AAL-2009-2-049, ALIAS
D4.3
Module for knowledge enrichment of event descriptions

Due Date of Deliverable	2011-07-01
Actual Submission Date	2011-09-30
Workpackage:	4.3
Dissemination Level:	Public
Nature:	Report
Approval Status:	Final
Version:	v1.0
Total Number of Pages:	41
Filename:	D4.3-EURECOM-KnowledgeEnrichment-v1.0.pdf
Keyword list:	LODE, EventMedia, instance matching, user interface, user-centered design

Abstract

We propose several extensions for events matching and we conduct experiments on some RDF datasets that confirm the reliability of our new metrics. Hence, we show how we can enrich the knowledge that one can associate to a known event. We present a web-based environment producing and consuming linked data to provide an explicit interlinking of event-related and up-to-date information. We propose interactive and user-friendly interfaces to visualize events with the aim to meet the user needs: relive experiences based on media, and support decision making for attending upcoming events.

History

Version	Date	Reason	RevisedBy
0.1	2010-06-28	created [EURECOM]	Raphaël Troncy
0.2	2011-09-23	updated after alignment experiments [EURECOM]	Raphaël Troncy
1.0	2011-09-30	final version [EURECOM]	Raphaël Troncy

Authors

Partner	Name	Phone / Fax / Email
EURECOM	Raphaël Troncy	Tel: ++33 4 9300 8242 Fax: ++33 4 9300 8200 Email: Raphael.Troncy@eurecom.fr
EURECOM	Houda Khrouf	Email: Houda.Khrouf@eurecom.fr
EURECOM	Ghislain Atemezing	Email: Auguste.Atemezing@eurecom.fr
CWI	André Fialho	Email: atfialho@gmail.com
CWI	Lynda Hardman	Email: lynda.hardman@cwi.nl

Table of Contents

1 Executive Summary

In this deliverable, we first rely on particular automatic alignment tools and we propose several extensions for events matching. We conduct experiments on some RDF datasets that confirm the reliability of our new metrics, and highlight some further developments. Hence, we show how we can enrich the knowledge that one can associate to a known event. Then, we present a web-based environment producing and consuming linked data to provide an explicit interlinking of event-related and up-to-date information. We propose interactive and user-friendly interfaces to visualize events with the aim to meet the user needs: relive experiences based on media, and support decision making for attending upcoming events.

2 Abbreviations and Acronyms

LODE	An ontology for Linking Open Descriptions of Events.
OWL	The Web Ontology Language is a knowledge representation language based on description logics. It has an RDF syntax and in its dialect OWL-Full (one the three flavors of OWL) includes the RDF/S semantics.
RDF	The Resource Description Framework is a knowledge representation language based on a triple model, and serves as foundation for other semantic web languages such as RDFS or OWL.
RDFS	The RDF Schema is a knowledge representation language that has an RDF syntax.
SPARQL	The Semantic Web query language.

3 Introduction

Many event-based web services host a substantial amount of data regarding past and upcoming events. However, the related information are generally all spread and locked in amongst these services providing overall limited event coverage. In a previous work, we generated the EventMedia dataset by gathering and semantifying event descriptions scraped from three public directories namely: Last.fm, Eventful and Upcoming. Since EventMedia has been published, we conduct an analysis on the linked data to unfold event-related information and we advocate the need for an event reconciliation strategy. The problem we address is cast as a data interlinking task, one of the Semantic Web key factors to add value and enhance data reuse.

In this deliverable, we first rely on particular automatic alignment tools and we propose several extensions for events matching. We conduct experiments on some RDF datasets that confirm the reliability of our new metrics, and highlight some further developments (Chapter 4). Then, we present a web-based environment producing and consuming linked data to provide an explicit interlinking of event-related and up-to-date information. We propose interactive and user-friendly interfaces to visualize events with the aim to meet the user needs: relive experiences based on media, and support decision making for attending upcoming events (Chapter 5). Finally, we give our conclusions and outline future work in Chapter 6.

4 Enriching Event Descriptions with Knowledge

Many online services provide functionalities for sharing one's participation and activities at real-world events. Web sites such as Last.fm, Eventful, Upcoming, Facebook and now Foursquare[1] are references to discover and explore event-related information. They cover worldwide events, such as Olympic games, to more specific ones, such as small conferences. Moreover, social media repositories offers a large amount of content captured at these events. However, all this information is often incomplete and always locked into those sites. Aggregating these heterogeneous sources of information by generating and consuming linked data has led to the construction of the EventMedia dataset [24]. Nevertheless, publishing data in compliance with Linked Data guidelines brings forward some challenges. A crucial task is interlinking the dataset with other related datasets, since creating cross-linkage enables a web application to enhance data usefulness and to ensure a higher information coverage. The most reliable way to generate proper `sameAs` links is to label them manually which is time-consuming and unfeasible for large datasets. To cope with the wide heterogeneity of published data, some alignment tools such as SILK [14], SERIMI [2] and RIMOM [27] perform an automatic instance matching with or without a prior knowledge of datasets schemas. But none of these tools is adapted for events reconciliation which need a set of specific metrics in compliance with the factual aspects characterising an event. In this work, we propose to extend one of these tools and we carry out event-oriented data reconciliation trying to cope with misspelled data.

This chapter is organized as follows. Section 4.1 presents an analysis of event-related information in the LOD cloud. Then, we discuss the related work for automatically aligning instances (Section 4.2). In section 4.3, we elaborate our methodology for events reconciliation and we propose two new similarity functions. In Section 4.4, we evaluate our new metrics and we expose our experimental results. Finally, we conclude this chapter by presenting the architecture of EventMedia live (Section 4.5).

4.1 Events in the LOD Cloud

In this section, we first give an overview of the EventMedia dataset, and we present some other LOD datasets relevant for event-related information.

[1] `http://blog.foursquare.com/2011/08/18/foursquare_events/`

4.1.1 EventMedia

Events are a natural way for referring to any observable occurrence grouping persons, places, times and activities that can be described and documented through different media [21]. The large dataset EventMedia [24] is obtained from three large public event directories (last.fm, eventful and upcoming) represented with the LODE ontology and media directories (flickr, youtube) represented with the W3C Ontology for Media Resource.

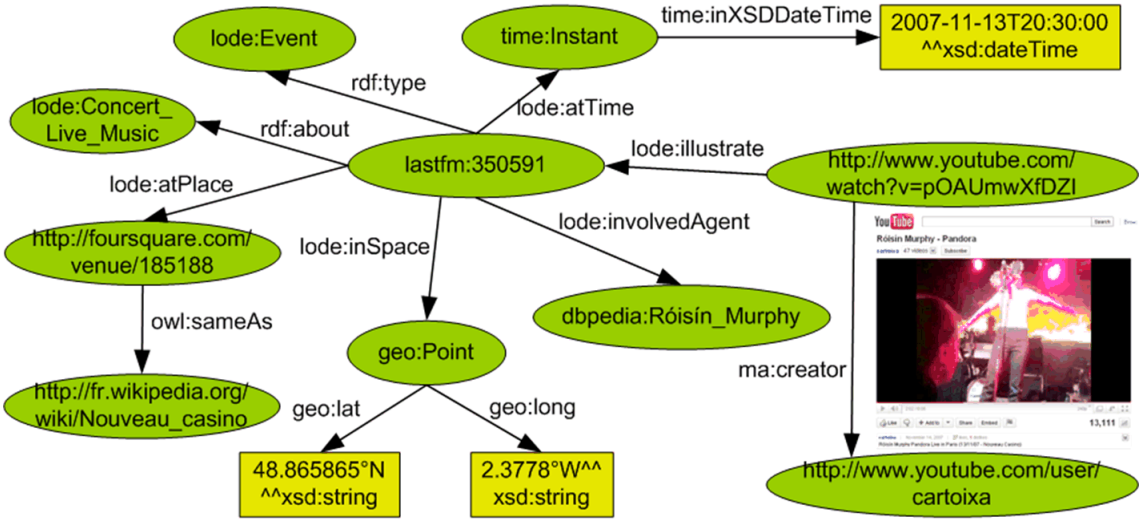

Figure 4.1: *Róisín Murphy at Nouveau Casino Paris* described with LODE

The figure 4.1 depicts the metadata attached to the event identified by `350591` on last.fm according to the LODE ontology. More precisely, it indicates that an event of type Concert has been given on `13th of July 2007 at 20:30 PM` in `Nouveau Casino` Paris featuring the artist `Róisín Murphy`. We have been able to convert the description of more than 1.7 million photos which are indexed by over 108.000 events (Table 4.1). We explored the overlap in metadata between three web sites (flickr, upcoming and last.fm) looking into explicit relationships between events and photos using machine tags.

