Linked Data for the Humanities: methods and techniques

Tutorial held in conjunction with DH2019 - Utrecht, 8th July, 2019.

Background and motivation

The primary goal of the digital humanities vision is the possibility to develop scholar studies at a large scale, where the sources of investigation can be reused among studies and arguments and methods to be developed and compared in a rigorous way. However, the underlying assumption is that sufficient data are available and that these data are properly described and distributed to the research community [Schöch, 2013].

The Linked Open Data (LOD) movement started within the Semantic Web (SW) research community with the objective of building a *Web of Data* [Heath, 2011] [Bizer, 2011]. So far, the impact of Linked Data in the Library and Cultural Heritage domain has been significant and testified by large scale efforts such as the one of Europeana [Haslhofer, 2011]. Related to that, LD has proven to be the framework for Open Science, linking knowledge with FAIR principles [Wilkinson, 2016] electing the Web URI as the natural method to *cite everything*. Indeed, Linked Data (LD) seems to match the Digital Humanities (DH) agenda perfectly. However, at a closer look, the impact of Semantic Web research on the Humanities has been discontinuous.

On one hand, digital humanities projects emerged having data as focal point of the research and as enabler of novel approaches to humanistic enquiry. Notable examples are the LED Project [Adamou, 2014], drawing on LOD published by The Open University [Daga, 2016], and Pelagios [Isaksen, 2014], aiming at building a hub of geospatial data about the ancient world. However, the potential for data reuse seems to remain largely unexploited. In many DH projects, data is linked through user interfaces as hypertexts with minimal support for RDF and SPARQL. Too often, the *data* side of LD is left out as future work.

On the other hand, Semantic Web researchers contribute tools to support scholars in dealing with LD (e.g. [Adamou, 2014] [Hyvönen, 2012] [Hoekstra, 2016], [Lodi, 2017]). The variety of use cases and endeavours is well testified in the proceedings of the WHiSe workshop [Adamou, 2016], reaching the third edition this year.[1] These span from cultural heritage [Bojars, 2016], historical datasets [Hoekstra, 2016], biographical data [Leskinen, 2017], the ancient world, music [Daquino, 2017] and musicology [Nurmikko-Fuller, 2016], on reflecting about methodological aspects [De Roure, 2016].

[1] Workshop on Humanities in the Semantic Web (WHiSe): http://whise.kmi.open.ac.uk/

Indeed, Humanities can be a land of opportunities for SW researchers, a space where the Semantic Web vision can be tested by challenging requirements coming from sophisticated, highly specialised domains of enquiry.
However, the interaction between the two communities has been occasional and as a result foundational techniques and methods developed by the SW community are still perceived as esoteric by many DH practitioners. In addition, more recent approaches have not been disseminated yet in the DH community and could contribute at enriching the toolkit available for leveraging Linked Data, from supporting the knowledge extraction phase (e.g. combining language processing with ontology engineering and deep learning) [Alam, 2017], to effectively build applications on top of SPARQL endpoints [Daga, 2015].
Adding semantics to DH has multiple effects: as an enabling technology (as exemplified above), as a vehicle to grow DH data as sharable and interoperable assets, and as an experimental platform to make innovative research in computer science by reverse engineering humanistic methods.
A plethora of use cases are emerging that can benefit of practical semantic platforms such as LOD and Knowledge Graphs (KG). A cartography of such cases can be built within the following areas (in an increasing degree of hybridisation between humanistic and computational techniques):

1. use or adaptation of computational tools for the collection, display or use of databases, corpora, collections;
2. creation of new exploration and discovery methods in humanities (both within works and collections, and in virtual reconstructions of places or artefacts);
3. implementation of software components supporting humanities;
4. creation of new knowledge (e.g., extraction of patterns from large corpora as in distant reading, availability of aggregated data for particular communities);
5. creation of new hybrid methods (e.g. cognitive computing, ontological engineering, defeasible logic, embodied or grounded semantics);
6. creative use of hybridisation to generate new works or collections (literary chatbots, painting morphisms).

This classification has been built out of a survey of DH projects from 40 academic centers, and shows the variety of data-oriented problems, approaches, and creative solutions: a “happy chaos” that can benefit from semantic methods.
With the above in mind, we propose a half-day tutorial on LD methods and techniques, having the following objectives:

- To present the theoretical and technical foundations of Linked Data and introduce basic methods for data production and publishing to students, researchers, and practitioners.

