Bringing Semantics into Historical Archives with Computer-aided Rich Metadata Generation

Bringing Semantics into Historical Archives with Computer-aided Rich Metadata Generation

Davide Colla
Dipartimento di Informatica, Università di Torino, Torino, Italy, davide.colla@unito.it

Annamaria Goy
Dipartimento di Informatica, Università di Torino, Torino, Italy, annamaria.goy@unito.it

Marco Leontino
Dipartimento di Informatica, Università di Torino, Torino, Italy, marco.leontino@unito.it

Diego Magro
Dipartimento di Informatica, Università di Torino, Torino, Italy, diego.magro@unito.it

Claudia Picardi
Dipartimento di Informatica, Università di Torino, Torino, Italy, claudia.picardi@unito.it

ABSTRACT

This paper relies on the idea that a semantically rich metadata layer is required in order to provide an effective, intelligent and engaging access to historical archives. However, building such a semantic layer represents a well-known bottleneck that can be overcome only by a hybrid strategy, integrating user-generated content and automatic techniques. The PRiSMHA project provides a contribution in this direction with the design and development of the prototype of an ontology-driven platform supporting users in semantic metadata generation. In particular, the main contribution of this paper is to show how automatic information extraction techniques (namely, Named Entity and Temporal Expression Recognition), and information retrieved from external datasets in the LOD cloud can support users in the identification and characterization of new entities to annotate documents with.

KEYWORDS

Artificial intelligence and archives, semantic metadata generation, linked data, ontologies, entity extraction, synergies between computational and human-based methods, semantic processing

1 Introduction

In recent years, the awareness of the great value of stories and memories stored in historical archives has increased, and a lot of efforts all over the world have been devoted to the digital curation of archival resources (Lee and Tibbo, 2011; Post et al., 2019), including document digitization, metadata generation, and access tools development. Many initiatives, for example, try to use storytelling techniques to provide a large public with an immersive experience in the interaction with historical memories (see, for instance, projects like CrossCult: www.crosscult.eu, or meSch: www.mesch-project.eu, see also (Battad et al., 2019; Underberg-Goode, 2017; Lombardo et al., 2016; Damiano and Lombardo, 2016; among many others).

However, many archives host a huge amount of undigitized material; digital catalogs often consist of quite poor metadata, usually describing resources at fonds or series level, while the actual content of single documents is neglected. However, a detailed knowledge about what individual archival documents talk about is precisely what is needed to provide an effective, intelligent and engaging access to historical archives. The idea that a semantically rich metadata layer is required in order to enhance the access to archival resources is shared in the Digital Humanities community; see, for instance, (Motta et al., 2000).

The PRiSMHA (Providing Rich Semantic Metadata for Historical Archives) project (Goy et al., 2017) aims at providing a contribution in this direction by designing an ontology-driven platform that supports semantic metadata generation, needed to offer an effective access to archival documents. PRiSMHA is a three-year (2017-2020) Italian project, funded by Compagnia di San Paolo Foundation and Università di Torino. It involves the Computer Science and the Historical Studies Departments of the same university, and the Fondazione Istituto piemontese Antonio Gramsci (www.gramscitorino.it), member of the Polo del '900 (www.polodel900.it), a cultural institution headquartered in Torino, hosting a very rich archive (www.polodel900.it/9centro).

The main idea underlying the PRiSMHA approach is that only a crowdsourcing model, coupled with automatic techniques, can enable the (collaborative) construction of the semantic layer required to guarantee a content-based access to historical archives. In order to test and demonstrate the feasibility of this approach, the PRiSMHA team developed a proof-of-concept prototype (see Section 3), running on a small set of Istituto Gramsci's collections (a little bit more than 200 documents), related to the students' and workers' protest during the years 1968-1969 in Italy (Goy et al., 2019a); besides a few number of pictures and newspaper articles, such resources are mainly represented by typewritten leaflets, often with handwritten annotations and drawings (see Figures 1 and 2).

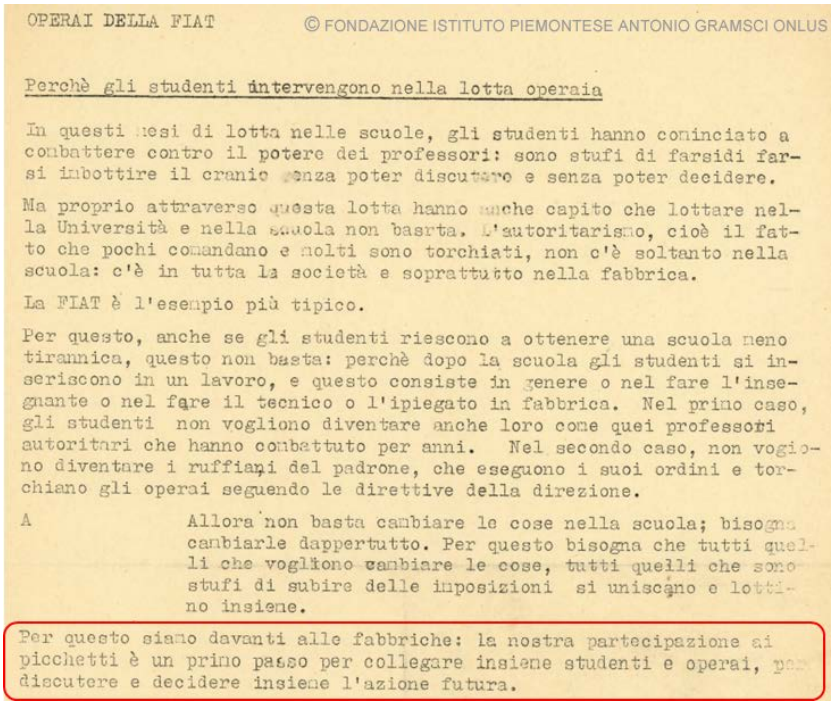

Figure 1: A leaflet about a strike at FIAT ***[copyright: Fondazione Istituto piemontese Antonio Gramsci Onlus]*****.**

Let us introduce a scenario with the goal of providing the reader with the motivations underlying the research in the PRiSMHA project. Consider Antonio, a school teacher who wants to enrich his lessons with information directly taken from original documents. He is talking to his students about protest actions that took place in Torino in 1968. In particular, he is searching a digital online archive system, looking for leaflets referring to strikes which both students and workers participated in. Even with a (very) good OCR tool, if the system is based on a keyword search mechanism, the results of a query for "sciopero" (strike) would not include the document shown in Figure 1: the document, in fact, does not contain the world "sciopero/i" (strike/es), although it actually talks about a strike, using the very specific word "picchetti" (picketings). Antonio also looks for leaflets mentioning specific people (e.g., Guido Viale, a leader of the '68 Movement in Torino) involved in protest actions. The results for a query for "Guido Viale" in a keyword-based system would probably include the leaflet in Figure 2. However, that leaflet would not be retrieved if the query also

contains (in AND) a keyword for the action (such as “sciopero”/strike, “manifestazione”/demonstration, and so on), since the protest action is not explicitly mentioned, although it is clear to a human reader that the document talks about a protest action.

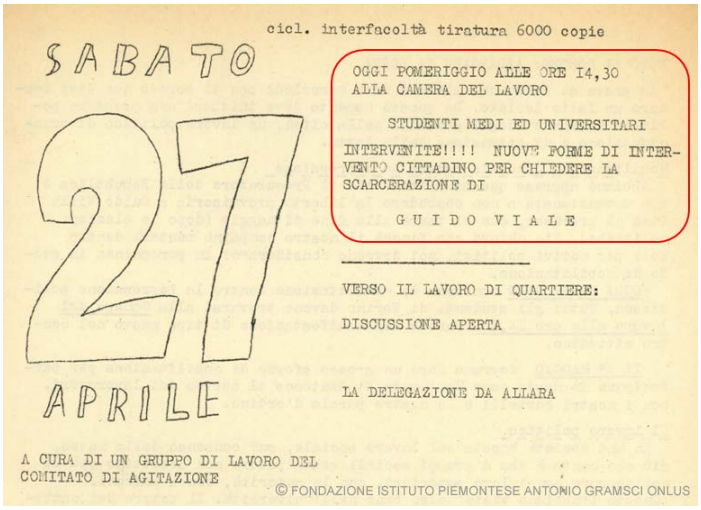

Figure 2: A leaflet mentioning Guido Viale *[copyright: Fondazione Istituto piemontese Antonio Gramsci Onlus]*.

Moreover, Guido Viale, although mentioned in the document, is not an active participant; the leaflet indeed says that he has been arrested and the demonstration is organized to ask for his release (he could be considered the “topic” of the protest action, or a participant in his release from prison, i.e., in the event representing the goal of the protest action).

If Antonio was searching for the active involvement of Guido Viale in protest actions, for example, he would need a tool enabling him to ask for all documents talking about any type of *protest action* which have a relation of type *active participation* with the *person* Guido Viale. A simplified sketch of the knowledge involved in such a query is shown in Figure 3, where rounded squares represent concepts (i.e., ontology classes), ovals represent individuals (class instances), and arrows represent relations (see Section 3).

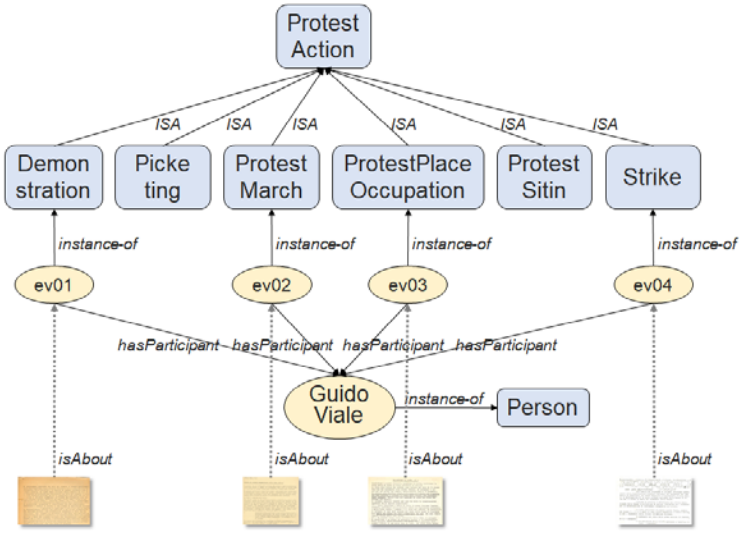

Figure 3: Simplified graphical semantic representation of participation of Guido Viale in protest actions, with links to archival documents.

