FACULTAD DE INFORMÁTICA
UNIVERSIDAD POLITÉCNICA DE MADRID

MASTER THESIS

MASTER IN ARTIFICIAL INTELLIGENCE RESEARCH

A PROVENANCE-ENABLED SERVICE FOR NEWS AND BLOG AGGREGATION

AUTHOR: DANIEL GARIJO VERDEJO
SUPERVISOR: OSCAR CORCHO GARCÍA
CO-SUPERVISOR: BORIS VILLAZÓN TERRAZAS

FEBRUARY, 2011

Acknowledgments

This work would not have been possible without the help of my supervisor Oscar Corcho García and my co-supervisor Boris Villazón Terrazas, who have suggested related bibliography, provided feedback and even explained some of the technologies used for the service design and implementation.

I would like to thank also Asuncion Gómez Pérez for providing ideas for the evaluation and additional comments, Raúl García for granting some motivational examples, Vanessa Barrero and José Cerezo for giving the access to the PRISACOM's files, María Poveda for answering my doubts about ontology design and my family (specially my mother, Mª Felisa Verdejo), colleagues from the OEG and my friends for their support during this year.

The research described in this work has been funded by the Spanish project WebN+1, under Plan Avanza I+D 2009 (TSI-020301-2009-24, Ministerio de Industria, Turismo y Comercio).

Resumen

Actualmente existe una gran cantidad de información publicada en la Web en forma de noticias online y blogs. Los usuarios que requieren usar dicha información para distintos propósitos (desde intereses personales a toma de decisiones profesionales) necesitan saber de su procedencia y su evolución con el fin de determinar su calidad y aplicarle un determinado grado de confianza.

Se han desarrollado diversos modelos de *provenance* con el propósito de representar y gestionar el historial de los contenidos en una gran cantidad de dominios distintos. Sin embargo aplicarlos en escenarios reales plantea retos metodológicos y tecnológicos aún no resueltos, tal y como ha señalado el *W3C Provenance Incubator Group*. En esta tesis definimos un marco de comparación para analizar las distintas propuestas, y justificamos la selección de la más apropiada, el "*Open Provenance Model*" (OPM), como referencia para modelar un escenario real de noticias online y blogs, en un contexto de turismo, para una de las empresas de comunicaciones y publicaciones más importantes de nuestro país: PRISACOM. Además, en este documento se define un servicio de anotación y recuperación de *provenance* usando el modelo previo como referencia, describiendo las decisiones de modelado y de diseño efectuadas en el contexto del uso y gestión de *provenance* para una serie de plataformas pertenecientes a la compañía. Finalmente, se presenta la evaluación realizada, con unos resultados prometedores que solucionan los retos planteados en los objetivos del proyecto. Señalar también que nuestro caso de uso es una contribución adicional, porque muestra cómo el OPM puede ser utilizado fuera de dominios científicos, que es donde ha sido comúnmente aplicado hasta el momento.

Durante el periodo en el que se ha realizado este trabajo, el autor ha sido miembro del *W3C Provenance Incubator Group* (realizando múltiples contribuciones a su informe final[1]), y ha participado en la discusión acerca de los *mappings* entre los modelos de provenance de mayor aceptación en la comunidad científica[2].

[1] "Final Report of the W3C Provenance Incubator Group." Yolanda Gil, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau, and Paolo Pinheiro da Silva. Report from the W3C Provenance Incubator Group, first release: November 30, 2010. Available from http://www.w3.org/2005/Incubator/prov/wiki/Final_Report_Draft

[2] "Provenance Vocabulary Mappings." Satya Sahoo, Paul Groth, Olaf Hartig, Simon Miles, Sam Coppens, James Myers, Yolanda Gil, Luc Moreau, Jun Zhao, Michael Panzer, and Daniel Garijo. Report from the W3C Provenance Incubator Group, first release: August 6, 2010. Available from http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

Abstract

There is a large amount of information published on the Web in the form of online news and blogs. Users willing to use this information for different purposes (ranging from personal interest to professional decision-making) need to know about its pedigree. This includes aspects like its origin, its evolution in time, etc., and it is the basis to determine their overall quality and their degree of trust on them.

Provenance models have been generated for the purpose of representing and managing this pedigree in a wide range of domains. However, their application in real case scenarios raise methodological and technological challenges, as the W3C Provenance Incubator Group has pointed out. In this thesis we define an evaluation framework for analyzing these models, and we select the most appropriate one, the Open Provenance Model (OPM), for modeling a real word scenario of online news and blogs, in the domain of travelling, for a major communication and editorial company in Spain: PRISACOM. Furthermore, we define a service for provenance annotation and retrieval using the previous model as reference, describing the modeling and design decisions made in the context of provenance usage and the main issues encountered during this development for a variety of platforms belonging to the aforementioned company. Finally, we present the evaluation, showing promising results and addressing the challenges proposed in the objectives of the project. Our use case is also an additional contribution, illustrating how OPM can be applied out of the scientific workflows and research fields, where it has been more commonly applied.

Along the period of this work the author has been a member of the W3C Provenance Incubator Group (with several contributions to its Final Report[3]), and has participated in the discussion of the mappings between several popular provenance models[4].

---

3 "Final Report of the W3C Provenance Incubator Group." Yolanda Gil, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau, and Paolo Pinheiro da Silva. Report from the W3C Provenance Incubator Group, first release: November 30, 2010. Available from http://www.w3.org/2005/Incubator/prov/wiki/Final_Report_Draft

4 "Provenance Vocabulary Mappings." Satya Sahoo, Paul Groth, Olaf Hartig, Simon Miles, Sam Coppens, James Myers, Yolanda Gil, Luc Moreau, Jun Zhao, Michael Panzer, and Daniel Garijo. Report from the W3C Provenance Incubator Group, first release: August 6, 2010. Available from http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

Table of contents

Figure List

Table List

Chapter 1

Introduction

The term "Provenance" has been traditionally used in the context of art, in order to determine the authenticity of a certain work or piece: which galleries have had ownership over the piece, restoration processes applied to it, main features of the piece, etc. In recent years, this term has also been applied to other areas like business (tracking of the financial and legal processes, ordering and shipping products, etc.), scientific research (processes and descriptions related to the outcome of an experiment to allow its reproduction) or the web (creation and publication processes of web resources, description, linking and reuse of those resources, etc.).

1.1 Motivation

Nowadays provenance has become a necessity in a wide range of social and scientific applications. In 2009, for example, the green revolution took place in Iran. Users from the Twitter platform[5] started to post and reference tweets about the army moving in Teheran against protestors, but the huge number of tweets and the constant saturation made the threads very hard to follow to know the source of the news[6]. It also made unclear which was the source publishing the original thread, because of the amount of people tweeting the news. To solve it, Twitter created the notion of "retweeting"[7], which allowed to clarify the source of the information, but which was obviously designed to track the source within the Twitter platform. Tweets can be referenced from many other applications too (from Facebook[8] to other blogging platforms), so being able to track the contents outside the platform is key to track the source of a piece of information.

Another example that shows the necessity for provenance is the recent information exposition by the Wikileaks web page[9]. The information was indeed leaked to many newspapers, but since it was not treated with any provenance model (associating metadata to each message), every time a journalist wanted to search news produced in a certain date or generated by a certain senator, the journalist would have to review the whole collection.

[5] http://twitter.com/
[6] http://danbri.org/words/2009/06/16/415
[7] http://evhead.com/2009/11/why-retweet-works-way-it-does.html
[8] http://www.facebook.com/
[9] http://213.251.145.96/

If this information had been treated with a provenance model, a simpler query process would allow the journalists to know, for example, which cables where sent by the US senator in Holland to the United States the day that some important event took place.

In the context of blogging and online news publication, the one we focus on in this work, information is produced and consumed by users (also known as prosumers). Prosumers usually want to know where a piece of information comes from, how it was created, who modified it or how it evolved, so as to be able to determine the quality of that information and generate a trust measure meant for it.

1.2 Provenance and the W3C

Given the increased interest in provenance in the Semantic Web and in the Web community areas, on September 2009 the W3C established the Provenance Incubator Group[10]. Its charter[11] was to provide a state-of-the-art understanding and to define a roadmap for development and possible standardization (developing a set of requirements for representing provenance information of Semantic Web Resources, describing a set of use cases for accessing and dealing with provenance information or identifying starting points for provenance representations).

1.2.1 Definition of Provenance

One of the first activities of the group focused in collecting the different provenance definitions from the participants in the context of the Web[12]. Three definitions resulted from that process:

1. "*Provenance refers to the sources of information, including entities and processes, involving in producing or delivering an artifact* ".- Yolanda Gil.
2. "*Provenance is a description of how things came to be, and how they came to be in the state they are in today. Statements about provenance can themselves be considered to have provenance.*".-Jim Myers.
3. "*Conceptually, the provenance of a piece of data is the process that led to that piece of data. Concretely, provenance is represented by*

[10] http://www.w3.org/2005/Incubator/prov/wiki/Main_Page
[11] http://www.w3.org/2005/Incubator/prov/charter
[12] http://www.w3.org/2005/Incubator/prov/wiki/Final_Report_Draft#What_is_provenance

asserted documentation".-Paul Groth, enhancing the original definition presented in [14].

The three definitions are similar, although the second one is more general and introduces the concept of metadata provenance. After several discussions, the group came up with the definitive definition (summarizing the previous ones), which will be the one we use when referring to provenance in this work:

"*Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance*"[13].

1.2.2 Provenance scenarios

For accomplishing its charter, the W3C Provenance Incubator Group proposed three different common scenarios for the use of provenance: the News Aggregator Scenario[14] (focused on a news and blogs context), the Disease Outbreak Scenario[15] (centered in a scientific and workflow domain) and the Business Contract Scenario[16] (for capturing the legal agreements).

The News Aggregator Scenario defines the problem of a news agency that wants to publish information based on the contents published by different online sources (digital newspapers, blogs, etc). The main objective of the agency is to allow their readers to know which sources contributed to every piece of news and how, in order to produce high trustable publications. The main gaps identified for developing an approach for this scenario are the following:

1. The lack of provenance from the content creators.
2. No guidance for content providers about how they should expose the provenance records in their web applications or services.
3. The lack of a well defined way to expose who has asserted provenance records (key to solve conflicts among them).

[13] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/#A_Working_Definition_of_Provenance
[14] http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_News_Aggregator_Scenario
[15] http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_Disease_Outbreak_Scenario
[16] http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_Business_Contract_Scenario

4. The integration of unstructured contents (such as images or documents).
5. Verification of original sources.
6. The absence of a well-defined standard to link provenance across sites (to understand the provenance of different sources we would have to understand the different APIs of each platform).

1.3 Objectives

The main objective of this work is to address some of the gaps analyzed by the W3C Provenance Incubator Group in the News Aggregator Scenario, in particular number one, two, four and six of the list above, through a real case scenario of a major publishing Spanish company: PRISACOM group[17].

We have selected from existing PRISACOM sources a variety of platforms where multiple users can register and create blogs to post their own experiences and recommendations in a context of travelling. These contents are heterogeneous and without a common editing set of principles, referencing other posts, images, external sources and news from part of the major newspaper of the company: *El País*[18], in particular, *El Viajero*[19].

To address this provenance problem, we will follow an iterative approach, including the next steps:

1) State of the art on provenance models, in order to compare their potential and limitations.
2) Define a set of requirements for the news and blogs scenario.
3) Analyze and identify the most suitable model for the scenario, detecting its limitations.
4) Expand and enrich the model to satisfy all the requirements of the scenario (using other resource descriptive models if necessary).
5) Design and develop a service for annotation and retrieval.
6) Evaluate empirically the coverage and performance of the service according to the requirements.
7) Reflect on the whole process to project our findings in relation to the gaps, beyond the case study.

[17] http://www.prisa.com/
[18] http://www.elpais.com/
[19] http://elviajero.elpais.com/

1.4 Structure of the document

This document describes the work done as follows: in chapter 2 we analyze the state of the art in data provenance. First we identify a set of relevant features, allowing characterizing the different popular provenance models and existing tools. Our study includes a summary of the reviewed models and tools in terms of these features. Then we give the rationale and justification for selecting OPM as our domain model. In chapter 3 we define in detail the problem stating the scenario in which the service has been developed (along with the benefits of its modeling), specifying the scope of the work and providing a set of requirements to set the minimum functionality that our service should have. In chapters 4 and 5 we define respectively the design decisions taken and how the service has been implemented, while in chapter 6 we describe the experimentation to test and evaluate the functionality of the service, providing two representative examples and covering the requirements effectively with the real data from the PRISACOM's platforms. Finally, in chapter 7 we analyze the work done, pointing out our contribution and describing our future lines of work.

Chapter 2

State of the art

The importance of data provenance has risen in the last decade, due to the increasing metadata publication in a wide range of applications (ranging from social networking to scientific workflows). Being able to track how a piece of information has achieved its current state is a great help for any user to understand where that information comes from, or how a certain result from an experiment can be rebuilt. However metadata per se is useless; we need to know how it affects the data and how it relates to the different pieces of information to understand their origin.

In the 2006's International Provenance and Annotation Workshop (IPAW)[20], the scientific community decided that it was necessary to understand the different approaches in provenance representation, their common aspects and why they were different, in order to start working towards a standard. They created the "First Provenance Challenge" in June 2006, for this purpose.

To be able to compare the different approaches, they also provided a set of instructions to the assistants to the Challenge [16]:

- The representations of the system to track the different processes had to be documented in detail.
- Each participant had to comment all the results of the approach when asking the system any of the designed query questions.
- Each participant had to define what the scope of the provenance in that area was for the authors; despite the problems they might encounter.

The Challenge also created a common workflow with a set of "common queries" to be answered by each approach (Figure 2.1).

These ten queries [16] were designed to cover all the possible values a user can obtain from the provenance of the workflow, like all the objects involved in the construction of an artifact, its whole history, differences between two different processes, annotations made in a process or who was the creator of the processes occurred before a date.

[20] http://www.ipaw.info/ipaw06/

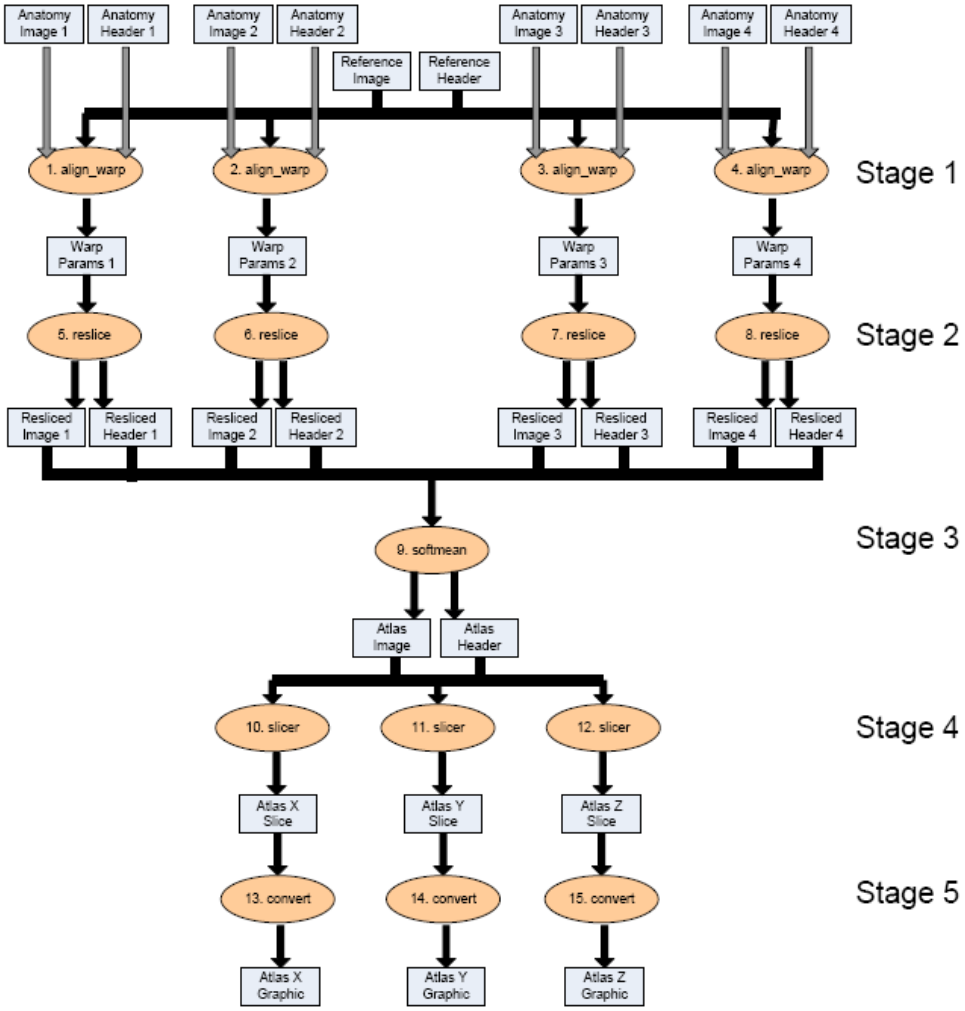

Fig 2.1: Workflow of the First Provenance Challenge

The purpose of this chapter is to describe the current popular provenance models and tools, pointing out the main differences between them. In section 2.1 we present the different features taken into account to distinguish between provenance models and tools, while in section 2.2 we describe briefly every approach, highlighting their pros and cons. Finally, in section 2.3 we present the main model we have chosen for the representation in the proposed scenario, and the rationale behind this decision.

2.1 A comparative framework for provenance approaches

Many of the models are designed with tools or frameworks that implement them. For this reason we identify in subsection 2.1.1 the different features relevant to the models, while in section 2.1.2 we deal with the ones related to the tools.

2.1.1 Features of the models

The First Provenance Challenge can be considered an inflexion point in the development and unification of provenance models. Therefore, the first feature we find relevant is the **date of publication** of the model. It will help us to know if the model participated in the Challenge, whether it was developed a long time ago, or it is recent and based on another popular approach.

The next important feature to take into account is the way the model **represents the information**. Most of the analyzed models use a **directed acyclic graph (DAG)** to this end, where the nodes represent the items or resources and the edges are the causal connections between them, allowing a notion of causality. From now on, when we refer to querying or retrieving content from the provenance graph, we refer to the DAG defined by each approach. However, they use **different formal languages** for this representation, from ontologies and vocabularies in the Resource Description Format (RDF)[21] or the Web Ontology Language (OWL)[22] to XML or SQL. Hence, we distinguish between the approaches that don't make use of ontological resources and approaches that do, because the latter ones are usually more domain generic and easier to expand and adapt.

When analyzing a provenance model, it is also important to know if it uses **another model as reference**. If many approaches are built from an existing provenance model, that provenance model is likely to become a standard. OPM, for example, appeared in 2008 after the Second Provenance Challenge with the idea of setting the basis for a common approach able to answer the queries proposed in previous Challenges. Despite determining a common model to use is a huge advantage, developers should provide an **implementation** too. We consider it

[21] http://www.w3.org/RDF/
[22] http://www.w3.org/TR/owl-features/

valuable in this analysis, because it is likely to find new problems when building it, problems that can lead to abandon the model.

When reviewing the main aspects of the model, the first relevant feature is its **domain dependence**: if the model is very specific, it can't be reused. On the other hand, if it is too abstract and takes too much time to adapt it to a certain domain, developers won't use it as baseline. The approaches also handle the data at different **granularities**: collections, files, tables, folders, etc. They vary depending on their capabilities to store provenance from applications: the provenance for databases usually store items at cell level, while provenance for workflows usually manages files or collections. Other feature related to granularity is how to deal with **confidential information**. For example, if the approach is designed to collect metadata from the patients of a hospital, it has to ensure secure information retrieval, because it is private information. The last feature related to the model captures whether it is **time oriented or event oriented**: a provenance model does not have to include time (if we know where does data come from and how it has been derived, we also know which data was produced first). However, having a temporal reference is always helpful, for example, when we need to compare processes from two different workflows.

The need to preprocess the data or how the approach models the data **updates** and **modifications** have been taken in consideration too. The first one because having to treat separately every piece of information before storing it would make the system more inefficient, and the second one because some approaches give the option to change directly the data, while others record the modification as a new whole process.

Finally, we identify the weaknesses of the approach, such as whether it just models **past executions** or tries to **infer future behaviors**, **it is able to answer all the queries** of the First Provenance Challenge, etc.

2.1.2 Features of the tools

The implementation of each of the approaches is what in the end will determine if the model is usable or not. For example, having to annotate by hand every entry of the provenance graph is a work that not every user is willing or able to do. A system with **automatic provenance collection and management** will end up being more popular (but the trade-off is the cost of developing such a system).

When reviewing how the system interacts with the environment, the first feature to analyze is whether the approach has been **designed to be**

integrated in another system or it **works on its own**. If it is integrated in another system, the developers would need to understand it in order to expand the tool or adapt the model to a new scenario, which requires more workload than if it is independent.

The **query capacity** of the tool is also a very relevant feature, since it limits which type of information we can retrieve from the model. We will distinguish three different levels, further described in [21]. The first level refers to the provenance information retrieval of a given entity, which is the most common type of query. The second level is a bit more demanding, requiring the provenance model to answer queries with some restrictions imposed by the user (e.g. a date, a certain creator, etc.); and the last level is for queries related to the modification and comparison of provenance graphs, such as data collections or provenance subgraphs. Any model which answers all the Challenge's queries satisfies the two first levels and certain aspects of the last one, but not many approaches accomplish to satisfy completely the three levels.

Finally, we evaluate if the approach addresses the problem of **incomplete provenance**: when creating provenance graphs from logs or other applications, we may find incomplete or corrupted data. The provenance systems have to be capable to deal with this kind of errors, either ignoring them (but not failing to process the rest) or trying to infer the missing data.

Table 2.1 summarizes the complete set of features:

Number	**Name**	**Type**
1	Date of publication (DP)	Model
2	Representation (Rep)	Model
3	Formal language (FL)	Model
4	Use of ontological resources (OR)	Model
5	Uses other model as reference (Ref)	Model
6	Existing implementation (EI)	Model/Tool
7	Domain dependency (DD)	Model
8	Granularity (G)	Model
9	Private information management (PI)	Model
10	Driven by time/events (T/E)	Model
11	Weaknesses (W)	Model/Tool
12	Automatic provenance collection (APC)	Tool
13	Independent from other systems (I)	Tool
14	Query capacity (QC)	Tool

15	Incomplete provenance management (IP)	Tool

Table 2.1: Features of the evaluation framework

2.2 Review of the approaches

In this section we characterize briefly each model (and the tool that uses it) based on the previous set of features. We identify three main groups of approaches depending if they are based in 1) ontological resources (to represent the main concepts and relations of the model), 2) layered models (which treat the provenance at different levels depending on the layer) or 3) none of the previous. There are also hybrid approaches (such as WINGS/Pegasus and KOPE), which are layered and use ontological resources, although most of the models are grouped in one of the previous categories.

2.2.1 Ontological resource driven approaches

Among the approaches using ontological resources, we can distinguish two main groups: resource centric approaches and process centric approaches. The former ones report the provenance of a resource, while the latter ones aim to describe the process which led to the specific resource.

2.2.1.1 Resource centric approaches

We include in this group a variety of RDF vocabularies and ontologies for describing resources in different domains. We have decided to include them for the analysis due to their simplicity and popularity in the scientific community, but actually they don't accomplish many of the main features described in 2.1.1. Nevertheless these vocabularies are useful to extend, adapt and complement the process centric vocabularies to different specific domains.

Dublin Core (DC)

Dublin Core Metadata Terms [1] provide a means to describe resources such that others will be able to interpret those descriptions. In particular, it includes a common vocabulary of core terms which can act as metadata keys, qualifications of those terms for specific applications, definitions of data types for the values of resource metadata, and so on. Among the terms available are many which relate to the provenance of the resource: who

created it, when it was changed, when it was created, who holds the rights for the resource, etc. This vocabulary belongs to the Dublin Core Metadata Initiative (DCMI)[23], organization which has been working to provide standards to share and manage the information since 1995. It is very popular in many communities (e.g. scientific community, librarian community, etc.) for describing the attribution of resources (usually documents).

PREMIS

Stands for "PREservation Metadata: Implementation Strategies" [8]. It is a data dictionary for supporting long-term preservation, defining a core set of semantic units that repositories should know in order to perform their preservation functions. It focuses on the provenance of the archived, digital objects (files, bitstreams, aggregations), not on the provenance of the descriptive metadata.

Semantic Web Applications in Neuromedicine (SWAN)

Ontology[24] for modeling scientific discourse, developed in the context of building a series of applications for biomedical researchers, as well as extensive discussions and collaborations with the larger bio-ontologies community. It has been created by the Swan Project initiative, since 2006.

Semantic Web Publishing Vocabulary (SWP)

An RDF-Schema vocabulary[25] developed for expressing different degrees of trust towards information and for representing digital signatures. It is intended to be used in contexts where information is passed through multiple intermediaries, who sign the information to enable trustable exchanges. It uses the Named Graph data model[26].

