A Knowledge-Based Framework for Events Representation and Reuse from Historical Archives

Marco Rovera(✉)

Department of Computer Science, Università di Torino, Turin, Italy
rovera@di.unito.it

Abstract. Thanks to the digitization techniques, historical archives become source of a considerable amount of biographical, factual and geographical data, that need to be structured in order to be usable in higher-level applications. In this paper we present the project of an ontology-based framework aimed at formally representing and extracting historical events from archives; this process should serve to the purpose of (semi)-automatically building narratives that allow users to explore the archives themselves. This proposal refers to a Ph.D. research at early stage and is part of a wider project, Harlock'900, established between the Computer Science department of the University of Turin and the Istituto Gramsci, a cultural foundation promoting research in contemporary history.

Keywords: Historical events · Ontology modeling · Events extraction

1 Introduction and Motivation

The progressive digitization process of historical archives makes documentary and archival contents spread beyond their traditional boundaries – the archives themselves – and opens new possibilities of employment for a significant amount of biographical, factual and geographical data available in the archival resources. On the other hand, the Semantic Web and Linked Open Data paradigm provides a stable technological layer for representing such information and for making it reusable in a wide range of applications. In such a context, how to extract and represent this knowledge, in order to make it accessible and usable for the end user, becomes an issue.

In this paper we propose to use a knowledge-based approach for extracting and representing historical events gathered from archives and to make them available for higher-level applications in the form of narrative chains. The resulting framework should provide a reusable ICT solution for the historical domain.

2 State of the Art

This project is situated at the intersection of different research areas; in this section we give a concise account of the most relevant works in such domains, highlighting interesting results that have been achieved, according to our purpose and scope.

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 845–852, 2016.
DOI: 10.1007/978-3-319-34129-3_53

Historical events and cultural heritage. In general, a comprehensive survey on the adoption of Semantic Web technologies in the historical research can be found in [20]. Our project shares some of the goals and, partially, the application domain of the Agora project, developed at the VU Amsterdam, aimed at providing a context for cultural heritage collections (in particular, museum objects), connecting them to relevant historical events [26]. The Agora project, in turn, was tightly related to the Semantics of History project, aimed at the extraction of historical events from textual resources. Further results of the two projects, particularly in providing semantically-based access to museum resources, can be found in [7]. An example of semantic, event-based access to cultural resources is depicted in [8] while in [4] events are first class concept in an ontology-based application that employs narratives for museums and exhibitions.

Event ontologies and ontology modeling. For a synthetic introduction to the concept of "event" in the philosophical thought, the "Events" item in the Stanford Encyclopedia of Philosophy can be taken as a reference [2]. As a first step of this Ph.D. project, an analysis of the existing ontologies has been carried out (and it is still ongoing) to assess how each model covers the concept of "event"; among the existing ontologies and models, the Simple Event Model (SEM) [27], the Linked Open Description of Events (LODE) [24], the Event Ontology [10] and the F Event Model [22] are explicitly designed for representing events, though they show different purposes and levels of expressiveness and reusability. SEM is a domain-independent model, aimed at representing events on the web. It explicitly models the concepts of *event*, *actor*, *time*, *space*, *role* and *authority* and it builds on a loose (inclusive) definition of "event". LODE is a lightweight and property-based ontology and it is the result of an alignment between different ontologies, namely ABC, CIDOC-CRM, Event Ontology and DOLCE + DnS Ultralite (DUL). LODE aims at describing the aspects of an event that result from the answer to the four w-questions: what is happened, where, when and who is involved. The Event Ontology (EO) is a minimal event model, it represents the basic features of an event and, though it can be considered domain-independent, it was designed having in mind the description of music performances (like concerts). The F-Event Model is an upper ontology extending DOLCE + DnS Ultralite (DUL); by means of a rich conceptualization and a quite complex formalization, it allows to describe events in time and space, participation, structural relations between events (like mereology, causality, correlation), documentary support about an event and interpretation.

Other ontologies, like CIDOC-CRM [19] and the Europeana Data Model [9], provide a modeling of the concept of event within the wider framework represented by Cultural Heritage-related resources. An interesting classification of events (at upper level) can be found in [18], where the authors present a formal characterization of events, based on order-sorted logic. Among the foundational ontologies, DOLCE [1] implements in its backbone taxonomy one of the possible classifications of events discussed in [2], distinguishing between *activities*, *accomplishments*, *achievements* and *states*. The analysis also suggested that there are ontological problems yet not tackled in the representation of historical events [See Sect. 3].