In this paper we will present the results of the PRiSMHA project, which aims at providing Antonio with the tool he needs. More specifically, in order to both provide users with the possibility of posing such queries, and be able to answer them, the system needs a *semantic layer* over archival documents, containing a formal machine-readable representation of their content. The conceptual vocabulary of such a layer is represented by *computational ontologies*: the ontology used in PRiSMHA will be briefly described in Section 3, where the languages and tools used to implement the semantic layer are also presented.
A formal machine-readable semantic layer could also be exploited by third party applications to access rich semantic metadata. Applications, ranging from education-oriented games to citizen services and history-aware tourist guides, could exploit knowledge about places, people, and organizations involved in the narrated events, offered by the semantic layer. Moreover, these rich semantic metadata are linked to the archival resources, thereby providing applications with access to the original documents, and offering archival institutions a great opportunity to turn their heritage into a visible and live matter.
However, given the conceptual vocabulary (i.e., concepts and relations), who builds the potentially extremely huge semantic knowledge base containing the formal representation of the content of archival documents?
We believe that the bottleneck represented by the population of the semantic knowledge base of systems like PRiSMHA can only be overcome by implementing a hybrid strategy, that integrates user-generated content and automatic techniques (Foley et al., 2017). In particular, the approach described in this paper builds on the work presented in (Goy et al., 2020), where we described our solution to provide users with an ontology-driven user interface enabling them to build the semantic layer, by "annotating" archival documents with semantic representations of their content. In this paper we show how this activity can be supported by exploiting automatic information extraction techniques (namely, Named Entity Recognition), and linking to external resources (such as Wikidata).
Therefore, the **research question** of the work presented in this paper is the following:

Given an ontology-driven web-based system enabling users to build the formal semantic representations of archival document content, can automatic text mining techniques (such as Named Entity Recognition), and entity linking to external resources (Linked Open Data) provide users with an effective support in the annotation activity?

The main contribution of this paper is to provide an answer to this research question by describing the PRiSMHA approach and presenting the results of a user evaluation that supports it.
The rest of the paper is organized as follows. Section 2 discusses the most relevant related works; Section 3 introduces the PRiSMHA architecture with its main modules, briefly describes the underlying ontology and overviews the main steps of the interaction with the annotation platform. Section 4 is devoted to the presentation of the details about the automatic support provided by Information Extraction and linking to external datasets in the LOD cloud. Finally, Section 5 presents the results of a qualitative evaluation of the mentioned support, while Section 6 concludes the paper by sketching the main future work directions.

2 Related Work

Several research areas are related to the PRiSMHA project, namely, at least: ontology-based annotation, ontology-based data access, and partially ontology integration.
In the work presented in this paper, we also use NLP techniques to extract relevant entities from texts. However, since we rely on state-of-the-art tools, without the ambition to provide enhancements in the field, we do not review related work about NLP in this section, but we simply provide an overview in Section 4.1.

The closest research area that needs to be mentioned is ontology-based annotation.
Ontologies are conceptual representations of the world, and several ontologies have been developed for the Cultural Heritage domain. In this perspective, the contribution of an ontology is three-fold: (i) it represents a conceptual vocabulary providing rich semantic knowledge; (ii) it defines a schema that can be integrated and

interconnected with heterogeneous ontologies already developed; (iii) it provides a base to software tools whose aim is a semantic exploration of digital collections (Alma'aitah et al., 2020).
Several studies have been conducted on developing ontologies in order to annotate documents related to the Cultural Heritage domain, according to a semantic model. Laclavik and colleagues (Laclavik et al., 2006) developed a tool for ontology-based text annotation called OnTeA. The OnTeA tool processes text documents by employing regular expressions and detects equivalent semantics elements (i.e., elements that share the same meaning) according to the defined domain ontology. Lana and colleagues (Lana et al., 2014) built an ontology, in the framework of the Geolat (https://geolat.uniupo.it/) project, to make the Latin literature accessible. Garozzo and colleagues (Garrozzo et al., 2017) developed CulTO, a tool relying on an ontology for the Cultural Heritage domain. CulTO is specifically designed to support annotation of photographic data and text documents related to historical buildings. The developed ontology has been modeled at a high abstraction level, enabling the connection to other Cultural Heritage ontologies. Carboni and De Luca (Carboni and De Luca, 2018) developed the Visual and Iconographical Representations (VIR) ontology to record statements about the physical and conceptual nature of the heritage. The developed ontology enables the association of assertions with 2D/3D representations.
According to Adrews and colleagues (Andrews et al., 2012) the structural complexity of annotations influences the way in which data can be displayed and what kind of data can be displayed. The authors distinguish between four types of annotation: tags, attributes, relations and ontologies. Tags are keywords assigned to a resource; attributes are pairs in the form *<AN, AV>* where *AN* is the attribute name and *AV* is the attribute value; relations are pairs in the form *<Rel, Res>*, where *Rel* is the name of the relation and *Res* is another resource; finally, ontology annotation (or semantic annotation) refers both to the process and to the resulting annotation, consisting of aligning a resource to an ontology: The ontology annotation model enables annotators to associate a resource with a semantic rich tag relying on the vocabulary provided by an ontology. The framework proposed by Andrews and colleagues assumes that the ontology completely describes the domain of the examined resources: users are requested to provide a link from a resource to the ontology. For example, consider a document about a strike by FIAT employees. The aforementioned framework assumes that the ontology already contains the characterization of the FIAT company and its employees, and users are called to link the document with both entities.
This approach is the closest to PRiSMHA, although the PRiSMHA point of view on annotations and role of users is slightly different. In PRiSMHA, the contribution of users is two-fold: (i) populating a semantic knowledge base, where entities of different types (events, persons, organizations, places, etc.) are represented, and (ii) linking archival documents to entities in the knowledge base, describing their content. The role of the annotators is central, since the knowledge contained into documents has to be interpreted and expressed in a semantic rich format. The user herself (supported by the automatic tools described in Section 4) builds the semantic description of entities and events mentioned in the examined document by exploiting the web platform (see Section 3). Let us consider again the document about the strike by FIAT employees. According to the PRiSMHA model, an annotator should provide the description of the FIAT company (if it has no characterization in the knowledge base), and then to link the FIAT characterization to the examined document.
Cultural Heritage ontologies are often paired with a high-level tool for digital content exploration (Ghiselli et al., 2005; Schreiber et al., 2008; Garrozzo et al., 2017; Tommy et al., 2017). From the PRiSMHA perspective, ontologies play a key role in both equipping documents with semantic metadata and building a software for semantic exploration of digital collection. Employing an ontology as a descriptive schema for the content of historical documents is a challenging task, since the ontology layer is fundamental to provide documents with semantically rich annotations, but its complexity has to be invisible to the end user of the platform (Goy et al., 2020). However, this is not always possible: The limited knowledge of users about the semantic schema makes metadata difficult to be exploited for both enrichment and exploration (Elsweiler et al., 2011); see Section 5 for a brief discussion on this topic.

The second research area that needs to be mentioned is ontology-based data access. Dealing with Cultural Heritage collections often involves coping with domain-specific terminology. Users are not always aware of

such a vocabulary, and this prevents them from fully exploiting the potential of the system providing access to collections (Walsh and Hall, 2015).
Several energies have been invested into making metadata resources more accessible for users (Kollia et al., 2012; Tonkin and Tourte, 2016; Windhager et al., 2016). Such projects aim at making resources, already enriched with semantic metadata, more accessible and readable to users, while in PRiSMHA the issue of accessibility is mainly considered in the context of data acquisition, where the user has to deal with the annotation tool to equip documents with semantic metadata: The goal of the PRiSMHA project is to provide users with an ontology-based platform that supports them in the semantic annotation process, lowering formal complexity, while, at the same time, offering them the expressive power provided by the ontological definition of domain concepts.

In the last few decades, many efforts have also been spent integrating and connecting heterogeneous ontologies. The so-called mapping operation aims at building a generic and shared schema that plays its role as an interface between syntactically and semantically heterogeneous metadata (Alma'aitah et al., 2020). The CIDOC-CRM ontology is probably the best known project aimed at providing a mechanism to perform integration, interchanging and connection of heterogeneous sources of Cultural Heritage knowledge (Crofts et al., 2003): The developed tool is specifically designed to integrate information in the Cultural Heritage domain.
Other projects are worth mentioning. Doulaverakis and colleagues (Doulaverakis et al., 2005) developed the REACH project, aimed at defining an ontology- based representation in order to provide enhanced unified access to heterogeneous distributed Cultural Heritage digital databases, mainly focused on Greek and Roman Antiquity. The system includes: (i) a Cultural Heritage web portal for unified access to the information and services; (ii) a digitalization system for the efficient digitalization of artwork and collections; (iii) an ontology to describe and organize Cultural Heritage content; (iv) a multimedia content-based as well as ontological-based search engine; (v) an e-Commerce section for the commercial exploitation of the portal.
Hyvönen and colleagues (Hyvönen et al., 2005) have successfully carried out the MuseumFinland project, with the objective of building a single access-point to more than 15 museum collections. The software combines databases transforming them into a shared XML format, thus obtaining syntactic interoperability. Semantic interoperability is obtained by translating from XML to RDF, exploiting seven domain ontologies.
Daquino and colleagues (Daquino et al., 2016) developed two ad-hoc ontologies to describe Zeri Photo Archive catalogue. The developed ontologies were mapped to CIDOC-CRM, as well as to Historical Context Ontology (HiCO) (Daquino and Tomasi, 2015), Publishing Roles Ontology (PRO) (Peroni et al., 2012), and FRBR-aligned Bibliographic Ontology (FaBIO) (Peroni and Shotton, 2012).
Dragoni and colleagues (Dragoni et al., 2016) present a general architecture for knowledge management platforms, together with an implementation (MOKI-CH), in the Cultural Heritage domain. The authors identify a number of requirements for the presented architecture, among which the most interesting for the approach presented in this paper is *data exposure and data linking*, which represent a goal of the PRiSMHA project (as mentioned in Section 1 and explained in Section 3 and 4).
Finally, two initiatives should be mentioned: Data for History (dataforhistory.org), which offers a web-based platform supporting ontology development and alignment with CIDOC-CRM, and ArCo (wit.istc.cnr.it/arco), whose goal is to develop a knowledge graph for Italian Cultural Heritage by aligning top-level models and ontologies characterizing properties and actions for Cultural Heritage curation.

