# A Provenance-Aware Linked Data Application for Trip Management and Organization

Daniel Garijo
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
dgarijo@delicias.dia.fi.upm.es

Boris Villazón-Terrazas
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
bvillazon@fi.upm.es

Oscar Corcho
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
ocorcho@fi.upm.es

## ABSTRACT

We present *El Viajero*, an application[1] for exploiting, managing and organizing Linked Data in the domain of news and blogs about travelling. *El Viajero* makes use of several heterogeneous datasets to help users to plan future trips, and relies on the Open Provenance Model[2] for modelling the provenance information of the resources.

## Categories and Subject Descriptors

E.2 [**Data storage representations**]: Linked Representations

## General Terms

Design, Experimentation

## Keywords

Linked Data, provenance, news, blog

## 1. INTRODUCTION

News content providers rarely include provenance information about the resources that they generate and publish, as already pointed out the W3C Provenance Incubator Group[3] in the News Aggregator Scenario Gap Analysis[4]. This provenance information is critical to determine whether a resource can be trusted or not, since it allows knowing about the source, references and process followed to create the resource.

This paper describes how we exploit the provenance information of guides, images, videos, trips and posts in order to help other users to reuse and explore existing contents to plan their own trips around the world. The application manages the data retrieval transparently to the users, and organizes the results in a map through a graphical interface. It relies on a dataset that we[5] have recently published as Linked Data, containing the provenance information of the editorial resources belonging to *El Viajero*[6], which is part of the spanish newspaper *El País*[7]. The dataset aggregates contents created by different kind of users, ranging from journalists and expert bloggers to common travellers wanting to share their experiences.

The rest of the paper is organized as follows: Section 2 describes in detail the scenario for which we have developed our application, explaining how the modelling of the provenance information is possible by reusing existing provenance vocabularies. Furthermore, the section explains the functionality of the application and the potential uses for the users. Finally, Section 3 exposes the conclusions of our work and the planned improvements for the application.

## 2. EL VIAJERO PLATFORM

Our scenario is related to the general context of travelling, where travelers want to share, read and reuse experiences in blogs and online news items. The platform on top of which we build our application aggregates content from a variety of newspapers and digital platforms owned by the Prisa Digital Group[8]: "Suplemento El País", "Guías Aguilar", "Canal Viajar" and "Prisa Digital", but it is also open for recommendations from users (with more than 1000 published), pictures (more than 2000 uploaded and visible in the web) and posts (from around 600 different blogs). The activity in the web accounts to more than 590000 unique visitors per month reading or commenting the contents produced, with more than 5000 registered users.

### 2.1 Modelling the scenario as Linked Data

The provenance of the published information has been modelled following a layered approach. On the bottom layer we have the Open Provenance Model (OPM) [3], a domain independent provenance model result of the Provenance Challenge Series[9]. OPM models the resources as artifacts (immutable pieces of state), processes (action or series of actions performed on artifacts), or agents (controllers of processes); and represents their relationships in a provenance graph with five causal edges: Used (a process used some

[1] http://webenemasuno.linkeddata.es/browser_en.html
[2] http://openprovenance.org/
[3] http://www.w3.org/2005/Incubator/prov/wiki/Main_Page
[4] http://www.w3.org/2005/Incubator/prov/wiki/Analysis_of_News_Aggregator_Scenario#Gap_Analysis
[5] http://www.oeg-upm.net/
[6] http://elviajero.elpais.com/
[7] http://www.elpais.com/
[8] http://www.prisa.com/
[9] http://twiki.ipaw.info/bin/view/Challenge/OPM


*I-SEMANTICS Triplification Challenge* 2011 Graz, Austria


**Figure 1: Overview of the Trip Management Linked Data Application.**

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).