Semantics for diachronical geography. As illustrated in Sect. 3, one of the goals of this Ph.D. proposal is to support a diachronical perspective on the geographical knowledge involved in the representation of historical events and concepts; in this field, valuable experience can be provided by projects like Pleiades [21], a gazetteer of places of ancient times, used for educational purposes, and GeoLat [3], a system allowing users to explore the geography of ancient times in Latin literature texts.

Event extraction. The task of extracting events from textual corpora has been performed according to different approaches, that can be roughly classified in data-driven or knowledge-driven [16]. While the former employ quantitative, statistical techniques and require to be trained on large textual corpora, the latter rely on domain knowledge and perform text mining by means of rules and patterns. A set of publicly available Information Extraction tools (some of which including event extraction), combining NLP and Semantic Web technologies, has been reviewed in [11]. However, as far as historical events are concerned, few examples are available (see [6, 23]), but there seems to be no benchmark to refer to. An interesting result has been achieved in [15], though in that case the input was represented by Wikipedia pages (i.e., a single type of source, providing semi-structured content) rather than by heterogeneous plain texts.

3 Problem Statement and Contributions

The Ph.D. proposal partially originates within Harlock'900, a three-year project established in December 2015 and involving the Department of Computer Science of the University of Turin and the Istituto Piemontese Antonio Gramsci, a non-profit institute promoting research on contemporary history topics. An overview of the approach that will be adopted in Harlock'900 can be found in [13]. The project's main goal is that of employing a set of resources from the Istituto Gramsci's archive and enriching the existing metadata with information describing the content of the resource itself; such information will be expressed using formal representations based on ontologies. The goal of the enrichment process is that of allowing the reuse and exploration of information for application (e.g. educational, touristic) purposes. The enriched metadata represent a semantic layer that will ensure a content-based access to archival resources.

The semantic layer should take into account the following aspects:

- supporting a diachronical perspective on geography, connecting the representation of a place in history with the actualized representation of the same place;
- connecting the representation of factual/historical entities at different granularity levels, allowing, for example, both a biographical and a general history perspective to be mutually connected and aligned;
- supporting the use – at application level – of narrative structures by giving a formal representation of events, their factual components and their complex relations.

This Ph.D. research will become part of such a wider project, taking on specific responsibility for the following **research questions**:

- what are the full requirements of the semantic layer (from an ontological point of view), accordingly to the previously stated aspects?
- Are the currently available event models adequate to fulfill the mentioned aspects?

- If not, what extensions to existing ontologies or integrations between different ontology modules are required for the purpose?
- Which tools can be used to automatically or semi-automatically perform event mining from text and annotation based on the model?
- What contribution can be given by Linked Open Data (LOD) (as input) in such a process?
- What contribution can the resulting framework give to the LOD (as output)?

The **research problem** at the basis of this proposal can be further decomposed into three main sub-problems:

- from the analysis performed so far it seems that the current available event models are not able to accurately account for all the conceptual requirements in the historical domain: for example, the reviewed models do not seem to provide the needed means to accurately represent collective entities (e.g. a political party); secondly, it seems difficult, using such models, to give a diachronical representation of places, asserting for example that a certain place, named in a certain way in 1920 and having a certain role at that time, was named differently in 2005 and had a different purpose (a military base became a school, a private house became an hotel, and so on). Third, in the event extraction process it could be beneficial to have an articulated taxonomy of historical events available, to classify the extracted events. In wider terms, it seems that, when individually taken, none of the available event models is able to cover all the conceptual aspects involved in a rich representation of (historical) events.
- an event-based framework supporting narrative is needed in order to put the domain information enclosed in historical archives at the service of different sorts of applications; this problem could impose backward requirements, especially on the event model;
- despite remarkable improvements in Natural Language Processing (NLP) and Semantic Web methodologies during the last decade, and even if some state-of-the-art tools are available, there still is no standard for event extraction from text; this project does not have as main goal that of improving NLP methodologies themselves; nevertheless, the availability of a domain event model, designed for the purpose of representing historical events, could allow us to treat our archival resources as a valuable test bed for such technologies.

The Ph.D. proposal aims at providing an event-based framework able to make contents provided by historical archives reusable for original applications. The underlying **research hypothesis** is that none of the available event models, individually taken as they are, provides a covering of all the aspects that characterize the concept of (historical) “event”, while their integration (and their enrichment, where necessary) would allow us to succeed both in extracting and in representing the semantics of events in historical archives in a form that is suitable for their exploitation in novel contexts (such as tourism or education applications) provided by the Harlock’900 project.

