Detection, Representation, and Exploitation of Events in the Semantic Web

Workshop in conjunction with the
10th International Semantic Web Conference 2011
Bonn, Germany, 23 October 2011

Edited by:
Marieke van Erp
Willem Robert van Hage
Laura Hollink
Anthony Jameson
Raphaël Troncy

DeRiVE 2011 is sponsored by:

_textkernel

Introduction

The goal of this workshop is to strengthen the participation of the Semantic Web community in the recent surge of research on the use of *events* as a key concept for representing knowledge and organising and structuring media on the web. The workshop call for papers invited contributions to three central questions, and the discussion at the workshop itself will aim to formulate answers to these questions that advance and reflect the current state of understanding. Each paper accepted for presentation at the workshop addresses at least one question explicitly, and several are accompanied by a system demonstration. The workshop concludes with a challenge competition in which systems that may address any of the three main questions make use of RDF datasets of event-related media such as EventMedia[1]. The challenge prize sponsored by _textkernel[2].

Why the Topic Is of Particular Interest Now

In recent years, researchers from several communities involved in aspects of the web have begun to realise the potential benefits of assigning an important role to *events* in the representation and organisation of knowledge and media—benefits which can be compared to those of representing entities such as persons or locations instead of just dealing with more superficial objects such as proper names and geographical coordinates. While a good deal of relevant research—for example, on the modelling of events—has been done in the semantic web community, a lot of complementary research has been done in other, partially overlapping communities, such as those involved in multimedia processing and information retrieval. The goal of this workshop is to advance research on this general topic within the semantic web community, both building on existing semantic web work and integrating results and methods from other areas, while focusing on issues of special importance for the semantic web.

[1] `http://thedatahub.org/dataset/event-media`
[2] `http://www.textkernel.com/`

Questions Addressed

The intended outcome of the workshop is to advance understanding of three high-level questions about the role of events in the semantic web. Below we reproduce each of the three main questions (and associated more specific questions) that were included in the call for papers for the workshop. We then indicate how the papers accepted for presentation in the corresponding sections of the workshop address the respective questions.

Question 1: How can events be detected and extracted for the semantic web?

More Specific Questions

- How can events be recognised in particular types of material on the web, such as calendars of public events, social networks, microblogging sites, semantic wikis, and normal web pages?
- How can the quality and veracity of the events mentioned in noisy microblogging sites such as TWITTER be verified?
- How can a system recognise when a newly detected event is the same as a previously detected and represented event?
- How can a system recognise a complex event that comprises separately recognisable subevents?

Contributions of Accepted Papers

One of the core obstacles for using events is that they are often difficult to detect. In text, one can describe and refer to events in a myriad of ways. In video, it is difficult to discern which frames denote interesting or significant events and which are merely fillers. For the event detection track, we received submissions that address a variety of issues in event detection. The papers we have accepted can be divided into two types: automatic event detection approaches (for text) and crowdsourcing approaches (for video and images).

The paper *An Overview of Event Extraction from Text*, by Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak and Franciska de Jong, provides a thorough overview of event detection approaches from text and makes recommendations for choosing the right approach for different problems. An example of a data-driven event detection approach is presented in *Using Semantic Role Labeling to Extract Events from Wikipedia*, by Peter Exner and Pierre Nugues. By using standard text mining tools in a cascaded event detection pipeline, the authors show how they can extract event elements with reasonable precision and recall.

As image and video processing have yet to reach a state where they can be used for event detection, the papers about detecting events from videos and images rely on crowdsourcing. *Crowdsourcing Event Detection in YouTube Videos*

by Thomas Steiner, Ruben Verborgh, Rik Van de Walle, Michael Hausenblas and Joaquim Gabarro Valles describes a three-tiered approach that uses visual processing combined with users' clicking behavior as well as the textual metadata that accompanies the video to identify different events. *Clues of Personal Events in Online Photo Sharing*, by Pierre Andrews, Javier Paniagua and Fausto Giunchiglia, identifies events by classifying how users organize their photos in albums. By classifying album titles, the authors show it is possible to identify photos about trips or different types of celebrations.

Question 2: How can events be modelled and represented in the semantic web?

More Specific Questions

- How can we improve the interoperability of the various event vocabularies such as EVENT,[3] LODE,[4] SEM,[5] and F?[6]
- How can aspects of existing event representations developed in other communities be adapted to the needs of the semantic web?
- What are the requirements for event representations for qualitatively different types of events (e.g., historical events such as wars; cultural events such as upcoming concerts; personal events such as family vacations)?
- To what extent can/should a unified event model be employed for such different types of events?

Contributions of Accepted Papers

The term "event" has several meanings. It is used to mean both phenomena that have happened (e.g., things reported in news articles or explained by historians) and phenomena that are scheduled to happen (e.g., things put in calendars and datebooks). Events are also a natural way for referring to any observable occurrence grouping persons, places, times and activities that can be described. Hence, a number of different RDFS+OWL ontologies providing classes and properties for describing the "factual" aspects of events (*What* happened, *Where* did it happen, *When* did it happen, and *Who* was involved) have been proposed and compared.

The papers we have accepted can again be divided into two types: the ones that have been applied in practical applications such as museum narratives or e-Science and the ones who present more theoretical work for representing relationships between events. Paul Mulholland, Annika Wolff, Trevor Collins and Zdenek Zdrahal in *An event-based approach to describing and understanding*

[3] `http://motools.sourceforge.net/event/event.html`
[4] `http://linkedevents.org/ontology/`
[5] `http://semanticweb.cs.vu.nl/2009/11/sem/`
[6] `http://isweb.uni-koblenz.de/eventmodel`

museum narratives presents the Curatorial Ontology (CO) for describing curatorial narratives. This ontology draws on structuralist theories that distinguish between story (i.e. what can be told), plot (i.e. an interpretation of the story) and narrative (i.e. its presentational form). Lianli Gao and Jane Hunter in *Publishing, Linking and Annotating Events via Interactive Timelines: an Earth Sciences Case Study* describe two ontologies: Event, Timeline, Annotation and TemporalRelation for relationships between events. They also developed a semantic annotation system that enables the discovery, retrieval and ontology-based markup of such event data via interactive timelines.

Ilaria Corda, Brandon Bennett and Vania Dimitrova in *A Logical Model of an Event Ontology for Exploring Connections in Historical Domains* describe a formal model for representing events and comparing temporal dimensions as the backbone for drawing connections and exploring relationships between happenings. Stasinos Konstantopoulos in *Using On-the-Fly Pattern Transformation to Serve Multi-Faceted Event Metadata* proposes the SYNC3 Ontology which is based on both the DOLCE Ultralite ontology and the F model and contains a number of conversion rules to the common LODE ontology.

Question 3: How can events be exploited for the provision of new or improved services?

More Specific Questions

- How can event representations be better exploited in support of activities like semantic annotation, semantic search, and semantically enhanced browsing?
- What application areas for semantic technologies can benefit from an increased use of event representations?
- How can we improve existing methods for visualising event representations and enabling users to interact with them in semantic web user interfaces?
- What requirements for event detection and representation methods (Questions 1 and 2 above) are implied by advances in methods for exploiting events?

Contributions of Accepted Papers

The four accepted papers for this part of the workshop mostly contribute new ideas about forms of exploitation and application areas, though there is also some attention to interaction design and visualisation.

Linked Open Piracy, by Willem R. van Hage, Véronique Malaisé, and Marieke van Erp, shows in detail how formally represented events can be used to support the creation of mashups and visual analytics. Referring to the specific application goal of analysing pirate attacks on shipping, the authors show how piracy reports intended for human reading can be augmented with semantic representations that in turn make possible a variety of visualisations and statistical analyses.

A different application area—web archiving—is discussed in *Using Events for Content Appraisal and Selection in Web Archives*, by Thomas Risse, Stefan Dietze, Diana Maynard, Nina Tahmasebi, and Wim Peters. The authors address the goal of archiving material from the web in a relatively structured and selective way, aiming to capture material related to events (and other entities) in a way reminiscent of a "community memory", exploiting the wisdom of the crowd. A good deal of the paper discusses strategies for overcoming the challenges for event extraction and detection that arise when this goal is pursued.

An application in the area of cultural heritage is presented in *Hacking History: Automatic Historical Event Extraction for Enriching Cultural Heritage Multimedia Collections*, by Roxane Segers, Marieke van Erp, Lourens van der Meij, Lora Aroyo, Guus Schreiber, Bob Wielinga, Jacco van Ossenbruggen, Geertje Jacobs, and Johan Oomen. The authors show how linking cultural artifacts to explicitly modelled events (and other entities) can support new forms of browsing and searching. The paper also discusses the challenges involved in extracting the relevant historical events from texts.

More attention to new forms of interaction with event representations is found in the paper *New Forms of Interaction With Hierarchically Structured Events*, by Sven Buschbeck, Anthony Jameson, and Tanja Schneeberger. The user interface presented differs from the more familiar timelines in that (a) it supports interaction with arbitrarily deep hierarchies of events linked via a "subevent" relation and (b) it offers functionality inspired by mind mapping applications to enable flexible browsing, searching, and media curation in a repository of events and associated media.

Challenge Competition

For the challenge part of the workshop, a dataset was made available consisting of over 100,000 events from the EventMedia LOD dataset (including events from Last.fm, Eventful, and Upcoming). Next to events, they contain artists, venues and location, description and time information. Some links between the instances of these three sources are provided.

This challenge dataset is intended to encourage participation by researchers who do not have an event dataset at their disposal and to increase shared understanding of the issues involved in working with data of this type. The application that makes best use of the provided datasets was awarded *The DeRiVE 2011 Challenge Prize*, which was sponsored by _textkernel. Submissions are judged by their (a) scientific contribution and (b) societal impact (e.g., how much the work contributes to useful applications by providing data or services).

Contributions of Contesters

The three accepted competition entries deal all with event background knowledge in some way. Two of them build new links to related concepts while one investigates how complex queries that use these relations to background knowledge can be executed in real time.

Kristian Slabbekoorn, Laura Hollink and Geert-Jan Houben study the problem of linking data to large, heterogeneous Linked Data sets in their paper *Domain-aware matching of events to DBpedia*. They use DBpedia Spotlight to create a baseline of matches between the artists in the EventMedia dataset and DBpedia resources. They show that knowledge of the domain in terms of relevant DBpedia categories and classes can increase the quality of the matches, and that this domain knowledge can be automatically derived. The resulting 19,840 links to DBpdia are made available for download.

In *Events Retrieval Using Enhanced Semantic Web Knowledge*, Pierre-Yves Vandenbussche and Charles Teissèdre demonstrate the benefit data enrichment in a retrieval system. They link the events to several external sources: city, country and address information, images associated to the events, and links to people and bands in DBpedia. They build a retrieval system that parses natural language queries containing agents, places, and complex temporal expressions. The resulting events and their images are visualised on a timeline.

In *Fusion of Event Stream and Background Knowledge for Semantic-Enabled Complex Event Processing*, Kia Teymourian, Malte Rohde, Ahmad Hassan and Adrian Paschke present research on how to apply reactive semantic complex event processing to event streams. By means of query pre-processing given a static knowledge base their Prova-based system is able to answer complex queries about events in real time.

Programme Committee

The following colleagues kindly served in the workshop's program committee. Their joint expertise covers all of the questions addressed in the workshop, and they reflect the range of relevant scientific communities.

- Jans Aasman, Franz Inc.
- Klaus Berberich, Max Planck Institute for Computer Science, Germany
- Fausto Giunchiglia, University of Trento, Italy
- Christian Hirsch, University of Auckland, New Zealand
- Ramesh Jain, University of California, Irvine, USA
- Krzysztof Janowicz, Pennsylvania State University, U.S.A.
- Jobst Löffler, Fraunhofer IAIS, Germany
- Marco Pennacchiotti, Yahoo! Labs, U.S.A.
- Yves Raimond, BBC Future Media & Technology, UK
- Ansgar Scherp, Universität Koblenz-Landau, Germany
- Nicu Sebe, University of Trento, Italy

- Ryan Shaw, University of North Carolina, U.S.A.
- Michael Sintek, DFKI, Germany
- Alan Smeaton, Dublin City University, Ireland
- Nenad Stojanovic, Forschungszentrum Informatik, Germany
- Denis Teyssou, AFP, France

Organising Committee

This workshop was organised by:

- Marieke van Erp, VU University Amsterdam, The Netherlands
- Willem Robert van Hage, VU University Amsterdam, The Netherlands
- Laura Hollink, Delft University of Technology, The Netherlands
- Anthony Jameson, DFKI, Germany
- Raphaël Troncy, EURECOM, France

Contents

Challenge

An event-based approach to describing and understanding museum narratives

Paul Mulholland, Annika Wolff, Trevor Collins, Zdenek Zdrahal

Knowledge Media Institute, The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK
{p.mulhollandla.l.wolfflt.d.collinslz.zdrahal}@open.ac.uk

Abstract. Current museum metadata tends to be focused around the properties of the heritage object such as the artist, style and date of creation. This form of metadata can index a museum's collection but cannot express the relations between heritage objects and related concepts found in contemporary museum exhibitions. A modern museum exhibition, rather than providing a taxonomic classification of heritage objects, uses them in the construction of curatorial narratives to be interpreted by an audience. In this paper we outline how curatorial narratives can be represented semantically using our Curate Ontology. The Curate Ontology, informed by a detailed analysis of two museum exhibitions, draws on structuralist theories that distinguish between story (i.e. what can be told), plot (i.e. an interpretation of the story) and narrative (i.e. its presentational form). This work has implications for how events can be used in the description of museum narratives and their associated heritage objects.

Keywords: Cultural heritage, curation, story, plot, narrative, event, ontology.

1 Introduction

Currently, museum metadata and content management systems focus predominantly on museum collections that comprise the heritage objects for which the museum acts as custodian. Museum metadata tends to be built around the objects that comprise the collection, indexing them, in terms of properties such as the artist, style, its date of creation, location and the materials used in its construction. In contemporary museum practice, an exhibition is constructed to tell a story that makes use of the displayed heritage objects but expresses relationships beyond the indexing used for collection management. Understanding and describing curatorial narratives involves going beyond the classification of heritage objects toward their interconnection in alternative conceptual and presentational structures.

This work is being carried out within the DECIPHER project, funded by the EU 7th Framework Programme. An objective of DECIPHER is to develop intelligent tools for assisting museum curators and visitors in presenting digital heritage objects within an overall coherent narrative. Within this, our current work is concerned with understanding and formally describing curatorial narratives and their construction.

Some previous research has been carried out related to building conceptual structures and presentations that span multiple heritage objects. These generally make use of event-based ontologies and metadata schemes such as CIDOC CRM [1] to conceptually interconnect heritage objects. Bletchley Park Text [2, 3] used historical interviews described according to CIDOC CRM event-based metadata to assemble an online newspaper in response to a query. Interviews were grouped according to the common people, places and objects mentioned in their constituent events. Hyvonen et al [4, 5] used event-based metadata to assemble related heritage objects around another heritage object that acted as a hub or backbone to the presentation. In one case a movie about the ceramics process was represented as events and linked to other resources related to concepts (e.g. people objects) featured in the events [4]. In the other case, event structures were used to generate links within a poem and also to external resources giving additional information [5].

Wang et al [6, 7] use content metadata and user preferences to suggest related heritage objects of interest. Van Hage et al [8] combine this with a real-time routing system to provide a personalized museum tour guide creating a conceptual path across a number of heritage objects. The personalized tour guide developed by Lim and Aylett [9] associated heritage objects with a metadata structure they termed a story element that comprised events, people, objects, museum location and causal relationships to other story elements. Recommendations were made based on casual relationships and shared items contained in story elements. Finally, van Erp et al [10] describe a prototype system for event-driven browsing. The system suggests related heritage objects based on their associated events. By selecting related heritage objects the user can create a pathway through the heritage objects.

All of these systems aim to go beyond the presentation of a single heritage object by connecting multiple heritage objects within a single conceptual graph. All make interconnections based on common terms or concepts included in metadata schemas associated with the heritage objects. Additionally, Lim and Aylett [9] have an explicit causal property connecting story elements associated with heritage objects. However, none of these systems have an explicit representation of the curatorial narrative, the story it tells, or how heritage objects are employed in the telling of this story.

Our aim is to propose a conceptual model for curatorial narratives that specifies the structure and types of relationships found within them. This model could then be used to capture the decisions and interpretation implicit in a curator-produced narrative. In the next section we introduce two exhibitions that were analyzed to inform the development of the model. The bulk of the paper outlines the Curate Ontology[1], drawing on examples from the exhibitions we have studied. Finally, we discuss how the work relates to the objectives of the workshop and outline ongoing work.

2 Investigating the curatorial process

The Curate Ontology, our model of the curatorial process, has drawn on an analysis of two exhibitions. Our investigation looked at how the exhibitions were constructed, the

[1] http://decipher.open.ac.uk/curate

conceptual structures within them and the use made of heritage objects. The two exhibitions were *The Moderns – The Arts in Ireland from the 1900s to the 1970s* (shown at the Irish Museum of Modern Art) and *Gabriel Metsu – Rediscovered Master of the Dutch Golden Age* (shown at the National Gallery of Ireland).

The Moderns explored Irish art from around 1900 to 1970 [11]. The exhibition, which ran from October 2010 to February 2011, looked at modernity in art, the introduction of continental ideas to Ireland and the development of new art forms. The Moderns exhibition surveyed a large number of artists over a relatively long time period. The exhibition included works in a number of different media including film and photography.

The Gabriel Metsu exhibition ran from September to December 2010 [12]. Unlike the Moderns that surveyed a broad range of artists, the Gabriel Metsu exhibition was monographic, concentrating on the work of a single artist. Gabriel Metsu was a genre painter, specializing in scenes of daily life. He lived and worked during the Dutch Golden Age of the 17th Century.

These two exhibitions were chosen because they differed in terms of their themes, scope, and the nature of the exhibited works. Both were also recent exhibitions held by partners of the DECIPHER project; the Irish Museum of Modern Art and National Gallery of Ireland. This provided first-hand access to how the exhibitions were developed, the range of people involved and the array of supporting materials associated with the exhibition.

Our analysis drew on a visit to the exhibition (in the case of *The Moderns*), discussions with museum staff, analysis of a range of resources (including visitor booklets, museum panels, audio guide transcripts) and participation in workshops organized by the museum partners. A one-day workshop was held at each of the museums focusing on one of the two exhibitions. The first half of each day was devoted to presentations by museum staff whose work had contributed to the exhibition. The functions covered in the presentations included the research and curatorial design of the exhibition space; the design of activities and resources around the exhibition, such as teaching plans, learner resources audio guides and visitor booklets; outreach to other local gallery spaces; and how the museum provides support for museum professionals and others to conduct research related to the exhibition.

For the second half of each workshop we provided a set of scenarios exploring different ways in which technology developed in the DECIPHER project could create new visitor or learner experiences and also support the work of museum curators and researchers. Findings from the workshop were interpreted in terms of existing work related to the nature of narrative and the use of narrative in museums. In the next sections we outline the Curate Ontology drawing on observations from the two exhibitions.

3 The curatorial process as story, plot and narrative

Our analysis of curatorial narrative drew on two working hypotheses that helped guide our interpretation. First, we hypothesized that curatorial presentations are in the

form of narratives and therefore contain the properties found in other types of narrative such as novels and films. This led us to consider how structuralist accounts of narrative [13] in general could inform the study of curatorial narratives. Second, we hypothesized that curatorial narratives are not only a presentation but also the product of a process of inquiry, in which heritage objects and other materials are sources of evidence. Narrative inquiry suggests how research can be conducted that makes use of, or produces, narratives [14].

Structuralist theories identify story, plot and narrative discourse as components of narrative. Chatman [13] distinguishes between story (what can be told) and narrative (a way of telling the story). One story may be realised in many different narratives. Both story and narrative discourse have their own time. Story time is the actual chronology of the events and narrative time is the order in which the events are revealed to the reader.

Structuralist theorists such as Tomashevsky [15] also make a distinction between story and plot. The story (or fabula) and plot (or sjuzhet) contain the same events. In the story, the events are ordered chronologically. In the plot the events are reorganized in order to explain the relationships between them and structure them as a coherent whole. The plot therefore transforms a pure chronology of events to a form that highlights for example the conflicts in the story, how they came about and how they are resolved by the characters. A similar distinction is found in narrative inquiry in which the process of research, in particular historical research, can involve imposing some interpretation on the chronology of events [14] and then presenting the result as a narrative. Story, plot and narrative are therefore not only types of description but also stages in a narrative-based process of research.

Hazel [16] argues that story, plot and narrative discourse constitute three primary elements of narrative in which a story constitutes the events, the plot is their organization that imposes some interpretation on events, and the narrative discourse (or narrative) is the communication of the story and plot to the reader.

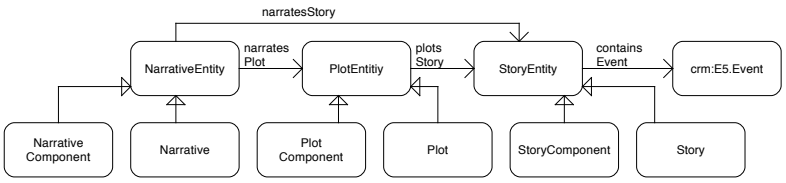


Fig.1. The relationships between narrative, plot, story and event.

As will be described later, our analysis of curatorial narrative has characteristics that can be usefully interpreted as story, plot and narrative. This distinction between story, plot and narrative allows us to introduce the first part of the Curate Ontology (see figure 1), in which a narrative narrates a plot and story, and a plot plots a story. A story contains events, which we illustrate here with the event class (E5) from the CIDOC CRM ontology. Finally, narratives (and plots and stories) can be divided into components. For example, a narrative (in the form of a book) may be divided physically into chapters, a plot can have sub-plots, and the story itself can be divided into components (as we shall discuss in section 5 on story structure).

4 Heritage object narratives and curatorial narratives

From the workshops, discussions with museum staff and analysis of materials it became clear that we needed to distinguish two types of narrative: heritage object narratives and curatorial narratives.

A heritage object narrative tells a story about a heritage object. Narratives can be found in the descriptions accompanying a heritage object when included in an exhibition. These may be, for example, in the exhibition catalogue, on a label displayed in the physical or virtual museum space, or in the audio guide description of the object. The Metsu exhibition website [17] shows some examples of what can be interpreted as heritage object narratives.

A heritage object may have multiple heritage object narratives. These heritage object narratives may draw on different aspects of the heritage object such as how the object was created, some insight it gives about the life of the artist, what is depicted in the heritage object or who has owned it. Heritage object narratives can also draw on different metaphorical uses of the heritage object. Pearce [18] gives an example of how an army jacket can be used to tell stories related to the Battle of Waterloo, in which it was worn or the Peterloo massacre in which the same jackets were worn.

Heritage object narratives may also be prepared for different audiences. For example, as part of The Moderns exhibition specially written descriptions of some of the included heritage objects were provided for older school children that matched their school curriculum.

These multiple narratives associated with individual heritage objects already start to move beyond schemas and management systems oriented around collections and start to provide some interpretation for the object, even situating it in the context of other objects in the same exhibition.

The second form of narrative identified is the curatorial narrative. We propose that a curatorial narrative threads across a number of heritage object narratives to create a narrative for the exhibition or some part of the exhibition space. Rowe et al [19] distinguish big and little narratives told by the museum to the visitor. An experience in the life of an individual could be a small narrative within the big, overarching narrative of the museum exhibition. Peponis et al. [20] in investigating the spatial design of science museums identify a narrative that makes conceptual relationships across a set of exhibits, yielding more complex insights than could be made from the exhibits individually.

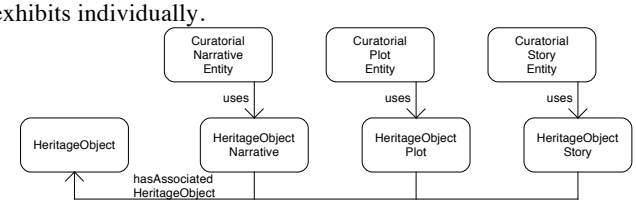


Fig. 2. The relationships between curatorial narratives, heritage object narratives and heritage objects.

In Gabriel Metsu, The Moderns and other exhibitions, examples can be found that can be interpreted as curatorial narratives. For example, in The Moderns, textual

narratives were associated with particular rooms or sub-sets of rooms within the exhibition. These constructed narratives concerned with, for example, Irish women modernists, that spanned a number of heritage objects and their individual narratives. The exhibition itself constitutes a narrative of which the narrative concerned with Irish women modernists is a component.

As heritage object narratives and curatorial narratives are both types of narratives they both have associated plots and stories. This provides us with the relationships in figure 2 where the curatorial narrative, plot and story layers make use of the heritage object narrative, plot and story layers, which in turn are associated with heritage objects.

5 Stories as conceptual organizations of events

As described earlier, a story is a collection of events that can be told within a narrative. Polkinghorne [14], in his study of narrative inquiry, describes how a story starts off as a chronological ordering of events (i.e. fabula, see section 3). A story can then be further organized into a storyline were the events are also classified according to specified themes, such as the type of activity or its location. This allows the story author to perceive the nature and frequency of different events over time. This is the definition of story adopted in the Curate Ontology.

This type of organization into a storyline could be seen from the two exhibitions investigated and the processes through which they were constructed. While the story is reflected in the final narrative it is not necessarily completely explicit and was therefore clarified in discussion with the curators.

Thematic and chronological organizations of the story were found in the two exhibitions. In the Gabriel Metsu exhibition, the story was divided into a number of sub-components that were organized chronologically, thematically or both. The first part of the exhibition was devoted to Metsu's early works. These were organized chronologically to show his progression as an artist. The other components were organized primarily according to themes. Some themes related to topics depicted in the works such as "taverns", "ladies and gentlemen"). One theme related to the use of the Amsterdam fine painting technique. Another set of works formed a group responding to Vermeer who was Metsu's contemporary.

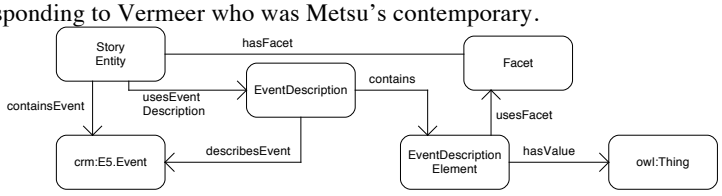


Fig. 3. Describing an event according to facets of the story.

Each of these story structures can be interpreted as a set of events organized by time and other dimensions. In some cases, such as the chronological organization of early works, the events of interest are those concerned with the creation of the heritage objects. For the story component responding to Vermeer, a broader set of

events is of relevance, concerned with how they may have influenced each other and changes in their relative standing as artists.

Within the Curate Ontology, we represent the organization of a story or story component as shown in figure 3. As in Event-Model-F [21] a distinction is made between an event and descriptions of that event. An event description contains event description elements that associate values according to different facets that have been assigned to the story or story component. Facets could be for example time, themes or location. The structure of the event description is, therefore determined by the facets of the story with which it has been associated.

When considered in combination with how heritage object stories are represented (section 4) we see that the relationship between a heritage object and an event is mediated by the heritage object story. This plays the role of the *illustrate* property in the LODE ontology [22, 23] that associates an object with an event. The mediating role of the heritage object story allows us to represent through which story the event is associated with the object.

6 Interpretation as emplotment of a story

Within the story, interpretation is limited to the selection and organization of events by time or other specified themes. Emplotment (the process of imposing a plot on the story) identifies a significant network of relationships between the events [24]. The plot is therefore more subjective and controversial than the story, placing a particular perspective on the events. A story could therefore be emplotted in multiple ways. Hazel [16] describes the plot as charting a path across the events of the story. The structure charted by the plot may be of different types such as tragedy, comedy and satire [14]. A plot also has a premise, moral or point that draws together the elements of the plot [19].

We have identified three types of plot element that express relationships between events, between story components, or between both events and story components. These will be considered in turn.

Plot relationships are expressed between events in order to place the events of the story into a coherent whole in which each included event has a role to play in the overall progression of the narrative. Possibly the most widely reported plot relationship between events is cause-effect, where the events of the story are organized into a causal sequence [25]. However, within narratology there is a recognition that plot relationships between events are not purely cause-effect. Chatman [13] highlights "happenings" that have no cause within the narrative. Many of the relationships identified in the two exhibitions were subtler than cause-effect. A good example is the part of the Metsu exhibition that explored the relationship between the work of Metsu and Vermeer. The reputation of the two artists has fluctuated wildly over the last 300 years and this has been reflected in varying accounts offering complex relationships between the artists and events in their lives, more nuanced than *cause-effect* or a general *influence* relationship between the artists.

In expressing plot relationships between events we make use of the Event-F-Model design pattern [21] used to express, for example, causal and correlational relationships

between events. As shown in figure 4, a plot contains plot descriptions. The subclass EventRelationDescription classifies events from within the story. A justification can be provided for each plot description.

We employ a similar pattern to express relationships involving components of the story. Within The Moderns exhibition there were story components related to the works of two brothers; the artist J. B. Yeates and the poet and playwright W. B. Yeates. Although plot relationships could be expressed between events in each of those two components, it was also useful to express a broader comparison relationship between the two story components, indicating their role within the overall story.

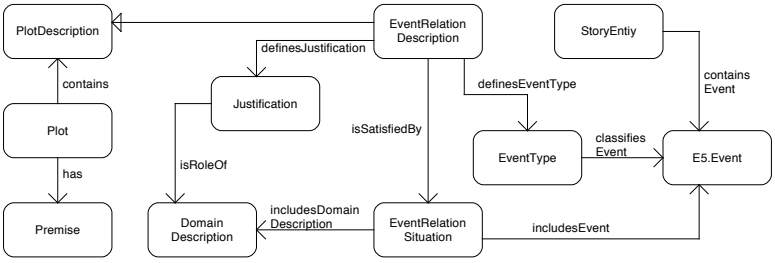


Fig. 4. Specifying plot relationships between events.

Finally, a similar pattern can be used to express relationships involving both events and story components. This could be used for example to express how an event was pivotal between two story components related to different time periods. For example, in The Moderns exhibition, the summer J. M. Synge spent on the Isle of Aran is seen as transforming the later representation of Irishness, which is taken up in other components of the story.

7 Narrative presentation of a story and a plot

The Curate Ontology can also be used to describe the contents of the narrative and its relationship to the underlying plot and story. This allows us to capture variations between the underlying conceptual structure and the narrative presentation in physical or digital form. A curatorial narrative within a physical museum space may vary considerably from the underling story due to different types of physical constraint. First, differences may be due to the fixed structure of the museum space. For example, the exhibition space at IMMA is made up of a number of relatively small rooms and interconnecting doors and corridors. This can result in a story component spanning a number of physical spaces, with the organization of heritage objects and interpretation panels across those spaces being as much determined by aesthetic and size constraints as the conceptual organization of the story.

Some differences between story and narrative organization may result from preservation constraints of the exhibits. For example, pencil sketches need to be displayed in darker conditions than are used for displaying paintings, therefore need to be separated in a physical museum space. Another obvious difference is that

heritage objects can be duplicated in the story space but not in the physical museum space. A number of examples were found of heritage object narratives that referred to not only to other works in the same physical area but also to works some distance away in the exhibition. Some, not due to preservation constraints, could be seen as reflecting alternative story structures that were not privileged in the physical space.

The Curate Ontology can represent the structure of the narrative again using a pattern similar to the Event-F-Model [21] though this time to classify components of the narrative and provide a justification for the structure. Example structures that can be defined include a linear structure (to represent a sequence of rooms in a physical or online gallery) or a hub and spoke structure (in which a central space has a number of offshoots). Work on rhetorical patterns in hypertext [26] indicates a number of candidate structures that can be described.

8 Discussion and further work

We have discussed our work developing the Curate Ontology, drawing on narratology, narrative inquiry, an analysis of museum exhibitions and event modelling research. Our work addresses the themes of the workshop in the following ways:

(i) We have identified how heritage objects can be associated with events mediated by the heritage object stories that can be told around a heritage object. Heritage object stories may highlight different perspectives such as the artist, how the object was made or what it depicts, or what has happened to it since its creation.

(ii) We have outlined how curatorial narratives can be described, distinguishing the presented narrative from the conceptual structure of the story and the role of events within that conceptual structure.

(iii) Our approach to representing event descriptions is consistent with existing patterns and shows how these descriptions can be tied to story entities and facets to create the storylines found in narrative inquiry research.

(iv) We have described how plots can be represented within museum narratives and how this builds on existing research related to the formal description of causal or correlational relationships between events.

Our current work is focussed on testing the Curate Ontology against cases offered by our museum partners. To facilitate this we have been developing an API and web interface to the Curate Ontology using the Drupal CMS. This makes mappings from content types and fields of the Drupal CMS to classes and properties of the ontology, similarly to Corlosquet et al [27]. In testing the model we are particularly interested in elaborating the types of story, plot and narrative structure required to express the decisions made in curatorial practice.

Acknowledgements

This work was supported by the DECIPHER project (270001), funded by the EU 7th Framework Programme in the area of Digital Libraries and Digital Preservation.

References

1. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (eds.): Definition of the CIDOC Conceptual Reference Model, (2010).
2. Collins T., Mulholland, P. and Zdrahal, Z.: Semantic browsing of digital collections. In: ISWC (2005).
3. Mulholland, P., Collins, T. and Zdrahal, Z.: Bletchley Park Text. In: Journal of Interactive Media in Education, http://jime.open.ac.uk/2005/24 (2005).
4. Hyvönen, E., Makela, E., Kauppinen, T., et al: CultureSampo: A national publication system of cultural heritage on the semantic Web 2.0. In: ESWC (2009).
5. Hyvönen, E., Palonen, T., Takala, J.: Narrative semantic web - Case National Finnish Epic Kalevala. In: 7th Extended Semantic Web Conference (2010).
6. Wang, Y., Aroyo, L. and Stash, N.: Interactive User Modeling for Personalized Access to Museum Collections: The Rijksmuseum Case Study. In: User Modeling (2007).
7. Wang, Y., Aroyo, L. and Stash, N., et al: Cultivating Personalized Museum Tours Online and On-site. In: Interdisciplinary Science Reviews, 32 (2), pp. 141-156, (2009)
8. van Hage, W. R., Stash, N., Wang, Y., Aroyo, L.: Finding Your Way through the Rijksmuseum with an Adaptive Mobile Museum Guide. In: ESWC (2010).
9. Lim, M. Y. and Aylett, R.: Narrative construction in a mobile tour guide. In: International Conference on Virtual Storytelling (2007).
10. van Erp, M., Oomen, J., Segers, R., van den Akker, C. et al: Automatic Heritage Metadata Enrichment with Historic Events. In: Museums and the Web (2011).
11. Arnold, B., Cass, B., Dorgan, T. et al: The Moderns - The Arts in Ireland from the 1900s to the 1970s. Irish Museum of Modern Art (2011).
12. Waiboer, A. E.: Gabriel Metsu, Rediscovered Master of the Dutch Golden Age. National Gallery of Ireland (2010).
13. Chatman, S.: Story and Discourse: Narrative structure in fiction and film. Cornell U. (1980).
14. Polkinghorne, D.: Narrative knowing and the human sciences. State Univ. NY Press (1988).
15. Tomashevsky, B.: Thematics. In: Lemon, L. T., Reis, M. J. (eds.): Russian Formalist Criticism: Four Essays. University of Nebraska Press (1965).
16. Hazel, P.: Narrative and New Media. In: Narrative in Interactive Learning Environments (2008).
17. Gabriel Metsu: Selected Works, http://www.gabrielmetsuexhibition.com/galleryIndex.php
18. Pearce, S. M.: On collecting: An investigation into collecting in the European tradition. Routledge: London (1995).
19. Rowe, S., Wertsch, J., Tatyana, K.: Linking Little Narratives to Big Ones: Narrative and Public Memory in History Museums. Culture and Psychology, 16 (2), pp. 96-112 (2002).
20. Peponis, J., Dalton, R., Wineman, J., Dalton, N.: Path, theme and narrative in open plan exhibition settings, International Space Syntax Symposium (2003).
21. Scherp, A., Franz, T., Saathoff, C., Staab, S. F—A Model of Events based on the Foundational Ontology DOLCE+DnS Ultralite. In: K-CAP (2009).
22. Shaw, R., Troncy, R., Hardman, L.: LODE: Linking Open Descriptions of Events. In: Asian Semantic Web Conference (2009).
23. Troncy, R., Malocha, B., Fialho, A.: Linking events with media. In: I-Semantics (2010).
24. Roberts, G.: The history and narrative reader. Routledge (2001).
25. Allen, R. B. Acheson, J.: Browsing the structure of multimedia stories. In: ACM Digital Libraries (2000).
26. Bernstein, M.: Structural patterns in hypertext rhetoric. In: ACM Computing Surveys, 31 (4) (2000).
27. Corlosquet, S., Renaud, D., Clark, T., Polleres, A., Decker, S.: Produce and Consume Linked Data with Drupal! In: International Semantic Web Conference (2009).

Publishing, Linking and Annotating Events via Interactive Timelines: an Earth Sciences Case Study

Lianli Gao[1] , Jane Hunter[1]

[1] School of ITEE, The University of Queensland,
Brisbane, Australia
{ l.gao1, j.hunter}@uq.edu.au

Abstract. Events are a critical entity for documenting information within many domains - and yet they are one class of information that, to date, has been relatively neglected with regard to both publishing on the Semantic Web and semantically annotating. In this paper we describe how we enable the interoperable integration, annotation and linking of information about major events from the earth sciences domain, by adopting a Linked Data approach to major events (earthquakes, tsunamis and volcanic eruptions) and the timelines and annotations that capture additional domain-expert knowledge. Firstly we describe the common *Event*, *Timeline*, *Annotation* and *TemporalRelation* ontologies that we use to enable interoperability and exchange of information about events and the relationships between them. We then harvest data describing major geological events from multiple authoritative sources, map it to our model(s) and publish it as RDF triples to the Web of Linked Data. We then describe the semantic annotation system that we have developed that enables the discovery, retrieval and ontology-based markup of such event data via interactive timelines. The resulting annotations significantly enhance the discovery and re-use of information about major geological events. More importantly these annotation tools enable scientists to document, share and discuss their hypotheses about the temporal relationships between such events.

Keywords: events, timelines, linked data, semantic annotations, geosciences

1 Introduction

Understanding the temporal relationships between historical events is often a critical step in predicting the occurrence of future events. The focus of this paper is on tools to assist scientists to improve their understanding of the temporal relationships between major geological events (earthquakes, tsunamis and volcanic eruptions). More specifically this paper describes the ontologies and semantic annotation system that we have developed that enables earth scientists to visualize, annotate and analyse temporal relationships between major geological events (earthquakes, tsunamis and volcanic eruptions) using interactive timelines. In the process of developing this system, we have also developed a Linked Data approach to such events that retrieves relevant data from a number of disparate authoritative

sources, integrates the datasets via a common *Event* model and publishes it as RDF triples (via Atom feeds or to a Linked Data Hub). Moreover to facilitate the interoperability, exchange and re-use of both the geological events and user-specified aggregations and annotations, we have also developed common, extensible ontologies to describe *Timelines, TemporalRelations*, and *Annotations*. The details and source of these ontologies are also described in this paper. The outcome is a set of tools for reasoning about the temporal relationships between major geological events that will hopefully lead to better models for predicting such potentially catastrophic events in the long run.