Semantically-Interlinked Online Communities (SIOC)

Ontology that aims to capture the social provenance produced in the online community information. It has achieved significant adoption through its usage in a variety of commercial and open-source software applications, and is commonly used in conjunction with the FOAF vocabulary for expressing personal profile and social networking information (such as blogs or forums).

[23] http://dublincore.org/
[24]http://swan.mindinformatics.org/spec/1.2/pav.html
[25] http://www4.wiwiss.fu-berlin.de/bizer/WIQA/#WebVocab
[26] http://www.w3.org/2004/03/trix/

Friend of a Friend (FOAF):

FOAF[27] is a simple technology created to track the social provenance about people and their activities (e.g. photos, personal pages, etc.), to link information between Web sites, and to automatically extend, merge and re-use it online. It has become popular within the community for its simplicity, and is commonly used in other vocabularies (e.g. SIOC, DC, etc).

2.2.1.2 Process centric approaches

The set of approaches described in this subsection deal with the evolution of the resources and aim to describe how they have achieved its current state rather than describing the main features of each resource. Most of them have been motivated by the First Provenance Challenge, so they incorporate tools for answering answer the queries.

Open Provenance Model (OPM):

The Open Provenance Model [15] was designed as a result of the First and Second Provenance Challenges (2005-2007). Its granularity is at the level of the data collections and their members, which are represented as an acyclic directed graph, independent of a specific domain. It is also defined in a technology-agnostic manner: the purpose of the model is to help developers to build tools using the model as a reference, but it doesn't impose any restrictions for the implementation.

At first there weren't any vocabularies or ontologies implementing OPM, but recently (during 2010) an "OPM Ontology" (OPMO)[28] and an "OPM Vocabulary" (OPMV)[29] describing the model in OWL and RDF respectively have been developed. Despite the model was designed to deal with workflows, it can be adapted to work at a finer granularity in order to manage the provenance in databases. Its representation in accounts (which are views of the graph at different levels of granularity) allows the provenance to be represented in different layers if needed, ensuring secure retrieval when dealing with confidential data.

Also, many applications have started incorporating the model to their systems (like OurSpaces [20] or Tupelo[30]), or creating plugins to be able to

[27] http://www.foaf-project.org
[28] http://openprovenance.org/model/opmo
[29] http://open-biomed.sourceforge.net/opmv/ns.html
[30] http://tupeloproject.ncsa.uiuc.edu/node/2

handle the provenance graphs represented by the model (as in Taverna[31], Karma [23] or eBioFlow[32]).

But OPM has its limitations too: some of the inferences proposed in the model haven't been fully developed in any tool yet (although it is a work in progress).

Provenance Ontology:

It is an OWL Ontology[33], which uses "Files", "Service Executions" and "Workflow Executions" as main concepts to represent the Challenge's workflow. However, according to the authors, it can be extended or modified to fit in other specific domains.

This ontology was used to develop an approach [6] that successfully answers all the queries proposed in the First Provenance Challenge, storing the provenance with a granularity of a file level, and having a query capacity level three, since it uses SPARQL[34] to retrieve the provenance graphs and subgraphs. It uses a time oriented model, storing the instant of every creation or modification of files (and representing past executions).

Provenance Vocabulary:

RDF Vocabulary [7] oriented to capture the metadata produced in the Web by the content providers (with special focus on the evolution of the Linked Data resources).

It was created on 2009 and designed to work both with digital resources and databases, using a time oriented representation which allows rebuilding the time line of the objects stored in the provenance graph. It is independent from the domain and manages the construction of the provenance graph automatically from the data found in databases, but its implementation uses various other tools to publish this information in the Web (like Pubby[35] or D2R[36]).

The implementation of the model depends on the Tripifly[37] system to convert the data to RDF (which is the format in which it is stored). It represent past executions, with a query level capacity of three (due to the use of SPARQL as query language), and it has some common points with OPM when modeling the information (similar representation of the agents

[31] http://www.mygrid.org.uk/dev/wiki/display/provenance/Taverna+Provenance+Howto
[32] http://sourceforge.net/projects/e-bio-flow/
[33] http://provenance.mindswap.org/provenance.owl
[34] http://www.w3.org/TR/rdf-sparql-query/
[35] http://www4.wiwiss.fu-berlin.de/pubby/
[36] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[37] http://triplify.org/Overview

and processes, with a similar notion of causality in the edges of the provenance graph). Thus, aligning the model with OPM was one of the future lines of work of the authors, accomplished in the recent provenance Vocabulary Mappings[38] .

Provenir Ontology:

Ontology designed in 2009[39] as a means to model the provenance information available in scientific domains (biology, marine sciences, and astronomy), and addressing the geospatial location from the metadata of the experiments. It is presented as a common provenance model along with a framework for provenance management (PrOM) [21].

The model is independent from any domain, and the developers can extend it to address other domains in a specific way. The query capacity of the framework is the third level, enabling to answer the queries appeared in the First Provenance Challenge and being able to retrieve and compare provenance graphs.

PrOM uses the RDF language for representation of the concepts and relations of the model and the SPARQL language for the querying and retrieving the provenance information. It also addresses the retrieval of incomplete provenance using the OPTIONAL SPARQL function, which can be used to specify query expression patterns that can succeed with partial instantiation.

Provenance Aware Service Oriented Architecture (PASOA):

Architecture developed in 2007 [12] with the objective of creating, storing, querying and managing the provenance of electronic document data in workflows.

It makes use of the "Concept Maps" [17], ontological resources in which a general overview of the concepts and their relationships can be specified, ensuring the approach to be independent from a specific domain (we could always expand or modify the current concept map, as in an ontology). It is driven by events, defining interactions in which agents play roles and exchange messages, but it doesn't provide the means to access the provenance graph at different levels (and therefore, it doesn't address the secure access to confidential or private data).

The query capacity offered by the tool which implements the architecture (also named PASOA) is level two. It is not part of other tools or systems, and it doesn't deal with the incomplete provenance.

[38] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings
[39] http://knoesis.wright.edu/library/ontologies/provenir/provenir.owl

Proof Markup Language (PML):

PML [18] [11] is an interlingua for representing and sharing explanations generated by various intelligent systems such as hybrid web-based question answering systems, text analytic components, theorem provers, task processors, web services, rule engines, and machine learning components. The interlingua is split into three different OWL ontologies (provenance, justification, and trust relations) to reduce maintenance and reuse costs. While the provenance of PML in the semantic web is clear in the choice of names (e.g. InferenceEngines), there are numerous examples where PML has been applied to non-text data and non-logic-based processing and thus the term definitions do not appear restrictive.

The main limitation we could find when applying this vocabulary is that it is not easy to understand and to adapt to new scenarios, compared to other approaches like OPM, Provenir or Provenance Vocabulary.

Table 2.2 summarizes the previous set of approaches. The letters in brackets refer to the properties of table 2.1

Feature	**OPM**	**Provenance Ontology**	**Provenance Vocabulary**	**Provenir**	**PASOA**	**PML**
1 (DP)	2008	2006	2009	2009	2007	2004
2 (Rep)	DAG[40]	DAG	DAG	DAG	DAG	DAG
3 (FL)	OWL	OWL	OWL	OWL	XML/ OWL	OWL
4 (OR)	Yes	Yes	Yes	Yes	Yes	Yes
5 (Ref)	No	No	No	No	No	No
6 (EI)	Yes	Yes	Yes	Yes	Yes	Yes
7 (DD)	No	Yes	No	No	No	No
8 (G)	Collections and its members	File	Collections and its members	Collections and its members	-	File
9 (PI)	Yes	No	No	No	No	-
10 (T/E)	Time	Time	Time	Time	Events	Time
11 (W)	Models just past executions	Models just past executions	Models just past executions	Models just past executions	-	Models just past executions
12 (APC)	No	-	Yes	No	No	No
13 (I)	Yes	Yes	Yes	Yes	Yes	Yes
14 (QC)	3	2	3	3	2	3
15 (IP)	No	No	No	Yes	No	No

Table 2.2: Features of the process centric vocabularies

[40] Directed Aciclic Graph

2.2.2 Layered approaches

The set of approaches described in this subsection introduce multiple abstraction layers to separate and manage the provenance at different levels, normally with the objective of separating the representation of the workflow from the storing and query system. Some of them also use ontological resources.

Redux

Model [2] designed in 2006-07 with the purpose of answering the queries of the First Provenance Challenge. It proposes to divide the provenance in four different layers. The first one for the general queries, a second one for specific activities; the third one to represent the executions of the instances specified in the previous layer and the last one for representing the information of the current execution.

This approach doesn't use any ontological resource, and manages automatically the creation of the provenance graph while the workflow is still being executed, being transparent to the user.

The implementation of the model uses relational databases for storing the data, and the SQL language to answer correctly all the queries of the Challenge (achieving a query capacity level two). To represent the workflows, the system relies on the Windows Workflow Foundation engine.

Wings/PEGASUS

Multilayer model [10] developed to participate in the Challenge (2006-07). It is divided in three different layers, which are the different stages of the creation of the workflow. The first one is a template of the workflow (to guarantee the domain independence and generality of the model), the second layer represents the specific data from the instances of the workflow, and the third layer stores all the data relevant to the execution of the workflow. The first two layers are supported by the Wings system, while the third one, by the PEGASUS system.

The approach uses a granularity at a level of collections and their members, and the query capacity of the system is three, since it supports comparisons between different provenance graphs (using SPARQL as query language). Furthermore, this approach introduces the nested collections of elements, which add the capacity to express new subtasks within a workflow and add a major level of detail.

Despite the provenance collection is not automatic from the logs, the system can infer part of the future executions thanks to the templates given

in layer one. To represent the provenance information and the constraints on file collections and their elements, constraints on inputs and outputs of each component and global constraints between multiple components, the Wings system makes use of an OWL ontology. The ontology depends on the data managed by the workflow, being saved in the "*workflow library*" (for each different type of workflow).

Vistrails

Model [22] created in 2006-07 to compete in the Challenge with the aforementioned approaches. Its main feature is that it uses a mechanism based in changes, where each node of the provenance graph represents a version of the workflow, and each edge corresponds to the change which transforms the previous node into the new one. It is organized in three layers. The first layer (workflow evolution layer), captures the different relationships between the series of workflows. The second layer, the workflow layer, consists on individual workflow specifications, and the last layer, the evolution layer, stores information about the executions and modules of the workflows

It deals with a granularity at a file level (a very fine grain), and it is event oriented, where the events are the changes related to each version. Despite being domain independent, it doesn't manage the secure access to the stored provenance.

The tool implementing the model has its own query language for retrieving the provenance, the VisTrails provenance query language (vtPQL, also described in [22]). With this language, it manages to answer all the queries from the challenge successfully. The management and creation of the provenance graph is made automatically from the saved logs of the workflows, implying that Vistrails may represent past executions of the workflows and not their current state (depending on the date of the log). It also relies on the AIR[41] and FSL[42] systems for representing the workflow.

Knowledge-Oriented Provenance Environment (KOPE):

Stand-alone system [5] developed in 2007 for analyzing the provenance in logs. The system builds the provenance automatically, using the Problem Solving Methods libraries (PSM) to separate the process representation in three different views. The first view represents the interaction view, which describes a PSM in terms of its inputs and outputs

[41] http://air.bmap.ucla.edu/AIR5
[42] http://www.fmrib.ox.ac.uk/fsl

(as in a "black box" perspective). The second view is the knowledge flow view, representing how the information is exchanged between subtasks; and the third view (the decomposition view) shows how PSMs decompose in subtasks until they reach the level of primitive actions.

The system is based on the PASOA architecture [12], and it participated in the Second Provenance Challenge to analyze the workflow proposed and check its interoperability with the PASOA system. It does not manage the querying of the resultant directed acyclic provenance graph, since the focus is to provide a reliable analysis of the workflow taken into account. Its only limitation depends on the degree of specification we require when demanding the subtask decomposition of the workflow (at a very low level it may be too difficult to match with the PSM libraries).

Table 2.3 summarizes the layered approaches:

Feature	**Redux**	**WINGS/ Pegasus**	**Vistrails**	**KOPE**
1 (DP)	2006-07	2006-07	2006-07	2007
2 (Rep)	DAG	DAG	DAG	DAG
3 (FL)	SQL	OWL	vtPQL	OWL
4 (OR)	No	Yes	No	Yes
5 (Ref)	No	No	No	PASOA
6 (EI)	Yes	Yes	Yes	Yes
7 (DD)	Yes	Yes	No	No
8 (G)	-	Collections and its members	Files	-
9 (PI)	-	No	No	No
10 (T/E)	Time	Time	Events	None
11 (W)	Requires workflow template	Requires workflow template	Models just past executions	Problems when demanding low level of detail.
12 (APC)	Yes	No	Yes	Yes
13 (I)	Yes	Yes	Yes	Yes
14 (QC)	2	3	3	-
15 (IP)	No	No	No	Yes

Table 2.3: Features of the layered approaches

2.2.3 Other approaches

In this section we define briefly the last non categorized set of approaches.

ZOOM

ZOOM [4] presents a generic provenance system, oriented to scientific workflows in the context of bioinformatics. It was developed in 2006-07, and despite not being based on OPM, it introduces the concept of sub-workflows or "user views", which are very similar to the OPM's accounts. They represent the subprocesses of a workflow, allowing having a better granularity detail if needed. In fact, ZOOM exploits this representation to deal with the confidential data, restricting the capacity to explore at a low level just to the users with the right permission.

It is a model driven by events, and the tool which implements it (also named ZOOM) answers successfully to all the questions of the Challenge. The provenance graph is constructed automatically from the logs of the workflows, so it represents the provenance of past executions. For storing the provenance graph, the tool uses Oracle and Java (JDBC) for the queries. Also, it is compatible with some other workflow representation systems like Kepler [3] or Taverna[43].

KARMA

Tool designed to collect and query the provenance from scientific workflows [23] like the one proposed in the Challenge. It is driven by events and it has a variable granularity, since it allows nesting workflows to describe them at different levels (although it is not considered to deal with the privacy issues when accessing the provenance data).

The provenance collecting and management is done automatically, keeping it up to date even when the workflow is being executed at the moment. As support, it uses the XBaya Workflow Composer[44] system, using BPEL (Bussiness Process Execution Language)[45] as workflow description language.

However, its main limitation is that it does not answer successfully to all the queries formulated in the Challenge, more specifically to query nine (related to the retrieval of annotations belonging to files with a specific type of annotation on it).

Provenance Aware Storage Systems (PASS)

[43] http://www.taverna.org.uk/
[44] http://www.extreme.indiana.edu/xgws/xbaya/index.html
[45] http://www.ibm.com/developerworks/library/specification/ws-bpel/

Systems that treat provenance as a first class object [9], collecting it automatically, storing it, managing it and providing a tool to query it in order to answer all the queries of the Challenge. A first prototype of PASS was designed between 2005 and 2006, but the Challenge forced to develop an ad-hoc query tool, since the creators of the system had been focusing in the storing and automatic management without taking into account what the final user was going to be able to retrieve.

It is oriented to be used in a workflow domain, and its query capacity level is two (it answers successfully to the queries of the Challenge, but it does not support retrieving and comparing graphs). It manages objects at a very low granularity, which is nice to track all the changes that happened in the workflow, but requires a bigger storage system.

Table 2.4 summarizes the last set of approaches:

Feature	**ZOOM**	**Karma**	**PASS**
1 (DP)	2006-07	2006-07	2006-07
2 (Rep)	DAG	DAG	DAG
3 (FL)	SQL	XML/SQL	new query (nq) [9]
4 (OR)	No	No	No
5 (Ref)	No	No	No
6 (EI)	Yes	Yes	Yes
7 (DD)	No	-	No
8 (G)	Collections and its members	Collections and its members	Files
9 (PI)	Yes	No	No
10 (T/E)	Events	Events	Time
11 (W)	Models just past executions	Does not answer all the queries of the FPC	Models just past executions
12 (APC)	Yes	Yes	Yes
13 (I)	Yes	Yes	Yes
14 (QC)	2	1	2
15 (IP)	No	No	No

Table 2.4: Features of the remaining approaches

In this section we have described briefly the approaches for modeling provenance according to our set of features. In the next section we will analyze which one is the right election for our scenario.

2.3 Selecting an approach for a news and blogs scenario.

As we have presented in the previous section, there already exist a high number of tools and models in provenance. However, some approaches fit better than others for the proposed news and blogs scenario. For this reason, when selecting a model among the previous ones we have given more relevance to the next two features:

1. **Process centric models**: We are interested in finding out the causes that led to a certain news or post, not just the description of its current state. Resource centric vocabularies can be used to complement the process centric one, but they can't be the main provenance model for the scenario.
2. **Use ontological resources**. We are not interested in modeling just the contents of the scenario according to a local template or description; we are trying to make the information as interoperable as possible, and, hopefully, publish it as Linked Data. For this purpose we need an ontology.

From the approaches that accomplish both requirements (namely OPM, Provenance Vocabulary, Provenance Ontology, Provenir, PML, PASOA and WINGS/Pegasus), we discarded Provenance Ontology (for not being still supported or available), and PASOA and WINGS/Pegasus (they don't propose a top level ontology for the modeling, using the ontological resources just to define the set of properties of the current domain).

Among the remaining models (OPM, Provenance Vocabulary, Provenir, PML), we have selected OPM as our provenance model for various reasons. Firstly, OPM comes from years of community effort and discussion, being adopted already by many applications (like OurSpaces, Tupelo, Taverna, Karma or eBioFlow). Secondly, OPM has been selected recently by the W3C provenance incubator group as the reference vocabulary to map popular existent provenance vocabularies, setting the first steps for working towards a definitive standard. The provenance graph produced by the model is also easily understandable by any user (unlike other approaches like PML), and as we will show in our work, it is adapted and extended to the scenario without a major effort.

However, in the proposed scenario we encounter heterogeneous resources (like photos, videos, guides or posts), so we need to complement OPM with resource centric vocabularies which capture the specific aspects of each type of resource according to its domain (duration of the video, title of the post, bytes of the image, etc). After searching for ontological

resources associated to each type of resource, we have selected SIOC for capturing the social provenance and relationships in the blogging platform, the MPEG-7[46] Ontology for describing the information about photos and videos, and the W3C GEO ontology[47] to add geolocalization.

In the next section we detail the scenario for which this approach was selected, stating the objectives, scope and requirements for a solution.

[46] http://metadata.net/mpeg7/
[47] http://www.w3.org/2003/01/geo/

Chapter 3

Problem statement: scenario and requirements

Our scenario is related to the general context of travelling, where travelers want to share and read experiences in blogs and online news items. The platform on top of which we build our system aggregates content from a variety of newspapers and digital platforms owned by the PRISACOM group: "*Suplemento El País*"[48], "*Guías Aguilar*"[49], "*Canal Viajar*"[50] and "*Prisa Digital*"[51], but it is also open for recommendations from users (with more than 1000 published), who can also upload their own pictures (more than 2000 uploaded and visible in the web). Users can also create their own blogs to post their travelling experiences (around 600 different blogs exist), apart from commenting the news posted by the reporters of the group. The activity in the web accounts to more than 590.000 unique visitors per month reading or participating on commenting the contents produced, with more than 5000 registered users (to make comments and opinions about the contents it is not necessary to be registered).

Other similar online platforms exist, where the work presented in this thesis could be also applicable. TripAdvisor[52], for example, exploits the same idea of using the content generated by users and their experiences to recommend different places. Users post opinions about hotels, flights or tours and can make reservations. In our scenario users upload general opinions about what they thought about the whole trip rather than a certain hotel or tour (although they can too). TripAdvisor is based mostly on user generated content, while in our scenario we have in addition the news and guides published by journalists from the PRISACOM group. Some of them are recommendations, but some other just highlight cultural aspects of certain villages or countries which may be interesting or have valuable

[48] http://www.elpais.com/suple/eps/
[49] http://www.elpaisaguilar.es/
[50] http://www.plus.es/esp/Canal_Viajar/otras_emisiones/_esp_/index.html
[51] http://www.prisa.com/etiqueta/prisa-digital/
[52] http://www.tripadvisor.es/

touristic aspects for some readers to visit. Finally, in TripAdvisor users can write critics or recommendations following a template (rating some of the features of the service or hotel which the critic is about), commenting and adding a maximum of three photos, whereas in our scenario, the blog entries can talk about anything related to the travelling world and submit as many photos and videos as the user considers necessary to illustrate the post.

Taking into account that (1) "*El Viajero*" is less restrictive than TripAdvisor when posting user generated content, (2) it has data aggregated from a variety of digital content platforms and (3) users frequently refer to contents between the news and blogs platforms, we have considered "*El Viajero*" as a good starting point for developing and testing our approach.

In this chapter we describe first the metadata available in the scenario in section 3.1, the data available from PRISACOM in section 3.2, the core objectives and specification of the scope in section 3.3 and the final set of requirements to be accomplished by the service in section 3.4.

3.1 Metadata available in the scenario

On the online news platform, the digital newspaper publishes guides and recommendations about places to visit. These pieces of news normally have images or videos to illustrate them and comments from other users together with their personal recommendations. In figure 3.1 we can see an example of an article posted on the online newspaper. The first metadata fact is at the top center of the image, which is the location that the document is talking about. Then we have the title, sections, the author, date of publication of the document, recommendations, "I likes" from Facebook and images or videos used as reference. All the referenced content is unlikely to come from other sources outside the PRISACOM's platform, because it would mean to provide advertisements to competitors (except for the links which refer to the place, service or activity described in the guide).

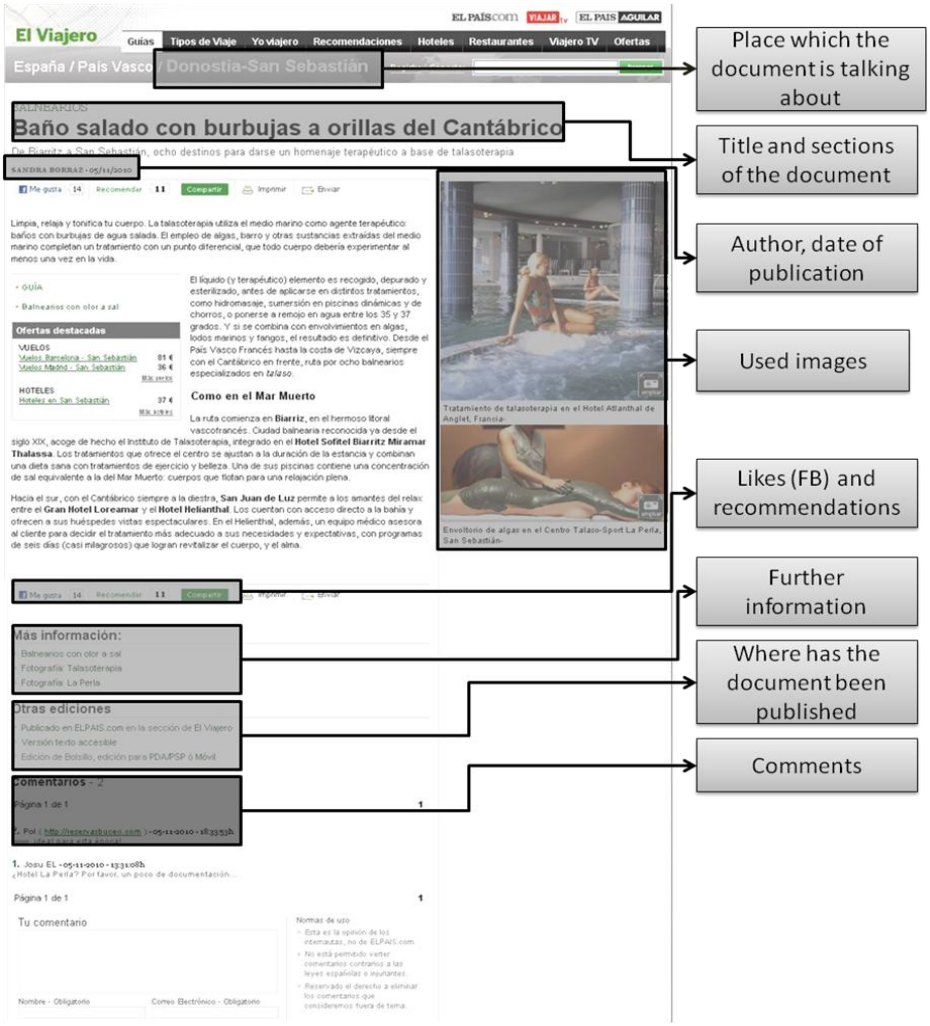

Fig 3.1: Meta information available in the online news platform

Finally, we have at the bottom of the image further information about the document, comments made by users and other places where the document has been published (normally in paper format, but sometimes in pdf. format).

The blogging platform allows users to create blogs to describe their personal trips over the world in posts, using references from other posts, news from the first platform or external resources. Blogs have categories and links updated by the creator, date of the creation and metadata about some of its structure.

Posts, as can be seen in Figure 3.2, have a permanent URL and can be commented by users, who can also reference them in microblogging and

social networking sites like Twitter or Facebook and share them in other popular social networks (MySpace[53], Yahoo[54], etc). The references don't have any metadata associated, but the ones coming from the newspaper are easy to recognize automatically. Finally, the creator of a blog has also a description, and may have friends, fans and followers.