Moreover, OPM adds an abstraction layer for all the resources of our domain, gathering guides, images, posts or videos as artifacts, and tracking their provenance without any distinctions. On the second layer we find the OPM profile for our domain, extending the OPM core ontology, OPMO[10], with domain specific relationships (additional types of processes, artifacts, roles and an adaptation of the geometry model proposed in [1]). Finally, on the top layer we have several vocabularies for describing the domain-specific properties of the artifacts in our domain, such as SIOC [11] for the posts and blogs (recording metadata like the date of creation and publication, the author account, etc.), MPEG-7 Ontology [12] for images and videos (bytes, size, color, etc.) and the W3C GEO[13] for adding geolocalization to the resources.

Provenance is attributed to each of the resources through an automatic process, parsing the files provided by Prisa Digital, transforming them to RDF according to OPMO constraints and inserting them into the repository[14]. Thanks to this procedure, new contents can be added each month without a major effort.

It is worth mentioning that we have enhanced the RDF locations of the resources pointing to other resources in the Linked Data cloud. To this end, we relied on the SILK framework[15] for discovering links to either GeoLinkedData[16] or DBpedia[17] resources.

## 2.2 Visualization of Trips, Guides and Blogs

Thanks to the publication of *El Viajero* as Linked Data, users have available more than 6600 resources about almost 1000 different locations around the world. With our application these users can explore, organize and reuse the existing contents of the platform.

The application is divided in two parts. The first one browses the contents from the newspaper and digital platforms, and locates in a map all the available resources to help users to find them. The second part allows to reproduce trips of other users and expert travellers step by step. The viewer is based on a modified version of map4RDF[18] [2], an extensible tool that uses Google's GWT framework to visualize and access to Linked Data resources.

### 2.2.1 *Accessing and filtetring contents around the world*

Many of the available resources belonging to the published dataset refer to specific locations around the world. With the objective of organizing them for the users, our application adds a small marker in the map whenever there is a resource about that location. When users click on one of these markers, a small window appears with additional information about all the resources referring to the location, so users can easily select one to read more. This data is recovered dynamically through a SPARQL query to the endpoint for speeding up the process, and it is available thanks to the publication of the provenance dataset.

We have also included year filtering for the guide retrieval. Users might be interested in recovering only the most recent guides, in order to ensure that what they read is accurate to what they might find when travelling there.

Some of the features displayed in the map for each resource are extracted from its provenance information (e.g. date of creation). However, since the majority of the users are not concerned about the evolution of the resource when planning a trip, we have simplified the interface by adding a link to access this data directly for those who are interested. Clicking on the RDF link leads to a new page pointing to our Pubby frontend[19] where all the aforementioned provenance information can be explored.

[10] http://openprovenance.org/model/opmo
[11] http://rdfs.org/sioc/spec/
[12] http://metadata.net/mpeg7/
[13] http://www.w3.org/2003/01/geo/
[14] http://webenemasuno.linkeddata.es/sparql
[15] http://www4.wiwiss.fu-berlin.de/bizer/silk/
[16] http://geo.linkeddata.es/
[17] http://dbpedia.org
[18] http://oegdev.dia.fi.upm.es/projects/map4rdf/
[19] http://www4.wiwiss.fu-berlin.de/pubby/

Figure 1 shows an overview of this part of the viewer. The "Facet" menu on the left allows to filter points of guides, trips or all resources, and on the map we can see all the current locations of the guides. Each line of the results refers to a different guide, and clicking on top of the title of the guide opens the guide from *El Viajero's* digital newspaper.

### 2.2.2 *Exploring and reusing trips*

The provenance of the resources plays a more important role when browsing and exploring the trips. Trips are resources that contain: itinerary, posts, comments, photos and videos from the user who has uploaded them previously. Moreover, trips are modelled as workflows by using OPM.

Trips can be also seen as interactive blogs, and the provenance information is key to capture their evolution whenever a user modifies them adding a new post, image or comment using other references. By querying the provenance information endpoint, we have created a timeline gathering the different versions of the trip, by reusing the SIMILE widget for GWT[20]. The timeline also contains the materials and comments uploaded by the user and their locations, so we can reproduce the itinerary followed by the traveller in detail and know when did the user do a certain tour or took a certain photo. An example can be seen in Figure 2, where we display the references, photos and posts belonging to the trip "Mi viaje a España". By clicking on each resource, we are able to obtain further information about it in a new window, and its location in the map.

