A Semantic Web Approach to Enable a Smart Route to Historical Archives

Annamaria Goy, Diego Magro and Alessandro Baldo

Dipartimento di Informatica, Università di Torino, Turin, Italy
E-mail: annamaria.goy@unito.it; diego.magro@unito.it; baldoalessandro@protonmail.com

Received 08 January 2019;
Accepted 06 June 2019

Abstract

In this paper we show that an ontology-based approach can be beneficial for enhancing the access to cultural resources, and in particular historical documents. The paper starts with an overview of our approach, aimed at providing online archival systems with a semantic layer based on Semantic Web standards (OWL 2 and RDF). Two projects are introduced, namely Harlock900 and PRiSMHA, carried out in collaboration with local cultural institutions owning rich historical archives. In particular, the paper describes the computational ontologies supporting the approach, and then focuses on two case studies showing that our framework provides better results if compared with standard access systems. The case studies show the enhancement provided by a semantically rich representation of time intervals and a detailed formal description of events and their participants.

Keywords: Semantic Web, Intelligent Web applications, Ontology-driven Web applications, Digital Humanities, Web-based access to historical archives.

Journal of Web Engineering, Vol. 18_4-6, 287–318.
doi: 10.13052/jwe1540-9589.18462

1 Introduction

In the era where almost everything seems to be based on Artificial Intelligence, the major role is played by Machine Learning (and often Deep Learning) approaches. However, we think that there are some cases in which "vintage" knowledge-based approaches can still be useful. In the domain of Cultural Heritage, for example, many cases can be found where sub-symbolic approaches can hardly be applied. Machine Learning and Deep Learning, in fact, require huge amounts of (annotated) data, which – in this domain – are often not available. In particular, when dealing with historical archives, metadata are usually very poor with respect to the needs of such techniques, and textual resources are frequently not obtainable (this is the case, for instance, of drawings, hand-written texts, old type-written documents that are very hard for OCR tools, etc.).

A couple of good objections to these claims could be the following: (a) There is a lot of work on OCR techniques, image processing, and hand-written text recognition (mainly based on Machine Learning and Deep Learning), which make (annotated) data actually available; see, for instance, the IAPR ICDAR (Int. Conf. on Document Analysis and Recognition) conference series (icdar2019.org). (b) Nowadays there are lots of datasets concerning Cultural Heritage available on the Web (see, for example, Europeana LOD Pilot [25]).

We have two major counter-objections: (a) The state-of-the-art tools actually available to cultural institutions that own historical archives are not yet aligned with state-of-the-art research prototypes, at least in many national or local contexts; (b) Using already available datasets keeps Cultural Heritage in a sort of "bubble", where the ICT-enabled cultural institutions gain new innovation opportunities, while those organizations that fell behind – which are those that mainly need to be involved in ICT projects – will continue lagging.

In this perspective, we believe that knowledge-based, symbolic approaches can still play an important role in supporting access to – and thus the exploitation of – cultural resources, and in particular documents stored in historical archives.

The major goal of our research activities is to build a web-based "smart archivist", able to provide users with: (a) Patterns and

meaningful connections between events, people, places archival documents "talk about"; (b) Relevant and useful documents. Exactly as a good human archivist, a digital system can be enabled to perform such tasks only if it is provided with a rich knowledge about the content of the documents stored in the archives.

The aim of the work described in this paper is to verify the following research hypothesis: A digital system providing access to historical archives, endowed with a *rich semantic layer* describing the content of documents, provides users with better results in terms of:

- Events – described in the documents – retrieved and displayed, in particular by referring to the time intervals they occur in (i.e., it provides and visualizes a larger and more accurate set of events with respect to standard timelines);
- Documents retrieved from the archives (i.e., it provides a larger and more precise set of documents with respect to a classical keyword-based retrieval system).

A *rich semantic layer* consists of metadata about the content of archival resources, based on a highly axiomatized ontology which characterize in details the domain (see Section 3.2).

In order to verify this research hypothesis, we developed a software prototype handling time expressions (within the Harlock900 project), and a mockup of a concept-based document retrieval system (within the PRiSMHA project), on which we performed two walk-through navigations, starting from two typical information needs (see Section 3).

The major contribution of this paper is to show that an approach based on a *rich semantic model* (in the sense specified above) can provide a significant enhancement in the quality of the access to archival resources. This result is relevant since semantic approaches in this field are usually based of very lightweight semantic models, which do not support the fine conceptual granularity proposed by our approach, and thus could not reach the same result quality. The comparison with existing semantic models is discussed in Section 2.2.

In the rest of the paper we provide: An overview of our approach, aimed at endowing archive metadata with a *semantic layer* based on

Semantic Web standards (Section 2.1); An account of the most relevant related work (Section 2.2); Two case studies, supplied by two projects, namely Harlock900 and PRiSMHA (Section 3.1). In particular, we describe the computational ontologies supporting the approach (Section 3.2), an evaluation of the TIME module of the core ontology (Section 3.3), and an evaluation of the EVENT modules of the core and domain ontologies (Section 3.4). Finally, we summarize the major conclusions and suggest some promising future research directions (Section 4).

2 Introducing a Semantic Layer Over Archives

2.1 Our Framework

Figure 1 shows the architecture of the overall system: the knowledge about the content of the documents is stored in the *Semantic Knowledge Base*, which is implemented as a *RDF triplestore*, following the principles of Linked Data [26]. In line with a well-consolidated approach [18, 40], the semantic representation of documents content is built around

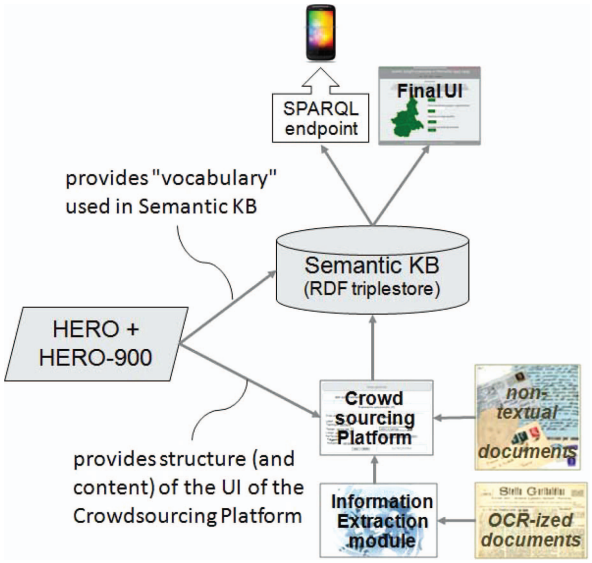


Figure 1 System architecture, showing the role of the semantic layer.

the notion of (*historical*) *event*, which represents a "glue" connecting places, time periods, people, organizations, and collective entities.

The vocabulary used in the Semantic KB is provided by the *HERO ontology*; in particular, HERO is a *core ontology* – thus providing general concepts and properties, such as *event*, *person*, *collective*, *organization*, *time interval*, and so on – and is coupled with a *domain ontology*, called HERO-900, that provides domain specific concepts and relations. As we will see in Section 3.1, the domain is represented by (subsets of) the Italian history of the 20th century. HERO and HERO-900 will be described in Section 3.2.