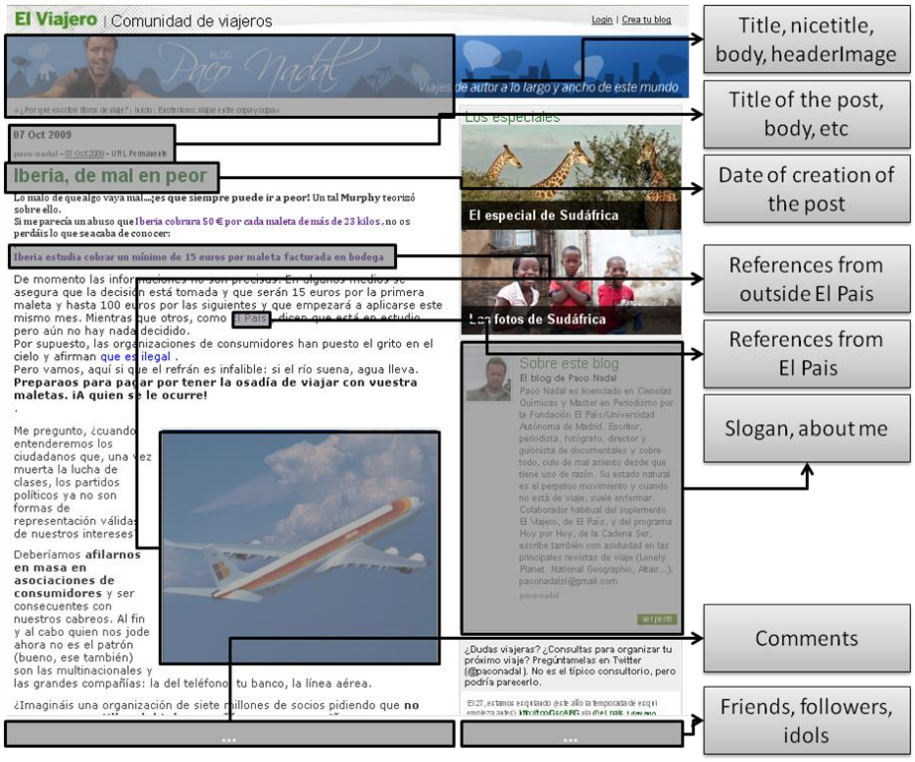

Fig 3.2: Overview of a blog with a post

There are several advantages related to the publication of provenance in the platform. From a company perspective it gives more reliability to the contents created by the platform, since users and third parties can access their history and timeline and check the quality or authority of the references (in particular in the blogging platform, where rumors may be initiated using doubtful sources). Furthermore, as the data is published in a standard format (RDF), there is no necessity to create specialized tools or parsers to understand the contents.

From a user perspective, it makes it easier to formulate complex searches in news and blogs, filtering by date (or period of time), author

[53] http://www.myspace.com/
[54] http://es.yahoo.com/

and place, or asking for contents created by an author and used somewhere else. Thus, keeping track of the latest activities of users in the platform enables their friends and followers to stay tuned on their updates.

From a third party perspective, the publication of provenance makes data more accessible. Instead of having to parse distributed XMLs to ask for certain metadata details, third parties can query the endpoint using SPARQL to retrieve provenance information.

3.2 Accessing the information

The previous sections analyze the data available in the platforms of the scenario. However, part of the information displayed does not belong uniquely to PRISACOM, either because third parties are behind its management and storing or because it is private information about the users.

PRISACOM has granted access and help to facilitate the next data from the platform:

- All the contents in "*El Viajero*" (which include guides, images and videos),
- Localization about some of the guides,
- Posts from some users (like Paco Nadal, one of the most active bloggers in the platform) and
- Comments belonging to those posts.

We have not been able to access (for the moment) the recommendations of the guides, "I likes" from Facebook, comments of the guides and the followers, friends and idols of a user, which will limit to show some of the functionalities that the model allows for the system. As future work we will incorporate this data into the system as soon as it is released.

3.3 Objectives and scope

The way to approach this scenario has been refined into the following objectives:

- Compare the different provenance models in order to determine which is the most suitable for the proposed scenario.
    - Determine a common set of features to compare the approaches.

  - Evaluate the models according to the set of features.
  - Justify the selection of an approach for the proposed scenario.
  - Define the requirements to cover in the scenario, in order to provide a minimum functionality.
- Build the provenance graph automatically from the different users' actions and generated content, reusing and developing annotation tools. We consider the graph a key structure to find out which references were used in a post or guide without having to navigate through every referenced resource.
- Provide the means necessary to retrieve the information modeled in RDF using standard methods (such as SPARQL).
- Expand the model to adapt it to the proposed scenario and evaluate the coverage of the extended model with real examples. The purpose of the work is not to create a new provenance vocabulary, but reuse one of the existing approaches.

The scope of the work done is within the Spanish project Web N+1[55], where one of its use cases involves the PRISACOM's platforms (although it aims to be able to expand to other blogging and news platforms if the appropriate information is available). Our provenance service provides the means to annotate, insert and retrieve contents, enabling many other applications, algorithms and a range of services to manage and deal with the modeled information (e.g. displaying the results in a more eye-catching way, developing advanced specific queries, illustrating with geolocalization where was certain piece of information published, created or modified; or, (with a major complexity) using the provenance of a resource for generating a trust measure for it). However, it is **out of the scope** of this work to develop such services, although some of them propose challenging research areas to explore as future work.

For the first objective, we have dedicated chapter 2 to deal with the first three points. Below, we specify the requirements.

3.4 Requirements

In the News Aggregator Scenario analysis, some key requirements are considered when dealing with blog information. Since our scenario will have some blog platforms in it and the guides we're dealing with in the news platforms are similar to a post, we have considered some of the proposed requirements for our system.

[55] http://www.webenemasuno.es/webnmas1/

- REQ 1: Determine all the contributors to an aggregated blog post or a guide. (News Aggregator (NA))
- REQ 2: Determine how a blog post or a guide has evolved, and which references have been used to create it. (NA)
- REQ 3: Provide the date and time when a blog post or a guide was generated. (NA)
- REQ 4: Allow users to see the entire provenance of an aggregated post or guide (NA)
- REQ 5: Allow users to see the provenance produced by other users and content providers of the platform.
- REQ 6: Reuse standard tools and languages for retrieval and querying information.
    - REQ 6.1: The time used for retrieving the information must be low (less than 3000 milliseconds).
    - REQ 6.2: The content providers must be able to query and insert new data through a REST API.
- REQ 7: Reuse existing provenance models to describe and develop the service in the scenario.
- REQ 8: The service must be able to parse the guides, posts and comments provided by PRISACOM.
- REQ 9: The system must answer all the competency questions specified in the Ontology Requirements Specification Document (ORSD) [19], created following the NeOn methodology [19], and using the data provided by PRISACOM (see appendix A1).

Chapter 4

Design

In this section we describe first how we have modeled the proposed scenario using OPM as the main provenance vocabulary in the news platform, discussing the design decisions taken step by step. Second, we explain the use of SIOC to adapt more meta-information in the blogging platform.

4.1 Modeling the scenario with OPM

OPM proposes a causal graph, where the nodes are either artifacts (immutable pieces of state), processes (action or series of actions performed on artifacts), or agents (controllers of processes); and the edges represent the causal relationships between the nodes: Used (a process used some artifact), WasControlledBy (an agent controlled some process), WasTriggeredBy (a process activated other process), WasGeneratedBy (a process generated an artifact) and WasDerivedFrom (an artifact was derived from another artifact). It also has the notion of accounts (partial subgraphs of the provenance graph), which are useful to represent multiple views of the same graph from different perspectives; and roles, which allow describing deeper some of the aforementioned causal relationships.

Two ontological approaches exist to model OPM: OPM Ontology (OPMO) and OPM Vocabulary (OPMV). The former is more complex and consists on an ontology that models the edges as an n-ary relationship pattern, while the latter is a lightweight ontology to assert the OPM concepts. OPMO is the one used to model the proposed scenario, because it allows adding extra metadata to the edges instead of having to use reification.

In Figure 4.1 we can see an overview of the OPMO representation. The artifacts are represented in the center boxes (Process, Artifact and Agent), while the edges are in the in the boxes WCB (WasControlledBy), WTB (WasTriggeredBy), WGB (WasGeneratedBy), Used and WDF (WasDerivedFrom). The figure also shows how are the boxes interconnected, according to the OPM specification: WasControlledBy has

as cause an Agent and as effect a Process, WasTriggeredBy and WasDerivedFrom have as cause and effect processes and artifacts respectively, Used has as cause an Artifact and as effect a Process and WasGeneratedBy has as cause a Process and as an effect an Artifact. Optionally, some of the edges could have a Role, or occur at an interval of Time (startTime, endTime).

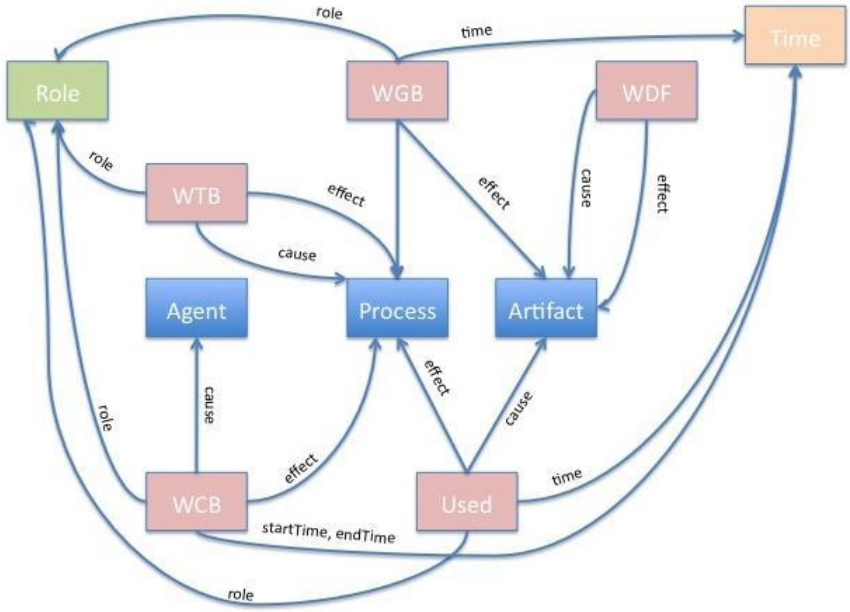

Fig 4.1: Relation of the nodes and edges in OPMO[56]

4.1.1 Versioning of artifacts:

Artifacts evolve as they are modified by different prosumers across their lifetime, and sometimes the modifications applied to them are meant to create replacements of themselves as new versions (e.g. in images, documents, guides, posts, etc).

OPM defines artifacts as immutable pieces of state, but when dealing with the real scenario we find news and posts identified by their own URIs, which may be modified by users in order to produce a new version.

There are two possible options to approach this problem: to save every version with a different identifier (as a snapshot) or to keep the same

[56] Extracted from http://openprovenance.org/model/dependencies.jpg

identifier for all the versions of an artifact (a newer version would automatically replace the older one).

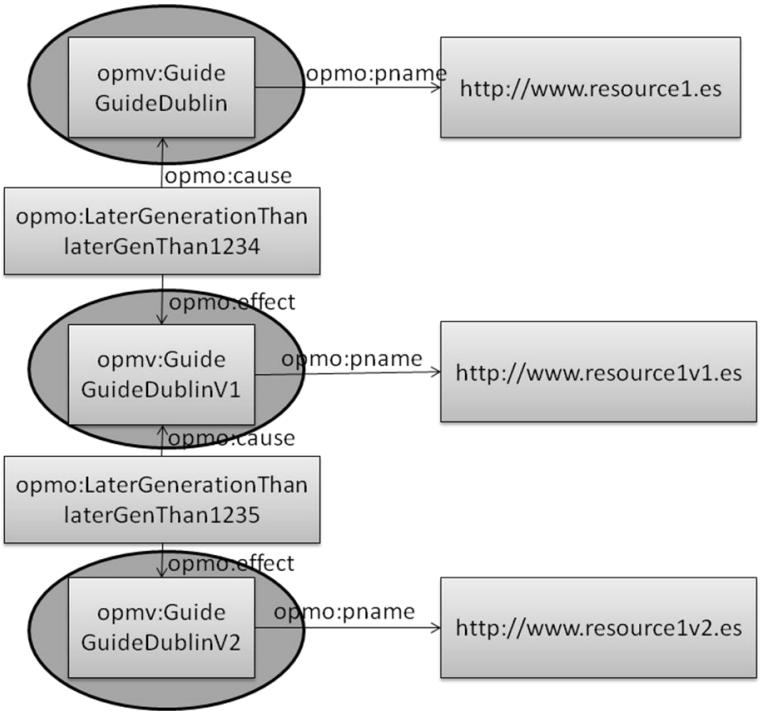

Fig 4.2: Versioning of an artifact saving every version

The former option is represented in Figure 4.2. It represents a guide about Dublin, which has suffered two modifications (GuideDublinV1 and GuideDublinV2), each of which have its own URL. This approach provides the real evolution of the artifact at any moment of its history and allows other prosumers to refer to the different versions separately (grey ovals in the figure). However, it would require more space to save the snapshot of every version, so this option may not be desirable in applications with a high number of versioning artifacts.

The latter option is represented in Figure 4.3. It shows the same example as Figure 4.2, but in this case every version points to the same URL, which is the pointer to the latest version. It requires less space per artifact, since we only store the latest version, but the evolution of the artifact is much more difficult to follow. In this option, the only referenced artifact is the original (represented in the figure in a grey oval), because in the end they all refer to the same version.

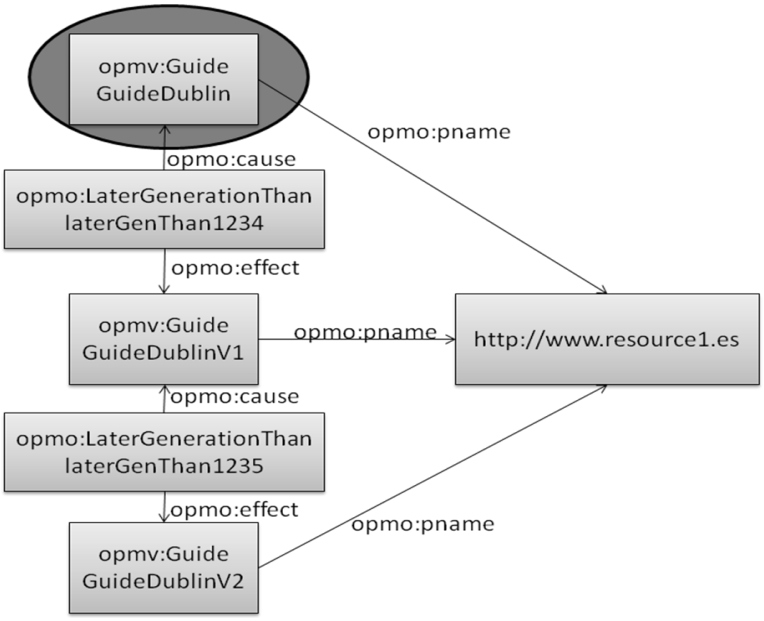

Fig 4.3: Versioning of an artifact keeping just the last version

In both options, to preserve the causality of the provenance graph it would be necessary to add a subclass of the WasDerivedFrom edge called laterVersionThan, as proposed in [13] (since it is an edge more specific than WasDerivedFrom).

In order to deal with our scenario, we have taken the second approach because of the way in which the data is already stored in the platform (leaving the same URL for a post or guide after each modification). Despite the fact that each version is unavailable, it is still useful to have the history of the modifications or updates done to a post, who did them and which references were used. However, if each version is available, it is possible to adapt the first approach too.

4.1.2 Accounts and OPM graphs

The next design decision is how to handle OPM's accounts. In OPM, two accounts can represent the same graph in different detail (or granularity) and overlap. An OPM graph has one or more accounts, but it may have dependencies on certain edges and nodes. The relations between graphs, nodes, edges and accounts can be seen in Figure 4.4.

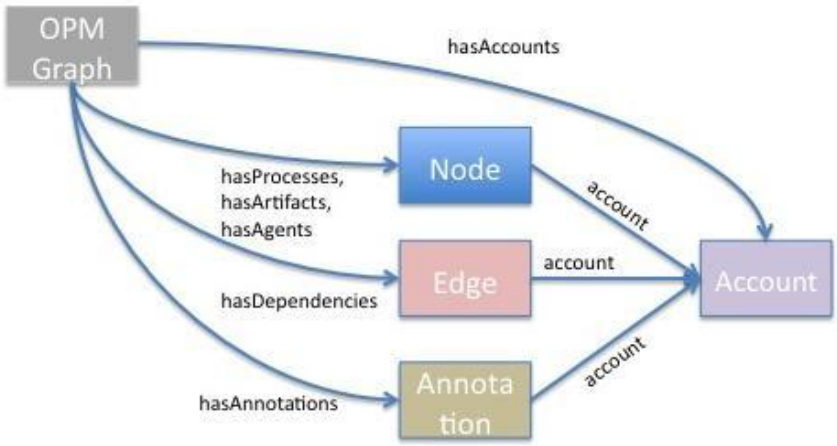

Fig 4.4: Account representation in OPMO[57]

We have chosen to link an account to every action an agent does, leaving the account not linked to a single evolution of an artifact, but to all the evolutions in which the agent has participated or been involved. We have made this decision because it fits better in the context of the proposed scenario, since it allows an easy retrieval of all the actions the user has done across the platforms. If we linked the account to every new generated artifact, we would have a huge amount of accounts monitoring a single process, which is not desirable. The only problem that our approach could have is related to the privacy of the user (other users could track their actions), but we want to generate provenance as open and accessible as possible, and, if needed, we can always restrict a user to query about certain aspects of other users' generated provenance. To obtain the evolution of a single artifact, we rely on the "WasDerivedFrom" edge, which collects the artifacts used to produce the artifact, but has to be asserted explicitly from the content creator (if a process uses an artifact and generates another one, it doesn't imply that the new artifact was derived from the first one).

4.1.3 Overview of the OPM Ontology (OPMO)

Figure 4.5 represents the OPM Ontology extended to fit in our scenario (see appendix A.3 for a further description).

[57] Extracted from http://openprovenance.org/model/accounts.jpg

Fig 4.5: Overview of OPMO

In the figure, both nodes and edges are classes on the ontology. The reason behind this is that the edges are modeled as n-ary relationships, to allow linking not just the two nodes that are supposed to connect, but also allowing specifying the role of the edge, the time associated to the process or artifacts involved. Figure 4.5 also shows how the "Role" concept has been extended to fit in our scenario, adding all the possible roles which the agents and used references can play on it: AditionalInformationRole (a reference is used as additional information in a guide or post), ContentModifierRole (an agent modifies some content), ReferenceRole (an artifact is used as reference in a guide or post), IllustrationRole (an image is used to illustrate some guide or post), SummaryRole (an artifact is used to summarize another), ContentCreatorRole (an agent creates some content) and ContentPublisherRole (an agent publishes some content). The ovals in grey represent the defined classes of the ontology

(classes which have necessary and sufficient conditions or are defined by an equivalentClasses statement).

Figure 4.6 shows how the nodes have been expanded to accomplish covering the necessities of the scenario. On the top right of the figure, Video, Image, Guide, Post and ExternalResource (which refers to the links pointing outside the platform) have been added as subclasses of opmo:Artficat. They are the *"immutable pieces of state"* managed by PRISACOM's platforms.

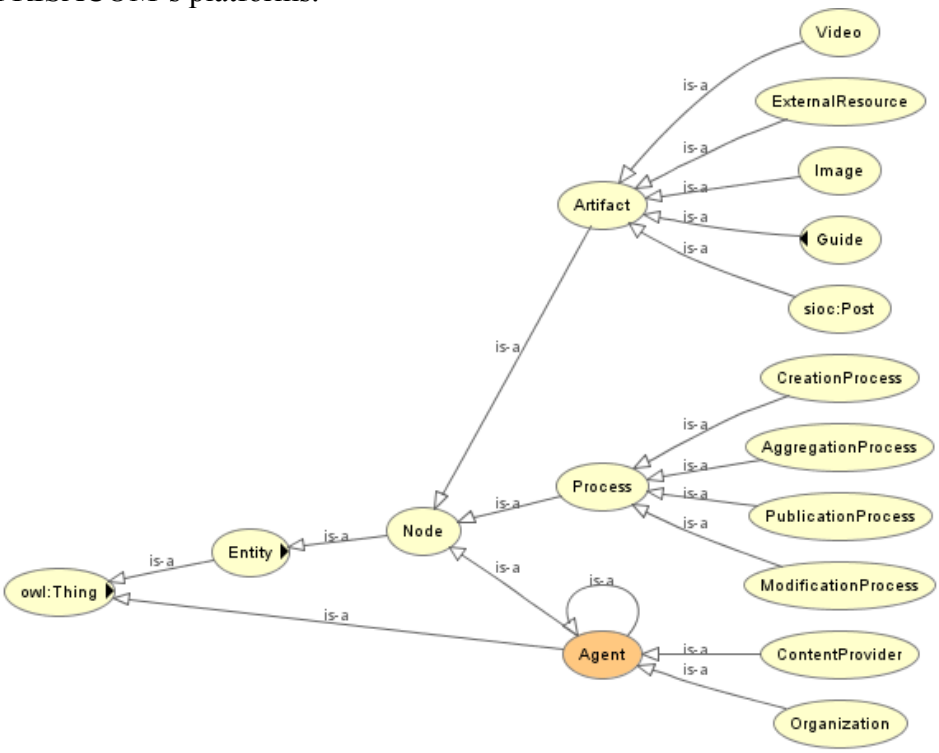

Fig 4.6: Extension of the nodes to fit in the proposed scenario

On the right center of the figure we represent the different processes that content providers can carry out: CreationProcess (for creating new contents), AggregationProcess (when the generated artifact is an aggregation of existing artifacts), PublicationProcess (for publishing existing or new contents) and ModificationProcess (for evolution and versioning of artifacts). Finally, in the bottom right of the figure we can see the two different types of agents in the system: ContentProvider and Organization. Agent has an "is a" edge to itself because of the mapping to foaf:Agent, made by default by the ontology. The edges of the ontology

have not been expanded (except for the LaterGenerationThan), because they are enough to model the actions of the prosumers.

4.1.4 Example of modeling

In figure 4.7 we can see an example of the reference between different news on the digital newspaper. On the right of the figure there is a guide ("*Picoteo dublinés*") which uses an image as illustration (bottom center) and points to another extended guide ("*Niños y otros duendes*") where it has been derived from (top center). This referenced guide was written after interviewing an expert in travels, summarizing part of his experiences as advice to the readers (left image in the figure).

Fig 4.7: Example: News referencing other news and images

Figure 4.8, on the other hand, shows how the whole previous process is modeled with the concepts of OPM (with some extension of the artifacts and roles to fit them in our scenario). On the top of the figure, in blocks 1 and 2, we have represented the steps that led to each of the references: the publication process of the image (block 1), showing how agentJames takes the photo and publishes it with the help of agentPrisa (in a process which could have other inner steps like a quality process) and the creation of the guide from the interviewing process (where the interviewer, the publisher and the expert participated with different roles). In block 3, the creation process of the derived guide is represented. It "uses" the previous two references, and it is "derived from" the top right resultant artifact (which was the general guide, in block 2). Each block belongs to the account of the agents who have controlled the process, so according to the figure block 1 would belong to agentJames, block 2 would belong to agentJulia

and agentJose, block 3 would belong to agentPaco and the union of the three previous blocks would belong to agentPrisa, the publisher of all the contents.

