EventMedia: a LOD Dataset of Events Illustrated with Media

Editor(s): X
Solicited review(s): Y
Open review(s): Z

Houda Khrouf [a] and Raphaël Troncy [a]
[a] *Multimedia Communications Department,
Eurecom Institute,
2229, route des Crêtes,
06560 Sophia Antipolis, France
E-mail: {houda.khrouf,raphael.troncy}@eurecom.fr*

Abstract. An ever increasing amount of event-centric generated knowledge is spread over multiple social services, either materialized as calendar of past and upcoming events or illustrated by cross-media items. This opens an opportunity to create an infrastructure unifying event-centered information derived from event directories, media platforms and social networks using the RDF data model. EventMedia aims at creating such an infrastructure that requires seamless aggregation and integration of disparate data sources, some of which overlap in their coverage. In this paper, we present the EventMedia knowledge base composed of events descriptions together with media descriptions associated with these events and interlinked with the larger Linked Open Data cloud. We describe how the data has been extracted, converted, interlinked and published following the best practices of the Semantic Web community.

Keywords: Events, Linked Data, Media, LODE Ontology

1. Introduction

In their daily life, people naturally organize their personal data according to occurring events: holiday, wedding, birthday party, concert, etc. Events are indeed a natural way for referring to any observable occurrence grouping persons, places, times and activities [1]. Events are also observable experiences that are often documented by people through different media. Nowadays, social platforms host very large amount of information about events, illustrative media and the social connections between participants. However, this information is often spread and locked in amongst different services providing limited event coverage and no interoperability of the description [2]. Aggregating these heterogeneous sources of information into one unified platform is the aim of the EventMedia project leveraging on the benefits of Semantic Web technologies.

One of the vision of the Web of Data is to organize and interconnect data silos in a structured way that can be understood by machines, and easily exploited by humans. It requires to use common vocabularies for the integration of fragmentary information into a logically coherent knowledge base. A growing number of RDF datasets have been published in the Linked Data Web covering a multitude of diverse domains such as digital libraries, government, health, media, geography or more generally encyclopedic data. Our goal is to create an event-domain knowledge base, so that we can explore the information with the flexibility and depth afforded by semantic web technologies. Furthermore, we will investigate the underlying connections between events to allow users to discover meaningful or surprising relationships amongst them. We also explore the intrinsic connection between media and experiences so that people can search and browse through content using a familiar event perspective. The EventMedia dataset is obtained from four large public event directories (Last.fm, Eventful, Upcoming and Laynrd) and from large media directories (Flickr and Twitter). The data is retrieved using the respective site API and converted into a structured knowledge base accessible via a REST API and a SPARQL endpoint.

The remainder of this paper is structured as follows. We explain how the data is scraped and aligned (Section 2), and then converted into an RDF model (Section 3). We present an overview of the EventMedia dataset in Section 4, and we describe how we interlinked it with other LOD datasets in Section 5. We showcase two user interfaces in Section 6 and we outline future work in Section 7.

2. Crawling and Aggregating Data

In this section, we describe how data from event and media directories is crawled and interlinked either statically using a REST-based crawler or dynamically using a live extractor.

2.1. REST-based data Crawler

The advent of the Social Web rises a need to provide tools that ensure a seamless and flexible way for crawling data from multiple social services. Such tools should be able to deal with many tasks such as policy management, requests chaining, data integration or merging responses schemas. We propose a framework that support those tasks to crawl data from event related services and to unify information source into a meaningful data model. The framework is illustrated in Figure 1. It is composed of two main components: the Unified REST Module and the Scraping Processor. The first module is based on RESTful service that allows for the unification of various Web APIs exploiting their commonality in terms of described methods, objects and input parameters. Each source API is attached to one JSON serialized file that describes the related server rules such as root URL and API key, and a set of query objects. Each query object describes an API method and the input parameters. To manage the requests chaining, we define two levels of query objects: a global level related to first order methods used for retrieving events, and a second level related to methods used for fetching additional information. Aiming for simplicity, we map some of the API methods to one newly defined method that searches events and photos taking as input a set of arguments such as source (e.g. last.fm, eventful), category, location, date and keywords. A user can specify many sources into one query so that requesting data from many web sites in parallel is possible. The RESTful service is flexible enough so that new methods can conveniently be created, and new similar REST-inspired Web APIs can be integrated by adding their JSON file descriptors.