One of the major objection to knowledge-based approaches is the bottleneck represented by knowledge acquisition: transferring the human archivist competence into a digital system, i.e. building rich RDF datasets describing the content of the documents, is an extremely demanding work, which can threaten the sustainability of the approach. To face this challenge, we decided to rely on a *Crowdsourcing Platform*, where users can access archival documents and collaborate in building the Semantic KB. Through a User Interface driven by the underlying ontology, such a platform populates the RDF triplestore. Moreover, when full-text OCR-ized documents are available, an *Information Extraction module* identifies Named Entities, which are then used by the Crowdsourcing Platform to support users with suggestions about people, places, organizations, time intervals, and events occurring in texts. Both the Crowdsourcing Platform and the Information Extraction module are work in progress. Relevant references can be found in Section 2.2.

The semantic layer, i.e. the Semantic KB based on the ontology, can be exploited by a *Final User Interface*, enabling a smart and flexible access to resources from historical archives [4, 11, 18]. The advantages represented by the semantic layer, when exploited by such a UI, will be discussed within two case studies, presented in Sections 3.3 and 3.4.

Moreover, the RDF triplestore can be made available also through a *SPARQL endpoint* [21], thus offering third parties the possibility of exploiting the knowledge about historical events, narrated in archival documents, to develop innovative applications in different fields, such as education, tourism, or entertainment.

2.2 Related Work

In recent years, the effectiveness of semantic approaches (including ontologies and Linked Data best practices) in publishing, connecting, and searching history and cultural heritage datasets has been largely documented [32, 34]. The interest for Semantic Web oriented approaches in the cultural heritage domain is proven by projects such as Europeana (www.europeana.eu) – the European Union digital library providing access to cultural heritage digitized contents of hundreds of European galleries, libraries, archives – and CIDOC Conceptual Reference Model (www.cidoc-crm.org) – an ISO standard that has been used in many projects about cultural heritage; see, for instance, WarSampo (seco.cs.aalto.fi/projects/sotasampo/en), a project aiming at publishing datasets about the Second World War in Finland as Linked Open Data, or PAPYRUS (www.ict-papyrus.eu), an EU project enabling users to query digital libraries to discover cross-domain relationships between concepts.

CIDOC-CRM provides a quite rich characterization of the specific types of events involving cultural heritage (e.g., the transfer of custody of an item in a museum), but its level of granularity is not enough to model in details historical events (although we aligned the top level of HERO with CIDOC top level corresponding notions for interoperability reasons: see Section 3.2).

Many projects in this research area provide semantic models centered on the notion of *(historical) event*, e.g., the Europeana Data Model (i.e., the metadata model adopted within Europeana) [28], the Event Ontology (purl.org/NET/c4dm/event.owl), LODE [38], SEM [24] – used within projects such as Agora [2] and DIVE [5], supporting users in event-centric browsing of cultural heritage items.

The Europeana Data Model provides a basic notion of event, enabling the representation of "who does what when and where", but it lacks any further semantic characterization of historical events. The Event Ontology (designed having in mind the description of music performances) is a minimal model representing the basic features of an event. LODE is a lightweight property-based ontology, resulting from the alignment of different ontologies (including CIDOC-CRM and Event Ontology); it aims at describing the classical four aspects of

an event: what is happened, where, when and who is involved. SEM is a domain-independent model, aimed at modeling the concepts of event, actor, time, space, role and authority.

All these ontologies, if compared with our model (described in Section 3.2), present a general very light axiomatization, which implies a loosely formal characterization of the involved concepts. In particular, they shows:

- A weak (or absent) characterization of domain-specific types of events (see, for instance, the events involved in the second case study of this paper, Section 3.4);
- A very simplified representation of time intervals (and space regions);
- A weak (or absent) characterization of properties and relations between events (e.g., causal relations), or between events and entities (e.g., event participants: see [22]).

As far as the User Experience is concerned, several projects could be mentioned, both outside the cultural heritage domain (e.g., [37]) and within it; see, for instance: Europeana Space (www.europeana-space.eu); Europeana Creative (www.europeanacreative.eu); AXES (www.axes-project.eu), that provides advanced multimodal search and access to audiovisual digital resources (search for spoken words in audio, for images within video sequences, for images depicting specific kinds of objects or similar to other images); the Atlas of Nazi and Fascist massacres (www.straginazifasciste.it/?lang=en); Memorie di Guerra (War Memories: www.memoriediguerra.it), where advanced search functionalities are offered to explore World War textual document [7]; ALCIDE [33], enabling users to select the time span to search for and display keywords and entities mentioned in the retrieved documents. None of the analyzed projects actually offers a full-fledged conceptual access to documents content.

Moreover, two fields mentioned in Section 2.1 are relevant for our research activity: Crowdsourcing and Information Extraction. Crowdsourcing is becoming a fundamental approach for providing rich meta-data describing resources that cannot be automatically processed, such as old pictures, handwritten documents, and so on [3];

see, for instance, the Scribe (www.nypl.org/blog/2015/11/23/scribe-framework-community-transcription) and Micropasts (crowdsourced.micropasts.org) projects. Furthermore, a significant research effort has been devoted to the task of automatically extracting information about events from texts; see, for instance, [12, 27, 39]. In particular, several works showed that historical texts represent a peculiar domain, where IE/NER tools show quite low performances compared to other domains; see, for example [7, 14, 33, 36]. We do not analyze the related work within these two fields in depth, since these aspects fall outside the scope of this paper.

3 Case Studies

3.1 The Projects and the Two Case Studies

The research activity sketched in Section 2.1 is carried out within two national projects, namely Harlock'900 (di.unito.it/harlock900) and PRiSMHA (di.unito.it/prismha).

The main goal of Harlock'900 is to exploit semantic technologies (computational ontologies and Linked Open Data) to make cultural heritage stored in historical archives accessible and usable in innovative ways, by users with different skills and interests. The project, running 2016–2019 and funded by University of Torino, is carried out by the Computer Science Dept. within the framework of a collaboration with the Fondazione Istituto piemontese Antonio Gramsci (www.gramscitorino.it). It aims at enabling a content-based access to archival resources by metadata enrichment. The selected use-case domain is represented by the events occurred between 1943 and 1945 during the "Resistenza" (the local partisans struggle against the Nazi occupation and the Fascist regime) in Piemonte (North-West of Italy).

An overview of the approach adopted in Harlock'900 can be found in [18]; an early-stage prototype is described in [11]. In [4] a full pipeline, from rough texts up to the final User Interface is presented, focusing on time expressions and showing the process of building and exploiting the semantic layer, i.e., the semantic representations

of documents content. We will describe the user experience enabled by the User Interface based on such a pipeline in Section 3.3.

PRiSMHA (Providing Rich Semantic Metadata for Historical Archives), running 2017–2020, is funded by Compagnia di San Paolo and University of Torino. It involves the Computer Science and Historical Studies Departments and relies on the collaboration with the Polo del '900 Foundation (www.polodel900.it), a cultural center, headquartered in Torino, involving nineteen cultural institutions hosting an extremely rich set of historical testimonies. In particular, PRiSMHA exploits documents from the archives and library of the Fondazione Istituto piemontese Antonio Gramsci, which is the major "contributor" (25%) of the Polo del '900 archives.