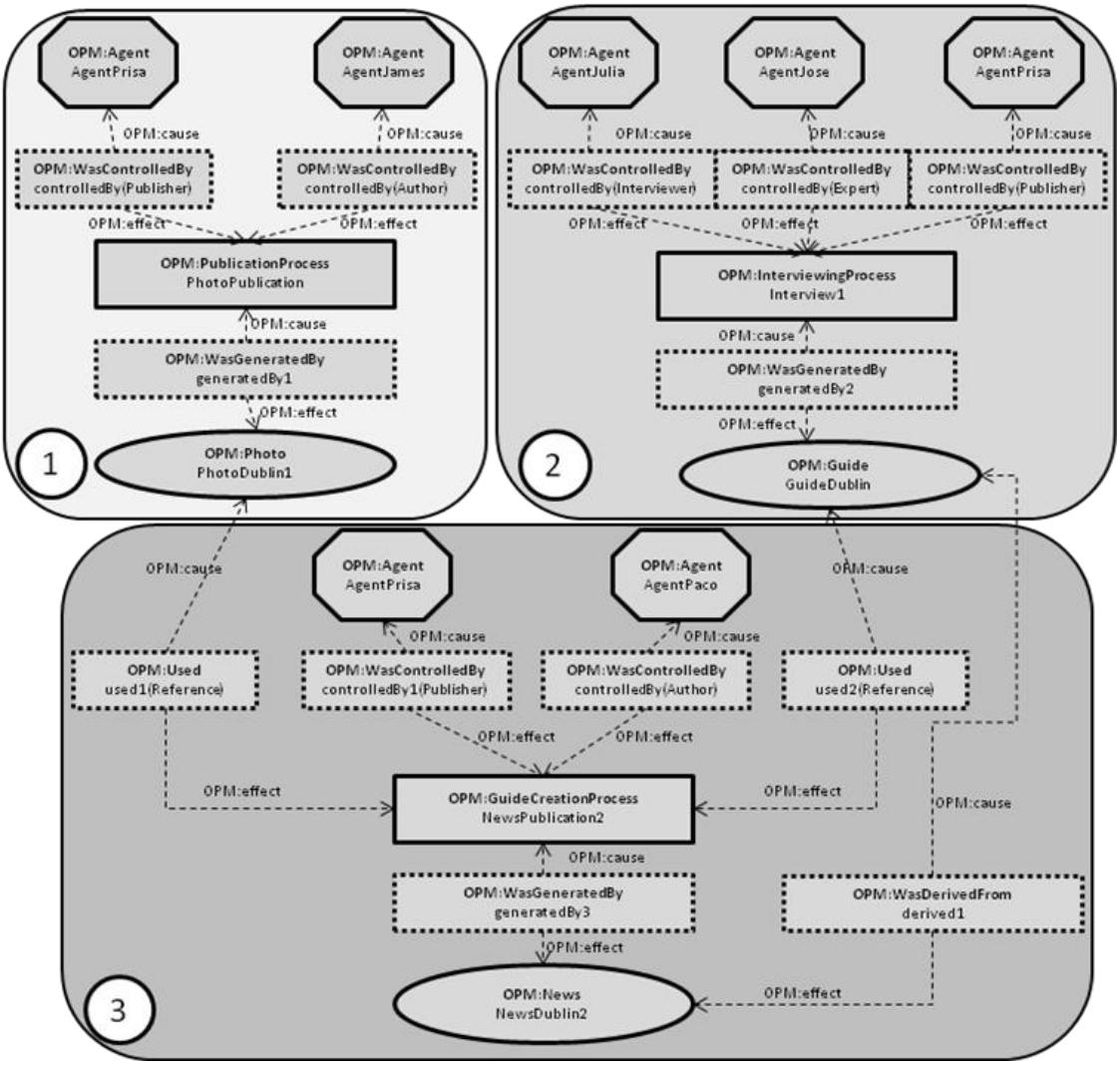

Fig 4.8: Using OPM to model the example of figure 4.7.

To facilitate the reading, in this figure we have removed other metadata associated to the graph, such as dates, times, geolocalization, etc. The roles have been included in brackets when necessary (publisher, creator, reference, modification, etc).

4.2 Ontology network for online news and blogs

OPM is intended to be used as a layered model (figure 4.9), having on the first layer the core model (block 1 in the figure), on an upper layer the required adaptations to each scenario (block 2 in the figure), and on the top layer the different vocabularies which describe the metadata of specific types of artifacts (in this case images, videos, news and posts) (block 3 in the figure).

This architecture allows managing the provenance in the platform at two different levels. The first one would be represented by the resource-centric vocabularies (SIOC, MPEG-7 and GEO), which describe the artifacts in detail at a low level (size, duration, location, title, etc.), while their evolution (which modifications were made, which other artifacts were used, etc) would be captured by the process-centric vocabulary OPMO.

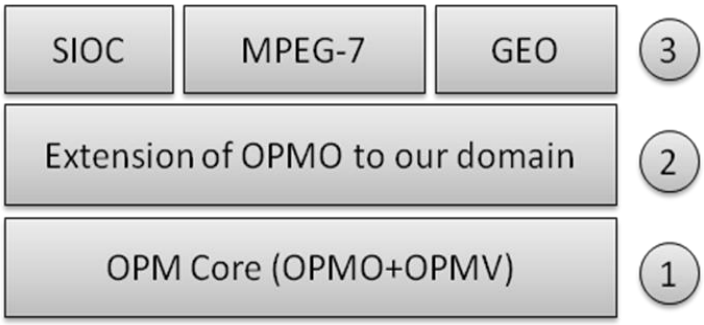

Fig 4.9: OPM as a layered model

The objective in this work has been to reuse existent provenance vocabularies rather than creating new ones, and for such a task we studied the scenario in depth, looking at what could and could not be done in the existing platforms, which data from the PRISACOM group was available and how users could interact which each other. To formalize the study, we created an Ontology Requirements Specification Document with a set of 23 competency questions to be answered in the scenario and a glossary of terms that were relevant (see appendix A.1); and we searched for specific ontological resources to reuse, such as SIOC, MPEG-7 and GEO ontologies, further discussed below.

4.2.1 SIOC Ontology

SIOC is an ontology designed to describe information from online communities (such as blogs or forums), and it is used with OPM in the

blogging platform. SIOC fits perfectly for this task, since it was designed for this purpose. It has Containers, Items and Posts to model the posting activity in the blogs, tracks the followers and subscribers of a user and can even deal with different versions of a post (pointing to previous versions). It also models the comments of a post, the RSS feeds or belonging to a group and links some of these relationships to other popular vocabularies such as FOAF (Friend of a friend) or DC. It is a resource-centric vocabulary centered in the domain of online communities. An overview of the ontology can be seen in figure 4.10:

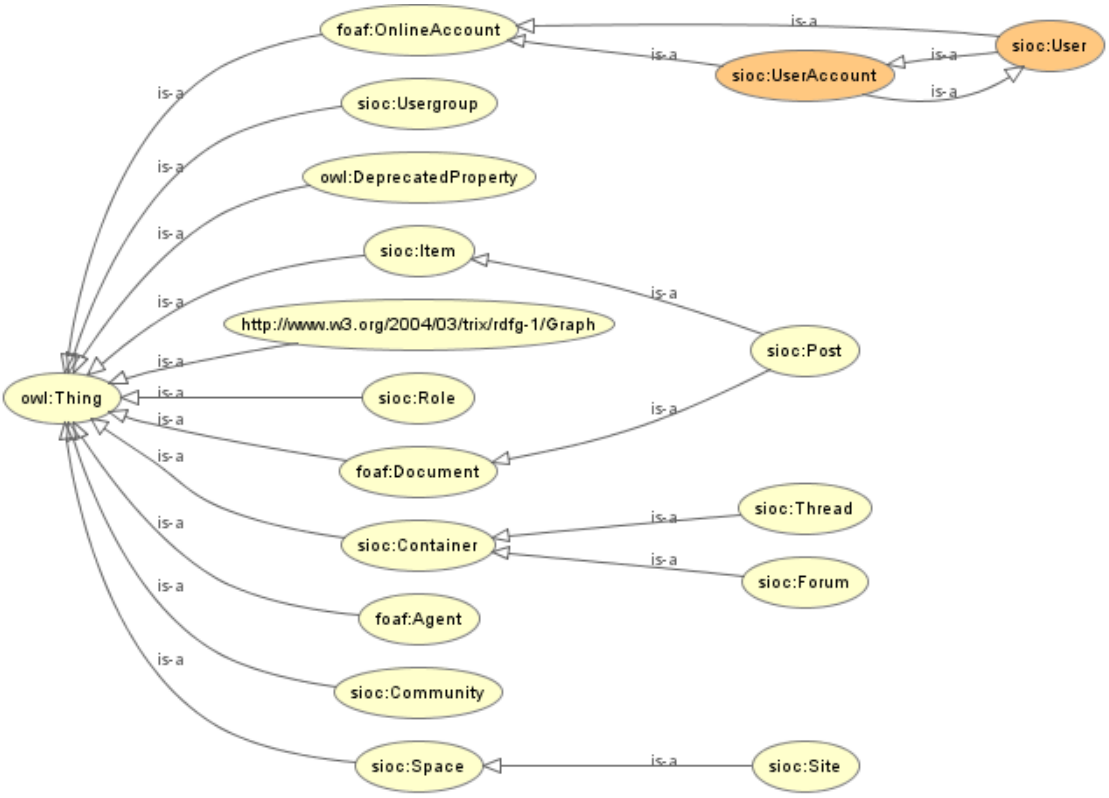

Fig 4.10: Overview of the SIOC Vocabulary

However, SIOC is not rich enough to capture the evolution of the information (its process) across different platforms. SIOC has the property "links_to", which are the references used in a post, but it skips key metadata such as the role of those references, the date they were used, or what type of resources are they (images, news, other posts, etc), so it is very hard to know more details about them without having to investigate them first. Also, SIOC does not deal with some of the versioning features

that we have currently in our scenario: if the only available link is the latest version, sioc:previous_version would link to itself and we would only be available to see the metadata of the latest version.

An idea to address these problems is to link opm:Artifact to sioc:Post as a superclass (artifacts are a broader concept than posts) using its edges and properties to model how the post was obtained (references and versions). Thus, OPM would be used to query the evolution of the piece of information we are seeing in a post, and if more information about the post is needed, we can query the model using SIOC's properties to know which blog it belongs to, which replies does the post have, etc. The owner of the online account which stands for the activities of a user in SIOC (foaf:agent) is mapped to opm:Agent to assure consistency in the graph.

Finally, since guides are not exactly posts, they have been linked as a subclass of sioc:Item (as in figure 4.11). This mapping allows tracking the comments posted by the users and still benefit from all the properties of the OPM's Artifacts.

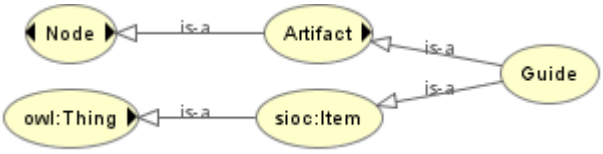

Fig 4.11: Guide as a subclass of opmo:Artifact and sioc:Item

The mappings have been done using the owl:equivalentClass and rdfs:subClassOf properties. Another possible approach would have been to use the SKOS vocabulary as in the W3C Provenance Vocabulary Mappings, but as it is said in the conclusion ("*While we decided to use the SKOS vocabulary to describe the mappings it would have been possible to even use more precise relationships (e.g. rdfs:subClassOf, rdfs:subPropertyOf, owl:equivalentClass) ...*"), our approach is valid too.

4.2.2 MPEG-7 Ontology

The MPEG-7 ontology is a transformation of the MPEG-7 standard into OWL-Full, allowing descriptions on every detail of an image, video or audio file: size, duration, color, decomposition in segments, etc. It has been used in our scenario to annotate the metadata of part of the contents provided by the users (the ones which refer to images or video).

In the next figure (figure 4.12), we can see the portion of the ontology that contains the main concepts mapped as subclass of opmo:Artifact: mpeg7:Video and mpeg7:Image.

Fig 4.12: Overview of the resused part of MPEG-7 Ontology

4.2.3 W3C GEO Ontology

This simple ontology[58] is used for describing the location of spatial things in coordinates (latitude, longitude and height) and places (Madrid, Barcelona, Ireland, etc). According to the specification, a spatial thing is "*Anything with spatial extent*", so we have included in that definition the edges of the OPM graph. Figure 4.13 provides an example: since every edge represents an action (dotted rectangles in the figure), to see that a process generated an artifact at some coordinates would mean that the artifact was created at that location (in the figure, the process was controlled in Madrid, and the resultant artifact was generated in Madrid).

[58] http://www.w3.org/2003/01/geo/

Artifacts could have the location in their description too, but with a different meaning, as it would refer to the content of the artifact rather than the coordinates where it was generated (in the figure, the GuideDublin artifact refers to Ireland). Adding this information is useful in our scenario, because we can track actions like where has a user posted in the last month (if the user is on a trip and we want to know the route), where has he taken all the uploaded pictures, etc.

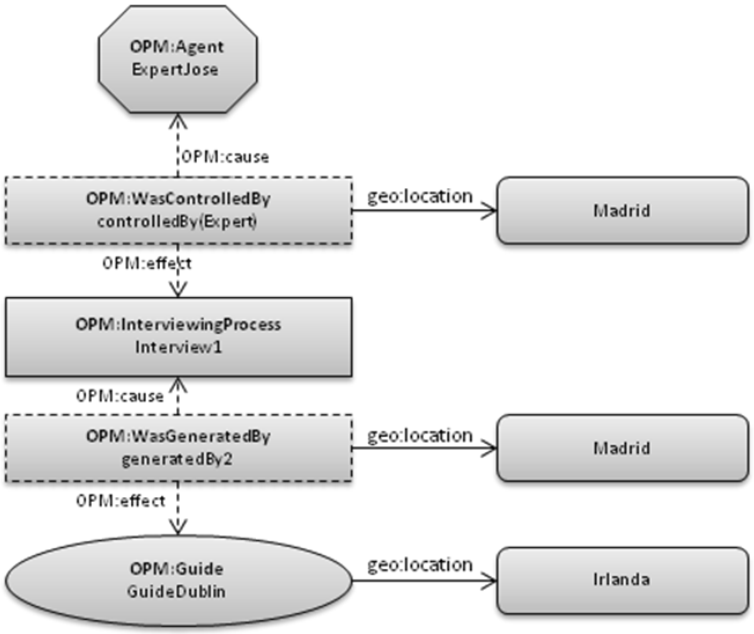

Fig 4.13: Location in the edges and descriptions of the artifacts

In summary, we use the process centric OPM as reference ontology to track the evolution of the artifacts and to ensure interoperability, while SIOC, MPEG-7 and GEO are resource centric vocabularies selected to describe the concrete artifacts in detail.

Chapter 5

Implementation of the service

In order to validate the design proposed in chapter 4, we have created a first version of the service. In this chapter we provide a small overview of the service in section 5.1, we then explain how have we have incorporated the data from PRISACOM to our system in section 5.2 and we describe the annotation and query interfaces in section 5.3 and section 5.4 respectively.

5.1 Overview of the service

The aim of the created service is to provide to users and third parties the means necessary to query and insert the contents from the provenance graph. The metadata modeled by our service can be originated from three different kinds of sources. The first one is PRISACOM, who has provided all the contents specified in section 3.2 (and the core contents of the platform), while the other two are the content providers and applications.

Figure 5.1 provides insights into the overview of the service. Both users and applications (top right in the figure) send their HTTP POST and GET requests to the server, where they are handled by the REST API. Depending on the chosen operation, the query follows different paths:

- If the request intends to add metatadata into the repository, the REST API redirects it to the parsing module (top left in the figure), where the XML parser extracts the concepts according to the templates (see appendix A.4 for a template example). Once completed, the parser provides the contents to the annotation interface, which creates the provenance graph and stores it in the repository (bottom left in the figure).
- If the request is an SPARQL query, it is redirected towards the Jena repository, which answers in a matter of milliseconds. The RDF response is then redirected to the user, content provider or application.
- Finally, if the request intends to make use of the query interface (bottom right in the figure), the REST API redirects the query to it depending on the operation. In this case, the query interface will

develop the complete SPARQL query and send it to the repository as happened in the previous step.

To parse the data from the sources provided by PRISACOM, the server and the REST API are not necessary. The parser processes the data directly from the hard drive and uses the annotation interface to insert the modeled data into the repository.

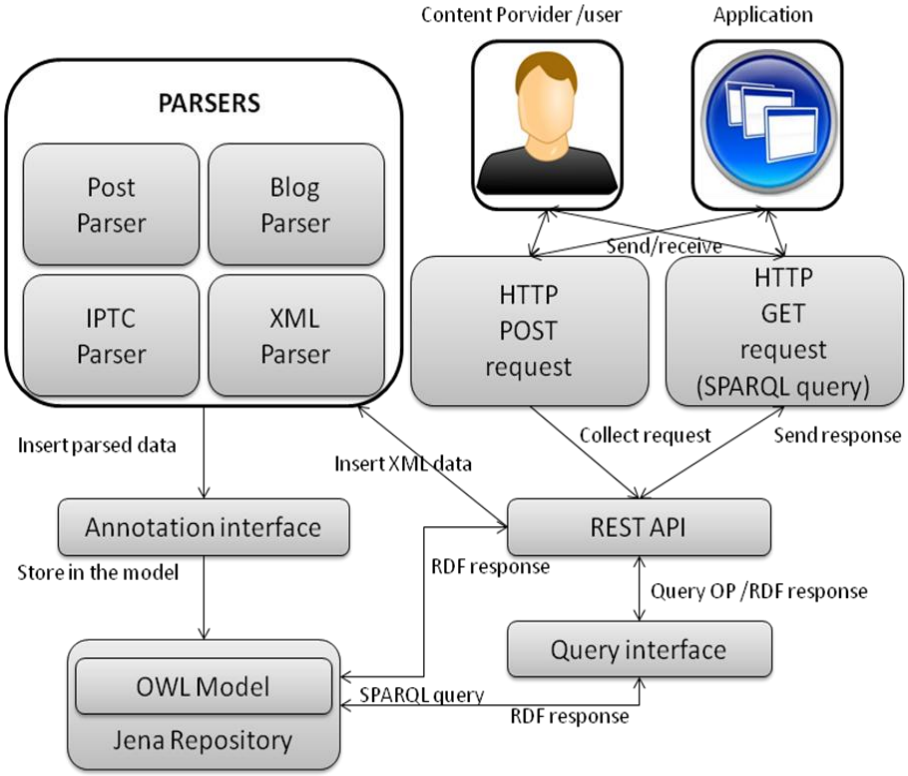

Fig 5.1: Overview of the service

The service has been developed using the JENA Framework[59], combined with TDB[60] (repository with a low time response per query, (a few milliseconds) and able to manage collections up to 1.7 billion triples[61]) and accessible through REST interfaces. The query language for information retrieval is SPARQL[62] (standard language to manage queries in RDF).

[59] http://jena.sourceforge.net/
[60] http://openjena.org/TDB/
[61] http://seaborne.blogspot.com/2008/06/tdb-loading-uniprot.html
[62] http://www.w3.org/TR/rdf-sparql-query/

5.2 Parsing PRISACOM's data

In the proposed scenario, we have to adapt the existing data from the news items and from the blogs, which are stored in the company databases in different types of files. The news are stored using an extension over the International Press Communications Council (IPTC)[63] Standard called News Industry Text Format (NITF), which is also used by other major publishing agencies in Europe, such as AFP[64], ANSA[65] or AP Digital[66]. The data is separated from the metadata in different kinds of files: one for the structure of the news, another for the content of each piece of news, another one for the metadata of the sources of the news like images or videos and another one for storing statistics about the news (such as views, date of the last view, etc). The blogs, posts and comments, on the other part, are stored in three different private databases: one with metadata about the posts (creator, number of comments, body, date of creation, etc), other one with metadata about the blogs (title, header image, description, owner, etc) and other one with the comments of each post (creator, date of creation, etc.). To adapt the data stored to our model and to publish it, we have parsed it and stored it in the repository using our framework.

5.2.1 IPTC Data

For adding all the guides' metadata to our repository, PRISACOM has provided a parser with a set of functions to extract the available parameters from each IPTC file. We have successfully incorporated the parser to our system and reused the next set of operations:

- `getAutorNombre()`: Name of the author of the guide, image or video.
- `getAutorFuncion()`*:* Role of the author in the creation process.
- `getGeoLoc()`*:* Geolocalization of the creation of the guide, image or video.
- `getFechaPublicacion()`*:* Date of publication of the IPTC file.
- `getArea()`*:* Area of creation of the IPTC file.
- `getFechaCreacion()`*:* Date of creation of the IPTC file.
- `getPagina()`*:* Page which the guide occupied in the newspaper.

[63] http://www.iptc.org/site/Home/
[64] http://www.afp.com/
[65] http://www.ansa.it/
[66] http://www.apdigitalnews.com/

- `getLongitud()`*:* Number of words, bytes, etc of the IPTC file.
- `getUnidadMedida()`*:* Unit of measure of the longitude(word, byte, etc.).
- `getAreaEdicion()`*:* Area of publication of the IPTC file.
- `getXref()`*:* Identifier of the IPTC file in PRISACOM's databases. It is not equivalent to the persistent name to the artifact.
- `getRelacionesSistemaEditorial()`*:* Relations with other PRISACOM's guides or images.
- `getIptcMediaType()`*:*MediaType of the IPTC.
- `getIptcMimeType()`*:* MimeType of the IPTC.
- `getTitulo()`*:* Title of the IPTC.
- `getTitularCopyright()`*:* Rights owner of the guide, image or video.
- `getLenguaje()`*:* Language of the guide.
- `getSubtitulo()`*:* Subtitle of the guide, video or image.

However, the parser does not analyze the html content of each file, being unable to find the links to other guides and external resources used as reference. For this task we have processed the html content using the Java String functions, searching the links, posts and referenced guides in the text, as we show in the next fragment of Java code:

```
    String parrafoActual = parrafos[i];
    String   []   parrafoSplit   =   parrafoActual.split("<a
href=\"");
    if(parrafoSplit.length>1){
      for(int j = 0; j<parrafoSplit.length;j++){
          String s =parrafoSplit[j];
          if( s.startsWith("http")){
             s = s.substring(0,s.indexOf("\""));
       referenciasUsadas.add(this.dameUrlNoCanonica
(this.dameId(s)));
          }
      }
    }
```

The "*dameUrlNoCanonica()*" method returns the final persistent name of the artifact as it is explained in section 5.2.4.

5.2.2 Data from posts and coments

In this case, the data provided by PRISACOM was in an Excel format, without any tool to extract the metadata of the posts. To be able to perform

the extraction, we made use of the Apache POI API for Microsoft Documents[67]. The API allowed us to access every column and row of the excel file automatically, which contained, for each post:

- Creation date of the post.
- The blog ID of the blog which belonged to.
- The author ID of the post
- Title of the post
- "Nice title" of the post (used to create its persistent name as it will be explained in 5.2.4.1)
- Body of the post.
- Date of publication of the blog in the web.

For the comments of each post, the metadata available is:

- Date of creation of the comment.
- Creator of the comment (alias).
- Message of the comment (messages have not been further parsed, since the references available in them were very few).

As it happened before, to detect the used references, we have had to parse the html body of the post in a similar way of the IPTCs (in this case, we search for metadata in the images and videos that were unavailable in the body of the news, such as the dimensions of the referenced object).

5.2.3 Localization Data

The localization associated to some of the guides was provided by PRISACOM on a separate file. Note that when we have extracted the "area" in the IPTC parser, it refers to the area where the guide was created, while the location stored in the current file refers to the place which the guide is talking about (a guide about Ireland could have been created in Madrid).

This data follows a simple XML schema:

```
<ROW>
          <nw_xref>20030712elpviavje_8.Tes</nw_xref>
          <upo_id>8185</upo_id>
          <upo_descripcion>México</upo_descripcion>
          <tupo_id>2</tupo_id>
          <tupo_descripcion>País</tupo_descripcion>
</ROW>
```

To parse this metadata, we decided to use the generic Javax parses, included in every Java Development Kit (JDK).

[67] http://poi.apache.org/

5.2.4 URI convention for the Artifacts

Identifying the resources in the platforms has been one of the key questions to ensure the final functionality of the service. If a guide is referenced in the IPTCs with some internal identifier and in a post is referenced by its URL, we have to find the way to know that we are dealing with the same resource, in order to connect it properly in the provenance graph. Same thing happens with the posts.

5.2.4.1 Building the posts' persistent name (pname)

The post's URL is built from the different fields available in the Excel file that PRISACOM provided. For example, the post of the blogger with alias "paco-nadal", who created the post on the 11-08-2010, with the nice-title: para-cenar-lengua-flamenco-y-talon-camello-mi-vida-como, would have the next link:

http://lacomunidad.elpais.com/paco-nadal/2010/11/08/para-cenar-lengua-flamenco-y-talon-camello-mi-vida-como

Posts which had errors in the nicetitle (or nicetitle unavailable) have not been processed, but they are less than 3% of the total number of posts.

5.2.4.2 Building the guides' pname

Unlike posts, the guides, images and videos can have multiple urls pointing to the same artifact (e.g. "http://elviajero.elpais.com/articulo/20101106elpviavje_3/Tes" and "http://elviajero.elpais.com/articulo/viajes/Picoteo/dublines/elpviavia/20101106elpviavje_3/Tes" refer to the same artifact (20101106elpviavje_3.Tes)).

For this reason, after consulting the content manager in PRISACOM, we decided to use the IPTC identification for pointing to the instances of the artifacts, and the non canonical URL for referring to their persistent name. The non canonical URL is built as follows:

Guides:

root/articulo/GuideID. Example:

http://elviajero.elpais.com/articulo/20101217elpepuvia_1/Tes

Images:

root/fotografia/ImageID. Example:

http://elviajero.elpais.com/fotografia/20101217elpepuvia_1/Ies

Videos:

root/video/?video=VideoID. Example:

http://elviajero.elpais.com/video/?video=20100728elpviavje_1.Ves

Thus, for every parsed link that references an image, video or guide, we have had to extract the identifier (which is always the last name in the

URL), and rebuild the non canonical URL according to the aforementioned process.

