Semantic Data Lakes for Knowledge Extraction in the Humanities: A Case Study on Bernard Berenson's Network of Acquaintances

Spinaci, Gianmarco
gspinaci@itatti.harvard.edu
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Grillo, Remo
grilloremo@gmail.com
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Klic, Lukas
lklic@itatti.harvard.edu
I Tatti - The Harvard University Center for Italian Renaissance Studies, Italy

Bonora, Paolo
paolo.bonora@unibo.it
Alma Mater Studiorum - Università di Bologna

Problem description and research question:

It is frequent practice in Digital Humanities (DH) studies to create Knowledge Bases (KB) limited to specific domains of interest. This leads to the creation of a plethora of highly specialized data silos [1]. As a result, the extraction of knowledge dispersed across KBs can be challenging. We argue that an extended KB — a Semantic Data Lake [2] (SDL) — obtained by aggregating heterogeneous vertical KBs could help define a set of heuristics to support transversal Knowledge Extraction [3]. We propose a case study based on a set of vertical KBs with converging content based on the correspondence of Bernard Berenson (1909-1960), the diaries of Mary Berenson (1879-1935), and their personal photographic archive.

As these semantic silos contain converging content, we aim to demonstrate that their union could support the discovery of original information. Through inductive reasoning, we aim to analyze data looking for graphs in order to assess the likelihood of a relationship among two or more actors. We aim to reconstruct a network of acquaintances by analyzing these paths within a consolidated dataset. If the proposal will be effective in identifying and qualifying the relationships among actors in the Berenson Circle, Art historians and domain experts will evaluate the relevance of these relationships.

Methodology:

The data lake-based approach does not require a prior alignment of different ontologies within KBs. In other words, sources do not necessarily need to use the same ontological framework or re-use the same modelling patterns, as long as each source is modeled consistently. KBs are then selected exclusively on the basis of their contents. The only prerequisite is that they are represented in RDF. In order to take into account a broader set of available KBs, we adopted an ontology-agnostic methodology. Knowledge Extraction begins with the analysis of the structure of the paths connecting target entities, in our case Berenson's acquaintances. This requires a reconciliation of actor identifiers and harmonized space/time dimensions. Regarding the entities that have not been already qualified, we used a NER algorithm to extract them and, together with qualified entities, we used Wikidata for a semi-supervised reconciliation process.

Paths are then identifiable through the data lake's composite graph, both from a quantitative (i.e. shortest paths) and a qualitative (i.e. semantics of the resulting paths' graph patterns) approach. Among this set of identified paths, the most relevant ones will be selected for answering the research question.

The extraction process will be organized into the following steps:

1. The SDL is built from source KBs partitioned within their own named graph
2. Reconciliation of target entities (i.e. people);
3. Harmonization of space/time coordinates (i.e. toponyms disambiguation and georeferencing);
4. Extraction of paths between instances of people;
5. Analysis and selection of paths that allow for relationships among people to be inferred;
6. Formalization of these paths as SPARQL Property Paths or SWRL rules;
7. Validation of extracted paths across KBs;

Sources:

This case study is designed around a strong convergence of sources with a known set of interactions between actors across space and time through common references to events, people, and places. Sources comprise letters exchanged between Berenson and Yukio Yashiro [4] (115 texts written from 1925 to 1960), letters sent from Belle Da Costa Greene to Berenson (470 texts written from 1909 to 1949), diaries of Mary Berenson (30 annual diaries written from 1879 to 1935), and metadata from historical photographs housed in the Berenson archive. We are focusing on the 1925 to 1935 period when we have the best overlapping of the corpora.

Expected results and validation:

In the correspondence between Yukio Yashiro and Bernard Berenson, numerous references are made regarding meetings with unspecified art historians. Diary entries by Mary Berenson report the guests at their residence for most days of the year, including Yashiro. By cross-referencing these sources, we can reconstruct a network of acquaintances between Yashiro, the Bereson's, and others. The numbers of extracted entities are: more than 1k names tagged as persons of which 20 qualified, and almost 300 toponyms of which 150 qualified. Once this information has been extracted through the proposed methodology and formalized in new assertions, new validation criteria should be adopted in order to refine the data quality and assess their effectiveness. These criteria need to consider metrics such as frequency count, recall and precision, and accuracy analysis by reconciliation with sources.

Conclusions:

We presented a SDL methodology that allows for a lightly supervised, model agnostic, data-driven approach to Knowledge Extraction from heterogeneous KBs. We expect that the experimentation performed on Berenson's network of acquaintances will help determine the feasibility of Knowledge Extraction from highly coherent and focused SDLs. This would motivate a further experimentation of this methodology with a broader scope and less converging sources. Moreover, it would help define a quality assessment framework for the extracted information.

Bibliography

[1] Nichols, Stephen G. "Time to Change Our Thinking: Dismantling the Silo Model of Digital Scholarship." Ariadne, no. 58 (30 January 2009)http://www.ariadne.ac.uk/issue58/nichols/

[2] Dibowski, Henrik, et al. "Using Knowledge Graphs to Manage a Data Lake", 2021.

[3] Darmont, Jérôme, et al. "Data lakes for digital humanities." Proceedings of the 2nd International Conference on Digital Tools & Uses Congress, October 15, 2020, 1–4.

[4] Takagishi, Akira. "A Twentieth-Century Dream with a Twenty-First -Century Outlook: Yashiro Yukio, a Japanese Historian of Western Art, and His Conception of Institutions for the Study of East Asian Art," in Asian Art in the Twenty-First Century, Williamstown (Mass.): Sterling and Francine Clark Institute 2007, 138–48.

Commercial crowdsourcing in digital humanities: prospects and ethical issues

Suviranta, Rosa
rosa.suviranta@helsinki.fi
University of Helsinki, Finland

Hiippala, Tuomo
tuomo.hiippala@helsinki.fi
University of Helsinki, Finland

This presentation discusses key issues in using commercial crowdsourcing in digital humanities. Traditionally, digital humanities have engaged volunteers for tasks like digitising and organising information (Dunn and Hedges, 2013; Carletti et al., 2013). However, not all fields in digital humanities can benefit from volunteer-based crowdsourcing. I argue that commercial crowdsourcing is a viable alternative for fields that cannot attract volunteers, provided that crowdsourcing is used in an ethically sustainable way. To do so, I propose solutions to a range of ethical issues related to fair pay and hidden labour on commercial crowdsourcing platforms. I also discuss linguistic and epistemic asymmetries between task requesters and the global crowdsourced workers and argue for the need to develop crowdsourcing methods that balance the needs of both ethics and data quality. To this end, I draw on examples from an ongoing project that uses crowdsourcing to create multimodal corpora and show how a combination of pedagogically motivated training, paid exams and multimodal instructions can mitigate these issues.

Crowdsourcing is a participatory method in which an individual, an institution, organisation or company can request an undefined group of workers with varying knowledge and number to perform a task through an open call (Carletti et al., 2013: 1–2). Crowdsourcing can take place on commercial and non-commercial platforms, and the tasks can range from data labelling to content creation (Dunn and Hedges, 2013). How crowdsourcing is understood depends on the field of research. Computer vision, for example, uses crowdsourcing to create training data for algorithms by decomposing complex tasks into piece-meal work and distributing this effort among paid non-expert workers on online platforms (Kovashka et al., 2016).

In digital humanities, crowdsourcing is often associated with the galleries, libraries, archives and museums (GLAM)