The majority of ontological models somehow related to the Cultural Heritage domain, including CIDOC-CRM along with the other mentioned ones, are mainly designed having curation in mind. This implies that they usually include a fine-grained characterization of cultural resource types, properties, and actions that can be taken with them. The goal of the ontology we developed within the PRiSMHA project is to model the concepts and relations representing the *content* of cultural resources. In this perspective, the major part of our ontology is not about Cultural Heritage, but it is about the *domain* Cultural Heritage "talks about". Only a very small part of our ontology, in fact, is used to model archival documents as such, fragments they are composed of, and the relation between fragments and semantic representations of their content.

3 The Architecture, the Ontology, and the Annotation Platform

Figure 4 shows the architecture of the PRiSMHA prototype. As already stated, the goal of the project was the design and development of the *Crowdsourcing Platform*, with its User Interface (*Crowdsourcing Platform UI*), which is driven by the *ontology* (*HERO* and *HERO-900*) and aims at populating the *Semantic KB* (implemented as a *RDF triplestore*). The *Semantic KB*, together with the *ontology*, represents the above mentioned *semantic layer*, which contains a formal machine-readable representation of the content of archival documents. As stated in Section 1, PRiSMHA implements a hybrid strategy, by integrating user-generated content and automatic techniques: user-generated content is provided through the *Crowdsourcing Platform UI*, while automatic techniques are represented by the *Information Extraction* (IE) module, which identifies relevant entities within textual documents, and the *LOD linking* module, which supports the connection of the *Semantic KB* with datasets in the LOD cloud.
The ontology-driven *Crowdsourcing Platform* and its UI, enabling users to annotate documents with semantic representations of their content, are described in detail in (Goy et al., 2020). In this paper, we focus on the role of the *IE* and *LOD linking* modules, which support users of the *Crowdsourcing Platform* in their activity. These modules are described in detail in Section 4.

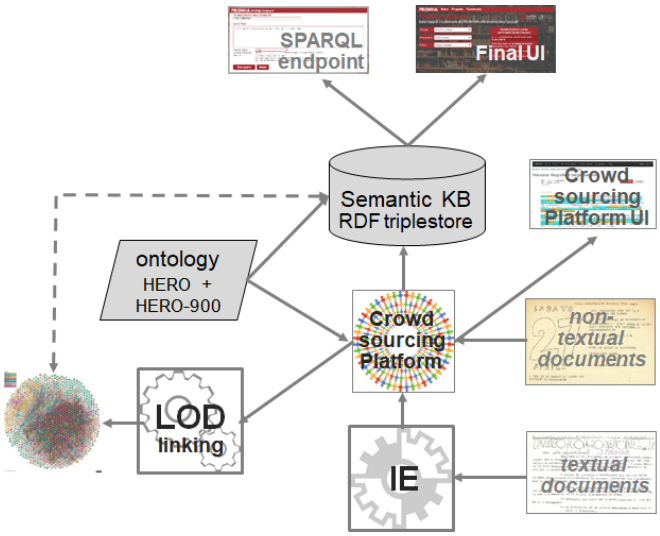

Figure 4: PRiSMHA prototype architecture.

In the rest of this section, we briefly describe the *HERO* and *HERO-900* ontologies and the structure of the *Semantic KB*; we then overview the most relevant aspects of the *Crowdsourcing Platform UI*, focusing on the work of a single user annotating a single document, being the collaborative aspects out of the scope of the present paper. This will lay the ground for the presentation of the additional features, represented by the *IE* and *LOD linking* functionalities, in Section 4 and 5.

HERO (Historical Event Representation Ontology) is a reference ontology that provides classes and properties useful to characterize historical events. In particular, HERO offers the conceptual vocabulary for specifying event types (e.g., a strike), the places events occur in (e.g., Milan), the date or time frame events take place in (e.g., November 1968), and -- most important for this paper -- the participants in the events, i.e., people, organizations, objects, etc. involved in events with different roles.
The upper-level module, HERO-TOP (https://w3id.org/hero/HERO-TOP), provides the most general concepts, inherited by all the other modules. It is grounded in the basic distinctions defined in DOLCE (Borgo and Masolo, 2009), i.e., *perdurants*, *objects*, and *abstract entities*. Perdurants can be *states* (e.g., being wounded) or *events* (e.g., a strike). Among objects, HERO distinguishes between *physical objects* (e.g., persons, buildings) and *non-physical objects*, among which *social objects* play a major role in the historical domain; examples of social objects are *organizations* (e.g., trade union) and *social roles* (e.g., student).

One of the most relevant relationships between *objects* and *perdurants* is *participation* (e.g., Sergio Garavini, a *person* -- thus an *object* -- participated in a strike, an *event* -- thus a *perdurant*).
The HERO-EVENT module (https://w3id.org/hero/HERO-EVENT) accounts for the mentioned distinction between *states* and *events*, and, among events, between *actions* (i.e., intentional events) and *phenomena* (i.e., non-intentional events). Moreover, this module offers properties that are useful for describing states and events, in particular *thematic* (or *semantic*) *roles*, expressing the modalities in which objects (e.g., persons, organizations) participate in events or states (e.g., agent, patient, instrument); see (Goy et al., 2018). HERO-PLACE (https://w3id.org/hero/HERO-PLACE) defines concepts and properties relevant for the characterization of places, while HERO-TIME (https://w3id.org/hero/HERO-TIME) provides the notions for expressing time.
With respect to the work presented in this paper, the most important module is HERO-ROCS (https://w3id.org/hero/HERO-ROCS), which defines the formal instruments for describing *organizations* (e.g., political parties, companies), *collective entities* (e.g., students, workers), and *social roles* (like professions, for example).
HERO-EVENT-900, HERO-PLACE-900, and HERO-ROCS-900 are domain modules refining the corresponding HERO modules by introducing concepts and properties useful to describe the history of the 20th century. The current versions of these modules cover the concepts and properties needed to describe the historical events -- and the involved entities -- considered in the PRiSMHA project, i.e., the students' and workers' protests during the years 1968-1969 in Italy. In particular, HERO-ROCS-900 offers a set of specific organization types (e.g., various types of trade unions, various types of organizations in the political sphere), a set of specific collective entity types (e.g., social classes, political-based collective entities), and a set of specific role types (e.g., various types of workers, various types of students).
Within the PRiSMHA project, we developed an application version of HERO, encoded in OWL 2 DL (www.w3.org/OWL), containing 429 classes, 378 properties, 79 individuals and nearly 4,500 logical axioms. This is the ontology that underlies the *Crowdsourcing Platform UI*, sketched below.
The *Semantic KB* is implemented as a RDF triplestore (www.w3.org/RDF), containing RDF triples of the form <*s*, *p*, *o*>, where *s* is an entity in the Semantic KB, *p* is a property (defined in HERO/HERO-900, or belonging to RDF itself -- e.g., *rdf:type*) and o can be either an entity in the Semantic KB, a literal (e.g., a string, a number), or a class defined in HERO/HERO-900 (e.g., *Organization*). Each triple represents an assertion stating that the entity *s* has the value *o* for the property *p*. Entities, properties, and classes are represented in the triplestore by URIs. Data stored in the Semantic KB can be accessed through a *SPARQL endpoint* -- thus making the Semantic KB available to third party applications -- or navigated through the *Final User Interface* (see Figure 5), which is currently available as a mockup (the description of which is out of the scope of the present paper).
The *Crowdsourcing Platform* prototype is a web application accessible through a browser. Its implementation is based on the AJAX model and exploits JQuery 3.3.1 (jquery.com) and Bootstrap 3.3.7 (getbootstrap.com/docs/3.3/). The implementation of the *Crowdsourcing Platform* application logic relies on the Spring Boot 1.5.10 framework (spring.io/projects/spring-boot), while data is stored in a MySQL 5.6.38 (www.mysql.com) relational database. The OWLAPI 5.1.0 (owlcs.github.io/owlapi) library supports the interaction with the ontology, and an RDF triplestore, implemented by means of Jena TDB 3.6.0 (jena.apache.org/documentation/tdb), stores the semantic representations.

Before entering the description of the User Interface of the *Crowdsourcing Platform*, a few words should be devoted to the prospective users of such a platform. By means of informal interviews with users and employees of the library and archives of the Polo del '900, we identified the potential users of the PRiSMHA *Crowdsourcing Platform*: such users can be historians, archivists, students, researchers or simply enthusiasts and people interested in the history of the 20th century, participating in the PRiSMHA community with the role of experts, or simply trusted users, motivated in spending time and effort in the semantic annotation process. Despite the efforts to reach a good level of usability -- see (Goy et al., 2020) -- the interaction with the *Crowdsourcing Platform UI* remains a challenging task, that requires some learning and training, as well as some knowledge about the domain (basically, the Italian history of the 20th century).

Figure 5: A document on the PRiSMHA Crowdsourcing Platform (the biography of Emilio Pugno).

Figure 5 shows a textual document, the biography of Emilio Pugno (an Italian trade union leader), accessed through the *Crowdsourcing Platform*. Users enabled to work on this document can identify textual units that can be annotated, called *fragments* (highlighted in cyan). By clicking on a fragment, users can see, on the right-hand bar, the annotations for that fragment, and by clicking on them, a modal window shows the details. For example, Figure 6 shows the semantic representation of the entity *Partito Comunista Italiano*, which at least one fragment in the biography of Emilio Pugno is annotated with. Such a representation contains the label for the entity (in the upper left corner), the type (class) label (*Partito politico* -- *Political party*), the corresponding entity in Wikidata, if any (see Section 4.2), the value for the properties (in this case, the value for the data-property *name*, i.e., PCI), the URI in the Semantic KB, and the list of documents containing fragments annotated with this entity.

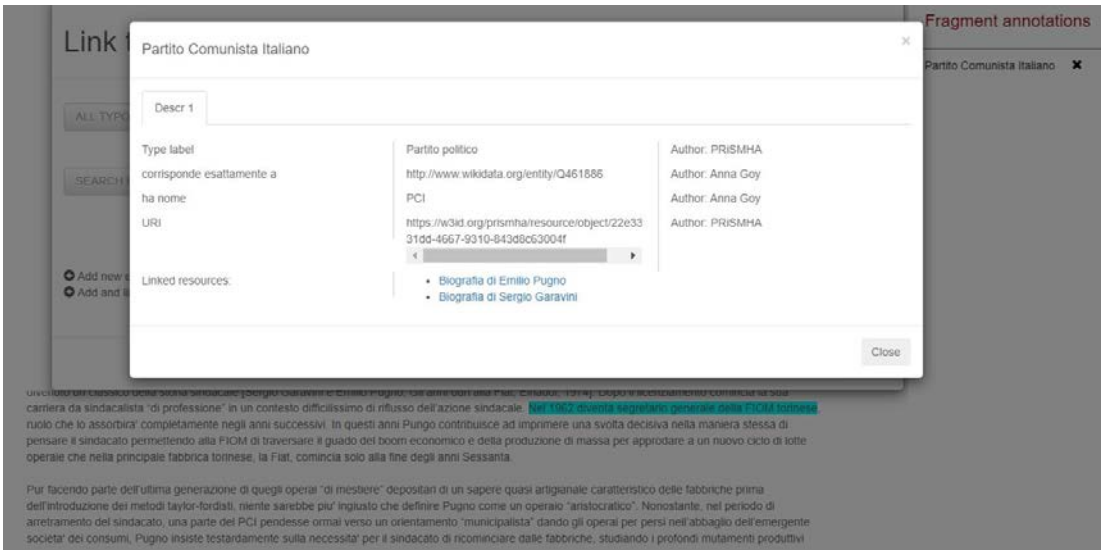

Figure 6: Semantic representation of the entity Partito Comunista Italiano (Italian Communist Party).