5.3 Annotation Interfaces

Manually annotating and generating the provenance graph with OPM requires a lot of work from the content provider or the prosumer. Only for a single process creation they would have to create the used artifacts, create the edges, assert the properties of cause and effect, create the roles the agents play in the process, the roles of the used artifacts, the dates of use and localization, etc. If the workload required for a single publication is that high, content providers will end up not using the model and preferring simpler approaches. Besides, if we provide a graphical tool, users will have to understand it and learn how to use it, which is not desirable either.

We want to make the provenance as transparent to prosumers and content providers as possible. For this purpose we have designed and developed a framework through which PRISACOM's platforms and users can annotate and create the provenance graph automatically.

For annotating, the framework proposes some common basic operations focused on the creation or the modification of the contents in both blogs and news platforms (and accessible from the REST API):

- `addBlogMetadata(…)`: adds the metadata about a certain blog to the repository, using mainly the SIOC ontology.
- `addComment(…)`: adds metadata about the comment of a post, using SIOC ontology.
- `addFollower(…)`: adds a single follower to an accountID. It uses the SIOC ontology.
- `addFollowers(…)`:adds a group of followers to an accountID.
- `addFriend(…)`: adds a Friend to an accountID. It uses the SIOC ontology.
- `addFriends(…)`: adds a group of friends to an accountID
- `addLocationToGuides(…)`: adds location to a guide, usind GEO ontology.
- `addNewsMetadata(…)`: adds metadata about a guide into the repository, using OPMO, SIOC and GEO.
- `addPhotoVideoMetadata(…)`: adds metadata about a video or an image, using the MPEG-7 ontology.

- `addPostMetadata(…)`: adds metadata about a post, using the SIOC ontology.
- `addRecommendation(…)`: states that a user has recommended some content, using SIOC ontology.
- `addSuscriber(…)`: adds asuscriber to an accountID.
- `addSuscribers(…)`:adds a group of subscribers to the same accountID.
- `anadirAgentesProcesoResultadoArtefactos(…)`: models, using the OPMO ontology, that a new artifact (or artifacts) have been created or modified, using or not some other artifacts as reference, each of which could have a role, and be used at a certain time. The process can be controlled by cero or more agents, which play a role when controlling it, and that maybe were located at certain coordinates at some time. Finally, the generated artifacts could have been derived from some of the used artifacts.
- `anadirIndividuo(…)`: adds an instance of a class of each of the ontologies included in the repository.
- `anadirPropiedad(…)`: adds a property between two instances of any of the ontologies included in the repository.
- `anadirPropiedadDeDatos(…)`: adds a data property between an instance of any of the ontologies included in the repository and a constant.
- `createContent(…)`: states that a creation process has been prpoduced. It relies on the "anadirAgentesProceso ResultadoArtefactos" operation to insert the contents in the repository.
- `modifyContent(…)`: sates that certain content has been modified. It also uses the anadirAgentesProceso ResultadoArtefactos operation.
- `modifyVersion(…)`: similar to the previous operation, but the generated artifact is a new version of one of the used artifacts.
- `multiStepDerivation(…)`: operation to assert that an artifact has been derived from another using multiple steps.
- `multiStepGeneration(…)`: operation to asser that an artifact has been generated from a process after multiple steps.
- `multiStepUse(…)`: operation to assert that an process has used an artifact after multiple steps.

- `publishContent(…)`: States that a new content has been published. It also uses the "anadirAgentesProceso ResultadoArtefactos" operation.

(We have omitted the parameters of each operation to facilitate the reading).

The framework is supposed to be used as a service called by the platforms involved in the scenario, being highly recommended to make use of the vocabulary we propose in order to guarantee a proper retrieval later.

5.4 Query Interfaces

We have also developed a query framework which includes the basic query operations in the system (like retrieving all the derived artifacts of another artifact) in order to make easier the development of more complicated queries afterwards. The main query operations, available through the REST API, are defined as follows:

- `influencedArtifacts(…)`: artifacts that have been derived or have used the current artifact.
- `guideMetadata(…)`: returns all the metadata from a guide.
- `postMetadata(…)`: returns all the metadata from a post.
- `imageVideoMetadata(…)`: returns all the metadata from an image or video.
- `completeProvenance(…)`: returns the complete provenance of the artifact: which sources have been used to create it and how.
- `modifiedAtLatLong(…)`: returns the artfifacts modificed at certain coordinates.
- `modifiedAtLocation(…)`: returns the artifacts modified at a certain Location.
- `agentsInvolvedinArtifact(…)`: returns the agents that have somehow controlled the process which led to the current artifact.
- `query(…)`: queries the repository with an SPARQL query.

Both interfaces (query and insertion) aim to be transparent to the users. Users should be able to query the SPARQL endpoint or post some data via the REST API if they wanted, but it is intended to be used by other applications which act as a bridge between the user and the framework (for example, applications that present the results to the users in a better eye-

catching way than RDF, or that provides more information to the provenance graph based on the actions of the user).

Chapter 6

Experimentation

In this chapter we explain how we have evaluated the proposed service taking into account the requirements defined in section 3.4. Thus, we firstly analyze two complete representative examples in sections 6.1 and 6.2, retrieving the process that led to the creation of the artifacts step by step, and secondly we answer all the competency questions specified in the ORSD (see appendix A.1) in section 6.3, measuring the time for each retrieval. In section 6.4 we analyze new functionalities for the platform available thanks to our service.

All the queries have been executed in a PC with processor Intel Core 2 Duo at 2GHz, with 4 GB of RAM. The OS in which we have run the system is Windows 7 (64 bits).

As for the size of the data modeled, we have processed the guides, images and videos published in "*El Viajero*" since September of 2001 (around 36000 IPTC files approximately (6523 including location)), and all the posts created by Paco Nadal in his blog (527 posts) along with the comments of each post (a total of 9582 comments).

6.1 Example 1: Guide referencing guides

In "*El Viajero*", guides describing a place often refer to other guides, images and videos to illustrate and give more information to the reader (posts are rarely mentioned). In this example, represented by Figure 6.1, we have as the generated artifact a guide which title is "*El desierto perfecto*" (top left in the figure). This guide uses an image about the topic it is talking about (block 1 in the figure), it references two different guides in its main body (blocks 2 and 3 of the figure), and it adds a third a third link as "more information" (represented in block 4). Since most of the referenced artifacts represent guides, they also make use of some references as well: the guides in block 2 and 4 of the figure use an image as reference.

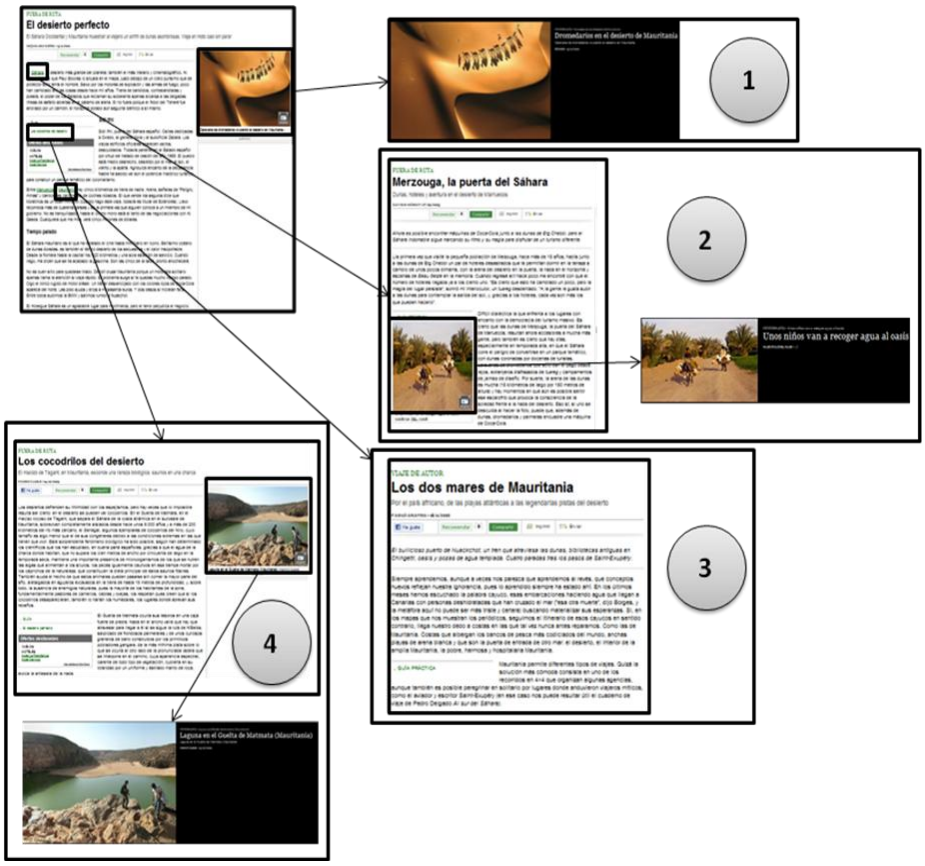

Fig 6.1: Guide referencing other news from the platform

Now we retrieve from the system all the previous process step by step. We will omit presenting all the metadata associated to the artifacts to make the process easy to understand.

First we query the system for the identifier of a guide which title is "*El desierto perfecto*" (namespaces have been removed to clarify the answers):

Query	SELECT distinct ?guide WHERE { ?guide sioc:title "El desierto perfecto"}
Answer	<20101113elpviavje_4.Tes>
Time	641 ms

Table 6.1: Retrieval of the identifier

Now we select the references used for the generation of the artifact.

Query	SELECT distinct ?result WHERE { ?gen opmo:effect opmv:20101113elpviavje_4.Tes. ?gen opmo:cause ?proc.

	?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result. ?used a opmo:Used.}
Answer	<20030927elpviavje_10.Tes> <20061118elpviavje_7.Tes> <20101113elpviavje_8.Ies> <20091024elpviavje_9.Tes> <elviajero.elpais.com/guia/marruecos>
Time	695 ms

Table 6.2: References used for the artifact

As we can see, we have obtained 5 results, four of them are the artifacts shown in figure 6.1 (blocks 1, 2, 3 and 4), and the last one is a link pointing to a page with all the guides about Marruecos (and without an IPTC file associated), which has not been represented in the figure).

Now we query for the used references of each of the obtained artifacts, starting with the image (block 1):

Query	SELECT distinct ?result WHERE { ?gen opmo:effect mpeg7:20101113elpviavje_8.Ies. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result. ?used a opmo:Used.}
Answer	<20101113elpviavje_4.Tes>
Time	590 ms

Table 6.3: References used for the image

As result, we obtain the reference to the original guide. Since it could not be used for the creation of image or reference, we query the system to know the role of the original guide in the image publication process:

Query	SELECT distinct ?rol WHERE { ?gen opmo:effect mpeg7:20101113elpviavje_8.Ies. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result. ?used a opmo:Used. ?used opmo:role ?rol}
Answer	<AditionalInformationRole>
Time	670 ms

Table 6.4: Role of the guide in the publication process

The role is the AditionalInformationRole, so it was not required for its creation.

Next we proceed with the references of the second identifier, which corresponds to block 2 in the figure (the second guide):

Query	SELECT distinct ?result ?rol WHERE { ?gen opmo:effect opmv:20030927elpviavje_10.Tes. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result. ?used a opmo:Used.}
Answer	<20030927elpviavje_3.Ies>
Time	731 ms

Table 6.5: References of the guide (block 2)

As we can see, it uses an image as reference. We can make a simple query to find out the title of the image:

Query	SELECT distinct ?result ?rol WHERE { mpeg7:20030927elpviavje_3.Ies mpeg7:title ?result}
Answer	<Unos_niños_van_a_recoger_agua_al_oasis>
Time	704 ms

Table 6.6: Title of the image used as reference in block 2

The second referenced guide (block three), doesn't use any references. If we query the system to know its title, we see that it is:

Query	SELECT distinct ?result ?rol WHERE { mpeg7:20030927elpviavje_3.Ies mpeg7:title ?result}
Answer	"Los dos mares de Mauritania"
Time	709 ms

Table 6.7: Title of the guide in block 3

The third referenced guide (block 4) uses one image as reference, so we check the answer from the system:

Query	SELECT distinct ?result WHERE { ?gen opmo:effect opmv:20091024elpviavje_9.Tes. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result.

	?used a opmo:Used.}
Answer	<20091024elpviavje_4.Ies>
Time	707 ms

Table 6.8: References used in the guide on block 4.

The answer was correct once again, so we can conclude that after reviewing the first example, the system has successfully retrieved the whole process according to how it occurs in the platform. Also, the time per query has not been greater than 800 ms, accomplishing requirement 6.1.

Average time per query: **605.8 ms**.

Maximum answer time: **731 ms**.

Minimum answer time: **590 ms**.

6.2 Example 2: Post referencing posts and guides:

The second representative example captures the references of a post. Posts often refer to other posts and guides, and in this example, which can be seen in figure 6.2, both news and other posts are referenced in the process which leads to the resultant post (top left in the figure). The post talks about how the bikes are common in cities like San Francisco, while in Madrid can hardly be seen in the city. It refers to a link which talks about the biking activities in Madrid (block 2 on the figure), and also uses some photos to illustrate how popular is the sport in San Francisco (block 1 in the figure).

The post refered in block 2 refers to a guide of "*El Viajero*", where the topic is treated in further detail (block 4). It also provides images to the post (block 3). Finally, the referenced guide links to another guide (block 6), and provides more information about the image used (block 5). The images represented in blocks 3 and 5 are similar, but they are two different artifacts in the PRISACOM's platforms. Also, the image in block 5 is an IPTC file, having more metadata associated to it than the image in block 3. In the image, references to external resources have been removed to clarify the process.

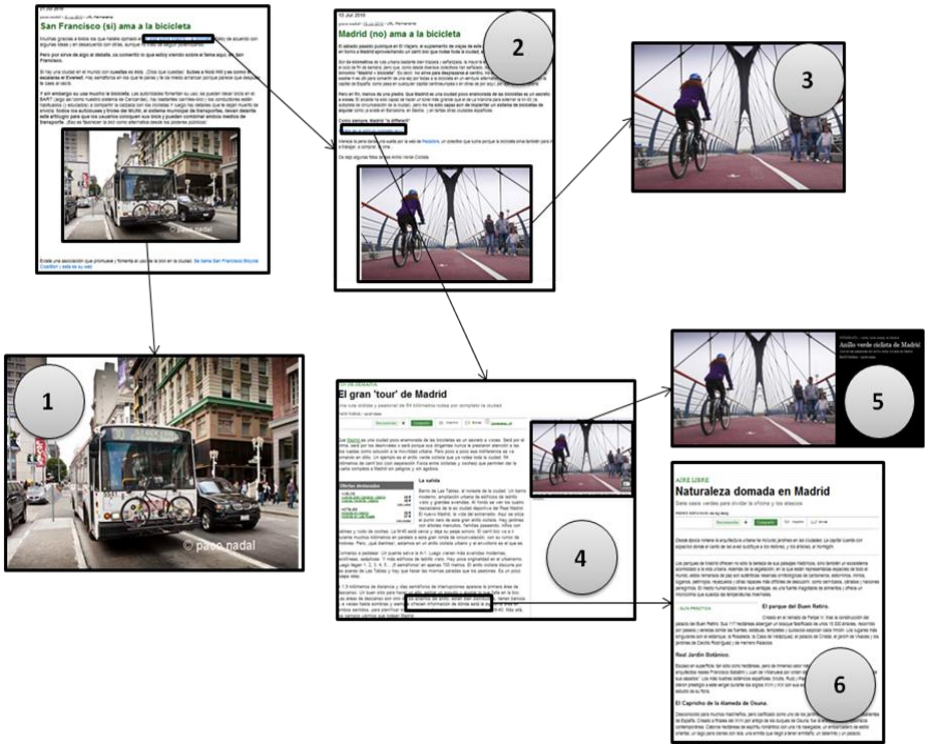

Fig 6.2: Post referencing posts and guides

Now we retrieve from the system all the previous process step by step. First we query the system for the identifier of the post which title is "*San Francisco (si) ama la bicicleta*": (We have removed the namespaces from the results to clarify the answers)

Query	SELECT distinct ?post WHERE { ?post sioc:title "San Francisco (si) ama a la bicicleta".}
Answer	<lacomunidad.elpais.com/paco-nadal/2010/7/21/san-francisco-si-ama-la-bicicleta>
Time	684 ms

Table 6.9: Retrieving the identifier of the post

Next we select the references used for the generation of the artifact, like we did in the previous example:

Query	SELECT distinct ?result WHERE { ?gen opmo:effect <lacomunidad.elpais.com/paco-nadal/2010/7/21/san-francisco-si-ama-la-bicicleta>. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc.

	?used opmo:cause ?result. ?used a opmo:Used.}
Answer	<lacomunidad.elpais.com/paco-nadal/2010/7/15/madrid-no-ama-la-bicicleta> <www.sfbike.org> <www.sfbike.org/main/seven-hells-of-san-francisco-ride> <lacomunidad.elpais.com/blogfiles/paco-nadal/SanFrancisco4.jpg> <lacomunidad.elpais.com/blogfiles/paco-nadal/Dan1.jpg>
Time	699 ms

Table 6.10: References used in the original post

Along with the external links, we can see that the images and the reference to the other post are in the results (named *madrid-no-ama-la-bicicleta*). The images and external links do not refer to IPTC files. Hence, they don't have any metadata associated.

After querying the references obtained for the second post (block 2 in figure 6.2), the results are:

Query	SELECT distinct ?result WHERE { ?gen opmo:effect < lacomunidad.elpais.com/paco-nadal/2010/7/15/madrid-no-ama-la-bicicleta>. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result. ?used a opmo:Used.}
Answer	<20100710elpviavje_4.Tes> <www.pedalibre.org> <lacomunidad.elpais.com/blogfiles/paco-nadal/Anilloverdecilcista-1.jpg> <lacomunidad.elpais.com/blogfiles/paco-nadal/Anilloverdecilcista-2.jpg> <lacomunidad.elpais.com/blogfiles/paco-nadal/Anilloverdecilcista-3.jpg> <lacomunidad.elpais.com/blogfiles/paco-nadal/Anilloverdecilcista-4.jpg> <lacomunidad.elpais.com/blogfiles/paco-nadal/Anilloverdecilcista-5.jpg>
Time	694 ms

Table 6.11: References used in "*Madrid (no) ama la bicicleta*"

In the references we find the guide from *El Viajero*, an external link and five images used to illustrate the post, which don't have any further metadata linked.

When retrieving the resources used for the guide, the results are:

Query	SELECT distinct ?result WHERE { ?gen opmo:effect opmv:20100710elpviavje_4.Tes. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result.

	?used a opmo:Used.}
Answer	<elviajero.elpais.com/guia-ciudad/madrid> <20030322elpviavje_11.Tes> <20100710elpviavje_3.Ies>
Time	673 ms

Table 6.12: References of *20100710elpviavje_4.Tes*

As we can see, it happened just as represented in figure 6.2: block 4 (guide *20100710elpviavje_4.Tes*) makes use of blocks 5 (the image *20100710elpviavje_3.Ies*) and 6 (the *guide 20030322elpviavje_11.Tes*). It also references other page from the platform (*elviajero.elpais.com/guia-ciudad/Madrid*), which doesn't refer to a particular guide; it redirects to all the available guides from the mentioned city.

As for the image referenced, we retrieve its description to check that it has been identified correctly (it doesn't use any further references):

Query	SELECT distinct ?result WHERE { mpeg7:20100710elpviavje_3.Ies mpeg7:abstract ?result}
Answer	<Una_de_las_pasarelas_del_anillo_verde_ciclista_de_Madrid.>
Time	619 ms

Table 6.13: Description of image 20100710elpviavje_3.Ies

Finally, we check the references of the guide in block 6 (*20030322elpviavje_11.Tes*):

Query	SELECT distinct ?result WHERE { ?gen opmo:effect opmv:20030322elpviavje_11.Tes. ?gen opmo:cause ?proc. ?gen a opmo:WasGeneratedBy. ?used opmo:effect ?proc. ?used opmo:cause ?result. ?used a opmo:Used.}
Answer	none
Time	711 ms

Table 6.14: References used for *20030322elpviavje_11.Tes*

To check that the guide is well referenced and it exists, we query the system for its associated metadata:

Query	SELECT distinct ?prop ?result WHERE { opmv:20030322elpviavje_11.Tes ?prop ?result .}
Answer	<Property> | <Value> | <type> | <Guide> | <account> | <AccGrupo_PRISA> | <account> | <AccPACO_NADAL>

	| <account> | <AccPedro_Retamar> | <IPTCMediaType> | "text" | <IPTCMimeType> | "text" | <#language> | "es" | <longit> | "864" | <longitMeasure> | "word" | <page> | "9" | <pname> |"http://elviajero.elpais.com/articulo/20030322 elpviavje_11 /Tes" | | <subtitle> | "Siete oasis verdes para olvidar la oficina y los atascos" | <has_creator> | <Pedro_Retamar> | <created_at> | "20030322" | <title> | "Naturaleza domada en Madrid" | <rightsHolder> | "Diario_El_País_S.L."
Time	614 ms

Table 6.15: Metadata associated to *20030322elpviavje_11.Tes*

Among the results we can distinguish the accounts of the agents involved in the publication of the guide: *Grupo_Prisa* is the publisher, Pedro Retamar is the creator and Paco Nadal used the guide as reference for another guide (block 4 in figure 6.2).

Again, we analyze the performance of the system for the queries executed in the example:

Average time per query: **670.5 ms.**

Maximum answer time: **711 ms.**

Minimum answer time: **614 ms.**

6.3 Answering the competency questions

The previous examples cover the most common functionality of the system, but in order to cover all the requirements, we have also answered all the competency questions in the ORSD document, analyzing the time required for each answer (see appendix A.2 for more details).

It is important to note that some of the competency questions refer to data that is currently unavailable in the system. They have been answered (since the functionality is implemented), but the result is always the empty set.

In summary, we have been able to answer all the competency questions available for the scenario, and retrieve the response under the maximum time specified in the requirements. As we have shown in tables 6.1 to 6.15, the queries are not complex and can be formulated with a minimum knowledge of the ontologies used for modeling the scenario.

6.4 New functionalities for the enhanced platform

Our approach provides new functionalities of the platform which are helpful from an end user point of view, adding new uses that cannot be achieved with the currently used IPTC XMLs (either because it is not well captured or because it is too hard to retrieve the data from them), such as:

1. Adding a description to each post or guide, allowing the user to see (if needed) which sources have contributed to it and how, without having to navigate through them (it would be like a time line of the post or news).
2. Advanced and complex searches of content. Now the search is made by keyword, which is fine when you are looking for a concrete content, but may not be reliable enough when you are trying to combine searches (for example, images of a user which were used in the guides created by another one, the guides created by a user a certain year, guides which have photos of a place, guides close to another place, etc).
3. On the blogging platform, the provenance model allows tracing the latest actions of a certain user: where he has commented, what he has recommended, etc., so it would allow a registered user to know which of his friends, subscribers or followers have been more active lately and inform him about their activities.
4. If the contents are uploaded from a mobile platform, the geolocalization can be used to follow the tour which a certain user is doing in his travel to somewhere (for example, showing his location in a minimap in his blog).

The first two new uses can be achieved by querying the provenance graph with OPM concepts, while for the third it would be required to combine it with SIOC too. The fourth use would require using GEO concepts in combination with OPM's.

Chapter 7

Contributions and future work

In this work we have analyzed the different existing provenance models, selecting a set of features in order to characterize them and compare them in terms of domain independence, interoperability, performance or granularity among others. As a result, we have selected the most appropriate approach for the proposed news and blog scenario: OPM.

We have used OPM as a reference to model the evolution of the contents produced in a variety of platforms belonging to PRISACOM, adapting it and expanding it according to the needs of the scenario. Besides, we have complemented the model with the SIOC ontology for describing the social relationships of the blogging platform, the MPEG-7 ontology for adding metadata about videos and images and the W3C GEO ontology to provide geolocalization in case the provider wants to. We have also developed a query framework with a set of methods to retrieve the provenance metadata easily from the repository and we have defined an annotation framework to build the graph automatically and transparently to the user. Finally, we have tested the functionality of the service, answering successfully the set of competency questions specified in the ORSD (with the available data) and presenting the results. The SPARQL queries required for this task are not complex, and manage to be resolved in no more than one second.

The work described in this document is beneficial for the PRISACOM group, since it helps to spread the contents of their platforms across the web, querying our endpoint to retrieve the modeled data in order to add more functionality to their platforms (reusing geospatial information to know available guides from places close to the one we are looking, advanced search features, allowing to show an explanation of where did the information of a post came from without having to investigate ourselves click by click, etc). Furthermore, it is beneficial to other content providers too, which can query and track some of the sources in order to reuse them. The modeling decisions and parsing tools used for our service are also beneficial for other companies in the news and blogs domain publishing data in the IPTC format (e.g. AFP, ANSA or AP Digital), since their contents could be adapted to the proposed model without much effort.