The main goal of the project is twofold, since it aims at demonstrating that a content-based access to archival resources, based on semantically rich metadata: (a) is sustainable, despite the overload imposed by knowledge acquisition; (b) provides a significant enhancement in the possibilities of exploitation of archival resources. To achieve the first goal, PRiSMHA is investigating the application of a crowdsourcing collaborative model, where trusted users participate in building the semantic layer, guided by an ontology-driven User Interface and supported with suggestions provided by automatic information extraction techniques (see the *Crowdsourcing Platform* and the *Information Extraction module* in Figure 1). To face the second challenge – i.e., evaluating the impact of a rich formal semantic representation of documents content on the effectiveness of archival search – we developed a mockup of the final User Interface and evaluated it on a subset of the available documents. This study is reported and discussed in Section 3.4. The domain selected for testing the approach is the students and workers protest during the years 1968–1969 in Italy. An overview of the main research issues and challenges in PRiSMHA can be found in [19].

In the following we present two case studies aimed at demonstrating that a content-based access to historical archives, based on a rich semantic layer, represents an advantage for the user looking for information and original documents. For the first case study, we annotated 120 pages extracted from biographies and testimonies about the "Resistenza"

in Piemonte, from Istituto Gramsci's library; in our approach, an "annotation" is a link between a document fragment (possibly the whole document) and a set of RDF triples describing events referred to in the fragment, together with their properties (basically, time, place, and participants with their roles). The extraction of temporal information from texts has been supported by HeidelTime [41], while the extraction of events has been performed manually (see [11] and [4]). For the second case study, we manually annotated 50 original documents about 1968–69 movements in Italy (newspaper articles, letters, leaflets, pamphlets, etc.) from Ist. Gramsci's archives.

The case studies focus on two perspectives that are particularly relevant in the historical domain, represented by *time* and *events*. In particular, the first case study (Section 3.3) is focused on the role played by a semantically rich representation of *time intervals*: the navigation possibilities supported by the semantic layer are compared with those offered by a standard timeline. The second case study (Section 3.4) exploits a detailed formal description of *events* and their *participants*: in this case, the study shows the advantages of our approach in retrieving archival documents, compared with a classical keyword-based search engine. Both cases rely on the vocabulary provided by HERO and HERO-900, thus, in the next section, we briefly describe the structure of such ontologies, mainly focusing on those features directly involved in the evaluation.

The case studies are built as walkthrough navigations starting from two example information needs: a navigation in the Semantic KB (first case study) and a search in the document index (second case study) are simulated exploiting real documents and actual formal representations in the RDF triplestore.

3.2 The Ontologies: HERO and HERO-900

HERO (Historical Events Representation Ontology) and HERO-900 are two OWL 2 ontology suites composed of different modules. The HERO suite is a *core modular ontology*, which provides the basis for representing things that happen in the physical or social world at the scale of human affairs. It contains five modules: HERO-TOP, HERO-EVENT, HERO-PLACE, HERO-ROCS and HERO-TIME.

HERO-TOP is the upper level module in HERO and it accounts for general notions, which represent the topmost level of each HERO module. It is rooted on basic ontological distinctions provided by DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) [6]. HERO-TOP distinguishes between *objects* (i.e., entities – either physical or non-physical – that are in time, such as buildings, persons, nations, laws, etc.), *perdurants* (i.e., entities that happen in time, such as elections, wars, earthquakes, etc.) and *abstract entities* (i.e., entities that are outside space and time, such as sets, time intervals, etc.). The basic relation between perdurants and objects is the *participation* relation, representing the general involvement of objects in perdurants (e.g., persons may participate in elections, buildings may participate in earthquakes, etc.). Among non-physical objects, *social objects* (i.e., objects that generically depend on a community of agents, such as organizations, laws, concepts, social roles, etc.) play an important role in human affairs. Both physical and non-physical objects can be structured into *parts*.

According to the framework provided by the DnS ontology [17], HERO-TOP puts at disposal the basics for talking about (reified) concepts (such as institutional and social roles, political identities, etc.). In particular, it offers the basic relation of *classification* between concepts and ground entities (e.g., the concept of parliamentarian "classifies" people that are members of a Parliament) and the *subsumption* relation between concepts (i.e., the concept of Italian parliamentarian subsumes that of Italian deputy). Moreover, HERO-TOP provides the mechanism for expressing both qualitative and (approximate or exact) quantitative *measurements* (specifying both the *unit of measure* and the *measurement value*) and for comparing measurement expressions (e.g., that event had a longer duration than that other event, in a battle participated more soldiers than in another, etc.).

HERO-EVENT refines HERO-TOP for what it concerns perdurants. Among perdurants, it distinguishes between *events* (*Event* class) and *states* (*State* class). The former represents perdurants producing changes in some state of affairs (e.g., person births, meetings, elections, etc.), while the latter represents a persistence of some state of affairs (e.g., being married, working at a company, etc.). Events can

be *actions* (*Action* class), representing intentional events. Perdurants happen in some *places* and in a *time interval*. HERO-EVENT provides different properties for expressing relations between perdurants and time intervals (e.g., *hasTimeSpan*, *hasInitialTime*, *hasFinalTime*).

Moreover, entities may *participate* in perdurants in different ways (e.g., as *agents* intentionally performing an action, as *patients* that are affected by an event or state, as *instruments*, etc.). Such different ways of participation in perdurants are represented by *thematic roles* and are included in HERO-EVENT as subproperties of the *hasParticipant* property, linking perdurants to objects or set of objects (i.e., *hasAgent*, *hasPatient*, *hasInstrument*, etc.) [20, 22].

Moreover, any entity may *participate in a perdurant as* an entity of a certain kind or playing a certain role. It is worth noting that the ways of participation (represented by *thematic roles*) and the *participation as* are two orthogonal aspects. Let us consider, for instance, the murder of the king of Italy Umberto I by Gaetano Bresci in 1900: Gaetano Bresci was agent (thematic role) in that event, in which he participated as an anarchist, while Umberto I was patient (thematic role) and he participated as a king (see also [22]). Events and states may *influence* (e.g., they may be causal factors of) other events and states and they may be structured in *sub-events* and *sub-states*.

HERO-PLACE refines HERO-TOP by characterizing *geographic features*. In this context, a geographic feature is any place in which an event or state may occur. Places may be *natural* (e.g., mountains, forests, etc.) or *artificial* (e.g., buildings, streets, etc.) and they may be the *location of* human organizations, such as companies, countries, etc.

HERO-ROCS mainly accounts for the notions underlying social and institutional reality, in particular *roles*, *organizations* and *collective entities*. According to Masolo and colleagues [30], a *role* is an anti-rigid [42] and well-founded concept. Examples of roles are: Head of State, French Head of State, French citizen, workman, workman at FCA, etc. Roles are *played by* entities, e.g. Mr. Emmanuel Macron is currently playing the French Head of State and the French citizen roles.

A second central notion defined in HERO-ROCS is that of *organization*. Organizations are modeled starting from their ontological analysis discussed by Bottazzi and Ferrario [9]. Organizations should

be intended in a broad sense, as social objects encompassing companies, hospitals, political parties, countries, criminal organizations, etc. Organizations can be *located in* physical places, e.g. a company is located in one or more buildings, a country is located in one or more areas on the earth, etc. Moreover, organizations *involve* entities. For instance, a company involves both industrial facilities and employees. Involved agents are, more specifically, *affiliated with* the organization. For instance, the employees of a company, being agents, are affiliated with the company. Organizations may have *sub-organizations*. For example, a company may be structured in departments; a nation may be subdivided into states, etc.

