provenance design patterns that can be applied to different scenarios. Out of these three scenarios, the first one was implemented as a provenance annotation in JSON format for the water Modeling Platform. Section 4.2 presents the details of the provenance visualization structure. Chapter 5 presents the details of

a survey performed on the workflow execution provenance design patterns and its results. The ontology terms are depicted with `Courier New font` for the rest of this document.

4.1.1 Workflow execution provenance design pattern

This provenance design pattern combines together the basic minimal components of a workflow namely: `Workflow`, `Steps`, `Input`, `Parameter Collection`, and `Variable Collection`. Figure 7 shows that all concepts use the standard symbols for PROV-O as Activity, Entity, and Agent. PROV concepts and relations are shown with the `prov` namespace, whereas the central proposed provenance design pattern concepts and relations are shown with `wprov` namespace. All concepts and relations in the provenance design pattern are aligned to PROV starting point terms, few expanded terms, and some qualified terms as well.

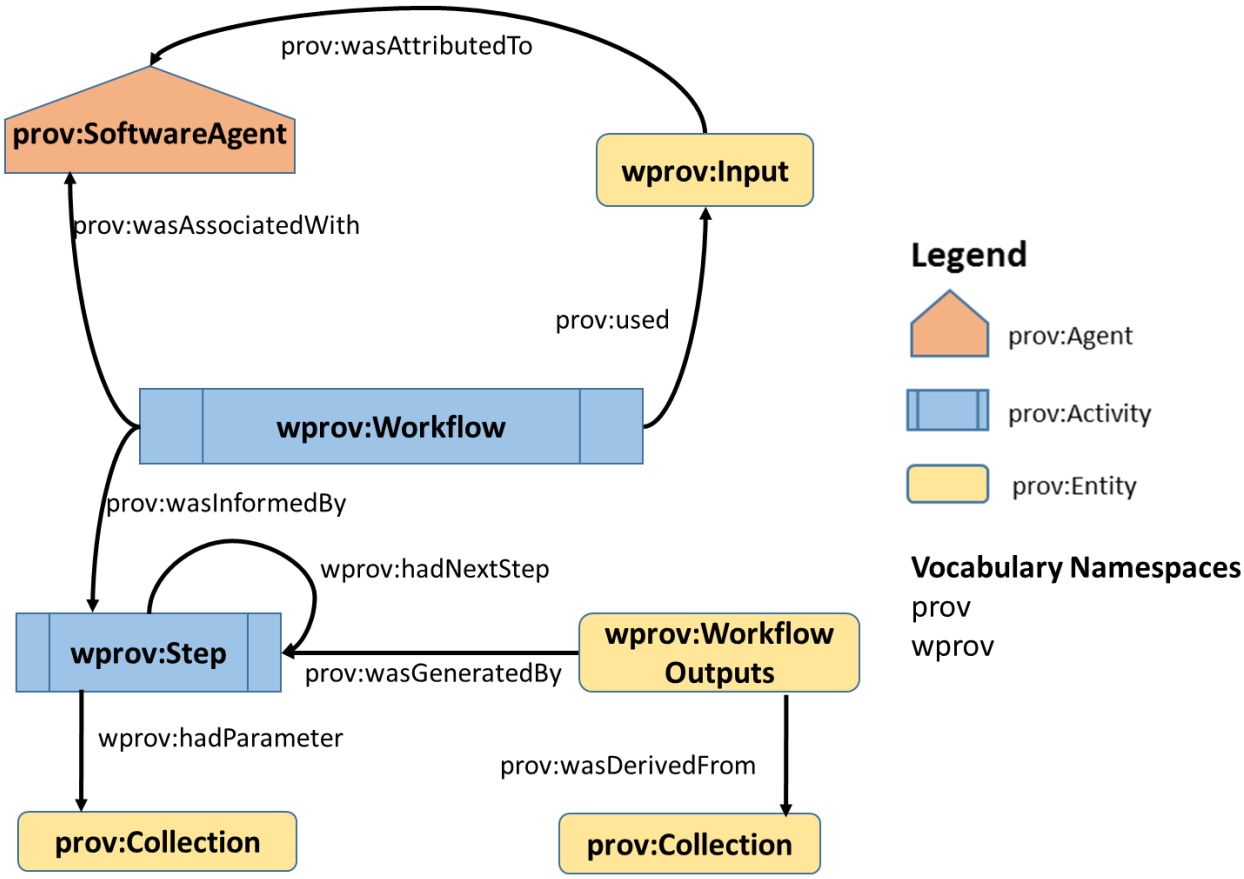

Figure 7: Workflow execution provenance design pattern aligned to PROV concepts and relations

As seen in Figure 7, for a workflow that is composed of many sub-processes, the provenance design patterns above consists of `wprov:Workflow` as an activity which starts at a certain date/time and ends after all sub-processes have been completed. The workflow activity is `prov:wasAssociatedWith` an agent, which is usually a software method controlled by a software agent. In order to get started with the workflow it needs some initial data seen above as `wprov:Input` that undergoes some manipulations through the constituent steps. Since, the workflow activity uses the data as input, these two have been connected by the `prov:used` relation.

As the workflow is a composition of workflow steps and the workflow steps contribute towards goal of the workflow, workflow and subsequent steps are linked by `prov:wasInformedBy` relation. Each `wprov:Step` has a succeeding step that follows it until the last step. This link between the workflow steps is illustrated with `wprov:hadNextStep` relation. Each step can have a relation to set of input parameters that may hold fixed values or some preset default ones that can be modified as needed. Since there are multiple parameters at play in workflows, this set of parameter collection is represented by `prov:Collection`. After the last workflow step executes, the `wprov:WorkflowOutputs` is generated as shown by `prov:wasGeneratedBy` relation.

The outputs of a workflow could be processed data, images, or some graphs all of which are concluded from some variables. As each workflow can have a set of such output variables, we used the concept `prov:Collection` to denote them. Hence, each input parameter and output

variable can be linked as a member of `prov:Collection` through `prov:hadMember` relation.

Example of provenance design pattern usage:

Figure 8 shows the use of Workflow execution provenance design pattern in the context of Water Modeling process. The **user-workflow** represents the overall workflow of implementing the water model, and it represents an aggregation of all workflow steps that needs to be completed to generate workflow results. The **user-workflow** uses water-model, in order to execute the different steps of workflow. We use the relation `prov:wasInformedBy` between the individual workflow step and overall workflow. Each succeeding workflow step is connected to the previous step by the `wprov:hadNextStep` relation.

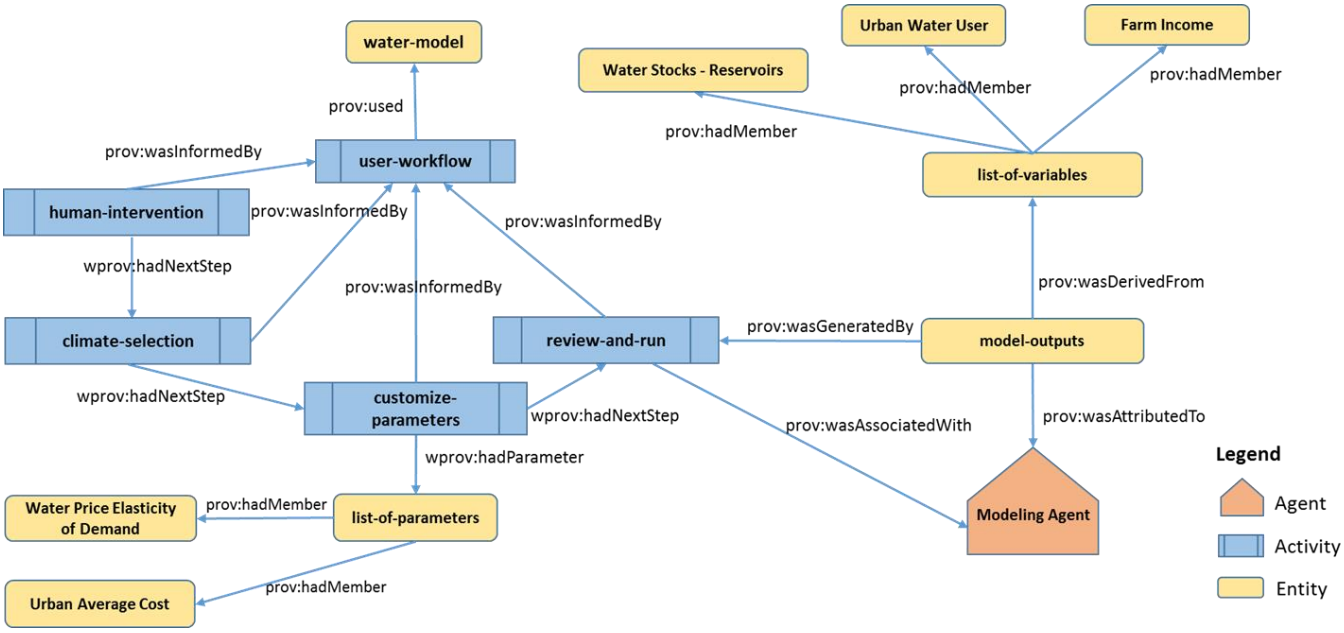

Figure 8. Workflow execution provenance design pattern used in Integrated Water Modeling Platform scenario aligned to PROV concepts and relations

In the water modeling process, the workflow step, **customize-parameters** is connected to **list-of-parameters** by the `wprov:hadParameter` relation. The **list-of-parameters** is an extension of `prov:Collection`, which resembles a collection of input parameters to the water

model. Hence, the individual parameters and parameter collection are linked with `prov:hadMember` as `prov:Collection prov:hadMember wprov:Parameter`.