By clicking on a fragment, users can also add new annotations (Figure 7): the system suggests to search the Semantic KB for a suitable entity among existing ones; if nothing satisfactory is found, the user can create a new entity and link it to the fragment in focus, by clicking on *Add and link new entity* (or *Add new entity*, to link it later on).

Figure 7: The modal window enabling users to add a new annotation.

If the user decides to create a new entity, she can characterize it. Figure 8 shows the window enabling her to describe an entity through its properties. First, the user selects a HERO class (*Entity type*) and enters a label for the entity. On the basis of the selected class, the system calculates the available properties, which are presented as a form, splitted into three tabs, corresponding to *important*, *useful*, and *other* properties; the details of the algorithm computing the compatible properties with respect to the selected class, as well as the criteria to split properties into tabs, can be found in (Goy et al., 2020).

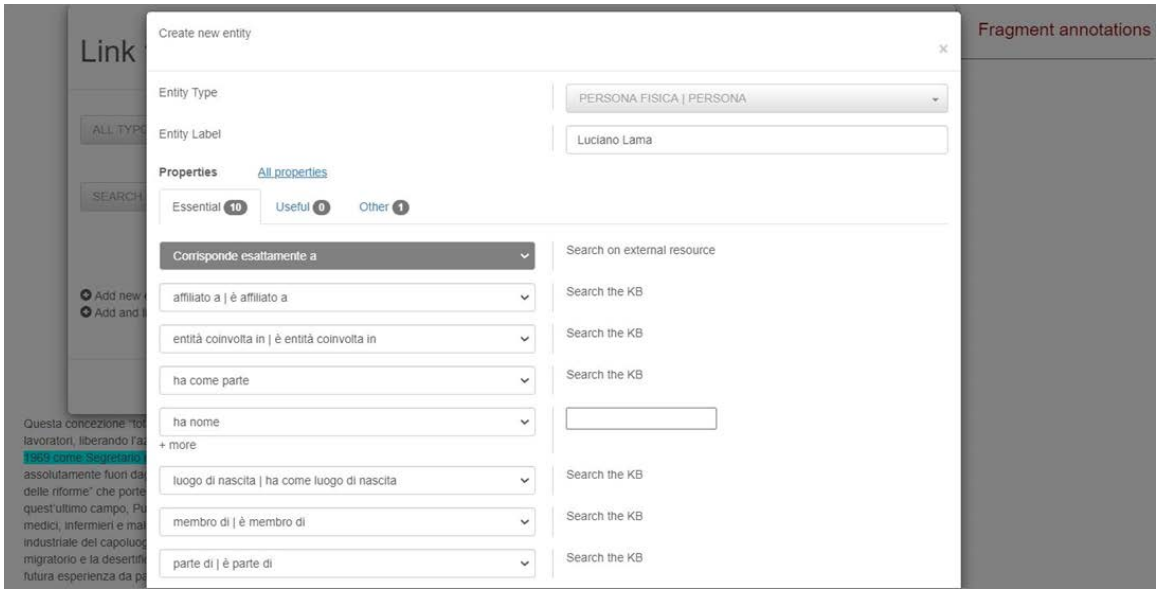

Figure 8: The window enabling users to describe an entity, by specifying its properties.

A detailed description of the properties is out of the scope of the present paper. We will focus on the first one (*Corrisponde esattamente a* -- *Exactly matching*, in the figure) in Section 4.2, when describing the link to

external datasets. For the other properties, it is worth saying that the system, again, suggests to start by searching the Semantic KB to find a filler; if no entity is suitable, the user can create a new one.
For the other functionalities of the *Crowdsourcing Platform UI* (e.g., the different privileges associated to different user types, or the visualization of the Semantic KB both in graphical and textual form), we suggest the reader to refer to (Goy et al., 2020). In Sections 4 and 5 we will describe the new features, representing the core of the work presented in this paper, i.e., the support provided to users by the *IE* and *LOD linking* modules.

4 Support to User Annotations

4.1 Automatic Information Extraction from Texts

Suppose that a user of the *Crowdsourcing Platform* is working on a text, obtained from an OCR-ized document or from an original textual source, as the one in Figure 5. By clicking on the *Show Named Entities and Temporal Expressions* link, Named Entities and Temporal Expressions are automatically extracted by the *IE* module (see Section 3) and highlighted in the text (see Figure 9, where the biography of Sergio Garavini, another well-known Italian trade union leader, is displayed).

Figure 9: A textual document (the biography of Sergio Garavini) with Named Entities and Temporal Expressions highlighted.

These entities are highlighted in two different ways, depending on whether or not they appear in a fragment (i.e., textual portions highlighted in cyan): entities that do not belong to fragments are represented with an orange font color and are not clickable, while entities that belong to fragments are represented with an orange background and are clickable. In the first case, the entities could help users to recognize interesting fragments, while in the second case they could be useful to recognize relevant entities representing the content of the fragment to be annotated (see the discussion in Section 5). By clicking on a Named Entity or Temporal Expression that occurs in a fragment, the system shows some information items hopefully helping the user in describing the entity (Figure 10). In particular, besides the entity name (corresponding to the expression identified in the text), the system suggests the HERO ontology class to be associated with the entity (*Elemento Geografico* -- *Geographic Feature* -- in this example), and it proposes potentially available links to external resources, such as the semantic networks BabelNet (Navigli and Ponzetto, 2010) and DBPedia (Auer et al., 2007). The system also tries to identify entities, already available in the *Semantic KB*, that refer to the Named

Entity or Temporal Expression recognized in the textual fragment (*Triplestore URI(s)* in the figure), in order to avoid duplicates. Such entities can be used to annotate the current fragment by clicking on *Link entity to fragment* (if the entity is already linked to the fragment in focus, the link can be removed). If no entity in the *Semantic KB* corresponds to the one recognized in the text fragment, the user can add a new one by clicking *Add new entity* or *Add and link new entity*: a user interface similar to the one in Figure 8 is shown, pre-compiled with the label and the ontology class automatically assigned by the system.

Figure 10: The window describing the recognized entity Torino, shown by clicking on "Torino" in Figure 9.

The *IE* module consists of two sub-modules:

- *Named Entities Recognition*, recognizing persons, organizations, and places;
- *Temporal Expressions Recognition*, recognizing hours, days, months, years, seasons, and centuries.

The *Named Entity Recognition* sub-module is based on the approach described in (Carducci et al., 2019), adapted to the Italian language. It consists of two integrated components:

- A component based on a machine learning approach, provided by the TINT Named Entity Recognition Module (Aprosio and Moretti, 2016), based on the Stanford CoreNLP NER module (Manning et al., 2014), which in turn relies on CRF (Conditional Random Field) classifiers (Lafferty et al., 2001). The classifier is trained on the Italian Content Annotation Bank (I-CAB) corpus (Magnini et al., 2006), containing around 180,000 words taken from the Italian newspaper *L'Adige*.
- A component that exploits a semantic-based approach, using the Word-Sense Disambiguation and Named Entities Recognition technique provided by BabelFy (Moro et al., 2014), that, in turn, employs the semantic network BabelNet (Navigli and Ponzetto, 2010) as source of information. The use of BabelFy is particularly relevant since it supports word-sense disambiguation, i.e., it selects the most promising sense within the set of candidates provided by BabelNet. Consider, for example, the sentence "Cavour è nato a Torino nel 1810" ("Cavour was born in Turin in 1810"); in this case, BabelFy recognizes the verb "nascere" ("to born") and two Named Entities ("Cavour" and "Torino"), assigning a particular sense to each of them. Assigning the correct sense to each element recognized in the text is not a trivial task: For example, if we look for the verb "nascere" ("to born") in BabelNet, the semantic network suggests ten possible senses (babelnet.org/search?word=nascere&lang=IT). Moreover, considering the entity "Cavour", the algorithm needs to discriminate between the sense representing the city of "Cavour" in Piedmont and the one representing the politician "Camillo Benso conte di Cavour", while considering the entity "Torino" the algorithm has to decide if the meaning of the entity in the sentence refers to the city of "Turin" in Piedmont or to the "Torino Football Club". As we can see in Figure 11, BabelFy

chooses the correct sense for each considered element, thanks to its word-sense disambiguation module.

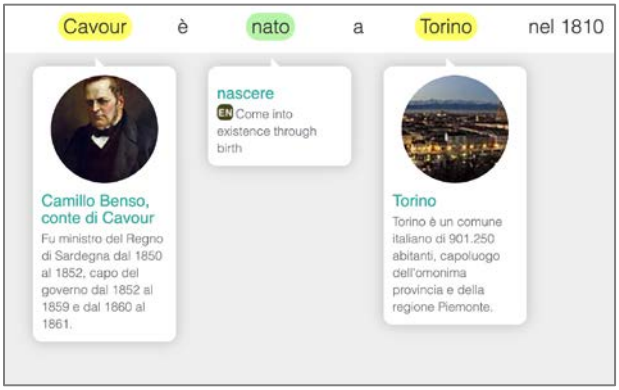

Figure 11: The results of the BabelFy disambiguation algorithm considering the sentence "Cavour è nato a Torino nel 1810" ("Cavour was born in Turin in 1810"), from http://babelfy.org/. Screenshot by authors.

The *Named Entity Recognition* sub-module can recognize three different types of entities, namely instances of the HERO class *PhysicalPerson*, instances of the HERO class *Organization*, and instances of the HERO class *GeographicFeature*, on the basis of the mappings shown in Table 1.

Table 1: Mappings between HERO, TINT, and BabelNet.

