Ontologies and historical archives: A way to tell new stories

Applied Ontology 00 (2015) 1–8
IOS Press

Ontologies and historical archives: a way to tell new stories *

Anna Goy **, Diego Magro and Marco Rovera

Università di Torino, Dipartimento di Informatica, Corso Svizzera 185, 10149, Torino, ITALY

E-mail: {annamaria.goy, diego.magro, marco.rovera}@unito.it

Abstract. Historical documentary heritage has a high potential for supporting citizens' awareness about their culture and identity. However, to exploit this potential, access tools are needed, which integrate data from heterogeneous sources and provide an effective user interaction. Moreover, historical archives can become a useful and attractive cultural resource, if they are exploited in popular contexts, like education and tourism: innovative ICT-based applications can employ documents, pictures, etc. to guide students and tourists in the discovery of interesting stories, connecting the present to the past, and providing a larger audience with a "live" access to original cultural heritage resources. In particular, we claim that this scenario can be enabled by providing ICT tools with a rich semantic layer, based on computational ontologies, in which the notions of *event* and *role* play a major role. Such a semantic layer can be further enriched by exploiting resources available in the Linked Open Data cloud.

Keywords: ontology-based applications, ontology of events, cultural heritage, content-based access, historical archives

*POSTPRINT VERSION. Cite as: Anna Goy, Diego Magro and Marco Rovera, Ontologies and historical archives: a way to tell new stories, Applied Ontology 10(3-4) (2015), pp. 331-338, IOS Press, DOI 10.3233/AO-150152

**Corresponding Author: Anna Goy, Dipartimento di Informatica, Corso Svizzera 185, 10149, Torino, ITALY. E-mail: annamaria.goy@unito.it

1. Introduction

Documentary heritage about social, political, and cultural history has a high potential for creating and supporting the awareness of citizens about their history, culture, and ultimately their identity. This heritage is obviously highly relevant for students and researchers, social and political scientists and historians, but it can also be valuable for a larger and non-specialized audience, thanks to its capability of narrating social changes by means of documents where places, people and events are brought to life by pictures, newspaper articles, audiovisual clips, interviews, etc. This potential is enhanced by the increasing availability of digital catalogues and metadata, as well as by digitization processes, which ensure the availability of a huge amount of heterogeneous resources. However, such availability implies an increased need for effective and user friendly access tools, able to integrate data coming from different and remote archives and to orientate the audience in the extremely rich and heterogeneous universe of documents and testimonies.

Moreover, historical archives could be converted into a useful and attractive cultural resource by exploiting them in the context of everyday activities (e.g., education and tourism), where – thanks to the availability of suited interfaces (API) – innovative ICT-based (mobile) applications can be created, which employ documents, pictures, videos, etc. to guide people (e.g., students and tourists) in the discovery of interesting stories, interlinking the present with the past, thus providing a large and popular audience the opportunity to "play" with original resources belonging to our cultural heritage (see Section 2).

In this position paper, we claim that this is possible by providing ICT-based tools with a rich semantic layer, mainly based on computational ontologies (Guarino et al., 2009), in which the notions of *event* and *role* play a major role. In the following sections we will discuss this claim (see Section 3) and we will underline two important aspects: (a) The process of building the mentioned semantic layer can be supported by users themselves, but also by advanced automatic knowledge extraction techniques; (b) The Linked Open Data cloud (lod-cloud.net) could provide both a valuable source of data and a space where semantically enriched metadata are published, in order to gain visibility and popularity (see Section 4).

2. Scenarios

In this section we sketch a couple of scenarios, in order to concretely show what the proposed approach means for final users. In both scenarios, we will call "the App" a mobile application developed on the basis

of the approach presented in this paper, and "Federated Archives" an integrated set of historical archives. Moreover, we would like to stress that the sketched scenarios are fictitious, in the sense that, although plausible, they do not aim at reporting any actual knowledge stored in any real archive, but are simple outcomes of our imagination.