	Event	Agent	Location	Photos	User
`Last.fm`	37,647	50,151	16,471	1,393,039	18,542
`Upcoming`	13,114	-	7,330	347,959	4,518
`Eventful`	37,647	6,543	14,576	52	12

Table 4.1: Number of event/agent/location and media/user descriptions in the dataset

4.1.2 LOD Cloud Analysis

Event Media is a new hub[2] of the LOD cloud published in September 2010 [6]. It comprises the descriptions of events extracted from Last.fm, Eventful and Upcoming which have at least one associated photo published on Flickr. To analyse event-related information in the LOD cloud, we have executed some generic queries on the LOD cloud cache at (`http://lod.openlinksw.com/`) and on SPARQL endpoints of some selected LOD datasets. We find the class `Event` defined in some vocabularies, namely Yago, DBpedia, Uberblic, Cyc, Umbel, etc. Finally, we have selected some datasets containing information related to agents, locations and events:

- The class `Agent`: Last.fm, Eventful, MusicBrainz, DBpedia, Freebase, Uberblic, New York Times
- The class `Location`: Last.fm, Eventful, Upcoming, DBpedia, Freebase, Foursquare, Geonames
- The class `Event`: Last.fm, Eventful, Upcoming, DBpedia, Freebase, Uberblic

4.2 Related Work

There are some studies which focus on events detection and media illustration such as news video [29] or social documents [3]. Event recognition is a challenging problem aiming towards identifying events from visual or textual cues. Nevertheless, none of these works proposed a methodology to reconciliate events through their entire factual aspects. Some other works leverage the temporal feature to provide a fine-grained chronology of events from news [17] or narration [15]. They build a temporal semantic structure using expressions from natural language to identify the sequence of events. Despite their approach is different from our metric, it seems interesting to merge the two functions in order to tackle the heterogeneity of time representations in LOD datasets. In fact, analyzing the DBpedia dataset, we find some events in which the time is represented by natural language expression such as present, November, etc. As for the hybrid string distance, several studies have proposed hybrid functions and we observe the lack of a real study comparing their performances. One new hybrid similarity function was proposed in [26] where they compute the fuzzy overlap between two token sets. However, they only use the traditional edit distance function as character-based similarity while there are some other variants more efficient such as Jaro and JaroWinkler. Moreover, they do not support the weighting feature of frequent words which significantly improve the similarity results in our experiments.

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

4.3 Events Reconciliation

The instance matching task refers to detect similar real-world entities which can be represented by different schemas in heterogeneous datasets. The challenge is how to define a function that yield a high recall and precision, considering the syntactical and semantic similarity factors. This function mainly relies on string similarity metrics such as Jaro, Leveinshtein, and semantic similarity metrics using an external dictionary (e.g WordNet). In this work, we aim at creating correct `owl:sameAs` links between similar events. An analysis focused on event description in the LOD cloud point out the most shared event properties among datasets such as title, venue and date. We argue that events similarity refers to the high overlap in term of these factual properties. To perform the matching comparisons, we use the Silk Link Discovery Framework, a semi-automatic interlinking tool which has the advantage to support a flexible configuration language. We propose two similarity extensions that better comply with event properties as it will be confirmed by some experimental results.

4.3.1 Silk Link Discovery Framework

The central element of the Silk Link Discovery framework resides in the link discovery engine based on the Silk-Link Specification Language (Silk-LSL) [14]. A key feature of SILK is its declarative language (Silk-LSL) used to define the matching algorithm where a set of operators and similarity metrics can be specified. RDF datasets are accessible via SPARQL endpoint enabling the users to restrict the field of examined resources, for example, by specifying their type. It supports a blocking feature which is used when large datasets have to be compared as it partitions similar data items into clusters. A variety of matching metrics are provided comprising string comparisons, geographical and date similarity, and concept distance in a taxonomy. Multiple comparisons can be combined using a set of aggregation functions such as MAX, MIN, AVG, and transformation functions can be applied on literals such as tokenizing, concatenating and replacing regular expressions. After computing the comparison scores, SILK filters those which are above a specified threshold μ to deduce the correct matching. As data representation and accuracy vary from dataset to another, it is recommended to perform iterative interlinking process in order to discover the best heuristics yielding at the same time high recall and precision.

4.3.2 Similarity Metrics Extension

Initial experiments help us to deduce that the existing similarity metrics are not sufficiently adapted to meet our needs. We therefore propose to add some similarity functions in SILK which is implemented in Scala. We hereafter expose our motivation behind these extensions namely : temporal inclusion and token-wise metrics, and we detail their algo-

rithms.

Temporal Inclusion metric.

Accounting for the LODE ontology, an event has a start date and might have an end date. Intuitively, we consider that two events are similar if they share among others the same time or temporal interval. Hence, matching two events requires a temporal overlap detection. The Date Time metric provided by SILK only measures the difference between two dates which is normalized against a predefined threshold `MaxDays`. Indeed, the more the difference between two dates is inferior than `MaxDays`, the higher is the similarity score. This metric brings confusion on what convenient threshold `MaxDays` must be specified especially when dealing with over thousands of events. Moreover, it is not suitable to infer if two events share a common temporal interval or not. For instance, we manually detect a similarity between two events, coming from two different social media directories, which have a common label `Radiohead en Chile`, but the first event is given between `26th at 07:49:01 PM and 27th at 18:05:01 PM of March 2009`, while the second one is given on `27th March 2009 at 9:00 PM`. We hence propose to introduce a new temporal inclusion metric which detects if a specific date belongs to one temporal interval or if two temporal intervals have an overlap. In the aforementioned example, we observe a difference of nearly 3 hours between the end date of the former event and the start date of the latter one. Thus, a certain number of hours θ has to be tolerated when comparing two dates.

Given two events (e_1, e_2) which have respectively the tuple start date d_i and end date d'_i, (d_1, d'_1) and (d_2, d'_2) where $d_1, d_2 \neq 0$ and d'_1, d'_2 can be null, the temporal inclusion similarity is represented by the following function:

$$s\,(e_1, e_2) = \begin{cases} 1 & \text{if } |d_1 - d_2| \leq \theta \text{ where } (d'_1, d'_2) = 0 \\ 1 & \text{if } d_1 \pm \theta \in [d'_1, d'_2] \text{ where } d_2 = 0 \text{ (idem for } d'_1) \\ 1 & \text{if } min(d'_1, d'_2) - max(d_1, d_2) \geq 0 \text{ where } (d'_1, d'_2) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Token-Wise string similarity.

Comparisons based on string similarity have been widely exploited by many applications such as duplicate detection and data integration. Overall, they are divided into two main classes: character-based and token-based functions, in which the similarity score ranges from 0 to 1. The former class considers a string as a sequence of characters to compute the score. A typical example is the edit distance which is the minimum number of edit operations (i.e insertion, deletion, and substitution) to convert one string to another. One advanced variant which is not based on the edit distance model is the Jaro metric [13], in which the distance depends on the number and the order of common characters be-

tween two strings. For example, the Jaro similarity between "finn brother" and "brothers finn" is 0.6. We note that this score is low despite the high similarity between these strings which have similar words. Indeed, changing the order of tokens affects the sequence of characters which makes the character-based metrics sensitive to the tokens positions. The Jaro metric is mainly used to overcome misspelled data and to match short strings such as personal names [5]. For example, the Jaro distance between "brother" and "brothers" is 0.98. The second class of string distances is the token-based distance which considers a string as a sequence of tokens. They split strings into token sets (i.e. by white-space) called bag of words, and they compute the similarity based on these sets. A typical metric is the Jaccard distance [12] where the score is the ratio of intersection size and the union size of two token sets. For example, the token sets of "finn brother" and "brothers finn" are respectively ("finn", "brother") and ("brothers", "finn"). The Jaccard distance between these two strings is 0.3 which is a low score as the close similarity between "brother" and "brothers" is ignored. Indeed, the token-based functions only consider the exact overlap between the token sets and neglect the approximate tokens.

In this work, we have been interested to study the performance of these string similarity functions over the EventMedia data that consists of user-generated content featuring typos, misspelled and noisy data. We find some similar events wherein the labels have a low similarity in terms of character-based and token-based functions. For example, the Jaccard score between "Anna Plus Kristen Puferfish" and "Pufferfish for Anna and Kristen" is 0.28 while the Jaro score is 0.6. To overcome these limitations, we introduce a novel hybrid metric called Token-Wise combining the two string similarity classes. Our objective is to compute a fuzzy overlap between two token sets. We enhance our method by allowing to attribute a low weight for a list of words (stop-words) frequently used such as the prepositions and, for, in, from, etc. The Token-Wise metric proceeds as follows:

- The strings s and t are firstly split into a set of tokens $s_1....s_k$ and t=$t_1....t_p$
- A list L_{sp} of stop-words and related weight can be specified. These stop-words are in general frequent words which differ from one context to another. For instance, analyzing the events titles in EventMedia help us to find out some frequent words such as festival, concert, music. etc. We can also attribute a specific weight for the remaining words which do not belong to L_{sp}.
- Using one of the character-based functions such as Jaro, Jaro-winkler and Levenshtein, the similarity scores are computed between each token pair from the two sets generating a set of triples $(s_i, t_j, sim'(s_i, t_j))$. We can keep the triples, in which the scores are larger than a given sim' threshold δ. Then, for each token, we keep its correspondent triple which contains the highest score.
- The token-wise score is based on Jaccard similarity pattern which computes the ratio of union size and intersection size of two token sets. Given two strings s, t and

their respectively token sets A and B, the token-wise score adopts the following pattern

$$\textit{pattern (s,t)} = \frac{|A \cap B|}{|A \backslash B| + |B \backslash A| + |A \cap B|} \tag{4.2}$$

Given $sim'(s_i, t_j)$ the score of the based-character similarity between two tokens, ws_i and wt_j the respectively weights of s_i and t_j, N the number of matched tokens (filtered triples), M the number of unmatched tokens where we set sim' = 0, the token-wise score can be written as

$$\textit{token-wise (s,t)} =$$

$$\frac{\sum_{i=1}^{N} sim'(s_i, t_j) \times ws_i \times wt_j}{\sum_{i=1}^{N+M}(1 - sim'(s_i, t_j)) \times (ws_i^2 + wt_j^2) + \sum_{i=1}^{N} sim'(s_i, t_j) \times ws_i \times wt_j} \tag{4.3}$$