The third pillar in HERO-ROCS are *collective entities*. Collective entities are collections of entities (their *members*); they have their own individuality and their properties or behavior cannot be (conveniently) reduced to properties or behavior of their members. The notion of collective entity (usually called "collection" or "collective") has been investigated from an ontological point of view by several researchers; see, for instance, [8, 16, 23, 43]. Examples of collective entities are the working class, the EU working class, Italian citizens, EU citizens, the naval fleet of Spanish Navy, etc. In HERO-ROCS, collective entities are non-physical objects.

In the historical domain, it is not always possible or appropriate to specify exact and detailed information about events and entities. For instance, a typical description of the atomic bombing of Hiroshima in August 1945 that we can read in many documents (books, newspapers, etc.) does not contain the list of all the victims, nor it specifies their exact number. Usually, it reports an approximate number of victims (without saying who actually they were). Besides qualitative and approximate measurement expressions (mentioned above), HERO-ROCS puts at disposals also *sets*, which can be used in order to specify information by abstracting from some details. In this way, it is possible to specify, for instance, that a set of people (whose actual members are not specified), with an approximate cardinality of 140,000, participated as victims in the atomic bombing of Hiroshima. Both collective entities and sets can be *described by* concepts, meaning that those concepts *classify* all their members. A typical case is that of a collective, such as *workers*, that

is "described by" the social role *laborer* (defined in HERO-900 as an instance of the HERO *Role* class, subclass of the *Concept* class), since all of its members are laborers (i.e., they play the social role of laborer).

HERO-TIME distinguishes between *time intervals* and their *expressions* and it provides the basics for representing *temporary relationships*. A *time interval* is an abstract entity consisting in a contiguous segment of the time line, as it is conceived in classical physics, and it is an instance of the *TimeInterval* class. In order to enable the reasoning with time intervals, HERO-TIME puts at disposal (and formally characterizes) the thirteen *relations of Allen's interval algebra* expressing all the possible (and mutually exclusive) relationships between two intervals (by means of these relations, one can, for instance, specify that a time interval precedes, meets, overlaps, finishes, etc. another time interval) [1]. Given their importance in the human affairs, HERO-TIME provides a specific characterization for those time intervals that can be specified by referring to clock/calendar conventions (i.e., time of days, specific days, months, years, etc.). In particular, it provides a taxonomy of *clock/calendar interval expressions*, encompassing dates and time, month, year and century expressions. Each of these expressions *identifies* a time interval, which is an instance of the *CalendarClockInterval* class; examples are: "1944", "September 1943", "April 25 1945".

Moreover, HERO-TIME supports the representation of time intervals that do not correspond to instances of *CalendarClockInterval*, but can be characterized as time intervals delimited by two *CalendarClockInterval* instances using the *intBeginsIn* and *intEndsIn* properties; examples are: "between December 1044 and January 1945", "from April 25 and May 2 1945", "between 22 pm and midnight of April 25 1945". HERO-TIME also provides support for those vague time interval expressions that we may find in documents talking about human affairs, such as "at about 10 am of January, 4 1944", "the evening of 2 May 1945", "at the end of April 1945", etc. These "fuzzy" time intervals are ordered with respect to calendar-based intervals by means of the above mentioned Allen's relations. For example, the time interval *ti*, denoted by the expression "at the end of April 1945" can be linked to the calendar/clock interval (instance of the class *Month*) *a45*, denoted by "April 1945" by means of the relation *intFinishes*,

which corresponds to the *finishes* relation in the Allen's Interval Algebra (and states that *ti* starts somewhere within *a45* end ends where *a45* ends). Finally, HERO-TIME also enables the expression of *temporary relations* such as *temporary parthood* (e.g., since May 8 2012, the balanced budget amendment rule is part of the Italian Constitution), *temporary measurement expressions* (e.g., the Italian population in 1914 was around 36 million people), etc.

The HERO-900 suite is a *domain modular ontology* that refines HERO by specifying notions relevant to the history of the 20th century. It does not aim at being the ontology of the world of the 20th century, but only at being an extensible ontological seed for that historical period. It currently covers the two fragments of the Italian history relevant to the "Resistenza" in Piemonte (1943–1945) and to the protests of 1968 in Italy that we have considered in the case studies (see Section 3.1). HERO-900 is composed of three modules: HERO-EVENT-900, HERO-PLACE-900 and HERO-ROCS-900.

HERO-EVENT-900 refines HERO-EVENT mainly by introducing specific kinds of events, states and thematic roles. It provides, for instance, taxonomies of *confrontational actions* (e.g., *military occupation*, *killing*, *massacre*, *imprisonment*, *sabotage*, *police charge*, *street clash*, etc.), of *protest actions* (e.g., *protest march*, *strike*, etc.), and so on. HERO-EVENT-900 also provides specific thematic roles for those kinds of events and states for which the general thematic roles provided by HERO-EVENT are not enough to accurately represent the actual participation ways [22]. For instance, in *elections*, we may have (at least) two different kinds of *agents*: *candidates* and *voters*.

HERO-PLACE-900 refines HERO-PLACE by introducing specific kinds of places and roles for places. It introduces, for instance, several kinds of places that we can find in populated areas (e.g., *building*, *street*, *bridge*, etc.), as well as some roles representing their designated use (e.g., *public building*, *residential building*, etc.).

HERO-ROCS-900 refines HERO-ROCS mainly by introducing specific kinds of roles, organizations and collective entities. It provides, in particular, the notions of *profession* (such as laborer, office worker, public sector worker, etc.), *public office* (encompassing *Head of State*, *parliamentarian*, etc.), political identity (such as communist, fascist,

etc.), *role in a political party*, etc. It also puts at disposal several organization taxonomies in different areas (e.g., the areas of *companies*, *trade unions*, *political parties*, *police forces*, *military organizations*, *public bodies*, etc.). Specific kinds of collective entities are also provided (e.g., *social class*, *collective entity based on nationality* or *on profession*, etc.).

The HERO and HERO-900 ontologies are highly axiomatized, in this way enabling interesting inferences (e.g., all the axioms relevant to the Allen's interval algebra that are expressible in OWL 2 have been introduced in HERO-TIME; some axioms in HERO-EVENT refine the characteristics of thematic roles for each type of event, stating for instance, that a riot can be performed by physical persons, or collectives of physical persons or organizations or sets of entities of these kinds, etc.). This is a first reason why HERO and HERO-900 can be considered semantically richer that the ontologies mentioned in Section 2.2, which offer lighter axiomatizations. Moreover, they provide a homogeneous and coherent framework for notions – both generic (such as those of collective entity, organization, participation in perdurants as entity of a certain kind or playing a certain role) and specific (such as those of protest action, voters, etc.) – that are not present in the ontologies mentioned in Section 2.2. However, for the sake of interoperability, whenever possible, mappings have been produced between the HERO and HERO-900 ontologies and the above-mentioned ones (stating, for instance, that there is a close match between the notion of event in HERO and both the one with the same name in Europeana Data Model [15] and the *E4_Period* in CIDOC CRM [13]).