- To provide a reference collection of reusable tools to boost an effective adoption of LD in DH projects.
- To showcase a set of innovative methods for extracting and linking data from texts.

Contents and target audience

The tutorial will be organised in three sessions: 1) Linked Data in a nutshell; 2) Producing and consuming Linked Data; 3) Hybrid methods for working with texts and LD. The content will be tuned to accommodate a wide range of participants, spanning from the student that is curious to hear more about LD to the humanist hacker that looks for a robust toolkit to apply to her research. One output of the tutorial will be an openly accessible and persistent registry of reusable resources for developing linked humanities applications.[2]

Linked Data in a nutshell

This first session offers an overview of the basic notions and technologies behind Linked Data, from the adoption of URIs, the Resource Description Framework (RDF), Semantic Web ontologies, and SPARQL, to their use for representing data as a KG. A catalogue of exemplary LOD sources will be used in the hands-on session.

Producing and Consuming Linked Data

The second session will be dedicated to illustrate a number of case studies from humanities domains such as library science, art, and musicology. The objective of this session is to show how Linked Data works “Under the hood” and to enable participants to orient themselves when deciding to start a project based on Linked Data. The content includes basic recipes for generating RDF from pre-existing data sources such as Excel/CSV files or Relational Databases, for reusing existing vocabularies and ontology design patterns (ODP) [Gangemi, 2010], for discovering links with pre-existing third-party datasets, and for publishing the result as 5-star Linked Data.[3] We will showcase the standard toolkit for consuming linked data as well as more advanced approaches that make it easier for developers to interact with SPARQL endpoints (e.g. BASIL [Daga, 2015]).

[2] http://purl.org/ld4dh/dh2019
[3] https://5stardata.info/en/

Hybrid methods for working with texts and LD

The last session is dedicated to hybrid methods for the identification, extraction, and linkage of data from texts. We will showcase methods and tools that leverage Semantic Technologies for discovering entities in texts (Entity Linking), and for automatically creating knowledge graphs from texts (Knowledge Extraction), and how LD can be exploited for characterizing the content of texts for similarity analysis and content recommendation. Off-the-shelf tools such as DBpedia Spotlight [Mendes, 2011], Tagme [Ferragina, 2014] and FRED [Gangemi, 2017] will be introduced to showcase the potential of hybrid methods.

About the Organisers

Enrico Daga

Enrico Daga has a PhD in Artificial Intelligence and has carried out research on Web Semantics and Ontology Engineering since 2006, first at the Italian National Research Council (CNR) and then at the Knowledge Media Institute of The Open University in the UK, where he leads the OU Linked Data initiative (data.open.ac.uk). He has played key roles in R&D projects related to the development of intelligent systems for Ontology Engineering (NeOn) and Smart Cities (MK:Smart). Currently, he is exploring the application of knowledge-based methods to support scholarship in the humanities (e.g. the LED project). A former student of Music and Performing Arts (University RomaTRE), Enrico is founder and chair of the WHiSe Workshop on Humanities in the Semantic Web.

Aldo Gangemi

Aldo Gangemi is full professor at University of Bologna, and associate researcher at Italian National Research Council, Rome. He has co-founded the Semantic Technology Lab (STLab) at ISTC-CNR. His research focuses on Semantic Technologies as an integration of methods from Knowledge Engineering, the Semantic Web, Linked Data, Cognitive Science, and Natural Language Processing. He has worked in many different domains, including Cultural Heritage, where he has designed ontologies and linked data (Luoghi della Cultura, ICCD's ArCo) for the Italian Ministry of Cultural Heritage. He has published more than 250 papers in international peer-reviewed journals, conferences and books, and seats as EB member of international journals. He has worked in several EU projects related to LOD such as WonderWeb, Metokis, NeOn, and IKS.

References

[Adamou, 2014] Adamou, Alessandro, et al. "LED: curated and crowdsourced linked data on music listening experiences." (2014): 93-96.

[Adamou, 2016] Adamou, Alessandro, Enrico Daga, and Leif Isaksen. "WHiSe 2016-Humanities in the Semantic Web." (2016).

