Extracting and Modeling Historical Events to Enhance Searching and Browsing of Digital Cultural Heritage Collections

Roxane Segers

Department of Computer Science, VU University Amsterdam
r.h.segers@vu.nl

1 Research Context and Problem Statement

Currently, cultural heritage portals limit their users to search only for individual objects and not for objects related to some historical narrative. Typically, most museums select objects for an exhibition based on the story they want to tell the public, but in digital collections this context can currently not be made explicit as the historical context is not part of the object annotations.

From previous experiences with cultural heritage portals such as Europeana[1] and CHIP[2], we observed that adding historical events to object descriptions is valuable for the grounding of cultural heritage objects in their historical context as events represent important changepoints in time and form the basic units of the historical narrative. Further, historical event descriptions are comprised of actors, locations and timestamps that are in some cases already present as facets of the object annotation. As such, adding events to object descriptions can enhance browsing and searching of cultural heritage collections as the events unify otherwise unrelated but historical relevant facets of object annotations.

The **problem** motivating this research is threefold: (1) There is no standard practice in cultural heritage organizations to include events within the object annotation, e.g. there is neither a shared vocabulary for the (historical) event descriptions, nor for different historical terms and concepts. The creation of such vocabulary is important in order to ensure alignment between the descriptions of historical events, which typically vary over time. (2) A multitude of different historical perspectives and interpretations of the same historical event exist. As a result, a variety of expressions can be used to describe the same event which implies problems for creating thesauri and vocabularies. (3) There is no consensus on the elements that make up an event and which could provide meaningful relationships between the events and pertaining objects.

This PhD research is situated in the **context** of two projects: (1) The *Semantics of History* project[3] with focus on the modeling and extraction of events and their perspectives from historical text documents; (2) *Agora*[4] with focus on

[1] http://www.europeana.eu
[2] http://www.chip-project.org/
[3] http://www2.let.vu.nl/oz/cltl/semhis/
[4] http://www.agora.cs.vu.nl

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 503–507, 2011.

searching, browsing and representing historical events online. Both projects use the digital collections of the Netherlands Institute for Sound and Vision[5] and the Rijksmuseum Amsterdam[6].

In my research, I focus on the following **research questions**:

1. What is an adequate event model to capture variances and invariances of historical events? Can we instantiate this event model (semi-)automatically?
2. What is an adequate organization of a historical ontology to be used for annotations for cultural heritage objects?
3. What is a suitable organization of a historical event thesaurus to allow for diverse interpretations and expressions for events?
4. What are relevant evaluation criteria for quality of the event model, the historical ontology and the thesaurus and their added value for exploration and search of cultural heritage objects?

Issues Related to the Research Questions

1. **The event model** provides the vocabulary for the event elements and their relations. However, the domain-specific requirements for modeling historical events in terms of *classes and properties* cannot be given beforehand. Additionally, the historical event model facilitates *relations between events*, e.g. causality, meronymy. However, these relations are not part of an event but exist as an interpretational element between two or more events and thus need to be modeled as a separate module.
2. **The historical ontology** serves as a semantic meta-layer to type historical events, independently of the expressions used. However, it is unknown to what degree an ontology can be used as an *unprejudiced meta-layer* for such typing as it might imply an interpretation. Ontologies typically represent a *time-fixed view on reality* which influences the modeling of objects that can only play a role in an event after a certain point in time. Additionally, the *expressivity and extensibility* of the ontology depends on the expressivity and extensibility of the event model and vice versa. It is critical to know how they interact, as incompatible properties can affect reasoning about events.
3. **The instantiation of the event model** needs to be based on different sources to capture the different perspectives and interpretations of events. Typically, event descriptions reside in unstructured text documents. Thus, portable information extraction techniques should be applied for detecting events and their elements in document collections of different style and topic.
4. **The event thesaurus** is a structured set of historical events used for event-based annotation of cultural heritage objects and for aligning different object collections. For the creation of such a thesaurus we need to know (1) how to identify and organize equal and similar event descriptions and (2) how to identify and structure multiple interpretations of the relations between events. Properties such as hasSubevent can become problematic for structuring the thesaurus, as some sources might only report temporal inclusion.

[5] http://portal.beeldengeluid.nl/
[6] http://www.rijksmuseum.nl/

2 State of the Art

This PhD work is related to four research areas. Here, we give a brief state of the art, a comprehensive overview of the related work can be found online[7].

Event models: Various event models exist, e.g. the Event Ontology[8], LODE [13], the F-Model [12], SEM [6] and CIDOC-CRM[9]. However, none were explicitly designed for historical events and each has various limitations concerning extending the model with domain-specific properties.

Model instantiation: Diverse information extraction (IE) techniques are used to instantiate models in a variety of domains, e.g. [1] and [4]. However, historical event extraction is emerging only recently.

Ontologies: Formal ontologies, e.g. DOLCE [10] and SUMO [11], and lexical databases, e.g. WordNet [5] exist that can partly be reused for historical ontology. Top-level ontologies pose modeling choices that may not compatible with the historical domain requirements. WordNet is language specific and not consistent in the modeling of synsets, which hampers the ontological soundness for an historical ontology.

Ontology learning can be seen as a subtask of information extraction that focuses on learning classes and relations between classes. Different techniques exist for ontology learning, e.g. [9], [2] but interpretational issues pertaining to historical events have not been addressed yet.

Related projects: A historical thesaurus [7] has been used in CultureSampo[10] to enhance searching and browsing of Finnish cultural heritage. It comprises event instances statically organized in a timeline and does not allow for various views on events. Modeling historical data and events has also been the focus of FDR/Pearl Harbor project [8] but no results have been published yet.

