# PRiSMHA (Providing Rich Semantic Metadata for Historical Archives)

Anna GOY[a,1], Rossana DAMIANO[a], Fabrizio LORETO[b], Diego MAGRO[a], Stefano MUSSO[b], Daniele RADICIONI[a], Cristina ACCORNERO[b], Davide COLLA[a], Antonio LIETO[a], Enrico MENSA[a], Marco ROVERA[a], Dunia ASTROLOGO[c], Bruno BONIOLO[c] and Matteo D'AMBROSIO[c]

[a] *Dip. di Informatica (Università di Torino, Italy)*
[b] *Dip. di Studi Storici (Università di Torino, Italy)*
[c] *Fondaz. Ist. Piemontese A. Gramsci (Torino, Italy)*

**Abstract.** In this paper we present the PRiSMHA project, whose main goal is to demonstrate that a rich semantic representation of the content of historical documents is useful - since it can significantly improve the access to archival resources - and sustainable - thanks to a crowdsourcing approach. This goal poses interesting research challenges, both for the semantic model definition and the user interaction. Such challenges range from the dialog between computer scientists and historians, to the design of an effective ontology-driven user interface; from the strategies to ensure the quality of the semantic metadata produced, to the application of Information Extraction techniques to support user annotation.

**Keywords.** historical events, ontologies, crowdsourcing, information extraction, cultural heritage



## 1. Introduction and Goal

For more than a decade, archive specialists have been aware of the huge gap between the resources contained in archives and the communities of users interested in them: "Organizations, even small, possess information worth millions, but cannot get it to the right people" (Making Information Work: the Dublin Core Way, 2007, DCMI). To address this gap, research in metadata has proposed flexible and richer schemata (such as Dublin Core), but only the advent of *semantic annotation* has brought about a change of paradigm in the description of resources. However, we think that in order to be effective, semantic annotation has to rely on a *rich semantic model*, enabling metadata to provide a *rich description* of the resource *content*. In fact, although the relative simplicity of currently available semantic models benefits processing, interoperability and sharing, it often prevents metadata to be actually useful.

A major flaw, however, still affects approaches based on rich semantic models, namely the difficulty of collecting the semantic annotations. The *crowdsourcing model* can contribute to overcome this obstacle by leveraging the work of users who do not share timing and location, through the mediation of specialized annotation platforms.

[1] Corresponding Author: Anna Goy, Dipartimento di Informatica, Università di Torino, C. Svizzera 185, 10149 Torino, Italia; E-mail: annamaria.goy@unito.it.

The PRiSMHA (Providing Rich Semantic Metadata for Historical Archives) project (lasting 2017-2019) was born in this perspective. It is a national project, funded by Compagnia di San Paolo and Università di Torino, involving the Computer Science and the Historical Studies Departments of the same university, and based on a close collaboration with the *Polo del '900* (www.polodel900.it), a cultural center headquartered in Torino (Italy) and co-funded by Compagnia di San Paolo, Comune di Torino and Regione Piemonte. It involves nineteen institutions engaged in research and cultural initiatives about the XX Century social, economic and political issues in Piemonte, and hosts a very rich set of documentary repositories. In particular, PRiSMHA will mainly rely on resources from the archives and library of the Fond. Ist. Piemontese A. Gramsci (www.gramscitorino.it), which – with its 2.5 km of documents – represents the major "contributor" (25%) of the Polo del '900 archives [10] (see also: www.gramscitorino.it/archivio.html, www.gramscitorino.it/biblioteca.html).

The main objective of PRiSMHA is to demonstrate how a rich semantic representation of the content of historical archival resources can both: **(a)** provide a significant enhancement in the possibilities of exploitation of archival resources; **(b)** be sustainable, as far as the overload imposed by knowledge acquisition (and the consequent bottleneck in the process) is concerned.

By "a rich semantic representation" we mean a formal representation based on established standards (e.g., OWL, RDF, Linked Open Data), grounded on well-founded reference ontologies (e.g., DOLCE; see Section 2), and supporting the detailed representation of historical events, including their location, temporal information, how the involved entities participate in them, and relations among events themselves. Such a representation would enable users, for example, to query historical archives about an event (e.g., the general strike on April 18, 1945 in Turin) and getting a set of references to pictures, texts, letters and historical accounts of the event, together with links to people and organizations involved in the strike; or to query archives about a certain typology of events (e.g., strikes) and getting also events belonging to related typologies (e.g., protests, marches, demonstrations).

The *sustainability* of the approach will rely on a *crowdsourcing collaborative model*, where experts and trusted users participate in the semantic annotation process. The interaction model will be driven by the underlying ontologies and supported with suggestions provided by automatic Information Extraction techniques. The project aims at demonstrating that the proposed crowdsourcing model overcomes the knowledge acquisition bottleneck thus making the overall approach sustainable.