The timeline allows to reproduce the trip to other users and give them ideas and possible points of interest from another perspective besides the guides. Thanks to the visualization on the map, users are able to change to another trip whenever they choose, and explore it in detail.

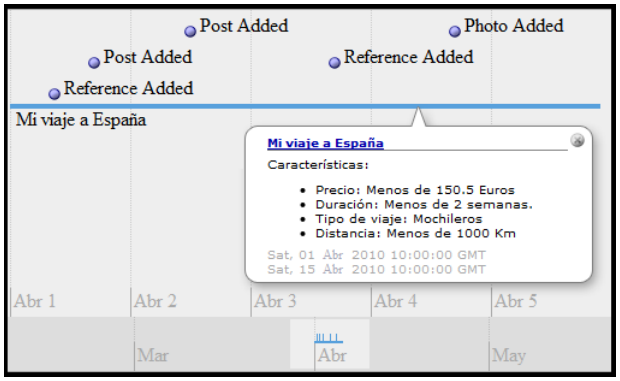


**Figure 2: Exploring the timeline of a trip.**

The application also allows drawing the whole itinerary of the trip, centering the map to be able to see all its points and offering customizable colors, in case the user wants to draw various itineraries at the same time.

## 3. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an application that makes use of several heterogeneous datasets published as Linked Data to provide a visualization of the different available resources from the Prisa Digital Group. We have retrieved and organized the provenance information from the resources referring to them in order to (1) display additional metadata from the resources, and (2) provide a dynamic timeline of the trips to help the browsing of the contents and updates made by their authors. This offers a fast way to read the guides from the places that users aim to visit, instead of doing it manually over thousands of travel guides. These guides are precisely what distinguishes this application from other trip organizers like TripAdvisor[21], because there is a large amount of data available from each destination and it does not rely uniquely in opinions from the users.

In this use case everything is retrieved and organized around the location of the resource, according to the necessities of the users in the domain. Thanks to the published provenance dataset we can retrieve the contents for other user needs through SPARQL queries (e.g. the references used for the development of the resource, the dates, etc.). Thus, additional functionality can be added very easily, due to the flexibility of the application.

At the moment our application can be considered mostly a visualization tool. Linking the data has demonstrated to be useful to retrieve additional metadata from the resources, helping to place them and group them in the map. Work in progress is to let users create and share their own new trips (not only browsing the existing ones), to be able to add ratings for recommendations of the resources by linking to the LUF dataset[22] and to enhance the application with the locations of popular hotels and restaurants often referred to in the guides, which belong to a dataset recently published as Linked Data[23]. Other ideas for leveraging and combining the data are to retrieve further descriptions from the places referred to in the guides querying DBPedia, to use meteoreological data like AEMET[24] to provide statistics about the weather at the locations users are planning to travel and to analyze the integration of the data from other travel sites like TripAdvisor.

## 4. ACKNOWLEDGMENTS

This work has been supported by the R&D project Webn+1. We would like to kindly thank Alexander de León and Miguel Angel García.

## 5. REFERENCES

[1] L. M. V. Blázquez, B. Villazón-Terrazas, V. Saquicela, A. de León, Ó. Corcho, and A. Gómez-Pérez. GeoLinked data and INSPIRE through an application case. In *GIS*, pages 446–449, 2010.

[2] A. De León, V. Saquicela, L. M. Vilches, B. Villazón-Terrazas, and F. Priyatna. Geographical linked data : a spanish use case. In *I-SEMANTICS 6th International Conference on Semantic Systems*, September 2010.

[3] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. Van den Bussche. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, July 2010.

[20] http://www.simile-widgets.org/timeline/

[21] http://www.tripadvisor.es

[22] http://soa4all.isoco.net/luf/about/

[23] http://santillana.linkeddata.es

[24] http://aemet.linkeddata.es