The tourist

Alice walks through Piazza Castello (Turin) and enters an ancient famous coffee bar. Thanks to the georeference feature, the App – by accessing the Federated Archives – tells her that AG, a famous Italian politician who lived in Turin at the beginning of the XX Century, was used to spend time in that bar. Alice accesses the info about AG provided by the App and finds that he was involved in a labor strike demonstration which took place in Piazza Castello at that time. The App provides some information about the event and also a link to a video which depicts the very same place at that time. After watching the video, Alice decides to further explore the information about the event, thus discovering the role played by AG, who was one of the protest leader; moreover Alice discovers that another participant in the event was AG's sister-in-law: focusing on this character, she can see that in the Federated Archive there are several pictures portraying AG with her. A couple of such pictures, together with some letters which show their relationships, are shown in a local museum, hosted in a building where AG had lived; Alice checks the opening hours and decides to visit it: the App guides her to the museum by showing an ancient map which appears as an overlay on top of the current city map: this further immerses Alice in the intriguing past story she is following.

The student

Young students, from a school which participates in a project for the enhancement of citizens' awareness about the local history, are walking around Turin. They can use the App to associate archival resources (e.g., documents, pictures) – retrieved from the Federated Archive – to specific places within the city. Their goal is to build interesting and appealing itineraries made up of a set of places in Turin linked to events and people belonging to the city history; such itineraries can be seen as stories about the city, where places are described by historical and social events that occurred there, together with people who participated in them and the role they played. Foreign students, visiting the school within a cultural exchange program, will be told such stories and they will be able to rate their experience: students from Turin getting the highest scores will

receive a prize: in this way, students are motivated to use the App and provide connections between places in the city and archival resources.

Roberto is walking through Piazza Carlina, and he sets a link between the NH Hotel and AG (a famous Italian politician who lived in Turin at the beginning of the XX Century): the hotel, in fact, is former AG's home. Moreover, the square has been the scene of the assassination of a famous bandit, PC: Roberto can thus link Piazza Carlina to that event, and to a set of documents available in the Federated Archive (some pictures of the involved people, among which the killer, which is tagged as such). At the end of his walk, Roberto ends up with a story, in which places in Turin are linked to historical events such as the assassination of the famous local bandit PC and his killer; thanks to the availability of information and pictures about these characters, the story that Roberto will be able to tell his foreign friends will be like a fascinating thriller.

3. The role of semantic technologies

With respect to the goal stated in Section 1, i.e., converting historical archives into a useful and attractive cultural resource exploitable in contexts like education and tourism, the integration of archive catalogues (i.e., resources metadata) is a prerequisite, and can be implemented with different degrees of interoperability. The first level is the integration of archival software systems, which can be achieved by implementing a Service-Oriented Architecture (SOA) (Alonso et al., 2004), acting as a *mediation layer*, following the *adapter pattern* (Gamma et al., 1995), in which each software component (i.e., each system used to manage an archive) interfaces with an *adapter* that communicates with the other SOA components according to the shared interfaces, and manages the interaction with the specific software system it wraps (see Fig. 1). The second level is the syntactic integration of metadata representation formats. The same SOA architecture could be exploited to address this problem: a common format, based on a shared standard, can be defined and the mapping between the shared format on the one side and the different formats used by each archive management system on the other side could be delegated to the adapters.

However, the most interesting aspect of the integration of metadata from different archives is faced at semantic level, since different archives can hold different "views" of the resources. For example, different archives may contain documents about Leonardo da Vinci, but they may use different expressions to refer to him ("Leonardo", "Leonardo di ser Piero da Vinci", "the painter of the Mona Lisa"), or they may classify

Leonardo into different categories (e.g. painter, inventor, artist), or they can link him to different contexts or events (i.e., the celebrations for the marriage between Gian Galeazzo Maria Sforza and Isabella d'Aragona, which Leonardo provided decorations for, or the project for the equestrian monument for Francesco Sforza).

