EventMedia Live: Exploring Events Connections in Real-Time to Enhance Content

Houda Khrouf, Vuk Milicić and Raphaël Troncy

EURECOM, Sophia Antipolis, France,
<firstName.lastName@eurecom.fr>

Abstract. An ever increasing amount of event-centric knowledge is spread over multiple social services, either materialized as calendar of events or illustrated by media items shared by people. Crawling data and mining in real-time events connection between these heterogeneous services is a key challenge to enhance events views. In this paper, we present EventMedia, a web-based environment that exploits real-time connections to deliver rich content describing events associated with media, and interlined with the Linked Data cloud. EventMedia exploits semantic web technologies and provides user-friendly interface with the aim to meet the user needs: relive experiences based on media, and support decision making for attending upcoming events. The reader is invited to watch http://eventmedia.eurecom.fr/demo.html before experimenting the live demo.

1 Introduction

In their daily life, people naturally organize their personal data according to occurring events: holiday, wedding, birthday party, concert, etc. With the advent of social web, they have been attracted by several services to create event proposals and to share illustrative media. Web services such as Eventful, Upcoming, LastFm or Flickr host an ever increasing amount of event-centric knowledge maintained by rich online social interactions. The problem is that this knowledge represents a large space of disconnected data fragments providing limited event coverage [2]. In this demonstration, we advocate the use of linked data technologies to connect event-centric information derived from event directories, media platforms and social networks. We consider events as a natural way for referring to any observable occurrence grouping persons, places, times and activities [4]. They represent observable experiences that are often documented by people through different media. As a result, the spatial-temporal dimension, the human participation and the illustrative media are meaningful components to achieve a complete overview of events. We propose to investigate the underlying connections between events and to reconcile the separated data fragments. In addition, the short life time of an event raises a challenge to maintain real-time data crawling and reconciliation, and to ensure a dynamic content enhancement. In this paper, we present EventMedia, a platform that aggregates and interlinks in real-time heterogeneous data sources leveraging on the benefits of Semantic Web technologies.

2 EventMedia Architecture

EventMedia is a hub in the Linked Data cloud since September 2010 [1]. It is obtained from three public event directories (Last.fm, Eventful, Upcoming) and from one large media directory (Flickr). It encapsulates media and events descriptions, enriched with background knowledge from external datasets such as DBpedia, MusicBrainz, BBC and Foursquare. The dataset used the LODE ontology and consists of more than 30 millions RDF triples. All URIs are dereferencable and served as either static RDF files serialized in N3 or as JSON by a RESTful API. The back-end of EventMedia consists of a Virtuoso SPARQL endpoint[1], a RESTful API powered by the ELDA implementation of the Linked Data API[2]. A complete architecture overview is depicted in Fig. 1. In the following, we describe the four central elements of our system: REST-based data crawling, RDF modeling, real-time reconciliation and finally the user interface.