To better understand the Token-Wise algorithm, we consider the following two event titles `"Island Treasre Music"` and `"Treasure Island"`. We compute the Jaro similarity between each token pair from their respective sets (`"Island"`, `"Treasre"`, `"Music"`) and (`"Treasure"`, `"Island"`). Then, we only take the highest score obtained for each token provided that this score is above a predefined character-based (sim') threshold δ. In this example, for the sake of simplicity, we set this threshold = 0 and we consider the word `"music"` as a stop-word with the weight w=0.1. Thus, we retain the following triples including the matched tokens and their Jaro scores: (1.0, island, island); (0.95, treasre, treasure). The word `"music"` is still unmatched, so its sim'=0. In this example, we attribute the weight w=1 for the remaining words. The score between s and t is:

$$\textit{token-wise (s,t)} = \frac{1\times1\times1+0.95\times1\times1}{(1-1)\times(1+1)+(1-0.95)\times(1+1)+(1-0)\times0.1+(1+0.95)} = 0.98$$

4.3.3 Alignment Methodology

An event is generally described by a title and occurs at a given place and time. Matching events based only on titles is unreliable since many events share the same title, but happen in different places and times. Likewise, the matching based on both title and place is not efficient since many events are recurrent. Thus, the optimal function must be an average of three comparisons based on: event label, date and venue label. This arises a question about the convenient weight to attribute for each comparison, which requires an iterative matching process. Intuitively, we perform an events matching based on simple average, in which all the weights are equal to 1. We sort the scores and we manually detect some high scores which correspond to wrong alignments. We then modify the weights trying to keep high the scores for correct matching, and decrease those for wrong alignments. Observing

the order of magnitude of the comparison scores, we note that the scores using the token-wise distance mostly range from 0.7 to 1 for similar strings, while the temporal inclusion returns either 0 or 1. Through the first matching, we find some high scores (i.e above 0.7) related to events pair having dissimilar titles and sharing the same temporal interval. In this case, the temporal inclusion score returns 1 which dominates the token-wise score, and increases the final score even if the labels are dissimilar. We hence decide to define a weight equal to 2 for comparing event titles. We propose empirically the following combination and we present our evaluation results in the next section.

$$sim(e_1, e_2) = 2 * \textit{t-wise}(t_1, t_2) + \textit{t-wise}(p_1, p_2) + \textit{TmpInc}([d_1, d'_1], [d_2, d'_2]) \quad (4.4)$$

where e_1 (resp. e_2) has a title t_1 (resp. t_2), venue label p_1 (resp. p_2), a start date d_1 (resp. d_2) and an end date d'_1 (resp. d'_2). The *t-wise* denotes the token-wise metric and the *TmpInc* denotes the temporal inclusion metric.

4.4 Experiments and Evaluation

4.4.1 Experiment Setup

In this section, we evaluate the performance of our proposed method on two types of resources: agent and event. First, we evaluate the token-wise metric using the ground truth provided by existing `owl:sameAs` links between Last.fm and MusicBrainz artists. We select 150 instances where the matched agents have slightly different names, and we compare them using Jaro and the token-wise metrics. Second, we evaluate our event alignment methodology on three event-based directories: Upcoming, Last.fm and Eventful. We build a manual ground truth asking 4 members from our group to evaluate the results. The ground truth contains 39 similar events between Eventful and Upcoming, and 492 similar events between Last.fm and Upcoming. These datasets are loaded into a Virtuoso open-source triple store where the data is accessible via its SPARQL endpoint[3]. The experiments have been executed on a Pentium(R) Dual-Core (2.50GHz) with 4 GB of RAM.

4.4.2 Evaluation Results

The first experiment applied on agents' names aims towards evaluating our hybrid string distance. These names are mostly short strings where the longest one contains only 28 characters. In record linkage literature, the Jaro function seems very efficient to compare short strings [5]. We therefore choose to compare the token-wise distance with Jaro function based on the ground truth of 150 matched instances between Last.fm and MusicBrainz. As parameters setting of token-wise, we define a list of stop-words including the most frequent prepositions such as *the, from, in, to, and, on, be, at*, where we set

[3]EventMedia SPARQL endpoint: `http://semantics.eurecom.fr/sparql`

the weight equal to 0 for these stop-words, and equal to 1 for all the other words. We use the Jaro metric as the character-based similarity and we set its threshold δ equal to 0. We compute the recall and precision of Jaro and Token-Wise matching for different thresholds μ as shown in Table 4.2.

	Jaro		Token-Wise	
μ	Recall(%)	Precision (%)	Recall (%)	Precision(%)
0.95	24	100	60	100
0.9	49	98	82	99
0.8	87	93	96	98
0.7	100	45	100	77

Table 4.2: Precision and Recall for Jaro and Token-Wise similarities

The results highlight the performance of our hybrid metric which generates more similar strings for a given threshold μ. Thanks to the stop-words penalization, many agents have been matched where some names contains an extra word such as `the` or `and`. An additional experiment using the Jaccard metric yield poor results with a recall of 12% and a precision of 100% whatever the threshold. The Figure 4.2 depicts a scatter plot of the

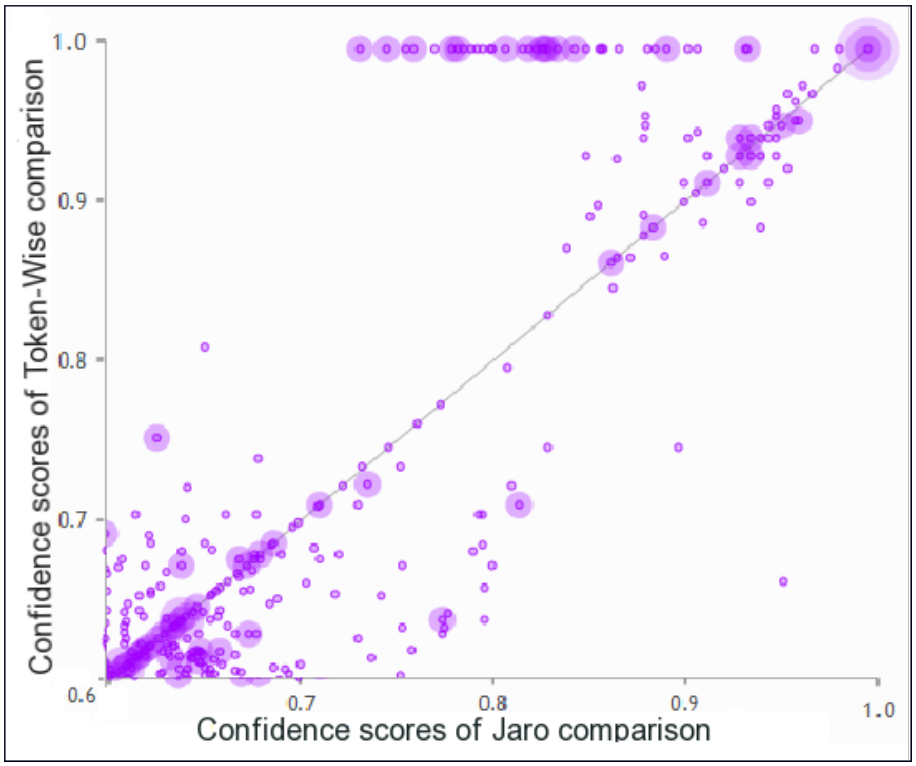

Figure 4.2: Distribution of Jaro and Token-Wise confidence scores

confidence scores ranging from 0.6 to 1 and generated by Jaro and Token-Wise distances applied on 150 agents' names. The points along the diagonal indicate that the confidence scores have not changed between the two matchings. We clearly observe that there are

some scores which are equal to 1 on token-wise axis, while they range from 0.7 to 1 on Jaro axis.

The second experiment aims at evaluating the proposed event alignment approach and the temporal inclusion contribution. We run our matching algorithm to compare events from Eventful and Upcoming, and from LastFm and Upcoming. Our ground truth contains 68 Eventful events compared with 104 Upcoming events, and 583 Last.fm events compared with 533 Upcoming events. We keep the last setting parameters of token-wise distance, and we run two evaluations varying the parameter θ (i.e. the number of tolerated hours) of temporal inclusion metric. The Table 4.3 shows the recall and precision evaluation of event alignment algorithms for different thresholds.

We observe that varying the parameter θ increases significantly the recall especially for high thresholds. This is due to the difference of data accuracy provided by event-based services. For example, we find some events which last two days according to Upcoming, but only one day according to Eventful. Moreover, the events schedules slightly differ from one site to another, where we often find a time span between similar events ranging from 1 to 3 hours. In addition, our results underline a critical point when the threshold is equal to 0.75. We observe a significant decrease of precision when the threshold range from 0.75 to 0.74. In fact, the score 0.75 represents the starting point to look up recurrent events which have similar titles and take place in the same location, but not at the same time. Indeed, our matching function returns 3/4 when there is no temporal overlap but when the event title and venue label are equal. For example, we detect at least 80 recurrent events derived from Eventful and Upcoming. Finally, we distinctly see best results for matching events derived from Last.fm and Upcoming, compared with ones from Eventful and Upcoming. The temporal inconsistency mostly detected in Eventful repository could be the source of this difference. We detect 18197 Eventful events, in which the time is not well defined and represented by `00:00:00`.