An integrated and smart access to the archives should make the information accessible in a simple and flexible manner, in order to enhance the content richness by offering related resources, links or suggestions, based on the relationships between events, historical periods, organizations, texts, images, artworks, people and places.

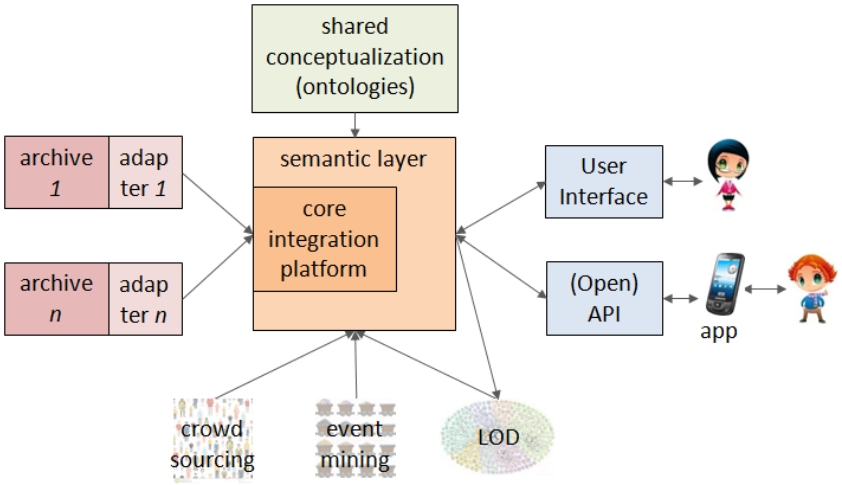

Fig. 1. Architecture.

To achieve semantic integration, a *semantic layer* should be defined and implemented (see Fig. 1), which takes archive metadata as input, links them to a *shared conceptualization* and enrich resource descriptions with new semantic connections among items; connected items can be archival resources (possibly belonging to different archives) or other real world entities, such as historical characters, places, battles, and so on, represented by means of the shared conceptualization. In this way, metadata, archive resources, and the historical knowledge they are related to become accessible and "understandable" to the integrated system (and thus ultimately to the end user). The shared conceptualization should be based on the integration of multiple computational ontologies into a unifying model, which explicitly represents the heterogeneous nature of the connections among archive resources and real world entities: events resources "talk about", people they refer to, roles played by the different actors in the identified events, and so on. W3C standards for the Web of Data, namely OWL 2 (Hitzler et al., 2012) and RDF(S) (Hayes & Patel-Schneider, 2014; Brickley & Guha, 2014), should be exploited in order to express the unifying semantic model.

The main claim of this position paper is to advocate the major importance of two concepts within the mentioned shared conceptualization, i.e., the notions of *event* and *role*. These concepts, in fact, enable the shared conceptualization to provide a structured representation of the domain, thus acting as a "glue" which connects contents stored in different archives to each other, and with real world entities such contents are related to. In particular, in the historical domain, events are a key concept to be modeled in order to achieve an effective representation of historical facts. Moreover, especially in this domain, the capability of distinguishing different types of events is essential: for example, there are events which are clearly limited in time and space or which have a limited and defined number of participants (e.g., the assassination of king Umberto I), while there are other events which have more fuzzy time and space boundaries, which do not have a clearly specified number of participants (e.g., the Second World War), and maybe can be decomposed into meaningful sub-events (e.g., the attack to Poland by Germany, the Italian armistice of September 8th 1943). Furthermore, there are relevant conditions, such as the illiteracy in Italy in 1945, which can be modeled as events, since they span over time, take place in space regions, involve people, may have effects, but they represent sorts of situations whose nature of "state" or "condition" make them rather different from the previously mentioned examples of events.