HERO	TINT	BabelNet
PhysicalPerson (https://w3id.org/hero/HERO-TOP#PhysicalPerson)	PER	BabelSynset representing the concept of human (https://babelnet.org/synset?word=bn:00044576n)
Organization (https://w3id.org/hero/HERO-ROCS#Organization)	ORG	BabelSynset representing the concept of company (https://babelnet.org/synset?word=bn:00021286n)
GeographicFeature (https://w3id.org/hero/HERO-TOP#GeographicFeature)	LOC	BabelSynset representing the concept of location (https://babelnet.org/synset?word=bn:00051760n)

TINT categories are natively supported by the TINT Named Entity Recognition Module, while the component based on BabelFy analyzes the ancestors of the entity in the BabelNet semantic network (following edges labeled is-a) in order to obtain the correct classification: For example, considering the BabelNet synset that represents the city of Turin, it is classified as a *city* (babelnet.org/synset?word=bn:03335997n), which in turn is a *settlement* (babelnet.org/synset?word=bn:00070724n), which in turn is a location; so, Turin is recognized as a *location*.

The results provided by the two components are merged into *S* (i.e., the set containing all the Named Entities that our system can retrieve) with the following strategy:

- If an entity is recognized only by the TINT-based component, it is added to *S* (associated with the corresponding class).
- If an entity is recognized only by the BabelFy-based component, it is added to *S* (associated with the corresponding class).

- As far as the entities that are recognized by both components are concerned, they are added to *S*, associated with the class identified by the TINT-based component, also in those cases in which the BabelFy-based component disagrees. This choice considers the fact that the accuracy of the classification is usually better for the approach based on TINT than for the one based on BabelFy.

The *Temporal Expressions Recognition* sub-module is based on the Heideltime library (Strötgen and Gertz, 2010), which, besides recognizing temporal expressions, also normalizes them. Normalization is essential in order to recognize the "prototype" of a particular temporal indication, whatever is the expression used in the text. As an example, both "2 giugno 2020" ("June 2nd 2020) and "2/6/2020" expressions are normalized as "2020-06-02". In this way, we have a unique representation of a temporal interval, independent from the natural language used in the text.
Heideltime recognizes temporal expressions using patterns represented as regular expressions, coded with the TimeML markup language (Pustejovsky et al., 2005). In particular, it can retrieve three types of temporal expressions:

- *explicit expressions*, such as "2 giugno 2020" or "2/6/2020";
- *implicit expressions*, such as "San Silvestro 2015" ("New Year's Eve 2015") or "Natale 2020" ("Christmas Day 2020"), normalized respectively as "2015-12-31" and "2020-12-25";
- *relative expressions*, that can only be normalized using the context in which they occur; for example, if it finds the expression "due anni dopo" ("two years later") and the previous lines provide the information that the story took place in 2017, Heideltime can normalize the expression as "2019".

Each recognized Temporal Expression is assigned one of the four TIMEX3 types (Saurí et al., 2006) available, namely:

- DATE, which describes a calendar time interval or subinterval (e.g., *a day*);
- TIME, which refers to a time frame within a day (e.g., *in the afternoon*);
- DURATION, which refers to explicit durations (e.g., *2 months*);
- SET, which describes a set of time intervals (e.g., *every two weeks*).

The PRiSMHA *Temporal Expressions Recognition* sub-module considers only the first two types. In particular, using a pattern-based approach which analyzes the normalized form (based on the Heideltime temporal tagger), we recognize five subtypes of entities belonging to the Heideltime DATE type, namely:

- days, i.e., instances of the HERO class *Day* (https://w3id.org/hero/HERO-TIME#Day);
- months, i.e., instances of the HERO class *CalendarYearMonth* (https://w3id.org/hero/HERO-TIME#CalendarYearMonth);
- years, i.e., instances of the HERO class *CalendarYear* (https://w3id.org/hero/HERO-TIME#CalendarYear);
- seasons, i.e., instances of the HERO class CalendarSeason (https://w3id.org/hero/HERO-TIME#CalendarSeason).

Entities associated by Heideltime with the TIME type are recognized as instances of the HERO class *DayTime* (https://w3id.org/hero/HERO-TIME#DayTime), representing time spans within a day, like hours.

4.2 Linking External Datasets

Fully describing an entity, from a semantic point of view, is not a trivial task, and the user that is adding a new entity to the *Semantic KB* often is not aware of all its features. In this case, she can ask the system for support, to obtain further information and possibly link the PRiSMHA entity to an external one. In particular, when facing the form for the characterization of a new entity, the user can click on the *Search on external resource* link (see Figure 8), thus activating the Wikidata explorer interface. Wikidata (Vrandečić and

Krötsch, 2014) is a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation and focused on items that represent topics, concepts, or objects using RDF as data model (www.w3.org/RDF). Each item is represented by a unique and persistent identifier, which is a positive integer prefixed with the upper-case letter Q, known as *QID*.
The Wikidata support page available in the PRiSMHA *Crowdsourcing Platform UI* (Figure 12) enables the user to specify the entity to search, by indicating a label for the entity (e.g., the name of the particular person to search) and its type. As for now, the user can search for persons, organizations, and places. In particular, the mappings shown in Table 2 have been defined.

Table 2: Mappings between HERO and Wikidata types.

HERO	Wikidata
PhysicalPerson (https://w3id.org/hero/HERO-TOP#PhysicalPerson)	human -- Q5 (https://www.wikidata.org/wiki/Q5)
Organization (https://w3id.org/hero/HERO-ROCS#Organization)	organization -- Q43229 (https://www.wikidata.org/wiki/Q43229)
GeographicFeature (https://w3id.org/hero/HERO-TOP#GeographicFeature)	geographic location -- Q2221906 (https://www.wikidata.org/wiki/Q2221906)

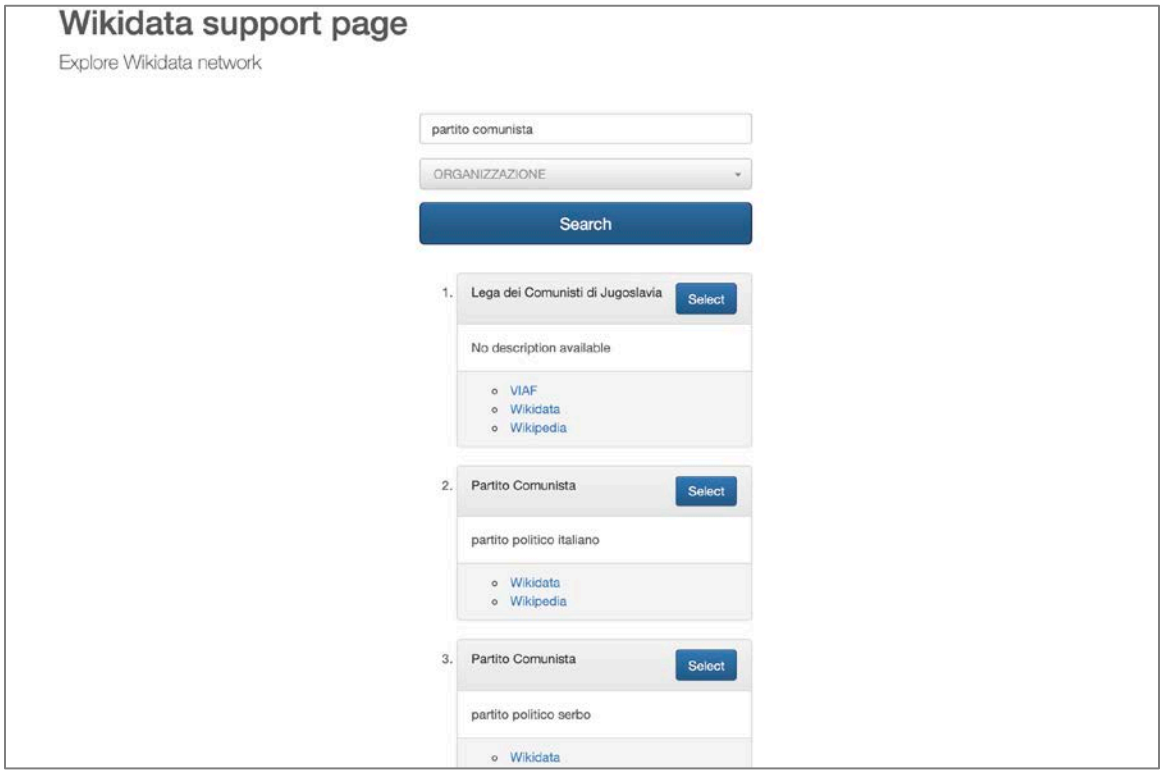

Figure 12: Wikidata support page.

When label and type have been specified, the user can click on the *Search* button: The system sends a query to Wikidata (see details below) and shows the results in the form of cards (Figure 12). For example, if the

user searches for an entity labeled “Partito comunista”, the system retrieves multiple candidates (*Partito Comunista Italiano*, *Partito Comunista Serbo*, *Partito Marxista-leninista americano*, etc…), among which the user can select the entity matching the one she has in mind by clicking on the corresponding Select button. Henceforth, the form for characterizing the new entity is shown (see Figure 8) filled-in with the link between the entity that the user is describing and the corresponding entry in Wikidata.
Actually, the user can select between two types of matching:

- *Corrisponde esattamente a* (*exactly matching*): it refers to the *skos:exactMatch* property in SKOS (Simple Knowledge Organization System) (Miles and Pérez-Agüera, 2007) and “indicates a high degree of confidence that two concepts can be used interchangeably across a wide range of information retrieval applications” (www.w3.org/TR/skos-reference/#L4858).
- *Corrisponde più o meno a* (*roughly corresponding to*): it refers to the *skos:closeMatch* property in SKOS and “indicates that two concepts are sufficiently similar that they can be used interchangeably in some information retrieval applications” (www.w3.org/TR/skos-reference/#L4858).