Our case of study is on its own a rich and complex scenario. Despite sticking to the contents of the platforms owned by PRISACOM, we consider our approach to successfully address some of the gaps analyzed by the W3C Provenance Incubator Group in the news aggregator scenario:

1. Gap one: we model the metadata of the platforms, adding relations among them and accessibility for third parties.
2. Gap two: we capture the actions of the prosumers automatically, and we provide the content providers with XML templates to insert new metadata in the repository.
3. Gap four: we integrate heterogeneous unstructured contents (images, videos, posts and guides).
4. Gap six: we use a popular provenance model (OPM) to make the contents public. These contents can be referenced from other sites and tools using the same reference model.

Our approach differs from systems like OurSpaces [20] (which also uses OPM as main provenance model), where all the artifacts belong to a centralized platform, since we track provenance from a variety of platforms in a context of blogging and online news rather than scientific workflows or research environments, showing that OPM can be applied in other contexts too (still being compatible with other applications using OPM).

As future lines of work, we will first focus on expanding the model for the remaining heterogeneous data from the platforms (friends, followers, idols, recommendations, blog descriptions, posts from other users, etc.), analyzing the coverage of the model answering the remaining competency questions and their complexity, in order to retrieve the results in time. Similarly, we will extend the current interface of the blogging platform to allow users describing better the references used in a post or a guide, and thus we will review the scalability and adaptability of the model with the new activities from the prosumers (which could provide cited text, external resources, detailed descriptions about the images or videos uploaded, descriptions about the process followed to create or publish the guide or post, etc.). It is an interesting line of work that would simplify adapting other scenarios similar to the proposed one, like TripAdvisor.

We will also focus on extracting additional metadata from the posts and comments created by the prosumers. These resources are not well structured following a standard IPTC (as it happens in the contents from "*El Viajero*"): they have to be processed with text extraction tools in order

to determine links, images, videos, context, etc. to enhance the resultant provenance graph.

Finally, we will analyze which resources are linkable to other data collections in the Linked Data cloud, like geoLinkedData[68] to add geolocalization to the news or DBPedia[69] (since many posts use external resources linking to Wikipedia). This will expand the provenance graph, gaining interoperability and letting the user navigate from contents belonging to different platforms.

Our provenance service sets the basis for trust algorithms to query the provenance graph for a resource in order to calculate a trust measure for it (especially in the blogs platforms, where the data is published by prosumers instead from journalists of "*El Viajero*"). Since this trust measure is relevant for the users in the proposed scenario, to calculate it automatically according to the users' preferences is an interesting and challenging line of work.

[68] http://geo.linkeddata.es/web/guest
[69] http://dbpedia.org/About

References

[1] Dublin core metadata initiative. *DLib Magazine*, 6(12), 2009.

[2] Roger S. Barga and Luciano A. Digiampietri. Automatic capture and efficient storage of e-Science experiment provenance. *Concurrency and Computation: Practice and Experience*, 20:419–429, 2008.

[3] Shawn Bowers, Timothy M. McPhillips, Bertram Ludäscher, Shirley Cohen, and Susan B. Davidson. A model for user-oriented data provenance in pipelined scientific workflow. In *IPAW*, 2006.

[4] Sarah Cohen-Boulakia, Olivier Biton, Shirley Cohen, and Susan Davidson. Addressing the provenance challenge using zoom. *Concurrency and Computation Practice and Experience*, 20(5):497–506, 2008.

[5] Jose Manuel Gómez-Pérez and Oscar Corcho. Problem-solving methods for understanding process executions. *Computing in Science and Engineering*, 10(3):47–52, 2008.

[6] Jennifer Golbeck and James Hendler. A semantic web approach to the provenance challenge. *Concurrency and Computation Practice and Experience*, 20(5):431–439, 2008.

[7] Olaf Hartig. Provenance Information in the Web of Data. In *Proceedings of the Linked Data on the Web (LDOW) Workshop at WWW*, Madrid, Spain, 2009.

[8] Sarah Higgins. Premis data dictionary for preservation metadata. *Preservation*, (March):1–3, 2010.

[9] David A. Holland, Margo I. Seltzer, Uri Braun, and K.-K. Muniswamy-Reddy. PASSing the provenance challenge. *Concurrency and Computation: Practice and Experience*, 20:531–540, 2008.

[10] Jihie Kim, Ewa Deelman, Yolanda Gil, Gaurang Mehta, and Varun Ratnakar. Provenance Trails in the Wings/Pegasus Workflow System. *Concurrency and Computation: Practice and Experience*, 20(5), 2008.

[11] Deborah L McGuinness, Li Ding, Paulo Pinheiro Da Silva, and Cynthia Chang. PML 2: A Modular Explanation Interlingua. In *Proceedings of the AAAI'07 Workshop on Explanation-Aware Computing*. Knowledge Systems Laboratory, Stanford University, 2007.

[12] Simon Miles. Electronically Querying for the Provenance of Entities. In *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, pages 184–192. Springer, 2006.

[13] Simon Miles. Mapping attribution metadata to the open provenance model. *Future Generation Computer Systems*, 2010.

[14] Luc Moreau. The foundations for provenance on the web. *Foundations and Trends in Web Science*, 2 (2–3):99–241, 2010.

[15] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, to appear(August 2007):1–30, July 2010.

[16] Luc Mureau, Bertram Ludäscher, Susan Davidson, Ewa Deelman, Luciano Digiampietri, Ian Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole Goble, Ilkay Altintas, Roger S. Barga, Shawn Bowers, George Chin, Shirley Cohen, Sarah Cohen-Boulakia, Ben Clifford, Jennifer Golbeck, Paul Groth, David A Holland, Sheng Jiang, Jihie Kim, Ales Krenek, Timothy McPhillips, Gaurang Mehta, Simon Miles, Dominic Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Karen Schuchardt, Margo Seltzer, Yogesh L. Simmhan, Peter Slaughter, Eric Stephan, Robert Stevens, Daniele Turi, Mike Wilde, Jun Zhao, and Yong Zhao. The first provenance challenge. 20(5):409–418, 2008.

[17] Joseph D. Novak. *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*, volume 69. Lawrence Erlbaum Associates, Inc., 1998.

[18] Paulo Pinheiro Da Silva, Deborah L McGuinness, and Richard Fikes. A Proof Markup Language for Semantic Web Services. *Information Systems*, 31(4-5):381–395, 2006.

[19] Asunción Gómez Pérez, Mari Carmen Suárez-Figueroa, and Boris Villazón. Neon methodology for building ontology networks : Ontology specification. *Methodology*, (February):1–18, 2008.

[20] Richard Reid, Edoardo Pignotti, Peter Edwards, and Adele Laing. ourSpaces:A linking provenance and social data in a virtual research environment. In *International World Wide Web Conference*, pages 1285–1288, 2010.

[21] Satya S. Sahoo, Roger Barga, Amith Sheth, Krishnaprasad Thirunarayan, and Pascal Hitzler. *PrOM : A Semantic Web Framework for Provenance Management in Science*. 2010.

[22] Carlos Scheidegger, David Koop, Emanuele Santos, Vo Huy, Steven Callahan, Juliana Freire, and Cláudio Silva. Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, 20(5):473–483, 2008.

[23] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Query capabilities of the Karma provenance framework. *Concurrency and Computation: Practice and Experience*, 20:441–451, 2008.

Appendices

A.1: ORSD

In this appendix we show the ORSD of the designed ontology. We include in it the competency questions, which have been answered by the service.

Ontology Requirements Specification Document	
	Purpose
	The purpose of this ontology is to model the provenance generated by users and contents in the context of "El Viajero" platform and to use OPM as reference provenance ontology.
	Objectives
	The objectives of this ontology are both to keep track the actions of the users in the platform (so other users can stay tuned to them), and to trace the contents produced by them.
	Language
	The ontology must be implemented in OWL or RDF(S) languages.
	End users
	User 1. A user from the platform that needs to query the metadata associated to a resource (e.g. guide, photo, post, etc) in order to check where it comes from (sources, authenticity, etc). User 2. An automatic system that needs to query the metadata associated to a resource (e.g. guide, photo, post, etc) in order to check where does it come from (sources, authenticity, etc). User 3. An external user (third party), that needs to check the sources of a piece of information which points to the platform as reference.
	Intended uses
	Use 1. Accessing and querying the metadata related to the description of a resource of the platform. Use 2. Accessing and querying the metadata about the evolution of a resource of the platform (e.g. photo, video, post, etc.). Use 3. Inserting new metadata in the platform.

Ontology Requirements

a. Non Functional Requirements

NFR1. The ontology must be written in English.

NFR2.The ontology must use OPM as a reference model. The purpose of this ontology is to reuse existent vocabularies rather than create new ones.

b. Functional Requirements. Set of Competency Questions

Nº	Competency Question	Answer
CQ1	Which references have been used in the post: "Iberia, de mal en peor"?	Reference to post iberia-siempre-haciendo-amigos, 322510_iberiademalenpeor.jpg, guide with id 20091007elpepueco_4/Tes and external link : www.europapress.es/economia/noticia-economia-empresas-ceaccu-insta-iberia-no-cobrar-equipaje-porque-ilegal-20091007144143.html
CQ2	What was the role of each of the references used in the post "Iberia, de mal en peor"?	ReferenceRole
CQ3	What is the provenance of the post "Iberia, de mal en peor" ? (Refering to the metadata associated directly to the resource).	All the metadata of the guide (link, creator, etc.)
CQ4	Where has the post with title "*Madrid (no) ama a la bicicleta*" been used?	Post with id san-francisco-si-ama-la-bicicleta
CQ5	Which are the references used in the guide "Picoteo dublinés"?	20100904elpviavje_1/Tes and 20101106elpviavje_4.Ies
CQ6	Who are the creators of the guide with title "Picoteo dublinés"?	Oscar Gutiérrez
CQ7	Who have been involved in the process of the creation of the guide with title "Picoteo dublinés"? How?(their role)	Grupo Prisa (publisher) and Oscar Gutierrez (creator)
CQ8	Who are the authors of the photographs used in the guide "Picoteo dublinés"?	Andrés Campos
CQ9	How many recommendations does the guide "*Picoteo dublinés*" have?	4
CQ10	Which guides have been written by Oscar Gutierrez in the last month	20091003elpviavje_2.Tes
CQ11	Which guides use photographs of Andrés Campos?	At least 20040313elpviapor_1.Tes (there are a lot)
CQ12	Which guides have a description of "Ireland" on them?	At least 20030614elpviavje_3.Tes (there exist a lot)

CQ13	Where has the guide with title "*Los cocodrilos del desierto*" been used as reference?	20091024elpviavje_4.Ie, 20091024elpviavje_8.Tes and 20101113elpviavje_4.Tes
CQ14	How was the guide "Picoteo Dublinés" modified in the last month?	None
CQ15	Which activities has the user "paco-nadal" carried out this month?	He has created the post cine-viajes-el-ocejon
CQ16	Which comments has the user "paco-nadal" made in any of his posts?	None with his alias
CQ17	Where has the post "Iberia, de mal en peor" been uploaded from?	There is not currently geolocalization metadata
CQ18	Which comments has the post "Iberia, de mal en peor"?	At least "por eso he dejado de volar con Iberia...." (there are more than 15)
CQ19	Which comments has the user "Blas" made in any of the posts of paco nadal?	At least Comment1294421522766 (there are more than 20)
CQ20	Which posts have been created by paco-nadal?	At least the one with id cine-viajes-el-ocejon (there are more than 200)
CQ21	How has the post from user "paco-nadal": "Iberia, de mal en peor" evolved	It has not been modified
CQ22	Who are the followers of user Paco Nadal?	At least NuriaNómada (there are more than 20)
CQ23	Who are the friends of user Paco Nadal?	At least NuriaNómada (there are more than 20)

Pre-Glossary of Terms

a. Terms from the Competency Questions

User, Guide, Post, Idol, Follower, Friend, Coordinates, Reference, Role, Blog, Upload, Comment, Author, Illustration, Creator, Date, Recommendation, Evolution

b. Terms from the answers

User, Post, Guide, Photo, Idol, Coordinates, Comment, Role, ReferenceRole, External Resource, Illustration, Reference.

c. Objects

iberia-siempre-haciendo-amigos, paco-nadal, lacomunidad.elpais.com/blogfiles/paco-nadal/322510_iberiademalenpeor.jpg, 20091007elpepueco_4/Tes, www.europapress.es/economia/noticia-economia-empresas-ceaccu-insta-iberia-no-cobrar-equipaje-porque-ilegal-20091007144143.html, Comment129442211168, 20101106elpviavje_4.Ies, Oscar Gutiérrez, Andrés Campos, 20091003elpviavje_2.Tes, 3, NuriaNómada.

Table A.1: ORSD of the ontology

A.2: Answers to the competency questions

Nº	**CQ1**
Competency Question	Which references have been used in the post: "*Iberia, de mal en peor*"?
SPARQL query	SELECT distinct ?reference WHERE { ?post sioc:title "Iberia, de mal en peor". ?gen opmo:effect ?post. ?gen opmo:cause ?process. ?used opmo:effect ?process. ?used opmo:cause ?reference. ?used a opmo:Used .}
Answer	<lacomunidad.elpais.com/paco-nadal/2009/9/29/iberia-siempre-haciendo-amigos> <es.noticias.yahoo.com/18/20091007/video/ves-iberia-cobrar-por-el-transporte-del-238f6cf.html> | <www.elpais.com/articulo/economia/Iberia/estudia/cobrar/facturar/maleta/bodega/elpepueco/20091007elpepueco_4/Tes> <www.europapress.es/economia/noticia-economia-empresas-ceaccu-insta-iberia-no-cobrar-equipaje-porque-ilegal-20091007144143.html>| <lacomunidad.elpais.com/blogfiles/paco-nadal/322510_iberiademalenpeor.jpg>
Time	668 ms
Nº	CQ2
Competency question	What was the role of each of the references used in the post "Iberia, de mal en peor"?
SPARQL query	SELECT distinct ?typeRole WHERE { ?post sioc:title "Iberia, de mal en peor". ?gen opmo:effect ?post. ?gen opmo:cause ?process. ?used opmo:effect ?process. ?used opmo:cause ?reference. ?used a opmo:Used .?used opmo:role ?role. ?role a ?typeRole}
Answer	<ReferenceRole> (All the references have the same role).
Time	616 ms
Nº	CQ3
Competency question	What is the provenance of the post "Iberia, de mal en peor" ? (Refering to the metadata associated directly to the resource).
SPARQL query	SELECT distinct ?relation ?any WHERE { ?post sioc:title "Iberia, de mal en peor". ?post ?relation ?any.}
Answer	|relation | any | |<type> | <Post> |<account>| <http://openprovenance.org/model/opmo# Accpaco-nadal> |<pname>| "http://lacomunidad.elpais.com/paco-nadal/2009/10/7/iberia-mal-peor" |<has_reply>|<Comment1294422111686>| |<has_reply>|<Comment1294422283875> |<has_reply>| <Comment129442 2285103> |<has_reply>| <Comment12944222 85730> |<has_reply>| <Comment12944222 85886> |<has_reply>| <Comment12944222 86064> |<has_reply>| <Comment12944222 86741> |<has_reply>| <Comment1294422 287048> |<has_reply>| <Comment1294422 287208> |<has_reply>| <Comment1294422 287519>

	|<has_reply>| <Comment1294422 287685> |<has_reply>| <Comment1294422 287831> |<has_reply>| <Comment1294422 287997> |<has_reply>| <Comment1294422 288175> |<has_reply>| <Comment1294422 288559> |<has_reply>| <Comment1294422 288720> |<has_reply>| <Comment1294422 288898> |<has_reply>| <Comment1294422 289136> |<has_reply>| <Comment12944222 89449> |<has_reply>| <Comment12944222 89620> |<has_reply>| <httpComment12944222 89798> |<has_creator>| <Accpaco-nadal> |<created_at> | "1254896940000" |<has_container>| <14644> |<title>| "Iberia, de mal en peor"
Time	644 ms
Nº	CQ4
Competency question	Where has the post with title "*Madrid (no) ama a la bicicleta*" been used?

SPARQL query	SELECT distinct ?result WHERE {
	?post sioc:title "Madrid (no) ama a la bicicleta".
	?used opmo:cause ?post.
	?used opmo:effect ?proc.
	?used a opmo:Used.
	?gen opmo:cause ?proc.
	?gen opmo:effect ?result.
	?gen a opmo:WasGeneratedBy.
	}

Answer	<lacomunidad.elpais.com/paco-nadal/2010/7/21/san-francisco-si-ama-la-bicicleta>
Time	681 ms
Nº	CQ 5
Competency question	Which are the references used in the guide "Picoteo dublinés"?
SPARQL query	SELECT distinct ?reference WHERE { ?guide sioc:title "Picoteo dublinés". ?gen opmo:effect ?guide. ?gen opmo:cause ?process. ?used opmo:effect ?process. ?used opmo:cause ?reference. ?used a opmo:Used .}
Answer	<elviajero.elpais.com/articulo/viajes/Ninos/otros/duendes/elpviavia/20100904elpviavje_1/Tes#despiece1> <20101106elpviavje_4.Ies>
Time	606 ms
Nº	CQ5
Competency question	Who are the creators of the guide with title "Picoteo dublinés"?
SPARQL query	SELECT distinct ?creator WHERE { ?guide sioc:title "Picoteo dublinés". ?guide sioc:has_creator ?creator.}
Answer	<ÓSCAR_GUTIÉRREZ>
Time	591 ms
Nº	CQ 6
Competency question	Who have been involved in the process of the creation of the guide with title "Picoteo dublinés"? How?(their role)

SPARQL query	SELECT distinct ?agent ?typerole WHERE { ?guide sioc:title "Picoteo dublinés". ?gen opmo:effect ?guide. ?gen opmo:cause ?process. ?controlled opmo:effect ?process. ?controlled opmo:cause ?agent. ?controlled opmo:role ?role. ?role a ?typerole. ?controlled a opmo:WasControlledBy .}
Answer	| agent | typerole | |<Grupo_PRISA>|<ContentPublisherRole> | |<ÓSCAR_GUTIÉRREZ>|<ContentCreatorRole>
Time	570 ms
Nº	CQ 7
Competency question	Who are the authors of the photographs used in the guide "Picoteo dublinés"?

SPARQL query	SELECT distinct ?photo ?creator WHERE {
	?guide sioc:title "Picoteo dublinés".
	?gen opmo:effect ?guide.
	?gen opmo:cause ?process.
	?used opmo:effect ?process.
	?used opmo:cause ?photo.
	?used a opmo:Used.
	?photo a mpeg7:Image.
	?photo mpeg7:creator ?creator.
	}

Answer	|<20101106elpviavje_4.Ies>|<ANDRÉS_CAMPOS>
Time	688 ms
Nº	CQ8
Competency question	How many recommendations does the guide "*Picoteo dublinés*" have?

SPARQL query	SELECT distinct ?prop ?result WHERE {
	?guide sioc:title "Picoteo dublinés".
	?guide opmo:recommendedBy ?result .
	}

Answer	None
Time	610 ms
Nº	CQ8
Competency question	Which guides have been written by Oscar Gutierrez in the last month? Note: needs a small additional processing, since the date is a literal, and SPARQL can't handle type castings. The date is in the format of the PRISACOM'S IPTC news.
SPARQL query	SELECT distinct ?guide ?date WHERE { ?guide a opmv:Guide. ?guide sioc:has_creator opmv:ÓSCAR_GUTIÉRREZ. ?guide sioc:created_at ?date.}
Answer	| guide | date | | <20091003elpviavje_2.Tes> | "20091003" | | <20090815elpviavje_9.Tes> | "20090815" | | <20101106elpviavje_3.Tes> | "20101106" |
Time	813 ms
Nº	CQ9
Competency question	Which guides use photographs of Andrés Campos?
SPARQL	SELECT distinct ?image ?guide WHERE {

query	?image a mpeg7:Image. ?image mpeg7:creator opmv:ANDRÉS_CAMPOS. ?used opmo:cause ?image. ?used opmo:effect ?process. ?used a opmo:Used. ?gen opmo:cause ?process. ?gen opmo:effect ?guide. ?guide a opmv:Guide. }
Answer	| image | guide | |<20040313elpviapor_2.Ies>|<20040313elpviapor_1.Tes> | |<20051217elpviavje_2.Ies>|<20051217elpviavje_3.Tes> | |<20051224elpviavje_1.Ies>|<20051224elpviavje_3.Tes> | |<20060902elpviavje_2.Ies>|<20060902elpviavje_4.Tes> | |<http20061014elpviavje_2.Ies>|<20061014elpviavje_4.Tes> | |<20070113elpviavje_5.Ies>|<20070113elpviavje_6.Tes> | |<20070609elpviavje_2.Ies>|<20070609elpviavje_4.Tes> | |<20070630elpviavje_6.Ies>|<20070630elpviavje_9.Tes> | |<20070825elpviavje_4.Ies>|<20070825elpviavje_7.Tes> | |<20070929elpviavje_4.Ies>|<20070929elpviavje_6.Tes> | |<20071006elpviavje_6.Ies>|<20071006elpviavje_10.Tes> | |<20071124elpviavje_3.Ies>|<20071124elpviavje_10.Tes> | |<20071229elpviavje_4.Ies>|<20071229elpviavje_7.Tes> | |<20080126elpviavje_13.Ies>|<20080126elpviavje_19.Tes> | |<20080301elpviavje_4.Ies>|<20080301elpviavje_6.Tes> | |<20080322elpviavje_5.Ies>|<20080322elpviavje_10.Tes> | |<20080412elpviavje_3.Ies>|<20080412elpviavje_4.Tes> | |<20080524elpviavje_5.Ies>|<20080524elpviavje_6.Tes> | |<20080607elpviavje_5.Ies>|<20080607elpviavje_7.Tes> | |<20081008elpviavje_4.Ies>|<20081004elpviavje_10.Tes> | |<20081004elpviavje_4.Ies>|<20081004elpviavje_6.Tes> | |<20081108elpviavje_5.Ies>|<20081108elpviavje_4.Tes> | |<20081129elpviavje_3.Ies>|<20081129elpviavje_3.Tes> | |<20081227elpviavje_3.Ies>|<20081227elpviavje_8.Tes> | |<20090725elpviavje_5.Ies>|<20090725elpviavje_4.Tes> | |<20090815elpviavje_2.Ies>|<20090815elpviavje_3.Tes> | |<20091024elpviavje_6.Ies>|<20091024elpviavje_12.Tes> | |<20100109elpviavje_2.Ies>|<20100109elpviavje_2.Tes> | |<20100814elpviavje_4.Ies>|<20100814elpviavje_3.Tes> | |<20101106elpviavje_4.Ies>|<20101106elpviavje_3.Tes> |
Time	706 ms
Nº	CQ10
Competency question	Which guides have a description of “Ireland” on them?
SPARQL query	SELECT distinct ?guide ?title WHERE { ?guide a opmv:Guide. ?guide geo:location geo:Irlanda. ?guide sioc:title ?title}
Answer	| guide | title | <20030614elpviavje_3.Tes> | "Irlanda" | <20040403elpviavje_5.Tes> | "Desayuno con Leopold Bloom" | <20050122elpviavje_7.Tes> | "Cork se zambulle en la fiesta" | <20050611elpviapor_1.Tes> | "La Irlanda de 'El hombre tranquilo'"
Time	844 ms
Nº	CQ11
Competency question	Where has the guide with title “*Los cocodrilos del desierto*” been used as reference?
SPARQL	SELECT distinct ?result ?p WHERE {

query	?post sioc:title "Los cocodrilos del desierto". ?used opmo:cause ?post. ?used opmo:effect ?proc. ?used a opmo:Used. ?used opmo:role ?role. ?role a ?p. ?gen opmo:cause ?proc. ?gen opmo:effect ?result. ?gen a opmo:WasGeneratedBy. }
Answer	Used in Role <20091024elpviavje_4.Ies> | <AditionalInformationRole> <20091024elpviavje_8.Tes> | <AditionalInformationRole> <20101113elpviavje_4.Tes> | <AditionalInformationRole>
Time	733 ms
Nº	CQ12
Competency question	How was the guide "Picoteo Dublinés" modified in the last month? (Again, extra processing is needed. The query will focus in extracting the derivations of the guide.)

SPARQL query	SELECT distinct ?guide WHERE {
	?guide sioc:title "Picoteo dublinés".
	?a opmo:cause ?guide.
	?a a opmo:WasDerivedFrom.
	}

Answer	None (it was not modified)
Time	571 ms
Nº	CQ13
Competency question	Which activities has the user "paco-nadal" carried out this month? (Processes which paco-nadal has controlled). It needs extra processing to do the filtering, as happened before.
SPARQL query	SELECT distinct ?result WHERE { ?wasControlled opmo:cause opmv:paco-nadal. ?wasControlled opmo:effect ?process. ?gen opmo:cause ?process. ?gen opmo:effect ?result.}
Answer	(Since there is a large set of results, we limit the set supposing that we have done the filtering). |<lacomunidad.elpais.com/paco-nadal/2010/11/12/cine-viajes-el-ocejon> |<lacomunidad.elpais.com/paco-nadal/2010/11/10/saharauis-como-mantener-dignidad-el-infierno> |<lacomunidad.elpais.com/paco-nadal/2010/11/9/el-27-estamos-esquiando> |<lacomunidad.elpais.com/paco-nadal/2010/11/8/para-cenar-lengua-flamenco-y-talon-camello-mi-vida-como> |<lacomunidad.elpais.com/paco-nadal/2010/11/5/diario-cayo-caementicius-mi-vida-como-romano-parte-ii-> |<lacomunidad.elpais.com/paco-nadal/2010/11/4/mi-vida-como-romano-parte-i-> |<lacomunidad.elpais.com/paco-nadal/2010/11/2/-de-verdad-van-desaparecer-guias-viaje-papel->
Time	586 ms
Nº	CQ 14
Competency question	Which comments has the user "paco-nadal" made in any of his posts?
SPARQL query	SELECT distinct ?post ?comment WHERE { ?wasControlled opmo:cause opmv:paco-nadal. ?wasControlled opmo:effect ?process. ?gen opmo:cause ?process. ?gen opmo:effect ?post.?post a sioc:Post. ?post sioc:has_reply ?comment.