One of the workflow steps is **review-and-run** activity which is handled by a **Modeling Agent**. Hence, the two concepts are linked with `prov:wasAssociatedWith` relation. Once the modeling agent completes the modeling process, certain **model-outputs** are generated. As these outputs are outcomes of **review-and-run** activity, the activity and the output are connected with `prov:wasGeneratedBy` relation. Also, as the generation of the outputs is a responsibility of the modeling agent, the **model-outputs** and **Modeling agent** are linked together via `prov:wasAttributedTo` relation. The **model-outputs** are `prov:wasDerivedFrom` **list-of-variables**, where **list-of-variables** is an extension of `prov:Collection`. Each individual output variable is a member of the list-of-variables, so these two are connected by `prov:hadMember` relation.

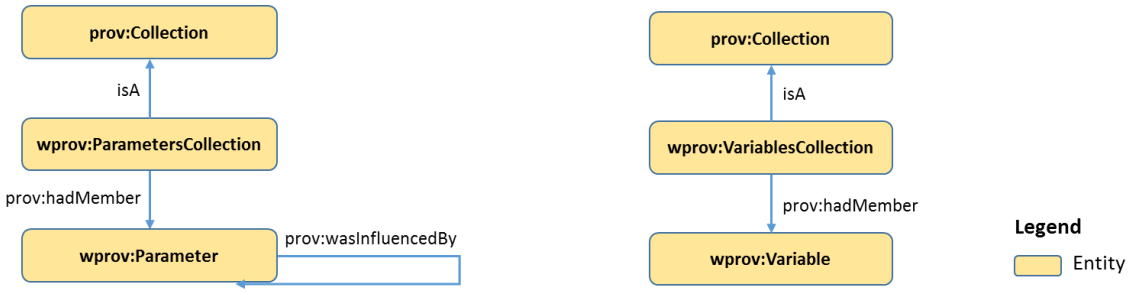

Figure 9: Membership provenance design pattern for input parameters, output variables, and dependence between parameters in integrated Water Modeling Platform scenario

The Workflow execution provenance design pattern has also been applied to Biodiversity Modeling scenario. Figure 10 shows the application of the workflow execution provenance design pattern to Biodiversity modeling domain for depicting the steps of Species distribution modeling. The **species-modeling-workflow** represents the overall workflow that is composed of multiple steps. The **species-modeling-workflow** used the **species-distribution-model** for completing the

modeling activity. We use the relation `prov:wasInformedBy` between the individual workflow step and overall workflow. Each succeeding workflow step is connected to the previous step by the `wprov:hadNextStep` relation. The first step is **environmental-data-region-selection**, which is an activity where a user selects a geographical region whose species occurrence data is available. The next step, **species-occurrence-set-selection** is another activity where a model user, is provided with a list of species whose data is available within the geographic region selected in the previous step. The next step, **modeling-algorithm-selection** is another activity where the model user chooses an algorithm for species distribution modeling. The chosen algorithm is connected to a **list-of-parameters** which is a `prov:Collection`, which resembles a collection of input parameters to the selected algorithm. Hence, the individual parameters and parameter collection are linked with `prov:hadMember` as `prov:Collection prov:hadMember wprov:Parameter`.

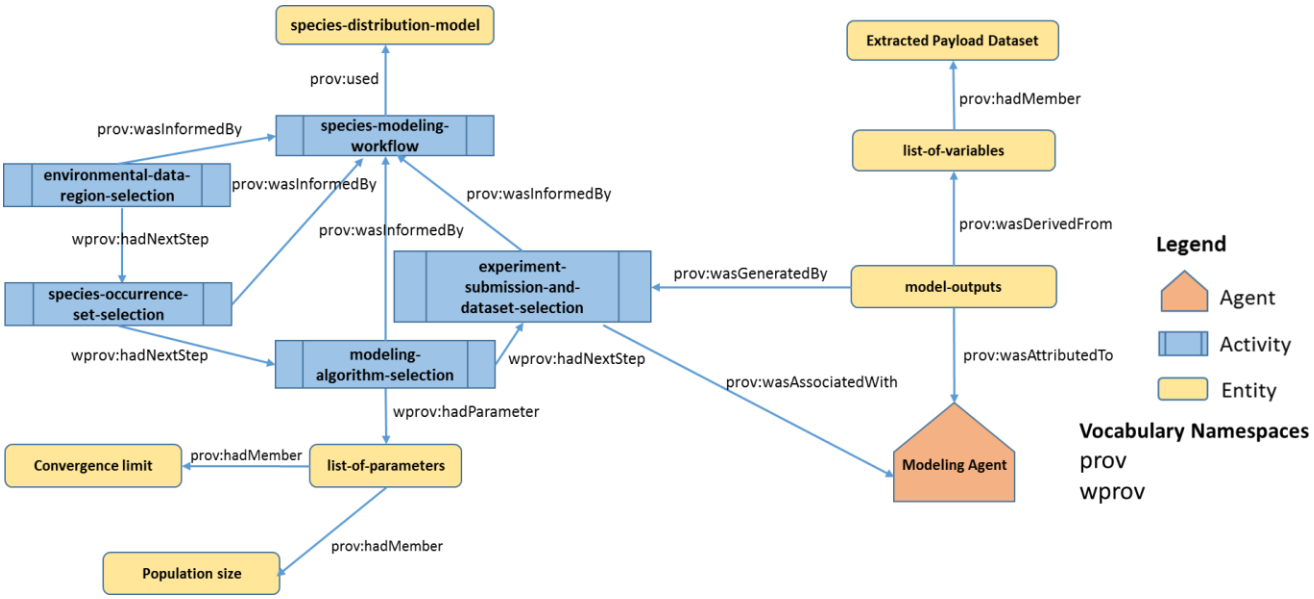

Figure 10: Workflow execution provenance design pattern used in Biodiversity Modeling scenario aligned to PROV concepts and relations

After the algorithm has been selected, the experiment needs to be submitted and model user is provided with a list of available datasets for the selected species within the geographic region. The model user then selects datasets to be used for the selected algorithm. The **experiment-submission-and-dataset-selection** activity generates the experiment results via a URL. The URL consists of the model outputs which is a `prov:Collection`. The model outputs are linked with **list-of-variables** with `prov:wasDerivedFrom` relation. The **list-of-variables**, through `prov:hadMember` relation is connected to the model outputs like **Extracted Payload Dataset**.

4.1.2 Data collection and processing provenance design pattern

This provenance design pattern deals with the process and sub processes that are carried out for data collection. The collected data acts as a data source or as an input to the workflow. The key elements for the provenance design pattern are the `Raw data, Observation` activity, and `sensors` involved in the observation. The provenance design pattern is aligned to the SSN ontology, hence the sensor concept includes sensing device and any person as well any organism (animals with attached sensors). This provenance design pattern helps in tracing the history of how the primary data was retrieved and who/what was responsible for the data collecting activity. When data sources are heterogeneous and are available in varying formats, it is important to understand the steps carried out in order to integrate them. Additionally, if the data are in different formats then some data transforming activities may need to take place such that the integrated data can be ingested by the destination activity.