In order to help the user to discriminate between different candidates, each one is represented as a card showing the following information:

- a Wikidata label (in Italian);
- a short description of the entity (in Italian);
- a set of available links to external resources representing the same entity; in particular, we selected the following external resources as relevant in our historical domain:
  Wikipedia (https://it.wikipedia.org/wiki/Pagina_principale)
  Enciclopedia Treccani (http://www.treccani.it)
  The Virtual International Authority File (VIAF, https://viaf.org)
  Catalogo del Servizio Bibliotecario Nazionale (SBN, https://opac.sbn.it/opacsbn/opac/iccu/free.jsp)
  Openpolitici (https://politici.openpolis.it/)
  Sito storico del Senato (http://www.senato.it/)
  Sito della camera (http://dati.camera.it/)
  Sito storico della camera (https://storia.camera.it/)
  Dizionario storico Treccani (http://www.treccani.it/)

When the user clicks on the *Search* button in the Wikidata support page (Figure 12), a SPARQL query is sent to the Wikidata Query Service (https://query.wikidata.org/), using the JENA API (McBride, 2002). Moreover, in order to retrieve the needed information, some additional services are used (Malyshev et al., 2018), namely:

- *wikibase:label* (https://en.wikibooks.org/wiki/SPARQL/SERVICE_-_Label): Wikibase (https://wikiba.se/) is used to obtain the label and the short italian description for the entity;
- *wikibase:mwapi* (https://en.wikibooks.org/wiki/SPARQL/SERVICE_-_mwapi): MediaWiki API Query Service (MWAPI service) is used to search for entities filtered by a particular type selected by the user.

For example, if the user is looking for “Sergio Garavini” (label) and provides *PhysicalPerson* as type, the following SPARQL query is executed:

```
PREFIX bd: <http://www.bigdata.com/rdf#>
        PREFIX wikibase: <http://wikiba.se/ontology#>
        PREFIX wdt: <http://www.wikidata.org/prop/direct/>
        PREFIX wd: <http://www.wikidata.org/entity/>
        PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
        PREFIX mwapi: <https://www.mediawiki.org/ontology#API/>
        PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
        SELECT DISTINCT ?item ?itemLabel ?itemDescription
        WHERE
```

```
{
        SERVICE wikibase:mwapi {
                bd:serviceParam wikibase:api "EntitySearch" .
                bd:serviceParam wikibase:endpoint "www.wikidata.org" .
                bd:serviceParam mwapi:search "Sergio Garavini" .
                bd:serviceParam mwapi:language "it" .
                ?item wikibase:apiOutputItem mwapi:item .
        }
        ?item wdt:P31/wdt:P279* wd:Q5.
        SERVICE wikibase:label { bd:serviceParam wikibase:language "it" }
}
ORDER BY asc(str(fn:lower-case(?itemLabel)))
```

Consider, in particular, the line `?item wdt:P31/wdt:P279* wd:Q5`: the Wikidata node *wd:Q5* (https://www.wikidata.org/wiki/Q5) represents the concept of *human* (mapped onto the HERO class *PhysicalPerson*), thus the system selects all the entities in Wikidata, retrieved by the MWAPI service, that are instances of human or of any of its sub-concepts.
Available links to external resources are found through another SPARQL query; for example, the external links related to the Wikidata node *Q338536* (https://www.wikidata.org/wiki/Q338536) representing Sergio Garavini are extracted with the following query:

```
SELECT ?treccaniURL ?viafURL ?sbnURL ?openPolisURL ?senateURL ?cameraDatiURL ?cameraStori
aURL ?storiaTreccaniURL ?wikipediaURL WHERE {
  wd:P3365 wdt:P1630 ?treccaniFormatter .
  wd:P214 wdt:P1630 ?viafFormatter .
  wd:P396 wdt:P1630 ?sbnFormatter .
  wd:P1229 wdt:P1630 ?openPolisFormatter .
  wd:P2549 wdt:P1630 ?senateFormatter.
  wd:P1341 wdt:P1630 ?cameraDatiFormatter.
  wd:P3935 wdt:P1630 ?cameraStoriaFormatter.
  wd:P6404 wdt:P1630 ?storiaTreccaniFormatter.
  optional {?wikipediaIRI schema:about wd:Q338536 ;  schema:isPartOf
<https://it.wikipedia.org/>}.
  optional {wd:Q338536 wdt:P3365 ?treccaniID} .
  optional {wd:Q338536 wdt:P214 ?viafID} .
  optional {wd:Q338536 wdt:P396 ?sbnID} .
  optional {wd:Q338536 wdt:P1229 ?openPolisID} .
  optional {wd:Q338536 wdt:P2549 ?senateID} .
  optional {wd:Q338536 wdt:P1341 ?cameraDatiID} .
  optional {wd:Q338536 wdt:P3935 ?cameraStoriaID} .
  optional {wd:Q338536 wdt:P6404 ?storiaTreccaniID} .
  BIND(str(?wikipediaIRI) as ?wikipediaURL) .
  BIND(REPLACE(?treccaniID, '^(.+)$', ?treccaniFormatter) AS ?treccaniURL).
  BIND(REPLACE(?viafID, '^(.+)$', ?viafFormatter) AS ?viafURL).
  BIND(REPLACE(?sbnID, '^(.+)$', ?sbnFormatter) AS ?sbnURL).
  BIND(REPLACE(?openPolisID, '^(.+)$', ?openPolisFormatter) AS ?openPolisURL).
  BIND(REPLACE(?senateID, '^(.+)$', ?senateFormatter) AS ?senateURL).
  BIND(REPLACE(?cameraDatiID, '^(.+)$', ?cameraDatiFormatter) AS ?cameraDatiURL).
  BIND(REPLACE(?cameraStoriaID, '^(.+)$', ?cameraStoriaFormatter) AS ?cameraStoriaURL).
  BIND(REPLACE(?storiaTreccaniID, '^(.+)$', ?storiaTreccaniFormatter)
AS ?storiaTreccaniURL).
  } LIMIT 1
```

5 Evaluating Suggestions by IE and LOD Modules

5.1 Evaluation Setting

In order to assess the effectiveness of *IE* and *LOD linking* modules in supporting PRiSMHA users, we carried out a qualitative evaluation, for which 30 subjects were recruited. Each participant was asked to perform twice a sequence of mini-tasks (in the following this sequence will be referred to as the *Main Task*), once with a version of the *Crowdsourcing Platform* prototype without IE and LOD support, and once with the full-fledged version. The subjects were split in two groups of 15 people, named Group O and Group W:

- Group O performed the Main Task first withOut IE and LOD support, and then with the full-fledged version of the prototype;
- Group W performed the Main Task first With IE and LOD support, and then with the stripped-down version of the prototype.

The two groups were necessary because it could be expected that, repeating the Main Task twice, the second execution would be experienced as "easier" by the participants (as we will see in the Section 3.5, this was indeed the case): We did not want them to incorrectly attribute their increased ease to our support tools, when it was indeed due to a better knowledge of the application. The Main Task included the following steps:

(A) Log into the prototype and find a given project.
(B) Find a specific document within the project and read it. The documents used for this step already included a few sample annotations, and the relevant fragments within them were highlighted.
(C) Find a few relevant entities within the text, check them out for possible existing annotations, and, if not, annotate the corresponding fragment as they saw fit.

Users were asked to annotate two different documents, one to be used in the first execution of the Main Task, the other in the second. The two documents were the same for everyone, but which was used in the first execution and which in the second was randomly chosen by the system.
The third sub-step was the active phase, where the IE and LOD support played its role. In the full-fledged version of the prototype the participants could benefit of the following functionalities:

- NER/TER: When the user is reading a document, by clicking on the Show Named Entities and Temporal Expressions link, the IE module (see Section 3) automatically identifies Named Entities and Temporal Expressions. The corresponding phrases are highlighted in the text (see Section 4.1), by using an orange background when they belong to a fragment, and an orange text color when they do not.
- INFO: By clicking on a Named Entity or Temporal Expression that occurs in a fragment, the system shows some information items describing the entity. In particular, it proposes links to external resources (namely, DBPedia and BabelNet, see Section 4.1), if available. The system also tries to identify those entities that are already available in the Semantic KB, in order to avoid duplicates. Such entities can be directly used to annotate the fragment in focus by clicking on Link entity to fragment (see again Section 4.1).
- LINK: When adding a new entity, in order to annotate a fragment with it, the user can specify a label and a type for the entity, and then ask the PRiSMHA platform to search for possible matches on Wikidata. The Wikidata query results are then displayed as a list of cards among which the user can select the one representing the entity s/he has in mind. Such an entity can be linked to the one in the PRiSMHA Semantic KB by means of the matches exactly or roughly corresponds to property (see Section 4.2).
- AUTOFILL: When the user selects a Wikidata entry as an exact or rough match, the PRiSMHA platform prompts her with the "Create new entity" form (see Figure 8 and Section 4.2); the system then automatically fills in the label and type fields thanks to the data retrieved from Wikidata.

After completing the assigned tasks, participants were asked to fill in a questionnaire which mainly focused on the difference between the two experiences with the two prototype versions. The results we present in the next section consist of the questionnaire answers we collected.

5.2 Results

All the subjects had the Italian equivalent of a BSc, a MSc or a PhD. In both groups, all of the subjects but one read their emails and browsed the Web on a daily basis. All of them also worked with standard office applications (text editors, electronic sheets, presentation software) at least on a weekly basis. All of them

declared a regular usage of both a personal computer and a smartphone. About half of them (7 out of 15 in group W, 8 out of 15 in group O) said they also regularly used a tablet.
The first part of the questionnaire was aimed at measuring the perceived complexity of the task. We asked the subjects to express their agreement with three statements, on a 5-point scale ranging from 1 (complete disagreement) to 5 (complete agreement). The statements were the following:

(A) The task was in itself complex;
(B) Even if the task was complex, it became easier once learned;
(C) I did what was asked of me quite easily.

Figure 13 shows the answers to these questions, split between the two groups W and O, by means of boxplots.

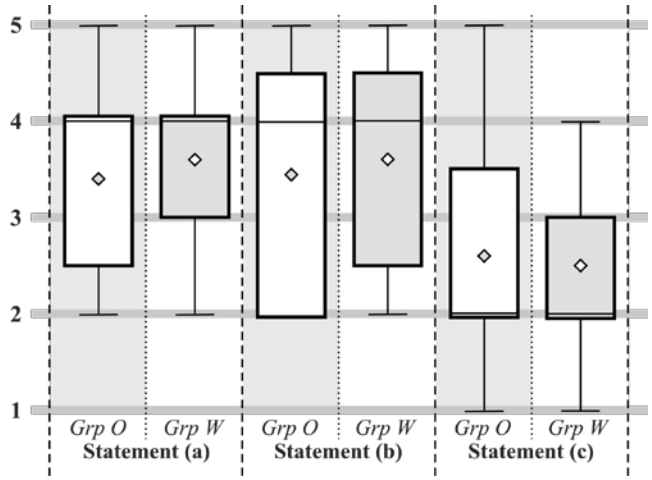

Figure 13: Boxplots for the answers to the first part of the questionnaire, evaluating the difficulty of the task. Participants were asked how much they agreed with statements (a), (b) and (c) described in the text. Answers were given on a 5-point scale (1 to 5), with 1 representing complete disagreement and 5 representing complete agreement. The boxplot themselves span from the 1st quartile Q1 to the 3rd quartile Q3, with a dividing line showing the median (2nd quartile Q2). Whiskers show the minimum and maximum values based on the interquartile range (respectively, the lowest value above Q1–1.5*IQR, and the highest value below Q3+1.5*IQR). The diamond inside each boxplot represents the mean value.