	?comment sioc:has_creator sioc:Accpaco-nadal.
Answer	none
Time	908 ms
Nº	CQ 15
Competency question	Where has the post “Iberia, de mal en peor” been uploaded from? (Note: no geolocalization is available yet in the posts).

SPARQL query	SELECT distinct ?lat ?long ?location WHERE {
	?post sioc:title "Iberia, de mal en peor".
	{?post geo:lat ?lat}
	UNION
	{?post geo:long ?long}
	UNION
	{?post geo:location ?location.}
	}

Answer	None
Time	581
Nº	CQ 16
Competency question	Which comments has the post “Iberia, de mal en peor”?

SPARQL query	SELECT distinct ?comment ?content WHERE {
	?post sioc:title "Iberia, de mal en peor".
	?post sioc:has_reply ?comment.
	?comment sioc:content ?content.
	}

Answer	| comment | content | | <Comment1294422111686> | "por eso he dejado de volar con iberia...." | <Comment1294422283875> | "Consumo en Baleares que les a dicho que de cobrar por el equipaje nada y que la ley es igual para todos.\r\n!Stop abusos a los consumidores!\r\nHay que unirse y denunciar" | <Comment1294422285103> | "Lo malo de la plebe, compañero, es que hay que formarla como individuo con criterio. Uy, nada menos ni nada más... criterio. Porque si el poder del Estado da Pánico. No sé yo no sé donde nos íbamos a tener que meter cuando los vahos de los vapores de los poderes fácticos y de la erótica del poder se les subieran a la cabeza a los presidentes de esas asociaciones de consumidores. Porque la idea del escritor de este blog está buena, es como de idealista, cosa que es adorable pero tengo la sensación de que mal remedio tiene todo porque lo corrupto está en el ser humano. Que no es depredador de cualquier modo, lo que es con materia gris. Vaya putada." | <Comment1294422285730> | "Cuando nos daremos cuenta de que hace tiempo que se anulo a la masa social? Quizas cuando Cristiano Ronaldo haga huelga...se movera la plebe!!!" | <Comment1294422285886> | "Cuando nos daremos cuenta de " | <Comment1294422286064> | "Cuando nos daremos cuenta de " | <Comment1294422286741> | "Nada Paco, esto es el principio, ya está en proyecto cobrar por usar el W.C., lo de las almohadas está en vigor en algunas aerolíneas de tramo corto y pronto venderán billetes de a píe como en los campos de fútbol de antes e iremos todo el viaje parados y agarrados a una barra de hierro del techo. Es el progreso ¿no te parece?" | <Comment1294422287048> | "Ostras!! En referencia a lo que ha dicho k-RoL, yo tb me envié las maletas cuando me fui de viaje a los Alpes! Además, como íbamos a esquiar pues tb nos enviamos los skis... Te recogen en casa y te entregan donde tu digas. Además está bastante bien de precio... Vale la pena hacer el boicot que dices!! Os dejo el link, podéis haceros vuestro propio presupuesto en un momento:\r\n\r\nwww.sinmaletas.com\r\n\r\n" | <Comment1294422287208> | "Sí, ojalá existiera tal asociación, pero de momento... Toca aguantarse con las medidas de Iberia. Tarde o temprano acabaremos pasando por el aro si ellos quieren. Yo de momento, ya ni facturo mis maletas. Allá dónde vaya me las envío con un servicio que he conocido hace poco. Soy estudiante universitaria y cuando me fui

	de Erasmus me envié todas las cajas con efectos personales a mi destino. La página se llama SINMALETAS.COM y tiene unos precios muy asequibles. Os la recomiendo! Así podemos empezar con un mini boicoteo a la facturación de maletas...\r\n\r\nUn saludo!" | <Comment1294422287519> | "Hola Paco, me imagino que cuando Iberia plantea esta posibilidad será porque bajará los billetes en otro tanto y que realmente ese sobrecoste lo pague el que use realmente el servicio.\r\nEn caso contrario, no lo entendería porque, desde el más elemental punto de vista marketiniano, cobrar más por lo mismo, cuando la competencia es aún más barata, es ir contra tus propios intereses.\r\nEstoy seguro de que, si esto sigue adelante, lo harán únicamente en aquellas rutas en las que sean la única compañía de tránsito o tenga muy escasa competencia.\r\nNo me imagino un vuelo Madrid-Bogotá con sobrecoste de maletas. Air Comet se forraría a costa de estos indocumentados de Iberia. Y no son tan tontos.\r\nUn abrazo." | <Comment1294422287685> | "Creo que las compañías aéreas (al igual que las de telefonía y similares), optan por una política de hechos consumados. Iberia ha causado este revuelo simplemente porque se ha filtrado la posibilidad de cobro por el equipaje, pero muchas low cost ya lo hacen. En España, mientras no cambie la legislación no pueden cobrar por el equipaje facturado, y ya hay demandas por ello, pero con la velocidad que caracteriza a nuestro sistema judicial seguramente se modificará la ley de aviación civil en favor de las compañías antes de que exista sentencia en contra de esta práctica.\r\nPor lo demás, como señalan más arriba, creo que los hábitos a la hora de hacer el equipaje también han cambiado; muchas personas optan por viajar solo con equipaje de cabina, y cuando quieran cobrarlo, pues iremos con las manos en los bolsillos; y cuando nos quieran meter en los aviones de pie, o en tumbados en fila en plan literas también tragaremos...\r\nSería genial boicotear a estas compañías, aunque solo fuera un día, perooo no creo que consigamos unirnos hasta ese extremo. Aunque soñar es bonito.\r\nSaludos!" | <Comment1294422287831> | "Estas aerolineas no se que esperan... yo viajo mucho, y muchas veces con estas low cost, y he notado que mi comportamiento al hacer la \"maleta\" (¡o mochila!) ha cambiado mucho: la mayoría del tiempo ni me molesto en empacar articulos de baño: los compro en el destino, y si sobra algo los dejo allí. Al final sale mas barato que arriesgarme a sobrepasarme el limite de peso. Al fin y al cabo las aerolíneas nos llevaran a todos a algo parecido: viajaremos solo con lo puesto, y al llegar al destino compraremos todo nuevo y luego lo dejaremos allí tirado al volver. Que viva el consumismo. " | <Comment1294422287997> | "Hola Paco:\r\nPersonalmente, no confío en absoluto en las asociaciones de consumidores. Mi experiencia: cuando Aerolíneas Argentinas rompió y perdió mi silla de ruedas. Pedí ayuda a la Unión de Consumidores de Aragón: porque entendí que mis derechos como consumidor fueron lesionados y podía afectar a miles y miles de personas con discapacidades diferentes que diariamente se ven perjudidados. RESPUESTA: es un problema particular y no podemos llevar un posible juicio (soy socio, pero pagando, sí podían encargarse).\r\nLamentable, ¿no creeis??\r\nUn abrazo\r\nMiguel\r\n----------\r\nwww.miguel-asaltodemata.blogspot.com" | <Comment1294422288175> | "Tienes razón Paco. ¿Pero es que no hay nadie que nos pueda defender?." | <Comment1294422288559> | "dónde firmo?" | <Comment1294422288720> | "Critico duramente el 'mamoneo' de estas compañías, multinacionales y similares. Aquí queda mi protesta contra ellas y, en concreto, contra 'Iberia'. Pero -por favor- no me hagas unirme a otros consumidores porque hay ¡¡¡cada elemento!!!, tan peligroso como las multinacionales, o más.\r\nUn abrazo." | <Comment1294422288898> | "YO NO CREO QUE SEA CIENCIA FICCIÓN TODOS DEBEMOS SER CONSECUENTES Y NO CRITICAR TANTO A CIERTAS MARCAS Y LUEGO VESTIR NUESTROS HIJOS CON ELLAS. MI PADRE SIEMPRE DICE QUE LA LEY EMPIEZA POR UNO MISMO. BUEN POST. BESOS" | <Comment1294422289136> | "¿Sabes lo que pasa? Que los de Iberia son seguidores de Michio Kaku, y se han dado cuenta de que en cuatro días (es que tal como lo cuenta...) vamos a ponernos todos unos megacinturones y no vamos a necesitar ni aviones ni ná para volar; además como mucho llevaremos con nosotros una maquinita de construcción molecular y qué necesitamos: ¿un pijama? Pues lo programamos y ya está, sobre la marcha. Luego además, el último grito, la teletransportación y casi ná para el optimismo: porque vivir ciento cincuenta años será lo normal y además superapañados, qué problema va a ver en esa society del futuro, donde el tiempo servirá para leer y siempre estar

	haciendo y pensando cosas edificantes. Entonces, lo que te decía, que Iberia ha caído en la cuenta y se ha propuesto hacer el agosto perpetuo, hasta que se le acabe el chollo. Yo creo que va a ser por algo así. Y no porque piensen que somos todos unos pringaos que como ovejas cantamos cuatro patas sí dos no y que nos den y que nos den y que nos vuelvan a dar. Por cierto hoy he soñado contigo y qué jdd calambre me despertó. Que el día que soñé con Michio no me pasó eso pero a cambio me sentí luego muy rara todo el día porque era la primera vez que me besaba un tío tan mayor. Como para no acordarme. Bss.\r\n\r\nhttp://www.youtube.com/watch?v=PEdE2TicKtg" | <Comment1294422289449> | "Esa asociación sería maravillosa, si fuésemos capaces todos de quitarnos las vendas de los ojos. Realmente los consumidores tenemos el poder. Si no compramos determinados productos somos capaces de condicionar a las empresas para que lleven a cabo prácticas menos abusivas.\r\nEstoy contigo.\r\nUn abrazo" | <Comment1294422289620> | "... o al autostop...." | <Comment1294422289798> | "A este paso volvemos a las diligencias....."
Time	643 ms
Nº	CQ 17
Competency question	Which comments has the user "Blas" made in any of the posts of paco nadal?
SPARQL query	?wasControlled opmo:cause opmv:paco-nadal. ?wasControlled opmo:effect ?process. ?gen opmo:cause ?process. ?gen opmo:effect ?post.?post a sioc:Post. ?post sioc:has_reply ?comment. ?comment sioc:has_creator sioc:AccBlas. }
Answer	(Limited to 14 results (there are too many)) |post | comment |<lacomunidad.elpais.com/paco-nadal/2010/11/10/saharauis-como-mantener-dignidad-el-infierno> | <Comment1294421522766> | |<lacomunidad.elpais.com/paco-nadal/2010/11/9/el-27-estamos-esquiando> | <Comment1294421524062> | <lacomunidad.elpais.com/paco-nadal/2010/11/2/-de-verdad-van-desaparecer-guias-viaje-papel-> | <Comment1294421536399> | |<lacomunidad.elpais.com/paco-nadal/2010/10/30/fin-fiesta-ahogado-vino> | <Comment1294421543011> | |<lacomunidad.elpais.com/paco-nadal/2010/10/29/el-garito-mas-canalla-bilbao> | <Comment1294421544611> | |<lacomunidad.elpais.com/paco-nadal/2010/10/28/un-paseo-gijon-y-sitio-cojonudo-comer-san-vicente> | <Comment1294421544834> | |lacomunidad.elpais.com/paco-nadal/2010/10/27/morirse-galicia> | <Comment1294421547387> | |<lacomunidad.elpais.com/paco-nadal/2010/10/26/valladolid-dia-que-enterramos-francino> | <Comment1294421550065> | |<lacomunidad.elpais.com/paco-nadal/2010/10/24/badajoz-existe> | <Comment1294421552821> | |<lacomunidad.elpais.com/paco-nadal/2010/10/21/un-tarzan-cordoba> | <Comment1294421555657> | | <lacomunidad.elpais.com/paco-nadal/2010/10/20/de-borrachera-toledo-con-bunuel>
Time	751 ms
Nº	CQ 18
Competency question	Which posts have been created by paco-nadal?
SPARQL query	SELECT distinct ?post WHERE { ?wasControlled opmo:cause opmv:paco-nadal. ?wasControlled opmo:effect ?process. ?wasControlled opmo:role ?role. ?role a opmo:ContentCreatorRole. ?gen opmo:cause ?process. ?gen opmo:effect ?post.

	?post a sioc:Post. }
Answer	(Limited to 14 results, since there are more than 100 different) |<lacomunidad.elpais.com/paco-nadal/2010/11/12/cine-viajes-el-ocejon> <lacomunidad.elpais.com/paco-nadal/2010/11/10/saharauis-como-mantener-dignidad-el-infierno> <lacomunidad.elpais.com/paco-nadal/2010/11/9/el-27-estamos-esquiando> <lacomunidad.elpais.com/paco-nadal/2010/11/8/para-cenar-lengua-flamenco-y-talon-camello-mi-vida-como> <lacomunidad.elpais.com/paco-nadal/2010/11/5/diario-cayo-caementicius-mi-vida-como-romano-parte-ii-> <lacomunidad.elpais.com/paco-nadal/2010/11/4/mi-vida-como-romano-parte-i-> <lacomunidad.elpais.com/paco-nadal/2010/11/2/-de-verdad-van-desaparecer-guias-viaje-papel-> <lacomunidad.elpais.com/paco-nadal/2010/10/30/fin-fiesta-ahogado-vino> <lacomunidad.elpais.com/paco-nadal/2010/10/29/el-garito-mas-canalla-bilbao> <lacomunidad.elpais.com/paco-nadal/2010/10/28/un-paseo-gijon-y-sitio-cojonudo-comer-san-vicente> <lacomunidad.elpais.com/paco-nadal/2010/10/27/morirse-galicia> <lacomunidad.elpais.com/paco-nadal/2010/10/26/valladolid-dia-que-enterramos-francino> <lacomunidad.elpais.com/paco-nadal/2010/10/24/badajoz-existe> <lacomunidad.elpais.com/paco-nadal/2010/10/21/un-tarzan-cordoba>
Time	657 ms
Nº	CQ 19
Competency question	How has the post from user “paco-nadal”: “Iberia, de mal en peor” evolved?

SPARQL query	SELECT distinct ?evolution WHERE {
	?post sioc:title "Iberia, de mal en peor".
	?evolution opmo:cause ?post.
	?evolution a opmo:WasDerivedFrom.
	}

Answer	None (it has not been updated!)
Time	633 ms
Nº	CQ 20
Competency question	Who are the followers of user Paco Nadal?

SPARQL query	SELECT distinct ?acc ?follower WHERE {
	?acc sioc:account_of opmv:paco-nadal.
	?follower sioc:follows ?acc.
	}

Answer	None (data not yet available).
Time	595 ms
Nº	CQ 21
Competency question	Who are the friends of user Paco Nadal?

SPARQL query	SELECT distinct ?acc ?follower WHERE {
	?acc sioc:account_of opmv:paco-nadal.
	?follower sioc:has_friend ?acc.
	}

Answer	None (data not yet available)
Time	612 ms

Table A.2: Answers to the competency questions

Average time per query: **728.9 ms**

Maximum response time: **908 ms**
Minimum response time: **570 ms**

A.3: Extended OPMO Ontology

In this appendix we show the extended OPMO ontology, after expanding it and adapting it to our scenario. We have divided in three different tables: classes, properties and datatype properties.

A.3.1 Classes

http://metadata.net/mpeg7/mpeg7.owl#Image Annotations rdfs:label: "Image" () rdfs:comment: "The class of images" () Superclasses MultimediaContent Subclasses StillRegion StillRegion3D Usage DataProperties Domain: height domain Image width domain Image
http://metadata.net/mpeg7/mpeg7.owl#Video Annotations rdfs:comment: "The class of videos" () rdfs:label: "Video" () Superclasses MultimediaContent Subclasses VideoSegment
http://openprovenance.org/model/opmo#AValue Annotations rdfs:label: "AValue" rdfs:comment: "The serial representation of an artifact Value" rdfs:isDefinedBy: "http://openprovenance.org/model/opmo" Superclasses Entity DataSomeValuesFrom content rdfs:Literal] DataSomeValuesFrom encoding xsd:anyURI]

Usage
DataProperties Domain:
content domain AValue
encoding domain AValue
ObjectProperties Range:
avalue range AValue

http://openprovenance.org/model/opmo#Account
Annotations
rdfs:comment: "The class representing an OPM Account."
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:label: "Account"
Superclasses
Entity
Usage
ObjectProperties Range:
account range Account
hasAccount range Account

http://openprovenance.org/model/opmo#AditionalInformationRole
Tree View
Role
AditionalInformationRole
Superclasses
Role

http://openprovenance.org/model/opmo#Annotable
Annotations
rdfs:label: "Annotable"
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:comment: "The set of OPM entities that can be annotated."
Superclasses
Entity
Equivalent Classes
[OR Account Annotation Edge Node OPMGraph Role]
Usage
DataProperties Domain:
datapropertyAbbreviation domain Annotable
label domain Annotable
pname domain Annotable
profile domain Annotable
type domain Annotable
ObjectProperties Domain:
annotation domain Annotable

http://openprovenance.org/model/opmo#Annotation
Annotations

rdfs:comment: "OPM class used to annotate Annotable entities."
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:label: "Annotation"

Superclasses

Entity

Usage

ObjectProperties Domain:

account domain Annotation
property domain Annotation

ObjectProperties Range:

annotation range Annotation

http://openprovenance.org/model/opmo#ContentCreatorRole

Superclasses

Role

http://openprovenance.org/model/opmo#ContentModifierRole

Superclasses

Role

http://openprovenance.org/model/opmo#ContentPublisherRole

Superclasses

Role

http://openprovenance.org/model/opmo#Edge

Annotations

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:label: "Edge"
rdfs:comment: "A (causal) relationship is represented by an arc and denotes the presence of a dependency between the source of the arc (the effect) and the destination of the arc (the cause)."

Superclasses

Entity

Subclasses

Used
WasControlledBy
WasDerivedFrom
WasGeneratedBy
WasTriggeredBy

Usage

ObjectProperties Domain:

account domain Edge
cause domain Edge
effect domain Edge

ObjectProperties Range:

effectInverse range Edge
hasDependency range Edge

http://openprovenance.org/model/opmo#Entity

Annotations

rdfs:label: "Entity"
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:comment: "The class of all constituents of an OPM graph."

Subclasses

AValue
Account
Annotable
Annotation
Edge
EventEdge
Node
OPMGraph
OTime
Property
Role

http://openprovenance.org/model/opmo#EventEdge

Annotations

rdfs:label: "EventEdge"
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:comment: "An EventEdge denotes an Edge associated with a time instant."

Superclasses

Entity

Equivalent Classes

[OR Used WasGeneratedBy]

Usage

ObjectProperties Domain:
time domain EventEdge

http://openprovenance.org/model/opmo#IllustrationRole

Superclasses

Role

http://openprovenance.org/model/opmo#LaterVersionThan

Annotations

rdfs:comment: "Edge to deal with the versioning issues when modeling mutable artifacts" ()

Superclasses

WasDerivedFrom

http://openprovenance.org/model/opmo#Node

Annotations

rdfs:label: "Node"
rdfs:comment: "Node is the class of nodes in an OPM graph. Nodes can be a source or effect of edges."

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Superclasses

Entity

Subclasses

Agent
Artifact
Process

Usage

ObjectProperties Domain:

account domain Node
effectInverse domain Node

ObjectProperties Range:

cause range Node
effect range Node

http://openprovenance.org/model/opmo#OPMGraph

Annotations

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:comment: "The class of all OPM graphs."
rdfs:label: "OPMGraph"

Superclasses

Entity

Usage

ObjectProperties Domain:

hasAccount domain OPMGraph
hasAgent domain OPMGraph
hasArtifact domain OPMGraph
hasConstituent domain OPMGraph
hasDependency domain OPMGraph
hasProcess domain OPMGraph

http://openprovenance.org/model/opmo#OTime

Annotations

rdfs:label: "OTime"
rdfs:comment: "Observed time."
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Superclasses

Entity

Usage

DataProperties Domain:

exactlyAt domain OTime
noEarlierThan domain OTime
noLaterThan domain OTime

ObjectProperties Range:

endTime range OTime
startTime range OTime

time range OTime

http://openprovenance.org/model/opmo#Property

Annotations

rdfs:comment: "Building block allowing for the construction of annotations. It consists of key-value pair."

rdfs:label: "Property"

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Superclasses

Entity

DataSomeValuesFrom value rdfs:Literal]

DataAllValuesFrom value rdfs:Literal]

Usage

DataProperties Domain:

value domain Property

ObjectProperties Domain:

key domain Property

ObjectProperties Range:

property range Property

http://openprovenance.org/model/opmo#ReferenceRole

Superclasses

Role

http://openprovenance.org/model/opmo#Role

Annotations

rdfs:comment: "A role designates an artifact's or agent's function in a process. Roles are constituents of "used", "was generated by", and "was controlled by" edges, aimed at distinguishing the nature of the dependency when multiple such edges are connected to a same process."

rdfs:label: "Role"

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Superclasses

Entity

DataAllValuesFrom value xsd:string]

Subclasses

AditionalInformationRole

ContentCreatorRole

ContentModifierRole

ContentPublisherRole

IllustrationRole

ReferenceRole

SummaryRole

Usage

DataProperties Domain:

value domain Role

ObjectProperties Range:

role range Role

http://openprovenance.org/model/opmo#SummaryRole

Superclasses

Role

http://openprovenance.org/model/opmo#Used

Annotations

rdfs:comment: "A “used” edge from process to an artifact is a relationship intended to indicate that the process required the availability of the artifact to be able to complete its execution. When several artifacts are connected to a same process by multiple “used” edges, all of them were required for the process to complete. Used is a class that encompasses all the properties defined by OPM for this kind of edge. It is a reification of the opmv:used property."

rdfs:label: "Used"

rdfs:seeAlso: "http://openprovenance.org/model/opmo#Used"

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Superclasses

Edge

[SOME causeUsed Artifact]

[SOME effectUsed Process]

[SOME role Role]

[ALL causeUsed Artifact]

[ALL effectUsed Process]

Disjoint Classes

WasControlledBy

WasDerivedFrom

WasGeneratedBy

WasTriggeredBy

Usage

ObjectProperties Domain:

causeUsed domain Used

effectUsed domain Used

role domain Used

ObjectProperties Range:

effectUsedInverse range Used

http://openprovenance.org/model/opmo#WasControlledBy

Annotations

rdfs:seeAlso: "http://purl.org/net/opmv/ns#wasControlledBy"

rdfs:label: "WasControlledBy"

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

rdfs:comment: "An edge “was controlled by” from a process P to an agent Ag is a dependency that indicates that the start and end of process P was controlled by agent Ag. WasControlledBy is a class that encompasses all the properties defined by OPM for this kind of edge. It is a reification of the opmv:wasControlledBy property."

Superclasses

Edge

[SOME causeWasControlledBy Agent]

[SOME effectWasControlledBy Process]
[SOME role Role]
[ALL causeWasControlledBy Agent]
[ALL effectWasControlledBy Process]

Disjoint Classes

Used
WasDerivedFrom
WasGeneratedBy
WasTriggeredBy

Usage

ObjectProperties Domain:

causeWasControlledBy domain WasControlledBy
effectWasControlledBy domain WasControlledBy
endTime domain WasControlledBy
role domain WasControlledBy
startTime domain WasControlledBy

ObjectProperties Range:

effectWasControlledByInverse range WasControlledBy

http://openprovenance.org/model/opmo#WasDerivedFrom

Annotations

rdfs:label: "WasDerivedFrom"

rdfs:comment: "An edge “was derived from” from artifact A2 to artifact A1 is a relationship that indicates that artifact A1 needs to have been generated for A2 to be generated. The piece of state associated with A2 is dependent on the presence of A1 or on the piece of state associated with A1 . WasDerivedFrom is a class that encompasses all the properties defined by OPM for this kind of edge. It is a reification of the opmv:wasDerivedFrom property."

rdfs:seeAlso: "http://purl.org/net/opmv/ns#wasDerivedFrom"

rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Superclasses

Edge
[SOME causeWasDerivedFrom Artifact]
[SOME effectWasDerivedFrom Artifact]
[ALL causeWasDerivedFrom Artifact]
[ALL effectWasDerivedFrom Artifact]

Subclasses

LaterVersionThan

Disjoint Classes

Used
WasControlledBy
WasGeneratedBy
WasTriggeredBy

Usage

ObjectProperties Domain:

causeWasDerivedFrom domain WasDerivedFrom

effectWasDerivedFrom domain WasDerivedFrom
ObjectProperties Range:
effectWasDerivedFromInverse range WasDerivedFrom

http://openprovenance.org/model/opmo#WasGeneratedBy

Annotations

rdfs:seeAlso: "http://purl.org/net/opmv/ns#wasGeneratedBy"
rdfs:label: "WasGeneratedBy"
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:comment: "A “was generated by” edge from an artifact to a process is a relationship intended to mean that the process was required to initiate its execution for the artifact to have been generated. When several artifacts are connected to a same process by multiple “was generated by” edges, the process had to have begun, for all of them to be generated. WasGeneratedBy is a class that encompasses all the properties defined by OPM for this kind of edge. It is a reification of the opmv:wasGeneratedBy property."