As seen in Figure 11, `wprov:Observation` is depicted as an activity with a start time and end time using PROV properties `prov:startedAtTime` and `prov:endedAtTime` respectively. The observation activity is associated with different types of sensor (physical sensing

device or any living organism) with the prov:wasAssociatedWith relation. Once the observation activity is completed, the raw data is generated from the activity, using the prov:wasGeneratedBy relation between wprov:RawData and wprov:Observation. Also, as the sensor was responsible for the raw data generation, the entity, wprov:RawData is attributed to ssn:Sensor using prov:wasAttributedTo relation.

After the raw data has been obtained, certain processes are employed so that the raw data can be cleaned or extracted. Such a process can have multiple sub-processes that perform manipulations on the raw data. Since, the process uses the raw data for further processing activities, the wprov:Process is linked with wprov:RawData using prov:used property. Each further manipulating process is linked to the wprov:Process using wprov:hadSubProcess relation. Once, the sub-processes have completed the manipulating activities, the refined data is represented as wprov:DerivedData. As this derived data is generated from a Process or a number of sub-processes, the connection is shown as wprov:DerivedData prov:wasGeneratedBy wprov:Process. And, if it is required to track the original source of the derived data, we can use the prov:hadPrimarySource relation to link between derived data and the raw data.

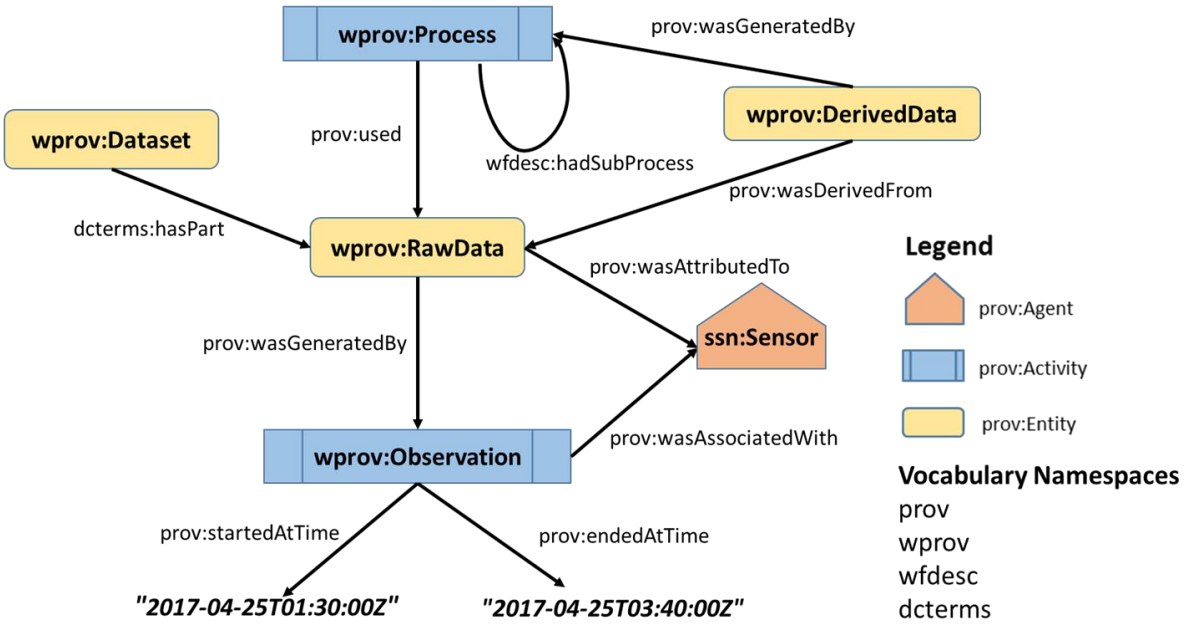

Figure 11: Data Collection and processing provenance design pattern aligned to PROV, SSN concepts and relations

Example of provenance design pattern usage:

The data collection provenance design pattern is utilized as a data measurement activity using **MODIS** sensing device and the data processing provenance design pattern has been illustrated above in the context of earth satellite imagery processing steps. The **data measurement activity** is the responsibility of **MODIS** which is a software agent. As a result of the activity, a **dataset band** with the measurement of certain resolution is produced. The relations between activity, entity, and agent are as shown in previous examples. The overall data processing has number of sub-processes such as: `Request data, Convert ASCII to TIFF, Extract and Reproject, Mosaic, Publish.`

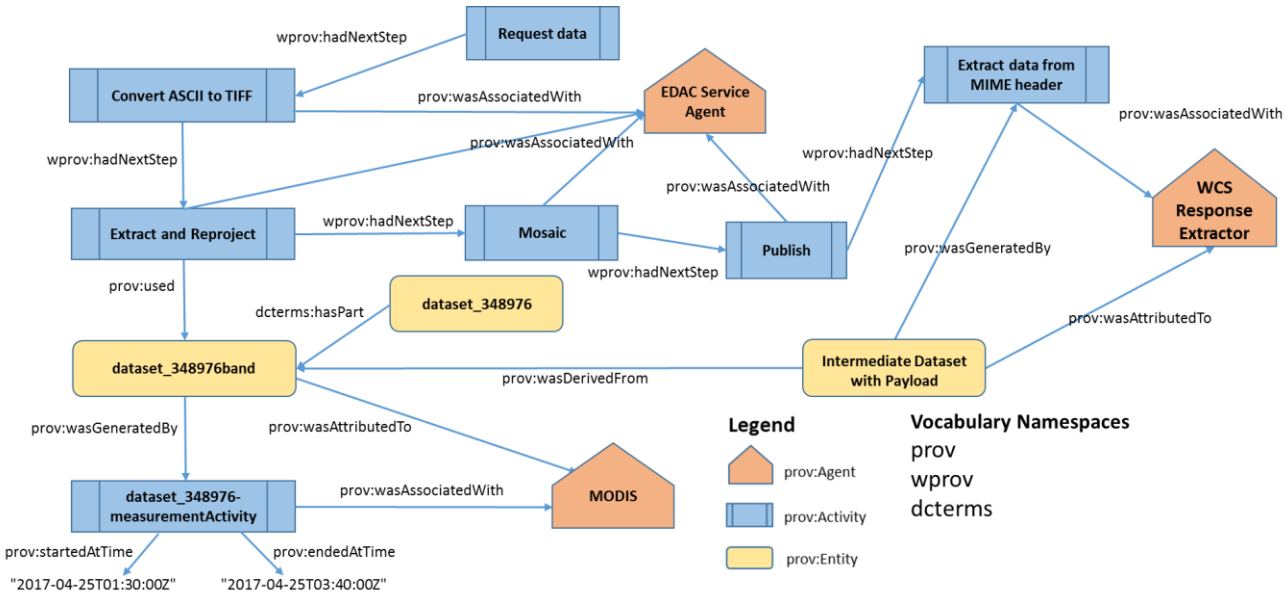

Figure 12: Data collection and processing provenance design pattern applied to Biodiversity Modeling scenario for raw data collection and processing them into processed data for ingestion by Biodiversity Model

All sub-processes are aggregated form the **Dataset Processing** activity. The first sub-process is **Request data** which sends a request for data to the EDAC. Next, the EDAC services perform format conversion as illustrated in Figure 12 from ASCII to TIFF format. Then, the **Extract and Reproject** activity uses data bands like **dataset_348976band**, and extracts the bands and reprojects the dataset from the data bands. EDAC services perform the **Mosaic** activity which includes merging of raster data bands and generating a dataset like **dataset_348976** (see Figure 12). As a result, EDAC publishes the mosaicked dataset via **Publish** activity. As the dataset with payload cannot be ingested by Lifemapper via **Extract data from MIME header** activity, the local Semantic Automated Discovery and Integration (SADI) service agent extracts the payload. Finally, SADI service agent publishes the dataset that can be used for Biodiversity modeling.

4.1.3 Data observation and sensor provenance design pattern

This provenance design pattern involves finding relationships and dependencies between Data Observation activities as `wprov:Observation` concept, the employed sensor as `ssn:Sensor`, and things that are being measured such as the property being measured as `ssn:Property`. The entity `ssn:FeatureofInterest` is the high level entity that affects the `ssn:Property`. The key elements of this provenance design pattern are: the feature of interest which symbolizes observation environment, the property that is being observed or measured, and the representation of these measured values to be used as input parameter towards a process (Figure 13).