In order to build the semantic model, we will couple the study of existing ontologies (see Section 2), with the analysis of published material, mainly books containing memoirs that narrate events related to the same domain (period and place) the selected archival resources refer to. Such kind of texts can also be objects of user annotation in the crowdsourcing platform, since they typically offer detailed information about the events (e.g., they narrate a strike) and thus they represent a valuable source for building the semantic representations that can be connected to the archival resources (e.g., pictures illustrating the same strike).

## 2. Background and Related Work

The PRiSMHA semantic model will be centered around the notion of (historical) *event*, which typically plays a major role in the historical domain [11], [19], [22]. Existing

ontologies defined and used in the Cultural Heritage domain will be analyzed: when possible, the PRiSMHA ontology will integrate (part of) them; when full integration is not possible, formal interoperability will be granted. The most important semantic model that will be taken into account is EDM (Europeana Data Model), defined within the Europeana framework [9]. Other ontological models are relevant, such as the EO [8], SEM [21], LODE [16], and CIDOC-CRM [4], just to mention a few. Some scientific projects could provide useful insights to PRiSMHA, too: for example (among many others) HOPE [15], that defines best practices for the Social History Domain, and Agora [1], funded by The Netherlands Organisation for Scientific Research. With respect to these projects, PRiSMHA aims at developing a system with a deeper "understanding" of the content of archival resources, thanks to a semantically richer formalization of the notion of *event*, grounded in foundational ontologies such as DOLCE [17], also supporting the representation of roles played by participants [12], [6]. A similar modeling attempt, though not explicitly tailored to historical events but to “narrative” ones, is described in [5].

The other relevant field whose state-of-the-art should be considered is represented by crowdsourcing projects in the Cultural Heritage domain. Crowdsourcing is becoming a fundamental approach for providing rich meta-data describing those resources that cannot be automatically processed (e.g., old pictures, handwritten documents), and many international Institution are starting crowdsourcing projects - e.g., The Library of Congress, The British National Archives, New York Public Library, among many others [2]. These projects can thus be used as a starting point for PRiSMHA: for example, Scribe [20] is an open-source framework enabling users to set up communities aimed at transcribing documents that cannot be successfully processed by OCR tools (e.g., handwritten texts). Scribe supports the management of workflows, tasks, and consensus. Micropasts [18] aims at providing a support for gathering quality meta-data over historical resources (e.g., the accurate location of photographed scenes, the identification of topics referred to in historical archives, the transcription of letters). A major challenge for PRiSMHA is the design of a workflow where the crowdsourcing model is integrated with the automatic and semi-automatic methodologies to obtain reliable, consistent descriptions of resources: the key to integration can be provided by formally specified, comprehensive meta-data schemata [14], [7] embedded in the crowdsourcing platform, whose role is to keep the contribution of each component aligned with the semantic model assumed by the project.

## 3. Research Issues and Challenges

PRiSMHA has been conceived as gathering cultural needs, objectives and technological efforts from different disciplines and approaches, in a *multidisciplinary* perspective. The research issues it will face include the definition of: **(a)** best practices, grounded in consolidated approaches of the historical research, aimed at supporting the specification of both the user requirements and the semantic model needed in the project; **(b)** a semantic model (ontology) tailored to the domain of historical events narrated by archival resources; **(c)** an innovative interaction model, grounded in collaborative approaches and Human-Computer Interaction best practices, in order to enable (selected) users to provide high-quality semantic metadata. These issues pose interesting challenges, which are presented in the following.

### *3.1. Challenges for the semantic model definition*

- **Definition of a rich and interoperable semantic model, emerging from the dialog with historians**. The definition of the semantic model should follow the best practices of the applied ontology community [13] and take into account existing ontologies. Moreover it should be based on domain expert knowledge and on the analysis of published material from libraries. In this perspective, a major challenge is the implementation of a fruitful dialog between computer scientists and historians, that starts from setting a common language and has the goal of defining (at least): the notions of event and sub-event, the issue of event granularity, the relations between events, the notion of participation. As emerged in [19] and [22], reaching an agreement between historians and computer scientists about these definitions is a particularly challenging – thus interesting and promising– issue.
- **Publication of good quality datasets**. Semantically enriched metadata produced by PRiSMHA should be published on the Linked Open Data (LOD) cloud, in order to make them largely accessible and to promote their innovative use (e.g., in applications in the tourism or education domain).