2 Objectives

Events are a critical information entity within many domains and yet they are overlooked when it comes to publishing as Linked Data. They are often hidden or encapsulated within databases, Web pages or timelines which prohibit their independent discovery, re-use, annotation or linking. The aim of the work described in this paper is to illustrate the benefits that are possible by treating events as first-class information objects and publishing them on the Semantic Web - where they can be annotated, interpreted and linked to other related datasets or events. More specifically, we demonstrate these benefits in the context of the earth sciences domain to enable scientists to analyse temporal relationships between past earthquake, tsunami and volcanic events (commonly known as "geochange" events) [1, 2].

Our first objective is to define a common data model for describing geochange events that enables interoperability between such events and a standardized model for publishing such events as Linked Data. Given this common model, we can then extract relevant geochange event data (about volcanic eruptions, earthquakes and tsunamis) from multiple authoritative Web sources (online databases, Web sites and timelines), represent it in our Event model and publish it as RDF triples, Atom feeds and/or to a Linked Data Hub.

One of the most common methods used to document, describe, aggregate, publish and visualize real world events on the Web is via *Timelines*. They provide a graphical representation of a chronology or sequence of events displayed along a time axis. One challenge to sharing event data is the multitude of timeline software systems and the lack of interoperability between them. When an event is published via a timeline (built using specific timeline software), the individual event data is not accessible, discoverable or re-usable – it is part of the "deep web", locked inside the particular timeline software and format. It is necessary to decouple *events* from *timelines* and the timeline rendering software – so that both events and timelines are discoverable and re-usable independently. Hence our second objective is to describe a common, interoperable model for *timelines* – that incorporates links to the contained events and that can also be published to the Web of Linked Data. To evaluate our timeline model, we identify a number of existing geochange timelines and show how the encapsulated data can be mapped to our models, without loss of information.

Given the resulting availability of both geochange events and timelines on the Web as Linked Data, our third objective is to develop semantic annotation tools for events – that enable researchers or the general public to add semantic markup to critical events

to enable additional interpretations and knowledge to be captured and to facilitate further reasoning across the events. Such knowledge includes the identification of temporal or causal relationships between events, which will lead to better predictability and early warning systems. As part of this objective, we also aim to develop an ontology of temporal relationships between events and methods for annotating relationships between events within the same and different timelines. Finally we plan to evaluate these ontologies and services in the context of the earth sciences domain by applying them to *geochange* events.

3 Related Work

There has been significant past research within numerous domains that aims to develop *a common event model* to support information integration. For example, the ABC model [3] was defined to document events (primarily in the information domain) that capture the provenance of documents undergoing change across multiple systems and platforms. The CIDOC/CRM [4] focuses on an interoperable model to support metadata exchange within cultural institutions. The Event (and associated Time) Ontology [5] was originally defined to describe events in the music and performance domain but has since been applied more generally. Other upper ontologies in which events are key entities include DOLCE+DnS Ultralite [6], the F Event model [7] and OpenCYC [8]. Shaw et al [9] provides a comparison of some of these existing event models in an effort to provide an interlingua model – the LODE ontology. Our approach is to adopt a simplified version of the LODE ontology (which is described in Section 5.1) and to apply it to specific types of geochange events (earthquakes, tsunamis and volcanic eruptions).

There also exists a vast number of Web-based tools for authoring and editing timelines. Examples include: SIMILE[1], My Timelines, Timeline Builder, xtimeline, Time Morph, Timelinr, Preceden and TimeGlider[2]. All of these systems rely on different sets of attributes and metadata to document the information contained within each timeline. There is little to no interoperability between the many different timeline software tools. The Timeline ontology developed by Raimond and Abdallah for the music domain is the most relevant previous work that aimed to develop a standardized RDF/OWL model for timelines [5]. However this timeline ontology is specifically aimed at the music domain to support mappings between time scales. In our Timeline model, we re-use a simplified sub-set of their classes and properties (described in Section 5.2). In addition, we define a "references/referencedBy" relationship between *timelines* and *events*, each of which are uniquely identifiable via persistent URIs. We also apply and evaluate the model using geochange data.

Existing approaches to annotating "events" have primarily involved proprietary approaches in which the annotations are locked inside the specific timeline tool or system. For example, Google's Interactive Charts, enable users to attach annotations to interactive timelines/charts that are rendered using Flash[3]. The annotations are not Web resources and are only accessible through the Google javascript API used to

[1] http://www.simile-widgets.org/timeline/
[2] http://www.shambles.net/pages/school/timelines/
[3] http://code.google.com/apis/chart/interactive/docs/gallery/annotatedtimeline.html

generate the timelines/charts. *RecordedFuture*[4] is an example of a browser-based temporal analytics tool that enables users to explore and visualize time-based data. *RecordedFuture* enables users to add annotations to events in the Timeline view. It also enables users to share event visualizations through Facebook, Twitter or a newly generated URL. However it is a commercial product that only allows the sharing of timelines between users who have purchased RecordedFuture. Other examples of timeline-based tools that support annotations include ChronoViz [10] and Chronozoom[5] – but again, neither system publishes the annotations as independent Web resources with unique persistent URIs that are discoverable, independent of the events. SemaTime [11] combines a timeline visualization interface with semantic annotation tools to annotate relationships between entities – but the focus is on visualizing semantic relationships that change over time (e.g., married_to).

4 Methodology

4.1 Case Study

Volcanic eruptions, earthquakes and tsunamis are strongly related to each other – both spatially and temporally. Both earthquakes and volcanoes occur at the boundaries of the tectonic plates that comprise the Earth's surface. Earthquakes are caused by pressure built up when plates collide, move apart, or slide past each other or over each other. Volcanoes form when the magma that is generated at plate boundaries rises to the surface. The movement of magma within a volcano or the adjustment of plates under volcanoes, causes earthquakes. Tsunamis are caused by the occurrence of earthquakes in oceanic or coastal regions. Understanding the temporal relationships between these geochange events will help scientists to develop better predictive models and early warning systems that may save lives of communities living in endangered zones. Hence our objective is to provide geologists and earth scientists with Web-based tools that enable them to aggregate disparate data sets describing geochange events and to document, analyse and interpret the temporal relationships between such events.

4.2 Process

Our approach can be divided into the five stages:

1. Firstly we developed common ontologies/data models for describing *events*, *timelines*, *annotations* and *temporal relationships;*
2. Next we harvested geochange event data from multiple Web sites and timelines (NOAA's National Geophysical Data Center's (NGDC) Natural Hazards Data[6], USGS Earthquake Database[7] etc), and represented it in our event and timeline models. We stored this data in our own RDF triple store, generating HTTP URIs for each event and timeline – but we also generated an Atom feed and published the RDF triples to the the Comprehensive Knowledge Archive Network (CKAN) [12] Linked Data Hub;

[4] https://www.recordedfuture.com/
[5] http://eps.berkeley.edu/~saekow/chronozoom/
[6] http://www.ngdc.noaa.gov/hazard/
[7] http://earthquake.usgs.gov/earthquakes/eqarchives/epic/database.php

3. We then developed a SPARQL interface to our RDF triple store to enable users to search, retrieve and display events based on metadata fields and display them on a Simile Widget Timeline [13] ;
4. Next we developed the SAFE (Semantic Annotation For Events) Firefox plugin that enables users to: annotate a single event on single timeline; annotate multiple events on single timeline or within a time period; annotate multiple events on different timelines displayed simultaneously; annotate relationships between events on same timeline or different timelines. The annotations are stored on an annotation server – using our OAC-based annotation model [14] – but we can also publish/share them as RDF via HTTP URIs and Atom feed. Users can also search and retrieve events via the annotations.
5. Finally we evaluated the system through user feedback and performance measures.

5 Ontologies

This section describes the ontologies that we've developed to support the publishing, linking and annotation of geological events. We have drawn on existing vocabularies and terms from the namespaces listed in the table below.

Prefix	XML namespace	Description
dc	http://purl.org/dc/elements/1.1/	Dublin Core
event	http://purl.org/NET/c4dm/event.owl#	Event ontology
tl	http://purl.org/NET/c4dm/timeline.owl#	Timeline ontology
time	http://www.w3.org/2006/time#	OWL-Time
geo	http://www.w3.org/2003/01/geo/wgs84_pos#	WGS84 GeoPositioning

5.1 Event Ontology

There exists a wide variety of existing *Event* ontologies from which to draw on, including the ABC ontology, the CIDOC CRM ontology, the Event Ontology, DOLCE+DnS Ultralite, OpenCYC and LODE [9]. Based on an analysis of these existing ontologies, as well as the requirements of our application, we defined a new *GeochangeEvent* class which is a subclass of the Event ontology's Event class [4] and defined the following properties in our GeochangeEvent class:

dc:identifer – HTTP URI for this event;
dc:title – title of the event;
dc:description – literal describing the event;
dc:source – HTTP URI of the source of the event data (e.g., USGS Web site);
event:time – range = time:TemporalEntity;
region – name of region where it occurred;
country – name of the country where it occurred;
geo:lat – coordinates in decimal degrees;
geo:long – coordinates in decimal degrees;
isReferencedBy – URIs of *Timelines* that reference this event.

We also defined a set of sub-classes of the GeochangeEvent class, that are specific to the geochange domain: *Earthquake*, *VolcanicEruption*, *Tsunami*. These three sub-classes each have additional specific properties. Earthquake events have the additional properties of: *magnitude* (0.0-9.9), *intensity* (0-12), *focalDepth* (0-700 km) and *numberOfDeaths*. Tsunamis have the additional properties of *waterHeight* (0-525 ms) and *numberOfDeaths*. VolcanicEruptions have the additional properties of *volcanoName*, *volcanoType* (Caldera, CinderCone, Lava, Mud, Pumice, Pyroclastic, Shield etc), *volcanicExplosivityIndex* (0-8) and *numberOfDeaths*.

5.2 Timeline Ontology

Our Timeline ontology was developed by analyzing the attributes used to describe existing timelines (e.g., the SIMILE widget/timeline) and also by drawing on terms from the Timeline ontology [5]. A critical addition to our model is the "references/referencedBy" relationship:

dc:identifier – HTTP URI for this timeline;
dc:creator – author of the timeline;
dc:title – literal title for the timeline;
dc:description – decription of the timeline;
dc:date.created – date the timeline was created;
tl:beginsAtDateTime – the start of the timeline;
tl:endsAtDateTime – the end of the timeline;
intervalUnit – the unit to be displayed on the axis intervals e.g. 1 hour, 1 year;
references – URIs of *Events* that are contained within this timeline..

5.3 Annotation Ontology

We chose to base our annotation ontology on the Open Annotation Collaboration (OAC) ontology [14] – which was specifically designed to enable the publishing and linking of annotations on the Web of Linked Data. The OAC ontology is ideal because: it is designed to support annotations in which the *body* and the *target* of the annotation may be of any media type (e.g. the body might be a seismograph); the annotation, body and target are all identifiable via HTTP URIs; and multiple targets are supported. This last aspect is particularly relevant as we want to support the annotation of temporal relationships between multiple events in different timelines.

Figure 1 below illustrates the OAC model corresponding to the annotation (A-1) of a "causal" relationship (B-1) between the Honshu Earthquake (E-1) in Japan in March 2011 contained in the Global Earthquakes timeline (TL-1) and the tsunami in the Miyako province (E-2) contained in the Global Tsunamis timeline (TL-2).

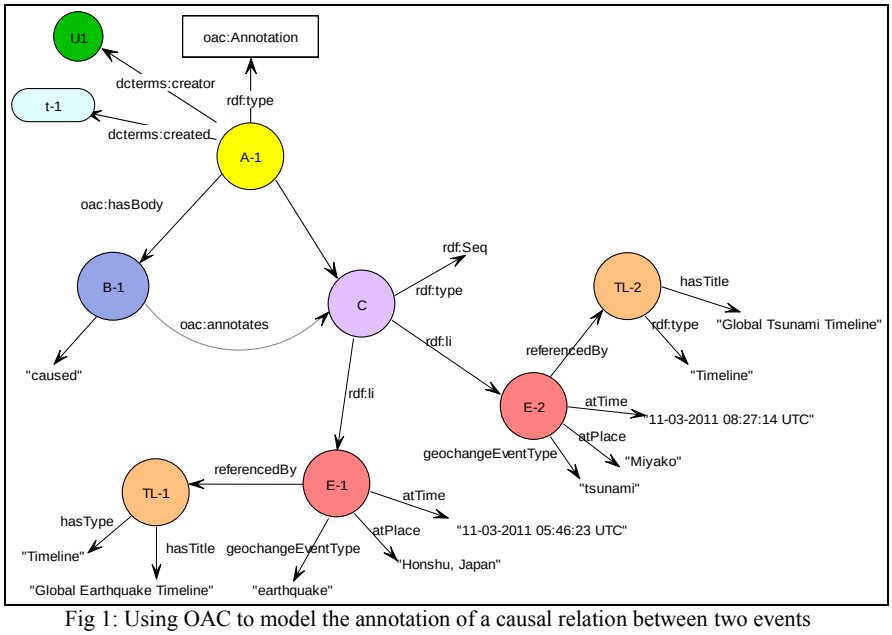


Fig 1: Using OAC to model the annotation of a causal relation between two events

5.4 TemporalRelation Ontology

The role of the TemporalRelation Ontology is to provide a set of controlled terms that can be used to tag relationships between different types of geochange events. The list of terms/properties that apply are listed below. This list is adapted from the list of temporal relations defined in the OWL Time ontology [15]:

time:before/time:after – one event precedes/follows another;
time:intervalOverlaps – the duration of two events overlaps;
time:intervalEquals – the start and end times of two events coincide;
time:intervalMeets – the end of one event coincides with the start of another event;
time:intervalContains – one event starts and finishes within the duration of a second event.

5.5 OtherRelations

Apart from the temporal relations described above, there were three other relationships that we defined:

isRelatedTo – one event is related to another – but the precise relationship is unclear;
causes/causedBy – one event causes/triggers another event (subPropertyOf isRelatedTo);
requires/requiredBy – one event cannot occur unless the other event has already occurred (subPropertyOf isRelatedTo).

6 Semantic Annotation Prototype

6.1 System Architecture

Figure 2 illustrates a high level view of the system architecture for our Semantic Annotation For Events (SAFE) service. A large collection of geochange event data is harvested from multiple authoritative Web sites and timelines including: USGS Earthquakes Database and the the NOAA NGDC natural hazards database. This data is mapped to our Event and Timeline models and stored as RDF triples in a Sesame RDF triple store. A user interface was developed that enables users to search and retrieve events from the Sesame triple store via titles, descriptions, dates/date ranges and/or keywords, using SPARQL. The retrieved events are dynamically displayed via a browser-based interactive timeline built using the SIMILE timeline widget. The SAFE annotation client is a Firefox sidebar built using XUL (XML User Interface Language), AJAX (Asynchronous JavaScript) and JavaScript. The annotation server is implemented using the Apache Tomcat. The various components of the system are accessible via the project Web Portal[8].

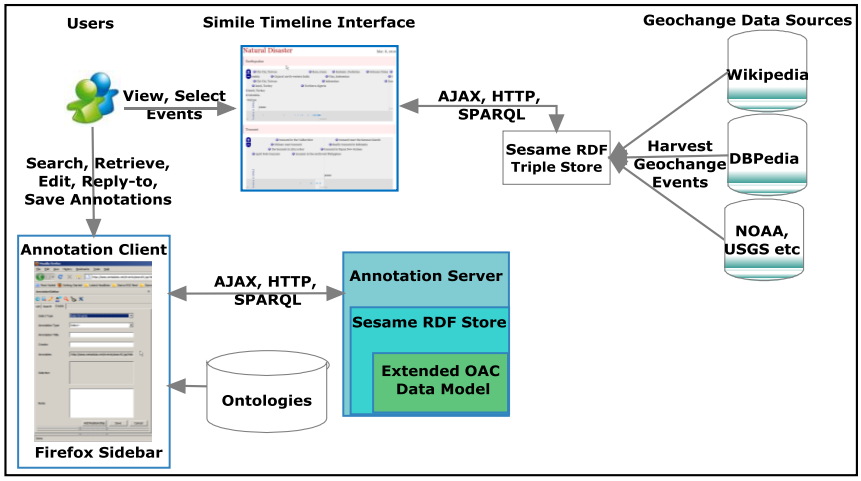


Fig. 2. High-level Architecture of the Semantic Annotation System

6.2 User Interface

The objective of the SAFE annotation system is to enable users to interactively:

- Annotate a single event on single timeline;
- Annotate multiple events on the same or multiple timelines, by selecting them individually or by interactively specifying a time period;
- Annotate relationships between events on the same or different timelines;
- Search and retrieve annotations and associated events based on the annotation metadata. Examples of is: “give me all causal relationships and associated events”, “give me all of the Timelines that reference this event”.

[8] http://seas.metadata.net/events/

Figure 3 illustrates the SAFE Firefox user interface being used to annotate a relationship between two events on different timelines. In the top LHS of Fig 3, is the sidebar for creating a new annotation - users specify the annotation type, title, creator, and body (e.g., controlled term, free text or URI). Users can also search, browse, edit and delete annotations via this sidebar. On the RHS are displayed one or more Simile timelines. The annotation client communicates with the Simile timeline(s) to extract and record the time range and the selected events that are associated with each annotation. Clicking on an existing annotation in the sidebar, causes the timeline to jump to the annotated event, highlight the event and display the associated annotation. To annotate a relationship between multiple events, users open one or more timelines, select the events of interest, and create a new annotation – the body of which is chosen from a pull-down menu. The corresponding RDF graph is displayed in the lower left hand corner of the sidebar.

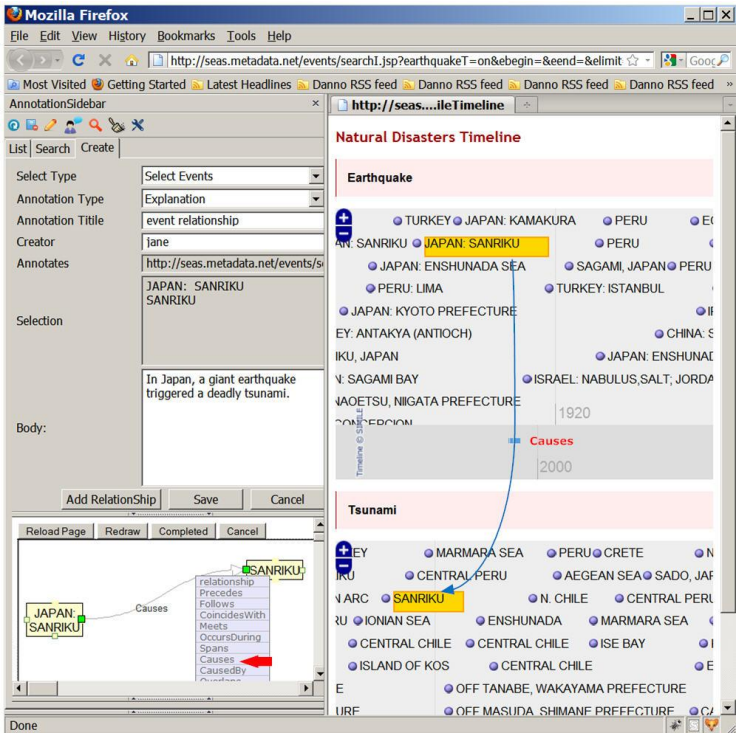


Fig 3: Using the SAFE Plugin with Firefox to annotate relations between events on 2 timelines

7 Evaluation and Discussion

The evaluation of the event, timeline and annotation ontologies described in Section 5 was based on their ability to support the mapping of harvested datasets and timelines from authoritative sources on earthquakes, tsunamis and volcanic eruptions. This mapping exercise illustrated that the property extensions to the sub-classes of geochange events enabled accurate descriptions to be captured e.g., the magnitude associated with an earthquake or the water height associate with a tsunami. An

additional sub-class of events that was identified as missing from the geochangeEvent class was “runup”. Runups are a consequence of tsunamis that occur when the water level rises onshore at multiple locations along the coastline, and of interest to geo-scientists. The relationships ontology needs to be extended to support related rules and restrictions. For example, a “causal” relationship between two events is only possible if the start time of the first event occurs before the start time of the second event. Currently the system does not check for such prerequisite temporal relationships but this would be relatively easy to implement as a validation process within the client annotation tool before saving the annotation. The OAC model was ambiguous in the context of annotating relationships between multiple events. The OAC model recommends the use of ore:aggregations for annotating multiple targets – however if they are ordered (e.g., sequential/list) then perhaps a blank node which is an rdf:Seq or rdf:List is a better approach. The other potential disadvantage associated with the OAC approach is the need to generate URIs for the annotation, body and target. This may well lead to a URI management problem in the long term – as well as a scalability problem as the number of annotations becomes very large and SPARQL querying struggles with the size of the RDF triple store.

User feedback to the SAFE annotation service (via a questionnaire) was mixed. Users found the Firefox sidebar easy to download, install, configure and the annotation interface intuitive and user friendly. Users liked the integration of the sidebar and timeline within the single browser and the speed of synchronization between the two panels. Users requested the ability to open more than two timelines simultaneously and to tag relationships between events contained within three or more timelines. They also requested the ability to attach a *certainty* measure to relationship tags. For example, they might tag a particular tsunami event as being *causedBy* a particular earthquake event, but the author’s confidence in this assertion is only 75%. Finally, a significant number of users requested the ability to specify both geo-spatial and temporal relationships simultaneously via a combined mapping and timeline interface (such as TimeMap [16]). Related to this was the additional request to enable interactive specification of more sophisticated querying and inferencing rules. For example, “find all tsunami events that fall within a 1000 km radius and within 18 hours of a particular earthquake event and tag them as *causedBy* the earthquake”..

8 Future Work and Conclusions

We have identified a number of future work directions that we would like to pursue. Firstly, we plan to integrate the timeline with a mapping interface to enable the annotation and visualization of both spatial, temporal and spatio-temporal relationships between geochange events. We are also planning to implement (SWRL) inferencing rules that enable users to reason across the data based on the tagged or inferred temporal relationships. For example, if someone tags *earthquake-E1->causes->tsunami-T1*, and someone else tags *tsunami-T1->causes->runups-R1, R2, R3, R4*. Then because the *causal* property is transitive, the system can infer that *earthquake-E1->causes->runups-R1, R2, R3, R4*. Users can then ask queries such as “what is the total numberOfDeaths caused by earthquake-E1?”.

In conclusion, we have described a set of services that enable information about geological events (that was previously hidden in databases, Web sites and timelines) to be exposed on the Web as Linked Data. Given the availability of these rich datasets on earthquakes, tsunamis and volcanic eruptions, we then developed a set of timeline-based annotation services that enable users to document and share their ideas and hypotheses about the temporal relationships between such events. The outcome is an extensible framework and a robust foundation for future more advanced temporal reasoning - not only about geological events, but about events more generally, from many domains and disciplines.

References

1. Foerster, T., Trame, J., Remke, A.: Web-based GEONETCast Data for Geochange Research. In: Hennebohl, K., Vinhas, L., Pebesma, E., Camara, G. (eds.) GIScience for Environmental Change Symposium Proceedings, vol. 40, pp. 1-6. (2010)
2. Devaraju, A., Kauppinen, T.: Geo-Processes and Properties Observed by Sensors: Can We Relate Them? In: GeoChange 2010 – GIScience for Environmental Change. (2010)
3. Lagoze, C., Hunter, J.: The ABC Ontology and Model. Journal of Digital Information (JoDI) 2(2) (2001)
4. Doerr, M., The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine 24(3), 75–92 (2003)
5. Raimond, Y., Abdallah, S., Sandler, M., Giasson, F., Sandler, M.: The Music Ontology. The 8th International Conference on Music Information Retrieval (ISMIR 2007), pp. 417-422, Vienna, Austria (2007)
6. Gangemi, A., Mika, P.: Understanding the Semantic Web through Descriptionsand Situations. In: International Conference on Ontologies Databases and Applications of SEmantics ODBASE, pp. 689-706. Springer, (2003)
7. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F—A model of events based on the foundational ontology DOLCE+DnS ultralight. In: The Fifth International Conference on Knowledge Capture (K-CAP 2009), pp. 137-144. (2009)
8. Matuszek, C., Cabral, J., Witbrock, M., Deoliveira, J.: An introduction to the syntax and content of Cyc. Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, pp. 44--49 (2006)
9. Shaw, R., Troncy, R., Hardman, L.: LODE: Linking Open Descriptions of Events. In 4th Annual Asian Semantic Web Conference (ASWC'09), vol. LNCS 5926, pp. 153-167, Shanghai, China (2009)
10. Fouse, A., Weibel, N., Hutchins, E., Hollan, J.D.: ChronoViz: A system for supporting navigation of time-coded data. The 2011 annual conference extended abstracts on Human factors in computing systems, pp. 299-304. ACM, Vancouver, BC, Canada (2011)
11. Stab, C., Nazemi, K., Fellner, D.W.: SemaTime - Timeline Visualization of Time-Dependent Relations and Semantics. Lecture Notes in Computer Science, vol. 6455, pp. 514-523, Los Angeles, USA (2010)
12. Open Knowledge Foundation, CKAN - the Open Source Data Hub http://ckan.net/
13. Simile Widgets Timeline, http://www.simile-widgets.org/timeline/
14. Open Annotation Collaboration (OAC): Beta Data Model Guide, 10 August, 2011 http://www.openannotation.org/spec/beta/
15. W3C, Time Ontology in OWL, Eds. J.Hobbs, F.Pan, W3C Working Draft , 27 Sept 2006 http://www.w3.org/TR/owl-time/
16. TimeMap Javascript library http://code.google.com/p/timemap/

A Logical Model of an Event Ontology for Exploring Connections in Historical Domains

Ilaria Corda, Brandon Bennett, and Vania Dimitrova

School of Computing, University of Leeds
ilaria,brandon,vania@comp.leeds.ac.uk

Abstract. Exploring connections between events is paramount to any historical investigation. In the course of human occurrences, historians have been always interested in unveiling connections between events for the purpose of establishing the significance of certain happenings and measure their impact. The paper describes a formal model for representing events and comparing temporal dimensions as the backbone for drawing connections and exploring relationships between happenings. The approach is illustrated in a case study from the Astronomical Revolution, a sub-domain of History of Science.

1 Introduction

Historical information is not just a collection of the most significant happenings, treated as distinct and unchained entities. It tells a story, forms a narrative which describes a chronological order and also suggests deeper connections. Hence, the ability to represent events and reason about their temporal relationships are paramount requirements when building a framework for exploring connections between historical occurrences. Understanding historical facts requires knowledge of many aspects of events such as: when and where an event happened, what events preceded or succeeded it, and whether its participants are involved in other events. Whereas ontological approaches are already established within subjects such as Biology and Medicine, domain ontologies for modelling historical domains, e.g. History or Philosophy, are still a relatively unexplored area. This may be attributed to a number of factors: historical domains tend to be both complex and loosely structured, they involve a wide variety of different kinds of entity and relation including temporal, conceptual and physical entities. There is clearly a need for a well-founded and general ontology applicable across historical domains which rigorously characterises the notion of events and formalises their key role within temporal information.

The remainder of this paper is organised as follows. First, we will describe the modelling decisions underpinning our model of an *Event Ontology* and temporal framework. In Section 4, we will illustrate a formal model of an *Event Ontology*, which includes vocabulary, domain, syntax and rules. Furthermore, in Section 6 the notion of semantic links will be introduced and exemplified as a means to construct sequences of semantically-related information. Finally, we will review related works and outline application domains in which our model can be employed.

2 Modelling Events

Events are situated occurrences incorporating complex and rich information which normally refers to the *5W*: *Who* (subject of the event), *What* (object), *When* (temporal dimension), *Where* (spatial dimension) and *Why* (causes and effects). We have developed a generic approach, applicable across historical domains, for modelling historical events and comparing time between them. This was inspired by Davidson's theory of events [5], which lays on the idea that each event-forming predicate is enriched with an extra argument-place to be filled with a variable ranging over event-tokens, which stands for particular dated occurrences. The main advantage is the ability to associate multiple properties to events, such as time, location, and other additional information, thereby avoiding adding extra relations to handle different event dimensions:

$$(\exists e)(\text{born}(\text{Galileo Galilei}, e) \wedge \text{Time}(e, 1564) \wedge \text{Place}(e, 1564))$$

Davidson's theory of events enabled us to deal with a wide range of historical events, such as scientific events, e.g. observation, discovery, human and social happening, e.g. births, deaths, cooperations and conflicts. In many cases, references to event tokens are hidden within the verbs that are used to describe them and, as in the above example, an additional event token variable is required to articulate the logical form. However, in the historical domain there are also cases where an event token is referred to directly by a naming phrase (what philosophers usually call a definite *description*). For instance wars and battles often have a specific name such as the "battle of Hastings", and historical periods are also referred to in this way, e.g. "Early Modern", and "Scientific Revolution". In such cases a term of the form named_de("Scientific Revolution") is used to refer directly to an event token.

$$\begin{aligned} &\text{named\_e}(\textit{"Scientific Revolution"}) \wedge \\ &\text{Time-start}(\text{named\_e}(\textit{"Scientific Revolution"}), 1543) \wedge \\ &\text{Time-end}(\text{named\_e}(\textit{"Scientific Revolution"}), 1750) \wedge \\ &\text{Place}(\text{named\_e}(\textit{"Scientific Revolution"}), \text{Europe}) \end{aligned}$$

In the next section, we will discuss the issues of dealing with temporal information in historical domains and present our modelling decisions in that respect.

3 Modelling Time

Temporal information in events has been embedded employing a calendar structure consisting of year, month and day in the form of YYYY-MM-DD. Temporal entities are represented as *time grains* which correspond to particular years, months, and days within the Gregorian calendar structure, also known as a Western calendar. In historical domains, temporal information can be missing due to the fact that historical sources

cannot fully reconstruct when exactly a given event occurred, and because of that time dimensions are only partially provided. *Time grains* refer to temporal entities that are considered as atomic, with respect to the temporal granularity with which information can be specified within the historical knowledge base. They correspond to particular time periods embedded within a calendar structure. More specifically, they refer to particular years, months or days within the calendar structure. We have mostly dealt with years as a minimum requirement and months. Instead, the finer day granularity is unusual in our domain. For instance, we are generally aware of the date of birth and death of a scientific figure, e.g. Isaac Newton died the 20th of March 1727, whereas it is quite unusual to hold complete information for events such a conducted experiment, e.g. Galileo Galilei conducted the experiment of falling bodies during 1604. Hence, the granularity in which the temporal information is expressed can vary, and our model needed to allow representing both coarse and fine-grained time dimensions. This particular modelling challenge has been taken into account when defining the semantics of ordering relations over the domain for comparing temporal information in events holding different time granularity. For instance, the time point 1564 is potentially coincident with 1564-04 as both occurred within the temporal span of that year. Comparing time points of different granularity was possible by introducing a weaker form of time inclusion based on the idea of *incidents*. *Incidents* define events that are temporally subordinated or included within a main event and can be applied between different levels of granularity. 1610-10 refines 1610 meaning that 1610-10 is incident within 1610. Hence, the first *time grain* is temporally within the second. In [1] a theory of time which takes intervals as primitives is presented, however the interval relations can be specified in terms of ordering constraints on their end points. We have employed Allen's vocabulary of interval relations to describe temporal relation between events on the basis of their start and end points. All 13 relations, including the converses, have been represented within our model. For instance, the relation *meet*(e_1, e_2) holds when the end point of e_1 is equal to or incident within the beginning e_2, as follows:

$$\text{Meet}(e_1,\ e_2),\ \text{Time-end}(e_1, t_2) =\ \text{Time-end}(e_2, t_4) \text{ or } \text{refines}(e_1,\ e_2)$$

In the next section, we will illustrate our *Event Ontology Model*, which includes vocabulary, domain, syntax and a set of inference rules.

4 An Event Ontology Model: Vocabulary and Domain

An Event Ontology is a logical structure such that:

$$\Omega = \langle \mathcal{V}, \mathcal{D}, \Phi, \preccurlyeq, begin, end, location, \delta \rangle$$

where: $\mathcal{V}$ is a vocabulary of symbols; $\mathcal{D}$ is a domain representing all entities in the real world; Φ is the set of all asserted and inferred formulae; $\preccurlyeq$ is an order relationship

over the domain $\mathcal{D}$; *begin* and *end*, *location* are functions over the domain; δ is an interpretation structure.

The **vocabulary** $\mathcal{V}$ specifies the sets of non-logical symbols:

$$\mathcal{V} = \langle \mathcal{V}_{\mathsf{c}}, \mathcal{V}_n, \mathcal{V}_t, \mathcal{V}_h, \mathcal{V}_r, \mathcal{V}_v \rangle$$

where $\mathcal{V}_c$ is the set of concept symbols; $\mathcal{V}_n$ is the set of name symbols; $\mathcal{V}_t$ is the set of *time grain* symbols; $\mathcal{V}_h$ is the set of symbols associated with event tokens (happenings); $\mathcal{V}_r$ is the set of binary relation symbols; $\mathcal{V}_v$ is the set of event-verb symbols.

The **domain** $\mathcal{D}$ specifies the objects from the real world and includes three distinct sub-domains

$$\mathcal{D} \;=\; I \cup E \cup T$$

where I is the set of all individuals. For instance, these can include particular people, places, physical objects and so forth; E is the set of all event tokens. These correspond to particular instances of events, which happen over a particular interval of time. Each event token has been defined following our adaptation of Davidson's theory of events. Event tokens are associated to particular event verbs which bind pairs of individuals known as subject and object of the relation; T is the set of all *time grains*. *Time grains* are particular years, months or days within the calendar structure and may be expressed in terms of any of these different levels of granularity. For example, the year 1066 is considered to be a *time grain* as is June 1965 and 1st April 2020. T consists of the union of all individuals from the three types of temporal entity:

$$T \;=\; Y \cup M \cup D$$

where Y is the set of all years; M is the set of all event months; D is the set of all days. We can define ordering relations on each of the sets of Y, M and D using the order relation $\preccurlyeq$. For instance, Y is a totally ordered set $(Y, \preccurlyeq)$ such that:

$$\forall y_1, y_2 \in Y : \; y_1 \preccurlyeq y_2 \;\vee\; y_2 \preccurlyeq y_1$$

Each *time grain* in T is a tuple including at least an element from Y. There are three possible combinations:

$$\langle y \rangle \;\text{ or }\; \langle y-m \rangle \;\text{ or }\; \langle y-m-d \rangle \;\; \textit{where} \;\; y \in Y, \;\; m \in M, \;\; d \in D$$

We define two temporal functions *begin* and *end* to map happenings from E to *time grains* from T, as follows:

$$\textit{begin} : E \rightarrow T$$

$$\textit{end} : E \rightarrow T$$

where for every event token $e \in E$ $\textit{begin}(\mathbf{e})$ is the *time grain* when e started and $\textit{end}(\mathbf{e})$ is the *time grain* when e ended; $\textit{begin}(\mathbf{e})$ always precedes $\textit{end}(\mathbf{e})$.

Similarly, we define the spatial function *location* to map happenings from E to individuals from I, as follows:

$$\textit{location} : E \rightarrow I$$

where for every event token $e \in E$ *location*(**e**) is the place where e occurred.

The **interpretation structure**

$$\delta = \langle \delta_c, \delta_n, \delta_t, \delta_h, \delta_r, \delta_v \rangle$$

interprets the non-logical symbols from the vocabulary by mapping them to the semantics:

- $\delta_c : \mathcal{V}_c \rightarrow 2^I$ assigns to each concept symbol a subset of individuals in I;
- $\delta_n : \mathcal{V}_n \rightarrow I$ assigns to each name symbol an individual from I;
- $\delta_t : \mathcal{V}_t \rightarrow P$ assigns to each *time grain* symbol a time point from P;
- $\delta_h : \mathcal{V}_h \rightarrow E$ assigns to each event token symbol an event token from E;
- $\delta_r : \mathcal{V}_r \rightarrow 2^{I \times I}$ assigns to each binary relation a subset of pairs from I;
- $\delta_v : \mathcal{V}_v \rightarrow ((I \times I) \rightarrow 2^E)$ assigns to each event-verb symbol a mapping from the set of pairs of individuals $I \times I$ to a subset of event tokens from E.

Example
We illustrate δ_c, δ_r and δ_h:

δ_c(astronomer) = {GALILEO, PTOLEMY, BRAHE ...}

δ_r(explain) = {⟨'GALILEAN THEORY OF TIDES', TIDE⟩, ⟨'KEPLERIAN MOON THEORY', TIDE⟩...}

δ_v(observe) = {⟨⟨GALILEO, SUNSPOT⟩, {GAL_OBSERVE_SUNP1, GAL_OBSERVE_SUNP2}⟩,
⟨⟨BRAHE, SUPERNOVA⟩, {BRAHE_OBSERVE_SUP1, BRAHE_OBSERVE_SUNP2}⟩, ⟨⟨BRAHE, STAR SPOT⟩, {}⟩, ...}

5 An Event Ontology Model: Syntax

Our syntax consists of atomic terms and propositions. The terms include Individuals $\mathcal{V}_n = \{a, b, c, \ldots\}$; Time points $\mathcal{V}_t = \{t_1, t_2, t_3, \ldots\}$; Concepts $\mathcal{V}_c = \{C_1, C_2, C_3, \ldots\}$; and Event tokens $\mathcal{V}_h = \{e_1, e_2, e_3, \ldots\}$. The **propositions** are either atomic propositions or propositional constructs. We have defined four types of declared propositions: Concepts and Individuals Propositions, Binary Relations Propositions, Time Propositions and Event Propositions.
Concepts and Individuals Propositions. Concepts and Individuals propositions include atomic propositions which deal with concepts and individuals from the domain.

- $\mathbf{C}_1 \sqsubseteq \mathbf{C}_2$ where $\mathbf{C}_1, \mathbf{C}_2 \in \mathcal{V}_c$;
- $\mathbf{C}_1$ (**a**) where $\mathbf{C}_1 \in \mathcal{V}_c$ and **a** $\in \mathcal{V}_n$;
- **a** = **b** where **a**, **b** $\in \mathcal{V}_n$.

Binary Relations Propositions. Binary relations propositions include binary relations between individuals over the domain.

- **R**(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$;

- t(**R**)(**a**, **b**) where t(**R**) $\in \mathcal{V}_r$ is a transitive relation where **a**, **b** $\in \mathcal{V}_n$;
- inv(**R**)(**a**, **b**) where inv(**R**) $\in \mathcal{V}_r$ is an inverse relation where **a**, **b** $\in \mathcal{V}_n$;
- sym(**R**)(**a**, **b**) where sym(**R**) $\in \mathcal{V}_r$ is a symmetrical relation where **a**, **b** $\in \mathcal{V}_n$;
- Binary relations introducing a lattice order between individuals. Lattice binary relations resemble the general binary relations between individuals, although they are used to cluster individuals that stand in a hierarchy based on their conceptual generality and specificity. The complete list of lattice relations have been defined, as follows:
  - sub_field(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$;
  - sub_phenomenon(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$;
  - sub_theory(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$;
  - sub_law(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$;
  - sub_doctrine(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$;
  - sub_historical_period(**a**, **b**) where **R** $\in \mathcal{V}_r$ and **a**, **b** $\in \mathcal{V}_n$.