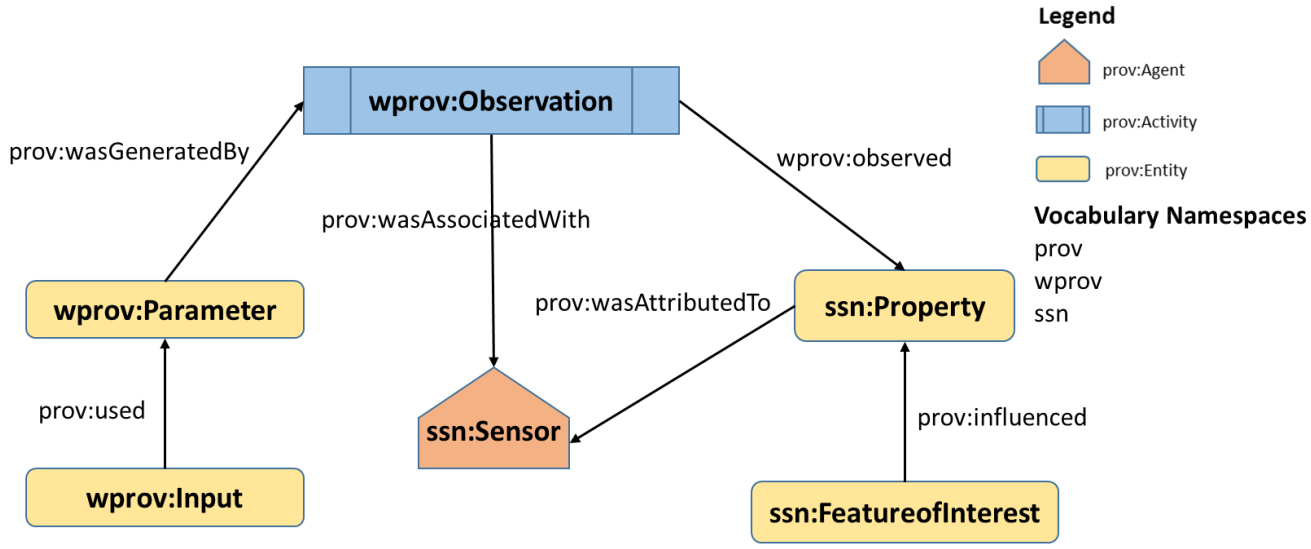

Figure 13: Data Observation Sensor provenance design pattern aligned to PROV, SSN concepts and relations

Example of provenance design pattern usage:

In the context of the smart cities scenario as explained in Section 3.3, the `wprov:Observation` concept is illustrated as an activity of measuring in Figure 13. The measurement activity observed the `ssn:property` – **light** of `ssn:FeatureofInterest` – **environment**. Furthermore, the **Measurement** and **light** are linked by `wprov:observed` relation. The surrounding **environment** influences the measured property – **light**. The luminosity sensor is responsible for the measurement activity, hence, the sensor and measurement are connected via `prov:wasAssociatedWith` relation. As the outcome of the measuring process, the **Measured variable** is used as `wprov:Input` for a process.

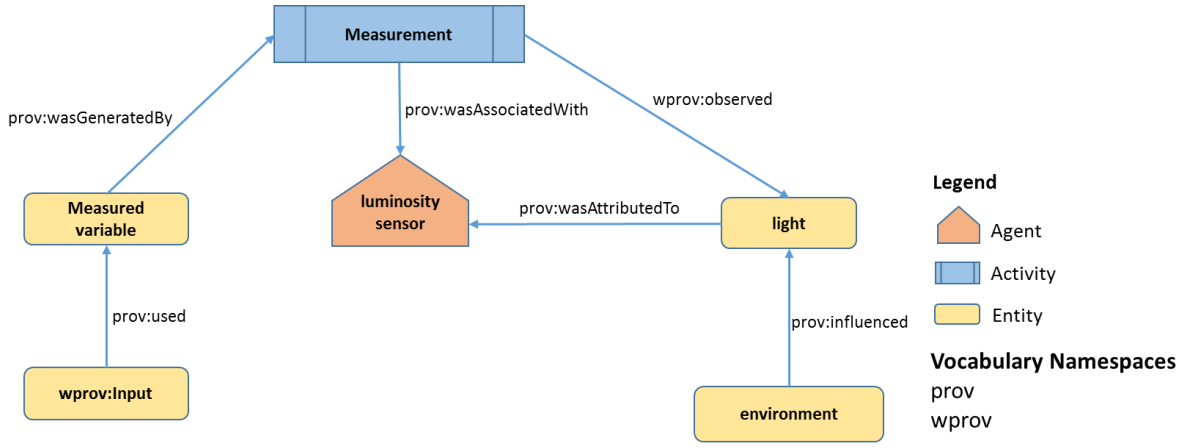

Figure 13: Data observation and sensor provenance design pattern applied to Smart Cities scenario to show the relation between measurement activity (observation), the measuring sensor (luminosity sensor), and measured property (light)

The Data observation and sensor provenance design pattern has been applied to illustrate the data observation activity performed by MODIS sensing device agent in the Biodiversity Modeling scenario as mentioned in ***Section 3.2.*** Figure 14 illustrates the application of Data observation and sensor provenance design pattern shows **MODIS** as the agent responsible for performing **dataset_348976-measurementActivity** activity. The measurement activity observed the **temperature** property. The **surface-layer** is the feature of interest that the sensor observes to measure **temperature**. As a result of the measurement activity, the data band like

dataset_348976band (satellite imagery of land cover) at specific resolution are generated. A collection of such data bands form a dataset which is later used as a `wprov:Input` to Species Distribution Model.

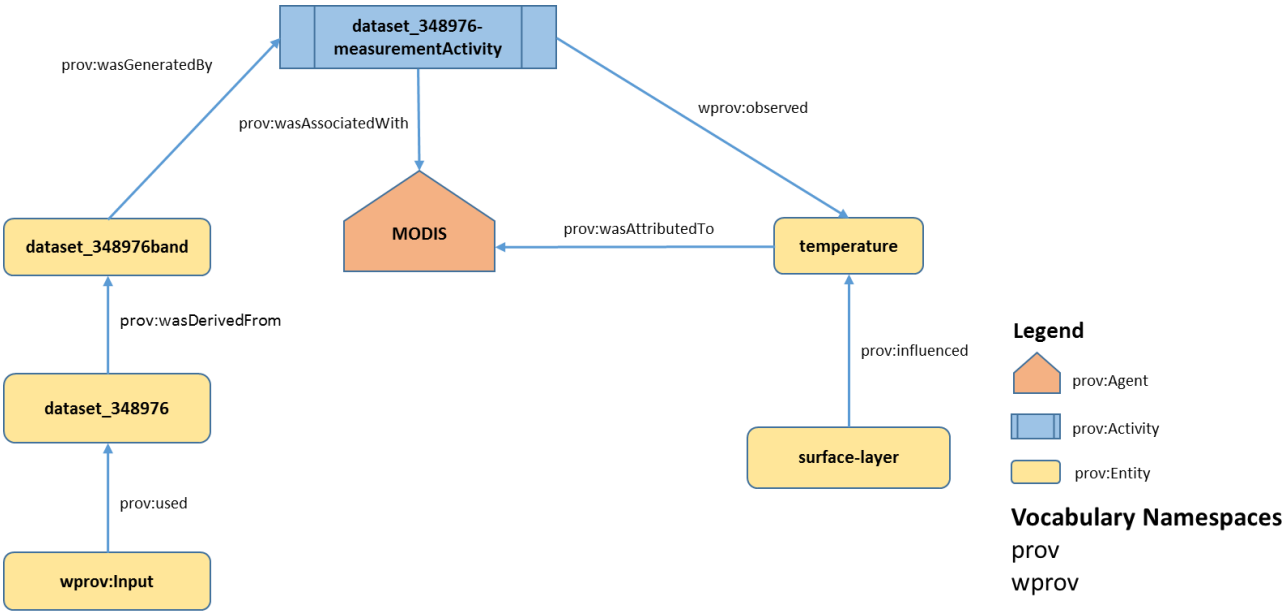

Figure 14: Data observation and sensor provenance design pattern applied to Biodiversity Modeling scenario to show the relation between dataset measurement activity (observation), the measuring sensor (MODIS), measured property (temperature)

4.2 Provenance Visualization

The Data collection and processing provenance design pattern was only applied to Biodiversity modeling scenario, as the Integrated Water Modeling Platform and Smart Cities scenario did not have any processes described in Data collection and processing provenance design pattern, at the time of writing this thesis.

In order to evaluate the Workflow execution provenance design pattern in the Integrated Water Modeling Platform, a graphical interactive representation was provided to the users (see Figure 4). After a water modeling interface user follows the steps needed to submit a water scenario, a provenance visualization shows all the history of the activities. Although, there are few

popular provenance visualization tools such as (Hoekstra & Groth, 2014) and (Anand, Bowers, & Ludäscher, 2010) a simple tool, a JSON-LD based open-source library was selected for this purpose.