3.3 Evaluation of HERO-TIME in Harlock900

Historical time is often represented in graphical form using *timelines*, i.e., geometrical mapping from time to space; when employed with uniform timescales timelines can provide insight on the relative duration of different events, and if used to display attributes other than event time span, they can ease the comparison between time and other dimensions (e.g., spatial location of events). Although this kind of representations

can be effective from a pedagogical and exploratory standpoint, they present also substantial drawbacks:

- A timeline using a uniform timescale can grow very quickly in dimensions when the number of events or the covered time period increase. This would not be a problem per se, because the timeline can adopt various zoom & pan solutions to only show a sensible portion at a time, but this solution is detrimental to the very advantages granted by a timeline based representation. As stated very clearly by Meirelles [31]: *sectioning the charts obliterates their significance, because the purpose is to provide historic context to the topics depicted. The charts make sense only when viewed as a whole: It is the broad view that communicates the historic content and context* [31, p. 92].
- When temporal data have different granularities and the time interval an event occurs in is not reported in a precise way – as it often happens in historical texts – events do not have comparable durations and thus a simple geometrical mapping can be misleading.
- To represent time intervals specified in a vague way (e.g., "the end of April"), a timeline-based UI forces the mapping of the interval boundaries into specific arbitrary timeline points, depending on the chosen granularity (e.g., when does the end of April starts?).

In order to provide users with a more effective temporal-based access to events narrated in historical documents, we developed a prototype User Interface (UI) focusing on the exploration of the historical events described in the Semantic KB using their temporal properties [4]. In particular, the UI enables the user to ask the system for events occurred in a given time span, specified by a start and end time intervals, expressed at the granularity decided by the user herself (from hours to years); see Figure 2.

As described in Section 3.2, time intervals expressed in terms of HERO can be both standard calendar-based intervals, corresponding to *CalendarClockInterval* instances (e.g., days denoted by expressions like "April 25 1945") and intervals – corresponding to generic *ProperTimeInterval* instances – defined with respect to calendar-based

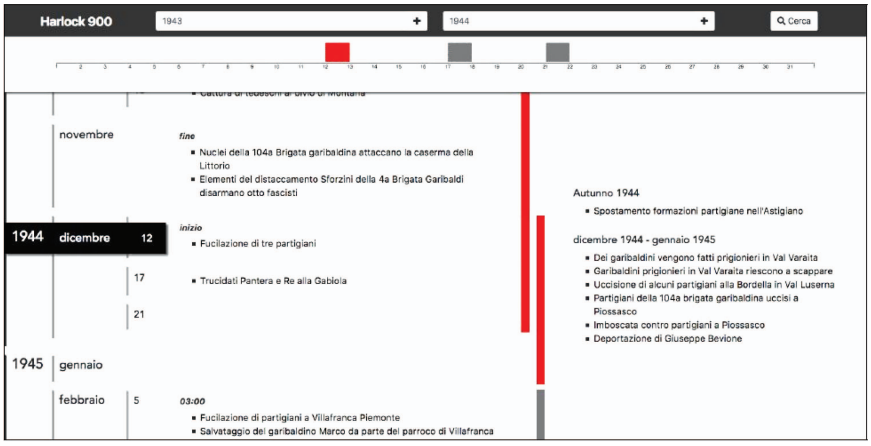


Figure 2 A fragment of the results page for the query asking for events occurred in the period 1943–1945 (source: [4]).

intervals through a HERO property (e.g., "between December 1944 and January 1945", "the end of April 1945"). In order to allow an efficient retrieval based on temporal range queries, including the latter type of time intervals, we developed a Query Support Module that builds and manages an indexing data structure ordering (also) intervals that cannot be placed on a calendar reference system by stating their relation with calendar-based intervals. As we will see, thanks to the Query Support Module, the result of a temporal range query will contain both calendar-based intervals and intervals specified using HERO relations. See [4] for more details the way the indexing data structure works.

Moreover, besides intervals completely falling within the range, results should include also intervals only partially overlapping it. For example, if the user selects a range spanning from April 15 and May 15 1945, the time interval denoted by an expression such as "the spring of 1945" has a partial overlap with it, and it is possible (although not sure) that an event occurred in such a time interval have occurred during the time range selected by the user. The Query Support module can retrieve these events and thus they can be included in the results.

Let's imagine Leo, a young researcher in Historical Studies who is preparing a PhD Thesis on the "Resistenza" in Piemonte. For his research, he is looking for documents related to events occurred

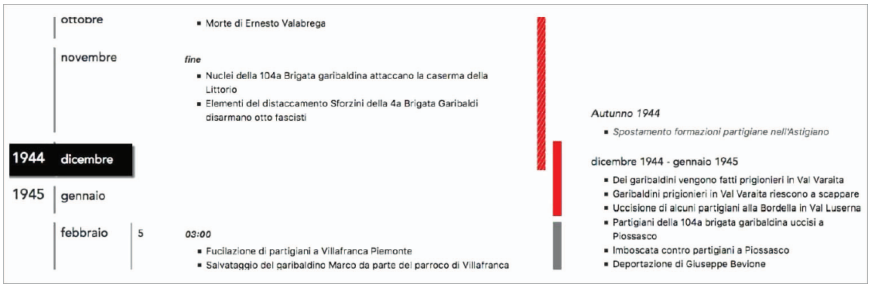


Figure 3 A fragment of the results page for the query asking for events occurred between October 1944 and January 1946.

between October 1944 and January 1946. He accesses the UI and selects October 1944 as start interval and January 1946 as end (he is not forced to select 1/10/44 and 31/1/46, because he can choose the granularity he prefers, months and not days in this case). The results are shown in Figure 3.

In the central area of the page Leo can find the chronological list of events occurred in calendar-based time intervals that fall within the query range (Ernesto Valabrega died in October 1944; at the end of November 1944 some Partisans from the 104a " Garibaldi" Brigade attacked the Littorio barracks; at 3 o'clock in the morning of February 5 1945 some Partisans have been shot in Villafranca Piemonte; and so on); the corresponding time intervals are shown on the left. By vertically scrolling the page, Leo can navigate through the list of events and when a calendar element crosses the middle of the viewport, it is highlighted (e.g., December 1944 in the figure).

On the right-hand side of the page, Leo can see a graphical representation of time intervals that do not directly refer to a calendar entity. These intervals are displayed as bars; the top and bottom edges are aligned with the start and end calendar dates delimiting them, when available, while they are shaded in case of "fuzzy" boundaries. When a bar intersects the middle of the viewport, the events occurred in that time interval are shown on the right (between December 1944 and January 1945 some Partisans are imprisoned in Val Varaita; in the same place some "Garibaldini" managed to escape; others are killed in Val Luserna and in Piossasco; etc.).

Moreover, some of the events displayed on the right occurred in a time interval only partially overlapping the query interval (in the example, during the Autumn 1944 some groups of Partisans moved to the Asti area): the time interval bar has a striped texture and the text describing the event is in italic.

We can conclude that the UI based on Harlock'900 semantic layer, compared to a timeline-based UI, allows:

- To correctly retrieve events occurred in time intervals denoted by expressions that do not directly refer to a calendar entity and by "fuzzy" expressions, with boundaries that cannot be precisely mapped into calendar dates (events in plain text on the right in Figure 3).
- To correctly retrieve events occurred in time intervals only partially overlapping the query interval (events in italic on the right in Figure 3).
- To link events to the time intervals mentioned in texts, without forcing choices for calendar-based definitions: if a text says that an event occurred at the end of April 1945, nobody is forced to place it in a time interval arbitrary starting in a precise day in April.