Time Propositions. Time propositions model temporal relations between *time grains*.

- $\mathbf{t}_1 = \mathbf{t}_2$ and $\mathbf{t}_1 \approx \mathbf{t}_2$ where $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{V}_t$. They define different types of *time grains* equality;
- $\mathbf{t}_1 \leq \mathbf{t}_2$ and $\mathbf{t}_1 \lesssim \mathbf{t}_2$ where $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{V}_t$. They define different types of order relation;
- $\mathbf{t}_1 \geq \mathbf{t}_2$ and $\mathbf{t}_1 \gtrsim \mathbf{t}_2$ where $\mathbf{t}_1, \mathbf{t}_2 \in \mathcal{V}_t$. They have been defined as the converse of $\mathbf{t}_1 \leq \mathbf{t}_2$ and $\mathbf{t}_1 \lesssim \mathbf{t}_2$;
- begin(**e**, **t**) and end(**e**, **t**) where **e** $\in \mathcal{V}_h$ and **t** $\in \mathcal{V}_t$. begin and end satisfy the condition that for any **e**, where begin(**e**,$\mathbf{t}_1$) and end(**e**,$\mathbf{t}_2$), $\mathbf{t}_1$ and $\mathbf{t}_2$ are of the same granularity.

Event Propositions. Event propositions include event verb relations and associated properties such as location and the equality relation between event tokens. Similar to begin and end, event properties are defined as functional properties mapping an event token **e** to an individual from the class of places, respectively.

- token (**e**, **V**(**a**, **b**)) where **e** $\in \mathcal{V}_h$, **V** $\in \mathcal{V}_v$ and **a**, **b** $\in \mathcal{V}_n$;
- location(**e**, **a**) where **e** $\in \mathcal{V}_h$ and **a** $\in \mathcal{V}_n$. Begin, end and location are generic functional relations across historical domains.
- $\mathbf{e}_1 = \mathbf{e}_2$ where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$.

Propositional Constructs Propositional constructs hold a newly introduced proposition name and combine one of more atomic propositions. They include the complete set of Allen's thirteen relationships which defines all possible relations that two distinct *time grain* can have. Six pairs of the event-token propositions are converses.

- precede($\mathbf{e}_1$, $\mathbf{e}_2$) and preceded by($\mathbf{e}_2$, $\mathbf{e}_1$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$;
- start ($\mathbf{e}_1$, $\mathbf{e}_2$) and started by($\mathbf{e}_2$, $\mathbf{e}_1$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$;
- finish($\mathbf{e}_1$, $\mathbf{e}_2$) and finished by($\mathbf{e}_2$, $\mathbf{e}_1$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$;
- meet($\mathbf{e}_1$, $\mathbf{e}_2$) and met by($\mathbf{e}_2$, $\mathbf{e}_1$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$;
- contain($\mathbf{e}_1$, $\mathbf{e}_2$) and during($\mathbf{e}_2$, $\mathbf{e}_1$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$;
- overlap($\mathbf{e}_1$, $\mathbf{e}_2$) and overlapped by($\mathbf{e}_2$, $\mathbf{e}_1$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$;
- equal($\mathbf{e}_1$, $\mathbf{e}_2$) where $\mathbf{e}_1, \mathbf{e}_2 \in \mathcal{V}_h$.

In addition, further propositional constructs can be defined to link elements from the domain D. For example, we have included the following:

- participate($\mathbf{a}$,$\mathbf{e}$) where $\mathbf{a} \in \mathcal{V}_n$ and $\mathbf{e} \in \mathcal{V}_h$;
- instrument($\mathbf{a}$,$\mathbf{e}$) where instrument $\in \mathcal{V}_r$ and $\mathbf{a} \in \mathcal{V}_n$ and $\mathbf{e} \in \mathcal{V}_h$

The **semantic evaluation** of each proposition is defined using the interpretation structure δ and standard set theory. For instance, $C_1 \sqsubseteq C_2$, $\mathbf{t}_1 \approx \mathbf{t}_2$ and participate($\mathbf{a}$,$\mathbf{e}$) are evaluated as:

$[\![\mathbf{C}_1 \sqsubseteq \mathbf{C}_2]\!]$ = true if $\delta_c(\mathbf{C}_1) \subseteq \delta_c(\mathbf{C}_2)$, otherwise = false

$[\![\mathbf{t}_1 \approx \mathbf{t}_2]\!] = \mathit{true}\ \ \mathit{if}\ \ \delta_t(\mathbf{t}_1), \delta_t(\mathbf{t}_2), (\mathbf{t}_1 = \mathbf{t}_2$ or $\mathit{refined\text{-}time}(\mathbf{t}_1, \mathbf{t}_2))$, otherwise = false

$[\![\text{participate}(\mathbf{a},\mathbf{e})]\!] = \mathit{true}\ \ \mathit{if}$ token($\mathbf{e}$,$\mathbf{V}$($\mathbf{a}$,$\mathbf{b}$)) or token($\mathbf{e}$,$\mathbf{V}$($\mathbf{b}$,$\mathbf{a}$)), otherwise = false

We use a set of **rules** in the form of $\varphi_1, \varphi_2 \Rightarrow \varphi_3$ classified in three main modes:

- *Concept-based mode* includes rules that determine direct and indirect concept-individual inheritance. For instance: $\mathbf{C}_1(\mathbf{a}), (\mathbf{C}_1 \sqsubseteq \mathbf{C}_2) \Rightarrow \mathbf{C}_2(\mathbf{a})$
- *Relation-based mode* includes rules which define transitive, symmetrical inverse relationship closures as well as transitivity on lattice relations. For instance: trans(R)($\mathbf{a}$,$\mathbf{b}$), trans(R)($\mathbf{b}$,$\mathbf{c}$) $\Rightarrow$ R($\mathbf{a}$,$\mathbf{c}$) where R is a transitive relation (e.g. influence).
- *Event-based mode* includes rules which define reasoning upon events. For instance: precede($\mathbf{e}_1$,$\mathbf{e}_2$), contain($\mathbf{e}_2$,$\mathbf{e}_3$) $\Rightarrow$ precede($\mathbf{e}_1$,$\mathbf{e}_3$)

Rules can be used to derive new knowledge on the basis of established information. In our framework, we needed to derive implicit information from facts which are explicitly declared in our historical knowledge base. For example, from the lattice binary relation sub_field(classical physics, mechanics) and sub_field(mechanics, physics), we might be interested to infer that classical physics is a sub field of physics, by applying transitive closure on the sub_field relation.

6 Semantic Links

Semantic Links are the formal specifications of association patterns that we use to make explicit the links between events and entities on the basis of both factual information and structure of the ontology. Semantic Links follow the form of

$$semantic\_link(link\_type, \chi_1, \chi_2) \Rrightarrow \Omega(\chi_1, \chi_2)$$

χ_1, χ_2 are variables referring to elements in the *Event Ontology Model* Ω and link_type denotes specific connections between those variables, e.g. sub-concept relation. $\Omega(\chi_1, \chi_2)$ is a constraint linking χ_1 and χ_2 expressed in terms of a set of formulas of the Ontology language. *Semantic Links* can also make reference to common elements occurring in facts, e.g. the same person participating in two or more events.

The set of pairs of ontology elements related by a semantic link of type *link_type* will be referred to by $\delta_l(link\_type)$.

Semantic Links are classified in three main modes:

- *Semantic Links associated with Atomic Propositions*. These are links that correspond directly to atomic propositions asserted in the ontology. For instance, we define a link corresponding to the primitive sub-concept relation:

  $$semantic\_link(subclass, \chi_1, \chi_2) \Rrightarrow \{\chi_1 \sqsubseteq \chi_2\}$$

- *Semantic Links associated with Inference Rules*. These are links that correspond to relations that can be inferred from the explicit facts in Ω by logical inference rules. For instance:
  $semantic\_link(indirect\_sub\_concept, \chi_1, \chi_2) \Rrightarrow \{\mathsf{indirect\_sub\_concept}(\chi_1, \chi_2)\}$

- *Semantic Links associated with a condition involving a common element*. These are links that correspond to relations between two elements from Ω depending on their relation to a third intermediate element of Ω. For instance, two events may be linked by having a common participant:

  $semantic\_link(common\_participant, \chi_1, \chi_2), \Rrightarrow \{\mathsf{participate}(\xi, \chi_1), \mathsf{participate}(\xi, \chi_2)\}$

  For instance:
  $\delta_l(common\_participant) = \{\langle \mathsf{Gal\_Improve\_Tel}, \mathsf{Gal\_Publish\_Sidereus} \rangle,$
  $\langle \mathsf{Har\_Observe\_Sunsp}, \mathsf{Gal\_Observe\_Sunsp} \rangle, \ldots\}$

  This indicates that the events of Galileo improving on the invention of the telescope and Galileo publishing Sidereus Nuncius have a common participant, namely Galileo; and the events of Harriot observing the sunspots and Galileo observing the sunspots also have a common participant (the phenomenon of sunspots).

Sequences of *Semantic Links* form our notion of *Semantic Trajectories*, semantically significant paths, which are derived from the *Event Ontology Model* by applying rules to construct paths constituted from relational links among entities and events. *Semantic Trajectories* support exploratory navigation of historical information, as introduced in [2].

7 Related Work

Modelling of events is increasingly gaining widespread attention in the knowledge representation community [15, 17]. There are mainly two kinds of event models: those which facilitate interoperability in distributed event-based systems [12] or enhance accessibility to museum-related information [6], and those developed for specific applications [9] or domains [10]. In particular, there is a lack of event-centred approaches, which provide formal syntax and semantics for modelling domain ontologies [7]. On the other hand, domain-independent formal models of events [14] [12] are not often adequate when modelling specific domains or families of domains, e.g. historical domains. Event-centred approaches in historical domains are often associated with enhancing access to Cultural Heritage collections [8, 16] and representing the underlying semantics of bibliographic records [6]. In [13], events are extracted from various textual data and an event model (SEM) is employed to interlink collection objects along the event dimensions. In [11] and [6] event-based models are employed for describing resources across domains and facilitate semantic interoperability of metadata. Our logical model is based on the event-token reification method as presented by [5], but also provides a formal syntax and semantics for representing relationships between entities and events which integrates our temporal representation. The resulting formal model of an Event Ontology has the ability to make explicit connections between events and entities.

8 Conclusion and Application Domains

We have illustrated a logical model of an Event Ontology, which includes formal syntax, semantics and reasoning rules for defining a generic approach applicable across historical domains. Our approach for representing events was inspired by Davidson's theory of events [5], an event-token reification method which enables linking properties (e.g. location, scientific instrument, and temporal information) to historical events. The logical model of an *Event Ontology* enables one to make explicit links between events and entities on the basis of both factual information and structure of the ontology. We have envisioned that our logical model can be employed in a number of application domains:

- Support search and browsing activities. The event ontology model would serve as a resource gateway for retrieving information associated to each semantic link. A prototypical implementation of the model has been presented in [3].
- Support essay writing. The event ontology model would help students discover key ideas and elicit their connections to support essay writing.
- Construct narratives for museum collections. The event ontology model would assist exploration in collections by generating historical narratives which describe the *contextual reference space* [4] associated to each artefact.

We are currently using our event ontology model to facilitate knowledge discovery for supporting essay writing in the History of Science domain.

References

1. James F. Allen. Time and time again: the many ways to represent time. *International Journal of Intelligent Systems*, **6**:341–355, 1991.
2. Ilaria Corda. Discovering connections in historical domains: an approach based on semantic trajectories. In *Proceeding of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning, Doctoral Consortium Poster, Toronto, Canada*, May 9 - 13 2010.
3. Ilaria Corda, Brandon Bennett, and Vania Dimitrova. Interacting with an ontology to explore historical domains. In *ONTORACTŐ8: Proceedings of the 2008 First International Workshop on Ontologies in Interactive Systems*, pages 65–74, Washington, DC, US, 2008. IEEE Computer Society.
4. Ilaria Corda, Vania Dimitrova, and Brandon Bennett. Personalised support to examine context dependency between history of science events. In *Proceeding of the Workshop on Personalized Access to Cultural Heritage (PATCH 2008)*, 2008.
5. Donald Davidson. *Essays on Actions and Events*. Oxford University Press, 1980.
6. Martin Doerr, Christian-Emil Ore, and Stephen Stead. The cidoc conceptual reference model. In *International Conference on Conceptual Modeling*, 2007.
7. Tim Fernando. Observing events and situations in time. *Linguistic and Philosophy Journal*, 30:527–550, 2008.
8. Eero Hyvnen, Eetu Mkel, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppl, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkil, Panu Paakkarinen, Joonas Laitio, and Katariina Nyberg. Culturesampo – a national publication system of cultural heritage on the semantic web 2.0. In *Proceedings of the 6th European Semantic Web Conference (ESWC '09), Heraklion, Greece*, May 31 - June 4 2009. Springer-Verlag.
9. Ai Kawazoe, Hutchatai Chanlekha, Mika Shigematsu, and Nigel Collier. Structuring an event ontology for disease outbreak detection. *BMC Bioinformatics*, (9), 2008.
10. Yves Raimond and Samer Abdallah. The event ontology. Available on line at: http://motools.sf.net/event., 2007.
11. Tuukka Ruotsalo and Eero Hyvönen. An event-based approach for semantic metadata interoperability. pages 409–422. 2008.
12. Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. A model of events based on a foundational ontology. Technical report, Department of Computer Science, 02/2009, University of Koblenz-Landau, ISSN (Online) 1864-0850, 2009.
13. Roxane Segers. Extracting and modeling historical events to enhance searching and browsing of digital cultural heritage collections. In *ESWC (2)'11*, pages 503–507, 2011.
14. Glenn Shafer, Peter R. Gillett, and Richard Scherl. The logic of events. *Annals of Mathematics and Artificial Intelligence*, 28:315–389, 2000.
15. Ryan Shaw, Raphael Troncy, and Lynda Hardman. Lode: Linking open descriptions of events. Technical report, School of Information, UC Berkeley, 2009.
16. Chiel van den Akker, Susan Ległne, Marieke van Erp, Lora Aroyo, Roxane Segers, Lourens van Der Meij, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. Agora digital hermeneutics: Online understanding of cultural heritage. In *Proceeding of of the 3rd International Conference on Web Science (WebSci11)*, Koblenz, Germany, 2011.
17. Willem Robert van Hage, Veronique Malaise, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128 – 136, 2011.

Using On-the-Fly Pattern Transformation to Serve Multi-Faceted Event Metadata

Stasinos Konstantopoulos

Institute of Informatics and Telecommunications,
NCSR 'Demokritos', Greece
konstant@iit.demokritos.gr

Abstract. In this paper we present an extension of DOLCE UltraLite and Event Model F developed for the SYNC3 repository, storing semantic information about news content and the world events that such content documents. In the SYNC3 ontology we introduce a conceptualization of web documents and also propose an alternative mereological hierarchy for text document that the one in DOLCE. Finally, we introduce the idea of using inference to provide multi-faceted querying access to the data.

1 Introduction

In this paper we present the ontology developed for the SYNC3 data store, and we introduce the idea of using inference to provide multi-faceted querying access to the data. The SYNC3 data store manages and serves relations between world events and the news content that documents them, as well as metadata such as events' thematic category, location and time, participating named entities, related events, and the sentiment expressed in news content towards them.

The SYNC3 ontology extends the DOLCE UltraLite and Event Model F models, and is the schema of a large-scale triple store holding automatically extracted data, generated at a rate of roughly 40 million triples per month. In Section 2 we discuss some key decisions and points of divergence from these foundations, and motivate the decision to diverge.

In Section 3 we proceed to present a novel approach to multi-faceted querying that enables using different (previously coordinated) ontological schemas to query the same data. Our approach uses inference to dynamically generate data in different facets, avoiding the reduplication of data at such a large scale. This approach is discussed in Section 4, where future work is also outlined.

2 The SYNC3 Ontology

The SYNC3 domain is that of news and events described in news articles and blog posts, so that the concepts of a text *document* and of a news-worthy *event* reported in it are prominently situated in the SYNC3 model.

We shall not delve into the details of the linguistic processing pipeline of SYNC3 [1]; it suffices to say that at the end of this processing, the following information about documents and events has been extracted:

- Document metadata, including title, date of publication, and source.
- A breakdown of documents into *segments*, each comprising consecutive syntactic elements of the document which document the same event. Besides extracting events, the *sentiment* (if any) expressed in each segment towards the event is also extracted.
- The resolution of the *abstract domain entity* that each concrete term, pronoun or other anaphora in the document refers to.
- The geographical and temporal grounding of an event, as well as a numerical valuation of the level of participation of domain entities in each event.

2.1 Extending DUL/F

The SYNC3 ontology[1] is based on DOLCE+DnS UltraLite (DUL),[2] a modular foundational ontology which is the Description Logic-compatible subset of the DOLCE ontology [2].

DUL is a lightweight foundational ontology for modelling both physical and social contexts, extensions of which have been successfully applied in several domains. Most DOLCE modules have been ported to DUL, but particularly pertinent to SYNC3 are:

- *Descriptions and Situations* (DnS), conceptualizing social entities such as relations, roles, contexts, situations, and parameters; and
- *Information Objects*[3] (IOLite), covering expressions and meaning, logical and physical documents, and reference.

Event Model F [3] extends DUL+DnS to represent the participation of agentive, temporal, spatial, and other entities in events, as well as temporal, causal, and generic correlative relationships between events. Furthermore, Event Model F supports event composition and alternative interpretations of the same event.[4]

Most pertinent to the work described here is Event Model F's *participation pattern* that links an event participation description (kinds of participation) with specific objects (participants). This approach offers Model F the flexibility to have participant instances assume different roles in different event patterns without the need to define new sub-properties of the `f:hasParticipant` relation, but rather by populating the ontology with event role instances.

The IOLite concept of `io:InformationRealization` is specialized to web content as the `sync3:DigitalDocument` concept: the class of information realizations that occupy a `sync3:WebArchive` region. A `sync3:WebArchive` is the

[1] `http://www.sync3.eu/rdf/sync3` abbreviated hereafter to `sync3:`
[2] `http://www.loa-cnr.it/ontologies/DUL.owl` abbreviated hereafter to `dul:`
[3] `http://www.loa-cnr.it/ontologies/IOLite.owl` abbreviated hereafter to `io:`
[4] `http://events.semantic-multimedia.org/ontology/2009/4/15/model.owl` abbreviated hereafter to `f:`

subclass of `dul:SpatioTemporalRegion` that specifies a particular web location at a particular time. Its instance's properties carry crawling meta-data, such as URL and time of crawling, as well as a key that identifies a specific web document from a specific crawl stored in the SYNC3 multimedia repository. Furthermore, the spatial component of `sync3:WebArchive` instances can optionally have the `sync3:startsAt` and `sync3:endsAt` properties, restricting the region to the fragment between these two token indexes.

What should be noted is the distinction between the localization of the information object and its realization: the spatio-temporal specification of information objects refers to the time and place where the object was authored, as extracted from the object itself. The `sync3:WebArchive` instances that specify the, one or more, realizations of this objects conceptualize that a concrete digital object was retrieved by the system from a certain URL at a certain time point.

2.2 A New Mereology of Information Objects

SYNC3 extends the IOLite `io:LinguisticObject` pattern to represent meta-data and named-entity extraction results; linguistic objects are information objects where information is expressed in natural language.

The IOLite mereological organization of `io:Text`, `io:Sentence`, `io:Phrase`, `io:Word` was deemed inappropriate for SYNC3 because its axiomatization forbids the omission of any of its levels. Since SYNC3 processing follows a bag-of-words approach, `io:Phrase` and `io:Sentence` instances are not extracted. Furthermore, named-entity recognition in SYNC3 extracts multi-word references to an entity, a significant level between `io:Phrase` and `io:Word` that is missing from IOLite. For these reasons, the SYNC3 ontology defines its own mereology of linguistic objects, comprising `sync3:Text`, `sync3:Segment`, and `sync3:DomainTerm`, linked in a mereology by the `dul:hasComponent` relation. In this model:

- a `sync3:Text` instance represents a complete document,
- a `sync3:Segment` instance represents the *maximal semantically homogeneous* fragment of a `sync3:Text` instance, such that is a *single* `dul:Entity` can fill its `dul:expresses` property . In SYNC3, this filler is a `dul:Event` so that `sync3:Segment`S represent the maximal fragments of the article such that all entities mentioned in them are participants in the same event.
- a `sync3:DomainTerm` instance represents the *minimal* `sync3:Text` fragment that has a semantics and can be linked to a `dul:Entity` via `dul:expresses`. In SYNC3, `sync3:DomainTerm` instances are (possibly multi-word) references to domain entities (persons, organizations, locations, etc.)

We believe this mereology to be not only more appropriate for the SYNC3 application, but a model that is *generally* better suited to modern information extraction systems, as well as more flexible than the rigid sentence/phrase/word hierarchy imposed by IOLite. For example, applications where full-depth syntactic analysis is used to extract the full compositional semantics of the text could represent the complete analysis as a tree of appropriate specializations the

sync3:Segment class, where the dul:Entity expressed by each is an expression that composes the semantics of the sub-segments of this segment. From this perspective, sync3:DomainTerm is the sub-class of sync3:Segment that has semantics that cannot be further decomposed but are references to semantic units in the domain of discourse.

3 Pattern Transformation as Inference

The SYNC3 repository is implemented within the OpenRDF Sesame framework[5] and its architecture of *Storage And Inference Layers* (SAILs). Sesame SAILs are 'stackable' components that infer implicit RDF triples from the (explicit or also implicit) data they receive from the SAIL immediately below.

3.1 The LODE SAIL

We have implemented a SAIL that infers data following the *Ontology for Linking Open Descriptions of Events* (LODE)[6] given Event Model F data.

Both event models annotate events with a dul:Location and a spatio-temporal dul:Region using the dul:hasLocation and dul:hasRegion properties. These do not require any transformation. More interesting is the transfer of data about event participants: The two event models are different but compatible, in that both make a distinction between the participation of dul:AgentS and the participation of other dul:ObjectS in an event, and that both use a sub-property to denote that the participation of dul:AgentS is a special kind of participation.

$$\frac{\textit{sss}\ \texttt{rdf:type}\ \texttt{f:ParticipationSituation} \quad \textit{sss}\ \texttt{dul:includesEvent}\ \textit{eee} \quad \textit{sss}\ \texttt{dul:includesAgent}\ \textit{xxx}}{\textit{eee}\ \texttt{lode:involvedAgent}\ \texttt{xxx}}$$

$$\frac{\textit{sss}\ \texttt{rdf:type}\ \texttt{f:ParticipationSituation} \quad \textit{sss}\ \texttt{dul:includesEvent}\ \textit{eee} \quad \textit{sss}\ \texttt{dul:includesObject}\ \textit{xxx}}{\textit{eee}\ \texttt{lode:involved}\ \textit{xxx}}$$

However, among non-agentive participants, LODE can only model the participation of dul:Object instances, as lode:involved is restricted to range over dul:Object.

The more generic relation dul:isSettingFor between a situation and any dul:Entity instance is not transferred, even if its filler falls under one of the cases above. The rationale is that there is no axiom in DUL that forces a dul:isSettingFor relation between a situation and a dul:Object to assume the semantics of the more specific dul:includesObject property.

[5] See http://www.openrdf.org

[6] http://linkedevents.org/ontology hereafter abbreviated as lode:

In the SYNC3 ontology, news content is modelled as `dul:InformationObject` patterns, using `dul:isAbout` to link then to the `f:ParticipationSituation` that are reporting. In the LODE model this link is more direct, since instances of `dul:InformationObject` link directly to the `dul:Event` instance. The rule below infers this link:

$$\frac{\textit{io}\ \texttt{rdf:type}\ \texttt{dul:InformationObject} \quad \textit{sss}\ \texttt{rdf:type}\ \texttt{f:ParticipationSituation} \quad \textit{io}\ \texttt{dul:isAbout}\ \textit{sss} \quad \textit{sss}\ \texttt{dul:includesEvent}\ \textit{eee}}{\textit{io}\ \texttt{lode:illustrate}\ \textit{eee}}$$

3.2 Discussion

What can be observed is that these rules do more than simply re-writing property names, as there is a significant change of perspective between the two models: in LODE the `dul:Event` instance assumes a more 'central' position in the pattern, being the only instance that is directly linked to all other instances in the pattern. Event Model F, on the other hand, is more closely adhering to the DUL foundation by extending the generic Description/Situation pattern, so that the Situation instance is the 'central' element of the pattern.

Another important point is that the transfer of data between different event models is achieved by Java code specific to DUL and LODE. In order words, the *ontology coordination* knowledge about the correspondences between two ontological schemas is encoded as Java code rather than in a knowledge representation formalism.

On the positive side, most of this code deals with model-idependent tasks such as retrieving the statements that make up the source pattern, recursively applying the transformation to their property fillers, etc. The part that maps triples can be easily generalized to read the mapping from a knowledge base that encodes knowledge about the coordination of the two schemas.

The decision to encode the model-specific knowledge in the implementations of the event interface was taken based on the absence of a stable and generally-accepted schema for representing ontology coordination knowledge; all code design decisions were taken in anticipation of such a schema that will enable the development of a generic implementation of the event interface.

4 Conclusions

The main contributions of this paper are the extension and deployment of the DUL and Event Model F foundational ontologies on a large-scale application in the domain of world events and the on-line news content that reports them; and the development of a novel approach to multi-faceted access to data that dynamically generates facets without the need to reduplicate information in order to serve data under a different perspective.[7]

[7] The code pertinent to DUL/Model F is published as part of the TransOnto knowledge management and transformation system, `http://transonto.sourceforge.net`

In Section 2.2 we have also identified a problem in the DUL/IOLite conceptualization of text and its fragments, where the rigidness of the proposed structure makes it inappropriate for modelling the results of modern information extraction applications. The proposed solution is both better suited for SYNC3 and similar systems, but can also be extended to models equivalent to the current DUL/IOLite conceptualization.

As RDF repositories become larger, dynamically generating alternative facets from coordinated conceptualizations of the same data will become a key enabling technology, avoiding the need to reduplicate information at a large scale. In the case of the SYNC3 system, for example, extracting roughly 38Mtriples per month, statically storing alternative facets would impose a prohibitive burden. In Section 3 we propose a novel approach for using the customized inference architecture available in many modern RDF frameworks in order to dynamically generate alternative facets.

Future research plans involve developing a vocabulary of OWL annotation properties that can provide meta-information about the ontological schema, such as identifying the 'central' instance of a pattern that links to every other instance and the property (or chain of properties) through which this instance reaches every other instance in the pattern. This will enable the development of a generic mechanism of serving facets without reference to any two particular ontologies.

Acknowledgements

The author wishes to acknowledge the support of the European FP7-ICT project SYNC3, `http://www.sync3.eu`

References

1. Sarris, N., Potamianos, G., Renders, J.M., Grover, C., Karstens, E., Kallipolitis, L., Tountopoulos, V., Petasis, G., Krithara, A., Gallé, M., Jacquet, G., Alex, B., Tobin, R., Bounegru, L.: A system for synergistically structuring news content from traditional media and the blogosphere. In: Proceedings of the eChallenges e-2011 Conference, Florence, 26–28 October 2011. (2011)
2. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In Gómez-Pérez, A., Benjamins, V.R., eds.: Proc. 13th Intl. Conf. on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (EKAW 2002), Siguenza, Spain, 1–4 Oct. 2002. LNCS 2473, Springer Verlag, Berlin/Heidelberg (2002)
3. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F – a model of events based on the foundational ontology DOLCE+DnS Ultralite. In: Proc. 5th Intl Conf. on Knowledge Capture (K-CAP 2009), Redondo Beach, California, 1–4 September 2009, ACM (2009)

Code specific to the SYNC3 ontology is part of public SYNC3 Deliverable 4.2, September 2011. Please see `http://www.sync3.eu` or contact author.

Using Semantic Role Labeling to Extract Events from Wikipedia

Peter Exner and Pierre Nugues

Department of Computer science
Lund University
peter.exner@cs.lth.se
pierre.nugues@cs.lth.se

Abstract. Although event models and corresponding RDF vocabularies are becoming available, the collection of events still requires an initial manual encoding to produce the data. In this paper, we describe a system based on semantic parsing (SRL) to collect automatically events from text and convert them into the LODE model. Furthermore, the system automatically links extracted event properties to the external resources DBpedia and GeoNames. We applied our system to 10% of the English Wikipedia and we evaluated its performance. We managed to extract 27,500 high-confidence event instances. Although SRL is not an error-free technique, we show that it is an effective tool, as the definition of the arguments (or roles) used in our analysis and the event properties are, most of the time, nearly identical. We evaluated the results on a randomly selected sample of 100 events and we report F-measures of up to 73. The extracted events are available online from a SPARQL endpoint[1].

1 Introduction

Event models, such as EVENT[2] [1], LODE[3] [2], and SEM[4] [3], share common features to represent the agents, time, and place involved in an event. Such models are interesting because they attempt to reconcile disparate theories and standardize their representations using RDF vocabularies; for a review and a discussion, see [3]. Ideally, they should enable a variety of providers to publish any kind of events in distributed repositories, where clients would gain a uniform access to data. Applications could then embed more easily event-related information and processing. However, actual mentions of events in source materials, such as history books, newspapers, encyclopedia, etc., rarely comply with such representations. Before we can get access to wide-coverage and standardized event repositories, we need to find ways to automate their collection – i.e. their detection and extraction – as well as the identification of their properties from these source materials.

Event calendars available from websites such as Last.fm, upcoming.yahoo.com, and eventful.com, are accessible through application programming interfaces (API). These

[1] http://semantica.cs.lth.se/
[2] http://motools.sourceforge.net/event/event.html
[3] http://linkedevents.org/ontology/
[4] http://semanticweb.cs.vu.nl/2009/11/sem/

calendars provide events in a structured format in which the majority of the event properties, such as time and place, has already been extracted. Transforming these calendar data to a given event model can be done through the mapping of their format to the properties of the selected model as in [4]. Extracting events from natural language, as found on blogs and ordinary web pages, poses a greater challenge since the events are inherently unstructured.

In this paper, we introduce a system to extract events automatically from natural language using semantic parsing. We built a processing pipeline that takes raw text as input and extracts predicate–argument structures from the sentences. We used a semantic role labeler (SRL) to identify the predicates together with their core arguments or roles, such as the agent or the theme, in the sentences. The predicate arguments also include modifiers, such as temporal, locational, and manner adjuncts.

Semantic role labeling [5] is a generic technique to parse predicate–argument structures, where most of the semantic role labelers for English use statistical models trained on either Framenet [6] or Propbank [7]. Although they can reach acceptable levels of performance in terms accuracy [8, 9], semantic role labelers are often too slow to be applied to large corpora as is, or lack specificity to be used in dedicated information extraction tasks. In the context of SRL, the extracted predicate–argument structures are often called *propositions*. We will use this term in the rest of the paper.

To gather a significant set of events, we used the English Wikipedia[5] as the source material. In addition to being sizable and easy to access, Wikipedia has a large coverage of historical and cultural events that, we believe, cannot be matched by other corpora. To cope with the size of this corpus, we extended the core SRL system with a database to store the propositions and backlinks to their location in the source text. Conceptually, the extraction of events comprises four main stages:

1. Semantic parsing of Wikipedia (SRL);
2. Event selection: argument identification and property extraction;
3. Disambiguation and linking of the time and location phrases to external resources;
4. Mapping of the predicate–argument structures onto an event model.

We evaluated the performance of our system to identify and extract events. We show that SRL is an effective tool, as the definition of the Propbank arguments (or roles) used in our analysis and the event properties as described in the event models are, most of the time, nearly identical.

2 System Architecture

The architecture of our event extraction system is a pipeline of components. It consists of four main modules (Figure 1):

1. The parsing module, Athena, is a framework for large-scale parsing of text written in natural language;
2. The argument identification module that associates the predicate–argument structures extracted by the first module and relates them to a restricted set of VerbNet roles;

[5] `http://www.wikipedia.org/`

3. The property extraction and linking module that associates agent, time, and location phrases to GeoNames and DBpedia entries;
4. The conversion module that maps the structures to the LODE event model.

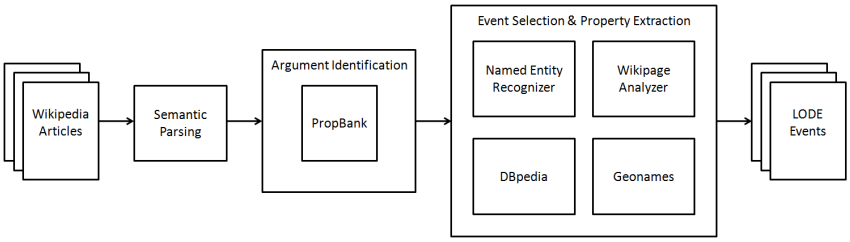


Fig. 1. Overview of the event extraction pipeline.

3 Semantic Representation of Sentences

Semantic Roles and Event Models. There are many linguistic theories on the semantic representation of sentences. Frame semantics [10] is one of the most productive that assumes that the meaning of a sentence is represented by a set of predicates and arguments. Framenet [6] and Propbank [7] are two projects that applied this theory to annotate corpora, respectively the British National Corpus and the Wall Street Journal with their predicate-argument structures. Predicates can have different senses, where each sense is associated with a specific set of arguments.

The argument annotation goes beyond the traditional subject and object and includes modifiers of the predicate, such as the temporal, locational, and manner adjuncts. These modifiers are crucial in the extraction of events since all the event models contain properties to hold the time and the place.

Figure 2 shows the predicate and the arguments contained in the sentence *In 1953, John Desmond opened the first architectural firm in Hammond.* annotated using the Propbank style. The predicate *open.01* uses the suffix 01 to denote its sense that corresponds to *open*. This differs from *open.02*, which means to *begin*. The A0 argument, *John Desmond* and the A1 argument, *the first architectural firm*, have the meanings *opener* and *thing opening* respectively for this predicate sense. The phrases *in Hammond* and *In 1953* correspond to locational and temporal modifiers, AM-LOC and AM-TMP, respectively. An ideal mapping would assign the core arguments A0 and A1 as well as the modifiers AM-LOC and AM-TMP to the agent, time, and place properties of an event model. In addition, proper nouns can be extracted, disambiguated, and linked to external resources.

The LODE Event Model. We chose the LODE event model to represent our extracted events because LODE is independent of the event domain, does not force aspect or agentivity, and makes a distinction between a named place and a geospatial space. We

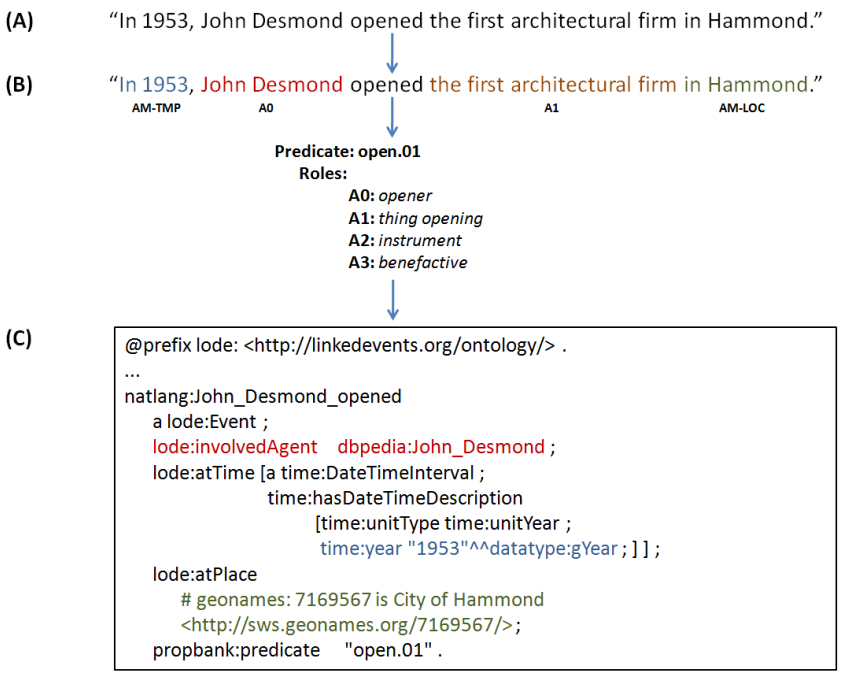


Fig. 2. (A) The starting sentence. (B) The sentence after parsing with SRL. (C) An ideal conversion to the LODE event model.

believe these features necessary for representing the wide diversity of the events found in Wikipedia. Compared to SEM, it is a minimal model that fits generally well with the conceptual nature of information contained in natural language sentences. Figure 2 shows an example of an ideal transformation of a predicate–argument structure to the LODE event model.

Semantic Parsing. As core parser, our system uses a high-performance multilingual semantic role labeler that obtained top scores in the CONLL-2009 shared task [9].

Even if SRL has made significant progress during the last ten years, it is still prone to errors especially with phrases involving proper nouns and adjuncts. This makes it more difficult to apply it to event extraction as events contain inherently more proper nouns than other sentences, both in the description of the event and the place. In addition, proper nouns like *Research in Motion* may contain words where the parser can misclassify a word as a verb, here *Research*, and then lead to a wrong predicate extraction. Furthermore, time expressions and time intervals are still not perfectly identified. Although predicate–argument extractions are not done with 100% accuracy, we still believe that SRL can be a very useful tool for event extraction and we also propose workarounds to increase the quality of the extractions.

4 Athena

Athena is a parsing framework intended to cope with large-scale multilingual information extraction. It consists of several components that fill a specific task in reading the Wikipedia text including both the English and Chinese versions, extracting, analyzing, and transforming knowledge. By using trained parsing models, the framework can be adapted to new languages without the need of reworking the extraction algorithms or patterns.

Athena reads articles from a Wikipedia database, filters, parses, and then stores the data in a semantically annotated structure. The task of parsing the entire database is parallelized using scripts, which subdivide a range of articles and launch parsing jobs applied to smaller ranges. Athena builds the proposition database by gathering the multiple small databases created during parsing and assembles them into one large database. With the use of a statistics module, the proposition database can be queried to provide statistics such as the number of and redundancy of propositions.

In our experiments, we used a subset of 10% of the English edition of Wikipedia consisting of 378,453 articles. We extracted all the sentences of all the articles and we parsed them. It resulted into 13,428,114 sentences and 53,694,899 propositions. We believe this size to be large enough to provide a significant number of propositions and events and at the same time enable us to carry out a sequence of try-and-fail experiments with an acceptable cycle time.

Mapping Predicates onto Events. Although predicates and events, such as in Probbank and LODE, have a similar structure, they are not identical. A major difference is that a set of arguments in Propbank is specific to one predicate sense, for instance the arguments of *open.01* are A0, *opener*, A1: *thing opening*, A2: *instrument*, and A3: *benefactive*, while LODE has only two universal properties, **involved** and **involvedAgent**, that correspond to these Propbank's core arguments. To cope with Propbank's diversity, a converter is necessary to map the predicate–argument structures onto the selected event model. [11] is an example of this that uses hand-generated rules or rules induced from manually-filled event templettes.