Superclasses

Edge
[SOME causeWasGeneratedBy Process]
[SOME effectWasGeneratedBy Artifact]
[SOME role Role]
[ALL causeWasGeneratedBy Process]
[ALL effectWasGeneratedBy Artifact]

Disjoint Classes

Used
WasControlledBy
WasDerivedFrom
WasTriggeredBy

Usage

ObjectProperties Domain:
causeWasGeneratedBy domain WasGeneratedBy
effectWasGeneratedBy domain WasGeneratedBy
role domain WasGeneratedBy
ObjectProperties Range:
effectWasGeneratedByInverse range WasGeneratedBy

http://openprovenance.org/model/opmo#WasTriggeredBy

Annotations

rdfs:label: "WasTriggeredBy"
rdfs:comment: "An edge “was triggered by” from a process P2 to a process P1 is a causal dependency that indicates that the start of process P1 was required for P2 to be able to complete. WasTriggeredBy is a class that encompasses all the properties defined by OPM for this kind of edge. It is a reification of the opmv:wasTriggeredBy property."
rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
rdfs:seeAlso: "http://purl.org/net/opmv/ns#wasTriggeredBy"

Superclasses

Edge

[SOME causeWasTriggeredBy Process]
[SOME effectWasTriggeredBy Process]
[ALL causeWasTriggeredBy Process]
[ALL effectWasTriggeredBy Process]

Disjoint Classes

Used
WasControlledBy
WasDerivedFrom
WasGeneratedBy

Usage

ObjectProperties Domain:
causeWasTriggeredBy domain WasTriggeredBy
effectWasTriggeredBy domain WasTriggeredBy

ObjectProperties Range:
effectWasTriggeredByInverse range WasTriggeredBy

http://purl.org/net/opmv/ns#Agent

Superclasses

Node

Subclasses

ContentProvider
Organization

Usage

ObjectProperties Domain:
hasRecommended domain Agent

ObjectProperties Range:
causeWasControlledBy range Agent
hasAgent range Agent
recommendedBy range Agent

http://purl.org/net/opmv/ns#AggregationProcess

Superclasses

Process

http://purl.org/net/opmv/ns#Artifact

Superclasses

Node

Subclasses

Image
Video
ExternalResource
Guide
Post

Usage

DataProperties Domain:
artifactDescription domain Artifact

ObjectProperties Domain:
avalue domain Artifact

effectWasDerivedFromInverse domain Artifact
effectWasGeneratedByInverse domain Artifact
recommendedBy domain Artifact
wasDerivedFromStar domain Artifact
wasGeneratedByStar domain Artifact

ObjectProperties Range:

causeUsed range Artifact
causeWasDerivedFrom range Artifact
effectWasDerivedFrom range Artifact
effectWasGeneratedBy range Artifact
hasArtifact range Artifact
hasRecommended range Artifact
usedStar range Artifact
wasDerivedFromStar range Artifact

http://purl.org/net/opmv/ns#ContentProvider

Superclasses

Agent

http://purl.org/net/opmv/ns#CreationProcess

Superclasses

Process

http://purl.org/net/opmv/ns#ExternalResource

Superclasses

Artifact

Disjoint Classes

Guide
Post

http://purl.org/net/opmv/ns#Guide

Superclasses

Artifact
Item

Disjoint Classes

ExternalResource

http://purl.org/net/opmv/ns#ModificationProcess

Superclasses

Process

http://purl.org/net/opmv/ns#Organization

Superclasses

Agent

http://purl.org/net/opmv/ns#Process

Superclasses

Node

Subclasses

AggregationProcess
CreationProcess

ModificationProcess
PublicationProcess

Usage

ObjectProperties Domain:

effectUsedInverse domain Process
effectWasControlledByInverse domain Process
effectWasTriggeredByInverse domain Process
usedStar domain Process

ObjectProperties Range:

causeWasGeneratedBy range Process
causeWasTriggeredBy range Process
effectUsed range Process
effectWasControlledBy range Process
effectWasTriggeredBy range Process
hasProcess range Process
wasGeneratedByStar range Process

http://purl.org/net/opmv/ns#PublicationProcess

Superclasses

Process

http://www.w3.org/2003/01/geo/wgs84_pos#Point

Annotations

rdfs:comment: "A point, typically described using a coordinate system relative to Earth, such as WGS84." ()

rdfs:label: "point" ()

rdfs:comment: "Uniquely identified by lat/long/alt. i.e. spaciallyIntersects(P1, P2) :- lat(P1, LAT), long(P1, LONG), alt(P1, ALT), lat(P2, LAT), long(P2, LONG), alt(P2, ALT). sameThing(P1, P2) :- type(P1, Point), type(P2, Point), spaciallyIntersects(P1, P2)." ()

Superclasses

SpatialThing

http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing

Annotations

rdfs:label: "SpatialThing" ()

rdfs:comment: "Anything with spatial extent, i.e. size, shape, or position. e.g. people, places, bowling balls, as well as abstract areas like cubes." ()

Subclasses

Point

Usage

DataProperties Domain:

alt domain SpatialThing
lat domain SpatialThing
long domain SpatialThing

ObjectProperties Range:

location range SpatialThing

Table A.3: Classes of the extended OPMO

A.3.2 Properties

http://openprovenance.org/model/opmo#account

Annotations

- rdfs:label: "account"
- rdfs:comment: "Object Property to express the member of an OPM entity to some Account."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains

Ranges

- Account

http://openprovenance.org/model/opmo#annotation

Annotations

- rdfs:label: "annotation"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "Object property to associate an Annotable entity and an Annotation."

Domains

- Annotable

Ranges

- Annotation

http://openprovenance.org/model/opmo#avalue

Annotations

- rdfs:comment: "Denotes a serialization of an application value associated with an Artifact. Such serialization should have a type (expressed in a type system suitable for the serialization). Serialization technologies include XML, JSON, and ntriples."
- rdfs:label: "avalue"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains

- Artifact

Ranges

- AValue

http://openprovenance.org/model/opmo#cause

Annotations

- rdfs:comment: "The cause of an Edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "cause"

Sub Object Properties

- causeUsed
- causeWasControlledBy
- causeWasDerivedFrom
- causeWasGeneratedBy
- causeWasTriggeredBy

Domains
- Edge

Ranges
- Node

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#causeUsed

Annotations
- rdfs:label: "causeUsed"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "The cause of a Used edge."

Super Object Properties
- cause

Domains
- Used

Ranges
- Artifact

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#causeWasControlledBy

Annotations
- rdfs:label: "causeWasControlledBy"
- rdfs:comment: "The cause of a WasControlledBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- cause

Domains
- WasControlledBy

Ranges
- Agent

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#causeWasDerivedFrom

Annotations
- rdfs:comment: "The cause of a WasDerivedFrom edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "causeWasDerivedFrom"

Super Object Properties
- cause

Domains
- WasDerivedFrom

Ranges
- Artifact

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#causeWasGeneratedBy
Annotations
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "The cause of a WasGeneratedBy edge."
- rdfs:label: "causeWasGeneratedBy"

Super Object Properties
- cause

Domains
- WasGeneratedBy

Ranges
- Process

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#causeWasTriggeredBy
Annotations
- rdfs:comment: "The cause of a WasTriggeredBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "causeWasTriggeredBy"

Super Object Properties
- cause

Domains
- WasTriggeredBy

Ranges
- Process

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#effect
Annotations
- rdfs:label: "effect"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "The effect of an Edge."

Sub Object Properties
- effectUsed
- effectWasControlledBy
- effectWasDerivedFrom
- effectWasGeneratedBy
- effectWasTriggeredBy

Domains
- Edge

Ranges
- Node

Inverse Object Properties
- effectInverse

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#effectInverse

Annotations
- rdfs:comment: "Convenience class introduced to describe the inverse of an effect. It is used to express property chains."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "effectInverse"

Sub Object Properties
- effectUsedInverse
- effectWasControlledByInverse
- effectWasDerivedFromInverse
- effectWasGeneratedByInverse
- effectWasTriggeredByInverse

Domains
- Node

Ranges
- Edge

Inverse Object Properties
- effect

http://openprovenance.org/model/opmo#effectUsed

Annotations
- rdfs:comment: "The effect of a Used edge."
- rdfs:label: "effectUsed"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- effect

Domains
- Used

Ranges
- Process

Inverse Object Properties
- effectUsedInverse

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#effectUsedInverse

Annotations
- rdfs:comment: "The cause of a Process by means of a Used edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "effectUsedInverse"

Super Object Properties

- effectInverse

Domains
- Process

Ranges
- Used

Inverse Object Properties
- effectUsed

Characteristics
- isInverseFunctional

http://openprovenance.org/model/opmo#effectWasControlledBy

Annotations
- rdfs:comment: "The effect of a WasControlledBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "effectWasControlledBy"

Super Object Properties
- effect

Domains
- WasControlledBy

Ranges
- Process

Inverse Object Properties
- effectWasControlledByInverse

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#effectWasControlledByInverse

Annotations
- rdfs:comment: "The cause of a Process by means of a WasControlledBy edge."
- rdfs:label: "effectWasControlledByInverse"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- effectInverse

Domains
- Process

Ranges
- WasControlledBy

Inverse Object Properties
- effectWasControlledBy

Characteristics
- isInverseFunctional

http://openprovenance.org/model/opmo#effectWasDerivedFrom

Annotations
- rdfs:label: "effectWasDerivedFrom"
- rdfs:comment: "The effect of a WasDerivedFrom edge."

- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- effect

Domains
- WasDerivedFrom

Ranges
- Artifact

Inverse Object Properties
- effectWasDerivedFromInverse

http://openprovenance.org/model/opmo#effectWasDerivedFromInverse

Annotations
- rdfs:label: "effectWasDerivedFromInverse"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "The cause of an Artifact by means of a WasDerivedFrom edge."

Super Object Properties
- effectInverse

Domains
- Artifact

Ranges
- WasDerivedFrom

Inverse Object Properties
- effectWasDerivedFrom

Characteristics
- isInverseFunctional

http://openprovenance.org/model/opmo#effectWasGeneratedBy

Annotations
- rdfs:label: "effectWasGeneratedBy"
- rdfs:comment: "The effect of a WasGeneratedBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- effect

Domains
- WasGeneratedBy

Ranges
- Artifact

Inverse Object Properties
- effectWasGeneratedByInverse

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#effectWasGeneratedByInverse

Annotations
- rdfs:label: "effectWasGeneratedByInverse"

- rdfs:comment: "The cause of an Artifact by means of a WasGeneratedBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties

- effectInverse

Domains

- Artifact

Ranges

- WasGeneratedBy

Inverse Object Properties

- effectWasGeneratedBy

Characteristics

- isInverseFunctional

http://openprovenance.org/model/opmo#effectWasTriggeredBy

Annotations

- rdfs:label: "effectWasTriggeredBy"
- rdfs:comment: "The effect of a WasTriggeredBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties

- effect

Domains

- WasTriggeredBy

Ranges

- Process

Inverse Object Properties

- effectWasTriggeredByInverse

Characteristics

- isFunctional

http://openprovenance.org/model/opmo#effectWasTriggeredByInverse

Annotations

- rdfs:label: "effectWasTriggeredByInverse"
- rdfs:comment: "The cause of a Process by means of a WasTriggeredBy edge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties

- effectInverse

Domains

- Process

Ranges

- WasTriggeredBy

Inverse Object Properties

- effectWasTriggeredBy

Characteristics

- isInverseFunctional

http://openprovenance.org/model/opmo#endTime

Annotations
- rdfs:label: "endTime"
- rdfs:comment: "The time at which the agent ended controlling a process."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains
- WasControlledBy

Ranges
- OTime

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#hasAccount

Annotations
- rdfs:label: "hasAccount"
- rdfs:comment: "Property that denotes the constituency relationship between an OPM graph and an account, meaning that the object of this property is an account of the subject."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- hasConstituent

Domains
- OPMGraph

Ranges
- Account

http://openprovenance.org/model/opmo#hasAgent

Annotations
- rdfs:comment: "Property that denotes the constituency relationship between an OPM graph and an agent, meaning that the object of this property is an agent of the subject."
- rdfs:label: "hasAgent"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties
- hasConstituent

Domains
- OPMGraph

Ranges
- Agent

http://openprovenance.org/model/opmo#hasArtifact

Annotations
- rdfs:label: "hasArtifact"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "Property that denotes the constituency relationship between an OPM graph and an artifact, meaning that the object of this property is an artifact of the subject."

Super Object Properties

- hasConstituent

Domains

- OPMGraph

Ranges

- Artifact

http://openprovenance.org/model/opmo#hasConstituent

Annotations

- rdfs:label: "hasContituent"
- rdfs:comment: "(Abstract) Property that denotes the constituency relationship between an OPM graph and one of its constituent entity, meaning that the object of this property is a constituent of the subject."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Sub Object Properties

- hasAccount
- hasAgent
- hasArtifact
- hasDependency
- hasProcess

Domains

- OPMGraph

http://openprovenance.org/model/opmo#hasDependency

Annotations

- rdfs:label: "hasDependency"
- rdfs:comment: "Property that denotes the constituency relationship between an OPM graph and an edge, meaning that the object of this property is an edge of the subject."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties

- hasConstituent

Domains

- OPMGraph

Ranges

- Edge

http://openprovenance.org/model/opmo#hasProcess

Annotations

- rdfs:comment: "Property that denotes the constituency relationship between an OPM graph and a process, meaning that the object of this property is a process of the subject."
- rdfs:label: "hasProcess"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Object Properties

- hasConstituent

Domains

- OPMGraph

Ranges
- Process

http://openprovenance.org/model/opmo#hasRecommended

Domains
- Agent

Ranges
- Artifact

Inverse Object Properties
- recommendedBy

http://openprovenance.org/model/opmo#key

Annotations
- rdfs:label: "key"
- rdfs:comment: "The key of a Property."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains
- Property

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#property

Annotations
- rdfs:label: "property"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "Object Property that associates an Annotation instance with a (set of) Propery(ies)."

Domains
- Annotation

Ranges
- Property

http://openprovenance.org/model/opmo#recommendedBy

Domains
- Artifact

Ranges
- Agent

Inverse Object Properties
- hasRecommended

http://openprovenance.org/model/opmo#role

Annotations
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "role"
- rdfs:comment: "The role of an edge."

Domains

Ranges
- Role

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#startTime

Annotations
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "startTime"
- rdfs:comment: "The time at which the agent began controlling a process."

Domains
- WasControlledBy

Ranges
- OTime

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#time

Annotations
- rdfs:comment: "A piece of timing information associated with an EventEdge."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "time"

Domains
- EventEdge

Ranges
- OTime

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#usedStar

Annotations
- rdfs:comment: "A multi-step used property."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "used*"

Sub Object Properties
- used

Domains
- Process

Ranges
- Artifact

http://openprovenance.org/model/opmo#wasDerivedFromStar

Annotations
- rdfs:comment: "the transitive closure of wasDerivedFrom" ()
- rdfs:label: "wasDerivedFrom*"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Sub Object Properties
- wasDerivedFrom

Domains

• Artifact Ranges • Artifact Characteristics • isTransitive
http://purl.org/net/opmv/ns#used Super Object Properties • usedStar
http://purl.org/net/opmv/ns#wasDerivedFrom Annotations • rdfs:seeAlso: "http://openprovenance.org/opmo#WasDerivedFrom" () Super Object Properties • wasDerivedFromStar
http://purl.org/net/opmv/ns#wasGeneratedBy Super Object Properties • wasGeneratedByStar
http://purl.org/net/opmv/ns#wasTriggeredBy
http://rdfs.org/ns/void#exampleResource Usage • opmo exampleResource Used
http://www.w3.org/2003/01/geo/wgs84_pos#location Annotations • rdfs:comment: "The relation between something and the point, or other geometrical thing in space, where it is. For example, the realtionship between a radio tower and a Point with a given lat and long. Or a relationship between a park and its outline as a closed arc of points, or a road and its location as a arc (a sequence of points). Clearly in practice there will be limit to the accuracy of any such statement, but one would expect an accuracy appropriate for the size of the object and uses such as mapping ." () • rdfs:label: "location" () Super Object Properties • based_near Ranges • SpatialThing
http://xmlns.com/foaf/0.1/based_near

Table A.4: Properties of the extended OPMO

A.3.3 Datatype Properties

http://openprovenance.org/model/opmo#IPTCMediaType Annotations

- rdfs:comment: "Media Type of the artifact (if any)" ()

Super Data Properties

- artifactDescription

http://openprovenance.org/model/opmo#IPTCMimeType

Annotations

- rdfs:comment: "Mime Type of the artifact (if any)" ()

Super Data Properties

- artifactDescription

http://openprovenance.org/model/opmo#artifactDescription

Annotations

- rdfs:comment: "Descriptions of a specific artifact. Can be replaced with a resource-centric description of the artifact (from another vocabulary)" ()

Sub Data Properties

- IPTCMediaType
- IPTCMimeType
- language
- longit
- longitMeasure
- page
- subtitle

Domains

- Artifact

http://openprovenance.org/model/opmo#content

Annotations

- rdfs:label: "content" ()
- rdfs:comment: "Content of the serial representation of Artifact."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains

- AValue

Ranges

- rdfs:Literal

Characteristics

- isFunctional

http://openprovenance.org/model/opmo#datapropertyAbbreviation

Annotations

- rdfs:comment: "The OPM Ontology offers several data properties that are abbreviation of common properties defined by OPM."
- rdfs:label: "datapropertyAbbreviation"
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Sub Data Properties

- label
- pname
- profile
- type

Domains

- Annotable

http://openprovenance.org/model/opmo#encoding

Annotations

- rdfs:comment: "Denotes how a serialization was constructed. For instance, using the Java bean serializer to create an XML document, by applying a specified transformation to the application data, e.g. anonymisation, by passing a reference to the actual value, or by creating a set of RDF triples."
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:label: "encoding" ()

Domains

- AValue

Ranges

- xsd:anyURI

Characteristics

- isFunctional

http://openprovenance.org/model/opmo#exactlyAt

Annotations

- rdfs:comment: "The event occurred exactly at this time. This is meant to be an abbreviation for an OTime where noEarlierThan and noLaterThan are identical. exactlyAt should be disjoint from noLaterThan U noEarlierThan."
- rdfs:label: "exactlyAt" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains

- OTime

Ranges

- xsd:dateTime

Characteristics

- isFunctional

http://openprovenance.org/model/opmo#label

Annotations

- rdfs:label: "label" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "This property provides a human-readable version of an OPM entity."

Super Data Properties

- datapropertyAbbreviation

Domains

- Annotable

Ranges

- xsd:string

http://openprovenance.org/model/opmo#language

Annotations

- rdfs:comment: "language of the artifact (if any)" ()

Super Data Properties

- artifactDescription

http://openprovenance.org/model/opmo#longit

Annotations

- rdfs:comment: "longitude of the artifact" ()

Super Data Properties

- artifactDescription

http://openprovenance.org/model/opmo#longitMeasure

Annotations

- rdfs:comment: "longitude measure unit of the artifact (e.g. word)" ()

Super Data Properties

- artifactDescription

http://openprovenance.org/model/opmo#noEarlierThan

Annotations

- rdfs:label: "noEarlierThan" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "the event occurred no earlier than this time"

Domains

- OTime

Ranges

- xsd:dateTime

Characteristics

- isFunctional

http://openprovenance.org/model/opmo#noLaterThan

Annotations

- rdfs:comment: "the event occurred no later than this time"
- rdfs:label: "noLaterThan" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Domains

- OTime

Ranges
- xsd:dateTime

Characteristics
- isFunctional

http://openprovenance.org/model/opmo#page

Annotations
- rdfs:comment: "The page of the artifact (if any)" ()

Super Data Properties
- artifactDescription

http://openprovenance.org/model/opmo#pname

Annotations
- rdfs:label: "pname" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "The persistent name of an entity. Denotes a persistent name that can be used by OPM graph queriers to compare OPM entities across graphs. The scope of this name is intended to be global."

Super Data Properties
- datapropertyAbbreviation

Domains
- Annotable

Ranges
- xsd:anyURI

http://openprovenance.org/model/opmo#profile

Annotations
- rdfs:comment: "This property applies to an OPM graph and denotes a profile that is supported by that graph."
- rdfs:label: "profile" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Data Properties
- datapropertyAbbreviation

Domains
- Annotable

Ranges
- xsd:string

http://openprovenance.org/model/opmo#subtitle

Annotations
- rdfs:comment: "Subtitle of the artifact" ()

Super Data Properties
- artifactDescription

http://openprovenance.org/model/opmo#type

Annotations
- rdfs:comment: "The subtype of an entity."

- rdfs:label: "type" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Data Properties

- datapropertyAbbreviation

Domains

- Annotable

Ranges

- xsd:anyURI

http://openprovenance.org/model/opmo#type

Annotations

- rdfs:comment: "The subtype of an entity."
- rdfs:label: "type" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"

Super Data Properties

- datapropertyAbbreviation

Domains

- Annotable

Ranges

- xsd:anyURI

http://openprovenance.org/model/opmo#value

Annotations

- rdfs:label: "value" ()
- rdfs:isDefinedBy: "http://openprovenance.org/model/opmo"
- rdfs:comment: "the value of a role or a property"

Domains

Characteristics

- isFunctional

http://purl.org/dc/terms/created

Usage

- opmo created 2010-10-12

http://purl.org/dc/terms/modified

Usage

- opmo modified 2010-10-13

http://purl.org/dc/terms/title

Usage

- opmo title The OPM OWL Ontology"@en

http://www.w3.org/2003/01/geo/wgs84_pos#alt

Annotations

- rdfs:comment: "The WGS84 altitude of a SpatialThing (decimal meters above the local reference ellipsoid)." ()
- rdfs:label: "altitude" ()

Domains

- SpatialThing

Characteristics • isFunctional
http://www.w3.org/2003/01/geo/wgs84_pos#lat Annotations • rdfs:comment: "The WGS84 latitude of a SpatialThing (decimal degrees)." () • rdfs:label: "latitude" () Domains • SpatialThing Characteristics • isFunctional
http://www.w3.org/2003/01/geo/wgs84_pos#long Annotations • rdfs:comment: "The WGS84 longitude of a SpatialThing (decimal degrees)." • rdfs:label: "longitude" () Domains • SpatialThing Characteristics • isFunctional

Table A.5: Datatype properties of the extended OPMO

A.4 XML template example

XML template for generic operations depending on the process and the resultant artifacts there may be more than one agent or used artifacts:

```
<?xml version=\"1.0\" encoding=\"UTF-8\"?>
<root>
    <Agentes>
        <Agente>
            <ID>AgenteId1</ID>
            <Tipo>TipoAgenteId1</Tipo>
            <Rol>
                <ID>IdRol1</ID>
                <Tipo>TipoRol1</Tipo>
            </Rol>
            <CoordsControl>
                <Lat>-40.1234</Lat>
                <Long>3.1231234</Long>
            </CoordsControl>
            <HoraControl>
                <HInicio>1212342342123</HInicio>
                <HFin>1212342342123</HFin>
            </HoraControl>
        </Agente>
```

```
      </Agentes>
      <Proceso>
          <ID>IdProceso1</ID>
          <Tipo>TipoProceso1</Tipo>
      </Proceso>
      <Uso>
          <Artefacto>
              <ID>IdArtefUsado1</ID>
              <Tipo>TipoArtefUsado1</Tipo>
              <Rol>
                  <ID>IdRolUso1</ID>
                  <Tipo>TipoRolUso1</Tipo>
              </Rol>
              <HoraUso>
        <Hora>12312432</Hora>
              </HoraUso>
          </Artefacto>
      </Uso>
      <Generacion>
          <Artefacto>
              <ID>IdArtefGenerado1</ID>
              <Tipo>TipoArtefGenerado1</Tipo>
              <Rol>
                  <ID>IdRolGenerado1</ID>
                  <Tipo>TipoRolGenerado1</Tipo>
              </Rol>
              <CoordsGeneracion>
                  <Lat>-40.1234</Lat>
                  <Long>3.1231234</Long>
              </CoordsGeneracion>
              <HoraGeneracion>
                  <Hora>123456789</Hora>
              </HoraGeneracion>
          </Artefacto>
      </Generacion>
</root>
```

Table A.6: XML template example