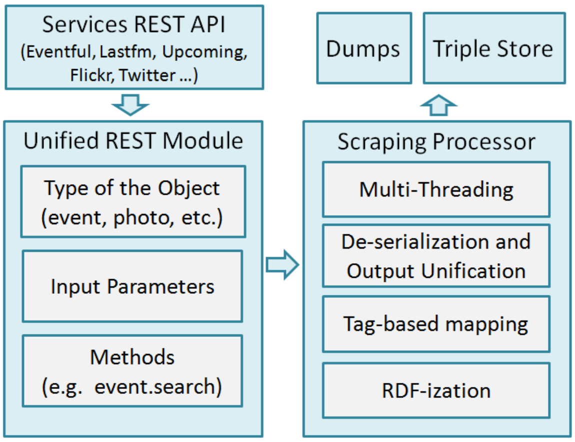
Services REST API
(Eventful, Lastfm, Upcoming, Flickr, Twitter ...)
Dumps
Triple Store
Unified REST Module
Type of the Object (event, photo, etc.)
Input Parameters
Methods (e.g. event.search)
Scraping Processor
Multi-Threading
De-serialization and Output Unification
Tag-based mapping
RDF-ization

Fig. 1. The Rest-based Crawler Architecture

Besides the REST service, the Scraping Processor manages four important tasks. The first task handles a limited number of multi-thread requests, and many sources can be specified into one request. This will contribute to save a tremendous amount of time usually required to fetch information from web services. The other tasks mainly deal 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. Then, we use tag-based mapping consuming some metadata, not only to establish links between events and media resources but also to enrich their descriptions with additional information from external datasets. Finally, to provide an easy handling data collector, we design a user interface composed of two main views: collect and statistics views. The collect mode provides a set of graphical widgets that allows for selecting query arguments. The statistics mode provides additional information about the number of collected resources per source and within a period of time. The web dashboard is available online at http://eventmedia.eurecom.fr/dashboard, but the data collecting functionality is only enabled for administrators.

2.2. Tag-based Mapping

We explore the overlap in metadata between four popular web sites, namely Flickr as a hosting web site for photos and videos, and Last.fm, Eventful and Upcoming as a rich documentation of past and upcoming events. Note that explicit relationships between events and photos exist using machine tags such as `lastfm:event=XXX`. Hence, we have been able to convert the description of more than 1.7 million photos which are indexed by nearly 140.000 events. We further leverage these machine tags to interlink events that share the same photos. There are nearly around 23.000 photos indexing similar events derived from different directories. Some other mappings have also been established to various directories exploiting the machine tags such as `foursquare:venue=XXX` used to link venues descriptions with Foursquare[1], a location based social network, and `musicbrainz:artist=MBID` used to link artists descriptions with MusicBrainz[2], an open music database. Pursuing the same method, we also benefit from the overlap between Twitter and Lanyrd, a social conference directory providing the hashtag associated with each conference. We have been able to convert the descriptions of more than 530.000 tweets which are indexed by nearly 1.167 conferences.

2.3. Live Data Extraction

New events are taking place everyday and people keep sharing an ever-growing amount of media. This requires EventMedia to be a dataset that can dynamically evolve. Providing the latest views that reflect the descriptions of past and upcoming events is highly important due to the meaningfulness of temporality in the event domain. To address this problem, we implemented a live extractor module, which consumes the feeds provided by Flickr[3]. This ensures the retrieval of up-to-date streams of photos containing the tag "`*:event=`" used to fetch additional information from event-related repository. The module converts every fifteen minutes the feeds published by Flickr into RDF and updates the triple store accordingly. On an average week, we observe 1500 new photos and 130 new events are added to EventMedia. Similarly, we also consume the Lanyrd feeds[4] that provides conferences information serialized using a simple data model. We implemented a framework named *Confomaton* capable of aggregating in real-time social media shared by conference attendees and aligns it with event descriptions, as explained in [3].

3. RDF Modelling

In this section, we describe our approach to generate RDF triples of events and media descriptions using the LODE ontology, a variety of media vocabularies and a large SKOS taxonomy of event categories.