	Eventful-Upcoming				LastFm-Upcoming			
	$\theta = 0$		$\theta = 24\ H$		$\theta = 0$		$\theta = 24\ H$	
μ	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
0.8	43	100	71	100	41	100	87	100
0.75	48	100	74	93	43	100	90	100
0.74	74	25	79	26	74	81	94	85
0.6	100	24	100	26	100	75	100	75

Table 4.3: Precision(%) and Recall(%) for events alignment

4.4.3 Discussion

In this chapter, we tackle a data interlinking problem, one of the crucial challenges in the Semantic Web community. A focus on events reconciliation underlines some peculiarities related to such type of resources. We believe that creating high quality `owl:sameAs` links requires a deep study on the nature of examined instances. The obtained results

sustain our alignment methodology where relevant similarity metrics successfully detect most of similar events. We note that the proposed hybrid token-wise distance significantly outperforms the character-based and token-based similarity functions. It has an advantageous weighting feature where attributing a low cost to frequent words lead to improvements. Likewise, the temporal inclusion function successfully quantify the temporal overlap. Throughout our experiments, we have been able to evaluate how the events reconciliation is sensitive to the temporal tolerance expressed by the parameter θ. Overall, we obtained a high precision and recall, and we generated a clustering of recurrent events. Nevertheless, some alignments have been missed, in which events refer to popular carnival or festival taking place at nearby venues and at the same time. We therefore advocate the need for a spatial-inclusion function quantifying the spatial overlap between two geographical areas.

On the other hand, during the evaluation phase, we experienced some performance issues regarding the consumed memory and the elapsed matching time. One of our alignment task attempted to align events derived from Eventful and Last.fm which contains large amount of data. Although the use of blocking feature to reduce the number of comparisons, this task takes around four days to be achieved. After a waiting time, SILK generates an out of memory error during the writing of results in an output file. This highlights a shortcomings of SILK that do not saved results as soon as alignments are detected. Another known problem is the lack of powerful feature to scale since comparing every pair of instances is expensive in terms of time and memory. One promising solution is the use of candidates selection algorithm such as the fast-join method proposed in [26] and based on token-sensitive signature scheme. From a set of tokens, this method generates a set of signatures (i.e set of 2-grams tokens) which undergo a filtering scheme, and then it uses a traditional inverted index mechanism to select the potential candidates. Following this algorithm, we can filter the pair of instances, in which the labels are susceptible to be similar. We perform an additional evaluation to compare the performance of the fast-join method with SILK measuring the matching time. The test is a Jaccard comparison applied on nearly `21.000` agents' names derived from Last.fm and Musicbrainz. We observe that SILK takes 28 minutes, while fast-join takes only 2 minutes which means 14 times faster than SILK. We therefore advocate that introducing such a functionality in SILK will give rise to a powerful tool supporting a declarative configuration language, a variety of transformation and aggregation functions and an efficient scalability method.

4.5 EventMedia Live

Data freshness is considered as one of the key factors of web data quality. Since events and media grow continuously, we argue for the need of an architecture that keeps up-to-date the knowledge base. Inspired by DBpedia Live[4], we develop an architecture that con-

[4] `http://live.dbpedia.org/`

sumes the feeds provided by Flickr[5] as it supports filtering by a list of tags (Figure 4.3). This allows us to retrieve an up-to-date streams of photos that contain the machine tag `*:event=`. We then rdf-ize photos and events metadata description using a .NET module. At this stage, we harness further information from the metadata to interlink EventMedia instances with other datasets, namely: Foursquare and Musicbrainz. On an average week, we observe 1500 new photos and 130 new events which are added to EventMedia.

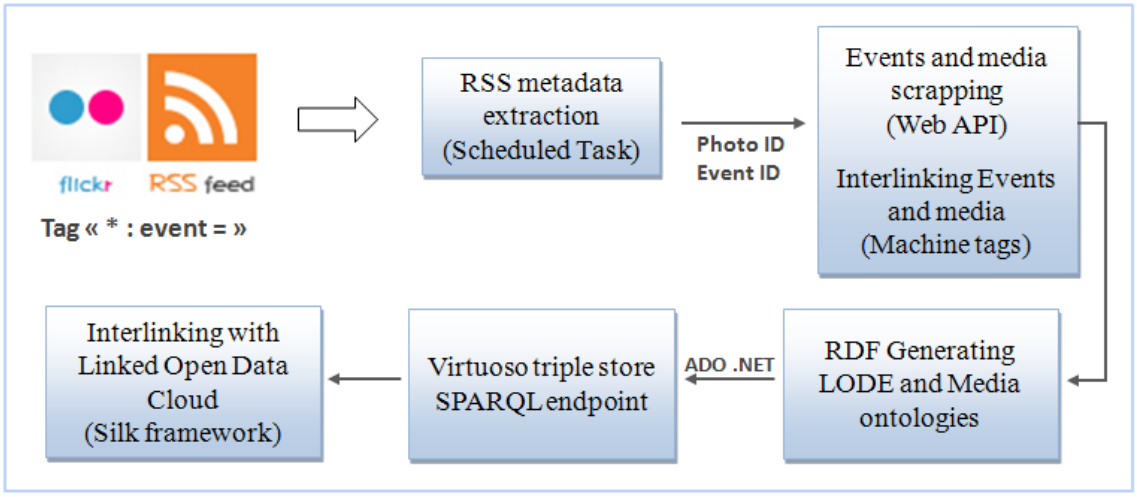

Figure 4.3: Event Media live architecture

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

5 Visualizing Events and Associated Knowledge and Media

As with all developing technologies, it is difficult to identify novel user needs that can be satisfied with emerging semantic web technologies. At the same time, it is difficult to develop the technology in specific directions without knowing what users are likely to want to do with the technology. In previous work we have identified comparison search tasks that can be supported using a combination of thesaurus-based linked data search and a modular user interface design [1], and also historical print annotation tasks [9] that can be supported using a combination of existing RDF data sets, semantic search functionality and task-oriented user interface.

In the context of ALIAS, we are exploring a similar method for designing an application that takes into account the "triple synergy" of users and their social networks, user-created content and metadata attached to this content in an application for supporting users in interacting with events. In the context of ALIAS, we target both the seniors but also their family members wishing to improve the intergenerational communication via interactive and user-friendly interfaces. Events are a natural way for referring to any observable occurrence grouping persons, places, times and activities that can be described [28, 21]. Events are also observable experiences that are often documented by people through different media (e.g. videos and photos). We explore this intrinsic connection between media and experiences so that people can search and browse through content using a familiar event perspective.

While wishing to support such functionality, we are aware that websites already exist that provide interfaces to such functionality, e.g. eventful.com, upcoming.org, last.fm/events, and facebook.com/events to name a few. These services have sometimes overlap in terms of coverage of upcoming events and provide social networks features to support users in sharing and deciding upon attending events. However, the information about the events, the social connections and the representative media are all spread and locked in amongst these services providing limited event coverage and no interoperability of the description. Our goal is to aggregate these heterogeneous sources of information using linked data, 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, entertaining or surprising relationships amongst them. We also use these connections as means of providing information and illustrations about future events, thus enhancing decision support.

The work reported in this chapter uses an explorative user-centered design approach, where users are asked about real-world tasks they would like to carry out, and then asked for their opinions on specific technologies that they are familiar with and how these might be used to support the tasks. This approach ensures making design decisions that con-

tribute towards an efficient, effective and satisfying user experience. Section 5.1 describes the method and the results of this user study, and presents the requirements for an event-based system for discovering and sharing media. We then select precise tasks we wish to support and we carried out a focus group study in order to ask users opinions and observe how they can realize these scenario using specific technologies that they are familiar with (section 5.2). We present our final application and we provide interfaces based on this dataset to illustrate the functionalities supported (Section 5.3).

5.1 User Need Assessment

We follow a user-centered design process consisting of an assessment of user needs and insights, identified through interaction with potential users at different stages of development. Our research starts by identifying who the users are, their interests, their goals and which tasks need to be supported in order to achieve these goals. We collect this information to define a first set of requirements and identify prospective scenarios that illustrate the environment task scope and a first design concept. The steps that follow consist of iterative cycles of re-design and evaluation until a satisfactory design is reached.

5.1.1 Method

The first step of our research was done in order to collect potential end-user experiences, opinions and interests while discovering, attending and sharing events, and user insights about potential web-based technologies that support these activities. We collected this initial input through an exploratory study with 28 participants (11 females). Participants were mostly students and researchers with ages varying from 23 to 47 years old. A similar study was performed by TUM-GSING, Youse and pme on a senior group [11] (Appending D). The study was done through an on-line survey with 8 questions divided into 2 sections. The same topics were then presented in discussion sessions with two groups of master students totalizing 35 additional participants: One discussion was done with students (n=10) from an Interactive Multimedia Systems course and the other with students (n=25) from a Human-computer interaction for the Web course. The results from these discussions were used to validate the survey responses and to extend it with other collected insights.

The first half of the survey aimed at identifying participants' personal experiences and behaviors. It invited them to recall memorable previously attended events (e.g. festivals, conferences, concerts, art galleries, exhibitions, gatherings) and to share their opinions and experiences regarding: (1) how events are discovered; (2) characteristics that support deciding rather or not to attend to an event; (3) how the event experiences are registered and shared; and (4) meaningful, surprising or entertaining relationships amongst events.