In the context of Water Modeling Platform, the model data, input, and output were all represented in JSON and stored in a document-based database. So, a JSON-LD based visualization tool (scienceai, 2016) was chosen for simplicity. Although, this visualization tool uses JSON-LD structure, we use JSON model, as the water model input and output information is available via database. This tool displays JSON content in the form of graphs. The tool appeared to be appropriate for displaying simple data flow that shows the order of activities and their relationship with other concepts. The workflow execution provenance design pattern was first serialized in JSON. It was then integrated as a part of the water modeling interface (Vargas-Acosta, 2016). The visualization tool displays the provenance triples as a graph, with different size and color of the nodes. However, for the purpose of clarity and simplicity, it was modified to display all nodes on one level of the tree to be of the same size. As the visualization was planned to be used for evaluating the ease of understanding presented provenance, it was kept as straight forward as possible.

Additionally, the two provenance design patterns: Workflow execution provenance design pattern and Data observation and sensor provenance design pattern were applied to two different scenarios. The Workflow execution provenance design pattern was applied to Integrated Water Modeling Platform as well as Biodiversity Modeling scenario. In both scenarios, the provenance design pattern has been used to illustrate the overall user performed workflow and related steps. These individual steps add more information to complete the workflow. One or more steps of the workflow are usually linked to parameters which affect the workflow outputs. As a result of the

steps and respective agent that perform those activities, the resulting model outputs are illustrated in Figure 8 and Figure 10.

Similarly, the Data observation and sensor provenance design pattern was applied to Smart Cities scenario and Biodiversity Modeling scenario. First, the provenance design pattern was applied to illustrate a light measuring activity. The relations of the provenance design pattern help to find the important aspects of the activity such as the sensor responsible to perform the measurement as well as the property to be measured resulting in a measured variable. The resulting variable can later be used as an input of a model or simulation program.

Chapter 5. Survey Evaluation

The main objective of the survey was to identify the key provenance elements required by scientist to understand scientific data and workflows and consider them for reuse, i.e., data sources and how data was manipulated. In addition, we evaluated the ease of understanding the provenance design pattern applied into a specific scenarios. The key provenance elements were represented in a visualization in the Integrated Water Platform.

The evaluation survey was done by scientists working on Water Modeling in the El Paso – Juarez border area from diverse backgrounds, including: Water resources, Hydrology, Geology, Environment Science, and Computer Science. The manipulation of data in the Integrated Water Modeling Platform was explained in a short demonstration during the Water Symposium 2017 (Sustainable Water Resources for Irrigated Agriculture in a Desert River Basin Facing Climate Change and Competing Demands: From Characterization to Solutions) at Tomás Rivera Conference Center UTEP Union Building. In addition, scientists were provided with a walk-through video and then asked to respond to the survey. This survey was approved by UTEP's Institutional Review Board (IRB) (See Appendix A for IRB document).

Most of the questions in the evaluation survey were five point Likert Scale questions along with demographic questions whereas some were open-ended. The survey was provided in print-outs as well as made available online. The participants of the Water Symposium could also access the survey via a QR code made available. For all the formats that were provided, participants were provided with a consent form to be signed before participating in the survey as shown in Appendix B. The consent form presented the risks and benefits of the survey as well as emphasized on the voluntary nature of the survey.

The survey responses included 36 responders. However, 4 responses were discarded due to lack of completion. The complete survey is Appendix C. Using five point Likert Scale, the options were provided on a scale of 1-5, where 1 represents "Strongly disagree" and 5 represents "Strongly agree". Questions focused on how much the users felt the key provenance elements they were presented would facilitate the reuse of user-defined scenarios. The users were also inquired about the usability of the provenance visualization presented. The open-ended questions, as mentioned in Appendix C, were mostly a proactive attempt to figure out what the users thought needs to be included in provenance design patterns to instill trust for the model and its data.

According to the survey results (see Figure 15), we found that most of the users placed importance (i.e., agree or strongly agree) on source of data (*It is important for you to know the source of the data used within the water model)* (88%) (See Figure 15a), data manipulation (*It is important for you to know how the data was manipulated to generate a water model (workflow provenance))* (88%) (See Figure 15b), and model run replication (*It would be easier for you to replicate a water model if the provenance of the data and the workflow is provided to you*) (85%) (See Figure 15c) respectively.

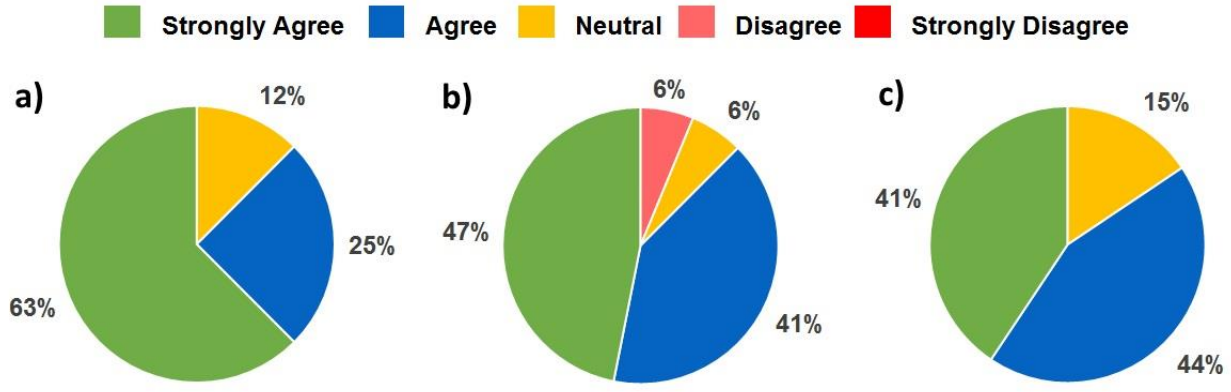

Figure 15: Percentage level of agreement on importance of a) knowing data source used for model, b) knowing data manipulation within model, and c) data and workflow provenance to reproduce a model run

Around 88% of the model interface users felt that knowing the source of the input parameters (*It would be easier for you to trust the water model projections if the sources for the model parameters were provided (Example: Links to paper or other related studies websites))* was important (see Figure 16a), 69% were willing to annotate data sources (*You will be willing to spend additional time annotating data sources and workflows so that other scientists could reuse them*) (see Figure 16b). On the criterion of user-friendliness of provenance visualization (*The provenance shown to you on the Integrated Water Modeling Platform website was easy to understand*), 69% users felt it was easy to follow (see Figure 16c).

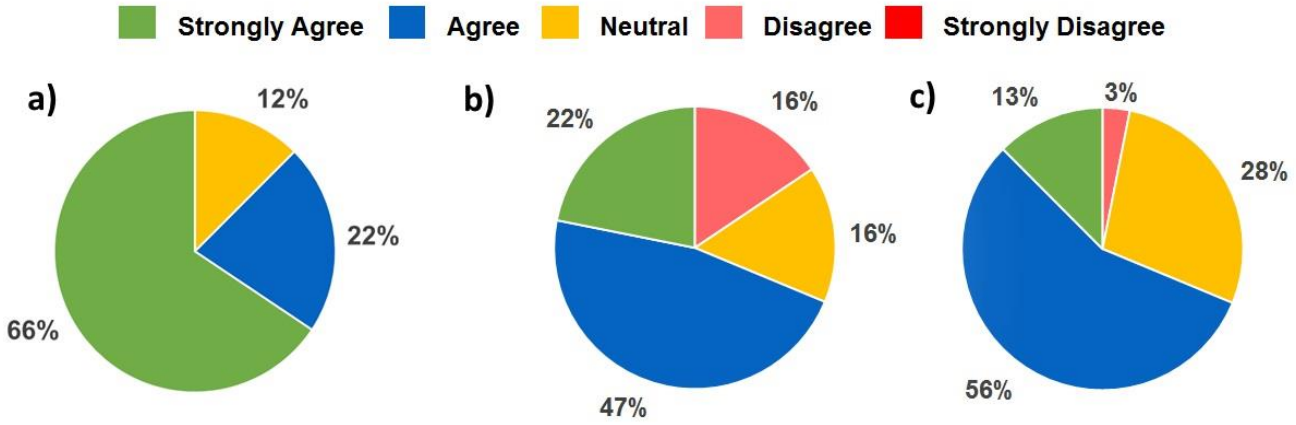

Figure 16: Percentage level of agreement on a) importance of model input parameter sources for trusting water model, b) will to annotate data for future reuse by other scientists, and c) ease of using and understanding provenance visualization

Although 60% users did not see the use of debugging abilities using displayed provenance (*It will be easier for you to find error in an experiment using the provided data and workflow provenance? (Example: Using an incorrect default value or missing a step in workflow))* (see Figure 17a), 81% users concluded that it increases the amount of trust they had on the water model (*The data and model provenance increase your trust to use or reproduce a water model*) (see Figure 17b).