4 Research Methodology and Approach

In order to achieve our research goals, we will adopt an iterative methodology, involving the following steps: (*a*) analysis of the existing event ontologies and, simultaneously, (*b*) requirements and data analysis (based on an appropriate selection of texts from the historical archives); (*c*) design hypothesis of the model and (*d*) resources annotation based on the model. "Outside" the iterative cycle, once the model will have become stable enough, the connection with the LOD will be established by contributing with the extracted data and, if necessary, by providing mappings of our model.

It is relevant to point out that, while the Harlock'900 project aims at covering different resource types, the Ph.D. proposal will focus only on textual digitized resources (hence images, posters, handwritten papers and similar are excluded from its scope). In particular, given the traditional distinction between primary and secondary historical sources (as recalled in [20]), the considered input data at application level will consist of primary ("narrative") sources like biographies, newspaper articles, chronicles. In the analysis stage, also secondary resources (e.g., history books) will be employed, as well as digital historical datasets (see for instance the one cited in [15] and available as API at http://www.vizgr.org/historical-events/) for hands-on testing of the existing event models and for creating a corpus of semantic representations of historical events to be used by the system. Since the Harlock'900 project aims at making accessible archival resources, which enclose much information about historical events (that are, in that form, often barely accessible), primary narrative sources are more interesting for the purpose of this work.

The described methodology will be instantiated as follows:

- requirements: a textual corpus from the archive will be selected and manually analyzed, extracting the ontological design requirements; simultaneously, the available event models are analyzed in order to identify the requirements fulfilled by each model, design patterns, lacks in coverage and differences between the models;
- hypothesis on the model: based on the previous step, an ontological analysis of the event domain will be carried out, in order to define the intended model; in doing so, particular attention will be paid to the reuse of existing (parts of) ontologies, integrating them as needed; in fact, the analysis will not necessarily produce a new ontological model (i.e., it is not aimed at creating a new ontology from scratch), but it should identify the classes and relations needed for the purpose of the project and the existing models that provide them. Only if needed, an extension in terms of classes or properties will be realized, but the current work is mainly conceived as an integration between different models. This approach will also guarantee a high degree of interoperability with existing ontologies. Furthermore, in the design process, state-of-the-art design methodologies will be employed like OntoClean [14];
- event extraction and annotation: based on the designed model, historical events must be extracted from textual resources and annotated; a set of tools will be implemented/ employed that make use of extraction techniques [11]; a second means for enriching resources (that will be taken into account in Harlock'900 but not in this Ph.D. project)

is that of allowing “trusted users” – i.e. domain professionals – to contribute by editing metadata in order to participate to their enrichment.

5 Current Work

Since the presented project refers to a Ph.D. at early stage (beginning: 1 October 2015), no significant result has been achieved so far. Current work is focused on the analysis of the literature and of the existing event ontologies, in order to identify requirements and design choices. Also, the ontology design and annotation experience gained in previous projects, like the one documented in [12], will represent a valuable starting point. The next steps in the analysis will be (a) a hands-on test of the models aimed at understanding their strength and weaknesses in the annotation of historical texts, and (b) an investigation of the LOD cloud to verify which models are employed to describe events, how many and what kind of events are represented.

6 Evaluation Plan

Within the described proposal, three main phases will be subject to evaluation. The *semantic event model* will be evaluated during each iteration cycle, identifying a suitable corpus of texts in order to asses the coverage and expressiveness of the designed model. The *extraction and annotation tool(s)* will be evaluated using traditional statistical performance measures (precision, recall, accuracy, for instance). Lastly, once the whole *framework* is available and operative, a user study will be set up to test its main functionalities with end users.

7 Conclusions

In this paper I outlined my Ph.D. proposal, aimed at extracting semantic knowledge from historical archives and making it available for application purposes. To this end, I stated the main research problem I intend to tackle, together with the approach and research methodology to be followed. Also, the steps to be evaluated and the evaluation approach have been sketched. Due to the early stage of this Ph.D., it has not been possible to provide achieved results yet.