[Alam, 2017] M. Alam, D. Reforgiato Recupero, M. Mongiovi, A. Gangemi, P. Ristoski. Event-based knowledge reconciliation using frame embeddings and frame similarity. Knowledge-Based Systems, 135, 192-203, Elsevier, 2017.

[Bizer, 2011] Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data: The story so far." Semantic services, interoperability and web applications: emerging concepts. IGI Global, 2011. 205-227.

[Bojars, 2016] Bojars, Uldis. "Case study: towards a linked digital collection of Latvian cultural heritage." WHiSe@ ESWC. 2016.

[Daga, 2015] Daga, Enrico, Luca Panziera, and Carlos Pedrinaci. "A BASILar approach for building web APIs on top of SPARQL endpoints." CEUR Workshop Proceedings. Vol. 1359. 2015.

[Daga, 2016] Daga, Enrico, et al. "The open university linked data–data. open. ac. uk." Semantic Web 7.2 (2016): 183-191.

[Daquino, 2017] Daquino, Marilena, et al. "Characterizing the Landscape of Musical Data on the Web: State of the art and challenges." (2017).

[De Roure, 2016] De Roure, David, Pip Willcox, and Alfie Abdul-Rahman. "On the description of process in digital scholarship." WHiSe@ ESWC. 2016.

[Ferragina, 2014] Francesco Piccinno, Paolo Ferragina: From TagME to WAT: a new entity annotator. ERD@SIGIR 2014: 55-62.

[Gangemi, 2010] A Gangemi, V Presutti: Ontology Design Patterns. In S. Staab: Handbook on Ontologies (2010).

[Gangemi, 2017] Gangemi, Aldo, Presutti, Valentina, Reforgiato Recupero, Diego, Nuzzolese, Andrea Giovanni, Draicchio, Francesco, Mongiovì, Misael: Semantic Web Machine Reading with FRED. Semantic Web 8(6): 873-893 (2017).

[Heath, 2011] Heath, Tom, and Christian Bizer. "Linked data: Evolving the web into a global data space." Synthesis lectures on the semantic web: theory and technology 1.1 (2011): 1-136.

[Hoekstra, 2016] Hoekstra, Rinke, et al. "An ecosystem for linked humanities data." International Semantic Web Conference. Springer, Cham, 2016.

[Haslhofer, 2011] Haslhofer, Bernhard, and Antoine Isaac. "data. europeana. eu-The Europeana Linked Open Data Pilot." (2011).

[Hyvönen, 2012] Hyvönen, Eero. "Publishing and using cultural heritage linked data on the semantic web." Synthesis Lectures on the Semantic Web: Theory and Technology 2.1 (2012): 1-159.

[Isaksen, 2014] Isaksen, Leif, et al. "Pelagios and the emerging graph of ancient world data." Proceedings of the 2014 ACM conference on Web science. ACM, 2014.

[Leskinen, 2017] Leskinen, Petri, et al. "An ontology and data infrastructure for publishing and using biographical linked data." Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). CEUR Workshop Proceedings (October 2017). 2017.

[Lodi, 2017] G. Lodi, V. Presutti, L. Asprino, A.G. Nuzzolese, D. Reforgiato, A. Gangemi, A. Orsini, C. Veninata. Semantic Web for Cultural Heritage Valorisation, in S. Hai-Jew (ed.), Data Analytics in Digital Humanities, Springer, Berlin, DOI: 10.1007/978-3-319-54499-1_1, 2017.

[Mendes, 2011] Mendes, Pablo N., et al. "DBpedia spotlight: shedding light on the web of documents." Proceedings of the 7th international conference on semantic systems. ACM, 2011.

[Nurmikko-Fuller, 2016] Nurmikko-Fuller, Terhi, and Kevin R. Page. "A linked research network that is Transforming Musicology." WHiSe@ ESWC. 2016.
[Schöch, 2013] Schöch, Christof. "Big? smart? clean? messy? Data in the humanities." Journal of Digital Humanities 2.3 (2013): 2-13.
[Volz, 2009] Volz, Julius, et al. "Silk-a link discovery framework for the web of data." LDOW 538 (2009).
[Wilkinson, 2016] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3 (2016).