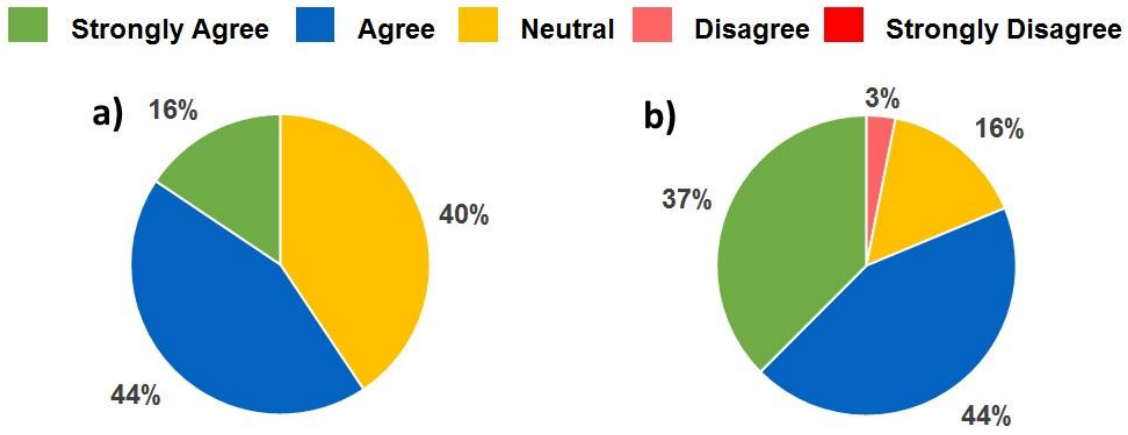

Figure 17: Percentage level of agreement on a) ease of workflow debugging through use of provenance provided, and b) increase in trust to use water model due data and model provenance respectively

Figure 18 shows survey responses based on the education level, the trend of cumulative responses were found to follow a similar design between the Master and Undergraduate level students. The PhD students provided few responses with "Strongly agree".

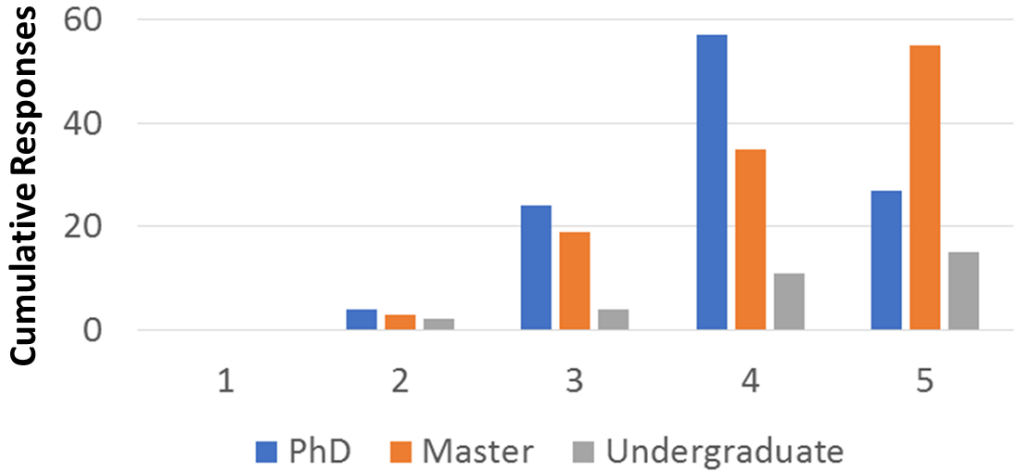

Figure 18: Cumulative Likert scale responses from responders of different education levels

Chapter 6. Discussion and Conclusion

The first objective of this work was to investigate the main provenance elements inspected by scientist when understanding and considering data and workflows for reuse. The most important provenance elements for scientists were surveyed through literature review and the survey questions: *It is important for you to know the source of the data used within the water model, It is important for you to know how the data was manipulated to generate a water model (workflow provenance)), It would be easier for you to trust the water model projections if the sources for the model parameters were provided (Example: Links to paper or other related studies websites))*.

The provenance elements: data sources, data manipulation, and input parameter source were included in the specialized provenance design patterns proposed in this work, namely - "Workflow execution provenance design pattern", "Data collection and processing provenance design pattern", and "Data observation and sensor provenance design pattern". These three provenance design patterns describe scientific data and workflow provenance using controlled vocabularies. Such designs act as basic building blocks (i.e., a pattern or template) to describe provenance concepts instead of describing them in terms of recommended W3C PROV standard from scratch. The outcome of the second objective of this work contributed in developing provenance design patterns that were specialized from PROV concepts and relations. The controlled vocabularies utilized by the current provenance-aware applications were incorporated in this research. In order to keep the designs reusable and extensible, the ontological commitments were considered in minimal capacity. The provenance designs were aligned to PROV vocabulary as well as other relevant controlled vocabularies. The three proposed designs were applied to three

distinct scenarios to achieve the third objective of this work, namely the Integrated Water Modeling Platform, Biodiversity Modeling, and Smart Cities.

Of the three provenance design patterns, two were applied to two different scenarios. The Workflow execution provenance pattern was applied to Integrated Water Modeling Platform scenario and Biodiversity Modeling. The Workflow execution provenance design pattern was implemented by annotating the provenance in JSON format for Water model provenance visualization. The Data observation sensor provenance design pattern was applied to Smart Cities sensor related observation provenance description as well as Biodiversity Modeling sensor description. The Data collection and processing provenance design pattern was only applied to Biodiversity Modeling scenario due to unavailability of processes that could be illustrated as provenance activities. This work illustrates the use of provenance design patterns for the purpose of identifying the key provenance elements for the scientists. As the proposed provenance design patterns were reused in different scenarios, this shows the applicability of such patterns to facilitate the understanding of scientific data and workflows.

The Workflow execution provenance design pattern was then implemented to Integrated Water Modeling Platform. The implementation of the first provenance design pattern proposed enabled the evaluation of this pattern by scientists across disciplines, states and countries working on water modeling issues.

Based on the survey responses received from these representative users of the Integrated Water Modeling Platform, it was found that majority of responders strongly agreed about the importance of knowing source data and their manipulation. Furthermore, Masters and Undergraduate students found that the developed provenance visualization was fairly easy to understand and explore than that for the PhD students. This fact indicated need of additional

include additional questions in the survey to get a more specific input about user's understanding of provenance. The survey responses to the open-ended questions indicated various usage scenarios of the proposed designs such as: finding related work in a convenient manner, testing the effect of changing parameter value in model outputs, and citing publication. The results shows the respondents interest of having a provenance trace when analyzing a scientific product but some respondents expressed disagreement on their willingness to invest additional time to learn and include provenance trace in their own systems that serve as data sources or scientific workflow systems.

Chapter 7. Future Work

This research was based on exploring the main information required by scientists to understand scientific data and workflows and evaluate them for potential reuse.

Based on the provenance representation to different interdisciplinary scenarios and the survey result evaluation, future work will require additional and ready-to-use design patterns and their serialization into common formats (e.g., JSON-LD, databases, RDF) so that the scientists are encouraged to annotate data sources in proactive manner. In order to make a visible impact on the usage of provenance standards, the provenance needs to be published as linked open data along with the data itself. This is a first step towards filling the gap between the creators of standards and controlled vocabularies and practitioners that require best practices, suggested provenance design patterns, and serialization of provenance traces for specific applications, e.g., provenance visualization.

Furthermore, provenance visualization was well received by survey respondents and work towards presenting provenance in different manner depending upon the role of the user in a specific scenario could be pursued. For example, if the user is a scientist display the visualization with model or experiment specific terminologies as input and output. On the other hand, if the user is a stakeholder unfamiliar with the model terminologies, display the key provenance elements in a simpler manner. Similarly, additional surveys needs to be done and provenance design patterns should be reevaluated with stakeholders from different level of expertise. Such surveys need different questionnaire that is suitable for a variety of stakeholders.