Instead of using rules that in any case would require significant manual work, we took advantage of the links between Propbank and VerbNet and we implemented a mapping module based on it. VerbNet [12] is a lexicon that builds on Levin's classification of English verbs [13]. Verb classes are described using a limited set of 23 roles used across all the lexicon and where each predicate role is constrained using selectional restrictions such as animate, comestible, etc. Although not complete, 11,500 arguments in Propbank have a correspondence with VerbNet thematic roles, making the conversion possible.

5 Selecting Event Propositions

We built our event set from the complete proposition output produced from Wikipedia. We considered that a proposition could fit an event if it contained a date, a place, and an agent. For a discussion on the aspects of event classification, see [14]. We used the links

associating the Propbank arguments to the VerbNet thematic roles and we extracted the propositions whose arguments matched a time, a place, and agents in the VerbNet structure. We used the following rules:

- We identified an agent from a Propbank argument when it could be associated with one the following VerbNet thematic roles: *Actor*, *Agent*, *Beneficiary*, *Experiencer*, *Recipient*, and *Theme*. If no such roles were found, we selected the *A0* argument as default.
- Similarly, we identified the places using the *Location* and *Source* VerbNet thematic roles. We also included the *AM-LOC* modifier.
- We could not find arguments in PropBank linked to the *Time* VerbNet thematic role. We therefore selected the arguments containing dates and times using the *AM-TMP* modifier.

These events were further filtered by selecting propositions having at least one extracted time, place, and agent property. Using a quick manual examination, we could observe that this very simple filtering enabled us to discard a large set of less reliable propositions.

6 Converting Propositions to Event Models

Following the argument identification, we extracted entities corresponding to the LODE ontology properties using regular expressions, a local subset of the DBpedia database [15], and the GeoNames web service[6].

Aspectual Verbs. We grouped pairs of predicates that begins with an aspectual verb, such as in *began working* or *stopped singing*. This grouping was performed when the second predicate together with all of its arguments formed a subset of the arguments of the first predicate. Figure 3 shows an example of it, where the arguments of the predicate *work.01* form a subset of the arguments of the predicate *begin.01*. Thus, the two predicates are grouped to form the event, *began working*.

Fig. 3. Parsing output showing an example of a sentence, where we group two predicates: *In the late 1990s NASA and Google began working on a new network protocol*. The semantic parser is accessible from `http://barbar.cs.lth.se:8081/`.

Single predicates and predicate groups are assigned to the *propbank* RDF property.

[6] `http://www.geonames.org/`

Converting Involved Agents. When possible, we linked the LODE arguments to DBpedia entries. This enabled us to integrate the data we produce with other types of structured information extracted from Wikipedia and from other sources. Eventually, this should improve interoperability of data sources and make it easier to build comprehensive applications.

To detect the entries, we applied a named entity tagger[7] [16] to the arguments extracted from the VerbNet thematic roles. We then selected entities representing organizations and persons as agent candidates. We used a subset of the DBpedia database containing infobox types, Wikipedia redirects, and Wikipedia page links to carry out the final name disambiguation.

Candidates are disambiguated and linked to their corresponding DBpedia entry by one of the following rules in this order:

1. When an infobox type matches the candidate phrase, we use this type. For instance the phrase *United Nations* is resolved directly to the DBpedia resource `<http://dbpedia.org/resource/United_Nations>`;
2. When a redirection is found for the candidate phrase, we use this redirection. As an example, the phrase *United States Supreme Court* is resolved to the DBpedia resource `<http://dbpedia.org/resource/Supreme_Court_of_the_United_States>` by using DBpedia page redirects;
3. When outgoing DBpedia resources from the originating Wikipedia article contain the candidate phrase, we use the most frequent resource. For example, if we wish to resolve the word *Loren* in the Wikipedia article `http://en.wikipedia.org/wiki/Carlo_Ponti` to a DBpedia resource, we start from the originating article and consider only outgoing DBpedia resources that contain the sought phrase. We find that the DBpedia resource, `<http://dbpedia.org/resource/Sophia_Loren>`, is mentioned three times in the article and we select this as the resolved resource for the word *Loren*;
4. When labels of outgoing Wikipedia links from the originating Wikipedia article also contain the candidate phrase, the corresponding targets are selected and resolved using the rules above. In this case, the outgoing DBpedia resources do not contain the candidate phrase and the labels of the outgoing Wikipedia links are searched instead. Using this technique, *Edith Somerville* is resolved to the DBpedia resource `<http://dbpedia.org/resource/Edith_Anna_Somerville>` in the Wikipedia article `http://en.wikipedia.org/wiki/Violet_Florence_Martin`;
5. When the title of the originating Wikipedia article contains the candidate phrase, it is selected and resolved using the rules above. For instance, *Weis* is resolved to the DBpedia resource `<http://dbpedia.org/resource/Weis_Markets>` in the Wikipedia article `http://en.wikipedia.org/wiki/Weis_Markets`.

If a DBpedia entry is found, it is assigned the *involvedAgent* property in the LODE model.

[7] `http://nlp.stanford.edu/ner/index.shtml`

Converting Place and Space. Similarly to the extraction of the involved agents, we identify arguments corresponding to the VerbNet thematic roles associated with locations. Following named entity extraction, entities representing locations are queried using the GeoNames web service. The first matching result is selected and the GeoNames identifier is assigned to the *atPlace* property. Entities representing organizations are identified and linked to DBpedia entries by using the methods 1 to 3 described in Section *Converting Involved Agents*.

Converting Time. The conversion to the LODE *atTime* property is carried out in 3 steps: We first identify the arguments containing date and time phrases; We then extract the time entities from the arguments using the named entity tagger; And we finally convert the time entities to the OWL DateTimeInterval format using of common date format patterns. In addition, we discard time phrases without an anchoring date expression, such as *Three days ago*. Our extraction module identifies the first occurring date expression and assigns it to the *atTime* event property.

Storing Events. The extracted events are saved to files in the Notation 3 format. The file names contain the Wikipedia article title, followed by the absolute line number of the sentence from which the event was extracted. This structure enables the backlinking from the event to the originating source material.

7 Experimental Results

In our evaluation, we sought to answer the questions: How much of the information in a sentence can be extracted and moved into an event model? And, which properties are the most difficult to extract? Since we did not have the precision or the recall of events in the source text, we omitted the evaluation of event identification and instead we focused on calculating the precision and recall of the identified and extracted events. We approached this task by computing the recall and precision of the individual properties of our extracted events and counting the error sources.

In total, we extracted 27,594 events from our subset of 378,453 English Wikipedia articles. We created our data set by randomly selecting a sample of 100 events from our extracted events. In order to calculate precision and recall, we calculated the number of retrieved atTime, atPlace, involvedAgent, and predicate properties in each sampled event. We examined the sentence corresponding to the sampled event to find the number of relevant properties.

We used two metrics to assess the properties using a strict and a relaxed criterion. We marked the atTime property as strictly correct if all the date components were extracted, and as relaxed correct, if the most significant date component was extracted. We marked the atPlace property as correct if the extracted reference to GeoNames or DBpedia was resolved to a correct entry. We made no distinction between strictly correct and relaxed correct for atPlace. Similarly, we marked InvolvedAgent as correct, if the property resolved to a correct DBpedia entry. Finally, we marked the predicate

property as strictly correct if both the corresponding verb and sense had the correct semantics and as relaxed correct regardless of the predicate sense. During evaluation, we also counted the properties causing the errors.

Based on these evaluations, we calculated the precision, recall, and the F1 score for our sample data set (Table 1, left). Table 1, right, shows the relative percentage of error sources categorized by extracted properties.

	Precision	Recall	F1
Strict	70.8	71.6	71.2
Relaxed	72.8	73.6	73.2

Error sources	
Agent	40.9%
Place	36.9%
Time	11.4%
Predicate	10.7%

Table 1. Left table: The precision, recall, and F1 score for the sampled events. **Right table**: Sources of errors.

From Table 1, we can observe that the largest percentages of error come from the agents and places. The reasons for these extraction failures can be attributed to the following causes:

- The arguments containing the agents were not found by the semantic parser.
- Ambiguity of the extracted proper nouns.
- Unresolved pronouns.
- Lack of DBpedia entry corresponding to the agent.

We believe that in many cases the ambiguity of the agents can be resolved by using a larger subset of DBpedia databases and thereby classifying the type of the agents. Together with a more explorative selection of arguments, we believe this may lead to a larger amount of correctly extracted agents.

Since we only extracted the first detected date, this caused the majority of failures in date extraction. We believe that date extraction can be improved using more extraction patterns.

8 Conclusions

In this paper, we investigated semantic parsing to extract events from text. We implemented a processing pipeline consisting of a high-performance semantic role labeler to extract predicate–argument structures and a converter using VerbNet thematic roles to produce events in the LODE RDF format. Using 10% of the English Wikipedia and simple filtering rules, we managed to sift more than 27,500 high-confidence instances. We evaluated the results on a randomly selected sample of 100 events and we report F-measures ranging from 71.2 to 73.2.

Misidentified agents are a frequent source of error. We believe such errors can be significantly reduced by improving the detection of proper nouns. This could be done by applying a preprocessing step to detect the named entities or using databases of proper

nouns and retraining the semantic parser on it. We could also improve the detection using a coreference solver that would tie pronouns such as *she*, *he*, or *it* to person or organization names. In the future, we also plan to parse the complete Wikipedia corpus in English and other languages.

An archive of extracted events is available for download as well as accessible from a SPARQL endpoint[8].

Acknowledgments. This research was supported by Vetenskapsrådet, the Swedish research council, under grant 621-2010-4800 and has received funding from the European Union's seventh framework program (FP7/2007-2013) under grant agreement 230902.

References

1. Raimond, Y., Abdallah, S., Sandler, M., Giasson, F.: The music ontology. In: Proceedings of the International Conference on Music Information Retrieval, Vienna (2007)
2. Shaw, R.B.: Events and Periods as Concepts for Organizing Historical Knowledge. PhD thesis, University of California, Berkeley (2010)
3. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (SEM). Journal of Web Semantics (2011) In press
4. Liu, X., Troncy, R., Huet, B.: Finding media illustrating events. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ICMR '11 (2011) 58:1–58:8
5. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics **28** (2002) 245–288
6. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: Framenet ii: Extended theory and practice. `http://framenet.icsi.berkeley.edu/` (2010)
7. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: an annotated corpus of semantic roles. Computational Linguistics **31** (2005) 71–105
8. Johansson, R., Nugues, P.: Dependency-based syntactic–semantic analysis with PropBank and NomBank. In: Proceedings of CoNLL-2008, Manchester (2008) 183–187
9. Björkelund, A., Hafdell, L., Nugues, P.: Multilingual semantic role labeling. In: Proceedings of CoNLL-2009, Boulder (2009) 43–48
10. Fillmore, C.J.: Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech **280** (1976) 20–32
11. Surdeanu, M., Harabagiu, S., Williams, J., Aarseth, P.: Using predicate-argument structures for information extraction. In: Proc. of the 41st Annual Meeting of the ACL. (2003) 8–15
12. Kipper-Schuler, K.: VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania, Philadelphia (2005)
13. Levin, B.: English verb classes and alternations: A preliminary investigation. University of Chicago Press, Chicago (1993)
14. Shaw, R., Troncy, R., Hardman, L.: LODE: Linking open descriptions of events. In: 4th Asian Semantic Web Conference. (2009) 153–167
15. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia—a crystallization point for the web of data. Journal of Web Semantics (2009) 154–165
16. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proc. of the 43rd Annual Meeting of the ACL. (2005) 363–370

[8] `http://semantica.cs.lth.se/`

An Overview of Event Extraction from Text

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

{fhogenboom, frasincar, kaymak, fdejong}@ese.eur.nl

Abstract. One common application of text mining is event extraction, which encompasses deducing specific knowledge concerning incidents referred to in texts. Event extraction can be applied to various types of written text, e.g., (online) news messages, blogs, and manuscripts. This literature survey reviews text mining techniques that are employed for various event extraction purposes. It provides general guidelines on how to choose a particular event extraction technique depending on the user, the available content, and the scenario of use.

1 Introduction

With the increasing amount of data and the exploding number of digital data sources, utilizing extracted information in decision making processes becomes increasingly urgent and difficult. An omnipresent problem is the fact that most data is initially unstructured, i.e., the data format loosely implies its meaning [9] and is described using natural, human-understandable language, which makes the data limited in the degree in which it is machine-interpretable. This problem thwarts the automation of for example vital information retrieval (IR) and information extraction (IE) processes – used for decision making – when involving large amounts of data.

Text Mining (TM) [15] is concerned with information learning from preprocessed text (e.g., containing identified parts of speech or stemmed words). By means of text mining, often using Natural Language Processing (NLP) [22] techniques, information is extracted from texts of various sources, such as news messages and blogs, and is represented and stored in a structured way, e.g., in databases. A specific type of knowledge that can be extracted from text by means of TM is an event, which can be represented as a complex combination of relations linked to a set of empirical observations from texts.

An example of an event is an acquisition. If we consider the representation `<Company> <Buy> <Company>`, words identified in text referring to companies are linked to the concept `<Company>`, and (conjugations of) verbs having the meaning of acquisition are associated with `<Buy>`. Representations of this event can be extracted from news message headers such as "*Google acquires Picnik*", "*Lala bought by Apple*", or "*Skype sold to Microsoft*".

Event extraction from unstructured data such as news messages could be beneficial for IE systems in various ways. For instance, being able to determine events could enhance the performance of personalized news systems [2, 10], as news messages can be selected more accurately, based on user preferences and identified topics (or events). Furthermore, events can be useful in risk analysis applications [3], monitoring systems [17], and decision making support tools [36].

Extracted events are also extensively applied within the medical domain [6, 38], where event parsers are utilized for extracting medical or biological events like molecular events from corpora. Another possible – but less researched – application of event extraction lies within the field of algorithmic trading, representing the use of computer programs for entering trade orders with algorithms deciding aspects like timing, price, and quantity of an order. Financial markets are extremely sensitive to breaking news [24]. Economic events like mergers and acquisitions [31], stock splits [14], dividend announcements [23], etc., play a crucial role in the daily decisions taken by brokers, where brokers can be humans or machines. Besides being able to process news faster, machines are able to deal with larger volumes of emerging news, having access to more information than we humans do, and thus making better informed decisions.

Given the promising potential for applications of event extraction, and assuming that the challenges of real-time extraction and combination of events can be tackled adequately, it is worthwhile to investigate which text mining techniques are appropriate for this purpose. The current body of literature is lacking a high-level survey on event detection in text. Therefore, the goal of this paper is to review existing approaches to event extraction from text. We aim for providing general guidelines on selecting the proper text mining techniques for specific event extraction tasks, taking into account the user and its context. For this, we strive for a similar overview of performance aspects and recommendations as has been developed for cross-lingual research systems [25]. The work presented herein is a first step, focussing specifically on event extraction from text. The recognition of the space and time event dimension in text is considered outside the scope of this paper.

Throughout this paper we evaluate event extraction approaches using several criteria. For this, we review data that are available in the literature and distinguish between the categories high, medium, and low. First of all, we investigate the amount of data needed for each approach. Moreover, the amount of required domain knowledge is evaluated, together with the required amount of user expertise. Finally, we also discuss the interpretability of the results.

This paper continues with an elaboration of approaches to event extraction in Section 2. Subsequently, Section 3 presents a discussion on the event extraction approaches introduced in this survey. Finally, Section 4 concludes the paper.

2 Event Extraction

We distinguish between three main approaches to event extraction, in analogy with the classic distinction that is made in the field of modeling. First, there

are data-driven approaches, described in Section 2.1, which aim to convert data to knowledge through the usage of statistics, machine learning, linear algebra, etc. Second, we distinguish expert knowledge-driven methods as discussed in Section 2.2, which extract knowledge through representation and exploitation of expert knowledge, usually by means of pattern-based approaches. Finally, the hybrid event extraction approaches elaborated on in Section 2.3 combine knowledge and data-driven methods.

2.1 Data-Driven Event Extraction

Data-driven approaches are commonly used for natural language processing applications. These approaches rely solely on quantitative methods to discover relations. Data-driven approaches require large text corpora in order to develop models that approximate linguistic phenomena. Furthermore, data-driven text mining is not restricted to basic statistical reasoning based on probability theory, but encompasses all quantitative approaches to automated language processing, such as probabilistic modeling, information theory, and linear algebra.

One could distinguish between many approaches, e.g., word frequency counting, ranking by means of the Term Frequency – Inverse Document Frequency metric, word sense disambiguation, n-grams, and clustering. Despite their differences, all approaches focus on discovering statistical relations, i.e., facts that are supported by statistical evidence. Examples of discovered facts are words or concepts that are (statistically) associated with one another. However, statistical relations do not necessarily imply semantically valid relations, nor relations that have proper semantic meaning.

Several examples of the usage of data-driven text mining approaches for event extraction can be found in literature. For instance, in their 2009 work, Okamoto et al. [27] elaborate on a framework for detection of occasional or local events, which employs hierarchical clustering techniques. While clustering itself could already yield promising results for event extraction, the authors of [21] make use of a combination of weighted undirected bipartite graphs and clustering in order to extract key entities and significant events from daily web news. Clustering techniques are also employed by Tanev et al. [34], who also aim for real-time news event extraction, but focus especially on violence and disaster events. The authors make use of automatic tagging of words and the presented framework is designed to automatically learn patterns from discovered events. Lastly, the authors of [19] also employ word-based statistical text mining in their work from 2005. The authors elaborate on a framework aimed at news event detection, based on support vector machines.

A drawback of the discussed data-driven methods to event extraction is that they do not deal with meaning explicitly, i.e., they discover relations in corpora without considering semantics. Another disadvantage of statistics-based text mining is that a large amount of data is required in order to get statistically significant results. However, since these approaches are not based on knowledge, neither linguistic resources, nor expert (domain) knowledge are required.

2.2 Knowledge-Driven Event Extraction

In contrast to data-driven methods, knowledge-driven text mining is often based on patterns that express rules representing expert knowledge. It is inherently based on linguistic and lexicographic knowledge, as well as existing human knowledge regarding the contents of the text that is to be processed. This alleviates problems with statistical methods regarding meaning of text. Information is mined from corpora by using predefined or discovered linguistic patterns, which can be either lexico-syntactic patterns [11, 12] or lexico-semantic patterns [2]. The former patterns combine lexical representations and syntactical information with regular expressions, whereas the latter patterns also make use of semantic information. Semantics are usually added by means of gazetteers, which use the linguistic meaning of text [7, 8], or by means of ontologies [10, 32].

Several attempts have been made for extracting events using pattern-based approaches to text mining. Both – mostly manually created – lexico-syntactic and lexico-semantic patterns are used; the former more often than the latter. For instance, in their 2009 work, Nishihara et al. [26] extract personal experiences from blogs by means of three keywords (place, object, and action) that together describe an event. For this, sentences are split using lexico-syntactic patterns. A similar approach can be found in [1], where the authors focus on pattern-based relation and event extraction. Here, lexico-syntactic patterns are employed in order to discover a wide range of relations and events in the domains of finance and politics. The authors of [38] elaborate on a methodology to extract events using a general-purpose parser and grammar applied to the biomedical domain. To this extent, lexico-syntactic patterns are employed that define the argumentation structures within texts. Hung et al. [13] elaborate on a framework that can be employed for mining the Web for event-based commonsense knowledge by using lexico-syntactic pattern matching and semantic role labeling. A large number of raw sentences that possibly contain target knowledge is collected through Web search engines. Web queries are formulated based on a set of lexico-syntactic patterns. After labeling the semantic roles, i.e., defining the relationships that syntactic arguments have with verbs, knowledge is extracted and stored in a database. A final example of the employment of lexico-syntactic patterns can be found in the work of Xu et al. [37]. Here, the authors envisage the usage of lexico-syntactic patterns in order to learn patterns from texts on prize award events, in the form of bootstrapping-oriented unsupervised machine learning, initialized with lexico-syntactic pattern seeds.

In pattern-based event extraction, concepts that have specific meanings and/or relationships are required, but either they are not available or they are not used due to the lack of pattern expressivity (i.e., in lexico-syntactic patterns). To solve this, lexico-semantic patterns are employed. These patterns are used for various purposes. In an attempt to discover event patterns from stock market bulletins, the authors of [20] analyze tagged corpora by means of gazetteering semantic concepts that are based on a (financial) domain. Cohen et al. [6] employ a concept recognizer on a biological domain in order to extract medical events from corpora, thus taking into account the semantics of domain concepts.

A similar approach is used by Vargas-Vera and Celjuska [35], who propose a framework for event recognition, focusing on Knowledge Media Institute (KMi) news articles. The framework aims for learning and applying lexico-semantic patterns. The extracted information is utilized to populate a knowledge base. Lastly, Capet et al. [3] present a methodology aimed at event extraction for an automated early warning system. The authors employ lexico-semantic patterns for concept matching using dependency chains enhanced using lexicons (word lists), so that concepts are matched whenever syntactically related chains of expressions conveying their constituent concepts occur within the same sentence.

Several advantages stem from the utilization of pattern-based approaches to event extraction. Firstly, pattern-based approaches need less training data than data-driven approaches. Also, it is possible to define powerful expressions by using lexical, syntactical, and semantic elements, and results are easily interpretable and traceable. Patterns are useful when one needs to extract very specific information. However, in order to be able to define patterns that retrieve the correct, desired information, lexical knowledge and possibly also prior domain knowledge is required. Other disadvantages are related to defining and maintaining patterns, as knowledge acquisition is made more difficult (e.g., in costs and consistency) when patterns need to be scaled-up to cover more situations [33] due to the fact that patterns are usually hand-tuned.

2.3 Hybrid Event Extraction

Despite the advantages of both data-driven and knowledge-driven approaches to event extraction, in practice, it is difficult to stay within the boundaries of a single event extraction approach. As both approaches have their disadvantages, combining the two methods could yield the best results. In general, an approach can be viewed as mainly data or knowledge-driven. However, there is an increasing number of researchers that equally combine both approaches, and thus in fact employ hybrid approaches. For instance, it is hard to apply solely pattern-based algorithms successfully, as these algorithms often need for instance bootstrapping or initial clustering, which can be done by means of statistics [29]. Hybrid approaches could emerge when solving the lack of expert knowledge for pattern-based approaches, by applying statistical methods [5]. Also, researchers can combine statistical approaches with (lexical) knowledge, e.g. to prevent unwanted results [28] or to reinforce statistical methods [30].I addition, you can also constrain the learning methods (i.e. data-driven approaches) by using expert knowledge so that a better model is learnt more easily.

In IE literature, many hybrid approaches to text mining are described for extracting events. Most systems are knowledge-driven methods that are aided by data-driven methods, and thus frequently solve the lack of expert knowledge or apply bootstrapping to boost extraction performances, e.g., in terms of precision and recall. For instance, Jungermann and Morik [16] combine lexico-syntactic patterns with conditional random fields (depicted as undirected graphs), in order to extract events from the minutes of plenary sessions of the German parliament. An example of bootstrapping lexical techniques with statistics is given

in [29]. Here, the authors bootstrap a weakly supervised pattern learning algorithm with clusters, in order to be able to extract violelence incidents from online news with high precision and recall, as well as storing these in knowledge bases. Chun et al. [4] extract events from biomedical literature by means of lexico-syntactic patterns, combined with term co-occurrences. Finally, aiming for ontology-based fuzzy event extraction for Chinese e-news summarization, the authors of [18] employ a grammar-based statistical method to text mining, i.e., part-of-speech tagging. However, tagging is based on domain knowledge that is stored in ontologies, thus making the event extraction a hybrid process.

In hybrid event extraction systems, due to the usage of data-driven methods, the amount of required data increases, yet typically remains less than is the case with purely data-driven methods. Compared to a knowledge-driven approach, complexity – and hence required expertise – increases due to the combination of multiple techniques. On the other hand, the amount of expert knowledge that is needed for effective and efficient event discovery is generally less than for pattern-based methods, because of the fact that lack of domain knowledge can be compensated by the use of statistical methods. As for the interpretability, attributing results to specific parts of the event extraction is more difficult due to the addition of data-driven methods. Yet, interpretability still benefits from the use of semantics. Disadvantages of hybrid approaches are mostly related to the multidisciplinary aspects of hybrid systems.

3 Discussion

Table 1 provides a summary of the methods discussed, by combining the results from the discussions in Section 2. Per approach elaborated on in this paper, the employed methods and the type of events that are discovered are summarized. Also, the minimum amount of required data and required domain knowledge and expertise are included, as well as the interpretability of the results.

From the results presented in this table, we derive that in terms of data usage, knowledge-driven event extraction methods require the least amount of data (i.e., experiments are performed on a couple of hundreds of documents or sentences). Data-driven methods on the other hand often make use of more than ten thousand documents. Hybrid methods generally report results on a maximum of ten thousand documents. As for interpretability, i.e., the ease with which the (intermediate) results can be translated to a human-understandable format, data-driven methods perform worst. Knowledge-driven methods on the other hand score higher on interpretability. Especially lexico-semantic pattern approaches have a high level of interpretability, as patterns can easily be translated into natural language, while lexico-syntactic patterns require more effort. Finally, when considering the amount of expert domain knowledge and expertise needed for each approach, data-driven methods require less of both than hybrid and knowledge-driven methods.

As a general guideline for selecting a suitable technique for event extraction, based on the results of our survey, we suggest the usage of knowledge-based

Table 1. Overview of the approaches discussed, displaying the method (*Method*) and the type of events that are discovered (*Events*). Also, the amount of required data (*Data*) is depicted, as well as required domain knowledge and expertise (*Know.* and *Exp.*, respectively), and the interpretability of the results (*Int.*). Note that the reported values in the last four columns are lower bounds.

Technique	Approach	Method	Events	Data	Know.	Exp.	Int.
Data	Okamoto et al. [27]	Hierarchical clustering	Local	Med	Low	Low	Low
	Liu et al. [21]	Graphs, clustering	News	High	Low	Low	Low
	Tanev et al. [34]	Clustering	Violent and disaster news	Med	Low	Low	Low
	Lei et al. [19]	Support Vector Machines	News	High	Low	Low	Low
Knowledge	Nishihara et al. [26]	Lexico-Syntactic	Personal experiences	Low	Med	High	Med
	Aone et al. [1]	Lexico-Syntactic	General	Low	High	High	Med
	Yakushiji et al. [38]	Lexico-Syntactic	Biomedical	Low	Med	High	Med
	Hung et al. [13]	Lexico-Syntactic	Commonsense knowledge	Low	Med	High	Med
	Xu et al. [37]	Lexico-Syntactic	Prize award	Low	Med	High	High
	Li et al. [20]	Lexico-Semantic	Financial	Low	High	High	Med
	Cohen et al. [6]	Lexico-Semantic	Biomedical	Med	High	High	High
	Vargas-Vera et al. [35]	Lexico-Semantic	KMi news	Low	High	High	High
	Capet et al. [3]	Lexico-Semantic	Early warning	Low	High	High	High
Hybrid	Jungermann et al. [16]	Lexico-Syntactic, graphs	German parliament	Med	Med	High	Med
	Piskorski et al. [29]	Lexico-Semantic, clustering	Violent news	High	Med	Med	Med
	Chun et al. [4]	Lexico-Syntactic, co-occurences	Biomedical	Med	Med	Med	Med
	Lee et al. [18]	Ontology-based Part-Of-Speech tagging	Chinese news	N/A	Med	Med	Low

techniques for casual users (e.g., students) that prefer an interactive, query-driven approach to event extraction, assuming domain knowledge and expertise to be readily available. Users can easily specify patterns in a language that is close to their own natural language, without being bothered with statistical details and model fine-tuning. On the other hand, users like (academic) researchers would benefit from both hybrid and data-driven approaches, as these are less restricted by, for example, grammars.

4 Conclusions

In this paper, we investigated the main approaches to event extraction from text that are elaborated on in the current body of literature. Overall, data-driven methods require many data and little domain knowledge and expertise, while having a low interpretability. Conversely, for knowledge-based event extraction little data is required, but domain knowledge and expertise is needed. These approaches generally offer a higher traceability of the results. Finally, hybrid approaches seem to be a compromise between data and knowledge-driven approaches, requiring a medium amount of data and domain knowledge and offering medium interpretability. However, it should be noted that the amount of expertise needed is high, due to the fact that multiple techniques are combined. As a guideline, we advise knowledge-driven techniques for casual and novice users, whereas data-driven are more suitable for advanced users.

References

1. Aone, C., Ramos-Santacruz, M.: REES: A Large-Scale Relation and Event Extraction System. In: 6th Applied Natural Language Processing Conference (ANLP 2000). pp. 76–83. Association for Computational Linguistics (2000)
2. Borsje, J., Hogenboom, F., Frasincar, F.: Semi-Automatic Financial Events Discovery Based on Lexico-Semantic Patterns. International Journal of Web Engineering and Technology 6(2), 115–140 (2010)
3. Capet, P., Delavallade, T., Nakamura, T., Sandor, A., Tarsitano, C., Voyatzi, S.: Intelligent Information Processing IV, IFIP International Federation for Information Processing, vol. 288, chap. A Risk Assessment System with Automatic Extraction of Event Types, pp. 220–229. Springer Boston (2008)
4. Chun, H.W., Hwang, Y.S., Rim, H.C.: Unsupervised Event Extraction from Biomedical Literature Using Co-occurrence Information and Basic Patterns. In: 1st International Joint Conference on Natural Language Processing (IJCNLP 2004). Lecture Notes in Computer Science, vol. 3248, pp. 777–786. Springer-Verlag Berlin Heidelberg (2004)
5. Cimiano, P., Staab, S.: Learning by Googling. SIGKDD Explorations Newsletter 6(2), 24–33 (2004)
6. Cohen, K.B., Verspoor, K., Johnson, H.L., Roeder, C., Ogren, P.V., Baumgartner, Jr., W.A., White, E., Tipney, H., Hunter, L.: High-Precision Biological Event Extraction with a Concept Recognizer. In: Workshop on BioNLP: Shared Task collocated with the NAACL-HLT 2009 Meeting. pp. 50–58. Association for Computational Linguistics (2009)

7. Cunningham, H.: GATE, a General Architecture for Text Engineering. Computers and the Humanities 36(2), 223–254 (2002)
8. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002). pp. 168–175. Association for Computational Linguistics (2002)
9. Ferrucci, D., Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering 10(3–4), 327–348 (2004)
10. Frasincar, F., Borsje, J., Levering, L.: A Semantic Web-Based Approach for Building Personalized News Services. International Journal of E-Business Research 5(3), 35–53 (2009)
11. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th Conference on Computational Linguistics (COLING 1992). vol. 2, pp. 539–545 (1992)
12. Hearst, M.A.: WordNet: An Electronic Lexical Database and Some of its Applications, chap. Automated Discovery of WordNet Relations, pp. 131–151. MIT Press (1998)
13. Hung, S.H., Lin, C.H., Hong, J.S.: Web Mining for Event-Based Commonsense Knowledge Using Lexico-Syntactic Pattern Matching and Semantic Role Labeling. Expert Systems with Applications 37(1), 341–347 (2010)
14. Ikenberry, D.L., Ramnath, S.: Underreaction to Self-selected News Events: The Case of Stock Splits. Review of Financial Studies 15(2), 489–526 (2002)
15. Ikonomakis, M., Kotsiantis, S., Tampakas, V.: Text Classification Using Machine Learning Techniques. WSEAS Transactions on Computers 4(8), 966–974 (2005)
16. Jungermann, F., Morik, K.: Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining. In: 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems (NLDB 2008). Lecture Notes in Computer Science, vol. 5039, pp. 335–336. Springer-Verlag Berlin Heidelberg (2008)
17. Kamijo, S., Matsushita, Y., Ikeuchi, K., Sakauchi, M.: Traffic monitoring and accident detection at intersections. IEEE Transactions on Intelligent Transportation Systems 1(2), 108–118 (2000)
18. Lee, C.S., Chen, Y.J., Jian, Z.W.: Ontology-Based Fuzzy Event Extraction Agent for Chinese E-News Summarization. Expert Systems with Applications 25(3), 431–447 (2003)
19. Lei, Z., Wu, L.D., Zhang, Y., Liu, Y.C.: A System for Detecting and Tracking Internet News Event. In: 6th Pacific-Rim Conference on Multimedia (PCM 2005). Lecture Notes in Computer Science, vol. 3767, pp. 754–764. Springer-Verlag Berlin Heidelberg (2005)
20. Li, F., Sheng, H., Zhang, D.: Event Pattern Discovery from the Stock Market Bulletin. In: 5th International Conference on Discovery Science (DS 2002). Lecture Notes in Computer Science, vol. 2534, pp. 35–49. Springer-Verlag Berlin Heidelberg (2002)
21. Liu, M., Liu, Y., Xiang, L., Chen, X., Yang, Q.: Extracting Key Entities and Significant Events from Online Daily News. In: 9th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2008). Lecture Notes in Computer Science, vol. 5326, pp. 201–209. Springer-Verlag Berlin Heidelberg (2008)
22. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, 1st edn. (1999)

23. Michaely, R., Thaler, R.H., Womack, K.L.: Price Reactions to Dividend Initiations and Omissions: Overreaction or Drift? Journal of Finance 50(2), 573–608 (1995)
24. Mitchell, M.L., Mulherin, J.H.: The Impact of Public Information on the Stock Market. Journal of Finance 49(3), 923–950 (1994)
25. Moreau, N.: Best Practices in Language Resources for Multilingual Information Access. Tech. rep., TrebleCLEF Consortium (2009), From: http://www.trebleclef.eu/getfile.php?id=255
26. Nishihara, Y., Sato, K., Sunayama, W.: Event Extraction and Visualization for Obtaining Personal Experiences from Blogs. In: Symposium on Human Interface 2009 on Human Interface and the Management of Information. Information and Interaction. Part II. Lecture Notes in Computer Science, vol. 5618, pp. 315–324. Springer-Verlag Berlin Heidelberg (2009)
27. Okamoto, M., Kikuchi, M.: Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries. In: 5th Asia Information Retrieval Symposium (AIRS 2009). Lecture Notes in Computer Science, vol. 5839, pp. 181–192. Springer-Verlag Berlin Heidelberg (2009)
28. Pakhomov, S.: Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). pp. 160–167. Association for Computational Linguistics (2002)
29. Piskorski, J., Tanev, H., Wennerberg, P.O.: Extracting Violent Events From On-Line News for Ontology Population. In: 10th International Conference on Business Information Systems (BIS 2007). Lecture Notes in Computer Science, vol. 4439, pp. 287–300. Springer-Verlag Berlin Heidelberg (2007)
30. Punyakanok, V., Roth, D., Yih, W.: The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. Computational Linguistics 34(2), 257–287 (2008)
31. Rosen, R.J.: Merger Momentum and Investor Sentiment: The Stock Market Reaction to Merger Announcements. Journal of Business 79(2), 987–1017 (2006)
32. Schouten, K., Ruijgrok, P., Borsje, J., Frasincar, F., Levering, L., Hogenboom, F.: A Semantic Web-Based Approach for Personalizing News. In: 25th Symposium On Applied Computing (SAC 2010). pp. 854–861. ACM (2010)
33. Su, K.Y., Chiang, T.H., Chang, J.S.: An Overview of Corpus-Based Statistics-Oriented (CBSO) Techniques for Natural Language Processing. Computational Linguistics and Chinese Language Processing 1(1), 101–157 (1996)
34. Tanev, H., Piskorski, J., Atkinson, M.: Real-Time News Event Extraction for Global Crisis Monitoring. In: 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems (NLDB 2008). Lecture Notes in Computer Science, vol. 5039, pp. 207–218. Springer-Verlag Berlin Heidelberg (2008)
35. Vargas-Vera, M., Celjuska, D.: Event Recognition on News Stories and Semi-Automatic Population of an Ontology. In: 3rd IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004). pp. 615–618 (2004)
36. Wei, C.P., Lee, Y.H.: Event detection from Online News Documents for Supporting Environmental Scanning. Decision Support Systems 36(4), 385–401 (2004)
37. Xu, F., Uszkoreit, H., Li, H.: Automatic Event and Relation Detection with Seeds of Varying Complexity. In: AAAI Workshop on Event Extraction and Synthesis (2006)
38. Yakushiji, A., Tateisi, Y., Miyao, Y.: Event Extraction from Biomedical Papers using a Full Parser. In: 6th Pacific Symposium on Biocomputing (PSB 2001). pp. 408–419 (2001)

Crowdsourcing Event Detection in YouTube Videos

Thomas Steiner[1], Ruben Verborgh[2], Rik Van de Walle[2],
Michael Hausenblas[3], and Joaquim Gabarró Vallés[1]

[1] Universitat Politécnica de Catalunya – Department LSI
08034 Barcelona, Spain
{tsteiner,gabarro}@lsi.upc.edu
[2] Ghent University – IBBT, ELIS – Multimedia Lab
Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium
{ruben.verborgh,rik.vandewalle}@ugent.be
[3] DERI, NUI Galway IDA Business Park, Lower Dangan Galway, Ireland
michael.hausenblas@deri.org

Abstract. Considerable efforts have been put into making video content on the Web more accessible, searchable, and navigable by research on both textual and visual analysis of the actual video content and the accompanying metadata. Nevertheless, most of the time, videos are opaque objects in websites. With Web browsers gaining more support for the HTML5 <video> element, videos are becoming first class citizens on the Web. In this paper we show how events can be detected on-the-fly through crowdsourcing (i) textual, (ii) visual, and (iii) behavioral analysis in YouTube videos, at scale. The main contribution of this paper is a generic crowdsourcing framework for automatic and scalable semantic annotations of HTML5 videos. Eventually, we discuss our preliminary results using traditional server-based approaches to video event detection as a baseline.

1 Introduction

Official statistics [26] from YouTube—owned by Google and one of the biggest online video platforms—state that more than 13 million hours of video were uploaded during 2010, and that 48 hours of video are uploaded every single minute. Given this huge and ever increasing amount of video content, it becomes evident that advanced search techniques are necessary in order to retrieve the few needles from the giant haystack. Closed captions allow for keyword-based in-video search, a feature announced in 2008 [7]. Searching for a phrase like *"that's a tremendous gift"*, a caption from Randy Pausch's famous last lecture titled *Achieving Your Childhood Dreams* [4], indeed reveals a link to that lecture on YouTube. If no closed captions are available, nor can be automatically generated [20], keyword-based search is still available over tags, video descriptions, and titles. Presented with a potentially huge list of results, preview thumbnails based on video still frames help users decide on the most promising result.

A query for—at time of writing—recent events such as the London riots[5] or the shooting in Utøya[6] reveals a broad selection of all sorts of video content, either professionally produced or, more often, shaky amateur videos taken with smartphones.