3.4 Evaluation of HERO-EVENT and HERO-EVENT-900 in PRiSMHA

The most common way to search a document repository, besides browsing categories, is keyword-based search. A typical search engine looks for the keywords specified by the user in the text of the indexed documents, or within the tags associated with each document. In turn, tags can be provided by the authors of the web site (e.g., this is often the case in systems offering an online access to archives, where tags are usually provided by the archivists themselves), or by users, through different types of crowdsourcing systems. In any case, tags – although possibly organized into *folksonomies* – are typically free texts labels, without any link to formal semantic structures. Despite advanced search support (that enables more complex queries, including – for instance – logical operators), the effectiveness of keyword-based search is based on the presence of the keywords in the engine index. The obvious main

Figure 4 A User Interface to express concept-based queries [source: [10]].

drawback of this approach is that it does not support a search based on *concepts* instead of keywords: if a user is looking for documents talking about protest actions, she has to query the engine for terms like "protest action", maybe coupled with keywords referring to more specific types of actions that could be mentioned (demonstration, march, strike, picket, sit-in, etc.), and with the verbs that could refer to a protest action.

In PRiSMHA, we designed a prototype UI enabling users to specify concept-based queries to access historical documents (see Figure 4). Such a UI is driven by the underlying ontologies (HERO and HERO-900) and enables users to select concepts corresponding to classes and relations defined in the ontologies. The query is then matched onto the Semantic KB (RDF triplestore) in order to retrieve the documents that refer to the concepts in the query. In the following we will see a concrete example, showing how such an approach can provide better results in terms of both recall (it retrieves documents that would not be retrieved by a keyword-based system) and precision (it does not

include documents that would be included by a keyword-based system, although irrelevant).[1]

Let's imagine Maria, a high-school teacher who is preparing a class about fights involving students during 1968. She would like to show her students some original documents narrating such events. She accesses the UI to search for archival documents referring to violent actions carried out by the police (or by other law enforcement agencies) against students during 1968. In a keyword-based system, she could write an advanced search like: *(caric* OR scontr*) AND (polizia OR poliziott* OR carabinier*) AND student**.[2] In the PRiSMHA UI, she is invited to select an event typology: Maria is supported by (a) auto-completion, suggesting her class names containing the string she starts writing; (b) a small button labeled "i" (for "information") that she can click to get a call-out explaining, in non-specialized language, the intended meaning of an ontology class or property. On the basis of the available explanation, in order to generalize the query, Maria selects *ConfrontationalAction* (which is a HERO-900 class with more specific sub-classes like *Beating*, *Firing*, *StreetClash*, *PoliceCharge*, etc.). On the basis of the characterization of such a class in HERO-900, the system invites her to select the (thematic) roles played by participants.

Firstly, Maria selects the *Agent* role: the underlying ontology constrains the types of entities that can play this role to: a person, an organization, a collective, a set; moreover, all these types of participants can participate in the event *as* entities playing certain roles (see Section 3.2). Guided by the system – again, in order to be general enough – Maria states that the *Agent* can be: a *person participating as lawEnforcementAgent* (i.e., a social role defined in HERO-900 subsuming roles such as *policeman* or "*carabiniere*"); an organization,

[1]The case study reported in this section is a walkthrough using real data, but it is not a quantitative test, therefore, here we use the notions of precision and recall in their "qualitative" meaning.

[2]*(charge OR clash) AND (police OR policemen OR carabinieri) AND student*; the star is used to capture both singular (e.g., *carica*) and the plural (e.g., *cariche*) of Italian nouns. In order to build a plausible query, we asked 10 users to write the queries they would have tried in order to satisfy the given information need. The case study can be easily modified taking into account slightly different queries.

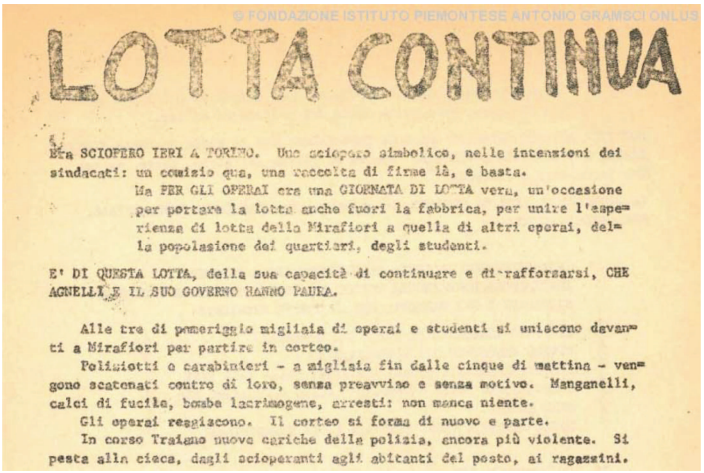
namely an instance of *LawEnforcementAgency* (which is a HERO-900 class, sub-class of HERO *Organization*); a set (e.g., "ten policemen") or a collective (e.g., "the policemen"), i.e., an instance of the class *Set* or *Collective*, *described by* the role *lawEnforcementAgent* (see Section 3.2). Secondly, Maria selects the *Patient* role: the underlying ontology constrains the types of entities that can play this role to the same as Agent, i.e., a person, an organization, a collective, or a set, all possibly participating in the event *as* entities playing certain roles. Guided by the system, Maria states that the *Patient* can be: a *person participating as student* (i.e., a social role defined in HERO-900); a set (e.g., "three students") or a collective (e.g., "the students"), i.e., an instance of the class *Set* or *Collective*, *described by* the role *student*. She can also select a time interval (*1968*).

Given the document repository represented by the 50 annotated original documents about 1968–69 movements in Italy from Ist. Gramsci's archives, mentioned in Section 3.1, the PRiSMHA system retrieves a set of documents linked to the semantic representations – stored in the Semantic KB – that match the concept-based query. In this specific case, this set includes:

- Five documents that contain the keywords and thus would probably be retrieved also by a keyword-based system; PRiSMHA is able to retrieve them because they are linked to semantic representations matching the concept-based query.
- Four documents that do not contain the keywords, but are linked to semantic representations matching the concept-based query; a keyword-based system would not retrieve these documents (see example below; Figure 6).
- Three documents that are stored as images (no text is available; see Figure 5); these documents are linked to semantic representations matching the concept-based query, but can hardly be retrieved by a keyword-based system.

These results demonstrate a (potential) higher recall (with respect to a generic keyword-based system).

Moreover, the PRiSMHA system does not retrieve documents where the keywords occurs with different meanings (e.g., documents talking about political responsibilities of the Police and the



LOTTA CONTINUA

Era SCIOPERO IERI A TORINO. Uno sciopero simbolico, nelle intenzioni dei sindacati: un comizio qua, una raccolta di firme là, e basta.
Ma PER GLI OPERAI era una GIORNATA DI LOTTA vera, un'occasione per portare la lotta anche fuori la fabbrica, per unire l'esperienza di lotta della Mirafiori a quella di altri operai, della popolazione dei quartieri, degli studenti.

E' DI QUESTA LOTTA, della sua capacità di continuare e di rafforzarsi, CHE AGNELLI E IL SUO GOVERNO HANNO PAURA.