On the second part of the survey, participants' were asked to share opinions regarding ex-

isting web technologies in the context of the aforementioned activities. To better address the triple synergy paradigm, we explored the concept of merging event directories, media directories and social networks. With that in mind, we asked participants to share their opinions regarding: (1) the perceived benefits and drawbacks of event directories (e.g. Eventful); (2) enabled possibilities, benefits and drawbacks of combining media sharing websites (e.g. YouTube and Flickr) with event directories; (3) enabled possibilities, benefits and drawbacks of combining social networks (e.g. Facebook and Twitter) with event directories; and (4) suggestions regarding desired and useful features.

Answers obtained from the survey were analyzed through affinity diagramming. The process consists of iterative clustering cycles which allow organizing the collected ideas into common themes, thus allowing to identify the most common opinions for each raised subject. The results from this first exploratory study are described in the following section.

5.1.2 Results

In this section, we present a summary of the results of our user study. The summary contains main reported experiences, interests and opinions around event related activities.

Past experiences. Concerning participants' experiences when discovering events, the vast majority reported to find out about events through invitations and recommendations from friends and colleagues. Traditional media such as posters, flyers, news articles and television ads seem to play a major role when discovering events. Social networks were also reported to be used, with specific reference to event posting and invitation features. More seldom participants use event directories (e.g. Livenation, local city event directories, Ticketmaster, last.fm) or participate in mailing lists, newsletters and forums to obtain updates. The use of search engines was reported, specifically when they knew what to look for. Moreover, participants also rely on previously attended events or venues as reference for finding new events. During the group discussions, participants seemed to rely more heavily on social networks in comparison to the survey responses.

When deciding whether or not to attend to an event, participants seem to prioritize background information. Location was often referred to for orientation and because of distance constraints. Price was commonly mentioned to allow identifying cost-benefit ratios and due to budget constraints. Time of the event was a main decision factor, followed by information about who else would be joining the event. and more specifically, which friends will attend. The content of the event itself (e.g. type, performer, topic) and subjective factors such as fun, relevance, interest, atmosphere, target audience and reputation were also mentioned. Students from the group discussions, preferred the event attendance ("who's joining?") and price constraints over all other characteristics.

Regarding how participants register the experience, they often take pictures for sharing after the event. Less commonly, participants record short videos. As for the how they share

the information, they most commonly talk to others, describing their experience. Participants share the collected media directly (e.g. file transferring, showing on the mobile) or use media directories and social networks such as Facebook, Flickr or YouTube.

Concerning relationships between events, the most referred to characteristics that motivate participants to look into related events were the event categories (e.g. type of event, topic, genre). Another important factor was the event attendees, to find other events they would attend. This could refer to groups of people (i.e. target audience, users with similar interests), but most importantly, individuals in their social networks. Other main event characteristics also mentioned were: location, performers, organizers and time/duration. Lastly, future events from repeated events was also seen as a strong relationship.

Existing technologies. Existing event directories (e.g. eventful) perceived benefit was clearly to be a single access point providing an overview of event information. Another reported benefit is that it supports opportunistic event discovery and facilitates exploration based on different contexts (e.g. location, popularity, categories). Other positive features include: social features (e.g. commenting, sharing events), notification of upcoming events, and shortcuts (e.g. ticket purchase). As for the drawbacks, the main reference was about the unreliability (i.e. unknown source) and incompleteness of information. In particular, low coverage of events and insufficient information for decision support (e.g. lack of location map, videos) have been mentioned. In contrast, the information overload was also seen as a potential drawback making it difficult to find specific events.

When presented the possibility of combining media and event directories, participants recognize benefits due to information enrichment. They claim it would help illustrate events with videos and pictures of past related events (i.e. past performances), other people's experiences, promotional (marketing) material, and so on. The main recognized value would be to give a better idea about the event's environment/atmosphere and provide visual information to support decision making. Participants said it would also support remembering and sharing past experiences. Drawbacks from the merger concern information overload and privacy issues while sharing personal media.

Regarding the possibilities afforded by merging social networks and event directories, some participants think that the main benefits are communication between users, and the sharing of more information (e.g. invitations, opinions, pictures). It was also said to facilitate viewing event attendance, identifying event popularity, and even provide an overview about friends' whereabouts. Live event information (e.g. real-time tweets and comments, live pictures/streams) updates were also seen as a positive afforded feature. Despite the benefits, some participants think the amount of information could clutter the service. Others pointed out that services such as Facebook already provide enough event sharing features. Suggestions included making use of existing social network profiles and/or extending these services.

Other features that users would appreciate having when dealing with events were broadly

described with little overlap. Some of these features are: recommendations (based on past attendance, preferences, and from people with similar interests); better visualizations for exploring and searching events (e.g. map integration); the potential to combine categories and attributes while browsing; obtain more information about events and users (e.g. opinions, price and availability).

Conclusions. The opinions gathered seem to support the development of an environment that merges event directories, social networks and media sharing platforms. Moreover, this information enrichment is thought to provide better means of supporting the decision making process. This assumption is based on the possibility of allowing users to better experience an event by viewing associated media. On the other hand, social information obtained implicitly (behaviors) and explicitly (comments, reviews and ratings) provide better judgments of events in terms of attendance, shared interests and reputation. A common concern about information overload suggests that the interface should avoid cluttering and provide only necessary information. Furthermore, there is a need to support different visualizations and better browsing possibilities depending on user interests and constraints. Lack of event coverage and information completeness is another important identified issue that can be addressed using and combining multiple information sources. These findings are correlated by the user study performed by WP1 which also points out the lack of a central source for getting information about past and upcoming events in the user's vicinity as a problem [11]. These issues, along with other identified user interests were translated into a set of requirements in order to guide the following steps of the environment design and development. The following section describes these requirements in more detail.

5.1.3 User Requirements

Based on this user study, we define a first set of requirements, translating user needs into functionalities that the system should support (Table 5.1.3). It is important to note that the requirements presented here are representative of users who participated in the previous studies. They should be complemented with other non-functional and functional requirements as described in existing design patterns and interface guidelines [8, 19].

5.1.4 Scenarios

Scenarios are informal narrative descriptions that allow exploration and discussion of context, user needs and requirements [4].For the purposes of our research, we created a number of scenarios, each covering a range of the aforementioned requirements and illustrating prospective goals and tasks supported by the system. To better emphasize the context and allow better interpretation and inference of user needs we created four personas. The personas were inspired by the different participants in the previous exploratory

Discovering
Provide a comprehensive coverage of past and upcoming events
Allow searching events based on tags (e.g. performer name, genre, title)
Allow opportunistic discovery by filtering and combining properties (e.g. categories, location, time, price)
Inspecting
Show complete background information about events (e.g. title, location, description, venue, performers, time, category, genre, availability, size)
Allow identifying subjective aspects of events (e.g. popularity, fun, atmosphere, reputation)
Show media associated to events for reliving experiences and for decision support
Show who is joining or joined the event (attendance)
Allow identifying related and repeated events
Visualizing
Rely on traditional media information display (e.g. posters, flyers, ads)
Show only the necessary information in a simple way
Allow different visualizations and browsing contexts (e.g. time, location, people)
Enriching
Allow creating events
Allow associating pictures and videos with existing events
Allow associating comments and opinions with existing events
Sharing
Make use of existing social networks (e.g. Facebook, Twitter)
Allow inviting and recommending events using existing services
Recommending & Preferences
Allow receiving recommendation about events based on personal interests and behaviors
Allow receiving recommendations based on other people's preferences and behaviors (collaborative filtering)
Identify interests and preferences based on past event attendance

Table 5.1: Requirements

study and describe attributes and background information about the actors involved in the scenario. Characteristics are representative of different age groups, professions, preferences, and commonly used event, media and social network sources.

We provide below four different scenarios, each from one different persona.

Scenario 1: *Johnny was invited to a party by a friend and receives a link providing information about this event. He wants to know when and where this event will be and who else was invited. More importantly, he wants to know whether his closest friends confirmed to attend the event or not.*

Scenario 2: *Julie would like to go to a play on her favorite theater. She wants to see a comedy, hopefully playing the upcoming week. She has only been to a few comedies, but she remembers one she specifically enjoyed. Julie would like to see if there is something similar playing and read what other people say about it.*

Scenario 3: *Jack recorded a video with his mobile phone camera while he was attending the Haiti Relief concert from Radiohead given on 24 January 2010 in Los Angeles. He thinks it was a really nice experience and wants to share it on-line. He would also like to see what other pictures and videos were captured during the concert and see how other people experienced the show.*

Scenario 4: *Jessica is going to Paris on her honeymoon and she would like to see what will be happening there during her stay. She wants to do many different things, but cannot decide yet, so she wants to put these things on a "maybe" list in order to decide later. If possible, she would like to see videos of these events to make sure it has a cozy and romantic atmosphere.*

5.2 Scenario-based user study

Based on the results of the exploratory user studies, we identified a set of potential use-cases. However, since participants relied on past experiences, it is debatable whether collected insights are representative of real user behaviors. In order to account for this and to allow a better understanding of behaviors in a well-defined scenario, a second study was performed in order to identify participants' strategies, information sources and behavior patterns while enacting four predefined scenarios:

1. Collect information about an event after receiving an invitation and decide whether or not to attend the event.
2. Use the information about an attended event, as well as other people's opinions (i.e. review, ratings) to identify similar events in the future.
3. Discover and decide about currently occurring events based on what other friends are currently doing (i.e. life streams).