[4] http://www.youtube.com/watch?v=ji5_MqicxSo
[5] http://en.wikipedia.org/wiki/2011_London_riots
[6] http://en.wikipedia.org/wiki/2011_Norway_attacks

Despite these and other differences, their thumbnails are typically very similar, as can be seen in Figure 1. These thumbnails are automatically generated by an unpublished computer vision-based algorithm [6]. From a user's point of view, it would be very interesting to see whether a video contains different shots. For example, a back-and-forth between a news anchorman and live images can be an indicator for professionally produced content, whereas a single shot covering the entire video can be an indicator for amateur-generated eyewitness footage.

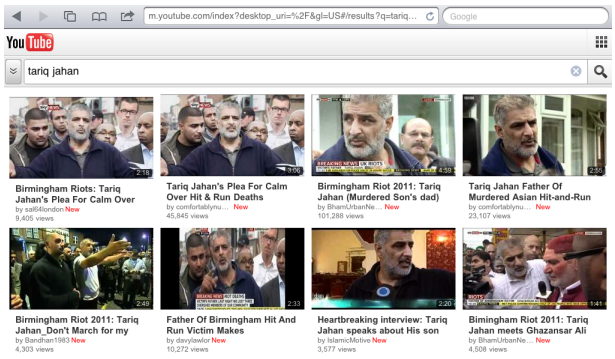


Fig. 1: YouTube search for "tariq jahan", father of a victim of the London riots.

In addition to the information provided by the separation of a video in shots, listing occurrences of named entities and their disambiguation can help users quickly decide whether a given video is of interest. For example, if a video about Utøya contains an occurrence of the Norwegian Prime Minister Jens Stoltenberg, or a video about the London riots contains an occurrence of the Prime Minister of the United Kingdom David Cameron, they can potentially be considered more trustworthy than other videos. It is up to the user to judge the trustworthiness aspect, however, the more context is available, the easier this decision gets.

While the detection of persons and their identification would be possible through face detection and face recognition techniques, this task is computationally expensive. As we have shown in [18], however, good results are possible through the analysis of the available textual metadata with Natural Language Processing (NLP) techniques, especially given the availability of (possibly automatically generated [20]) closed captions on YouTube. Finally, for videos that are longer than the attention span of a typical YouTube user, exploiting purposeful in-video navigation data can help determine points of interest within videos. For example, many users might skip the intros typically contained in professionally produced video content, or jump to spectacular shots directly.

We define three types of events: *visual events* in the sense of shot changes, *occurrence events* in the sense of the appearance of a named entity, and *interest-based events* in the sense of purposeful in-video navigation by users. In this paper, we report on a browser extension that enables crowdsourcing of event detection in YouTube videos

through a combination of *textual*, *visual*, and *behavioral* analysis techniques. When a user starts watching a video, three event detection processes start:

Visual Event Detection Process We detect shots in the video by visually analyzing its content [19]. We do this with the help of a browser extension, *i.e.*, the whole process runs on the client-side using the modern HTML5 [12] JavaScript APIs of the `<video>` and `<canvas>` elements. As soon as the shots have been detected, we offer the user the choice to quickly jump into a specific shot by clicking on a representative still frame.

Occurrence Event Detection Process We analyze the available video metadata using NLP techniques, as outlined in [18]. The detected named entities are presented to the user in a list, and upon click via a timeline-like user interface allow for jumping into one of the shots where the named entity occurs.

Interest-based Event Detection Process As soon as the *visual events* have been detected, we attach JavaScript event listeners to each of the shots and count clicks on shots as an expression of interest in those shots.

Fig. 2: Screenshot of the YouTube browser extension, showing the three different event types: *visual events* (video shots below the video), *occurrence events* (contained named entities and their depiction at the right of the video), and *interest-based events* (points of interest in the video highlighted with a red background in the bottom left).

Figure 2 shows the seamless integration of the detected events into the YouTube homepage. Contributions of this paper are the browser extension itself as well as the underlying crowdsourcing framework for automatic and scalable semantic annotations of HTML5 videos.

2 Related Work

Many different approaches to event detection in video exist. A first category is artificial vision, which tries to extract visual characteristics and identify objects and patterns. A second option is to reuse existing metadata and try to enhance it in a semantic way. Finally, using the combined result of collaborative human efforts can lead to data that is otherwise difficult or impossible to obtain.

2.1 Computer Vision Techniques

Searching through multimedia objects is inherently more difficult than searching through text. Multimedia information retrieval is still an active research topic with many challenges left to address [8]. One possibility is the generalization of text-based search to nontextual information [16], in which the query is posed as a multimedia object itself, the so-called query-by-example strategy. Another strategy is semantic indexing, *i.e.*, to annotate a multimedia item's content using textual or ontological means [9]. In this context, various feature extraction algorithms can be used, an interesting option being face detection [23] followed by face recognition [22].

2.2 Semantic Enrichment of Existing Metadata

In addition to automatically available metadata such as recording time and location, video creators can add metadata to their creations, such as title, textual description, and a list of tags. Also, YouTube automatically provides closed captioning in some cases. Unfortunately, these elements are not constrained to any framework or ontology, making automated interpretation difficult. Therefore, several efforts have tried to semantically enrich these existing metadata. Choudhury *et al.* [2] describe a framework for the semantic enrichment, ranking, and integration of Web video tags using Semantic Web technologies. They use existing metadata and social features such as related videos and playlists a video appears in. Gao *et al.* [4] explicitly model the visual characteristics of the underlying semantic video theme. This semantic model is constructed by finding the common features of relevant visual samples, which are obtained by querying a visual database with keywords associated with the video. Recently, Bræck Leer [1] also provided an interesting method to detect events in videos using semantic subtitle analysis. We previously described [18] a Web application that allows for the automatic generation of Resource Description Framework (RDF) video descriptions based on existing metadata. Textual information is enriched by extracting named entities via multiple Natural Language Processing Web services in parallel. The detected named entities are interlinked with DBpedia concepts. These entities are explicitly anchored to a point in the video thanks to the closed captions. In combination with a shot detection framework, entities can be anchored to shots instead, which is context-wise the better option.

2.3 Crowdsourced Annnotation Approaches

A radically different approach is to tackle the plethora of videos with the driving force behind it: an enormous community of users. The idea of crowdsourcing [3] is that, given the current limitations of automated vision and semantic analysis, we use human intelligence to perform those tasks in which humans currently excel. The aim is to make this task as easy and as less time-consuming as possible, in order to avoid disturbing a user's experience. Soleymani and Larson describe the use of crowdsourcing for annotating the effective response to video [17]. They discuss the design of such a crowdsourcing task and list best practices to employ crowdsourcing. The trade-off between the required effort versus the accuracy and the cost of annotating has been described by Vondrick *et al.* [24]. The quality of annotations generated by a crowdsourcing process has been assessed by Nowak and Rüger [14]. They conclude that a majority vote is a good filter for noisy judgements to some extent, and that under certain conditions the final annotations can be comparable to those of experts. Welinder and Perona [25] devise a model that includes the degree of uncertainty and a measure of the annotators' ability. It should be noted, however, that the usefulness of annotations also depends on their envisioned functional value, *i.e.*, what purpose they should serve in the application.

3 Crowdsourcing Event Detection in Videos

The term *crowdsourcing* was first coined by Jeff Howe in an article in the magazine Wired [11]. It is a *portmanteau* of "crowd" and "outsourcing". Howe writes: *"The new pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R&D"*. The difference to outsourcing is that the crowd is undefined by design. For our specific use case, any YouTube user with the browser extension installed could be part of that crowd.

Event detection in videos is an ideal candidate for crowdsourcing, as each video is an independent object in itself, *i.e.*, the whole set of all existing YouTube videos can be easily split into subtasks by just analyzing one video at a time. We store analysis results centrally, as outlined in Section 4. In the following, we explain for each event type the crowdsourced parts: for *visual* and *occurrence events*, shots and named entities in the video are detected once by whatever the first YouTube user that watches the video. Subsequent viewers can directly profit from the generated annotations. For *interest-based events*, acknowledging that points of interest within a video might change over time, we capture purposeful navigation events by all users. This allows for the generation of a heat-map-like overlay on top of the video shots, which results in an intuitive representation of popular scenes. Our advancement here is that we do not need write access to YouTube, but through our browser extension generate that metadata layer on top, while still creating a seamless and crowd-enriched experience for the user.

4 Implementation Details

We first provide an overview of the background technologies used in the framework and then explain how our browser extension works.

4.1 Background Technologies

Google Chrome Extensions Google Chrome extensions are small software programs that users can install to enrich their browsing experience with the Google Chrome browser. They are written using a combination of standard Web technologies, such as HTML, JavaScript, and CSS. There are several types of extensions; for this paper we focus on extensions based on so-called *content scripts*. Content scripts are JavaScript programs that run in the context of Web pages via dynamic code injection. By using the standard Document Object Model (DOM), they can modify details of Web pages.

Google Analytics Google Analytics is Google's Web analysis solution allowing for detailed statistics about the visitors of a website. The software is implemented by adding an unobtrusive snippet of JavaScript code on a website. This code collects visitor data through a request for an invisible image, during which the page and user data is reported back in the query part of the image's URL. The snippet also sets a first party cookie on visitors' computers in order to store anonymous information such as whether the visitor is a new or returning visitor, or the website the visitor came from.

4.2 Event Detection Processes

This paper is a first step in the direction of future work outlined in a prior publication [19]. Therein, we described the visual analysis-based shot detection algorithm in isolation and noted the potential of combining the visual results with textual analysis results following a method detailed in [18].

Visual Event Detection Process Our approach is based on HTML5 [12] JavaScript APIs of the `<video>` and `<canvas>` elements and falls in the family of histogram-based shot detection algorithms. The complete process has been detailed in [19]. We analyze the video frames' pixels tile-wise and calculate the local histograms in steps of one second. We then calculate the frame distances and finally split the video in shots wherever the frame distance is greater than the average deviation of all frame distances.

Occurrences Event Detection Process In [18], we document an interactive Web application that allows for the automatic annotation of YouTube videos in RDF based on title, description, tags, and closed captions. In the current implementation, we use `Factor`, `Product`, and `Agent` from the Event Ontology [15] to relate events to factors (everything used as a factor in an event), products (everything produced by an event), and agents (everything that can serve as an event agent). Listing 1 shows a sample video fragment annotated with the Event Ontology.

Interest-based Event Detection Process For each scene in a video, we generate a set of `<img>` elements. These sets get injected into the YouTube homepage's DOM tree, as can be seen in Figure 2. Each of the `<img>` elements has a registered JavaScript event handler that upon click triggers two actions: first, the video seeks to the corresponding time, and second, the shot is tracked as a point of interest in the video. We therefore use Google Analytics event tracking [5], logging the video ID and the video timestamp.

```
<http://gdata.youtube.com/[...]/9oWNcw8dits> event:Event :event.

:event a event:Event;
  event:time [
    tl:start "PT0.00918S"^^xsd:duration;
    tl:end "PT0.01459S"^^xsd:duration;
    tl:duration "PT0.00541S"^^xsd:duration;
    tl:timeline :timeline;
  ];
  event:factor <http://dbpedia.org/resource/David_Cameron>;
  event:factor <http://sw.opencyc.org/2008/06/10/concept/en/↩
      PrimeMinister_HeadOfGovernment>;
  event:factor <http://dbpedia.org/resource/Plastic_bullet>;
  event:factor <http://dbpedia.org/resource/Water_cannon>;
  event:product [
    a bibo:Quote;
    rdf:value """Prime Minister David Cameron authorized police
              to use plastic bullets and water cannons,"""@en;
  ] .
```

Listing 1: Exemplary extracted named entities from a YouTube video on the London riots.

4.3 Bringing It All Together

From a Linked Data [10] point of view, the main challenge with our browser extension was to decide on an as-consistent-as-possible way to model the three different event types of *visual events*, *occurrence events*, and *interest-based events*. We decided for a combination of two vocabularies: the Event Ontology [15] mentioned before, and the W3C Ontology for Media Resources [13], which aims to foster the interoperability among various kinds of metadata formats currently used to describe media resources on the Web. The ontology also allows for the definition of media fragments. For this purpose we follow the Media Fragments URIs [21] W3C Working Draft that specifies the syntax for media fragments URIs along several dimensions. The temporal dimension denotes a specific time range in the original media denoted by the `t` parameter. In our case, a media fragment is the part of a video spun by the boundaries of the shot that contains the frame that the user clicked. Listing 2 shows an exemplary semantic annotation of a 27s long video shot containing a *visual event* (the shot itself), an *occurrence event* (the DBpedia URI representing David Cameron), and an *interest-based event* (a point of interest spanning the whole shot).

5 Discussion of our Approach

Regarded in isolation, neither of our video event analysis steps is new, as detailed in Section 2. Our contributions are situated (i) in the scalability through crowdsourcing, (ii) in the on-the-fly HTML5 client-side nature of our approach, and (iii) in the combination of the three different event type annotations. Hence, we discuss our preliminary results

```
<http://gdata.youtube.com/[...]/9oWNcw8dits> event:Event :event1.

:event1 a event:Event;
  event:time [
    tl:start "PT0.025269S"^^xsd:duration;
    tl:end "PT0.05305S"^^xsd:duration;
    tl:timeline :timeline;
  ];
  event:factor <http://dbpedia.org/resource/David_Cameron>;
  event:product [
    a bibo:Quote;
    rdf:value """on camera. DAVID CAMERON, British prime
              minister: We needed a fight back, and a fight
              back is under way. [...] there are things that
              are badly wrong in our society. [...]"""@en;
  ];
  event:product ↩
      <http://gdata.youtube.com/[...]/9oWNcw8dits#t=25,53>.

<http://gdata.youtube.com/[...]/9oWNcw8dits#T=25,53> a ↩
    ma:MediaFragment.
```

Listing 2: Semantic annotation of a 27s long video shot (*visual event*) showing David Cameron (*occurrence event*) talk about the London riots. The shot is also a point of interest generated by a click of a YouTube user (*interest-based event*).

in contrast to a classic centralized approach. For *visual event* analysis, rather than detecting shots client-side with HTML5 JavaScript APIs, a centralized approach with low level video tools is superior in terms of accuracy and speed, as the video files do not have to be streamed before they can be processed. The crowdsourced approach is not necessarily more scalable, however, more flexible as it can be applied to any source of HTML5 video. For *textual event* detection, this is a task that necessarily runs centrally and not at the client due to the required huge text corpora. Finally, *behavioral event* detection by definition is only possible on the client. While most users are not aware that their navigation behavior can be used to detect points of interest and thus behave naturally, fraud detection is necessary to filter out spam pseudo navigation events.

In [3], Doan *et al.* introduce four questions for a crowdsourced system, the first being *how to recruit and retain users*. Our response is by seamlessly and unobtrusively enriching the user's YouTube experience. The user is not even aware that she is part of a crowdsourced system, and still profits from the crowd. Doan's next question is *what contributions can users make*. The response are annotations for the three event types defined earlier. The third question is *how to combine user contributions to solve the target problem*, with the target problem being to—in the longterm—improve video navigability, searchability, and accessibility. Our response is twofold: for *visual* and *textual events*, we consider only the first user's annotations, and for *behavioral events* we consider the annotations from all users by means of a heat map, as detailed in Section 3. The last question is *how to evaluate users and their contributions*. Our response is again

twofold. First, given that *visual* and *textual events* once detected are not questioned (as the outcome will always be the same), here the performance of individual users does not need to be evaluated. In contrast, the quality of *behavioral events* will simply improve by the combined wisdom of the crowd, always given proper fraud detection and future improvements mentioned in Section 6.

6 Future Work and Conclusion

Future work will focus on several aspects. First, given the streaming video nature, our approach inherits the speed and accuracy challenges from [19]; the solution here is to work with lower resolution versions of the video files in the background. Second, more elaborate interaction tracking for *interest-based events* is necessary. Facets like playing time after a navigational click can shine more light on the quality of the believed point of interest. If a user clicks on a supposedly interesting scene but then navigates away quickly afterwards, this is a strong indicator we need to consider. In the complete opposite, if a user never navigates within a video, this can be an indicator that the video is exciting from the first second to the last. Third, rather than just enriching the user experience for the current video, we will explore in how far the crowd-generated background knowledge gained on videos can be used for a more efficient video recommender system. This can be evaluated via A/B blind tests on clickthrough rates, where a user is randomly presented with a YouTube-generated related video recommendation, and a recommendation generated by the browser extension.

Concluding, our crowdsourced approach has shown promising results. The combination of *textual*, *visual*, and *behavioral* analysis techniques provides for high quality metadata that otherwise could only be generated through human annotators. Our framework is a scalable first step towards video event detection, with actionable steps ahead.

References

1. Bræck Leer, E.: Detecting Events in Videos Using Semantic Analytics of Subtitles. Master's thesis, University of Tromsø (Jun 2011)
2. Choudhury, S., Breslin, J., Passant, A.: Enrichment and Ranking of the YouTube Tag Space and Integration with the Linked Data Cloud. In: The Semantic Web – ISWC 2009, Lecture Notes in Computer Science, vol. 5823, chap. 47, pp. 747–762. Springer (2009)
3. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the World-Wide Web. Commun. ACM 54, 86–96 (April 2011)
4. Gao, Y., Zhang, T., Xiao, J.: Thematic video thumbnail selection. In: Proc. of the 16^{th} IEEE Int. Conf. on Image Processing. pp. 4277–4280. ICIP'09, IEEE Press, Piscataway, NJ, USA (2009)
5. Google: Google Analytics Event Tracking Guide, http://code.google.com/apis/analytics/-docs/tracking/eventTrackerGuide.html
6. Google Research Blog: Smart Thumbnails on YouTube (January 19, 2009), http://googleresearch.blogspot.com/2009/01/smart-thumbnails-on-youtube.html
7. Google Video Blog: Closed Captioning Search Options (June 05, 2008), http://googlevideo.blogspot.com/2008/06/closed-captioning-search-options.html
8. Hanjalic, A., Lienhart, R., Ma, W.Y., Smith, J.R.: The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? Proc. of the IEEE 96(4), 541 –547 (Apr 2008)

9. Hauptmann, A., Christel, M., Yan, R.: Video retrieval based on semantic concepts. Proc. of the IEEE 96(4), 602–622 (Apr 2008)
10. Hausenblas, M., Troncy, R., Raimond, Y., Bürger, T.: Interlinking Multimedia: How to Apply Linked Data Principles to Multimedia Fragments. In: WWW 2009 Workshop: Linked Data on the Web (LDOW2009). Madrid, Spain (2009)
11. Howe, J.: The Rise of Crowdsourcing. Wired 14(6) (2006), http://www.wired.com/wired/archive/14.06/crowds.html
12. HTML5: A vocabulary and associated APIs for HTML and XHTML. W3C Working Draft (August 2009), http://www.w3.org/TR/2009/WD-html5-20090825/, http://www.w3.org/TR/2009/WD-html5-20090825/
13. Lee, W., Bürger, T., Sasaki, F., Malaisé, V., Stegmaier, F., Söderberg, J.: Ontology for Media Resource 1.0. Tech. rep., W3C Media Annotation Working Group (06 2009), http://www.w3.org/TR/mediaont-10/
14. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proc. of the Int. Conf. on Multimedia Information Retrieval. pp. 557–566. MIR '10, ACM, New York, NY, USA (2010)
15. Raimond, Y., Abdallah, S.: The Event Ontology (October 25, 2007), http://motools.sourceforge.net/event/event.html
16. Sivic, J., Zisserman, A.: Efficient visual search for objects in videos. Proc. of the IEEE 96(4), 548–566 (Apr 2008)
17. Soleymani, M., Larson, M.: Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. In: Carvalho, V., Lease, M., Yilmaz (eds.) Proc. of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010). ACM SIGIR, ACM (Jul 2010)
18. Steiner, T., Hausenblas, M.: SemWebVid - Making Video a First Class Semantic Web Citizen and a First Class Web Bourgeois. In: Semantic Web Challenge at ISWC2010 (November 2010), http://www.cs.vu.nl/~pmika/swc/submissions/swc2010_submission_12.pdf
19. Steiner, T., Verborgh, R., Van de Walle, R., Brousseau, A.: Enabling On-the-Fly Video Scene Detection on YouTube and Crowdsourcing In-Video Hot Spot Identification. In: Proc. of the 2^{nd} IEEE Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (submitted). ARTEMIS 2011, IEEE (2011)
20. The Official Google Blog: Automatic Captions in YouTube (November 19, 2009), http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html
21. Troncy, R., Mannens, E., Pfeiffer, S., Deursen, D.V.: Media Fragments URIs. W3C Working Draft (December 8, 2010), http://www.w3.org/2008/WebVideo/Fragments/WD-media-fragments-spec/
22. Verstockt, S., Van Leuven, S., Van de Walle, R., Dermaut, E., Torelle, S., Gevaert, W.: Actor recognition for interactive querying and automatic annotation in digital video. In: IASTED Int. Conf. on Internet and Multimedia Systems and Applications, 13^{th} Proc. pp. 149–155. ACTA Press, Honolulu, HI, USA (2009)
23. Viola, P., Jones, M.: Robust real-time object detection. In: Int. Journal of Computer Vision (2001)
24. Vondrick, C., Ramanan, D., Patterson, D.: Efficiently scaling up video annotation with crowdsourced marketplaces. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol. 6314, pp. 610–623. Springer (2010), 10.1007/978-3-642-15561-1_44
25. Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, San Francisco, CA, USA (Jun 2010)
26. YouTube: Official Press Traffic Statistics (2011), http://www.youtube.com/t/press_statistics

Clues of Personal Events in Online Photo Sharing

Pierre Andrews, Javier Paniagua, and Fausto Giunchiglia

Dipartimento di Ingegneria e Scienza dell'Informazione
University of Trento, Italy
{andrews,paniagua,fausto}@disi.unitn.it

Abstract. There is currently a trend in media management and the semantic web to develop new media processing methods and knowledge representation techniques to organise and structure media around events. While this increased interest for events as the central aggregator when organising media is supported by strong research in the fields of knowledge representation and computer vision; it is not yet clear how the digital era users use events when sharing their personal media collection. In this paper, we explore how users share photos online and discuss the results of a preliminary automatic processing of the data collected. We show that while media sharing services do not support events as yet, users still share their media around personal events, either by providing explicit spatio-temporal metadata, or by using an event-centric vocabulary.

1 Introduction

With the increased availability of digital capturing devices, people now build large personal media collections; what is then done with these media has drastically changed in the last years with the emergence of popular photo sharing services like Flickr[1] or Picasa[2] and social networks sites such as Facebook[3]. Understanding how people organise and share their digital collections is key to building better tools that accommodate for the users' needs instead of forcing them to change their mental model to fit a fixed software workflow. This workflow has changed with the introduction of new technologies and ways to interact and share media online; the user's goal is now, not only to archive media for a personal use but also to share them with relevant contacts online. Therefore the issue is not only of organising the user personal media collection for better future search and retrieval, but also the one of organising shared media for visibility to the relevant people and future search and retrieval of these media not just for the author that built the collection but also by these relevant people.

Until recently, the "album" concept was one of the main metaphor for helping users to organise their personal collection, thus staying close to how photo prints

[1] http://www.flickr.com
[2] http://picasaweb.google.com
[3] http://www.facebook.com

where organised previously. However, new metaphors of organisation are now emerging to leverage more complex indexing and search. Flickr for instance has introduced a very loose organisation system, focusing on tags to group photos, and with the availability of GPS technology, media management services have introduced the possibility to "geotag" media and to browse and search them with location based services. Some have also introduced search and navigation services based on who is in the photo and when it was taken, using the metadata provided by the camera.

This use of media metadata is moving away from the physical photo album metaphor. However, there is still a semantic gap between the low level metadata, the high level information of who is in the photo or where it was taken and how people group their media for personal archiving and sharing. In fact, most popular media management services still provide the "album" metaphor[4] as people still have a need to group media together in ways that are more meaningful to them than just a location or time grouping. Researchers are thus focusing around the *event* metaphor to combine metadata seems to represent part of the higher level intent of the users when they group their media.

While this metaphor is backed by some early user studies, these were led before the large adoption of social media sharing services and there has been little recent research on how users actually use events digitally to organise and share their media. Discovering if this is the case is not an easy task, and in this paper we discuss an initial study of the sharing behaviour of users on Flickr and Picasa to see if they are currently using events when sharing their photos online. We first introduce the current work on event representation for media management and the current model of events (Section 2). In the following sections we discuss how we have collected data on Flickr and Picasa (Section 3), how people use event metadata when organising in the *album* metaphor of these sites (Section 4) and how, even when they do not explicitly use such metadata, they often share media by using an event-centric vocabulary (Section 5).

2 Events for Media Organisation

It seems that people take photos to archive important events and share within their close community. [16] and [11], found clues that people intentionally classify photos according to events in their lives. This fact has been observed even before the advent of digital photography in [2]; however, Chalfen argues that people do not share pictures per-se but use them to tell a story. More recently, Miller et al's user study [12] similarly concluded that users took photos primarily to archive important events and share within their community; however, at the time of their study, they found that layman users did not share photos actively online and preferred to use prints or email.

This organisation of photos "chronologically by event" eases the search and retrieval of specific photos in personal collections as it aligns with the way memory is structured. According to [21], humans identify activity boundaries at

[4] Flickr calls it a "Set".

points that correspond to a maxima in the number of changing physical features, thus aggregating memories around events. [10] states that the brain operates in this way to cope with the increased difficulty brought by indexing new information when it is dissimilar from the "current moment" beyond a certain threshold. Some are thus proposing a event-centric models to characterise media in terms of the events they are associated with [6,8,4].

Last.fm and Upcoming.org are services that already try to link media and event. They do so for public events such as concerts or conferences, but still do not allow users to share their personal events (e.g. weddings, birthdays). [3] presents a user study to elicit requirements for such services and interaction paradigms that help discover and enrich public events. While this approach is interesting for public events, it does not clarify if users naturally use personal events to organise and share their personal media collections.

[21] recognises subject, actors and causal properties as components of the human perception of and event, stressing the importance of the *temporal and spatial* aspects to build the event structure. [1] defines events as having a close link to their *spatio-temporal collocation* and to the things that constitute their subject (e.g., a sparrow in the event "a sparrow falls"). Inter-event relations are studied in [17] that states that events may be composed of sub-events that are temporally, spatially and causally connected. [7] explores use case scenarios to show possible ways in which untrained users may organise media in terms of events with complex *spatio-temporal structure*.

Practical models for events can be found in the IPTC G2 family of news exchange standards are provided. EventML[5] is one of these standards oriented at describing public events in a journalistic fashion, although support for media is limited, and this model is close to Chalfen's idea that media are only used to support a story. A set of requirements for a base model of events is presented in [20] that categorises all the properties and relations of an event into six aspects: *temporal*, *spatial*, informational, experiential, structural and causal. The F event model [18] specifically addresses most of these requirements, [19] also addresses the *temporal*, *spatial* and informational aspects by integrating different ontological models. The Simple Event Model is proposed in [5] to represent not only who did what, when and where, but also to model the roles of each actor involved, when and for how long this is valid and according to whom. MediAssist [14] organises digital photo collections using time and location information combining it with content-based analysis (face-detection and other feature detectors). The work in [15] uses time and latitude/longitude data to analyse tags and unstructured text from photos on Flickr to extract place and event semantics. VisR[6] is a smartphone application that detects events from photos and metadata available on the device. All these studies have in common the predominance of the spatio-temporal aspect of events as it is the one that helps users determine inter-event boundaries, recollect their memories and find

[5] `http://www.iptc.org/site/News_Exchange_Formats/EventsML-G2/`
[6] http://www.visrapp.com/

their media. Thus, events refers to "something that occurs in a certain place during a particular interval of time" [6].

3 Photo Sharing Websites: Data Collection

Event-centric services such as Upcoming.org or Last.fm are focused on public events such as concerts or conferences. While datasets [3] based on these websites already provide samples of media organised around event metadata, they do not represent personal events. That is, media of more personal events, such as a birthday or a holiday, are not shared on these websites. However, this kind of media can be found on photo sharing websites such as Flickr and Picasa where users share photos of personal happenings with their family and friends.

These websites do not provide a way to organise photos around events but provide a way to group photos in *albums*.These albums can only have a very small amount of metadata and are not presented as events to the user. On Picasa, albums can have a *title* and a *description*, and optionally a *date* and a *location*; on Flickr, sets can only have a *title* and a *description*.

We are interested in seeing how the users describe albums they share on Picasa and Flickr by using the title and description fields. Our hypothesis is that if they share media related to events, they will provide the event metadata in the fields that are available to them and we will find event references in the titles and descriptions of the albums. We are focusing on these two social sharing sites as they are some of the more popular available at the time of writing; while Facebook is also very popular, it provides very similar features (album based organisation of photos) and does not allow data collection.

We have thus collected a dataset of digital albums shared on Flickr and Picasa. To select users, we use the "explore" pages of each website that feature randomly selected photos; from these photos, we find a set of random users and collect all public albums that are shared by these users. For each album shared on Flickr we retrieve: (a) the *title* of the set, (b) the *user identification*, (c) the *URL* of the set and (d) the *number of photos and videos* within the set. For each album from Picasa we collect: (a) its *URL*, (b) the *date* specified for this album, (c) the *number of photos*, (d) the *title*, (e) the *description* and (f) the *user identification*.

Because both websites are international, many entries are not written in English. In this paper, we are only able to process metadata provided in English and thus want to filter out the other languages. The perl Lingua::Identify[7] module was used to identify the language of the title and description (when available) in each album entry. The algorithm provided by this module was trained on the EuroParl corpus [9]; we have performed a manual annotation of a subset of the automatically processed entries from the Picasa dataset and have found that the algorithm labels English albums with a 89% precision.

We are interested to see if users refer to locations when they describe albums and have thus automatically processed the dataset to find references to geo-

[7] `http://search.cpan.org/~ambs/Lingua-Identify/`

graphic locations. The Yahoo! Placemaker[8] service is used to perform this task. This freely available geoparsing service can identify place references in unstructured text. While Yahoo! does not provide information on the accuracy of their algorithm, from our manually annotated sample, we found that Placemaker is able to detect if there is at least one location reference in an English title with 81.2% accuracy.

References to dates are also of interest to us as time is a main attribute of an event. To detect such references, we analysed each title with a custom date parsing algorithm that detects full dates but also partial dates (e.g. "Paris'08") and periods (e.g. "40.5 miler in Sespe Wilderness April 2nd - 5th 2010"). On our manually annotated sample, this algorithm performed with 88.1% accuracy.

We have collected 32 168 sets from Flickr and 88 593 albums from Picasa over the month of July 2011[9]. We have kept only English albums, resulting in 5 339 (16.6%) sets from Flickr, and 11 355 (12.8%) albums from Picasa.

4 A Given Place and Time

According to the definition that we introduced in Section 2, the two main attributes defining an event are its location and when it happened. Thus, if users are to describe events using albums when sharing their photo, they will probably specify some of these attributes within the available attributes. We found that in the Picasa dataset, only 31% of the albums have a description and thus, in this paper, we focus on the title attribute of the albums as we do not have enough data to draw conclusions from the descriptions.

Table 1. Proportion of Albums with Titles Referring to Dates or Location

	Flickr	Picasa	Flickr+Picasa
Dates	33.9%	44.6%	41.2%
Locations	22.4%	26.7%	25.3%
Both	8.7%	12.9%	11.6%

Table 1 shows the proportions of albums where date or location references can be found, a test of equal proportion shows that Picasa and Flickr are comparable ($p < 0.01$) and we thus consider that there is no difference in users' behaviour between the two services in the factors we analyse.

The number of albums where an explicit date reference can be found in the title makes for more than a third of the dataset. We can thus see that people do like to share their albums with metadata about the date when the photos were taken. Note that while the title is set manually by the users, the *date* field

[8] `http://developer.yahoo.com/geo/placemaker/`

[9] the random crawling collected albums posted between 2006 and 2011.

on Picasa is filled automatically with the album creation date if the user does not specify any value explicitly. When there is a date in the title on Picasa, it is often not consistent with the album *date* field. It seems that while the users are ready and interested to share their photos around dates, they are not motivated to fill in an extra metadata field. The reason behind this might be a limitation in Picasa's interface or it can simply be because the users do not see the gain in filling this extra field.

While the date is an important attribute of events, albums with only a date reference are not always events according to our previous definition. In fact, from a preliminary manual annotation of the Picasa dataset, we can see that 27.3% of the albums with a date reference but no location reference are not really events. This is because there are catch-all albums for entire years or months, where users put photos of many different events in the album (e.g. "Misc. Apr. 2009"). The album is thus only a way to aggregate photos in a time range and not used to represent a specific event. This happens also when people share photos of their newborn child for milestone periods (e.g. "Jake - 9 months: March").

There are less albums with an explicit location reference, but it still makes for a fourth of the dataset. From the manual annotation, we can see that 78.7% of the albums with only a location reference are actually events. In the same way as with the dates, users use locations for catch-all albums where they put photos of a location they visited multiple times but not for any specific event (for instance for photos of their home-town).

In these two cases, we can see that the dates and locations are sometimes used only as aggregators for media that could be replaced by automatic metadata based services. However, it seems that the users are not aware of, or willing to use, these services on the studied websites.

96.8% of the albums with a date and a location together were annotated as being events by the manual validators. While these albums represent a small amount of the dataset, we can already see that when space and time are specified in the title, the users wanted to share an important event.

5 An Event Vocabulary

In the previous section, we have looked at how users might use album attributes to describe explicitly an event location or date. However, there are many events represented on Picasa and Flickr that do not include explicit dates or locations. For instance "Janet and Ian's wedding", "father's day", "Michelle's shower" or "Christmas Eve" are all titles of albums from our dataset that do represent important personal events with no explicit dates or locations. Thus, there might be more albums in this dataset that represent events than the previous section's analysis hinted.

In fact, if we look at the most popular words used in the titles (see Table 2), many of them are references to events (e.g. "party, "wedding", "trip") or time periods, without having explicit dates. Note that, while not shown in Table 2, the most popular words in the vocabulary are years, in fact on Flickr, 11.0%

of the vocabulary are numerals while on Picasa 17.5% of the words used are numbers. Table 2 reports figures in per-thousand, while the distribution of the vocabulary follows a very steep long-tail curve, the most popular words still do not cover a large part of the album vocabulary.

Table 2. Most Popular Words in Titles (‰ of the whole vocabulary)

Flickr	‰	Picasa	‰	Flickr+Picasa	‰
spring	5.88	**new**	4.72	**spring**	1.76
city	6.24	**trip**	5.82	city	1.87
day	7.80	**wedding**	8.30	**day**	2.34
wedding	12.78	**day**	11.73	**wedding**	3.83

While it is easy to see that in the most used words in the dataset there are concepts representing events, it is not an exhaustive view of the dataset and it would be interesting to see how many albums refer to events by using such vocabulary. However, it is not easy to exhaustively list manually the whole vocabulary that could be used to refer to events. We take a semi-automatic approach, using WordNet [13] as a thesaurus, to find all terms that might refer to a concept representing an event. To do so, we have listed all inherited hyponyms of the synset `event#n#1` – which include the words "wedding", "birthday", etc. – and of the synset `calendar_day#n#1` – which include the words "Christmas", "Thanksgiving", etc. This provides us with a list of 11 092 words and 14 304 concepts combined in 15 389 word-concept pairs[10] that we then searched in the titles of the albums in the dataset.

Table 3. Top Leaf Concepts Related to Events

Flickr			Picasa		
	Events %	**Overall %**		**Events %**	**Overall %**
`Sunday#n#1`	3.33	1.64	`marriage#n#3`	3.42	1.74
`Easter#n#1`	3.41	1.68	`Easter#n#1`	4.37	2.22
`Michigan#n#3`	3.41	1.68	`Halloween#n#1`	4.51	2.29
`Halloween#n#1`	4.28	2.11	`Christmas#n#2`	5.30	2.70

We found that around half of the albums (Flickr: 49.4%; Picasa: 50.9%) have a title with at least one word that represents an event according to WordNet. Of these albums, only 29.6% have a date or a location (or both) in the title. There are indeed many albums that describe events without providing either an explicit date or a location reference (e.g. "Katie's Swiss trip", "Field trip - Farm",

[10] Note that because of homography, the same word can appear under different concepts.

Table 4. Top Concepts Related to Events – cumulating the hyponyms occurrences

	Flickr		Picasa	
	Events %	**Overall %**	**Events %**	**Overall %**
`calendar_day#n#1`	24.2	11.9	27.5	14.0
`activity#n#1`	39.3	19.4	32.4	16.5
`act#n#2`	66.3	32.8	64.3	32.8
`event#n#1`	75.8	37.5	72.5	36.9

"Lily fathers Day"). From a preliminary analysis at these albums, it seems that many of them either refer to the third important attribute of an event: the participants; or to relative dates (e.g. "Father's Day", "My Birthday") or locations (e.g. "Trip Home"). In fact, we can see in Tables 3 and 4 that the `day#n#3` and `calendar_day#n#1` are among the most used concepts. This is in line with Jain's [6] definition of an event: "a significant occurrence or happening, or a social gathering or activity". However, relative location or participant references are hard to detect automatically and further work is required to check how these are used in the album vocabulary.

WordNet is a very detailed vocabulary and many terms that it declares as relating to the *event* concepts might not be used by the users to refer to events. Indeed, there is ambiguity in the vocabulary and we have taken a naive approach where we count the occurrence of all possible words without applying disambiguation. For instance, `Michigan#n#3` appears as one of the most popular leaf concepts for Flickr; however, this concept represents a card game called "Michigan" but might have been used by users in their album title as the location. The other top concepts however represent less ambiguously event references.

This confirms Chalfen's conclusions that people like to take photos around personal events that they then share with a community made of close relations ([2]). However, as we have discussed earlier, these photos are usually shared without description, and thus Chalfen's hypothesis that people use photo to tell a story might not be exact on photo sharing websites.

6 Discussion

The results we found, while preliminary, show that there is a tendency for users to share photos around places and location. While this is not a guarantee that they are sharing albums about specific personal events, it seems to align with the previous observations of Zacks et al and Kurby et al [10,21] who found that users like to segment their memories around time and space.

While most event models discussed in the state of the art (for instance [19]) represent events around dates and locations too, they do not seem to fit perfectly the behaviour of the users that we observed on the sharing sites. In particular, some users seem to aggregate media around date or location without describing events (e.g. the newborn album cases pointed out earlier). While this could be

done automatically from the metadata of the photos, there might be a higher semantic to this grouping when sharing. As Chalfen [2] discusses, even if it was for printed photo, people group photos together to support a story and not always just for the content of the photos per-se. That is, the grouping of photos of the "second month" of a baby is not a specific event according to most of the existing metadata models but is still an event of importance for the users that share them online.

In addition, in accordance to Chalfen's [2] and Miller et al. [12], people share photos around important personal events. These events (e.g. Christmas, trips, visits) are not always global events and their scope is limited to the close circle of personal relationships. This kind of sharing has probably a different purpose from the one of exploring concert or conference photos (for instance) as is discussed in [3], or from the news outlet use-cases for which the IPTC standards have been developed[11]. Therefore, we need custom model and services for layman users.

As was pointed out in [12], there is also a stronger issue of privacy and access control when dealing with the sharing of personal events. On Picasa and Flickr, we were able to crawl public albums – featured on the website main pages – that were of highly personal nature but are accessible to anyone online. While this is not the scope of this paper, we believe that there is a need for better privacy services directly integrated with the event models to deal with the personal media sharing use-cases.