3.1. The LODE Ontology

The LODE ontology[5] 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. "Factual" relations within and among events are intended to represent intersubjective "consensus reality" and thus are not necessarily associated with a particular perspective or interpretation. We enhance the LODE descriptions with properties for categorizing events and for relating them to other events through parthood or causal relations using the Descriptions and Situations approach of the Event-Model-F. LODE is not yet another "event" ontology *per se*. It has been designed as an *interlingua* model that solves an interoperability problem by providing a set of axioms expressing mappings between existing event ontologies. Hence, the ontology contains numerous OWL axioms stating classes and properties equivalence between models such as MO [4], CIDOC-CRM [5] and DOLCE to name a few. In addition, LODE can be enhanced with mappings to other vocabularies such as Schema.org or DBpedia. We use the LODE ontology together with properties from FOAF, Dublin Core

[1] https://foursquare.com/

[2] http://musicbrainz.org/

[3] http://api.flickr.com/services/feeds/photos_public.gne?tags=*:event

[4] http://api.lanyrd.com/conferences/

[5] http://linkedevents.org/ontology/

and VCard. Figure 2 depicts the metadata attached to the event identified by `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`. Using the machine tag of related media, an *owl:sameAs* link is discovered between this event and a similar one announced on Upcoming.

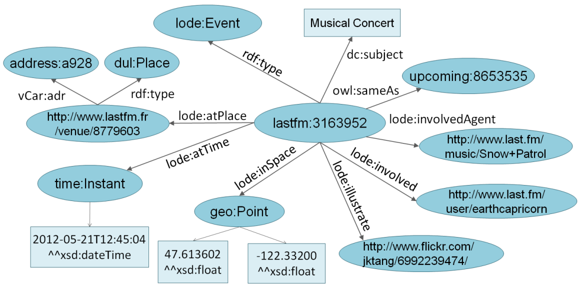

Fig. 2. The *Snow Patrol Concert* described with LODE ontology

3.2. Media Modelling

For describing media, we re-use two popular vocabularies: the W3C Ontology for Media Resources[6] for photos and videos, and SIOC[7] for tweets, status, posts and slides. The Ontology for Media Resource is a core vocabulary which covers basic metadata properties to describe media resources. It also contains a formal set of axioms defining mapping between different metadata formats for multimedia. The SIOC Core Ontology provides the main concepts and properties required to describe information from on-line communities (e.g., message boards, wikis, weblogs). We use those ontologies together with properties from SIOC, FOAF and Dublin Core to convert into RDF the photos, tweets and slides descriptions. The link between the media and the event is realized through the `lode:illustrate` property. Figure3 depicts the description of photos, tweets and slides related to the `ISWC 2011 Conference`.

3.3. Events Taxonomy

Events are generally categorized in lightweight taxonomies that provide facets when browsing event directories. We have manually analyzed the taxonomy used in various sites, namely facebook, eventful, upcoming, zevents, linkedin, eventbrite and ticketmaster, and used card sorting techniques in order to build a rich SKOS thesaurus of event categories. This SKOS thesaurus contains axioms expressing mappings relationships with these taxonomies while the terms are defined in our own namespace (`http://data.linkedevents.org/category/`). The top level categories are Sports, Music, Food, Arts, Movies, Family, Social Gathering, Community and Professional but alignment with other classification such as the IPTC News Codes for sports of the last.fm genres for music is also provided.

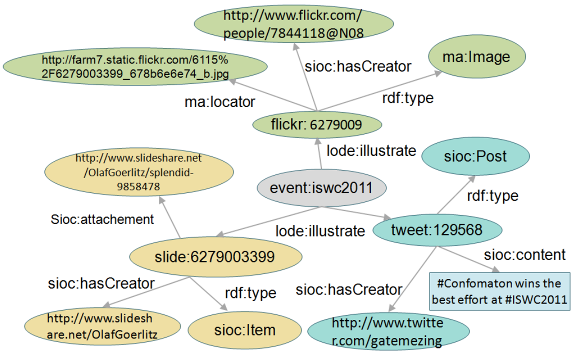
http://www.flickr.com/ people/7844118@N08
http://farm7.static.flickr.com/6115% 2F6279003399_678b6e6e74_b.jpg
sioc:hasCreator
ma:Image
ma:locator
rdf:type
flickr: 6279009
lode:illustrate
sioc:Post
http://www.slideshare.net /OlafGoerlitz/splendid-9858478
event:iswc2011
rdf:type
Sioc:attachement
lode:illustrate
tweet:129568
slide:6279003399
sioc:content
sioc:hasCreator
rdf:type
sioc:hasCreator
#Confomaton wins the best effort at #ISWC2011
http://www.slideshare.net/OlafGoerlitz
sioc:Item
http://www.twitter.com/gatemezing

Fig. 3. RDF Modelling of photos, tweets and slides associated with the ISWC 2011 Conference

4. EventMedia Dataset

EventMedia is a new hub[8] of the Linked Data cloud since the September 2010 snapshot [6]. We use the Last.fm, Eventful, Upcoming and Lanyrd APIs to convert each event description into the LODE ontology. We mint new URIs into our own namespace, for example, for events (`http://data.linkedevents.org/event/`).