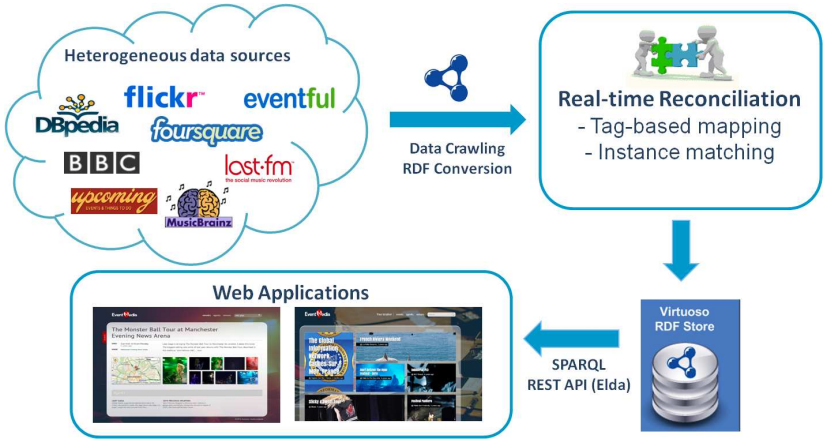

Fig. 1. EventMedia Architecture

2.1 Data Crawling

We crawl data using the various web API of Eventful, Upcoming, LastFm and Flickr. However, their specifications and heterogeneity makes complex and time-consuming the crawling task. A key idea is to leverage the commonalities of these different APIs, so that they could be exported into one unified restful service. Such service should be able to deal with many tasks such as policy management, requests chaining, data integration or merging response schemas. We propose a framework to support these tasks and convert events and media descriptions into the LODE ontology. Our framework is composed of a REST-based module that defines new methods and map them with the targeted web API. For instance, the

[1] http://eventmedia.eurecom.fr/sparql

[2] http://code.google.com/p/linked-data-api/wiki/Specification

method for collecting events takes as input a set of parameters such as source (e.g. eventful, upcoming, etc.), category, location, date and keywords. A user can specify with a single query multiple sources and request in parallel various information. In addition, a second module of our framework deals with data processing starting from JSON de-serialization to RDF conversion and loading into a triple store. More precisely, the data retrieved is de-serialized and exported into a common schema providing descriptions of events, venues, agents, attendees and photos. Finally, a web dashboard has been created to easily handle data crawling. It also provides an interface to reconcile in real-time the data collected and to have detailed statistics about EventMedia. The dashboard is available at `http://eventmedia.eurecom.fr/dashboard`.

2.2 Data Modeling

Once collected, data is converted into RDF triples providing descriptions of events using the LODE ontology and a large SKOS taxonomy of event categories. The descriptions of media use the W3C Ontology for Media Resources. LODE is a minimal model that encapsulates the most useful properties for describing events. The goal of this ontology is to enable interoperable modelling of the "factual" aspects of events, where these can be characterized in terms of the four Ws: What happened, Where did it happen, When did it happen, and Who was involved. The dataset contains a highly diverse set of categories, ranging from large festivals and conferences or exhibitions to small concerts and social gatherings. Figure 2 depicts the metadata attached to the event identified by

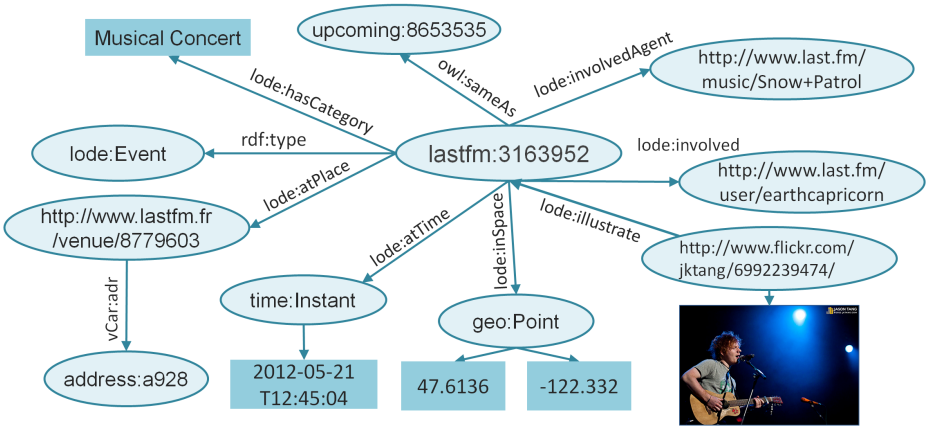

Fig. 2. The *Snow Patrol Concert* described with LODE ontology

`3163952` on Last.fm according to the LODE ontology. More precisely, it indicates that an event of type `Concert` has been given on the `21th of May 2012 at 12:45 PM` in the `The Paramount Theatre` featuring the `Snow Patrol` rock band, and one of attendees is the Last.fm user `earthcapricor`. This event is linked with similar one announced on Upcoming.