References

Alper, P., Belhajjame, K., Goble, C. A., & Karagoz, P. (2014). LabelFlow: Exploiting Workflow Provenance to Surface Scientific Data Provenance. In *Provenance and Annotation of Data and Processes* (pp. 84–96). Springer, Cham. https://doi.org/10.1007/978-3-319-16462-5_7

Anand, M. K., Bowers, S., & Ludäscher, B. (2010). Provenance browser: Displaying and querying scientific workflow provenance graphs. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)* (pp. 1201–1204). https://doi.org/10.1109/ICDE.2010.5447741

Belhajjame, K., Goble, C., Soiland-Reyes, S., & Roure, D. D. (2011). Fostering Scientific Workflow Preservation through Discovery of Substitute Services. In *2011 IEEE Seventh International Conference on eScience* (pp. 97–104). https://doi.org/10.1109/eScience.2011.22

Belhajjame, Khalid, Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., … Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*, *32*, 16–42. https://doi.org/10.1016/j.websem.2015.01.003

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 28–37.

Blomqvist, E., Hitzler, P., Janowicz, K., Krisnadhi, A., Narock, T., & Solanki, M. (2016). Considerations regarding ontology design patterns. *Semant. Web*, *7*(1), 1–7.

Cavner, J. A., Stewart, A. M., Grady, C. J., & Beach, J. H. (2012). An innovative Web Processing Services based GIS architecture for global biogeographic analyses of species distributions. *OSGeo Journal*, *10*(1), 11.

Community:ListPatterns - Odp. (n.d.). Retrieved April 19, 2017, from http://ontologydesignpatterns.org/wiki/Community:ListPatterns

Consoli, S., Mongiovic, M., Nuzzolese, A. G., Peroni, S., Presutti, V., Reforgiato Recupero, D., & Spampinato, D. (2015). A Smart City Data Model based on Semantics Best Practice and Principles (pp. 1395–1400). ACM Press. https://doi.org/10.1145/2740908.2742133

Corsar, D., Edwards, P., Baillie, C., Markovic, M., Papangelis, K., & Nelson, J. (2013). Short Paper: Citizen Sensing Within a Real-time Passenger Information System. In *Proceedings of the 6th International Conference on Semantic Sensor Networks - Volume 1063* (pp. 77–82). Aachen, Germany, Germany: CEUR-WS.org. Retrieved from http://dl.acm.org/citation.cfm?id=2874543.2874550

Cyganiak, R., Wood, D., & Lanthaler, M. (2014, February). RDF 1.1 Concepts and Abstract Syntax. Retrieved October 17, 2016, from https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#referents

Gangemi, A. (2005). Ontology Design Patterns for Semantic Web Content. In Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen (Eds.), *The Semantic Web – ISWC 2005* (pp. 262–276). Springer Berlin Heidelberg. https://doi.org/10.1007/11574620_21

Garijo, D., Gil, Y., & Corcho, O. (2014). Towards Workflow Ecosystems through Semantic and Standard Representations. In *2014 9th Workshop on Workflows in Support of Large-Scale Science (WORKS)* (pp. 94–104). https://doi.org/10.1109/WORKS.2014.13

Garijo, Daniel, & Gil, Y. (2012). Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. In *Proceedings of the 2nd International Workshop on Linked Science* (Vol. 951, pp. 0–0). Boston, USA: Facultad de Informática (UPM). Retrieved from http://ceur-ws.org/Vol-951/

Groth, P., & Beek, W. (2016). Measuring PROV Provenance on the Web of Data.

Groth, P., & Moreau, L. (2013, April). PROV-Overview. Retrieved April 19, 2017, from https://www.w3.org/TR/prov-overview/

Hoekstra, R., & Groth, P. (2014). PROV-O-Viz - Understanding the Role of Activities in Provenance. In *Provenance and Annotation of Data and Processes* (pp. 215–220). Springer, Cham. https://doi.org/10.1007/978-3-319-16462-5_18

Hu, Y., Janowicz, K., Carral, D., Scheider, S., Kuhn, W., Berg-Cross, G., … Kolas, D. (2012). A Geo-ontology Design Pattern for Semantic Trajectories. In *SpringerLink* (pp. 438–456). Springer International Publishing. https://doi.org/10.1007/978-3-319-01790-7_24

JSON. (n.d.). Retrieved November 21, 2016, from http://www.json.org/

Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: the state of the art. *The Knowledge Engineering Review*, *18*(1), 1–31. https://doi.org/10.1017/S0269888903000651

Koop, D., Santos, E., Bauer, B., Troyer, M., Freire, J., & Silva, C. T. (2010). Bridging Workflow and Data Provenance Using Strong Links. In *Scientific and Statistical Database Management* (pp. 397–415). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13818-8_28

Koop, D., Santos, E., Mates, P., Vo, H. T., Bonnet, P., Bauer, B., … Silva, C. T. (2011). A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers. *Procedia Computer Science*, *4*, 648–657. https://doi.org/10.1016/j.procs.2011.04.068

Kwasnikowska, N., Moreau, L., & Bussche, J. V. D. (2015). A Formal Account of the Open Provenance Model. *ACM Trans. Web*, *9*(2), 10:1–10:44. https://doi.org/10.1145/2734116

Lifemapper. (2017). Retrieved May 16, 2017, from http://lifemapper.org/

Mao, M. (2007). Ontology Mapping: An Information Retrieval and Interactive Activation Network Based Approach. In *The Semantic Web* (pp. 931–935). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-76298-0_72

Markovic, M., Edwards, P., Kollingbaum, M., & Rowe, A. (2016). Modelling Provenance of Sensor Data for Food Safety Compliance Checking. In *International Provenance and Annotation Workshop* (pp. 134–145). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-40593-3_11

Missier, P., Dey, S., Belhajjame, K., Cuevas-Vicenttin, V., & Ludäscher, B. (2013). D-PROV: Extending the PROV Provenance Model with Workflow Structure. Presented at the 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP 13). Retrieved from https://www.usenix.org/conference/tapp13/technical-sessions/presentation/missier

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

Moreau, L., & Groth, P. (2013). Provenance: An Introduction to PROV. Morgan & Claypool.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

OWL - Semantic Web Standards. (n.d.). Retrieved November 7, 2016, from https://www.w3.org/OWL/

Sahoo, S., Raymer, M., Henson, C., Sheth, A., & York, W. (2008). Ontology Driven Semantic Provenance for Heterogeneous Bionomics Experimental Data. *Kno.e.sis Publications*. Retrieved from http://corescholar.libraries.wright.edu/knoesis/122

Salayandia, L. (2012). *Ontologies for scientific data transformation* (Ph.D.). The University of Texas at El Paso, United States -- Texas. Retrieved from http://0-search.proquest.com.lib.utep.edu/pqdtglobal/docview/1294054342/abstract/247705612C524ECEPQ/1

scienceai. (2016). scienceai/jsonld-vis. Retrieved May 2, 2017, from https://github.com/scienceai/jsonld-vis

Shu, Y., Taylor, K., Hapuarachchi, P., & Peters, C. (2012). Modelling provenance in hydrologic science: a case study on streamflow forecasting. *Journal of Hydroinformatics*, *14*(4), 944–959. https://doi.org/10.2166/hydro.2012.134

Taylor, K., Janowicz, K., Phuoc, D. L., & Haller, A. (2017, January). Semantic Sensor Network Ontology. Retrieved May 1, 2017, from https://www.w3.org/TR/vocab-ssn/

The ProvONE Data Model for Scientific Workflow Provenance. (n.d.). Retrieved April 20, 2017, from http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html

Umuhoza, D., & Braun, R. (2012). Trustworthiness Assessment of Knowledge on the Semantic Sensor Web by Provenance Integration. In *2012 26th International Conference on Advanced Information Networking and Applications Workshops* (pp. 387–392). https://doi.org/10.1109/WAINA.2012.197

Vargas-Acosta, R. A. (2016). Model UI. Retrieved May 2, 2017, from https://water.cybershare.utep.edu/bucket_2/create#

Villanueva-Rosales, N., Rio, N. del, Pennington, D., & Chavira, L. G. (2015). Semantic Bridges for Biodiversity Sciences. In M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, … S. Staab (Eds.), *The Semantic Web - ISWC 2015* (pp. 310–317). Springer International Publishing. https://doi.org/10.1007/978-3-319-25010-6_20

Ward, F. A. (2016). Bucket Model UI. Retrieved April 30, 2017, from https://water.cybershare.utep.edu/bucket_2/home

Appendix A

THE UNIVERSITY OF TEXAS AT EL PASO
Office of the Vice President for Research and Sponsored Projects
Institutional Review Board
El Paso, Texas 79968-0587
phone: 915 747-8841 fax: 915 747-5931

FWA No: 00001224

DATE:	December 13, 2016
TO:	Natalia Villanueva-Rosales, Doctoral
FROM:	University of Texas at El Paso IRB
STUDY TITLE:	[998706-1] Data Presentation and Usability Testing for the Integrated Water Modeling Interface.
IRB REFERENCE #:	College of Engineering
SUBMISSION TYPE:	New Project
ACTION:	DETERMINATION OF EXEMPT STATUS
DECISION DATE:	December 12, 2016
REVIEW CATEGORY:	45 CFR 46.101(b)(2)

Thank you for your submission of New Project materials for this research study. University of Texas at El Paso IRB has determined this project is EXEMPT FROM IRB REVIEW according to federal regulations.