7 Conclusion

In this paper we present a preliminary study of a dataset of albums shared on Flickr and Picasa. As we show, while these two services have different interfaces and features, users tend to have the same behavior on both sites and we believe that this demonstrates some general intent of the users more than site-specific behaviors. While this is a raw analysis of the data and a more extensive manual annotation is required, we have found that a significant amount of users share media online illustrating personal events, and use time-location metadata to describe them. In fact, we have found that more than a third of the albums shared reference a date in their title and more than a fourth refer explicitly to a location. Users also seem to group their photos around important personal events (e.g. birthdays, wedding, festivities) without always specifying explicitly a location or date.

We are planning future work, in particular in analysing the user needs and habits directly with the users, it shows that they already try to use events when sharing media, even when the applicative workflow does not allow it explicitly. We are also planning to extend this work to study the current use of geo-tagging when sharing media. Therefore providing users with new interfaces and services using the event metaphor should improve their experience and the searchability of the media they share online.

[11] `http://www.iptc.org`

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o}248984 GLOCAL and n^{o}247758 EternalS.

References

1. Casati, R., Varzi, A.: Events. In: The Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab; Stanford, CA (2010)
2. Chalfen, R.: Snapshot Versions of Life. Bowling Green State University Popular Press, Bowling Green, Ohio (1987)
3. Fialho, A., Troncy, R., Hardman, L., Saathoff, C., Scherp, A.: What's on this evening? In: EVENTS'10 (2010)
4. Giunchiglia, F., Andrews, P., Trecarichi, G., Chenu-abente, R.: Media Aggregation via Events. In: EVENTS'10 (2010)
5. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L.: Design and use of the Simple Event Model (SEM). Journal of Web Semantics (Accepted (2011)
6. Jain, R.: EventWeb: Developing a Human-Centered Computing System. Computer 41(2), 42–50 (2008)
7. Jameson, A., Buschbeck, S.: Interaction design for the exchange of media organized in terms of complex events. In: EVENTS 2010 (2010)
8. Kim, P.: Event-based Multimedia Chronicling Systems. Computer Engineering pp. 1–12 (2005)
9. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5 (2005)
10. Kurby, C.A., Zacks, J.M.: Segmentation in the perception and memory of events. Trends in Cognitive Sciences 12(2), 72 – 79 (2008)
11. Lansdale, M., Edmonds, E.: Using memory for events in the design of personal filing systems. International Journal of Man-Machine Studies 36(1), 97–126 (1992)
12. Miller, A.D., Edwards, W.K.: Give and Take : A Study of Consumer Photo-Sharing Culture and Practice. In: CHI'07. pp. 347–356 (2007)
13. Miller, G.A.: WordNet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
14. O'Hare, N., Lee, H., Cooray, S., Gurrin, C., Jones, G., Malobabic, J., O'Connor, N., Smeaton, A., Uscilowski, B.: MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections. In: CIVR'06. pp. 529–532 (2006)
15. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR '07. p. 103 (2007)
16. Rodden, K., Wood, K.R.: How do people manage their digital photographs? In: CHI'03. p. 409 (2003)
17. Scheffler, U.: Events as shadowy entities. Philosophy 2, 35–53 (1994)
18. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F–a model of events based on the foundational ontology dolce+DnS ultralight. In: K-CAP'09 (2009)
19. Shaw, R., Troncy, R., Hardman, L.: Lode: Linking open descriptions of events. The Semantic Web pp. 153–167 (2009)
20. Westermann, U., Jain, R.: Toward a Common Event Model for Multimedia Applications. IEEE Multimedia 14(1), 19–29 (2007)
21. Zacks, J.M., Tversky, B.: Event structure in perception and conception. Psychological Bulletin 127(1), 3 – 21 (2001)

New Forms of Interaction With Hierarchically Structured Events

Sven Buschbeck, Anthony Jameson, and Tanja Schneeberger*
Sven.Buschbeck@dfki.de, jameson@dfki.de, tanjaschneeberger89@googlemail.com

German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

Abstract. Research in the semantic web community has given rise to some powerful methods for visualizing events and related media—mostly in terms of interactive timelines—and for enabling users to interact with these visualizations. The present paper aims to advance this line of research in two ways: (a) by developing interactive visualizations of event hierarchies of essentially arbitrary depth and size, which are more natural than timelines in the case of complex events that comprise subevents at various levels; and (b) by supporting forms of interaction that go beyond the usual activities of browsing and searching for events and related media, supporting additionally the sharing and annotation of media and the provision of interactive illustrations of narrative texts.

1 Introduction

1.1 Issues and Goals

This paper addresses the third of the three questions that were formulated in the call for papers for this workshop: "How can events be exploited for the provision of new or improved services?"

One focus of the paper is on new ways of visualizing events and associated media and user commentary, in particular in the case of a complex event (e.g., a soccer tournament) which consists of several levels of subevents (e.g., individual games in the tournament and events within the games). But we also show how the resulting interaction design makes possible novel forms of interaction with events and associated content.

Though the ideas presented have considerable generality, they are introduced here with reference to an implemented prototype that will be demonstrated interactively at the workshop. This prototype is being developed in the context of the European Integrating Project GLOCAL.[1] It is designed to enable professional and nonprofessional users to browse and search for media that are organized in terms of events; to comment

* The research described in this paper is being conducted in the context of the 7th Framework EU Integrating Project *GLOCAL: Event-based Retrieval of Networked Media* (http://www.glocal-project.eu/) under grant agreement 248984. We thank the anonymous reviewers for their perceptive suggestions, which have led to improvements in the final version of the paper.

[1] http://www.glocal-project.eu

on such media and on the associated events; and to upload their own media, introducing them into the event-based organization and thereby in effect indexing them on-the-fly.

To understand the application context, consider a website that presents a large amount of media and information about a complex sports event, such as the soccer World Cup of 2010. Assume that a great many events and associated media have been detected and represented using methods such as those discussed in other papers for this workshop: How can users be enabled to explore these media and contribute media of their own?

One now-familiar way of organizing the media would involve associating each event in the tournament—and hence its associated media—with a given location and showing these locations on an interactive map (see Figure 1). This method can be seen with this example not to supply the most relevant context: It is usually less important to know where the game between Nigeria and the Korean Republic occurred in South Africa than to know where it occurred in the hierarchical structure of the whole tournament. Similarly, showing the game on a timeline would only vaguely and indirectly suggest its significance within the tournament. Consequently, although the GLOCAL interface supports visualizations in terms of geographical maps and timelines, we focus in this paper on the more innovative visualization in terms of subevent hierarchies.

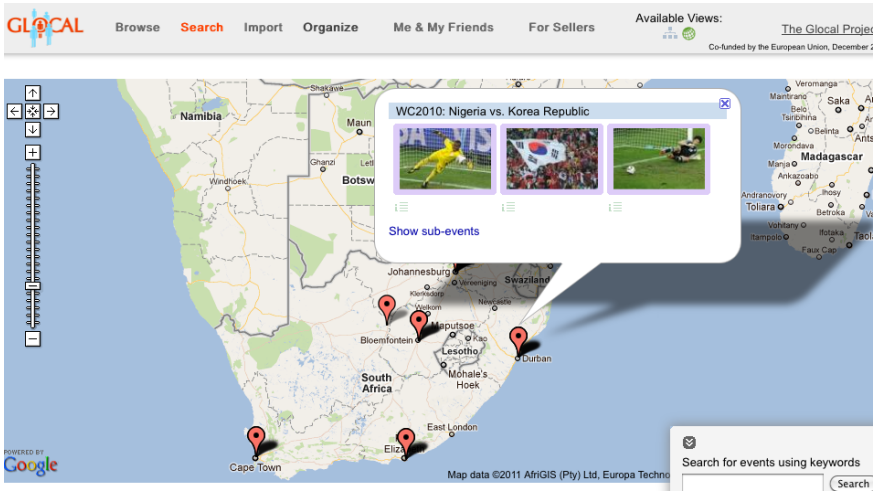


Fig. 1. Example of how events and associated media can be organized relatively conventionally in terms of a geographical map.

1.2 Background of Related Work

There is a fairly extensive tradition in the semantic web community of visualizing events in terms of timelines. Since the work up to 2007 has already been ably summa-

rized by André et al. ([1]), for reasons of space we will discuss just a few highlights, including the work of those authors.

Perhaps the most widely used tool of this sort is TIMELINE,[2] a web widget for visualizing temporal data. It was developed by David F. Hyunh as part of the SIMILE project. In a typical use of the widget, a number of related events (such as those involving the assassination of President Kennedy) are displayed on an interactive timeline (which may comprise a number of parallel lines in the vertical dimension to make it possible to represent events with high density). By clicking on an event in the timeline, the user can access a textual description, which includes further links, for example to a discussion page for that event.

TIMEMAP[3] builds on TIMELINE by integrating it with online mapping systems such as GOOGLE MAPS. Basically, a TIMEMAP display shows simultaneously (a) a timeline with events or intervals and (b) a map with corresponding locations. Clicking on an item in either the timeline or the map brings up additional information about the event in question. It is possible to include filtering functionality so as to show only the events that fulfill particular criteria.

The system CONTINUUM ([1]) likewise builds upon TIMELINE, advancing it in several ways. The way of interest for the present paper is CONTINUUM's ability to represent explicitly hierarchical relationships among events. For example, the lifetimes of classical composers can be shown as belonging to various eras, and each musical composition can be shown as a subevent in the life of its composer. Though this method may cover many types of event hierarchy elegantly, it cannot apparently deal with hierarchies of arbitrary depth and content. In the above example, the eras are represented as segments of the overall timeline; each composer's life is visualized in a box; and his or her compositions are listed in the box. It is not clear how further levels of hierarchy could be added to this visualization.

Outside of the semantic web area, some researchers developing new methods for media organization have explored the use of event hierarchies. For example, REMINISCING VIEW ([2]) enables users to organize photos in a way that is based on an underlying event hierarchy. But REMINISCING VIEW does not, as one might expect, present the media in a hierarchical layout; in fact, it does not visualize the events at all.

As we will see, an attempt to visualize event hierarchies (and the associated media, metadata, and user comments) explicitly requires a visual tree structure with various types of links. The solution presented below (in particular the use of *aggregator nodes*) was inspired in part by the work of Hirsch et al. ([3]) on visualization of large knowledge spaces.

Other aspects of the interface—such as expanding and collapsing subtrees, filtering, and focusing on subtrees— were inspired by functionality offered by mind-mapping tools (e.g., Mindjet's MINDMANAGER, which was used for a first semifunctional prototype). Although these systems have found widespread use on desktop computers, this functionality is seldom found in web-based systems (aside from web-based mind mapping tools such as MINDMEISTER[4].

[2] http://www.simile-widgets.org/timeline/
[3] http://code.google.com/p/timemap/
[4] http://www.mindmeister.com

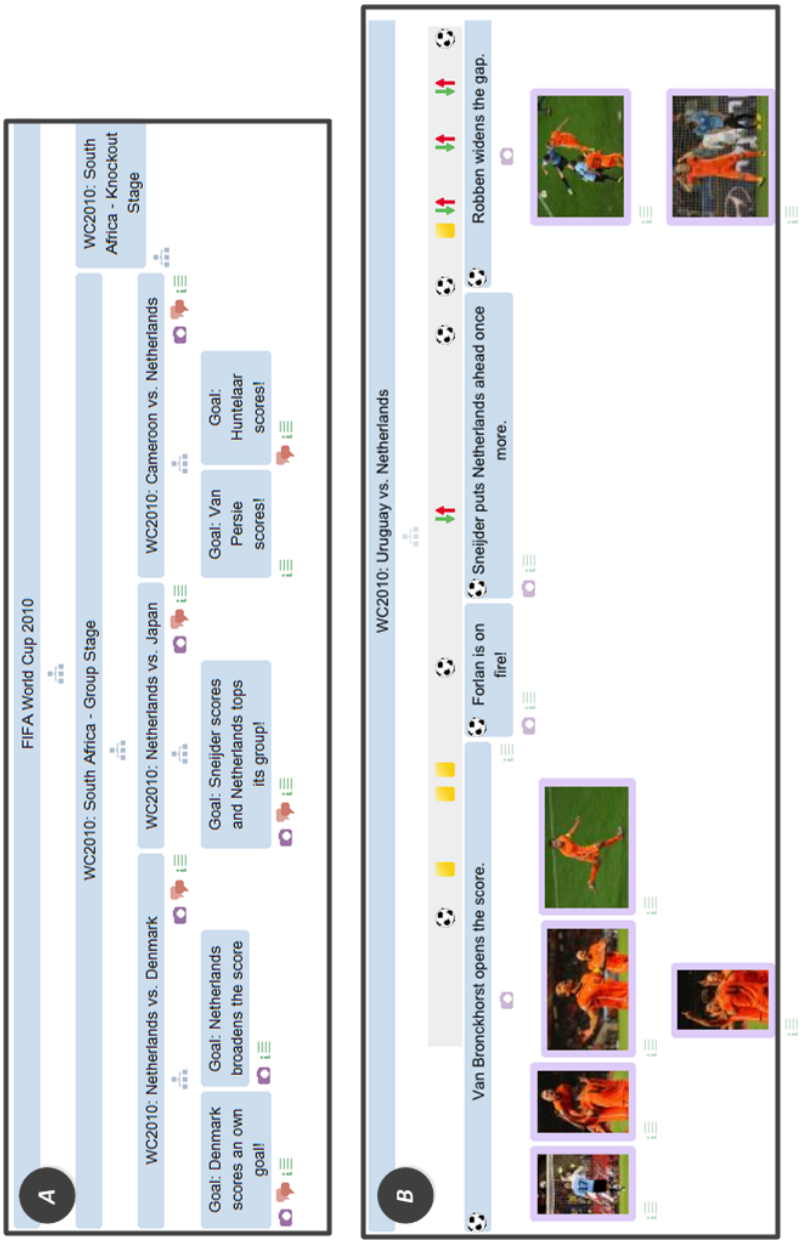


Fig. 2. A: A filtered and partly collapsed representation of the 2010 soccer World Cup as a hierarchy of events (explanation in text). B: The user has zoomed in on a single game and clicked on the "media" links for two goals, so as to be able to compare the associated media.

2 The GLOCAL User Interface

2.1 Hierarchical Browsing of Events

We will now introduce the GLOCAL interface, a web-based interface implemented with the GOOGLE WEB TOOLKIT[5]. It makes use of REST services to access media and event structures.

Figure 2A illustrates several characteristics of the interface's visualizations and functionality. First, the hierarchical structure of the football tournament can be seen. The subevent link (⁂) connects an event with its subevents: By clicking on this link, the user can cause the subevents to toggle back and forth between being hidden and being displayed. In this figure, the entire "Knockout Stage" on the right has been collapsed to a single node, since the user wants to focus on the Group Stage.

Even the Group Stage contains much too many events to display at once. But the interface offers a filtering functionality, like that found in some mind-mapping tools, which enables the user to specify that only events that fulfill certain criteria are to be displayed—along with their superevents. In this figure, the user has chosen to focus on goals scored by the Netherlands team.

If the user wants to focus on one (complex) event in the hierarchy, he or she can click on an icon within the event node[6] to cause it to become the root node of the hierarchy. For example, in Figure 2B the game between Uruguay and the Netherlands has become the root node. It is now feasible to display subevents at a finer-grained level, where applicable using domain-specific symbols that correspond to the types of the events in question. Also, in the other direction, the user can step upwards in the hierarchy to include the parent node in the current view—and hence also the sibling nodes and their descendants, insofar as they match the current filters.

2.2 Organization of Media

Figure 2B also illustrates how, by clicking on the "media" link (📷), a user can cause thumbnails of the media associated with a given event to be displayed (or a selection of these thumbnails, if the number is large).[7] Figure 2B illustrates how this organization makes it possible, for example, to compare media associated with two of the events even if there are a number of events between them.

[5] http://code.google.com/p/google-web-toolkit/

[6] The icon that the user clicks on, along with several analogous icons, becomes visible only when the cursor hovers over the node. In this way, rich functionality can be offered with a minimum of clutter in the interface.

[7] As is usual with thumbnails, it is possible to click on a thumbnail to have a larger version of the media item displayed elsewhere on the screen. Currently the media are photos and videos, but the GLOCAL project is also working with other types of media item, such as automatically generated transcripts of audio files.

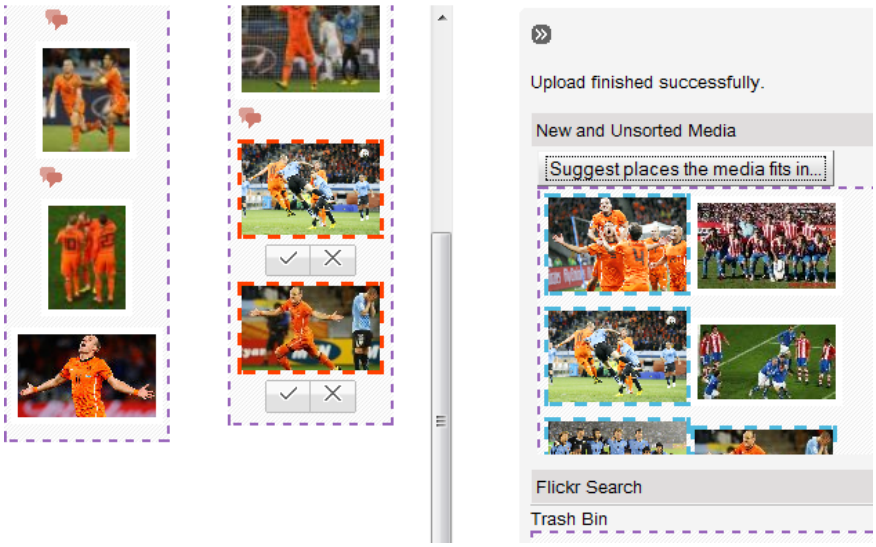


Fig. 3. The user has uploaded the media shown on the right and can now drag them into the appropriate boxes on the left to associate them with events.

(The media on the left that have checkboxes are ones that the system has tentatively aligned in response to the user's clicking on the suggestion button on the right.)

2.3 Importing and Aligning Media

Consider a user who has taken some photos and videos of the game between the Netherlands and Uruguay and wants to introduce them into the platform. Figure 3 shows how the user can upload media into a sort of inbox on the right-hand side of the screen. The system then displays a dotted box for each of the existing events, so that the user can drag a thumbnail into a box to indicate that it belongs to the event in question.

Since the system will often be able to make a good guess about the event that a given media item depicts, the interface also allows the user to ask the system to suggest an alignment of media to events (by clicking on the link "Suggest places ..."). [8] In the cases where the system has a recommendation, the system places a thumbnail tentatively in the dotted box for one of the events on the left. The user can then accept or reject the system's suggestion by clicking on one of two icons associated with the thumbnail.

Instead of a uploading media from a local computer, the user can import media from another site such as FLICKR by formulating a search query which is passed to FLICKR's

[8] Partners in the GLOCAL project are exploring various methods for suggesting alignments of media items to events, including image analysis and the use of spatial and temporal metadata. Whatever method is used, its accuracy is likely to be imperfect, making it natural to put users in the loop in some way.

API. The media retrieved in this way are placed in the user's inbox, where they can be subjected to the treatment just described.

Media contributed in this way become available for sharing with other users; they also provide the system with more information about the events and media that it already has.

2.4 Narrative Plus Visualization

Figure 4 illustrates a novel use of event representations that was suggested by our collaboration with the AGORA project.[9] Instead of merely commenting on individual events or media, a (professional or amateur) user can create a textual *narrative* and then provide illustrations of it by supplying links to relevant *views* of the event representation. Each view shows a subset of the events and the associated media and metadata which the author of the narrative can (a) specify interactively by applying filters and clicking on aggregator nodes and then (b) save with a bookmark (much as users of GOOGLE MAPS can save a view of a map with a bookmark, which can be embedded in a web page or emailed to another user). The reader of the narrative can interact with each view in the same way, in particular exposing information that was not visible in the view as specified by the author. In the example in Figure 4, a Netherlands fan might choose to have the cautions incurred by the Spanish team displayed.

This style of interaction is reminiscent of the increasingly popular trend of *media curation*, which is supported by platforms like STORIFY[10] and OURSTORY,[11] which supports the creation of timelines. A difference is that each view consists not of arbitrary content (e.g., photos or TWITTER feeds) that has been acquired from somewhere in the internet but rather of a specified view of a very large content repository. An important consequence of this difference is that the reader of a narrative is not restricted to contemplating the provided illustrations but can interact with the visualization.

We believe that this approach will (a) enable news agencies quickly to create interactive illustrations of their news stories (in particular, longer stories that cover a number of related events) and (b) enable amateur users to provide richer forms of user-generated content (including, for example, personal essays on complex events such as political campaigns and wars, supported by interactive visualizations). Once a large number of illustrated narratives of this sort exist, it should be possible to mine them in various ways to support new forms of searching and browsing.

3 Lessons Learned

3.1 Ongoing Evaluations

User testing of the GLOCAL interface has so far been formative rather than summative. Early mockups created with the MINDMANAGER software were tested with 2–3 users at a time, each test yielding feedback about the perceived usefulness of the

[9] http://agora.cs.vu.nl/
[10] http://storify.com
[11] http://www.ourstory.com/

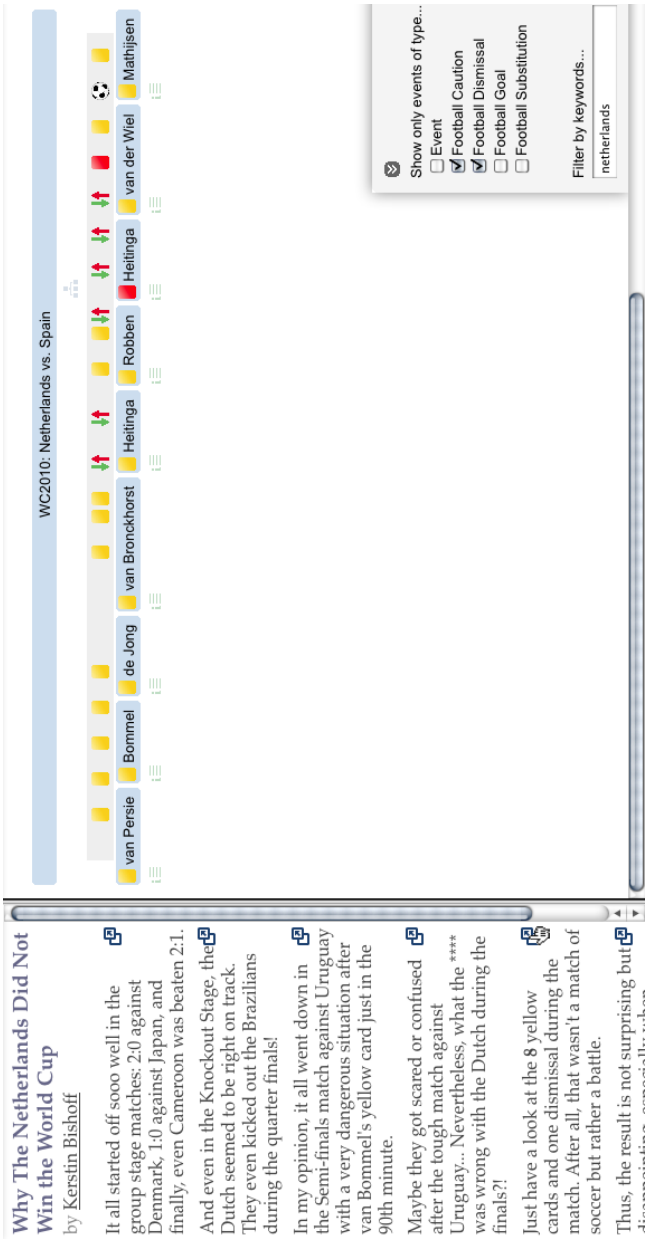


Fig. 4. Example of the use of the "narrative plus visualization" functionality offered by the interface.

(Each time the reader clicks on one of the hyperlinks to the right of the text, a previously specified view *of the events and associated media is shown in the right-hand panel—in this example, a visualization of the "8 yellow cards and one dismissal" incurred by the Dutch team.)*

functionality and ideas about how to improve the functionality and the visual representations. A more recent test was conducted with a version of the prototype much like the version described above. It took the basic form of a contextual inquiry ([4]): Four student-age users with typical amounts of experience with media exchange systems and social networks were observed and unobtrusively questioned as they performed various tasks with the interface. They were then interviewed about their experience.

One general result was that the representation of numerous events within a single hierarchy was seen to have advantages over the distribution of event representations over numerous hierarchically hyperlinked pages: Since the hierarchy remains basically visible at all times while the user is interacting with the complex event in question, the participants found it easier to remain oriented within the event structure. The participants stated that they could imagine using a similar interface for dealing with media and information concerning sports events, upcoming and past cultural events, politics and current events, and private events such as weddings.

The way in which controls for some functions remain invisible until they are hovered over implies that it takes a few minutes for users to become aware of all of the available functionality. But from then on, users know that they need only hover over a node of interest to be reminded of the available functions. Making these controls more conspicuous appears undesirable; in fact most of the suggestions made by the users of this and earlier versions concerned ways of reducing clutter in the interface.

At the time of this writing, a much larger-scale evaluation is being prepared in collaboration with GLOCAL partner AFP. It will involve interaction with representations of media about a set of current events: the recent uprisings in northern Africa. At the time of the workshop, it will be possible to provide some information about this evaluation study.

3.2 Overview of Contributions

This paper has aimed to contribute (at least to some extent) to each of the four questions of this workshop that concern the exploitation of events for the provision of new or improved services:

1. How can event representations be better exploited in support of activities like semantic annotation, semantic search, and semantically enhanced browsing?

We have illustrated how, when media are closely related to events, organizing the media within an interface in terms of events opens up new and improved possibilities for search, browsing, and annotation. Essentially, the benefits are analogous to those that come from organizing media in terms of geographical maps and/or timelines. The novel contribution of this paper is to show how additional functionality such as the support for interaction with event hierarchies and flexible filtering supports these activities.

2. What application areas for semantic technologies can benefit from an increased use of event representations?

To date, two application areas that are illustrated by the GLOCAL project are (a) the provision of news and media by news agencies (GLOCAL's partners include AFP and CITIZENSIDE); and (b) the exchange of media among nonprofessional users, as well as contributions by such users to media offerings of the type just mentioned. The way in which the GLOCAL interface encourages users to contribute media and comment on

existing media distinguishes it from previous methods used for exploiting event representations for interaction in the semantic web.

3. How can we improve existing methods for visualizing event representations and enabling users to interact with them in semantic web user interfaces?

Though the semantic web community has already produced impressive and useful techniques for visualizing events and supporting interaction with them, the GLOCAL user interface augments these approaches in several ways: The novel use of functionality typical of mind mapping applications introduces new ways of interacting with event hierarchies. The idea of enabling users to illustrate narratives with interactive event representations is a novel approach to event visualization that shifts some of the representational burden from the graphical visualization to natural language text. It is true that, with enough imagination and effort, it may be possible to visualize just about any relationship between two events, even if they are temporally and spatially far apart (see, e.g., the visualizations of this sort offered by [1]). But if statements about such relationships are subjective and intended only for consumption by a human user—not for automatic processing and inference—it may be most natural to have them expressed in natural language, reserving formal representation for the relations that form the backbone of the system's internal representations and external visualizations.

4. What requirements for event detection and representation methods are implied by advances in methods for exploiting events?

The main requirement introduced by the GLOCAL interface is the need to detect and represent the *subevent* relation. Representation is straightforward, and in some domains characterized by clearly structured complex events (e.g., sports tournaments, conferences), detection may also be easily automated. In domains where any hierarchical structure is not defined in advance but rather emerges as events evolve—for example, political and military uprisings such as those that have occurred in northern Africa in 2011—the identification of events and their subevents is likely to require sophisticated automatic analysis and/or human intervention. But this conclusion need not be discouraging, given that there exist many professionals and amateurs (e.g., journalists and historians) who are more than willing to apply their knowledge and skill to the analysis and interpretation of events.

References

1. André, P., Wilson, M.L., Russell, A., Smith, D.A., schraefel, m.: Continuum: Designing timelines for hierarchies, relationships and scale. In: UIST '07 Proceedings of the 20th Annual ACM symposium on User interface Software and Technology. (2007) 101–110
2. Chen, J., Hibino, S.: Reminiscing View: Event-based browsing of consumers photo and videoclip collections. In: Tenth IEEE International Symposium on Multimedia. (2008) 23–30
3. Hirsch, C., Hosking, J., Grundy, J.: Interactive visualization tools for exploring the semantic graph of large knowledge spaces. In: Proceedings of the Workshop on Visual Interfaces to the Social and Semantic Web (VISSW 2009), in conjunction with IUI 2009, Sanibel Island, FL (2009)
4. Beyer, H., Holtzblatt, K.: Contextual Design: Defining Customer-Centered Systems. Morgan Kaufmann, San Francisco (1998)

Linked Open Piracy

Willem R. van Hage[1], Véronique Malaisé[2], and Marieke van Erp[1]

[1] Department of Computer Science, VU University Amsterdam
{W.R.van.Hage,Marieke.van.Erp}@vu.nl
[2] Elsevier Content Enrichment Center (CEC)
v.malaise@elsevier.com

Abstract. There is an abundance of semi-structured reports on events being written and made available on the World Wide Web on a daily basis. These reports are primarily meant for human use. In this paper we present a new linked data set and a method for automatically adding such RDF metadata to semi-structured reports to speed up the creation of geographical mashups and visual analytics applications. We showcase our method on piracy attack reports issued by the International Chamber of Commerce (ICC-CCS). We show how the semantic representation makes it possible to easily analyze and visualize the aggregated reports to answer domain questions. Our pipeline includes conversion of the reports to RDF, linking their parts to external resources from the Linked Open Data cloud and exposing them to the Web.

1 Introduction

In this paper we present a new data set on the Web of Data, Linked Open Piracy (LOP), how it was constructed, and how it can be used to answer complex questions about piracy. We expose descriptions of piracy attacks at sea published on the Web by the International Chamber of Commerce's International Maritime Bureau (ICC-CCS IMB)[3] and the US National Geospatial-Intelligence Agency (NGA)[4] as Linked Data RDF[5].

LOP can be seen as an Open Government Data[6] initiative for intergovernmental data. The goal of Open Government Data is to reduce the time to do analytics and mashups with open government data. The piracy reports are, like most open government data, published in a human readable format[7]. We show how we can reduce the commonly acknowledged bottleneck of data preprocessing time in the workflow from question to answer. This format and type of publication (following a given pattern for a year of publication, daily update of the webpage) makes it an ideal test case for automatic RDF event extraction; the

[3] http://www.icc-ccs.org/home/imb
[4] NGA, http://www.nga.mil/portal/site/maritime/
[5] LOP, http://semanticweb.cs.vu.nl/lop
[6] http://data-gov.tw.rpi.edu/wiki/Open_Government_Data
[7] A notable exception is data.gov.uk where the data are exposed directly as machine friendly RDF.

topic of the reports is also of contemporary socio-economic concern and are related to research questions that go beyond what classic data mining can easily answer. We therefore chose to take this example as a showcase for the feasibility and usability of event extraction coupled with novel research question answering methods.

We represent LOP data in RDF with the Simple Event Model (SEM) [7] and demonstrate that an event model is not only an intuitive way of representing (inter)governmental data, but also a powerful tool for data integration. We evaluate the usefulness of SEM as a model for Open Government Data by answering complex domain questions derived from authorities in the domain of piracy analysis, namely UNITAR UNOSAT and the ICC-CCS IMB. We use SWI-Prolog[8] to extract event descriptions from the web, represent them in SEM and store them in a ClioPatria RDF repository [10] extended with the SWI-Prolog space package [8] for spatial and temporal indexing. The entire ICC-CCS data set is hosted as Linked Data, all URIs in the data set are resolvable. A SPARQL endpoint is available at `http://semanticweb.cs.vu.nl/lop/sparql/`.

This paper is organized as follows. In Section 2, we show how we created RDF event descriptions from web pages. In Section 3, we discuss the modeling of the events in SEM. In Section 4, we show example domain questions from UNOSAT that can easily be answered using our event representation. In Section 5, we discuss related work and in Section 6, we conclude with a discussion and future plans.

2 Screen Scraping

We start crawling of the ICC-CCS IMB webpage with the links to the yearly archives in the menu of the Live Piracy Map page. Figure 1 (top) shows what an ICC-CCS piracy report looks like. The reports are semi-structured, and concern seven predefined types of events: Hijacked, Boarded, Robbed, Attempted, Fired Upon, Suspicious (vessel spotted) and Kidnapped. The reports contains a field for the vessel type of the ship broadcasting the report; although the types of the vessels are often recurring, this field is filled manually, which gives rise to spelling variations (e.g., firedupon vs fired upon) and a lack of certainty in terms of coverage; a new ship type could be filled in any day. The description of the event itself is done in full text, without a specific formatting except that it is preceded, in the same field, by the geographic and temporal coordinates of the event. The geographic and temporal coordinates are repeated in an independent field each.

For each of these pages we follow all the links in the descriptions of the placemarks on the overview map, returning us one semi-structured description pages for each event. We fetch the various fields from these pages using XPath queries and Prolog rules for value conversion and fixing irregularities. In this way we fetch: (1) The IMB's attack number, which consists of the year and a

[8] SWI-Prolog, `http://www.swi-prolog.org/`

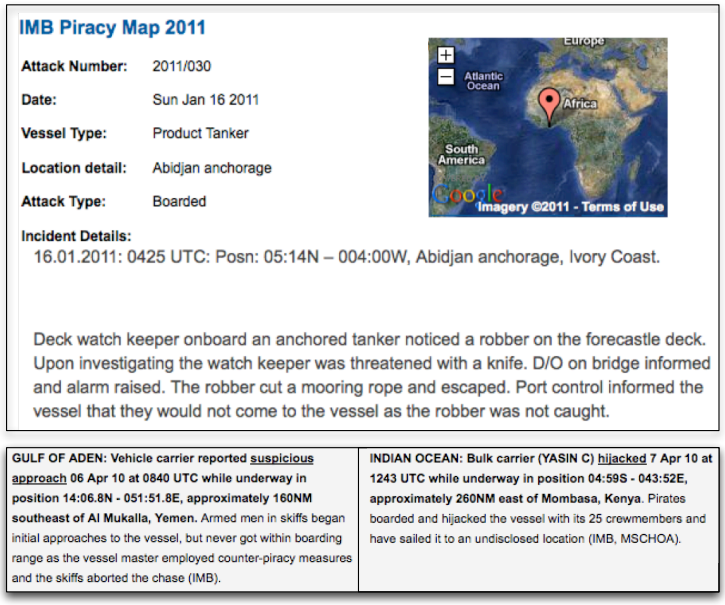


Fig. 1. Example of an IMB piracy report (top) and two NGA piracy reports (bottom)

counter. From this we generate an event identifier by prepending a namespace and by appending a suffix whenever there are duplicate attack numbers in a year; (2) The date of the attack, which we convert to ISO 8601 format; (3) The vessel type, which we map to URIs with rules that normalize a few spelling variations of the types. (4) The location detail, which we use as a label for the place of the event; (5) The attack type, which we map to URIs in the same way as the vessel type; (6) The incident details, which we convert to a comment describing the event itself. The first line is split into a time and place indication. These are used as backup sources to derive the date and location, should the parsing of fields 2, 4 and 7 fail; (7) The longitude and latitude of the placemark on the map insert. These are used as coordinates of a generated anonymous place (i.e., without a URI) for the event. The time fetched from the date (3) or narrative (6) field has a number of different representations in the source pages. Some time indications are in local time, while others are in UTC. Often there is no indication of the time zone. For many events the indicated time is 00:00 (midnight) to denote the time of attack is unknown. These inconsistencies in the time notation, in combination with the fact that there are few events on the same day, led us to the decision to use the date without a time indication whenever there is ambiguity about the time.

To demonstrate that representing extracted events in SEM aids the integration of data sources, we take another set of piracy reports and integrate these with the IMB reports. For this, we use the Worldwide Threat to Shipping re-

ports by the US National Geospatial-Intelligence Agency describing 36 piracy events between 26 March 2010 and 16 April 2010. 31 of these events overlap with the IMB reports. The remaining 5 come from other sources: Reuters (2)[9], UKMTO[10], MSCHOA[11], and ReCAAP[12]. These reports are (re)posted on many websites, some of which are plain-text representations of the reports, while others add some additional layout tags to separate the place, time, and state of the ship during the attack from the narrative. Two example NGA reports are shown in Figure 1 (bottom).

By changing the XPath and grammar rules to suit the different structure of the NGA reports we were able to recognize the same 7 attributes we got from the IMB website. The event terminology is nearly the same as on the IMB website, except there is a distinction between boardings and robberies. There is also some extra information in 34 of the 36 reports about the state of the ship during the attack, (e.g., moored or underway). For some of the events there are no explicit coordinates of the location of the event, but there is a textual description, for example, "approximately 150NM northwest of Port Victoria, Seychelles". For these events we look up the coordinates of Port Victoria using GeoNames[13], which returns RDF. From this location we use trigonometry along the geoid with the haversine formula in the specified direction. For example, in the case of 150NM northwest we compute the coordinates 150 minutes of angle at a bearing of 315 degrees. We treated time in the NGA reports in the same way as in the IMB reports, reducing them to an ISO 8061 date.

We match the NGA reports to the IMB reports by picking the nearest event that occurred on the same day that has compatible actor types, i.e., when the types are not the same, one has to be sem:subTypeOf the other. This enables us to automatically map 30 of the 31 overlapping reports correctly. We store these matches with an owl:sameAs property between the two matching events. We believe the single unmatched report was mistakingly identified as a distinct IMB report, because it is extremely similar to another report (the same date, place, time, victim vessel type, and similar narrative) which has a matching IMB report. Therefore, we believe there should only have been 30 overlapping reports, which we were all able to match.