4. Upload and share media for an attended event and explore other people's experiences through available media.

Two sessions were conducted where participants were requested to complete a set of scenarios, performing different tasks in order to achieve the presented goals. Fifteen participants (3 females) took part in the study at two research organizations. The mean age was 26 (range 23-49, SD=9.3). All participants used internet on a daily basis and were acquainted with related on-line services. On-line social network, media directory and event directory usage was measured on a 5pt Likert scale ranging from "never used" to "constantly used". Participants reported that they sometimes used social networks (M=3.3, SD=1.3) and media directories (M=3.1, SD=0.7), and never or rarely used event directories (M=1.3, SD=0.5). During each session of around 1.5 hours, participants role-played the four different scenarios while making use of internet access and a list of links to well-known social networking services, media and event directories. Each participant was requested to self-report his or her experience. After each scenario was completed, participants shared their expectations, strategies, and outcomes of their actions. In addition, collaborative affinity diagramming was conducted by the groups to organize the outputs for each scenario. The affinity diagramming was done using post-it notes with the self-reported actions. During the exercise, participants collaboratively clustered the notes into different action sets on a flip-chart, thus making the strategy and behavior patterns explicit.

5.2.1 Observations

While completing each scenario, participants used on average 5 different information sources (most-used first): search engine (Google), venue/event website, media directories (e.g. YouTube, Flickr, Picassa), social networks (e.g. Facebook, Twitter), event directories (Last.fm, local city event directories).

Seeking information. The majority of the participants started the scenarios by searching for events using a search engine (e.g. Google). Participants used general terms or information provided by the scenario, usually combining title, venue, performer and other information, such as city or time, to constrain the search. Results directed users to event directories or to specific venues/events. Some participants were acquainted with specific events or venues and tried to reach the website directly. Some participants also searched in social networking services, limiting their strategies to finding events through friends, but had little or no information about the specific events presented in the scenarios. Other participants used event directories (e.g. Eventful, Upcoming , Zevents) but were mostly dissatisfied with results from the event directories due to low coverage for the specific scenarios. However, local directories seem to provide better results. In many cases, participants ended up at specific venue/event websites. Regardless of the sources, participants usually performed several subsequent searches, mainly on search engines to obtain

further information (e.g. location on map, images, videos, user comments). While searching for related events, participants used event characteristics (e.g. type, genre, sub-genre, performers) as keywords. Alternatively, few participants relied on related event videos on YouTube or related event artists from Last.fm to identify related events.

Exploring Information. For participants who were redirected to the event website or were able to track the event on the venue website agenda, information about the event in terms of date, description, and performers was readily available as well as some images and videos providing a better illustration about the event. These seem to be the most complete information sources. Some used information from event directories which also provide factual data and few images from the event or the event performers. The directories provide also information about the event attendance as well as comments and reviews which was also appreciated by participants. Social networks such as Facebook were the favorite means for obtaining friends whereabouts, their event opinions and to see it they were attending the investigated event. Another few participants said they would rely on instant messengers or emails in order to contact their friends directly. Social networking services (e.g. Twitter and Facebook) was also said to be the best source of live information about the events. In most cases, however, participants still searched for other related images and videos on a search engine or on media directories (e.g. YouTube) in order to better convey the experience. Participants said that images and specifically videos allowed a better understanding of the event's atmosphere and environment (e.g. party, disco, cozy).

Sharing Information. Most participants chose to share their event experiences through images on media directories such as Picasa and Flickr. The directories were specifically good for posting whole sets of pictures. These directories are also the main source to find more pictures about an event. Some participants also referred to YouTube to post videos about attended events. Many participants pointed out that they would use social networks such as Facebook to post images and videos on their profile or on the Facebook event page. Contrary to image directories, media is selected more carefully when posted on Facebook. The most interesting, representative or funny pictures and videos are posted. Friends can then comment on these media and share their experiences. Few participants preferred to post media directly on their personal web-sites or blogs. Some participants rely more on face-to-face sharing of media where they can point out specific pictures and discuss them. Few others share some pictures directly through emails and MMS messages.

5.2.2 Discussion

After completing all scenarios, clustered actions were used as creative input for semi-structured group discussions. During the discussion, participants addressed their expecta-

tions, strategies, main challenges and recommendations for exploring and sharing events. The results from these discussions are described below.

Information is spread and decentralized. When exploring events, participants reported that there were too many different information sources. They recognized that in order to fulfill the scenarios, there was a need to access several different on-line services. One participant reported *I don't like always having to go from one site to another to find out things about the event*. Therefore, participants agreed it was easier to use a single search engine that has broader coverage of different information sources. One participant reported *There is so much information that it's difficult to prevent the immediate reaction to go to Google*. However, if the participants knew where to find the information, they would go directly to that information source using bookmarks or known website addresses. Specific venue or event web sites were seen as the best source for information overviews. The sources often provided all necessary integrated information, including media. Social networks such as Facebook were also reported to integrate available information (e.g. photos, attendance list, discussions) to some degree, but not sufficiently. Participants suggested integrating social networks with other services. However, participants also agreed that an ideal solution should not just be another information source. There should be some means of centralizing all available information. This can be summarized through one participant's comment: *It would be nicer to have a mash-up with the most important information from each website*.

Information seeking and decision making strategies. While searching for interesting events, there was no agreement on the search strategy. However, most participants reported that the most important information was location, time (date), type of event and popularity. A common strategy was to start constraining the factual properties of the event (e.g. type, location and date) in order to filter the available information. Participants showed an interest in specifying these constraints by defining ranges (e.g. max price). Searching by title (if known) and other information were amongst other options. Other few participants suggested alternative methods for identifying events, such as mobile location-based services to track nearby events. Participants also agreed that the strategies were dependent on type of event (e.g. concerts, parties, art exhibitions). Furthermore, social aspects such as people and friends who are attending could have priority over any other available information. Therefore, starting to search by friends was another potential starting point. One participant commented *If your close friends like it, it's more likely that you will enjoy it*.

Relationships and recommendations. Similarities among events that were said to be interesting include: location, date, event type and genre. However, participants agreed that the most valuable relationship was based on the common interests of people who attend the event. Recommendations were also seen as a potential feature, where users could receive interesting events based on popularity and ratings, people with similar in-

terests/behaviors, friends' attendance, or on the user's past attendance, by keeping a user history.

Participants agreed that the presentation of information about events should be sorted (e.g. time, relevance, popularity) according to their needs (customizable). While displaying large numbers of related media, these could be clustered depending on the event type, media owner (e.g. friends) or visual similarity. Another suggested option was to show only the most popular media or filter the media that belong to known friends. Participants indicated that the most important information is factual information (e.g. what-where-when) and that any other relevant information should be one click away. Most participants argued they would like to have an instant overview of the event through associated media. Participants also agreed that while conducting the scenarios, associated media was the best way to easily illustrate the whole experience. One participant commented *You can have an idea about how the event looks like and what kinds of people go there. It's kind of like a preview*. Excerpts of songs, for a musical event (audio or video), would also be highly appreciated. Additional information such as ratings for events, weather conditions, distance from current location, travel information (public transportation) and accessibility are also said to be very useful. Being able to see user attendance was mentioned. Additionally, user profile pictures can convey a better idea about the "type of people" or target audience of the event.

5.2.3 Conclusions

We are aware of a number of limitations with this study. The events selected for the scenarios may not have corresponded with users' interests. We feel, however, that our results are sufficiently valid for guiding the development of the functionality of the application. Our results from the studies suggest the need for services that combine information from different event directories, social networks and media sharing platforms. Since information is spread and locked in different services, users express the need for a single resource to explore to experience events. Although this benefit is recognized for existing event directories, their lack of event coverage and information completeness affects the user experience.

Users also pointed out the benefits of merging different information sources. Social factors are a strong component when identifying, deciding and sharing experiences about events. Participants rely on other people not only to receive invitations and recommendations about events, but also to decide whether or not to attend. Social information obtained by attending an event (e.g. "who else is going?") or by sharing experiences (comments, reviews and ratings) provide valuable support for the decision making process. Shared interests amongst users is also the main identified means for obtaining event recommendations. "A picture is worth a thousand words" and associated media were also described as the perfect means for representing events. Images and video provide a powerful means for identifying several event characteristics, to convey the experience and to provide de-

cision support. While participants request more information about the events, there is a common concern about information overload. This issue suggests that interfaces should avoid cluttered information and provide only timely and necessary information. Furthermore, there is a need to support different visualizations and improved browsing options that depend on user interests and constraints.

5.3 End-User Application

In the Deliverables D4.1 [23] and D4.2 [16], we have presented the EventMedia dataset, a new hub in the linked open data cloud [6] that share entities descriptions with many other datasets. In this section, we present the end-user application that makes use of this dataset that we have developed following the user centered design approach described above. We first show the original sketches of the application (Section 5.3.1) before describing the back-end (Section 5.3.2) and the front-end interfaces (Section 5.3.3).

5.3.1 Sketching The User Interface

The interfaces are represented through low-fidelity prototypes. The prototypes allow exploring, refining and validating prospective concepts along with interface and interaction aspects through small studies with potential end-users and usability experts. Unsurprisingly, the sketches below correspond to the basic properties defined in the LODE ontology.

Views and Perspectives.

What - One prospective view is media centered and allows to quickly illustrate the event through associated media. In this view we display events through a representative images and convey different event characteristics (e.g. relevance, rating, popularity, etc) with one and/or more of the image properties, i.e., size and transparency. This approach has been used in other applications[1] to represent clustered result sets or convey sorting by size on different contexts (Figure 5.1a).

When - Ordering can also be used to represent chronological event occurrence. In fact, the time centric view can be interpreted as the sorting of events chronologically (Figure 5.1b).

Where - A location centric view can be used to represent where the events occur geographically to orient the user and convey distance. The use of maps is commonly used to visualize such information (Figure 5.1c).