3 Approach

We propose the following novel approach for extracting and structuring knowledge of historical events from various text sources. First, we adapt an existing event model to meet the domain specific requirements. Next, we populate this model and learn a historical ontology using information extraction techniques.[11] For the creation of the event thesaurus we consider to use different reasoning techniques over both the instances and types of the modeled event descriptions. Following, we elaborate on the approach in relation to the research questions:

RQ1: We consider SEM[6] as a model to start from, as it is not domain-specific, represents a minimal set of event classes and includes placeholders for a foreign typing system.

[7] http://semanticweb.cs.vu.nl/agora/relatedwork
[8] http://motools.sf.net/event/event.html
[9] http://cidoc.ics.forth.gr/officialreleasecidoc.html
[10] http://www.kulttuurisampo.fi/
[11] see: http://semanticweb.cs.vu.nl/agora/experiments

RQ 2: We consider learning the ontology bottom up by using the facets of the extracted events as relevant terms in the domain[2]. WordNet is used as an external vocabulary to semantically organize the terms and determine the least common subsumer[3]. We consider to map the ontology to DOLCE to guarantee ontological soundness.

RQ 3: We consider to learn lexical patterns for extracting coarse-grained historical event descriptions from general Web documents and apply these to domain-specific text collections. These patterns are semantically rich and can be used to classify the extractions. The relevance scores for the patterns are used to determine the precision of the extractions. To boost the recall of the pattern-based extraction, we consider using the internal syntactic structure of the events as patterns.

RQ4: For the creation of the thesaurus, we consider temporal-spational reasoning methods to identify similar event descriptions. To identify explicit relations between events, we consider information extraction in the text documents. Further, implicit relations are inferred from the typing of the events.

4 Methodology

We apply the following iteration methodology in order to realize the approach in section 3, i.e. **Iteration I** is scoped on acquisition of basic models:

- Analysis of SEM classes for the information extraction process.
- Learn *patterns* to instantiate SEM classes, starting with the *event class*. Next, we extend to other classes and pertaining relations. We combine the results of three IE techniques: (1) pattern-based and (2) co-occurrancy based, both using Yahoo and Wikipedia and (3) lexical framing in newspaper collections. For each we evaluate the recall, precision and reusability.
- *Ontology*, version 1, based on the first extraction results.
- *Thesaurus*, version 1, with limited relations.
- Test and evaluate the *ontology* and *thesaurus* in the Agora demonstrator. We define new requirements from the evaluation.

In **Iteration II** we iterate all the RQs once again to extend the models with domain specific requirements:

- Extend the document collection to domain-specific texts, e.g. scopenotes with links to historical themes and historical handbooks. We scope the domain to two periods/themes of interest to the involved cultural heritage institutions. Apply the IE techniques and the extended event model. Creation of ontology version 2 with unprejudiced typing of events.
- Evaluate the ontology and thesaurus version 2 by applying the IE module and event model to another historical period/theme to ensure that the results are not over-fitting the data. Integrate the results in the Agora demonstrator.
- Define requirements for evaluating the thesaurus in the Agora demonstrator, e.g. added value in terms of links between objects (quantitative), added value in terms of relevant and coherent links (qualitative).

5 Achieved Results and Future Work

The PhD work is now entering the second year. Current work involves analysing the extracted events by the pattern-based IE. The results so far are:

- literature study on event models, requirements for an historical event model and best practices in the application of event models within different domains. (journal paper, accepted for JWS2010).
- study on historical events definition and modeling requirements; use case for event annotations of cultural heritage objects (Workshop Events2010).
- experiments with pattern-based event extraction (accepted abstract at CLIN'11).
- prototype of Agora portal for event-based searching and browsing of cultural heritage collections (demo accepted at Museums at the Web'11)

Future work will accomplish the steps in the approach. We also consider experiments on the portability of the results to other domains and languages.

References

1. Buitelaar, P., Cimiano, P., Magnini, B. (eds.): Ontology learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam (2005)
2. Cimiano, P.: Ontology Learning and Population from Text Algorithms, Evaluation and Application. Springer, Heidelberg (2006)
3. Cohen, W., Borgida, A., Hirsh, H.: Computing least common subsumers in description logics. In: Proceedings of AAAI 1992. AAAI Press, Menlo Park (1992)
4. de Boer, V.: Ontology Enrichment from Heterogeneous Sources on the Web. PhD thesis, VU University, Amsterdam, The Netherlands (2010)
5. Fellbaum, C. (ed.): Wordnet: An Electronical Lexical Database. MIT Press, Cambridge (1998)
6. van Hage, W., Malaisé, V., de Vries, G., Schreiber, G., van Someren, M.: Combining ship trajectories and semantics with the simple event model (sem). In: EiMM 2009, New York, NY, USA, pp. 73–80 (2009)
7. Hyvönen, E., Alm, O., Kuittinen, H.: Using an ontology of historical events in semantic portals for cultural heritage. In: ISWC 2007 (2007)
8. Ide, N., Woolner, D.: Historical ontologies. In: Words and Intelligence II: Essays in Honor of Yorick Wilks, pp. 137–152 (2007)
9. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems, 72–79 (2001)
10. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Schneider, L.: Wonderweb deliverable d18. Technical report, ISTC-CNR (2003)
11. Niles, I., Pease, A.: Towards a standard upper ontology. In: Proceedings of FOIS 2001, pp. 2–9. ACM, New York (2001)
12. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F—a model of events based on the foundational ontology dolce+dns ultralight. In: K-CAP 2009, Redondo Beach (2009)
13. Shaw, R., Troncy, R., Hardman, L.: LODE: Linking open descriptions of events. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 153–167. Springer, Heidelberg (2009)