3 Event Representation in SEM

We use the set of 7 elements (see Section 2) extracted per report to generate a semantic event description using SEM. We generate a URI for the event described in each report and a URI for the victim ship, which we represent as a sem:Actor, based on the IMB attack number (nr. 1). The date (nr. 2) is attached to the

[9] Reuters, `http://www.reuters.com/`
[10] UK Maritime Trade Operations,`http://www.mschoa.org/Links/Pages/UKMTO.aspx`
[11] The Maritime Security Center – Horn of Africa, `http://www.mschoa.org/`
[12] The Regional Cooperation Agreement on Combating Piracy and Armed Robbery against Ships in Asia, `http://www.recaap.org/`
[13] GeoNames search, `http://sws.geonames.org/search`

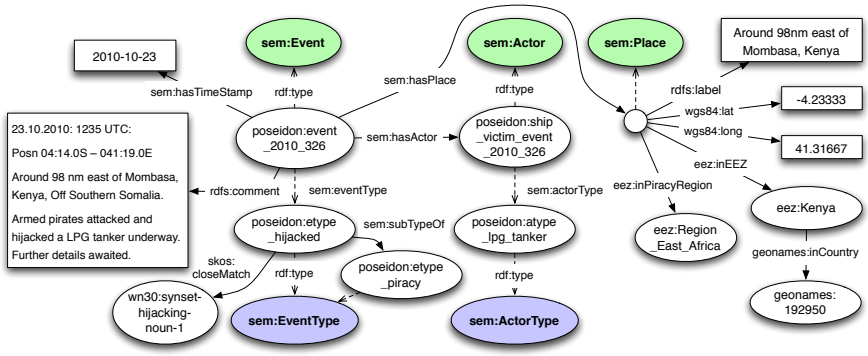


Fig. 2. The complete RDF graph of a piracy report modeled in SEM including mappings to types in WordNet 3.0, a VLIZ exclusive economic zone, its corresponding GeoNames country, and its Piracy Region.

sem:Event by means of the sem:hasTimeStamp property. The sem:hasTimeStamp datatype property was chosen over the sem:hasTime object property, because we do not need type hierarchies over time instances to answer our domain questions. The vessel type (nr. 3) is typed as a sem:ActorType attached to the victim ship sem:Actor with the sem:actorType property, a subproperty of rdf:type. The location detail (nr. 4) is made a rdfs:label of the blank node representing the location of the attack. We chose not to use the Exclusive Economic Zones (EEZs)[14] (usually defined as 200 nautical miles from the coast of the nearest state), or the GeoNames identifier of the nearest relevant place, as the URI of the location of the attack because this would have removed the distinction between the exact location of the attack and the more general region. We did use the EEZs for an initial partitioning of the world into regions (e.g. Gulf of Aden, Carribean). The remaining surface of the earth, including the international waters and inland seas is partitioned based on the nearest EEZ. The area nearest to an EEZ is assigned a new URI, e.g., the international waters off the coast of Liberia and closest to Liberia's EEZ (i.e., not closest to Ascension's, Côte d'Ivoire, Sierra Leone's, or Saint Helena's EEZs) is assigned the URI eez:Nearest_to_Liberia. Based on the distribution of the piracy events, we grouped particular sections of the world together. This grouping is only specific to the piracy event domain.

The attack type (nr. 5) is modeled analogously to the vessel type as a sem:EventType, which is attached to the event using the sem:eventType property. The event type *robbery* that we found in the NGA set was modeled as a sem:subTypeOf the IMB event type *boarding*. The *mooring* and *underway* vessel states are modeled as additional event types of the piracy event using sem:eventType properties attached to the event. All event types used in this data set are sem:subTypeOf the piracy event type, poseidon:etype_piracy. The narrative of the report (nr. 6) is attached to the event as a rdfs:comment. The WGS84

[14] http://www.vliz.be/vmdcdata/marbound/

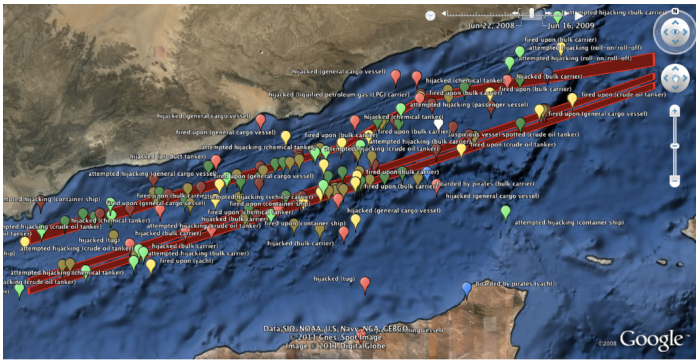


Fig. 3. Attacks plotted in Google Earth.

coordinates (nr. 7) are assigned to the blank node with the W3C WGS84 vocabulary. Additional ship names are attached to the sem:Actor using the ais:name property, a domain-specific label for ship names.

We create local URIs to represent the types of the extracted events and the types of their participants (e.g., poseidon:etype_hijacked or poseidon:atype_yacht). The SEM piracy events are aligned with WordNet 2.0[15], 3.0[16], OpenCyc[17] and Freebase[18]. WordNet gives us the advantage of relating different lexical variations to a unique URI e.g., mapping *highjacking* and *hijacking* to *hijacking*. This can also be used to automatically transform piracy descriptions to types. As WordNet has a hierarchy of hyponym relations between synsets (e.g., a *tankership* is a hyponym of *cargoship*) we can do hyponym inference.

We can not map all of our types to any one of these three vocabularies, but by mapping to all three of them we get a good coverage of our domain-specific type vocabulary. Our data set contains 73 ActorTypes and 26 EventTypes, which is too few to make it worthwhile to use an automatic mapping method, so we manually created the following mappings: 70 skos:closeMatch (24 to Freebase, 24 to OpenCyc, 25 to WordNet);10 skos:broadMatch (5 to OpenCyc, 4 to WordNet, 1 to Freebase); 33 skos:relatedMatch (13 to OpenCyc, 11 to WordNet, 9 to Freebase). A "related" relation hold for example between WordNet's *to fire* and the event type *fired upon*, because *to fire* only conveys part of the meaning.

4 Answering Domain Questions

In this section, we show how the SEM representation simplifies answering domain questions through visualizations and analyses. We first show how the enriched

[15] WordNet 2.0, `http://www.w3.org/2006/03/wn/wn20/`
[16] WordNet 3.0, `http://semanticweb.cs.vu.nl/lod/wn30/`
[17] OpenCyc, `http://sw.opencyc.org/`
[18] Freebase, `http://{www|rdf}.freebase.com/`

data could be used to recreate UNOSAT questions. Then we show the added value of the mappings and hierarchies in an additional set of domain questions.

4.1 Rebuilding UNOSAT Reports

The analysis performed and compiled for the UNOSAT reports [5] have mostly been carried out manually and sometimes with the aid of a GIS. The analyses are thorough and insightful, but do require painstaking manual sifting through the data because only the unprocessed attack reports are used. Human researchers then plot these data on maps, and assign attack types to them. With the RDF version and the mappings to the VLIZ economic zones and geospatial reasoning the analyses that require a combination of data sources can be sped up immensely. SPARQL and Prolog rules make many complex questions as simple as a graph query.

The conclusion of map 1 in the UNOSAT 2009 Q1 report, namely that the attacks have shifted southward and extended further east-west along the axis of the International Recommended Transit Corridor (IRTC)[19] can be reproduced by combining plotting the attacks on a map along with information about the IRTC. This is illustrated in Figure 3, a time animation in KML is available online[20]. Although more coastguard and marine vessels are present in the recommended corridor, pirates also know that there are more ships there, hence more chances of finding a victim.

4.2 Additional Questions

We start with an easy visualization of number of attacks per region per year (top left Figure 4). We can see that the most active regions are the Gulf of Aden, Indonesia, India and East Africa. The graph also shows that Indonesia used to be the most active region, but sometime in 2007 activity in the Gulf of Aden and East Africa have become the regions with most piracy activity.

Although the narrative section of each report are not split up and represented in RDF yet, we can give some ideas on differences in weapon use by comparing the number of occurrences of the terms "guns" and "knives" in the different reports. For instance, there are no reports that mention knives in the Gulf of Aden region at all, while there are 109 in the Indonesia region while there are 85 that mention guns in the Gulf of Aden and only 25 in Indonesia. The pie charts in Figure 4 show an overview of five weapons types. In order to properly analyse these we will use more sophisticated NLP techniques in future work.

If we further look into the four most active areas, we can use the ship type mapping to compare differences in ships attacked in different regions. The stacked bar chart in Figure 4 immediately highlights the difference between Indonesia and the other areas, namely that in the Indonesia region far more tugs

[19] `http://www.icc-ccs.org/news/163-coalition-warships-set-up-maritime-security-patrol-area-in-the-gulf-of-aden`

[20] `http://semanticweb.cs.vu.nl/poseidon/piracy_reports_2005-2010.kmz`

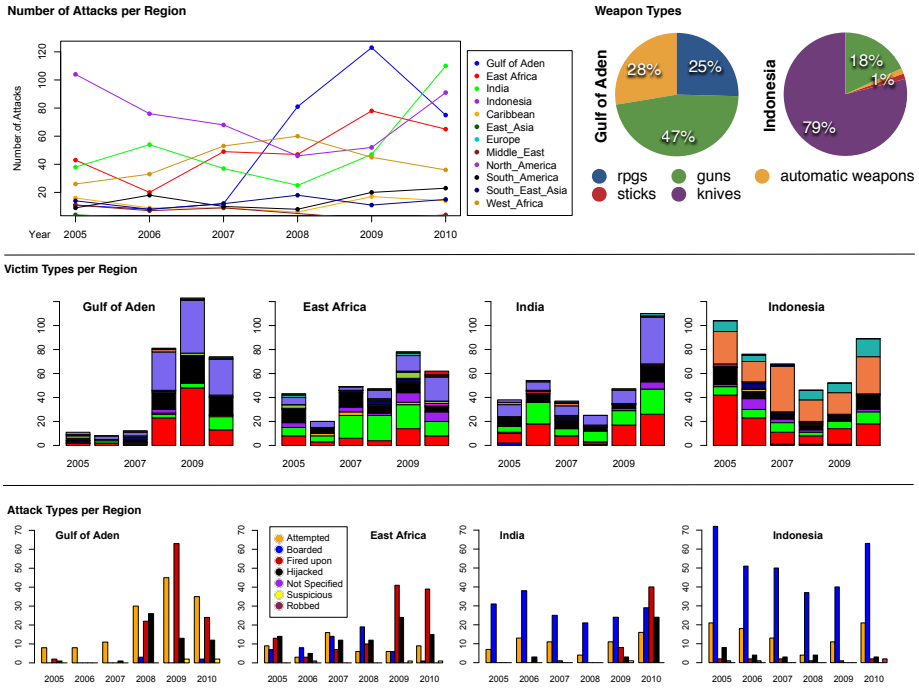


Fig. 4. Number of attacks reported per region per year, weapon types per region, victim types per region and attack types per region.

are attacked than in the other regions. In the Gulf of Aden, for a larger number of attacks the ship type of the victim is not known. Interestingly, the attacks on bulk carriers has been declining in the Asian regions until 2009, whereas it was on the rise in the African regions. In order to explain this, extra information is needed, for example on the number of ship movements in these areas. Unfortunately, such data is not openly available.

We can also split out the attacks by types of attack to see whether pirates take a different approach in different regions. Plotting these statistics in a graph, split out per region, has the advantage that one can quickly see the differences, whereas plotting these on a map still requires interpretation from the user. Here, the region clustering shows its merit. In the last series of charts in Figure 4, one can see that significant differences exist between the regions in the types of attacks. In Asia, for example, far more often ships are boarded (which often also means robbed) than in the African regions. In the Gulf of Aden attacks have become more aggressive and more often victim ships are fired upon. In the Gulf of Aden, also more attempted hijackings occur than elsewhere.

5 Related Work

This work essentially describes an Open Government Data project, like data.gov [2] and data.gov.uk [3], with the exception that data are intergovernmental. The case we present deals with scraping event description from web pages. In the past we have done similar work with different types of data sources, such as user ratings of museum pieces [9], historical events [6], and Automatic Identification System NMEA ship data for the recognition of ship behavior from trajectories and background knowledge from the Web [11]. This is accomplished with the SWI-Prolog space package [8], which is similar to Franz Inc.'s Common Lisp-based AllegroGraph system[21]. We use SEM to describe our events, because it is a simple but not spartan model. A very similar model is LODE, which has been used for the extraction of events from Wikipedia timelines [4]. Both SEM and LODE focus on the *"Who does what, where and when?"*, but LODE does not contain a typing system, whereas SEM does. An example of a much richer event model is part of the CIDOC-CRM. The purpose of CIDOC-CRM is the integration of meta data about (museum) artifacts. A description of an integration method that, like the work presented in this paper, also combines space, time and semantics, using CIDOC-CRM can be found in [1]. The SEM specification[22] contains mappings to LODE and CIDOC-CRM.

6 Conclusions and Future Work

We have shown that the ideas behind the Open Government Data initiative can also be applied to information sources from intergovernmental organizations without the need for changing their entire information workflow. Automatic conversion of online open data can bring their data to the Web and help these organizations with their business by making it easier to answer questions about their data. In this case study, the representation we use is the Simple Event Model, which helps to integrate spatio-temporal reasoning with web semantics. SEM has an appropriate level of abstraction for the integration of piracy event data: it is more general than the differences between the data sources taken into account in this paper, but still specific enough to answer domain-specific questions. This modularity of the flexible event extraction set allows us to combine data sources with relatively little change in the code base. We have shown that different data sources provide different aspects of an event, and their combination allows for interesting and serendipitous data analysis. As future work, we aim at doing further natural language processing on each report's content description in plain text in order to extract more information: the types of weapons used during the attack, the number of pirate boats and pirates, the intervention of a coalition war ship or helicopter, the outcome of the attack which would help to answer even more domain questions. Also, we would like to investigate the possibility to interlink the Linked Open Piracy data set with news items on the

[21] `http://www.franz.com/agraph/allegrograph/`
[22] SEM, `http://semanticweb.cs.vu.nl/2009/11/sem/`

World Wide Web. This would provide additional background information to the semantic event descriptions, but also a semantic description of the news articles on the Web.

7 Acknowledgements

This work has been carried out as a part of the Poseidon project and the Agora project. Work in the Poseidon project was done in cooperation with Thales Nederland, under the responsibilities of the Embedded Systems Institute (ESI). The Poseidon project is partially supported by the Dutch Ministry of Economic Affairs under the BSIK03021 program. The Agora project is funded by NWO in the CATCH programme, grant 640.004.801. We would like to thank Davide Ceolin, Juan Manuel Coleto, and Vincent Osinga for their significant contributions. We thank the ICC-CCS IMB and the NGA for providing the open piracy reports.

References

1. G. Hiebel, K. Hanke, and I. Hayek. Methodology for CIDOC CRM based data integration with spatial data. In *38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology*,
2. D. D. Li Ding, D. L. McGuinness, J. Hendler, and S. Magidson. The data-gov wiki: A semantic web portal for linked government data. In *8th International Semantic Web Conference (ISWC 2009)*, 2009.
3. T. Omitola, C. Koumenides, I. Popov, Y. Yang, M. Salvadores, M. Szomszor, T. Berners-Lee, N. Gibbins, W. Hall, M. C. Schraefel, and N. Shadbolt. Put in your postcode, out comes the data: A case study. In *7th Extended Semantic Web Conference (ESWC 2010)*, 2010.
4. R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In *4th Annual Asian Semantic Web Conference (ASWC'09)*,
5. UNOSAT. Analysis of somali pirate activity in 2009. `http://unosat-maps.web.cern.ch/unosat-maps/SO/Piracy/2009/UNOSAT_Somalia_Pirates_Analysis_Q1_2009_23April09_v1.pdf`, April 2009.
6. M. van Erp, J. Oomen, R. Segers, C. van den Akker, L. Aroyo, G. Jacobs, S. Legêne, L. van der Meij, J. van Ossenbruggen, and G. Schreiber. Automatic heritage metadata enrichment with historic events. In *Museums and the Web 2011*, 2011.
7. W. R. van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics*, 9(2):128–136, July 2011.
8. W. R. van Hage, J. Wielemaker, and G. Schreiber. The space package: Tight integration between space and semantics. *Transactions in GIS*, 14(2), 2010.
9. Y. Wang. *Semantically-Enhanced Recommendations in Cultural Heritage*. PhD thesis, Technische Universiteit Eindhoven, 2011.
10. J. Wielemaker, Z. Huang, and L. van der Meij. *SWI-Prolog and the web*, volume Theory Theory and Practice of Logic Programming. Cambridge, pages 363–392. Cambridge University Press, 2008.
11. N. Willems, W. R. van Hage, G. de Vries, J. Janssens, and V. Malaisé. An integrated approach for visual analysis of a multi-source moving objects knowledge base. *International Journal of Geographical Information Science*, 24(9):1–16, Sept. 2010.

Using Events for Content Appraisal and Selection in Web Archives*

Thomas Risse[1], Stefan Dietze[1], Diana Maynard[2], Nina Tahmasebi[1], and Wim Peters[2]

[1] L3S Research Center, Hanover, Germany
{risse|dietze|tahmasebi}@L3S.de

[2] University of Sheffield, Sheffield, UK,
{diana|w.peters}@dcs.shef.ac.uk

Abstract. With the rapidly growing volume of resources on the Web, Web archiving becomes an important challenge. In addition, the notion of community memories extends traditional Web archives with related data from a variety of sources on the Social Web. Community memories take an entity-centric view to organise Web content according to the events and the entities related to them, such as persons, organisations and locations. To this end, the main challenge is to extract, detect and correlate events and related information from a vast number of heterogeneous Web resources where the nature and quality of the content may vary heavily. In this paper we present the approach of the ARCOMEM project which is based on an iterative cycle consisting of (1) targeted archiving/crawling of Web objects, (2) entity and event extraction and detection, and (3) refinement of crawling strategy.

Keywords: Event Detection, Crawler Guidance, Web Archiving

1 Introduction

Given the ever increasing importance of the World Wide Web as a source of information, adequate *Web archiving* and *preservation* has become a cultural necessity in preserving knowledge. However, in addition to the "common" challenges of digital preservation, such as media decay, technological obsolescence, authenticity and integrity issues, Web preservation has to deal with the sheer size and ever-increasing growth rate of Web data. Hence, selection of content sources becomes a crucial task for archival organizations. Instead of following a "collect-all" strategy, archival organizations are trying to build *community memories* that reflect the diversity of information people are interested in. Community memories largely revolve around *events* and the *entities* related to them such as persons, organisations and locations. These may be unique events such as the first landing on the moon or a natural disaster, or regularly occurring events such as elections or TV serials.

* This work is partly funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270239 (ARCOMEM)

In this work, we refer to an event as a situation within the domain (states, actions, processes, properties) expressed by one or more relations. Events can be expressed by text elements such as:

- verbal predicates and their arguments ("The committee dismissed the proposal");
- noun phrases headed by nominalizations ("economic growth");
- adjective-noun combinations ("governmental measure"; "public money");
- event-referring nouns ("crisis", "cash injection").

Events can denote different levels of semantic granularity, i.e. general events can contain more specific sub-events. For instance, the performances of various bands form sub-events of a wider music event, while a general event like "Turkey's EU accession" has sub-events such as the European Parliament approving Turkey's Progress Report.

In this paper, we provide an overview of the approach we follow in the ARCOMEM[3] project. The overall aim is to create incrementally enriched Web archives which allow access to all sorts of Web content in a structured and semantically meaningful way. In addition to topic-centred preservation approaches, we are exploring event- and entity-centred processes for content appraisal and acquisition as well as rich preservation. By considering a wide range of content, a more diverse archive is created, taking into account a variety of dimensions including perspectives taken, sentiments, images used, and information sources.

To build a community archive from Web content, a *web crawler* needs to be guided in an intelligent way based on the events and entities derived from previous crawl campaigns so that pages are crawled and archived if they relate to a specified event or entity. While at the beginning of any crawl campaign the amount of information is very limited, the crawler needs to learn about the event incrementally, while at the same time it has to decide about following links. Therefore, our approach is based on an iterative cycle consisting of the following steps:

1. Targeted archiving/crawling of Web objects;
2. Entity and event extraction and detection;
3. Refinement of crawling strategy.

To this end, the main challenges are related to the *extraction*, *detection* and *correlation* of entities, events and related information in a vast number of heterogeneous Web resources. While *extraction* covers the identification and structured representation of knowledge about events and entities from previously unstructured material from scratch, *detection* refers to the detection of previously extracted events and entities. Therefore, in contrast to the extraction step, detection takes advantage of existing structured data about events and entities. Both processes face issues arising from the diversity of the nature and quality

[3] ARCOMEM - From Collect-All Archives to Community Memories, http://www.arcomem.eu/

of Web content, in particular when considering *social media* and *user-generated content*, where further issues are posed by informal use of language.

In the following section, we give an overview of related work, and introduce the ARCOMEM approach and architecture in Section 3. Section 4 provides an overview of the event detection mechanisms deployed by ARCOMEM, while we discuss some key challenges in Section 5.

2 Related Work

Since 1996, several projects have pursued Web archiving (e.g. [AL98]). The Heritrix crawler [MKSR04], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet Preservation Consortium (IIPC)[4], is a mature and efficient tool for large-scale, archival-quality crawling.

The method of choice for memory institutions is client-side archiving based on crawling. This method is derived from search engine crawl, and has been evolved by the archiving community to achieve a better completeness of capture and to reduce temporal coherence of crawls. These two requirements come from the fact that, for web archiving, crawlers are used to build collections and not only to index [Mas06]. These issues were addressed in the European project LiWA (Living Web Archives)[5].

The task of crawl prioritisation and focusing is the step in the crawl processing chain which combines the different analysis results and the crawl specification for filtering and ranking the URLs of a seed list. The filtering of URLs is necessary to avoid unrelated content in the archive. For content that is partly relevant, URLs need to be prioritised to focus the crawler tasks to crawl in order of relevancy. A number of strategies and therefore URL ordering metrics exist for this, such as breadth-first, back link count and PageRank. PageRank and breadth-first are good strategies to crawl "important" content on the web [CGMP98,BYCMR05], but since these generic approaches do not cover specific information needs, focused or topical crawls have been developed [CBD99] [MPS04]. However, these approaches have only a vague notion of topicality and do not address event-based crawling.

Entity and event recognition are two of the major tasks within Information Extraction, and have been successfully applied in research areas such as ontology generation, bioinformatics, news aggregation, business intelligence and text classification. Recognising events in these fields is generally carried out by means of pre-defined sets of relations, possibly structured into an ontology, which makes such tasks domain dependent, but feasible. Entity extraction in this case comprises both named entity recognition [CMBT02] and term recognition [BS09,MLP08].

The identification of relations between entities in text is generally performed by means of heuristic, rule-based applications using background knowledge from

[4] http://netpreserve.org/
[5] http://wiki.liwa-project.eu/

instantiated ontologies and lexico-syntactic patterns to establish links between textual entities and their ontological provenance [MFP09a], or a combination of statistical and linguistic techniques [MPB08]. Tools such as Espresso [PP06] and Text2Onto [CLS05] make use of predefined or automatically extracted text patterns in order to structure the domain in terms of classes and relations. Furthermore, shallow parsing techniques such as semantic role labelling [Gil02] characterise the relationship between predicates (relations) and their arguments (entities) on a semantic level by means of roles such as agent and patient. On the other hand, unsupervised machine learning techniques such as TextRunner[BE08] and Powerset[6] scale to the extraction of facts from hundreds of millions of web pages, but they use only very shallow linguistic analysis and may not be so accurate. While PowerSet, for example, uses advanced parsing and some NLP techniques, it does not understand word and phrase meanings in context. In this work, we position our event extraction approach somewhere between the very constrained template-filling approach used in MUC, and the open domain approach of finding new relations over the whole web, used by systems such as TextRunner and Powerset.

In addition, for representation of events and entities we consider Semantic Web and Linked Data-based approaches, as one of our fundamental aims is to expose the generated knowledge in an interoperable and reusable way. We consider in particular Linked Open Descriptions of Events, LODE [STH09], Event-Model-F [ASS09] and the Event Ontology[7]. While LODE and the Event Ontology follow a similar approach to and provide rather lightweight RDF schemas for event description, the Event-Model-F is a more formal OWL ontology which applies the DOLCE Descriptions and Situations pattern by using DOLCE+DnS Ultralight (DUL)[8] as an upper level ontology.

3 Approach and Architecture

3.1 Overall Approach

The goal for the ARCOMEM system is to develop methods and tools for transforming digital archives into community memories based on novel socially-aware and socially-driven preservation models. This will be done by leveraging the Wisdom of the Crowds reflected in the rich context and reflective information in the Social Web for driving innovative, concise and socially-aware content appraisal and selection processes for preservation, taking events, entities and topics as seeds, and by encapsulating this functionality into an adaptive decision support tool for the archivist.

Archivists will be able to trigger interactive and intelligent content appraisal and selection processes in two ways: either by example or by a high-level description of relevant entities, topics and events. Intelligent and adaptive decision

[6] http://www.powerset.com/
[7] http://motools.sourceforge.net/event/event.html
[8] http://www.loa-cnr.it/ontologies/DUL.owl

support for this will be based on combining and reasoning about the extracted information and inferring semantic knowledge, combining logic reasoning with adaptive content selection strategies and heuristics.

The system is built around two loops: content selection and content enrichment. The *content selection* loop aims at content filtering based on community reflection and appraisal. Social Web content will be analysed regarding the interlinking, context and popularity of web content, regarding events, topics and entities. These results are used for building the seed lists to be used by existing Web crawlers. Within the *content enrichment* loop, newly crawled pages will be analysed for topics, entities, events, perspectives, Social Web context and evolutionary aspects in order to link them together by means of the events and entities.

In the following we will focus on the *content selection loop*.

3.2 Architecture

The main tasks of a Web crawler are to download a Web page and to extract links from that page to find more pages to crawl. An intelligent filtering and ranking of links enables focusing of the crawls. We will combine a breadth-first strategy with a semantic ranking that takes into account events, topics, opinions and entities (ETOEs). The extracted links are weighted according to the relevance of the page to the semantically rich crawl specification. The general architecture is depicted in figure 1.

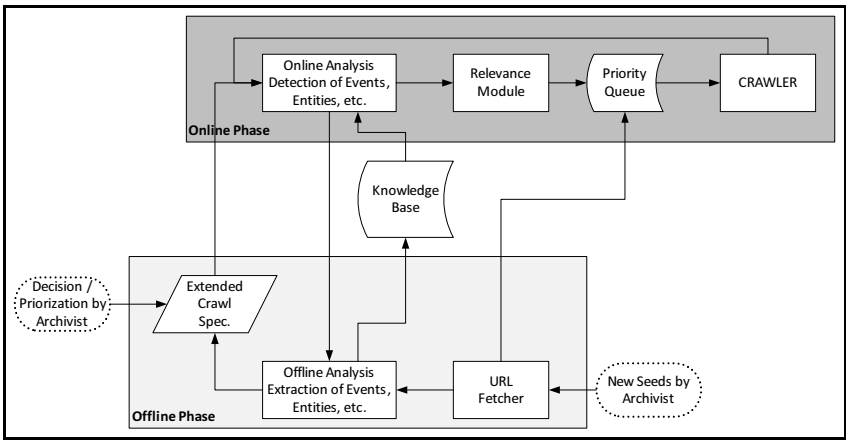


Fig. 1. Architecture for the Content Selection

The whole process is divided into an online and offline phase. The online phase focuses on the crawl task itself and the guiding of the crawler, while the

offline phase is used to analyze the crawl results and the crawl specification to setup a knowledge base for the online decision making.

Offline Phase To bootstrap a new crawl campaign, the archivist specifies a crawl by giving an initial seed list complemented with some information about events, entities and topics. e.g. [event: "Rock am Ring"], [Band: "Coldplay"] , [Location: "Nürburgring"]. The idea behind the following process is that the archivist is not able to give a full crawl specification as they cannot be fully aware of how the events, topics, etc. they are interested in are represented on the web. Therefore the crawler needs to help the archivist to improve the specification.

The initial seed list is used by the *URL Fetcher* to initiate a reference crawl. This reference crawl will be analyzed by the *offline analysis component* to extract ETOEs, which are used to derive an extended crawl specification. In this step the archivists need to assess the relevance of the extracted information to the envisioned crawl. They have the possibility to weight the information and also to explicitly exclude some of it from the crawl. The resulting *extended crawl specification* is handed over to the online phase.

In addition to the extended crawl specification, a *knowledge base* will be built, in order to provide additional information such as more detailed descriptions of events or entities, different lexical forms or other disambiguation information. The offline phase will be called regularly from the online phase to further improve the crawl specification and the knowledge base.

Online Phase The online analysis component receives newly crawled pages from the crawler and the extended crawl specification from the offline phase. Due to the necessary high crawl frequency, the processing time and decision making for a single page should take no longer then 2-3 secs. Therefore complex analysis like extracting new ETOEs is not possible. Instead, the analysis component will rely on the information in the knowledge base to detect the degree of relevance of a page to the crawl specification, to rank the extracted links and to update the priority queue of the crawler accordingly. The crawler processes the priority queue and hands over new pages to the online analysis.

4 Event Extraction

The event extraction method we adopt involves the recognition of entities and the relations between them in order to find domain-specific events and situations. As discussed in Section 2, in a (semi-)closed domain, this approach is preferable to an open IE-based approach which holds no preconceptions about the kinds of entities and relations possible. Building on the work of [MYKK05], we combine a number of different techniques, using two parallel strategies for event detection. The **top-down approach**, similar to a template-based IE approach as used in the Message Understanding Conferences [CHL93], consists of identifying a number of important events, based on analysis of the user needs and manual

inspection of the corpora. Here, the slots are known in advance and the values are entities extracted from the text. In our Rock am Ring use case, the following example depicts a band perfmance event:

```
Band:Coldplay  Relation: performed  Date: 3 June 2011
```

The technique consists of pre-defining a set of templates for the various relations, and then using a rule-based approach based on GATE [CMBT02] to identify the relevant slot values. First, we perform linguistic pre-processing (tokenisation, sentence splitting, POS tagging, morphological analysis, and verb and noun phrase chunking), followed by entity extraction, which includes both named entities and terms: for this we make use of slightly modified versions of ANNIE [CMBT02] and TermRaider[9] respectively. The third stage involves a semantic approach to finding the verbal expressions which represent the relations. We automatically create sets of verbs representing each relation, using information from WordNet and VerbNet to group verbs into semantic categories: for example, the relation "perform" might be represented by any morphosyntactic variant of the verbs "perform", "play", "sing", "appear" etc. We then develop hand-crafted rules to match sentences containing the relevant entities and verbs: for example, a rule to match the "performance" event described above should contain an entity representing a band name as the subject of a "perform" verb, and optionally a date and/or time within the sentence.

This kind of rule-based approach tends to be very accurate, achieving relatively high levels of precision (depending on how specific the rules are), but can suffer from low recall. On the other hand, a **bottom-up** technique involving open-domain IE can find previously unknown events and does not limit us to a fixed set of relations. This can be vital for discovering new information. By combining the high precision of the top-down method with the high recall of the bottom-up method, we can get the best of both worlds if done correctly.

The bottom-up approach we adopt is rather different from the machine learning approach adopted by e.g. [BE08], in that we still specify hand-coded rules. However, these rules are flexible and under-specified, making use of linguistic structure and semantic relations from WordNet [ME90] rather than prespecifying exact relations. We use the Noun Phrase and Verb Phrase chunker from GATE to identify certain linguistic patterns contextualising verb phrases, and then cluster these verbs into semantically related categories to find new relations. The participants in the relations can also be semantically clustered around similar relation types, such that an iterative development cycle can be produced. We also combine rules for ontology learning developed in SPRAT [MFP09b] which can be used to find patterns denoting relations between entities, such as hyponyms and properties. Preliminary experiments with news texts in English have found relations such as the following:

```
Mr Woerfel        represented    Daimler Benz-Aerospace
Gen Musharraf     reshuffled     two pro-Taliban generals
```

[9] http://gate.ac.uk/projects/neon/termraider.html

```
Gen Musharraf     appointed      Lt Gen Mohammed Yousuf
Mr Daoudi         was arrested   in Leicester
```

We do not only restrict ourselves to verbal relations, but also look for nominalisations. For example, "the arrest of Mr Daoudi in Leicester" is semantically equivalent to "Mr Daoudi was arrested in Leicester".

The work on event detection is still very much in progress, and it is clear that there are many difficult issues to solve. We do not use full parsing because it is very slow and because it does not work so well on social media where English is often not written correctly in full sentences. Related work on opinion mining from tweets [MF11] has proved that shallow linguistic techniques are, however, promising for extracting knowledge from this kind of noisy data, using backoff strategies and fuzzy matching where necessary.

5 Challenges

For the long-term availability and usage of Web content, it is important to preserve not only the content itself but also its context and interactions from relevant Web destinations. These include those that the content providers own (the main portal, channel portals or programme portals), those that they partner with (e.g. joint broadcaster portals), social media services or platforms, and both professional and user blogs/websites. This type of content is varied and comprises general content, commenting, rating, ranking and forwarding, while containing both structured data and unstructured free text.

To this end, it is a challenge to manage and correlate content from these information sources, differing in quality, form (e.g. both audiovisual and textual material) and structure. In order to achieve a focused crawl, it is necessary to identify semantically related objects, e.g. ones which discuss the same events or entities. However, the preservation and identification of correlations within such a diverse variety of Web sources poses a number of key challenges:

1. extraction of events and entities from heterogeneous and unstructured content;
2. detection of events and entities in heterogeneous and unstructured content;
3. targeted Web crawling.

Entity and event extraction from unstructured and heterogeneous Web data is one of the key challenges. This involves the use of natural language processing (NLP) techniques to extract events and entities from unstructured and heterogeneous text (as described in Section 4, and video analysis techniques to deal with audiovisual material. Although extraction is performed in the offline phase (see Fig. 1), there are still time requirements. Because the newly extracted entities and events are used in the online phase to focus the crawl, the extraction must be reasonably fast. To keep the crawl from becoming too diffuse, the results of the extraction must also be highly accurate, which provides an additional challenge.

In contrast to the extraction, the *detection of events and entities* needs to exploit the data captured in the knowledge base in order to automatically detect events and entities. Both NLP and video processing techniques need to be exploited here too, but with much less time for analysis: this means that the processing will be more shallow. Because the detection occurs in the online phase (see Fig. 1) and is in close interaction with the crawler, a key challenge is to perform the detection in a very short time frame and with limited time for deep, linguistic analysis.

Finally, the results of both processing phases in Section 3.2 are used for *targeted Web crawling*. This allows the crawling strategy to be gradually refined, based on the outcomes of the previous crawling, extraction and detection activities. It is a challenge to make appropriate use of these outcomes to create focused archives.

6 Conclusions

In this paper we have presented the approach we follow in the ARCOMEM project to build Web archives as community memories that revolve around events and the entities related to them. The need to make decisions during the crawl process with only a limited amount of information raises a number of issues. The division of online and offline processing allows us to separate the initial complex extraction of events and entities from the faster but shallower detection of them at crawl time. Furthermore, it allows learning more about the particular events and topics the archivist is interested in. However, the typically limited set of reference pages and the limited time to detect events during crawling are open issues to be addressed in the future. Moreover, the whole approach needs to be evaluated in real world scenarios: namely, crawling pages related to the election and to the upcoming Olympic Games.

References

[AL98] Allan Arvidson and Frans Lettenström. The Kulturarw Project - The Swedish Royal Web Archive. *Electronic library*, 16(2), 1998.

[ASS09] C. Saathoff A. Scherp, T. Franz and S. Staab. F-a model of events based on the foundational ontology dolce+dns ultralight. In *International Conference on Knowledge Capturing (K-CAP)*, 2009.

[BE08] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08*, 2008.

[BS09] C. Buckley and G. Salton. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 2009.

[BYCMR05] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a country: better strategies than breadth-first for web page ordering. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, pages 864–872, New York, 2005. ACM.

[CBD99] S. Chakrabarti, M. Van Den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Computer Networks*, pages 1623–1640, 1999.

[CGMP98] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 161–172, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[CHL93] N. Chinchor, L. Hirschman, and D.D. Lewis. Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3):409–449, 1993.

[CLS05] P. Cimiano, G. Ladwig, and S.Staab. Gimme' The Context: Context-driven automatic semantic annotation with C-PANKOW. In *Proceedings of the 14th World Wide Web Conference*, 2005.

[CMBT02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[Mas06] Julien Masanès. *Web archiving*. Springer, 2006.

[ME90] G. A. Miller (Ed.). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.

[MF11] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In *Proc. of MSM 2011: Making Sense of Microposts Workshop at 8th Extended Semantic Web Conference*, Heraklion, Greece, May 2011.

[MFP09a] D. Maynard, A. Funk, and W. Peters. NLP-based support for ontology lifecycle development. In *CK 2009 – ISWC Workshop on Workshop on Collaborative Construction, Management and Linking of Structured Knowledge*, Washington, USA, October 2009.

[MFP09b] D. Maynard, A. Funk, and W. Peters. SPRAT: a tool for automatic semantic pattern-based ontology population. In *Proc. of Int. Conf. for Digital Libraries and the Semantic Web*, Trento, Italy, September 2009.

[MKSR04] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, 2004.

[MLP08] D. Maynard, Y. Li, and W. Peters. NLP Techniques for Term Extraction and Ontology Population. In P. Buitelaar and P. Cimiano, editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. IOS Press, 2008.

[MPB08] A. Moschitti, D. Pighin, and R. Basili. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224, 2008.

[MPS04] F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.*, 4:378–419, Nov. 2004.

[MYKK05] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Aspects of the Semantic Web"*, Heraklion, Crete, 2005.

[PP06] P. Pantel and M. Pennacchioni. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 113–120, Sydney, Australia, 2006.

[STH09] R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *4th Asian Semantic Web Conference, ASWC-2009*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2009.

Hacking History: Automatic Historical Event Extraction for Enriching Cultural Heritage Multimedia Collections*

Roxane Segers[1] , Marieke van Erp[1], Lourens van der Meij[1], Lora Aroyo[1], Guus Schreiber[1], Bob Wielinga[1], Jacco van Ossenbruggen[12], Johan Oomen[3] and Geertje Jacobs[4]

[1] VU University Amsterdam
[2] Centre for Mathematics and Computer Sciences (CWI)
[3] Netherlands Institute for Sound and Vision
[4] Rijksmuseum Amsterdam

Abstract. Within cultural heritage collections, objects are often grounded in a particular historical setting. This setting can currently not be made explicit, as structured descriptions of events are either missing or not marked up explicitly. This poster reports a study on automatic extraction of an historical event thesaurus from unstructured texts. We also present a demo in which relations between events and museum objects are visualised to accommodate event- and object-driven search and browsing of two cultural heritage collections.

1 Introduction

Events have recently gained attention in the knowledge representation community as valuable constructs [4, 7, 8] that can help tie together relevant but yet unrelated elements of information. In the cultural heritage domain, knowledge about historical events is often concealed in textual descriptions that can only be accessed via keyword search. As such, the available knowledge can not be reused across collections as it is not part of the shared metadata and controlled vocabularies.

In this study, we investigate how historical events in unstructured text collections can be captured and modeled to create an event thesaurus for enriching metadata in cultural heritage collections. We adopt the SEM event model [8] to distinguish event types, actors, locations, and dates. We experiment with natural language processing (NLP) techniques to extract event names and their associated actors, dates and locations. Additionally, we show how this resulting preliminary event thesaurus is employed in a new platform for event- and object driven search and browsing of the collections of the Rijksmuseum Amsterdam (RMA) and the Netherlands Institute for Sound and Vision (S&V).