Who - Events are intrinsically bound to a social component. Users want to know who will

[1] See for example `http://www.jinni.com` or `http://www.ted.com/talks`.

be attending to an event when deciding to attend to it. In this context, a people centric view would be relevant to explore the relationships between users and events. Alternatively we can combine attendance information to other views such as location, allowing users to browse for friends on a map and identify their attended events. It could also be used to provide means of visualizing event popularity ,e.g. identify the cities hot-spots on a map, indicate visual cues of popularity according to number of attendees.

In order to allow users to relive experiences from events attended in the past, follow future confirmed events, and keep track of authored events, it is necessary to display events in the context of the users own attendance and ownership. For this reason we will support a "my events" feature with overall browsing possibilities. If several views are to be supported one challenge that can arise concerns transitions between these different views. This is specifically important for facet browsing, due to sudden disappearance of items during navigation [19]. Animated transitions could be used in order to allow the users to maintain orientation during such navigational changes.

Search Interface

When discovering events we believe users will also rely on browsing, which allow them to analyze large sets of event sets, and narrow them according to their interests and constraints. Overall, we believe users will have different information or browsing/search needs as follows:

- Navigational - when the intention is to reach a particular known event;
- Contextual Browsing - discover one or more events given a specific context (e.g. by location, performer, type, time);
- Entertainment Browsing - serendipitous and opportunistic discovery of events;

Since it is often easier to recognize a word or name than it is to think up that term, it is useful to prompt users with information related to their needs. Based on that principle, we will explore the use of hierarchical faceted metadata which will allow users to browse through multiple categories, each corresponding to different dimensions of the collection [8]. As a general guideline and given users' request, we will avoid empty results during search. Faceted browsing can avoid empty results by restricting the available filtering options in the given focus to only those which lead to non-empty results (poka-yoke principle) [19]. Consequently, the user is visually guided through an interactive query refinement process, while visualizing the number of results in different categories. Additionally, we will explore the information afforded by linked data to display results which are closely related to the user interests. For example, if during a search, no Jazz concerts are available in Amsterdam, we will show other events from nearby cities, time period or even other type of events closely related to Jazz music.

Figure 5.1: Interface views illustrating a set of events under: (A) media centric perspective; (B) a chronological perspective; and (C) location centric perspective

While trying to reach a specific event, traditional keyword search can be done through entry forms. Dynamic term suggestions or auto-completion can be used to provide rapid and effective user feedback by suggesting a list matching terms as the user types the message. Semantic auto-completion extends this method by providing means of clustering the terms according to different categories or facets [9]. Keyword search can also be integrated to faceted browsing and extend the defined classification options. In this context, it is important to indicate if the search will act as a keyword filter or if it will match the classification terms [19]. In regards to event attributes at initial search constraint definitions, time, place and event type seem to be the core indispensable inputs. A potential solution is to always display these attributes during the whole searching/browsing process to enable zooming-in and out from a search result set at any point. Since time period is a range variable input, a common solution is to use a timeline slider control input [19].

Event Representation

When representing an event instance, we show all information needed to support the decision making process (e.g. Figure 5.2). Since experiences are centered around media content, we wish to explore different media that better illustrate the event to end-users. Some information that can support decision making are the following.

- background information (e.g. performers, topic, genre, price, attendance list, etc)

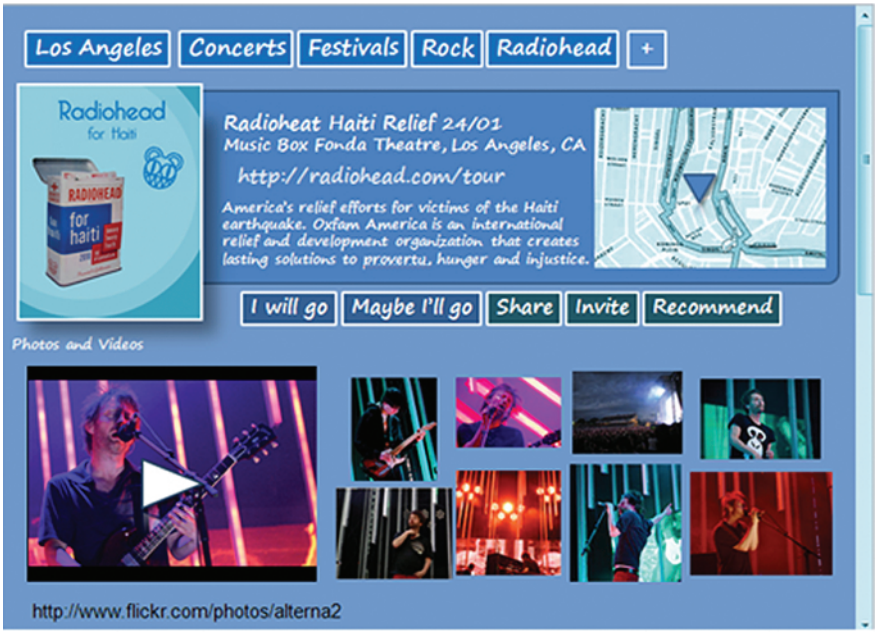

Figure 5.2: Interface illustrating an event instance view for a Radiohead concert

- subjective or computed attributes (e.g. reputation, fun, atmosphere, audience)
- user opinions, comments and ratings (strangers and friends)
- representative media (ads, media from past related events, media from the audience, etc.)

Apart from the inspection of the event instance, other conceptual classes (e.g. users, venues, performers, media) should also have accessible views, so that the user can obtain more information about these instances and explore events related to them. In future work we will also identify what are the relevant associated information and how to represent navigation from and to these nodes.

Enriching Information

Regarding event content enrichment, interfaces that allow users to add/upload information and assign such information to events will be investigated and explored in future studies.

One of the required enrichment features refers to assigning user attendance and keeping track of the users' previously attended events. This information can be used so that the user can easily access past experiences. Moreover, attended events may be used to identify user interests for recommendation and personalization of the facet-pears during search [19]. In order to keep track of events, we will give options to allow users to say if they were in a past event (e.g. *I was there*) or if they are attending to an upcoming event

(e.g. *I will go*). Another prospective option is to allow the user to select events that he is unsure if he will attend (e.g. *I might go*). This will allow adding multiple events to a "maybe" list for future decisions or even comparisons.

Finally, since users are likely to revisit information they have viewed in the past [8], we will also support simple history mechanisms, by saving a list of recently viewed events. History mechanisms can also be incorporated into the facet search to allow users to undo query filtering and return to a specific query set.

5.3.2 Back-end Architecture

The back-end of the system consists of a Virtuoso SPARQL endpoint, a RESTful API powered by the ELDA implementation of the Linked Data API, and a web server. All URIs minted in the dataset are dereferencable and are served as either static RDF files serialized in N3 or as JSON by the RESTful API. We implemented content negotiation in order to let clients decide about the desired representation. Clients requesting a JSON representation are redirected to the RESTful API, which is implemented using the Virtuoso configuration. Besides serving JSON representations of resources available in the dataset, the RESTful API also provides convenience methods exposed as additional resources, which are not explicitly represented in the dataset. Examples for such functionalities include search over the dataset using different parameters such as keyword, time, location. We also allow dataset updates, for instance, by being able to specify attendance information or link additional media to existing events from the front-end.

The Linked Data API[2] provides a configurable way to access RDF data using simple RESTful URLs that are translated into queries to our SPARQL endpoint. More precisely, we use the Elda[3] implementation developed by Epimorphics. Elda comes with some prebuilt samples and documentation which allow to build specification to leverage the connection between the back-end (data in the triple store) and the front-end (visualizations for the user). The API layer helps to associate 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. A URI is used to identify a single resource whose properties are to be retrieved or to identify a set of resources, either through structure of the URI or through query parameters. Listing 5.1 shows an example of the configuration file of the EventMedia API specifying the event, media and tweet viewers followed by the events and media properties access.

```
# define some properties of the endpoint
<#MyAPI> a api:API ;
  rdfs:label "EventMedia API"@en ;
  api:maxPageSize "1000";
  api:defaultPageSize "10" ;
  api:endpoint <#event>,<#media>,<#tweet>,<#agent>, <#venue>,<#user>,<#eventbyid>,<#mediabyid>,<#
       tweetbyid>,<#agentbyid>,<#venuebyid>,<#userbyid>;
  api:sparqlEndpoint <http://semantics.eurecom.fr/sparql> ;
```

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

[3] http://code.google.com/p/elda

```
  api:defaultViewer api:describeViewer
  .
# specification of the event viewer (all properties appear in the json file)
spec:eventViewer a api:Viewer ;
  api:name "ev";
  api:property "title","description","space.lat","space.lon","time.datetime","inagent.label",...
spec:mediaViewer a api:Viewer;
  api:name "mv";
  api:property "title","description","url","thumbnail.url", "user","phototime","illustrate","keyword
       ", "sameas","creator.label".
spec:tweetViewer a api:Viewer;
  api:name "tv";
  api:property "sameas","hashtag","content","created","user","creator.label","illustrate","creator.
       avatar".

<#eventbyid> a api:ItemEndpoint;
    api:uriTemplate "/event/{id}";
    api:itemTemplate "http://data.linkedevents.org/event/{id}";
    api:defaultViewer spec:eventViewer.
<#mediabyid> a api:ItemEndpoint;
    api:uriTemplate "/media/{id}";
    api:itemTemplate "http://data.linkedevents.org/media/{id}";
    api:defaultViewer spec:mediaViewer.
```

Listing 5.1: Example configuration file of the EventMedia API, specifying event, media, and tweet viewers followed by events and media properties access.