Exempt protocols do not need to be renewed. Please note that it is the Principal Investigator's responsibility to resubmit the proposal for review if there are any modifications made to the originally submitted proposal. This review is required in order to determine if "Exemption" status remains.

We will put a copy of this correspondence on file in our office.

If you have any questions, please contact the IRB Office at (915) 747-8841 or irb.orsp@utep.edu. Please include your study title and reference number in all correspondence with this office.

cc:

Appendix B

Data Presentation and Usability Testing for the Integrated Water Modeling Interface

Information and Consent Form

Introduction:
This research study is being supervised by Dr. Natalia Villanueva-Rosales, assistant professor at The University of Texas at El Paso. Please read this form and ask questions before you agree to form part of the research activities.

Background Information:
The goal of this study is to evaluate the Integrated Water Modeling Interface for the USDA sponsored Sustainable Water Resources Project. The evaluated interface forms part of the project website at https://water.cybershare.utep.edu/bucket_2.

The usability test objectives are:

- To determine design inconsistencies and usability problem areas within the user interface and content areas. Potential sources of error may include:
    - Navigation errors – failure to locate functions, failure to follow recommended modeling workflow.
    - Presentation errors – failure to locate and properly act upon desired information in screens, selection errors due to labeling ambiguities.
    - Control usage problems – improper toolbar or entry field usage.
- Collected data will be used to access whether usability goals regarding an effective, efficient, and well-received user interface have been achieved.
- Establish baseline user performance and user-satisfaction levels of the user interface for future usability evaluations.

The data presentation pattern test objectives are:

- To evaluate the ease of understanding of the data presentation patterns in helping the user to
    - Trust the inputs that are used within the water model.
    - Understand the data manipulation within the water model.
- Exercise the Water Modeling Data Presentation Pattern under controlled test conditions with representative users. Data will be used to access whether understanding goals regarding a clear and reproducible data presentation pattern have been achieved.

Procedures:
As a participant you will complete a set of assigned tasks as efficient and timely a manner as possible, and provide feedback regarding the usability and acceptability of the interface. You will be directed to provide honest opinions regarding the usability of the application, and to participate in a post-session subjective survey.

A laptop computer with the web application and supporting software will be assigned to you in a typical research environment. Your interaction with the web application will be monitored by the facilitator seated in the same room. Your face will not be recorded for the purpose of this study. The purpose of this study is on evaluating the application, rather than the facilitator evaluating you.

Risks and Benefits:
There is no risk associated with individual's participation during this research.

Compensation:
There is no compensation associated with this research.

Confidentiality:
Your participation in this research will be confidential. In any written reports or publications, no one will be identified or identifiable and only group data will be presented. Your name will not be recorded in any document. Only three demographic questions will be considered in the post-test survey: age, education level and area of study.

Voluntary Nature of the Research:
Your decision whether or not to participate will not affect your future relations with The University of Texas at El Paso in any way.

Contacts and Questions:
If you have any questions, please feel free to contact Dr. Natalia Villanueva, at **nvillanuevarosales@utep.edu**. You may ask questions now, or if you have any additional questions later, we will be happy to answer them.

You may keep a copy of this form for your records

Statement of Consent:

You are making a decision whether or not to participate on activities during the usability study session of The Integrated Water Modeling Platform. Your initials indicate that you have read this information and your questions have been answered. Even after signing this form, please know that you may withdraw from the research at any time.

_____

I consent to participate in the research.

_____

Participant Initials Date

_____

Signature of Researcher Date

Appendix C

The Integrated Water Modeling Interface

Usability and Data Presentation Pattern Testing

Survey

Demographics

1. Age:

2. Education Level:

3. Area of Study:

Background

On a scale of 1 to 5, 1 being "Strongly Disagree" and 5 being "Strongly Agree", provide an agreement level for the following statement:

4. I am very confident when it comes to the use of computers.
   - 1- Strongly Disagree
   - 2 - Disagree
   - 3 - Neutral
   - 4 - Agree
   - 5 - Strongly Agree

5. Did you have formal or self-training?

Feelings after using the interface

On a scale of 1 to 5, 1 being "Strongly Disagree" and 5 being "Strongly Agree", provide an agreement level for the following statements:

6. The use of the interface was easy most of the time, i.e. there was no confusing instructions.
   - ○ 1- Strongly Disagree
   - ○ 2 - Disagree
   - ○ 3 - Neutral
   - ○ 4 - Agree
   - ○ 5 - Strongly Agree

7. I felt in control when using the water modeling interface.
   - ○ 1- Strongly Disagree
   - ○ 2 - Disagree
   - ○ 3 - Neutral
   - ○ 4 - Agree
   - ○ 5 - Strongly Agree

8. The interface was user friendly.
   - ○ 1- Strongly Disagree
   - ○ 2 - Disagree
   - ○ 3 - Neutral
   - ○ 4 - Agree
   - ○ 5 - Strongly Agree

9. Workflow to submit a custom scenario was easy to follow.
   - ○ 1- Strongly Disagree
   - ○ 2 - Disagree
   - ○ 3 - Neutral
   - ○ 4 - Agree

- 5 - Strongly Agree

Aspects in the interface design

On a scale of 1 to 5, 1 being "Strongly Disagree" and 5 being "Strongly Agree", provide an agreement level for the following statements:

10. Names in menus and buttons are meaningful.
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

11. Error and instruction messages were helpful and easy to understand.
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

12. The established workflow order is appropriate for understanding and completing of the overall goal.
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

Final Evaluation

13. What would be your overall score for the graphical user interface?
    - Excellent
    - Good

- Average
- Fair
- Poor

14. Suggestions for improvement.

Feelings after using the Data Presentation visualization

On a scale of 1 to 5, 1 being "Strongly Disagree" and 5 being "Strongly Agree", provide an agreement level for the following statements:

15. It is important for you to know the source of the data used within the water model.
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

16. It is important for you to know how the data was manipulated to generate a water model (workflow provenance).
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

17. It would be easier for you to replicate a water model if the provenance of the data and the workflow is provided to you.
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree

- 5 - Strongly Agree

18. It would be easier for you to trust the water model projections if the sources for the model parameters were provided? (Example: Links to paper or other related studies websites)

    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

19. You will be willing to spend additional time annotating data sources and workflows so that other scientists could reuse them.

    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

20. The provenance shown to you on the Integrated Water Modeling Platform website was easy to understand.

    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

21. It will be easier for you to find error in an experiment using the provided data and workflow provenance? (Example: Using an incorrect default value or missing a step in workflow).

    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral

- 4 - Agree
- 5 - Strongly Agree

22. The data and model provenance increase your trust to use or reproduce a water model.
    - 1- Strongly Disagree
    - 2 - Disagree
    - 3 - Neutral
    - 4 - Agree
    - 5 - Strongly Agree

Area of improvement

23. In addition to trusting and reproducing a water model, in which other ways would you use the data and model provenance?

24. Is there any missing information that you would want to see in order to trust the water model presented on the Integrated Water Modeling Interface website? (If so, please list the missing information)

25. Do you have any suggestions for how we could improve the visualization of the data and model provenance on the Integrated Water Modeling Interface website?

Vita

Smriti Rajkarnikar Tamrakar was born in Lalitpur, Nepal. She graduated with her Bachelor's degree in Information Technology from Nepal College of Information Technology. She worked in the industry for five years gathering experiences with local companies as well as enterprises based in Netherland and Germany. She began her further studies at The University of Texas at El Paso in 2014 to pursue a Master's degree in Computer Science.

Contact Information: srajkarnikartamrakar@miners.utep.edu

This thesis was typed by Smriti Rajkarnikar Tamrakar.