Subsequently, we asked the subjects to express the degree to which each of the support functionalities mentioned above (NER/TER, INFO, LINK, AUTOFILL) helped them or rather hindered them. Again, the participants could answer on a 5-point scale, with 1 representing “significant hindrance” and 5 representing “significant help”.
The boxplots for the replies to these questions, from both group W and group O, are shown in Figure 14. In this phase users could also provide free-text comments, if they wanted to do so; the feedback received in these comments is discussed in Section 5.3.

5.3 Discussion

As commented in Section 3, the PRiSMHA *Crowdsourcing Platform* is aimed at a quite specific and competent type of users: Using the application is not easy, at least at first glance, and becoming acquainted with the underlying ontology requires some background knowledge about the domain and some interaction rounds. Although our 30 test subjects were reasonably tech-savvy, and with a good degree of academic education, they were using the application for the first time, and this can explain why most of them, in both groups, found the task rather complex, although less so once learned, and they did not find it easy to complete it, as the results in Figure 13 show. Such difficulty could of course impact on their evaluation of the support functionalities offered by the application, since appreciating them required a certain degree of knowledgeability on both the domain and the document annotation task itself.

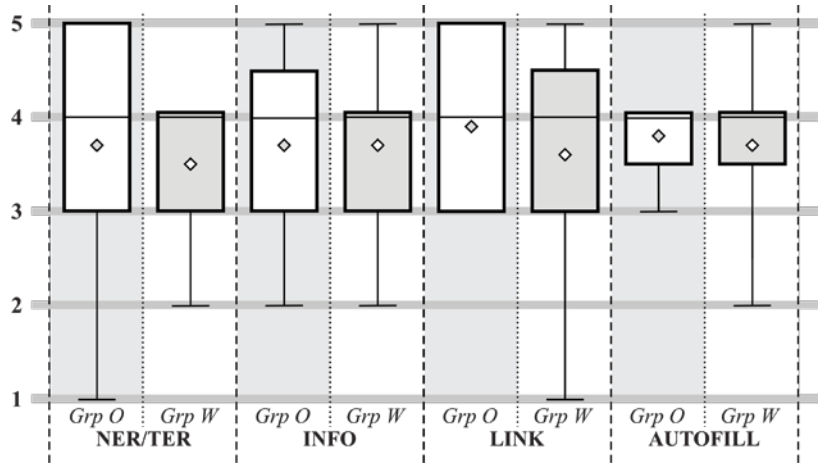

Figure 14: Boxplots for the answers to the second part of the questionnaire, evaluating the degree of help/hindrance experienced for each support functionality. Answers were given on a 5-point scale (1 to 5), with 1 representing significant hindrance and 5 representing significant help. The conventions for the boxplot representation are the same as those used for Figure 13..

Nonetheless, Figure 14 shows that both groups of users found the four support functionalities reasonably helpful. None of the boxplots falls in the bottom ("hindrance") half of the diagram, and for all four functionalities the median is 4, and the mean falls between 3.5 and 4.

It can be observed that answers from Group O were consistently slightly lower than those of group W. This mild difference between the two groups can be ascribed to the fact that, for group W, the lack of support in the second execution of the task was compensated by an increased familiarity with the application. Also, as some participants remarked in the free-text comments to the questionnaire, the support features added visual and interaction complexity for people new to the task (group W). The support features were in general easier to exploit for those that had already learned the basic use of the application, i.e. people in Group O. This is consistent with the fact that, as we can notice in Figure 13, group W deemed the task slightly more difficult, and less easy to learn, than group O.

While all the participants found the second execution easier than the first, some people in Group W actually blamed the support tools for this. As previously stated, we introduced the two groups to factor out a "false" positive bias from Group O ("the second time I did it, it was easier, thus the support tools were helpful"); apparently Group W actually "compensated" it with the opposite negative bias ("the second time I did it, it was easier, thus the support tools were a hindrance"). For this reason we appreciated in particular the overall positive evaluation given from the participants in Group W.

As stated above, participants had the possibility to add free text comments to their answers. These provided us with directions regarding possible areas of improvement as well as an assessment of the main advantages provided by the support tools. Let us briefly discuss them.

Regarding the NER/TER support, 18 people out of 30 provided a free text comment. Of these 18, 13 participants remarked positively on the helpfulness of the tool: they not only found it useful to identify examples of potential entities ("it helps recognizing which phrases correspond to entities", "it helps recognizing relevant entities"), but also saw it as an aid in reading the text, identifying key concepts and relevant portions ("it was useful in identifying keywords", "it made my job faster", "it helped me choose which parts of a fragment were relevant and which not"). The remaining 5 participants commented on possible improvements: 2 of these remarks concerned UI improvements ("the meaning of differences in text color/background was not immediately clear"; "I was expecting that the highlighted entities were not only suggested, but already added into the system"), while the remaining 3 concerned the IE module itself, which did not always identify the full "phrase" corresponding to the entity. This was particularly true in case of organizations with articulated names; for example, an organization named "Federazione giovanile del Partito socialista di unità proletaria" (Youth Federation of the Proletarian Unity Socialist Party) was recognized as

three separate entities (“Federazione giovanile”/Youth Federation, “Partito socialista”/Socialist Party, “unità proletaria”/Proletarian Unity).
Moving to the INFO support, i.e. the link to external resources provided when clicking on an entity recognized by NER or TER, we collected 13 free text comments. 8 participants expressed a positive remark (“it sped up the process of inserting simple entities, so that I could concentrate on more complex ones”, “the external links were useful to discover more about certain entities”) with some suggestions for improving the UI (“it would be helpful to include here a brief description of the entity taken from these external sources”). Other 5 participants expressed perplexity toward this feature, partly because the dialog presenting the information was not informative enough (“I did not understand these alphanumeric IDs”), partly because of unmet expectations (“I expected to be able to directly add the entity by using the external sources”). This last case is reported by 3 people; it is interesting to note that they lamented the absence of a feature that was made available at a later stage in the annotation process, i.e. the LINK support tool.
The LINK support received 9 free text comments. 6 people commented positively, stating that being able to search for possible correspondences in Wikidata helped them to discover more on the entity itself; besides being helpful for filling in the entity creation form (“It was very intuitive to use and it helped me discover facts concerning the entity I was exploring”), they also found it interesting as an enrichment of their knowledge on the topic (“I found unexpected connections”) and as a validation instrument (“I think that connecting to an external resource is an enrichment because it somehow validates the identity of the entity itself”). In this case, the 3 negative remarks concerned exclusively the UI: for 1 participant the labels used in the form were misleading (“It was not clear to me that selecting *exactly matches* would ensue in a search on Wikidata”); other 2 people complained that the search would not find anything, probably because it was not clear that both the label and type fields needed to be filled in for the search to work.
Last but not least, 9 people commented on the AUTOFILL tool. All of them provided positive remarks, highlighting how having these two fields already filled in partially compensated the complexity of the form, making it faster to fill it in, if not always easier (“Without it you would spend a lot of time finding the right category”, “When I had the automatic suggestion I felt less worried about making mistakes”).
On the whole, most of the problems with the support tools turned out to be related, not to the support itself -- that has been evaluated positively -- but to some awkwardness in the User Interface, which either did not sufficiently highlight how one could benefit from the support, or did not enforce the correct steps needed for the support to be effective. Nonetheless these remarks pointed out the main directions of improvement.
An interesting point is the request for a broader support in the annotation task itself: in the NER/TER support, some users suggest to go beyond simple identification of entities in the text, together with correspondence to external resources, and ask for an automatic annotation, or at least a suggestion for it. Also in the LINK functionality, some users suggest to automatically create the candidate annotation on the basis of the Wikidata entry.
These enhancements in the annotation support are not trivial (for example, they risk to produce lower-quality data within the Semantic KB), but they are clearly worth to be considered. Moreover, the fact that users found the AUTOFILL functionality useful, encourages us to plan to design a new version of the system where properties other than typology and label are automatically pre-filled (see Section 6).

6 Conclusions and Future Work

In this paper we have demonstrated that automatic techniques can be successfully exploited in order to support users in the semantic annotation of archival documents. More specifically, we have shown that Named Entity (and temporal Expression) Recognition techniques are useful when users of the PRiSMHA *Crowdsourcing Platform* have to identify new entities to annotate a document with. Moreover, when creating new entities to be added to the *Semantic KB*, exploiting information from external datasets such as Wikidata proved to be useful. These findings answer the **research question** introduced in Section 1: Given an ontology-driven web-based system enabling users to build the formal semantic representations of archival document content, automatic text mining techniques (namely, Named Entity and Temporal Expression

Recognition), and entity linking to external resources (Linked Open Data) actually provide users with an effective support in the (semantic) annotation activity.

Participants in the evaluation also suggested some promising improvements, the most interesting and challenging of which concerns the exploitation of the mentioned external datasets such as Wikidata. As commented in Section 5.3, when a Wikidata entry was selected, the PRiSMHA platform automatically fills in the label and type fields of the form for the characterization of the corresponding new entity to be added to the *Semantic KB*. The fact that users valued very much this feature encourages us to carry on the study aimed at designing a new version of the system where properties other than typology and label are automatically pre-filled on the basis of information retrieved from Wikidata, or from other datasets.

ACKNOWLEDGMENTS

This work has been supported by Compagnia di San Paolo Foundation and Università di Torino within the PRiSMHA project (CSTO168023). Thanks to all PRiSMHA collaborators, and special thanks to Rossana Damiano and Daniele Paolo Radicioni, for their valuable support in the project.

REFERENCES

Alma'aitah, W.Z., Talib, A.Z., and Osman, M.A. (2020). "Opportunities and challenges in enhancing access to metadata of Cultural Heritage collections: a survey". Artificial Intelligence Review, 53: 3621–3646.

Andrews, P., Zaihrayeu, I., and Pane, J. (2012). "A Classification of Semantic Annotation Systems". Semantic Web, 3(3): 223-248.

Aprosio, A. P., and Moretti, G. (2016). "Italy goes to Stanford: a collection of CoreNLP modules for Italian", arXiv preprint arXiv:1609.06204.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). "Dbpedia: A nucleus for a web of open data". In Aberer, K. et al. (Eds.) "The semantic web. ISWC 2007, ASWC 2007", LNCS 4825, Springer, Berlin-Heidelberg, 722-735.