2.3 Real-time Reconciliation

At the core of our system is the real-time reconciliation framework that aligns every incoming stream of overlapping but highly heterogeneous data sources. This will sustain a continuous content enhancement, a crucial task to cope with the dynamics of social services. The major gain is to bring context and expand the reach of an event at different stages. In fact, viewing an event page from one event-based service underlines an incomplete content that needs to be further enriched. For instance, we have always detected a real lack of involved agents and their descriptions in the Upcoming directory. While in Last.Fm, people seem to be more responsive to attend events, but only limited to musical concerts without consistent description. We believe that reconciling event-centric data will leverage the benefits of each service and achieve a better overview of an event. Hence, we first explore the connections between events and media using tag-based mapping. Then, we mine meaningful connections between event components, and interlink them with external datasets based on instance matching paradigm. In the following, we explain how we addressed the different challenges raised in four levels of semantic connections.

Events and Media connections. We explore the overlap in metadata between Flickr as a hosting web site for photos, and the event-based services. We interlink every incoming stream of photos with associated events using an explicit relationship materialized by a machine tag such as `lastfm:event=event_id`. Hence, we have been able to convert the description of more than 2 million photos which are indexed by nearly 160.000 events. Other machine tags have also been exploited to establish `owl:sameAs` links with Foursquare and Muscibrainz.

Events connections. We define events similarity as a mutual agreement in terms of their factual properties, namely: title, time, location and involved agents. In order to reconcile events in real-time, we should first select a set of instances of the source dataset by an issuance date. Then, a SPARQL query is performed for each event to fetch similar candidates filtering by some blocking keys. This raises a question about the convenient keys that could minimize the number of retrieved instances. The identification of these keys will obviously depend on the target dataset. In this context, we distinguish two kinds of task: the first task addresses the reconciliation of event-based services, while the second task aligns events with DBpedia. Each task is described separately as it has distinct challenges.

- *Event-based services connections.* Intuitively, the blocking keys could be the spatio-temporal patterns widely used for clustering events. However, the spatial dimension poses accuracy issues in our case since the imprecision of some geo-coordinates in social services confuses the setting of spatial window. Moreover, it is worth noting that discovering similar events from heterogeneous directories is a challenging task owed to some reasons. One reason is the heterogeneity of values caused by multiple representations of the same value (e.g. a title can contain a list of agents' names in one source, and a short name in another one). Another reason is the structural heterogeneity since some instance properties could be empty such as involved agents and description. To

overcome this problem, we define a TFIDF based function that quantifies and ranks the dependency between different properties. Using a ground truth of 300 matched events, TFIDF computation is performed for each pair of events to fetch the common unigrams between each pair of properties. The obtained results show the following high-score order: "$(title_1, title_2), (place_1, place_2), (agent_1, title_2), (description_1, agent_2), (agent_1, agent_2), (description_1, title_2)$". The second step is to identify the blocking keys that will be used in conjunction with the temporal dimension. To do so, we compute how many times an overlap (i.e at least one matched unigram) occurred for each properties dependency. Then, we retain the m-highest dependencies given that the sum of their overlaps frequency is equal to the size of matched events. As a result, we obtain two first dependencies $''\{(title_1, title_2); (place_1, place_2)$ that represent a minimal condition to fetch similar instances. Hence, after removing stop word unigrams (in English language), SPARQL queries are performed to retrieve events in temporal window, and in which the title match each unigram in "title_1" *(resp. for place)*. At the final step, based on the insights gained from the dependency between different properties, a set of string similarity and temporal inclusion metrics [3] are applied to refine the results. Using the ground truth, this approach yields high precision of about 96% and high recall of about 94%. Note that more statistics about our reconciliation results are available at `http://eventmedia.eurecom.fr/dashboard` (statistics tab).

- *Connections with DBpedia.* EventMedia and DBpedia are two datasets that encapsulate descriptions of events, but different in terms of data models and data granularity. EventMedia provides a fine-grained information detailing a spatio-temporal dimensions along with other properties. At the opposite, DBpedia keep a general level of description of well-known events without granular precision about time except for few of them. Regarding this fact, we decide to create `rdfs:seeAlso` links between EventMedia and DBpedia using SPARQL queries and label-based pattern-matching (applied on title).