The dataset 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 available at (`http://eventmedia.eurecom.fr/sparql`), a RESTful API available at (`http://eventmedia.eurecom.fr/rest/{resource}`) and powered by the ELDA implemen-

[6] `http://www.w3.org/TR/mediaont-10/`
[7] `http://rdfs.org/sioc/spec/`

[8] See also the description in CKAN `http://ckan.net/package/event-media`

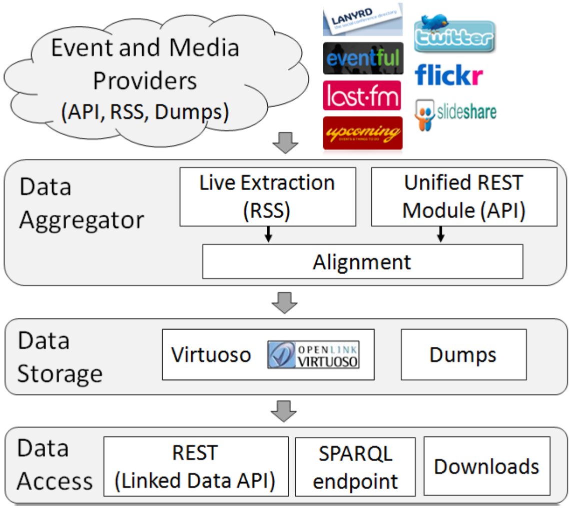

Fig. 4. Overview of the EventMedia components

	Event	Agent	Location	Media
Last.fm	57,258	50,150	16,471	1,425,318
Upcoming	13,114	0	7,330	347,959
Eventful	37,647	6,543	14,576	0
Lanyrd	1,167	0	439	537,091
Total	109,186	56,693	38,3816	2,310,368

Table 1

Number of resources per type and source in EventMedia

tation of the Linked Data API[9]. ELDA provides a configurable way to access RDF data using simple RESTful URLs that are translated into queries to our SPARQL endpoint. The API layer enables associating URIs with processing logic that extract data from the SPARQL endpoint using one or more SPARQL queries and then serialize the results using the format requested by the client. Figure 4 depicts the main components surrounding EventMedia dataset, and Table 1 provides an overview about the number of resources per type and source.

5. Interlinking

Event directories have overlap in their coverage and it is worthwhile to discover similar events so that one description can complement another. However, discovering similar events from these overlapping but separately constituted directories presents some challenges, well-known in instance matching. In addition, we also investigate the enrichment of EventMedia with additional information from open datasets. In our approach, we favour high precision rather than high recall since the cost of missed mapping is lower that the the cost of incorrect matching for the applications we envision.

5.1. Interlinking of Event Directories

We create an *owl:sameAs* link between events that reflect a high similarity in terms of their factual properties, namely: title, date, location and involved agents. It is worth noting that EventMedia is a challenging dataset for an instance matching task, due to the presence of some typographical errors and the high difference of data values detected sometimes between two similar properties. The interlinking is performed using the tools SILK [7], requiring a pre-configured specification by the user, and KnoFuss [8], based on semi-supervised genetic algorithm. We integrate into those tools two newly defined similarity functions, namely: a temporal inclusion metric and a string similarity metric described in [9]. The results obtained highlight the time-sensitivity of events reconciliation since the time expressed in some events is not sufficiently precise. Overall, we obtained high precision of about 95% but fair recall of about 75%. We have generated a linkset of 1103 matched events between Last.fm and Upcoming, and we plan to deeply explore the overlap of the other event directories in the future.