### *3.2. Challenges from the user interaction perspective*

- **Definition of user requirements of the crowdsourcing platform, based on the dialog with historians.** Again, a fruitful interaction between computer scientists and historians plays a major role in the identification of target users and in the definition of requirements of the crowdsourcing platform. The latter include, for example, the specification of the roles involved in the annotation processes, and the workflow required to collect and align the annotations in order to reach an acceptable agreement.
- **Ontology-driven User Interface.** The User Interface (UI) of the crowdsourcing platform will be driven by the underlying ontology. Given the claim that such an ontology should be semantically rich, the design and implementation of the UI requires a careful study to merge the requirements imposed by the ontology with those emerging from the platform users.
- **Quality of the semantic metadata produced.** Although we expect users of the PRiSMHA crowdsourcing platform to be selected and trusted, specific mechanisms should be designed and implemented to guarantee the quality of the semantic metadata produced. First steps in this direction are the selection of a pilot corpus and a set of data for testing the suitability of the semantic model and the design of the annotation workflow. Moreover, users themselves could be involved in the metadata quality assessment process.
- **Information Extraction approaches to support user annotation.** An important support to the user annotation activity could be provided by automatic techniques for Information Extraction, when full-text is available. As widely recognized [3], the automatic analysis of historical texts of the XX Century is particularly challenging and techniques aimed at obtaining good results in this domain are still an open research issue.

## 4. Conclusions

In this paper we presented the PRiSMHA project, by focusing on the main issues we are going to face, concerning the semantic model definition and the user interaction. Such challenges will be faced having in mind the goal of the project, i.e. demonstrating that a rich semantic representation of the content of historical documents can significantly improve the access to archival resources and can be sustainable, thanks to a crowdsourcing approach. We foresee that the results of the project will highlight the tight connection between the quality of the underlying ontology and the production of usable and useful semantic metadata, through user interfaces able to mirror the semantic model in an effective and user-friendly way. Moreover, the inclusion of an inference engine in the system architecture could be taken into account, in order to fully leverage the semantic metadata.

## References


[1] Agora: www.ghhpw.com/agora.php, accesses 16/7/2017.

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

[3] F. Boschetti, A. Cimino, F. Dell'Orletta, G.E. Lebani, L. Passaro, P. Picchi, G. Venturi, S. Montemagni, A. Lenci, Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II, in *LREC 2014 Workshop on Language resources and technologies for processing and linking historical documents and archives – Deploying Linked Open Data in Cultural Heritage* (2014), 70-75.

[4] CIDOC Conceptual Reference Model: www.cidoc-crm.org, accesses 16/7/2017.

[5] R. Damiano, A. Lieto, Ontological representations of narratives: a case study on stories and actions, *OASIcs-OpenAccess Series in Informatics*, **32**, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2013).

[6] M. Daquino, S. Peroni, F. Tomasi, F. Vitali, Political Roles ontology (PRoles): Enhancing archival authority records through semantic web technologies, *Procedia Computer Science*, **38** (2014), 60–67.

[7] E. Duval, W. Hodgins, S. Sutton, S. Weibel, Metadata principles and practicalities, *D-lib Magazine* **8(4)** (2002).

[8] Event Ontology: motools.sourceforge.net/event/event.html, accesses 16/7/2017.

[9] Europeana framework: www.europeana.eu/portal, accesses 16/7/2017.

[10] Fondazione Istituto Piemontese A. Gramsci, *Istituti Culturali di cui alla Legge 17.10.1996, n.534 (Art.8) - Scheda Descrittiva*. (2017).

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

[12] A. Goy, D. Magro, M. Rovera, 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, Heidelberg (2017), 123-126.

[13] N. Guarino, C. Welty, Evaluating ontological decisions with OntoClean, *Communications of the ACM*, **45(2)** (2002), 61-65

[14] R. Heery, P. Manjula, Application profiles: mixing and matching metadata schemas, *Ariadne* **25** (2000).

[15] Heritage Of the People's Europe: www.peoplesheritage.eu, accesses 16/7/2017.

[16] Linking Open Descriptions of Events: linkedevents.org/ontology, accesses 16/7/2017.

[17] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, WonderWeb Deliverable D18, Technical Report, CNR, 2003.

[18] Micropasts: crowdsourced.micropasts.org, accesses 16/7/2017.

[19] .F Nanni, S. P. Ponzetto, L. Dietz, Building Entity-Centric Event Collections, *Proc. 17th ACM/IEEE-CS on Joint Conference on Digital Libraries*, (2017), 199-208.

[20] Scribe: www.nypl.org/blog/2015/11/23/scribe-framework-community-transcription, accesses 16/7/2017.

[21] Simple Event Model: semanticweb.cs.vu.nl/2009/11/sem, accesses 16/7/2017.

[22] R. Sprugnoli, S. Tonelli, One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective, *Natural Language Engineering*, **23(4)** (2017), 485-506.