Alle tre di pomeriggio migliaia di operai e studenti si uniscono davanti a Mirafiori per partire in corteo.
Poliziotti e carabinieri - a migliaia fin dalle cinque di mattina - vengono scatenati contro di loro, senza preavviso e senza motivo. Manganelli, calci di fucile, bombe lacrimogene, arresti: non manca niente.
Gli operai reagiscono. Il corteo si forma di nuovo e parte.
In corso Traiano nuove cariche della polizia, ancora più violente. Si pesta alla cieca, dagli scioperanti agli abitanti del posto, ai ragazzini.

Figure 5 Example of a document from the archives of the Ist. Gramsci, stored as an image (no text available) [copyright: Fondazione Istituto piemontese Antonio Gramsci].

Government; in Italian: "... responsabilità politiche della Polizia e delle più alte cariche dello Stato"). This demonstrate a (potential) higher precision.

In order to provide a more concrete insight, Figure 6 shows a simplified graphical representation of the description of an event narrated in one of the retrieved documents (in particular, in one of the documents retrieved by PRiSMHA, but not by the keyword-based system).

PRiSMHA knows that the document "talks about" (*isAbout* relation) an event (*studenti aggrediti dai carabinieri – students attacked by "Carabinieri" policemen*); in particular, this event is an instance of *PoliceCharge*, which is a subclass of *ConfrontationAction*. PRiSMHA also knows that it happened on November 20 1968 (*hasTime* relation) and that different actors took part in the action, with different roles, namely:

- Carabinieri, i.e., a *LawEnforcementAgency*, subclass of *Organization*, in the role of *Agent* (*hasAgent* relation);
- a group of students (*studenti*), playing the role of *Patient* (*hasPatient* relation); in particular such a group is represented as an instance of the class *Set*, and it is "described by" the role *student* (*hasDescribingConcept* relation);

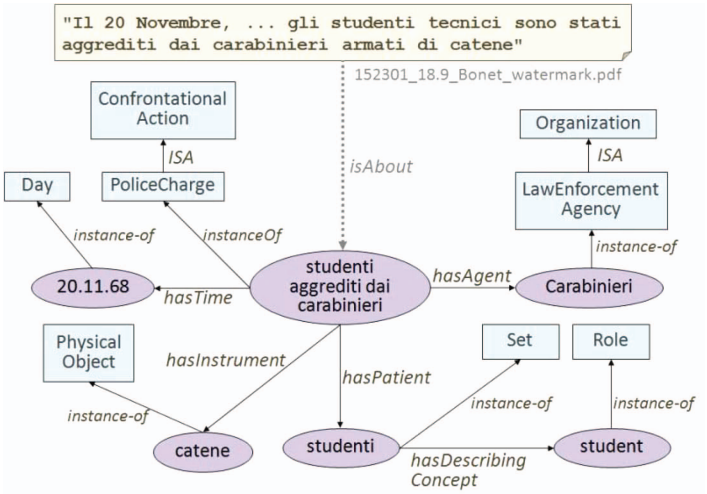


Figure 6 Simplified graphical representation of the description of an event narrated in a leaflet from the *Luciano Bonet* archive collection.

- some chains (*catene*), in the role of Instrument (*hasInstrument* relation).

When Maria queries confrontational actions where law enforcement agencies act as agents and students act as patient (i.e., they undergo the effects of the action), PRiSMHA searches the Semantic KB and finds the description of the event *studenti aggrediti dai carabinieri* (*students attacked by "Carabinieri" policemen*), shown in Figure 6. As a consequence, it can retrieve the document talking about such an event; since this document does not contain the keywords *caric* OR scontr**, a keyword-based search engine would not be able to retrieve it. Similar simulations of how PRiSMHA works can be done for the other documents that do not contain the keywords, but are linked to semantic representations matching the concept-based query.

4 Conclusions and Future Directions

This paper presented the core idea underlying two research projects (Harlock'900 and PRiSMHA): the answers of a system providing access to historical resources can be significantly enhanced by enabling

users to specify (complex) queries in terms of concepts. This core functionality is supported by a semantic layer, represented by a Semantic KB (RDF triplestore) containing formal descriptions of the content of the archival documents. Both the Semantic KB and the UI enabling concept-based queries are based on the underlying ontologies, which have been briefly described. Moreover, two case studies – the enhancement provided by a semantically rich representation of time intervals and by a detailed formal description of events and their participants, respectively – showed that this approach provides better results if compared with standard access systems (timelines and keyword search).

The research indicates several possible future directions for our research activity on these topics:

- The ontology-driven UI presented in this paper has been built with a user-centered approach, by involving user (through focus groups), since the first design phase. However, a usability evaluation [29] of the current prototype could provide important suggestions for its improvement.
- The domain ontology HERO-900 will evolve, to cover new parts of the domain (the Italian history of the 20th century). With respect to this aspect, the issue of defining a process to manage *ontology evolution* should be investigated [35].
- Finally, with a robust implementation of the crowdsourcing platform, the Semantic KB will be populated with a significant amount of data, that, in turn, will enable to couple the walkthrough evaluation provided in this paper with quantitative results.
- The knowledge stored in the Semantic KB could be made available through a RESTful service (more "friendly" and more efficient than a SPARQL endpoint) in order to enable third parties to build innovative applications that exploit the semantic representation of documents content, such as automatic constructions of document-supported narratives, biographical trajectories through events, etc. This possibility would improve the visibility of cultural organizations holding historical archives and would promote the engagement of a larger public.

References

[1] Allen, J. F., Maintaining knowledge about temporal intervals, *Communications of the ACM*, 26(11), 832–843, 1983.

[2] van den Akker, C., Aroyo, L., Cybulska, A., van Erp, M., Gorgels, P., Hollink, L., Jager, C., Legêne, S., van der Meij, L., Oomen, J., van Ossenbruggen, J., Schreiber, G., Segers, R., Vossen, P., Wielinga, B., Historical Event-based Access to Museum Collections, *Applied Artificial Intelligence*, 25, 2010.

[3] Ashenfelder M., *Cultural Institutions Embrace Crowdsourcing*, September 16, 2015 (blogs.loc.gov/digitalpreservation/2015/09/cultural-institutions-embrace-crowdsourcing).

[4] Baldo, A., Goy, A., Magro, D., A Pipeline Supporting a Smart Access to Historical Documents based on a Rich Semantic Representation of Their Content: A Case Study on Time Expressions, *Proc. WEBIST'18*. INSTICC SciTePress, 199–206, 2018.

[5] 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, *Journal of Web Semantics*, 35(3), 152–158, 2015.

[6] Borgo, S., Masolo, C., Foundational Choices in DOLCE, in S. Staab and R. Studer (Eds.), *Handbook on Ontologies*, Second Edition (pp. 361–381), Springer, 2009.

[7] Boschetti, F., Cimino, A., Dell'Orletta, F., Lebani, G. E., Passaro, L., Picchi, P., Venturi, G., Montemagni, S., Lenci, A., Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II, *Proc. LREC 2014 Workshop on Language resources and technologies for processing and linking historical documents and archives – Deploying Linked Open Data in Cultural Heritage*, 2014.

[8] Bottazzi, E., Catenacci, C., Gangemi, A. and Lehmann, J., From Collective Intentionality to Intentional Collectives: an Ontological Perspective, *Cognitive Systems Research – Special Issue on Cognition Joint Action and Collective Intentionality*, 7(2–3), 192–208, 2006.

[9] Bottazzi, E., Ferrario, R., Preliminaries to a DOLCE Ontology of Organizations, *Int. Journal of Business Process Integration and Management*, 4(4), 225–238, 2009.