Battad, Z., White, A., and Si, M. (2019). "Facilitating Information Exploration of Archival Library Materials Through Multi-modal Storytelling". In Cardona-Rivera, R. E., Sullivan, A., and Young, R. M. (Eds.) "International Conference on Interactive Digital Storytelling", LNCS 11869, Springer Nature Switzerland, 120–127.

Borgo, S., and Masolo, C. (2009) "Foundational choices in dolce", in Staab, S., and Studer, R. (Eds.) "Handbook on Ontologies", 2nd Ed., Springer, 361-381.

Carboni, N., and De Luca, L. (2017). "Towards a Semantic Documentation of Heritage Objects through Visual and Iconographical Representations". International Information & Library Review, 49: 207-217.

Carducci, G., Leontino, M., Radicioni, D. P., Bonino, G., Pasini, E., and Tripodi, P. (2019). "Semantically Aware Text Categorisation for Metadata Annotation". Proceedings of the Italian Research Conference on Digital Libraries, Pisa, Italy, 315-330.

Crofts, N., Doerr, M., and Gill, T. (2003). "The CIDOC Conceptual Reference Model A Standard for Communicating Cultural Contents". Cultivate Interactive, 9.

Damiano, R., and Lombardo, V. (2016). "Labyrinth 3D. Cultural archetypes for exploring media archives". Digital Creativity, 27(3): 234-255.

Daquino M., Mambelli F., Peroni S., Tomasi F., and Vitali F. (2016). "Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data". J. on Computing and Cultural Heritage, 10(4): 1-21.

Daquino, M., and Tomasi, F. (2015). "Historical context ontology (HiCO) - A conceptual model for describing context information of Cultural Heritage objects". In Garoufallou, E., Hartley, R. J., and Gaitanou P. (Eds.) "Metadata and Semantics Research", Vol. 544, Springer, Berlin, 424-436.

Doulaverakis, C., Kompatsiaris, I., and Strintzis, M.G.. (2005). "Ontology-Based Access to Multimedia Cultural Heritage Collections - The REACH Project". Proceedings of EUROCON 2005 - The International Conference on "Computer as a Tool", 151-154.

Dragoni, M., Tonelli, S., and Moretti, G. (2016). "A knowledge management architecture for digital Cultural Heritage". ACM Transactions on Applied Perception, 1(1).

Elsweiler D., Wilson M.L., and Kirkegaard Lunn B. (2011). "Understanding casual-leisure information behaviour". In Spink, A. and Heinstrom, J. (Eds.) "New Directions in Information Behaviour", Emerald Press, Bingley, 211-241.

Foley, J., Kwan, P., and Welch, M. (2017). “A web-based infrastructure for the assisted annotation of heritage collections”. J. on Computing and Cultural Heritage, 10(3): 1-25.

Garozzo, R., Murabito, F., Santagati, C., Pino, C. and Spampinato, C. (2017). “CULTO: An Ontology-based annotation tool for data curation in Cultural Heritage”. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 42(2/W5), 267-274.

Ghiselli, C., Bozzato, L., and Trombetta A. (2005). “Representation and management of ontologies in Cultural Heritage domain”. Proceedings of Semantic Web Applications and Perspectives (SWAP 2005), vol. 1338 of CEUR workshop proceedings. CEUR-WS.org.

Goy, A., Accornero, C., Astrologo, D., Colla, D., D'Ambrosio, M., Damiano, R., Leontino, M., Lieto, A., Loreto, F., Magro, D., Mensa, E., Montanaro, A., Mosca, V., Musso, S., Radicioni, D. P., and Re, C. (2019a). “Fruitful synergies between computer science, historical studies and archives: the experience in the PRiSMHA project”. Proceedings of the Int. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Vol. 3: KMIS, 225-230.

Goy, A., Colla, D., Magro, D., Accornero, C., Loreto, F., and Radicioni, D. P. (2020). “Building Semantic Metadata for Historical Archives through an Ontology-driven User Interface”, J. on Computing and Cultural Heritage, 13(3).

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., and D’ambrosio, M. (2017). “PRiSMHA (Providing Rich Semantic Metadata for Historical Archives”. Proceedings of the Contextual Representation of Objects and Events in Language, Bolzano, Italy.

Goy, A., Magro, D., and Baldo, A. (2019b). “A Semantic Web Approach to Enable a Smart Route to Historical Archives”. Journal of Web Engineering, 18(4-6): 287-318.

Goy, A., Magro, D., and Rovera, M. (2015). “Ontologies and historical archives: a way to tell new stories”. Applied Ontology, 10(3/4): 331-338.

Goy, A., Magro, D., and Rovera, M. (2018). “On the Role of Thematic Roles in a Historical Event Ontology”. Applied Ontology, 13: 19-39.

Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., and Kettula, S. (2005). “MuseumFinland - Finnish museums on the semantic web”. J. of Web Semantics, 3(2-3): 224-241.

Kollia, I., Tzouvaras. V., Drosopoulos, N., and Stamou, G. (2012). “A systemic approach for effective semantic access to cultural content”. Semantic Web J., 3(1): 65-83.

Lana, M., Ciotti, F., Magro, D., Peroni, S., Tomasi, F., Vitali, F. (2014). “Annotating texts with ontologies, from geography to persons and events”. Proceedings of Digital Humanities 2014.

Laclavik, M., Šeleng, M., Gatial, E., Balogh, Z., and Hluchý, Ladislav. (2006). “Ontology based Text Annotation - OnTeA”. In Duží, M., Jaakkola, H., Kiyoki, Y., and Kangassalo, H. (Eds.) Information Modelling and Knowledge Bases XVIII. IOS Press, Amsterdam, 311-315.

Lafferty, J., McCallum, A., and Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". Proceedings of the 18th Int.l Conf. on Machine Learning, Morgan Kaufmann; 282-289.

Lee, C. A., and Tibbo, H. (2011). “Where’s the archivist in digital curation? exploring the possibilities through a matrix of knowledge and skills”. Archivaria, 72: 123-168.

Lombardo, V., Pizzo, P., and Damiano, R. (2016). “Safeguarding and accessing drama as intangible Cultural Heritage”. J. on Computing and Cultural Heritage, 9(1): 1-26.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. (2006). “I-CAB: the Italian Content Annotation Bank”. Proceedings of the 5th Int. Conf. on Language Resources and Evaluation (LREC 2006), 963-968.

Malyshev, S., Krötzsch, M., González, L., Gonsior, J., and Bielefeldt, A. (2018). “Getting the most out of Wikidata: semantic technology usage in Wikipedia’s knowledge graph”. Proceedings of the Int. Semantic Web Conference, 376-394.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). “The Stanford CoreNLP natural language processing toolkit”. Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations, 55-60.

McBride, B. (2002). “Jena: A semantic web toolkit”. IEEE Internet computing, 6(6): 55-59.

Miles, A., and Pérez-Agüera, J. R. (2007). “Skos: Simple knowledge organisation for the web”. Cataloging & Classification Quarterly, 43(3-4), 69-83.

Moro, A., Raganato, A., and Navigli, R. (2014). “Entity linking meets word sense disambiguation: a unified approach”. Transactions of the Association for Computational Linguistics, 2, 231-244.

Motta, E., Buckingham Shum, S., and Domingue, J. (2000). “Ontology-driven document enrichment: principles, tools and applications”. Int. J. of Human-Computer Studies, 52(6): 1071-1109.

Navigli, R., and Ponzetto, S. P. (2010). “BabelNet: Building a very large multilingual semantic network”. Proceedings of the 48th annual meeting of the association for computational linguistics, 216-225.

Peroni, S., and Shotton, D. M. (2012). “FaBiO and CiTO - Ontologies for describing bibliographic resources and citations”. J. of Web Semantics, 17, 33–43-.

Peroni, S., Shotton, D., and Vitali, F. (2012). “Scholarly publishing and linked data: Describing roles, statuses, temporal and contextual extents”. Proceedings of the 8th Int. Conf. on Semantic Systems, ACM Press, 9-16.

Post, C., Chassanoff, A., Lee, C. A., Rabkin, A., Zhang, Y., Skinner, K., and Meister, S. (2019). “Digital Curation at Work: Modeling Workflows for Digital Archival Materials”. Proceedings of ACM/IEEE Joint Conf. on Digital Libraries, ACM Press, 39-48.

Pustejovsky, J., Ingria, R., Saurì, R., Castaño, J. M., Littman, J., Gaizauskas, R. J., Setzer, A., Katz, G., and Mani, I. (2005). “The Specification Language TimeML”. In Pustejovsky, J., Mani, I., and Gaizauskas, R. (Eds.) “The Language of Time-A Reader”, Oxford University Press, 545-558.

Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). “TimeML annotation guidelines”. www.timeml.org.

Schreiber, G., Amin, A., Aroyo, L. , van Assem, M., De Boer, V ., Hardman, L. , Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A. , Wielemaker, J., and Wielinga, B. (2008). “Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator”. Web Semantics, 6(4). 243-249.

Strötgen, J., and Gertz, M. (2010). “Heideltime: High quality rule-based extraction and normalization of temporal expressions”. Proceedings of the 5th Int. Workshop on Semantic Evaluation, 321-324.

Tommy, M., Véron, P., Halin, G. and De Luca, L. (2017). “An ontological model for the reality-based 3D annotation of heritage building conservation state”. J. of Cultural Heritage. 29, 100-112.

Tonkin, E. L., and Tourte, G. J. L. (2016). “Using the crowd to update Cultural Heritage catalogue”. Proceedings of Presented at involving the crowd in future museum experience design - CHI 2016 workshop, 1-6.

Underberg-Goode, N. (2017). “Digital Storytelling for Heritage across Media”. Collections: A Journal for Museum and Archives Professionals, 13(2): 103-114.

Vrandečić, D., and Krötzsch, M. (2014). “Wikidata: a free collaborative knowledgebase”. Communications of the ACM, 57(10): 78-85.

Walsh, D., and Hall, M.M. (2015). “Just looking around: supporting casual user’s initial encounters with digital Cultural Heritage”. Proceedings of the 1st Int. Workshop on Supporting complex search tasks, vol. 1338 of CEUR workshop proceedings. CEUR-WS.org

Windhager, F., Mayr, E., Schreder, G., Smuc, M., Federico, P. and Miksch, S. (2016). “Reframing Cultural Heritage collections in a visualization framework of space–time cubes”. Proceedings of the 3rd Histo-informatics workshop, vol 1632, 20-24.