Agents connections. Mining agents connections plays an important role to bring valuable context such as artists' discography, fine detailed biography and illustrative photos. We reconcile agents derived from event-based services, and with open datasets such as DBpedia, Musicbrainz and BBC. Each unigram of an agent's name is considered as blocking key to fetch similar candidates using SPARQL query. Then, a first string similarity metric applied on labels will discard irrelevant instances. However, the key challenge widely investigated in persons connections is to decide whether the candidate's name refers to the same person. To overcome this challenge, we introduce a contextual filter comparing the description of each agent using an enhanced Cosine similarity with Porter stemming. Hence, if the description is short or missing, we enrich it using surrounding elements such as tags and category of an event.

Venues connections. We finally explore the venues connections of event-based services, DBpedia and Foursquare. The process was straightforward thanks to the light and consistent description of venues. Overall, a string similarities

have been applied on venue properties such as label, address, locality and country.

2.4 User Interface: live your event

One of the challenges we want to address is how to enable fluid faceted navigation of a vast event-based space, and to create harmonious views of the interconnected datasets. Users wish to discover events either through invitations and recommendations, or by filtering available events according to their interests and constraints. We provide mechanisms to browse events by location or a period of time. Once an event is selected, media are presented to convey the event experience, along with background information such as category, agents, venues, attendance list, ticket, etc. A typical example is illustrated in Figure3. Apart from the inspection of the event instance, other conceptual classes (e.g.

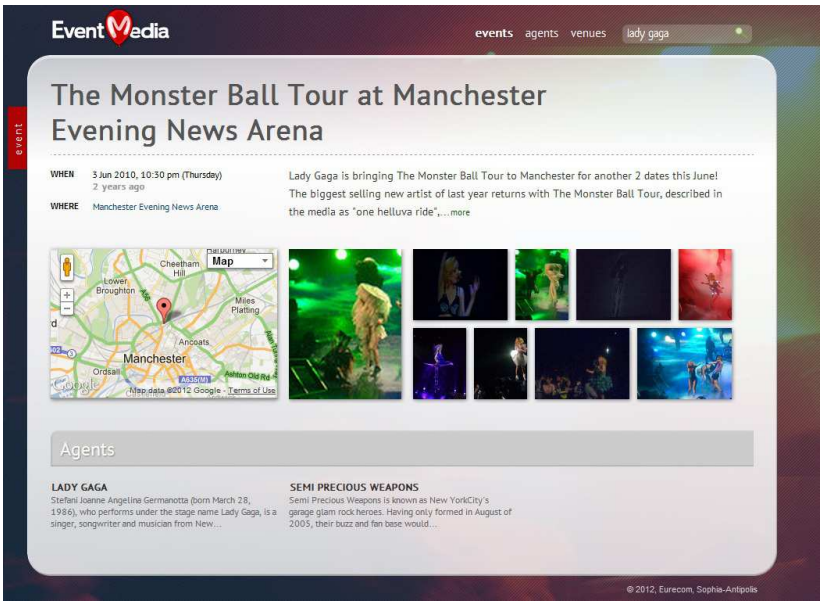

Fig. 3. Interface illustrating a concert of *Lady Gaga* in 2010

venues, agents, users) have also accessible views, so that the user can obtain more information about these instances and explore events related to them. We also leverage data from open datasets to be displayed in infobox separated from the main information, but some parts of this data are interwoven with the main data as well, or used to replace missing data. Finally, we incorporate recommendation mechanism sorting events by popularity. Intuitively, we consider that more attendees an event has, more popular it is. The demonstration of EventMedia is available at `http://eventmedia.eurecom.fr`. From the technical point of view, we have been based on Elda, a java implementation that enables a configurable way to

access RDF data using simple RESTful URLs that are translated into queries to a SPARQL endpoint. It provides a simplified XML and JSON representations of RDF data, suitable for use in the context of JavaScript Frameworks. We used a popular Backbone.js JavaScript framework[3] to facilitate developing the complex user interface. It is a simple but powerful MVC framework, providing Model, Collection, View and Router constructor, together with Event constructor for supporting Pub/Sub pattern. Moroever, it provides an elegant REST integration that make dealing with Elda REST implementation painless.

