The application of Semantic Publishing technologies in the Science of Science research domain

Ph.D Research project in Data Science and Computation, 2018, University of Bologna, Bologna

Ivan Heibi,
ivan.heibi2@unibo.it, ivan.heibi@studio.unibo.it

State of the art review

Understanding, quantifying and predicting the scientific researches and the resulting outcomes is the main object of the Science of Science study (SOS) [16]. This field of study has attracted the attention of scholars from different backgrounds, since this issue is related to almost all the scholarly domains. Improvements in SOS will lead to better solutions to many challenging problems. A particular important emphasis in SOS, should be also given to the multiplex structure, dynamics and evolution mechanisms of the entire scholarly literature, since these are the factors that have been less studied in the last years [1].
The advent of the Web lead to a much higher availability of unstructured, semi-structured, and structured data that need to be converted into knowledge formalised with appropriate tools so as to be understandable also to machines and interlinked in a way which is easier to move from one concept to another.
In the past 20 years, the Semantic Web technologies [17] have been used by humans so as to enable the definition of such machine-readable knowledge on the Web.
Taking in consideration SOS, these technologies can turn out to be very beneficial. A possible application of the Semantic Web technologies in the scholarly publishing is possible, and is highly used in different study domains [18]. Using these technologies potentials, could have a very significant impact in revolutionizing the management of the sparse scientific knowledge we have in literature [2], for instance, enriching articles with appropriate metadata open to automated processing and analysis, allowing enhanced verifiability of published information and providing the capacity for automated discovery and summarization. This specific application of the Semantic web, is also known as Semantic Publishing, which concern “the use of Web and Semantic Web technologies and standards for enhancing a scholarly work semantically so as to improve its discoverability, interactivity, openness and reusability for both humans and machines” [2].
The publishers, editors and authors all have primary roles and responsibilities in making the semantic publishing a successful mechanism which will bring substantial benefits into the

scholarly community. The semantic publishing technologies, and the positive impacts and improvements that the academic community will gain from adapting the semantic publishing politics, is well reviewed and documented in previous works like [2] [3] and [14].
On the last years, several projects like OpenAIRE [4] and OpenCitations [5] have used semantic publishing technologies for this purpose. Both these initiatives have emphasized the importance of having a solid open science data available to the researching community – where “open science” can be defined as an umbrella concept that embraces the ideas of different open movement such as open source, open access and open data, while embracing trends of open distributed collaboration, data-intensive science and citizen science. Adopting an open science approach will lead to an easier spread of knowledge, information and data that will more generally benefit a sustained progress in society [26].
OpenCitations represents a very good example of the use of Semantic Publishing technologies so as to create an open repository of scholarly citation data. A citation, links one document to another document. Citations and citation analysis processing have an essential part, as also described in [2] and [22]. The current instantiation of the OpenCitations Corpus (OCC), is hosted by the University of Bologna, and has ingested, processing and publishing bibliographic metadata and citation links from several scholarly articles available in the Open Access subset of PubMed Central.
In addition, these technologies are getting more popular also in the scientific publishing industry, e.g. SciGraph by Springer Nature [6] and the SCAR project [7] – which is a collaboration between the University of Bologna and Elsevier.
However, the studies done by all the aforementioned projects focus mainly on scientific disciplines (e.g. computer science, biology, chemistry, or medicine), and do not take into account other domains, such as the Social Sciences and the Humanities.
Social science is a notable domain example that has not been treated enough with semantic publishing technologies in the last years [21]. This fact might be due to the complexity of the related papers and materials of these fields, we therefore have a less structured format of the inner textual contents and a much more convoluted linguistic form.
An interesting thesis about the use of Semantic Publishing technologies in the Humanities [8] talk about the few efforts made for the integration of Semantic Publishing technologies in the social sciences.
A particular sub-branch of the Humanities, i.e. Digital Humanities, has recently started to work actively on adopting Semantic Publishing technologies. For instance, one of the few good examples introduces knowledge engineering models and tools to manage the digital scholarly publishing of manuscripts [9]. Other important steps have been made to concern and teach the semantic publishing in digital humanities, like the “Linked Data for Digital Humanities” [12], a workshop of the Oxford university which is aimed at any and all learners interested in gaining skills and expertise in the Linked Data publication paradigm in this domain.
Considering the potentials of the semantic publishing technologies, when applied to this domain, can gain important benefits on the future of the humanities community. Therefore, the Social Science and Humanities domains are a highly interesting fields to work with for a larger integration of semantic publishing technologies, since few studies have made it in the past [21].
A general study of some novel semantic publishing applications on this particular domain could possibly reveal interesting new patterns and dynamics still hidden to the scholarly

community. The overall need for new elaboration strategies regard social science and humanities materials are noticeable. For instance the Emilia Romagna region released some new financial supports for projects that aim on developing novel technologies, e.g: big data analysis and semantic technologies, on the elaborations of these specific domains [10].
The above mentioned issues and research themes involve also a much more general study of the citation function inside different contexts and scientific domains [22][23]. A citation function tries to answer the question "why this resource is getting cited?". An automated technique to detect and answer this question involve a large study of the characteristics and parameters that could possibly have an effect on this decision, together with establishing how results could possibly change when applied to different study domains.