* This work was previously presented as a poster at The Sixth International Conference on Knowledge Capture (K-CAP 2011)

Fig. 1. Screenshot of object page in the Agora Event Browsing Demonstrator

2 Event Extraction from Text

As no annotated historical document collections exist in Dutch, our approach is focused on extracting named events with minimal manual effort. For this study we selected 3,724 historical Wikipedia articles as a test set. The event extraction process consists of three steps: in the **first step**, we recognize *actor names* and *locations* using the Stanford Named Entity Recognition system [2] adapted for Dutch historical texts. Dates were recognized via regular expressions. This step resulted in 18,623 candidates for actors (F-measure of 0.77), 7,023 locations (F-measure of 0.66) and 7,981 dates. In the **second step**, we use a pattern-based method for recognizing *event names* such as *French Revolution*. We harvest patterns from the Web (e.g., *destroyed during the*, *before the*) using the Yahoo! search API [5] and a seed set of one hundred historical events. Patterns are ranked by frequency of co-occurrence with two or more seed events [6]. To retrieve event candidates, we applied the patterns to the Wikipedia corpus. The event candidates are then filtered, based on a threshold on the pattern score, resulting in a set of 2,444 unique events. The precision score of this set is 56.3%.

In the **third step**, we associate events with actors, locations and dates. We experiment with both redundancy and co-occurrence of data on the Web, inspired by the work of Geleijnse et al. [3] and Cilibrasi &Vitanyi[1]. Each combination of an event name and actor/location/date is sent to Yahoo! and for each pair a score is computed. We discovered 392 event names that were paired with an actor, a location and a date. Through manual evaluation we conclude the following: 71.9% (323) are correct event names, 45.6% (179) are correct actors, 41.1% (161) are correct locations and 51.5% (202) are correct dates.

[5] http://developer.yahoo.com/search

3 Enrichment by Events

The extracted events are linked to the RMA and S&V collections. In total 35 unique events provide direct relations from 435 S&V objects to 675 RMA objects. An additional 34 unique events provide links from 391 S&V objects to 362 RMA objects, but this link exists indirectly through the event instance (e.g., S&V object - Actor - RMA object). We hypothesize that these links are potentially useful for navigating cultural heritage collections.

4 The Agora demonstrator

The automatically generated event thesaurus is applied in a new historical event browser called Agora[6] which provides an integrated access route to museum objects and audio-visual material from RMA and S&V respectively. It is a first step towards a platform to investigate the added value of historical events and narratives for the exploration of integrated collections. For each event and object there is an automatically generated page that shows (1) all associated objects, e.g., museum and audio-visual objects; (2) all associated events and the type of their relationship, e.g., previous-in-time event, sub-event; (3a) the event descriptive metadata, e.g., actors, place, period; or (3b) object descriptive metadata organized in three groups, e.g., biographical, material and semiotic dimensions – see figure 1 for a screenshot –and finally (4) the navigation path. The current version of the event thesaurus will be extended further to accommodate searching for relations between events such as temporal inclusion, causality and meronymy.

5 Discussion

In this paper, we presented a modular pipeline for capturing knowledge about historical events from Dutch texts. Compared with previous approaches (i.e., [5]), it relies on a minimum of manual annotation and can be repurposed for other languages. To the best of our knowledge, this is the first work to extract events from unstructured Dutch text. Although our results are promising, more sophisticated techniques are necessary to obtain more fine-grained extractions and define measures for the historic relevance of the extracted events. Additionally, we also aim to find and represent relations between events such as causality, meronymy and correlation.

6 Acknowledgements

This research was funded by the CAMeRA Institute of the VU University Amsterdam and by the CATCH programme, NWO grant 640.004.801.

[6] `http://agora.cs.vu.nl/demo`

References

1. R. Cilibrasi and P. Vitanyi. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, 19(3):370–383, 2007.
2. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 2005.
3. G. Geleijnse, J. Korst, and V. de Boer. Instance classification using co-occurrences on the web. In *Proceedings of the ISWC 2006 workshop on Web Content Mining (WebConMine)*, Athens, GA, USA, November 2006.
4. N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Automatic event-based indexing of multimedia content using a joint content-event model. In *ACM Events in MultiMedia Workshop (EiMM10)*, Oct 2010.
5. N. Ide and D. Woolner. Exploiting semantic web technologies for intelligent access to historical documents. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, pages 2177–2180, Lisbon, Portugal, 2004.
6. E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI '99*, pages 474–479, 1999.
7. R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In *4th Annual Asian Semantic Web Conference (ASWC'09)*, 2009.
8. W. R. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Abstracting and reasoning over ship trajectories and web data with the Simple Event Model (SEM). *Multimedia Tools and Applications*, 2011.

Events Retrieval Using Enhanced Semantic Web Knowledge

Pierre-Yves Vandenbussche[1,2] and Charles Teissèdre[1,3]

[1] Mondeca, 3, cité Nollez, 75018, Paris, France
[2] INSERM UMRS 872 ÉQ.20, Ingénierie des Connaissances en Santé, Paris, France
[3] MoDyCo – UMR 7114 CNRS / Université Paris Ouest Nanterre La Défense, 200 av. De la République, 92001 Nanterre, France
{firstname.lastname}@mondeca.com

Abstract. In this article, we present an experimental end user application to query DeRiVE 2011 challenge dataset in an innovative and intuitive manner. After enriching the dataset with external sources of information, it is indexed in a way that enables users to submit queries combining keywords, location and temporal anchor, in a single search field. The goal is to ease event retrieval providing a simple user interface to query and visualize events over time.

Keywords: Events Retrieval, Semantic Web, Data Mashup

1 Introduction

While geolocation services have enjoyed strong progress, few initiatives take into consideration chrono-localization and temporal query processing for Information Retrieval over the Web [1]. As Linked Open Data grows, things are changing, since more and more temporally anchored data is available. However, processing temporal data remains a challenge from (i) a modeling point of view, (ii) for data acquisition, (iii) as well as in terms of querying and navigating through it.

In this article, we address the last issue of querying and navigating through temporal data. We describe a system using the RDF data provided along with the DeRiVE 2011 challenge[1]. The dataset describes entertainment events related to music, such as advertisements for concerts or festivals. It also provides some information about agents involved in these events and about their location. The main objective of the system we present here is to hide data complexity and make it simple to query, providing a single search field as a first step in events retrieval. The goal is to make DeRiVE dataset temporally browsable. The considered use case consists in finding events occurring at a given period of time at a specific location.

After a brief overview of how temporal information is handled in the context of Information Retrieval over the Web, we will describe the way we processed the dataset to enrich it with external sources of information and to index it. We will then describe the final application to query and browse the dataset.

[1] Dataset is available at: http://semanticweb.cs.vu.nl/derive2011/Challenge.html

2 Temporal Information Retrieval over the Web

Retrieving temporal information over the Web of Content (*i.e.* HTML-based Web) and in the Web of Data (aka the Semantic Web) are two different issues, though they may converge on some points.

Temporal Search within the Web of Content. Major search engines currently offer few temporal search services. One such service is Google timeline feature[2], which offers a way to visualize keywords frequency at different periods of time and to browse sentences where these keywords are associated with a date. However, temporal expressions are reduced to point in time with no duration extent, hence there is an important loss of information. Processing temporal information expressed in Web documents is a challenge from at least three different points of view: (i) modeling temporal references (models should be able to represent dates and intervals, but may also need to cope with approximate information (*e.g.* "*by the end of the 13th century*"), iterative occurrences (*e.g.* "*every day from 10am to 8pm*"), as well as deictics (*e.g.* "*yesterday*", "*two months ago*") and anaphorics (*e.g.* "*the day before*")); (ii) document annotation (it requires processing huge amount of documents with NLP techniques that necessarily have to deal with imperfect precision and recall rate) and (iii) relevancy ordering of the results from the temporal perspective (how to rank documents by relevance from the temporal perspective?).

Temporal Search within the Web of Data. While the modeling issue remains a difficulty, the acquisition process in this context is quite different, since the data to process is structured. Data acquisition however can be an issue as well. As for the querying process, the main querying language, SPARQL, allows filtering results in a timespan (*i.e.* intervals of well defined dates). This approach explains why generally only well defined temporal properties are effectively employed in LOD[3].

The three Web sites that provided data for the challenge relies on this process: Upcoming Yahoo!, Last.fm and Eventful all propose similar approach to event retrieval. The main search scenario, with little variation depending on the Web site, follows this path: user has to provide a location, then a type of event (concert/festival), then eventually a musical genre, a date filter, etc. Such rich faceted search scenario is not made possible, though, with the DeRiVE challenge dataset, since no information on the type of event is provided.

3 Processing DeRiVE 2011 Dataset

The application we present here is an experimental retrieval engine with the goal to query and browse events temporally in the simplest way possible. It can be used both

[2] URL for the query "revolution": http://bit.ly/relfGV

[3] Despite Time Ontology [2] capability to describe complex time knowledge representation, it is generally not used in all its' complexity.

in the context of the Web of Content [3] and the Web of Data, as it relies on indexing process and NLP resources for temporal references extraction which can analyze either a query or Web documents. For the DeRiVE challenge, in order to get enough information to enable users to submit queries combining keywords, location and temporal information, we first had to enrich the dataset. The DeRiVE 2011 dataset is composed of 107.874 events and related knowledge. Knowledge is originating from Upcoming Yahoo! (12.15%), LastFm (53.04%) and Eventful (34.81%). It has been transformed by EventMedia [4]. The dataset is made of more than 1.800.000 statements. Temporal information consists in either single dates or intervals of dates.

Event geo-location augmentation. 98.794 events (91.58%) have latitude and longitude information. The first knowledge augmentation process concerns events' geolocation. It tries to fetch city, country and address information from coordinates, using Google and Yahoo! reverse geocoding API. In our application this geolocation information is used during query processing to cope with countries or cities. It is also used to propose a map visualization using Google maps API.

Event Image augmentation. Images provide a simple way to ensure a pleasant way to experience event browsing. To associate images to events, we set up a strategy based on images information in the Semantic Web (via SPARQL queries on EventMedia and DBpedia) and on the Web (via Flickr API). As a result, at least one image was associated to 95.01% of events. SPARQL query example on EventMedia endpoint[4] using event URI:

```
SELECT distinct ?imageURI ?image
WHERE{
      ?imageURI                    <http://linkedevents.org/ontology/illustrate>
      <http://data.linkedevents.org/event/dba9e034-fea0-4d01-ba4c-
      fb0515b89051>.
      {?imageURI <http://www.w3.org/ns/ma-ontlocator> ?image. }
      UNION{?imageURI <http://www.w3.org/ns/ma-ont#locator> ?image. }
}
```

Agent Information augmentation. Information about agents involved in an event is valuable for our application users. By enriching the dataset with Wikipedia links that point toward articles concerning these agents, users can further their search. We collected these links thanks to SPARQL queries on DBpedia endpoint. We have been able to find Wikipedia links for 25.22% of the agents. SPARQL query example on DBpedia endpoint[5] using agent label:

```
SELECT distinct ?wikiLink
WHERE {
      {?s <rdfs:label> "Bob Dylan"@en.}
      UNION{?s <http://xmlns.com/foaf/0.1/name>
      "Bob Dylan"@en.}
      {?s a <http://dbpedia.org/ontology/Person>.}
      UNION{?s a <http://dbpedia.org/ontology/Band>.}
      ?s <http://xmlns.com/foaf/0.1/page> ?wikiLink.
}
```

[4] URL: http://semantics.eurecom.fr/sparql
[5] URL: http://dbpedia.org/sparql

4 An Experimental Temporal Search Engine to Retrieve Events

The system we have implemented is both a search engine and a tool to visualize and browse events[6]. Temporal query relies on the search engine developed by [3]. The search engine is able to process queries with approximate temporal conditions like "*around May 2007"*, even if this temporal expression does not exist in DeRiVE data. From the temporal perspective, event retrieval is based on an algorithm that calculates similarity scores between the temporal reference of the query and those that are associated to events. Based on Lucene and several modules to compute the dataset (see fig 1), the system can handle queries that may combine keywords, location and temporal information, such as "*rock in London in August 2008*" or "*Bonn by the end of 2007*". Temporal information, location information and event or agent description are indexed as different fields once the dataset is fully preprocessed.

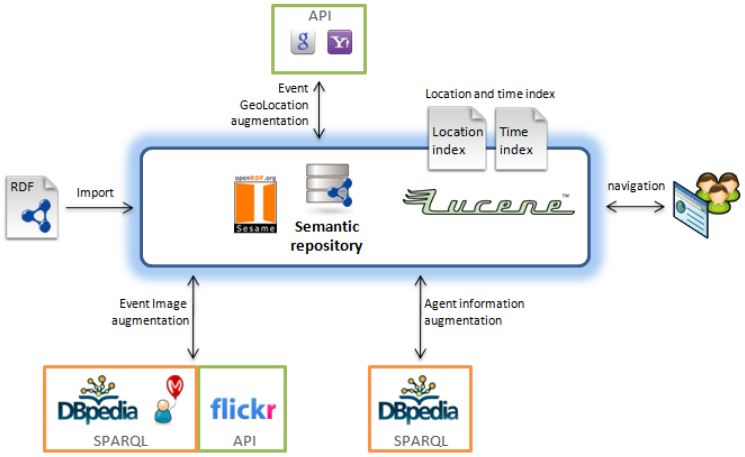


Fig. 1. System's architecture.

Queries are analyzed in such way that keywords, dates and location information are separated. Temporal data recognition in queries is performed thanks to an NLP module described in [5]. The location extraction is performed thanks to a dictionary built during the indexing process: the dictionary contains cities and countries entities collected during the event geolocation enrichment process. Any other information that may appear in a query is considered as simple keywords, on which no semantic analysis is performed.

The events returned by the system are presented on a SIMILE timeline[7] (see fig 2). The timeline on which results are displayed is fully browsable, which means that users can move over time: the system generates new queries on the fly as users move forward or backward in time.

[6] The system can be tested at the following address:
http://labs.mondeca.com/ChallengingTime/?locale=en&demo=eventMedia
[7] URL: http://www.simile-widgets.org/timeline/

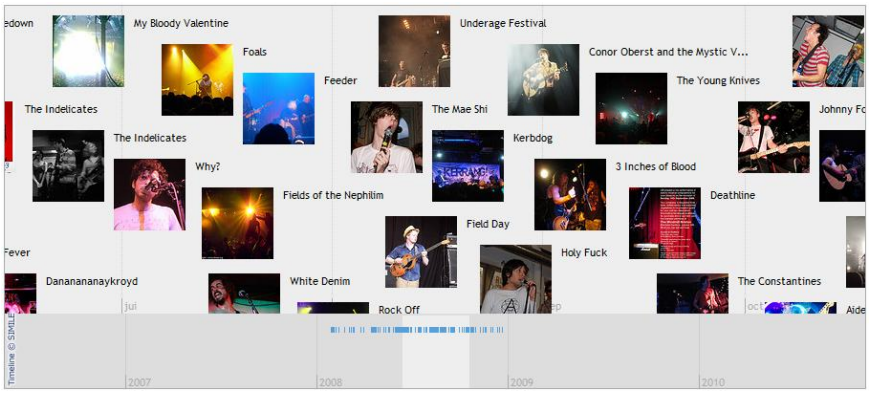


Fig. 2. Screenshot of the UI for the query "*rock in London in August 2008*".

5 Conclusion and Perspectives

The experimental application presented could be used as first step in events retrieval. Since the approach is generic and not tightly bound to DeRiVE dataset, it can be used in any other use case scenario where data is temporally anchored. If the dataset had contained information about musical genres, it could have been interesting to introduce faceted search with SolR tool, so that users could refine the results and eventually disambiguate query. Another interesting feature for possible improvement would be to synchronize a map for geolocation with the timeline, so as to present the results both in their temporal and geographic context.

Acknowledgments. This project is partially granted by Datalift ANR project (ANR-10-CORD-009) and Chronolines ANR project (ANR-10-CORD-010).

References

1. Alonso, O.; Gertz, M. & Baeza-Yates, R.: On the Value of Temporal Information in Information Retrieval. Proc. of ACM SIGIR Forum 41, no. 2 (December), 35-41 (2007)
2. Hobbs, JR & Pan, F.: An ontology of time for the semantic web.Proc. of ACM Transactions on Asian Language 3, no. 1 (March), 66-85 (2004)
3. Teissèdre, C; Battistelli, D. & Minel, J.-L.: Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires. Proc. of TALN 2011, Montpellier (2011)
4. Troncy, R.; Malocha, B. & Fialho, A. Linking events with media Proceedings of the 6th International Conference on Semantic Systems, 1-4 (2010)
5. Teissèdre, C.; Battistelli, D. & Minel, J.-L.: Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts. Proc. of LREC 10, Malta (2010)

Domain-aware Matching of Events to DBpedia

Kristian Slabbekoorn, Laura Hollink, and Geert-Jan Houben

Web Information Systems Group
Delft University of Technology, The Netherlands
k.slabbekoorn@student.tudelft.nl,{l.hollink, g.j.p.m.houben}@tudelft.nl

Abstract. In this paper, we present our work on the enrichment of the EventMedia dataset as provided by the DeRiVE data challenge with links to DBpedia. Our main contribution is an exploration into the use of domain knowledge in the matching process. As a starting point we take DBpedia Spotlight, an off-the-shelf tool for matching textual resources to DBpedia. We present a bootstrap method to automatically derive the needed domain knowledge from an initial set of high confidence matches, and compare this to a baseline method without any domain knowledge, and an 'oracle' method with perfect domain knowledge.

1 Introduction

In this paper, we present our work on the enrichment of the EventMedia dataset as provided by the DeRiVE data challenge with links to DBpedia. Tools and algorithms have emerged that automate the task of matching two ontologies or datasets. Most of these systems use string similarity measures and/or structural measures to determine the similarity between a pair of resources. However, little is known about how one can include the domain of the data into the matching process. In our case, we know that the EventMedia dataset is about events, performing artists and venues. The main contribution of this paper is an exploration into the use of this knowledge of the domain to produce better or more matches. In addition, the resulting matches are made publicly available for download.

As a starting point we take DBpedia Spotlight, an off-the-shelf tool for matching textual resources to DBpedia. DBpedia Spotlight has been shown to be able to compete with established annotation systems while remaining largely configurable [1]. The configurability allows us to include various forms of domain knowledge, and test the effect on the resulting matches. It also means, however, that we have to choose values for a relatively large number of parameters that potentially influence the results. To minimize this effect, we set the parameters systematically and transparently in section 2.1.

We present a bootstrap method to automatically derive the needed domain knowledge from an initial set of high confidence matches, and compare this to a baseline method without any domain knowledge, and an 'oracle' method with perfect domain knowledge. To explore the generalizability of the derived domain knowledge, we perform an evaluation in which we derive the domain knowledge from one dataset and use it to find matches in another dataset.

Several bootstrapping methods to derive links between Linking Open Data (LOD) datasets have been proposed previously. [2] matches concepts by finding candidates in DBpedia, then comparing classifications of their own concepts to the classes and categories of the DBpedia candidate concepts. In our case, we do not assume to have a classification of the source data available. BLOOMS+ [3] uses the Wikipedia category hierarchy to bootstrap the process of finding schema-level links between LOD datasets. We exploit the Wikipedia category hierarchy in a similar fashion; not to find matches directly but to find categories (and classes) that effectively describe our domain of interest.

1.1 Dataset and Reference Alignment

All experiments are performed on the EventMedia dataset provided as part of the DeRiVE data challenge, containing RDF statements about events, artists and venues from the websites Last.fm, Eventful.com and Upcoming.yahoo.com. We have chosen to focus on matching artists as they are more likely to have pages dedicated to them on Wikipedia than venues and events do - pages can be found for roughly 35% of the artists contained in the Last.fm dataset, and 45% of the artists contained in the Eventful dataset. Upcoming does not contain explicit mentions of artists. We evaluate our approach by comparison to a manually composed reference alignment of 1500 randomly picked artists (1000 from Last.fm and 500 from Eventful.com) to DBpedia resources.

1.2 DBpedia Spotlight

Throughout this work we have used DBpedia Spotlight [1], a powerful tool for automatically annotating natural language texts with links to DBpedia resources. It does so by first finding surface forms in the text that could be mentions of DBpedia resources (the 'spotting' function), then disambiguating to link to the right DBpedia resources based on context similarity measures (the 'disambiguation' function). Its results can be directed towards high precision or high recall by setting two parameters: a 'support' threshold for minimum popularity of the Wikipedia page (i.e. the number of inlinks from other Wikipedia pages) and a 'confidence' threshold for minimum similarity between source text and context associated with DBpedia surface forms. The latter has been normalized to a range of 0..1. In addition, Spotlight's 'black- and whitelists' allow one to filter the results to exclude/include only members of certain classes and categories that correspond to the domain of the source text.

2 Approach

In this section we present our approach to domain-aware matching. We compare our results to a baseline approach, where we run Spotlight without any domain filters, and an 'oracle' approach, where the optimal classes and categories are chosen as a domain filter, based on the best matching results in hindsight.

We do our matching in two passes. In the first pass we attempt to match the full label. We do this by marking the full `rdfs:label` of an artist for disambiguation and appending the `dc:description` value, if available, as context for Spotlight's 'disambiguate' function. We attempt to increase the number of links by running a second pass with Spotlight's 'spotting' function on artists not matched initially to search for surface forms 'hidden' inside the label (for instance, some labels include more than one artist).

Bootstrap approach: deriving domain filters. We bootstrap the selection of domain filters by first running DBpedia Spotlight without any knowledge of the domain, with parameters set towards high precision, to obtain an initial set of links from our data to DBpedia resources. From these resources we gather the associated DBpedia classes, YAGO classes and Wikipedia categories and use these as a domain knowledge filter to find further matches.

To get a set of classes and categories that concisely describes our domain, we first gather all DBpedia and YAGO classes of the matched DBpedia resources, including all super-classes up to the root of their respective hierarchies. For categories, we gather only up to 4 ancestors of each, as due to the size and messy structure of the Wikipedia category hierarchy the set will quickly become too large and broad to be useful.

Second, we select the appropriate classes and categories from this long list as follows. We count the number of occurrences of each class or category. We filter out all classes that occur less than r% of the total number of classes found. General (super-)classes will occur more frequently then specific (sub-)classes. Therefore, the higher the value of r, the more general our list of classes will be. For categories, this effect is less strong since we do not gather super-categories up to the root. Therefore we simply select the top t categories that occur the most. To avoid too much redundancy in the list of classes and categories, i.e. to avoid including a super-class plus all its sub-classes, we filter out all super-classes where the sum of the numbers of occurrence of their sub-classes is more than 90% of the number of occurrence of the super-class. The same procedure is followed for categories. The resulting list of DBpedia classes, YAGO classes and categories represents our domain filter.

2.1 Spotlight parameter optimization

In this paper we assume an application that values precision and recall equally, and therefore we optimize the parameters for high F-measure (F). To allow a fair comparison between the three approaches, we set the parameters of each approach independently to the values that are optimal for that approach.

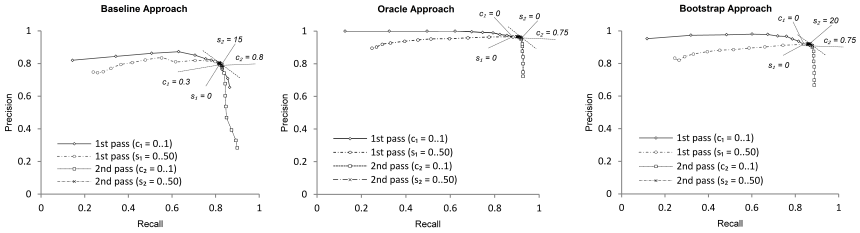
For each approach, we need to optimize 'confidence' c and 'support' s for pass 1 and pass 2, resulting in the four parameters c_1, c_2, s_1 and s_2. We determine the best setting of all parameters by evaluating the resulting matches against our reference alignment. First, we keep s_1 fixed at 0 and vary c_1 between 0 and 1 in steps of 0.1. Next, we take values around and including the value of c_1 that provided the highest F and vary s_1 between 0 and 50 in steps of 5. We

take this relatively low range of inlinks due to the nature of our dataset, which largely consists of obscure entities that are likely not often linked to. We settle on whichever combination of c_1 and s_1 gives the best F. To set the parameters c_2 and s_2 of the second pass, analogous experiments are performed, this time varying c_2 and s_2 respectively.

Baseline approach Figure 1a shows that the highest F-measure (F_{max}) is obtained when parameters are set as follows: $c_1 = 0.3$, $s_1 = 0$, $c_2 = 0.8$ and $s_2 = 15$.

Oracle approach Optimal parameters for this approach are $c_1 = 0.0$, $s_1 = 0$, $c_2 = 0.75$ and $s_2 = 0$. See figure 1b.

Bootstrap approach Our aim is to evaluate our bootstrap approach with varying amounts of classes and categories to specify the domain. The optimal parameters for each variant could be different and hence they need to be determined independently. For space reasons, figure 1c only shows the parameters for the approach that gave the best results: $c_1 = 0.0$, $s_1 = 0$, $c_2 = 0.75$ and $s_2 = 20$.



(a) The baseline approach. $F_{max} = 0.811$

(b) The oracle approach. $F_{max} = 0.939$

(c) The bootstrap approach ($r = 1, t = 0$). $F_{max} = 0.895$

Fig. 1: The effect of different parameter settings on precision and recall. The optimal values for each parameter are denoted in the graphs. The highest F-measure scores F_{max} correspond to the values at the top-right corners of the graphs.

3 Experimental results

Table 1 show the results of our experiments on the Last.fm artist dataset. We see that our domain-aware bootstrap method gives results that are better than the baseline, and close to the 'oracle'. The results are slightly better when we do not consider categories at all ($t = 0$). A reason for this is because it is difficult to detect and filter out overly general categories (for example, `Category:People` is always part of our result). There is also an expected inherent trade-off to be seen between precision and recall when we choose r as either 1 or 2. We additionally run the baseline, oracle and best-performing bootstrap approach on the Eventful artist dataset and evaluate based on a ground truth of 500 artist labels derived in a similar way to the Last.fm ground truth. These results again show the value

Table 1: Results for each approach sorted by maximum F-measure.

Approach	*Precision*	*Recall*	F_{max}
Oracle	0.966	0.914	0.939
Bootstrap: $r = 1, t = 0$	0.921	0.870	0.895
Bootstrap: $r = 2, t = 0$	0.869	0.916	0.892
Bootstrap: $r = 1, t = 10$	0.877	0.905	0.891
Bootstrap: $r = 2, t = 10$	0.846	0.902	0.873
Bootstrap: $r = 1, t = 5$	0.835	0.902	0.867
Bootstrap: $r = 2, t = 5$	0.828	0.902	0.863
Baseline	0.799	0.824	0.811

of domain knowledge, with $F_{max} = 0.726$ for the baseline, $F_{max} = 0.905$ for the oracle, and $F_{max} = 0.901$ for the derived approach.

The DBpedia links created with the oracle approach for both datasets are available for download[1]. Also included are interlinks between Last.fm and Eventful artists if they have all of their links in common. For Last.fm, we have 50120 entities in total and end up with 17116 DBpedia links. For Eventful, we have 6540 entities and 2724 links. There are 2450 interlinks made between datasets.

4 Discussion and Future Work

In this paper we presented a bootstrapping method to improve the matching of concepts within a particular domain to DBpedia resources. We found that our bootstrapping method performs better than a general domain-independent matching, and that the F-measure associated to our best derived model is consistent across two datasets. It is not yet clear to what extent our proposed method is applicable to other domains of a different nature. We are currently exploring how robust our method is against different sets of initial high confidence matches, varying the size as well as the domain.

We found that we often end up with rather generic classes/categories, such as YAGO class `LivingPeople` and category `People`, in our final selections for a domain filter. Our future work focusses on improving the class and category selection algorithm in order to filter out these general cases.

References

1. Mendes, P., Jakob, M., Garca-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In the Proc. of the 7th International Conference on Semantic Systems (I-Semantics). Graz, Austria, 7-9 September 2011. (to appear)
2. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Conections. Proc. of ESWC 2009, Heraklion, Crete.
3. Jain P., Yeh P. Z., Verma K., Vasquez R. G., Damova M., Hitzler P., Sheth A. P.: Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton. In G. Antoniou (Ed.), Proc. of ESWC 2011, Heraklion, Crete.

[1] http://wis.ewi.tudelft.nl/iswc2011/derive/

Fusion of Event Stream and Background Knowledge for Semantic-Enabled Complex Event Processing

Challenge Paper

Kia Teymourian, Malte Rohde, Ahmad Hasan, and Adrian Paschke

Freie Universität Berlin
Institute for Computer Science
Corporate Semantic Web Research Group
http://www.inf.fu-berlin.de/groups/ag-csw/
{kia, malte.rohde, ahmadhaidar, paschke}@inf.fu-berlin.de

Abstract. Usage of ontological knowledge about events and their relations to other concepts in the application domain can improve the quality of complex event processing (CEP). In this paper, we present a solution for knowledge-based event entity extraction using background knowledge bases. For our experiments we use the DeRiVE-2011 workshop dataset. We enrich the incoming event data stream with background knowledge from an external knowledge base, e.g., DB-Pedia so that our event processing engine has more knowledge about events and their relations to other concepts in application domain. [1]

Keywords: Complex Event Processing, Semantic CEP, Event Query Pre-Processing

1 Motivation

Semantic models of events can improve the quality of event processing by using event metadata in combination with ontologies and rules. The success of the Semantic Web research community in building standards and tools for semantic technologies such as formalized vocabularies/ontologies and declarative rules is opening novel research and application areas. One of these promising application areas is Semantic Complex Event Processing (SCEP), for which we previously proposed a new approach in [5, 4]. We claim that semantic models of events can improve the quality of event processing by using event data in combination with knowledge bases.

Existing methods for event processing can be categorized into two main categories, logic-based approaches and non-logic-based approaches [2]. One of the logic-based approaches is introduced in [1] which proposes a homogeneous reaction rule language for complex event processing.

In this paper, we describe a method for knowledge-based complex event processing to extract complex event entities from an event stream. Fusion of event data stream and background knowledge about events and other non-event objects in the application

[1] This work has been partially supported by the "InnoProfile-Corporate Semantic Web" project funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions.

domain can build up a complete knowledge about events and their relationships to other concepts. We use the workshop dataset to demonstrate how our method can be used for SCEP. In Section 2, we focus on use case scenarios and show which kind of complex events can be detected from the workshop data set using DBpedia[2] as a knowledge base. Section 3 describes our method for event processing which includes data fusion with the background knowledge base. Finally in Section 4, we describe our demonstration using SCEP.

2 Use Case Scenarios

The DeRiVE 2011 event dataset consists of over 100,000 events from music and entertainment websites. This dataset includes data about most popular concerts, festivals, kids events, sports events, and other social events. In the following, we describe three concrete complex situations where a person or an organizations defines a complex event query and is interested in detecting specific events from the upcoming event stream:

Scenario 1 - Specific Music Concerts Interests: Consider that Mr. Smith is interested in a special music type. For example he is especially interested in *"Canadian alternative rock music from the city Toronto"*. Mr. Smith lives in Europe and might be able to travel to different European cities when there are interesting concerts. Mr. Smith is married and has two children. For his travel plans, he has also to consider his family situation. He would like to travel only if there are some interesting kids or family events in the same city at the same time, so that they can travel together. His children like *"kids theater"*. If there are such upcoming events in combination, Mr. Smith would like to be informed in advance. This kind of application scenario can be seen as a real-time recommendation system, which may also trigger some automated reactions after the detection of a complex event.

Scenario 2 - Surveillance, Riot Prevention and Control: A security organization might be interested to know which concert types or which events are happening at the same time in the same cities, which might cause some potential conflicts between different fan parties. For example, if there are *"rock music concerts"* together with *"hip-hop music concerts"* in a small city at the same time, conflicts might arise if fans meet each other on the streets or in bars. In order to detect such potentially dangerous situations in advance, the security organization needs to be able to define its own high-level complex event queries, not only depending on the event data itself but also on background information, e.g. known hostility between fans of soccer clubs, etc.

Scenario 3 - Music Market Monitoring: A concert promoter company is eager to find out about gaps on the music concert market. It is interested to know in which European cities which types of concerts are organized and happening now. For example, when some outstanding *"60s hard-rock band"* goes onto its reunion tour, it might be the time to host a concert of a *"cheap cover band"*. Also, whenever a city is overwhelmed with concerts of a specific type of music, there may emerge a demand for other types of concerts. Thus, the concert promoter company needs to detect these event patterns early, so that they can organized concerts which are well attended by its fans.

[2] http://www.dbpedia.org, July 2011

3 Semantic Enabled Event Processing

The fusion of background knowledge with the data from an event stream can help the event processing engine to know more about incoming events and their relationships to other related concepts. We propose to use an external knowledge base which can provide background conceptual and assertional information about the events as it is shown in Figure 1. This means that events can be detected based on reasoning on their type hierarchy relationships, or temporal/spatial relationships. It can also be based on their connections to other relevant concepts from the domain, e.g., relationship of a concert event to the health situation of a band member. Some of the existing CEP systems [3] can integrate and access external static or reference data sources. But these systems do not provide any inferencing on external knowledge bases and do not consider reasoning on relationships of events to other non-event concepts.

The realization if SCEP is a challenging task, because it should provide real-time processing and high scalability. The naïve approach for SCEP might be a storage-based approach. This means to store all of the background knowledge in knowledge bases and start pulling the knowledge base, every time when a new event comes into the system, and then process the result from the external knowledge base with event data. This approach may have several problems when the throughput of the event stream is high, the size of background knowledge is high, or even when expressive reasoning should be done on the knowledge base.

Event Query Pre-Processing:
We propose to do an Event Query Pre-Processing (EQPP) before the event processing is down on the event stream. In this approach, the original complex event query can be pre-processed by use of a knowledge base and rewritten into a single *new query*. This *new query* is a query which can be syntactically processed only with the knowledge from the event stream and without an external knowledge base.

In this paper, we are addressing a simple pre-processing of event queries and illustrate the potential of such a pre-processing approach for SCEP. In our method the user query is pre-processed and rewritten into a single new query which has the same semantic meaning as the original one. The advantage of this method is that the user can define event queries in a high level abstraction view and does not need to care about some details, e.g., the user only defines *"alternative rock music band from the city Toronto"* and does not need to know all of the names of such music bands which might be a long list and might not be simple for humans to remember. One other advantage is that the SCEP system is able to provide real-time event processing as events arrive into the system because the external reasoning on knowledge base is done in advance. On the other side, one dist-advantage of this approach is that the query needs to be updated each time when the knowledge base is changed (or when a part of the KB is changed). We assume that in some of the use cases (like music concert events) the rate of background

[3] Several rule-based and storage-based event processing system are already proposed and developed, some of the commercial products are:
TIBCO BusinessEvents, `http://www.tibco.com/`, July 2011
Oracle CEP `http://www.oracle.com`, July 2011
Sybase CEP `http://www.sybase.de`, July 2011

knowledge updates is not very high as the rate of the main event stream, e.g., frequency of happening of music concerts compare to changes in a music band.

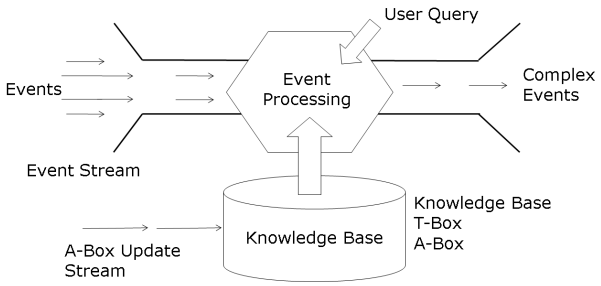


Fig. 1. High Level Architecture of Semantic-Enabled Complex Event Processing

4 Experiments and Demonstration

For our experiments, we use Prova[4] as reaction rule language formalization and as a rule-based execution which can be used as event processing engine. Prova uses reactive messaging[5], reaction groups and guards[6] for complex event processing. Multiple messages can be revived using revMult(XID, Protocol, Destination, Performative, Payload) ; XID, a conversation id of the message; Protocol, message passing protocol; Destination, an endpoint; Performative, message type; Payload, the content of message. Prova implements a new inference extension called literal guards. During the unification only if a guard condition evaluates to true, the target rule will proceed with further evaluation. We implemented the *sparql_select* built-in[7] to run SPARQL queries from Prova which can start a SPARQL query from inside Prova on an RDF file or a SPARQL endpoint. This buit-in can use results which come from the SPARQL query and use them inside Prova. It also provides the possibility to replace variables in SPARQL string which are starting with $ with variables.

In our experiments we use the Prova rule engine. We use the *sparql_select* built-in: the rule engine first sends the embedded SPARQL query to triple store, gets the results back and then waits for incoming event stream to process. It processes the sequence of

[4] Prova, ISO Prolog syntax with extensions `http://prova.ws` , July 2011

[5] Prova Reactive Messaging `http://www.prova.ws/confluence/display/RM/Reactive+messaging` , July 2011

[6] Event Processing Using Reaction Groups `http://www.prova.ws/confluence/display/EP/Event+processing+using+reaction+groups`, July 2011

[7] Source codes for Semantic Web extensions in Prova 3 can be found in `https://mandarax.svn.sourceforge.net/svnroot/mandarax/prova3/prova-compact/branches/prova3-sw/` , July 2011

events using the provided results from the knowledge base. In our experiments, we use a replicated version of part of DBpedia as an external knowledge base and the YAGO classification [3] for the classification of different music types and bands, e.g., a user can use the YAGO classification to express his music interest.

The complete pre-processing step should be updated on the knowledge base, whenever there is a change in the knowledge base, e.g., if new music bands are added to DBpedia our event query has no knowledge about them. In many use case like ours, the frequency of such updates can be considered to not be very high. Here, one useful approach is to implement the updates also in an event-based manner, if any relevant changes are done on the knowledge base a notification informs the event processing engine to update the event query.

Our experiments show clearly that the EQPP can achieve a better perofmance than the naïve storage-based approach (or pulling approach). They also show that the EQPP approach is an applicable approach for the above described use case. It shows also that the scalability of SCEP systems has five different dimensions; 1. Discharge rate of events 2. Number of rules in main memory 3. Number of triples in KB (amount of knowledge) 4. Rate of knowledge updates 5. Expressive level of reasoning on KB.

5 Conclusion and Outlook

We described our initial work on semantic event processing and semantic pre-processing of event queries, and illustrated the potential of this approach by use of a demonstration.

Our future steps are to work on semantics of event processing languages, and define which semantics can be adequate semantic for Complex Event Processing. Furthermore, we are working on an algorithm for rewriting of complex event queries to several simple queries which can be distributed on an event processing network to achieve high performance and scalability.

References

1. Adrian Paschke, Alexander Kozlenkov, and Harold Boley. A homogeneous reaction rule language for complex event processing. *CoRR*, abs/1008.0823, 2010.
2. Kay-Uwe Schmidt, Darko Anicic, and Roland Stühmer. Event-driven reactivity: A survey and requirements analysis. In *SBPM2008: 3rd international Workshop on Semantic Business Process Management in conjunction with the 5th European Semantic Web Conference (ESWC'08)*. CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073), June 2008.
3. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.
4. Kia Teymourian and Adrian Paschke. Semantic rule-based complex event processing. In *RuleML 2009: Proceedings of the International RuleML Symposium on Rule Interchange and Applications*, 2009.
5. Kia Teymourian and Adrian Paschke. Towards semantic event processing. In *DEBS '09: Proceedings of the Third ACM International Conference on Distributed Event-Based Systems*, pages 1–2, New York, NY, USA, 2009. ACM.