5.2. Enrichment with Linked Data

In order to enrich EventMedia, we perform several interlinking processes using SILK attempting to discover connection between the agents and the locations with Linked Data. For aligning the agents, we selected some relevant datasets namely: Musicbrainz, DBpedia, Freebase and Uberblic. We compare agents' names using Jaro, an efficient metric to match short strings and we keep a high threshold of about 0.96, so maintaining a high precision. Overall, we generated 30.235 links to MusicBrainz, 11.427 links to DBpedia and 14.472 links to Uberblic. Hence, the agent URI which has for label "Radiohead" is interlinked with the DBpedia URI (`http://dbpedia.org/page/Radiohead`) which provides additional information about this band such as its complete discography. Similarly, the datasets being selected to enrich the location descriptions are:

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

DBpedia, Foursquare (RDF provided by Uberblic) and Geonames hosting a large amount of geographical information. The similarity function combines the geographic distance with Jaro applied on labels. In total, we generated 305 links to DBpedia, 3507 links to Foursquare and finally 897 links to Geonames.

6. Event-Based Applications

6.1. EventMedia

EventMedia [10] is a Semantic Web application which uses the REST API of EventMedia to deliver to the user different event-centric views (what, where, when and who) and allows users to relive experiences based on media. We observed that people wish to discover events either through invitations and recommendations, or by filtering available events according to their interests. Therefore, the interface allows constraining different event properties (e.g. time, place, category). Mechanisms for providing the desired support include restricting a time period through a timeline slider control input and a map grouping markers. After an event is selected, all associated information is displayed. Media are presented to convey the event experience, along with social information to provide better decision support. The application is available online at `http://eventmedia.eurecom.fr`.

6.2. Confomaton

Confomaton [3], is a Semantic Web application that aggregates and reconciles real-time information such as microposts, slides, photos, and videos shared on social networks that could potentially be attached to a scientific conference. It also uses the REST API of EventMedia, and provides a live visual summary of the conference, enabling users to re-live the event afterwards and catch up with what they could have missed. The *Confomaton* application is available at `http://eventmedia.eurecom.fr/confomaton`.

7. Conclusion and Future Work

The integration of event-centric information from social services using linked data technologies gives rise to EventMedia, an open dataset continuously synchronized with recent updates. Several improvements could potentially enhance its quality and usability. Indeed, further interesting vocabularies could be incorporated such as the Ticket Ontology to add meaningful relationship between an event and the related ticket, or the Allen's vocabulary to express in fine-grained level the temporal relationships between events. Another enhancement is to enrich EventMedia using other services such as Youtube, Google+ or Facebook, so that we increase the dataset coverage and more connections between these web sites could straightforwardly be explored. In this context, we need to go beyond a simple tag-based mapping and investigate more advanced techniques aligning the different data fragments. Finally, it will be valuable to develop a live interlinking framework that ensures the instance matching of each incoming stream of updates with Linked Data.

References

[1] R. Shaw, R. Troncy, and L. Hardman, "LODE: Linking Open Descriptions Of Events," in 4^{th} *Asian Semantic Web Conference (ASWC'09)*, 2009.

[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 1^{st} *International Workshop on EVENTS - Recognising and tracking events on the Web and in real life*, (Athens, Greece), pp. 40–54, 2010.

[3] H. Khrouf, G. Atemezing, G. Rizzo, R. Troncy, and T. Steiner, "Aggregating Social Media for Enhancing Conference Experience," in 1^{st} *International Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS'12)*, (Dublin, Ireland), 2012.

[4] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, "The Music Ontology," in 8^{th} *International Conference on Music Information Retrieval (ISMIR'07)*, (Vienna, Austria), 2007.

[5] M. Doerr, "The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata," *AI Magazine*, vol. 24, no. 3, pp. 75–92, 2003.

[6] R. Cyganiak and A. Jentzsch, "Linking Open Data cloud diagram." LOD Community. (`http://lod-cloud.net/`), 2010.

[7] A. Jentzsch, R. Isele, and C. Bizer, "Silk - Generating RDF Links while publishing or consuming Linked Data," in 9^{th} *International Semantic Web Conference (ISWC'10)*, (Shanghai, China), 2010.

[8] A. Nikolov, V. Uren, E. Motta, and A. D. Roeck, "Handling Instance Coreferencing in the KnoFuss Architecture," in 1^{st} *International Workshop on Identity and Reference on the Semantic Web (IRSW'08)*, 2008.

[9] 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.

[10] H. Khrouf and R. Troncy, "EventMedia : Visualizing Events and Associated Media," in *Demo Session at the* 10^{th} *International Semantic Web Conference (ISWC'11)*, (Bonn, Germany), 2011.