The mentioned distinctions can be modeled thanks to existing ontologies, such as DOLCE (Borgo & Masolo, 2009) – which provides a rich and well-grounded framework useful to this purpose – or more specific semantic resources, such as Europeana Data Model (Isaac, 2013), Event Ontology (purl.org/NET/c4dm/event.owl), LODE (Shaw et al., 2009), CIDOC-CRM (Le Boeuf et al., 2015) – which provide basic notions of event enabling the representation of "who does what when and where". CIDOC-CRM also accounts for some other characteristics of events and provides a characterization of the specific types of events involving cultural heritage (e.g., the transfer of custody of an item in a museum); moreover, besides being an ISO standard, CIDOC CRM has been successfully adopted in several EU-funded projects like PAPYRUS (www.ict-papyrus.eu), among others.

As we mentioned above, an effective formalization of the notion of event within our context should allow us to specify also the roles played by actors participating in events. Moreover, it should be possible to represent different (possibly mutually inconsistent) perspectives on (interpretations of) a same event. This is considered of paramount importance also in that it underpins Digital Hermeneutics (van den Akker et al.,

2011). To these purposes, the Simple Event Model (Hage et al., 2011), which has already been successfully applied in other projects – e.g., Agora and DIVE (van den Akker et al., 2010; de Boer, 2014) – can be taken into consideration, as well as the Event-Model-F (Scherp et al., 2009). As far as the representation of roles is concerned, useful insights could come from the PRoles ontology (Daquino et al., 2014), and an in depth foundational analysis of the concept of role can be found in (Masolo et al., 2004).

The rich connections provided by the identified (historical) events and roles played by participants can be exploited by specific applications as those roughly depicted in the scenarios presented in Section 2.

4. Opening the semantic layer: user-generated content, event-mining, API, and Linked Open Data

Opening the input of the semantic layer: user-generated content and event-mining.

One of the most challenging aspects of the sketched approach is metadata semantic enrichment, i.e. the process of linking archive metadata to the shared conceptualization and creating new semantic connections among archival resources and real world entities (places, people, events), that strengthen the resource network representing the semantic layer of the envisioned system. In fact, this process can be costly and time consuming. There are two main directions in which a support to this process can be found (as depicted in Fig. 1):

- Metadata semantic enrichment can be supported by users themselves: trusted users (e.g., domain experts, researchers, students) can be enabled to edit metadata in the semantic layer, in order to enrich the semantic characterization of the available resources.
- Event-mining techniques can be exploited: automatic detection and extraction of events from texts (van den Akker et al., 2010) can provide a (semi-)automatic source of semantic knowledge, that can be used to enrich the system semantic layer. Event-mining can exploit lexical and semantic resources such as WordNet (wordnet.princeton.edu) or BabelNet (babelnet.org), or even the data available in the open datasets of the Linked Open Data (LOD) cloud, together with Named Entity Recognition tools and other NLP techniques. Obviously, automatic event mining can be used only where full texts are available, i.e., where written documents have been digitized and support a good quality OCR output. Moreover, as done in the Agora project (van den Akker et al., 2010), a thesaurus of historical events could be extracted from trusted and good quality external resources: such a thesaurus could then be used to support manual

enrichment of metadata referring to resources such as multimedia or old handwritten letters, which automatic extraction techniques cannot be applied to.

Opening the output of the semantic layer: Open API.

Besides a user interface enabling final users to access archival resources by navigating integrated metadata, the system should provide a set of API (see Fig. 1), providing a programmatic access to the semantic layer, which can be used by developers to create new applications exploiting integrated archive catalogues in specific contexts, such as tourism or education, as sketched in the scenarios presented in Section 2.

Opening both input and output of the semantic layer towards Linked Open Data (LOD).