References

1. Borgo, S., Masolo, C.: Foundational choices in DOLCE. Handbook on Ontologies. International Handbooks on Information Systems, pp. 361–381. Springer, Heidelberg (2009)
2. Casati, R., Varzi, A.: Events. Stanford Encyclopedia of Philosophy (2002). http://plato.stanford.edu/entries/events (Substantive revision 2014). Accessed 10 Dec 2015
3. Ciotti, F., Lana, M., Tomasi, F.: TEI, ontologies, linked open data: geolat and beyond. J. Text Encoding Initiative (8) (2014)

4. Collins, T., Mulholland, P., Wolff, A.: Web supported employment: using object and event descriptions to facilitate storytelling online and in galleries. In: Proceedings of 4th Annual ACM Web Science Conference, pp. 74–77. ACM, June 2012
5. Cybulska A., Vossen P.: Event models for historical perspectives: determining relations between high and low level events in text, based on the classification of time, location and participants. In: LREC (2010)
6. Cybulska, A., Vossen, P.: Historical event extraction from text. In: Proceedings of 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics (2011)
7. de Boer, V., Oomen, J., Inel, O., Aroyo, L., van Staveren, E., Helmich, W., de Beurs, D.: DIVE into the event-based browsing of linked historical media
8. Den Akker, C., van Aroyo, L., Cybulska, A., Van Erp, M., Gorgels, P., Hollink, L., Jager, C., Legene, S., van der Meij, L., Oomen, J., van Ossenbruggen, J., Wielinga, B.: Historical event-based access to museum collections. In: Proceedings of 1st International Workshop on Recognising and Tracking Events on the Web and in Real Life (EVENTS2010), Athens, Greece, May 2010
9. EDM Definition v. 5.2.6. http://pro.europeana.eu/page/edm-documentation
10. Event Ontology (EO). http://motools.sourceforge.net/event/event.html
11. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 351–366. Springer, Heidelberg (2013)
12. Goy, A., Magro, D., Petrone, G., Rovera, M., Segnan, M.: A semantic framework to enrich collaborative tables with domain knowledge. In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 3, pp. 371–381. KMIS (2015)
13. Goy, A., Magro, D., Rovera, M.: Ontologies and historical archives: a way to tell new stories. Appl. Ontol. (2015, in press)
14. Guarino, N., Welty, C.A.: An overview of OntoClean. Handbook on ontologies. International Handbooks on Information Systems, pp. 201–220. Springer, Heidelberg (2009)
15. Hienert, D., Wegener, D., Paulheim, H.: Automatic classification and relationship extraction for multi-lingual and multi-granular events from Wikipedia. In: Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012), vol. 902, pp. 1–10 (2012)
16. Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F.: An overview of event extraction from text. In: Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at 10th International Semantic Web Conference (ISWC 2011), vol. 779, pp. 48–57, October 2011
17. Hyvönen, E.: Semantic portals for cultural heritage. Handbook on Ontologies. International Handbooks on Information Systems, pp. 757–778. Springer, Heidelberg (2009)
18. Kaneiwa, K., Iwazume, M., Fukuda, K.: An upper ontology for event classifications and relations. In: Orgun, M.A., Thornton, J. (eds.) AI 2007. LNCS (LNAI), vol. 4830, pp. 394–403. Springer, Heidelberg (2007)
19. Le Boeuf, P., Doerr, M., Ore, C.E., Stead, S. (eds.): Definition of the CIDOC Conceptual Reference Model (Version 6.1). ICOM/CIDOC CRM Special Interest Group (2015)
20. Meroño-Peñuela, A., Ashkpour, A., Erp, M., Mandemakers, K., Breure, L.: Semantic technologies for historical research: a survey. Semant. Web J. **6**(6), 539–564 (2015)
21. PLEIADES project. http://pleiades.stoa.org/. Accessed 10 Dec 2015
22. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F–a model of events based on the foundational ontology DOLCE + DnS ultralight. In: Proceedings of 5th International Conference on Knowledge Capture (K-CAP 2009). pp. 137–144. ACM, New York, NY, USA (2009)

23. Segers, R., Van Erp, M., Van Der Meij, L., Aroyo, L., van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J., Jacobs, G.: Hacking history via event extraction. In: Proceedings of 6th International Conference on Knowledge Capture, pp. 161–162. ACM (2011)
24. 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)
25. Van Den Akker, C., Aroyo, L., Cybulska, A., Van Erp, M., Gorgels, P., Hollink, L., Wielinga, B.: Historical event-based access to museum collections. In: Proceedings of 1st International Workshop on Recognising and Tracking Events on the Web and in Real Life (EVENTS 2010), Athens, Greece (2010)
26. Van Den Akker, C., Legêne, S., Van Erp, M., Aroyo, L., Segers, R., van Der Meij, L., Van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J., Jacobs, G.: Digital hermeneutics: agora and the online understanding of cultural heritage. In: Proceedings of 3rd International Web Science Conference, p. 10. ACM, June 2011
27. Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (SEM). Web Semant.: Sci. Serv. Agents World Wide Web **9**(2), 128–136 (2011)