3 Usage Scenario

To highlight the benefits of EventMedia, we consider the following scenario: Alice wants to see what will happen and who will attend "*the Coldplay Concert given on the 8th of August 2012 at 07:00 PM in Chicago, IL*". A first insight into this concert in the event-based services underlines different descriptions as illustrated in Table 1. By collecting and interlinking this spread information, we attempt to deliver a more homogeneous and complete description. To see the result, a screencast of this scenario is available at `http://eventmedia.eurecom.fr/demo.html`.

	description	time	place	category	ticket	artists	attendees	photos
LastFm[4]		imprecise	✓			✓	✓	✓
Eventful[5]	✓	✓	✓	✓	✓			
Upcoming[6]		✓	wrong	✓			✓	

Table 1. Comparison between descriptions of Coldplay Concert in event-based services

Acknowledgments

The research leading to this paper was partially supported by the European Union's 7th Framework Programme via the projects LinkedTV (GA 287911), ALIAS - Adaptable Ambient Living Assistant (AAL-2009-2-049) and the activity EventMAP (Event-centric Multimedia content Access Platform, TFMC 12116) of the EIT ITC Labs.

References

1. Richard Cyganiak and Anja Jentzsch. Linking Open Data cloud diagram. LOD Community. (`http://lod-cloud.net/`), 2010.
2. A. Fialho, R. Troncy, L. Hardman, C. Saathoff, and A. Scherp. What's on this evening? Designing User Support for Event-based Annotation and Exploration of Media. In *1st International Workshop on EVENTS - Recognising and tracking events on the Web and in real life*, pages 40–54, Athens, Greece, 2010.

[3] `http://backbonejs.org/`

[4] `http://www.last.fm/event/3159427`

[5] `http://eventful.com/E0-001-050047180-4`

[6] `http://upcoming.yahoo.com/event/8634740`

3. H. Khrouf and R. Troncy. EventMedia Live: Reconciliating Events Descriptions in the Web of Data. In 6^{th} *International Workshop on Ontology Matching (OM'11)*, Bonn, Germany, 2011.
4. R. Shaw, R. Troncy, and L. Hardman. LODE: Linking Open Descriptions Of Events. In 4^{th} *Asian Semantic Web Conference (ASWC'09)*, 2009.

Minimal Requirements

- *The application has to be an end-user application.* EventMedia targets anyone who wants to relive past experiences and/or to attend upcoming events.
- *The information sources should be under diverse ownership, heterogeneous and contain real world data.* The data is retrieved from public event and media directories, and from various linked datasets. They are heterogeneous: different models and various representations of the same entities. They provide descriptions of real-world events.
- *The meaning of data has to play a central role semantic web technologies, manipulation/processing, alternative technologies* (i) Data is represented in RDF and described by well-known ontologies. (ii) Data is exploited to discover new semantic connections, to enhance content and to propose new recommendations. (iii) The Semantic modeling enables the discovery of new connections thanks to SPARQL and instance matching. It also enables an efficient semantic search for data visualization.

Additional Desirable Features

- *Attractive and functional Web interface.* EventMedia provides an interactive interface enabling fluid navigation and search capabilities.
- *Scalable application.* Our dataset contains solely more than 40 millions triples and is interlinked with other datasets.
- *Rigorous evaluations.* The reconciliation results have been partly evaluated. The user interface have gone through user-centered design.
- *Novelty.* We propose a new approach to discover in real-time events connections from social services and Linked data.
- *Clear commercial potential.* EventMedia provides a rich and valuable data that can be used to create interesting applications.
- *Contextual information is used for ratings or rankings.* We have use the attendance rate to rank events.
- *Multimedia documents are used in some way.* The event view is media-centered.
- *Use of dynamic data.* We crawl and reconcile in real-time data from social services and linked data.
- *Accurate results.* New enriched event views could be discovered on the interface.
- *Support for multiple languages and accessibility on a range of devices.* The interface has been designed to work also on small screens.