Thanks to the adoption of Linked Data best practices (Heath & Bizer, 2011) and standards – basically, interconnected RDF datasets accessible through the HTTP protocol or by means of more specific protocols and languages, such as SPARQL (Prud'hommeaux & Seaborne, 2008) – the system semantic layer can be linked to datasets available in the LOD cloud from two perspectives (shown in Fig. 1):

- LOD datasets can represent an input: in this perspective, LOD datasets are a very rich knowledge source, which can be connected to the system semantic layer in order to enrich it. Obviously, such a connection is not trivial, since issues concerning the mapping of the semantic model underlying external datasets onto the semantic model of the system have to be faced (Goy et al., 2015).
- The LOD cloud can represent a possible output: enriched metadata belonging to the semantic layer can be published as LOD, thus greatly improving archives visibility. To this purpose, interesting projects aimed at supporting the publication of metadata related to cultural heritage (e.g., museum catalogues) in the LOD cloud could be taken into account: see, for instance (Szekely et al., 2013), or more general frameworks, such as LOD2Stack (Auer et al., 2012).

5. Other interesting works to be taken into account

Besides the semantic resources already mentioned in Section 3, a lot of other works are relevant for the proposed approach. In the following we will mention a few of them, which we consider the most relevant, but the reader should be advised that they represent a small set of examples far from a complete and systematic survey.

"Classical" resources and models such as Functional Requirements for Bibliographic Records (IFLA, 1997), BibFrame (www.loc.gov/bibframe), Dublin Core (dublincore.org), and Simple Knowledge Organization System (Miles & Bechhofer, 2009) should be taken into account, mainly to guarantee interoperability with other systems. In this same direction, Europeana (www.europeana.eu) – by providing access to digitized cultural heritage of hundreds of European galleries, libraries, archives – represents a milestone: Europeana performs a semantic enrichment of metadata, through connections to external datasets (e.g., Geonames, DBpedia, the GEMET thesaurus), and a considerable portion of Europeana metadata are already available in the LOD cloud; see, for instance, the Europeana LOD Pilot (Haslhofe & Isaac, 2011).

In the last years, there is a remarkable interest in the application of Semantic Web technologies to history research both by Semantic Web researchers and historians, as documented by the survey reported in (Meroño-Peñuela et al., 2014). Many results considered in such a survey show the effectiveness of semantic formalisms (including ontologies), Linked Data best practices and NLP approaches in publishing and connecting historical datasets and in enhancing search, retrieval and classification. This holds, in particular, for archives and cultural heritage contents, a domain where semantic technologies and LOD principles are receiving more and more attention (Oomen & Belice, 2012).

6. Conclusions

In this paper we tried to show how semantic technologies, and computational ontologies in particular, can be useful in order to empower ICT tools with the knowledge needed to "understand" the content of historical archives. We claimed the importance of the concepts of *event* and *role*, and we sketched a possible exploitation of such a knowledge in order to bring the content of historical archives to a large audience and to connect it with everyday life, thus enhancing the citizens awareness about their history, their culture, and their identity.

Many issues have not been explicitly mentioned in the present paper, although they deserve attention. Among them it is worth mentioning intellectual property rights, which are a very important issue to be taken into account when working with historical material.

Acknowledgements

We would like to thank all the colleagues who participated in the discussions about the topics covered in this paper.

References

van den Akker, C., Aroyo, L., Cybulska1, A., van Erp, M., Gorgels, P., Hollink, L., Jager, C., van der Meij, L., Oomen, J., van Ossenbruggen, J., Schreiber, G., Segers, R., Vossen, P., & Wielinga, B. (2010). Historical Event-based Access to Museum Collections. Proc. 1st Int. Workshop on Recognising and Tracking Events on the Web and in Real Life (EVENTS 2010).

van den Akker, C., Legêne, S., van Erp, M., Aroyo, L., Segers, R., van der Meij, L., van Ossenbruggen, J., Schreiber, G., Wielinga, B., Oomen, J., & Jacobs, G. (2011). Digital hermeneutics: Agora and the online understanding of cultural heritage. Proc. 3rd Int. Web Science Conference (WebSci 2011).

Alonso, G., Casati, F., Kuno, H., & Machiraju, V. (2004). Web Services - Concepts, Architectures and Applications. Springer.

Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P. N., Van Nuffelen, B., Stadler, C., Tramp, S., Williams, H. (2012). Managing the Life-Cycle of Linked Data with the LOD2 Stack. Proc. 11th Int. Semantic Web Conference (ISWC 2012), Part II, 1-16.

de Boer, V., Oomen, J., Inel, O., Aroyo, L., van Staveren, E., Helmich, W., & de Beurs, D. (2014). DIVE into the Event-Based Browsing of Linked Historical Media. Proc. 13th Int. Semantic Web Conference (ISWC 2014).

Borgo, S., & Masolo, C. (2009). Foundational Choices in DOLCE. In S. Staab, R. Studer (Eds.). Handbook on Ontologies - Second Edition, Springer, 361-381.

Brickley, D., & Guha, R.V. (Eds.) (2014). RDF Schema 1.1. W3C.

Daquino, M., Peroni, S., Tomasi, F., & Vitali, F. (2014). Political Roles Ontology (PRoles): Enhancing Archival Authority Records through Semantic Web Technologies. Proc. 10th Italian Research Conference on Digital Libraries, in Procedia Computer Science, 38, 60-67.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Longman Publishing Co.

Goy, A., Magro, D., Petrone, G., Rovera, M., and Segnan, M. (2015). A Semantic Framework to Enrich Collaborative Tables with Domain Knowledge. Proc. Int. Conf. on Knowledge Management and Information Sharing (KMIS 2015).

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology?. In S. Staab, R. Studer (Eds.). Handbook on Ontologies - 2nd Edition, Springer, 1-17.

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

Haslhofe, B., & Isaac, A. (2011). data.europeana.eu - The Europeana Linked Open Data Pilot. Proc. Int. Conf. on Dublin Core and Metadata Applications.

Hayes, P. J., & Patel-Schneider, P. F. (Eds.) (2014). RDF 1.1 Semantics. W3C.

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (Eds.) (2012). OWL 2 Web Ontology Language Primer (Second Edition). W3C.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1997). Functional Requirements for Bibliographic Records. IFLA.

Isaac A. (Ed.) (2013). Europeana Data Model Primer. Creative Commons Licence.

Le Boeuf, P., Doerr, M., Ore, C. E., & Stead, S. (Eds.) (2015). Definition of the CIDOC Conceptual Reference Model (Version6.1). ICOM/CIDOC CRM Special Interest Group.

Masolo C., Vieu L. R., Bottazzi E., Catenacci C., Ferrario R., Gangemi A., & Guarino N. (2004). Social Roles and their Descriptions. In D. Dubois, C. Welty, M.A. Williams (Eds.). Principles of Knowledge Representation and Reasoning, AAAI Press, 267 - 277.

Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., & van Harmelen, F. (2014). Semantic Technologies for Historical Research: A Survey. Semantic Web Journal, 1-27.

Miles, A., & Bechhofer, S. (Eds.) (2009). SKOS Simple Knowledge Organization System Reference. W3C.

Oomen, J., & Belice, L. (2012). Sharing cultural heritage the linked open data way: why you should sign up. Proc. Conf. Museums and the Web.

Prud'hommeaux, E., & Seaborne, A. (Eds.) (2008). SPARQL Query Language for RDF. W3C.

Shaw, R., Troncy, R., & Hardman, L. (2009). LODE: Linking Open Descriptions of Events. Proc. 4th Asian Conf. on The Semantic Web (ASWC 2009), 153-167.

Scherp, A., Franz, T., Saathoff, C., & Staab, S. (2009). F--a model of events based on the foundational ontology DOLCE+DnS Ultralite. Proc. 5th Int. Conference on Knowledge Capture (K-CAP 2009), 137-144.

Szekely, P., Knoblock, C. A., Yang, F., Zhu, X., Fink, E. E., Allen, R., & Goodlander, G. (2013). Connecting the smithsonian american art museum to the linked data cloud. Proc. 10th Int. Conference ESWC 2013 – The Semantic Web: Semantics and Big Data, 593-607.