[10] Carretta L., *Comunicare l'innovazione negli archivi storici attraverso lo User Experience Design: la progettazione di un mockup per il progetto PRiSMHA*, Tesi di laurea Magistrale, Università di Torino, 2019.

[11] Caserio, M., Goy, A., Magro, D., Smart access to historical archives based on rich semantic metadata, *Proc. IC3K – KMIS'17*. INSTICC SciTePress, 93–100, 2017.

[12] Cybulska, A., Vossen, P., Historical Event Extraction from Text, *Proc. LaTeCH'11*, 39–43, 2011.

[13] Doerr, M., The CIDOC Conceptual Reference Model: An Ontological Approach to Semantic Interoperability of Meta data, *AI Magazine*, 24(3), 75–92, 2003.

[14] Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F., Diachronic evaluation of NER systems on old newspapers, *Proc. KONVENS'16*, 97–107, 2016.

[15] Europeana, *EDM Definition of the Europeana DataModel v.5.2.7*, 2016 (http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.7_042016.pdf).

[16] Galton, A., Wood, Z., Extensional and intensional collectives and the de re/de dicto distinction, *Applied Ontology*, 11(3), 205–226, 2016.

[17] Gangemi, A., Mika, P., Understanding the SemanticWeb through Descriptions and Situations, in Meersman R., Tari Z., Schmidt D.C. (eds), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (OTM 2003), LNCS 2888, Springer, 689–706, 2003.

[18] Goy, A., Magro, D., Rovera, M., Ontologies and historical archives: A way to tell new stories. *Applied Ontology*, 10(3–4), 331–338, 2015.

[19] Goy, A., Damiano, R., Loreto, F., Magro, D., Musso, S., Radicioni, D., Accornero, C., Colla, D., Lieto, A., Mensa, E., Rovera, M., Astrologo, D., Boniolo, B., D'Ambrosio, M., PRiSMHA

(Providing Rich Semantic Metadata for Historical Archives), *Proc. Contextual Representation of Objects and Events in Language* (CREOL 2017), 2017.

[20] Goy, A., Magro, D., Rovera, M., An ontological perspective on thematic roles, in P. Ciancarini, F. Poggi, M. Horridge, J. Zhao, T. Groza, M.C. Suarez-Figueroa, M. d'Aquin, V. Presutti (eds), *Knowledge Engineering and Knowledge Management*, LNAI 10180, Springer, 123–126, 2017.

[21] Goy, A., Magro, D., Conforti, F., Exploring RDF Datasets with LDscout, *Proc. IC3K – KMIS'18*. INSTICC SciTePress, 92–100, 2018.

[22] Goy, A., Magro, D., Rovera, M., On the Role of Thematic Roles in a Historical Event Ontology, *Applied Ontology*,13, 19–39, 2018.

[23] Guizzardi, G., Ontological Foundations for Conceptual Part-Whole Relations: The Case of Collectives and their Parts, *Proc. 23th Int. Conf. on Advanced Information System Engineering* (CAiSE 2011), 2011.

[24] van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G., Desing and use of the Simple Event Model (SEM), *Journal of Web Semantics*, 9(2),128–136, 2011.

[25] Haslhofe, B., Isaac, A. data.europeana.eu – The Europeana Linked Open Data Pilot, *Proc. Int. Conf. on Dublin Core and Metadata Applications*, 2011.

[26] Heath, T. and Bizer, C., *Linked Data: Evolving the Web into a Global Data Space*, Morgan and Claypool, 2011.

[27] Hogenboom F., Frasincar F., Kaymak U., de Jong F., An Overview of Event Extraction from Text, *Proc. DeRiVE'11 at ISWC 2011, Vol. 779*, 2011.

[28] Isaac A. (Ed.) *Europeana Data Model Primer*, Creative Commons Licence, 2013.

[29] Krug, S., *Web Usability: Rocket Surgery Made Easy*, Addison-Wesley, 2010.

[30] Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., Guarino, N., Social Roles and Their Descriptions, *Proc. KR2004*, AAAI Press, CA, 267–277, 2004.

[31] Meirelles, I., *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*, Rockport Publishers, 2013.

[32] Meroño-Peñuela, A. Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F., Semantic Technologies for Historical Research: A Survey, *Semantic Web Journal*, 6(6), 539–564, 2015.

[33] Moretti, G., Sprugnoli, R., Menini, S., Tonelli, S., ALCIDE: Extracting and visualising content from large document collections to support humanities studies, *Knowledge-Based Systems*, 111, 100–112, 2016.

[34] Oomen, J., Belice, L., Sharing cultural heritage the linked open data way: why you should sign up, *Proc. Museums and the Web Conference,* 2012.

[35] Rahnama A. and Abdollazadeh Barforoush A., A novel ontology evolution methodology, *Journal of Web Engineering*, Vol. 14, No. 3&4, 301–324, 2015.

[36] Rovera, M., Nanni, F., Ponzetto, S. P., Goy, A., Domain-specific Named Entity Disambiguation in Historical Memoirs, *Proc. CLiC-it'17, vol. 2006*. CEUR, 2017.

[37] Ruijgrok, P., Frasincar, F., VandicD., Hogenboom, F., OntoNavShop: An ontology-based approachfor web-shop navigation, *Journal of Web Engineering*, Vol. 17, No. 3&4, 241–269, 2018.

[38] Shaw, R., Troncy, R., Hardman, L., LODE: Linking Open Descriptions of Events, *Proc. 4th Asian Conference on The Semantic Web*, 153–167, 2009.

[39] Segers, R., van Erp, M., van der Meij, L., Hacking History via Event Extraction, *Proc. K-CAP'11*, 161–162, 2011.

[40] Sprugnoli, R., Tonelli, S., One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective, *Natural Language Engineering*, 23(4), 485–506, 2016.

[41] Strötgen, J., Gertz, M., Multilingual and cross-domain temporal tagging, *Language Resources and Evaluation*, 47(2), 269–298, 2013.

[42] Welty, C., Guarino, N., Supporting Ontological Analysis of Taxonomic Relationships, *Data and Knowledge Engineering*, 39, 51–74, 2001.
[43] Wood, Z., Galton, A., A taxonomy of collective phenomena. *Applied Ontology*, 4(3–4), 267–292, 2009.

Biographies

Annamaria Goy is a Researcher at the Computer Science Department of the Università di Torino (Italy), where she works in the area of web-based systems and semantic technologies. She obtained her Ph.D in Cognitive Science at the same university, with studies in the area of lexical semantics. She currently carries on her research and development activity applying knowledge representation, ontology modeling and human computer interaction approaches mainly to the Cultural Heritage and Digital Humanities areas. She teaches Web Technologies and Web Programming classes.

Diego Magro got the Master degree in Computer Science from the Università di Torino (Italy) in 1997. He currently works as a Researcher at the Computer Science Department of the Università di Torino.

His research interests are in the areas of Artificial Intelligence, Digital Humanities and Web of data. His current research activity is mainly focused on Knowledge Representation, Ontologies and Semantic Technologies. He teaches Programming, Algorithms, Databases and Ontology Modeling and Reasoning in undergraduate and postgraduate university courses.

Alessandro Baldo received his Bachelor and Master Degree in Computer Science at the Università di Torino (Italy). Since 2017 he works as a Software Architect at TomorrowData Srl, where he designs and maintains industrial IOT applications.