5.3.3 Final User Interface

Users wish to discover events either through invitations and recommendations, or by filtering available events according to their interests and constraints. Therefore, the interface allows constraining different event properties (e.g. time, place, category). Mechanisms for providing this desired support include restricting a time period through a timeline slider control input and a map grouping markers (Figure 5.3).

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. Media have different sizes to convey their popularity in terms of views count. We explore these different views according to the basic event properties defined in the LODE ontology (what, where, when and who). The demonstration is available at `http://semantics.eurecom.fr/eventmedia/`.

The Silverlight PivotViewer[4] control offers an easy interactive visualization of large amount of data. It supports a fast and fluid navigation where the gallery objects can be filtered and sorted, and it leverages a Deep Zoom technology enabling a smooth zoom in/out on images. We have created an EventMedia staged pivot collection where the photos and the events can be filtered using simple keywords and some properties such as the date, attendance, geo-coordinates, city and country. The Zoom in enables the user to focus on one photo where its correspondent panel shows some details such as the related event, the involved agents, the venue, etc. The figure 5.4 depicts two snapshots of EventMedia browsing using PivotViewer technology. The demonstration is available at `http://semantics.eurecom.fr/eventmedia/pivot.html`.

[4] `http://www.microsoft.com/silverlight/pivotviewer/`

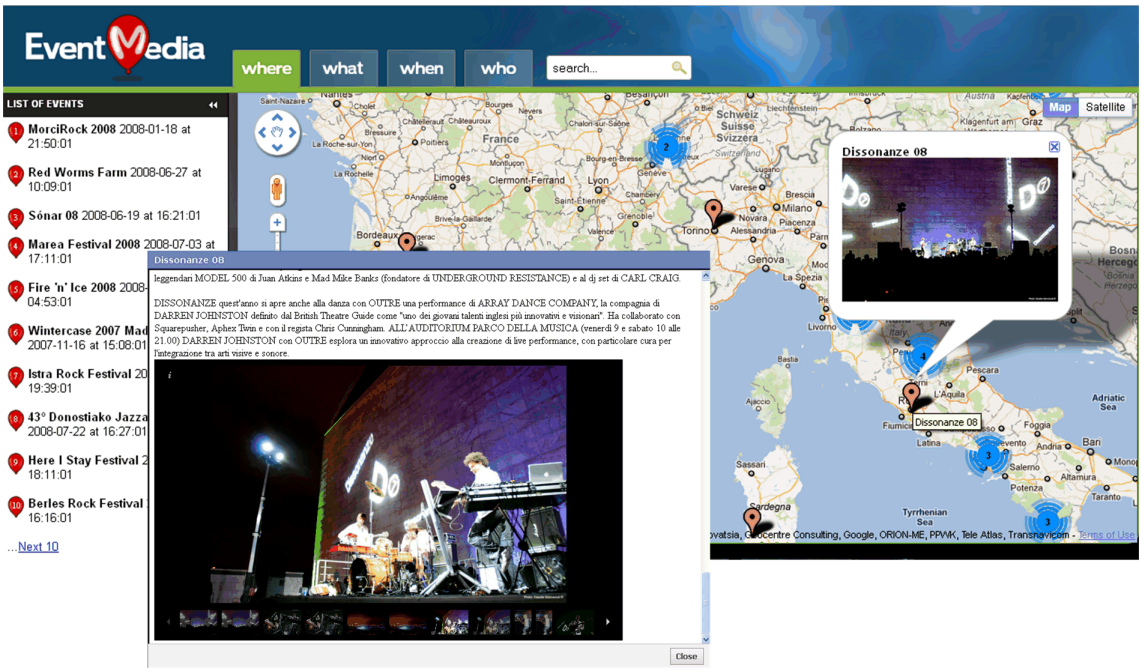

Figure 5.3: Interface illustrating the festival *Dissonanze* in 2008

Figure 5.4: EventMedia browsing using a PivotViewer control

6 Conclusion

The ever increasing amount of RDF datasets published in the Linked Data Cloud brings forward a high demand of data reconciliation. Identifying real-world similar instances is a challenging task accounting for the various descriptions and object peculiarities. In this deliverable, we deal with events reconciliation that requires specific methods to comply with their nature. We proposed a powerful string similarity distance by combining character-based and token-based distances with the aim to better overcome the variety of titles given to similar events. We also proposed a temporal inclusion metric to detect the temporal overlap between RDF instances. The evaluation applied on three different datasets led high recall and precision, consolidating the efficiency of our alignment methodology.

We also described two user studies, where users were asked about real-world tasks they would like to carry out. We observe and identify participants' strategies, information sources and behavior patterns while enacting predefined scenarios. We used and consumed linked data technologies for integrating information contained in event and media directories. Finally, we present the architecture and the user interfaces of the application available at `http://semantics.eurecom.fr/eventmedia/`.

For the future work, we plan to integrate a spatial inclusion metric in our methodology with a view to enhance events reconciliation. Furthermore, we will further develop EventMedia live, our real-time dataset instance where events and media are continuously updated based on RSS feeds of social media sites. We are interested to build an event-oriented live interlinking framework, in which each incoming stream of instances will be automatically interlinked with LOD datasets.

Bibliography

[1] A. K. Amin, M. Hildebrand, J. van Ossenbruggen, and L. Hardman. Designing A Thesaurus-Based Comparison Search Interface For Linked Cultural Heritage Sources. In *15th International Conference on Intelligent User Interfaces (IUI'10)*, pages 249–258, Hong Kong, China, 2010.

[2] S. Araujo, J. Hidders, D. Schwabe, and A. de Vries. SERIMI: Resource Description Similarity, RDF Instance Matching and Interlinking. Technical report, Delft University of Technology , July 2011.

[3] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*, pages 291–300, 2010.

[4] J. M. Carroll. *Making use: Scenario-based design of human-computer interactions*. MIT Press, 2000.

[5] W. Cohen, P. Ravikumar, and S. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *1st International Workshop on Information Integration on the Web (IIWeb'03)*, pages 73–78, Acapulco, Mexico, 2003.

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

[7] 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 (EVENTS'10)*, pages 40–54, Athens, Greece, 2010.

[8] M. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[9] M. Hildebrand, J. van Ossenbruggen, L. Hardman, and G. Jacobs. Supporting Subject Matter Annotation Using Heterogeneous Thesauri, A User Study In Web Data Reuse. *International Journal of Human-Computer Studies*, 67(10):888–903, 2009.

[10] J. Hobbs and F. Pan. Time Ontology in OWL. W3C Working Draft, 2006. http://www.w3.org/TR/owl-time.

[11] S. Ihsen, K. Scheibl, S. Niedermeier, S. Glende, F. Kohl, and P. Dinckelacker. Requirements list-needs and preferences of user groups. ALIAS, Deliverable 1.1, 2011.

[12] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[13] M. A. Jaro. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

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

[15] K. Lee. Processing and Representing Temporally Sequential Events. In *18th Pacific Asia Conference on Language, Information and Computation*, Tokyo, Japan, 2004.

[16] X. Liu, R. Troncy, and B. Huet. Module for cross-media linking of personal events to web content (v1). ALIAS, Deliverable 4.2, 2011.

[17] I. Mani and D. G. Wilson. Robust Temporal Processing of News. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, pages 69–76, Hong Kong, China, 2000.

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

[19] G. M. Sacco and Y. Tzitzikas. *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*, volume 25 of *The Information Retrieval Series*. Springer, 2009.

[20] S. Schenk, C. Saathoff, S. Staab, and A. Scherp. SemaPlorer - Interactive Semantic Exploration of Data and Media based on a Federated Cloud Infrastructure. *Journal of Web Semantics*, 7(4):298–304, 2009.

[21] R. Shaw, R. Troncy, and L. Hardman. LODE: Linking Open Descriptions Of Events. In *4th Asian Semantic Web Conference (ASWC'09)*, 2009.

[22] R. Troncy, A. Fialho, L. Hardman, and C. Saathoff. Experiencing Events through User-Generated Media. In *1st International Workshop on Consuming Linked Data (COLD'10)*, Shanghai, China, 2010.

[23] R. Troncy, H. Khrouf, R. Shaw, and L. Hardman. Specification of an event model for representing personal events. ALIAS, Deliverable 4.1, 2011.

[24] R. Troncy, B. Malocha, and A. Fialho. Linking Events with Media. In *6th International Conference on Semantic Systems (I-SEMANTICS'10)*, Graz, Austria, 2010.

[25] W. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). In *1st ACM International Workshop on Events in Multimedia (EiMM'09)*, Beijing, China, 2009.

[26] J. Wang, G. Li, and J. Feng. Fast-join: An efficient method for fuzzy token matching based string similarity join. In *27th International Conference on Data Engineering (ICDE'11)*, Hannover, Germany, 2011.

[27] Z. Wang, X. Zhang, L. Hou, Y. Zhao, J. Li, Y. Qi, and Y. Tang. RiMOM Results for OAEI 2010. In *5th International Workshop on Ontology Matching*, Shanghai, China, 2010.

[28] U. Westermann and R. Jain. Toward a Common Event Model for Multimedia Applications. *IEEE MultiMedia*, 14(1):19–29, 2007.

[29] D. Xu and S.-F. Chang. Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment. *IEEE Transaction on Pattern Analysis and Machine Intelligence (T-PAMI)*, pages 1985–1997, 2008.