Project description

Semantic Publishing in SOS can have different facets and a variety of applications [2], and we have a significant large number of possible strategies to follow. Therefore I would like to concentrate my work on some specific sub detailed research questions:

- Is there a way to formally categorize a common citing behaviour of authors belonging to a specific domain of study, through the application of semantic publishing technologies? In addition, which are the features – e.g. the citations context, the average number of citations for each paper, the number and general structure of the sections, or the citation function – that best define such “citing behaviour”?.
- How the way (i.e. the reasons why) a scholarly contribution is cited changes in time, when taking in consideration different initial constant parameters? – such as the number of citations made to external study-domain papers inside computer science works, compared to 10 years ago. A set of well defined considerations like this, can reveal different publishing attitudes and their impact in the scientific publishing evolution.
- What will happen to the previous research questions in case we take in consideration social science and humanities domain? Exploring and working with these domains can bring to light a large number of interesting intra- and inter-domain dynamics. Observing the citations can reveal interesting logical connections with other study domains, e.g. a social science paper talking about political issues might have a large number citations to statistical processing methods.

Since my work will be based on Semantic Publishing, and Semantic Web technologies, this will involve a general consideration of several components and concepts. During my Ph.D work I will analyse Semantic Web methodologies and different related aspects:
1) The data representations [24]: the usage of ontologies (a conceptual schemas for describing data) [18] to define vocabularies, concepts and relationships to describe and represent an area of concern. W3C offers a number of technologies to define different forms of vocabularies in standard formats: RDF (the data model used for expressing semantic data on the Web) and RDF Schemas, Simple Knowledge Organization System (SKOS, an ontology for defining taxonomies), Web Ontology Language (OWL, the language for defining ontologies on the Web).

2) Data querying: SPARQL is the main query language used on RDF data. In particular SPARQL implements a mechanism for querying RDF data similar to what SQL provides for common databases.
3) Data visualization: the analysis of techniques which answer the question “how to visualize graph based data?”. This part is crucial especially when trying to expose information to other users [25].
An important goal to be aware of while working on this Ph.D project, is to make sure that the large collections of scholarly data, information and knowledge used in the research will be available for a further use by other researchers of the community. In addition, I envisage that the outcomes of my project may have a positive impact also in the publishing industry.
My plans are to work with Social Science and Humanities data collections and try answering my research questions by using and extending them and by developing novel services. A notable interesting service which works with Social Science and Humanities data and makes it available, is ISIDORE [13], a platform and search engine which allows the access to such type of data openly to all, and especially to teachers, researchers, Ph.D students and other students, that relies with enrichments on the principles of semantic web and that provides access to data in open access. ISIDORE proposes more than 5 millions of ressources of the whole world. Therefore ISIDORE represent a good starting point for getting an open linked dataset on Social Science and Humanities domain.
During the development of my project, it could be also interesting to evaluate my work relevance following the collaborations with external communities and their feedbacks. Therefore, it's important to keep the sources, applications and services developed open for external contributions, following the Open Science politics: making scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.

Workflow and expected results

Working on this research project implies two different processing phases: one theoretical and the other applicative. The first phase of the project will be dedicated to a detailed theoretical study and background on the technologies to use, I will move on the next phase just after having established an extensive theoretical background. Here I present my Ph.D workflow timeline for the upcoming years.

First year

I will dedicate my first year reading up about the last strategies adopted in SOS, and what issues have been recently highlighted by the community. I will try to figure out how these issues can be addressed by using semantic publishing technologies, and I will deepen my analysis on the metadata associated with the published information to describe it by means of Semantic Web technologies. At this stage this background will be general without a particular dedication to a specific study domain.
This will obviously involve the study of the markup language and the developed ontologies, on different scholarly domains. In a more abstract view, this also means investigating the strategies used in the elaboration of the knowledge graph data in the scholarly literature, and what are the aspects that have been mostly taken in consideration when observing it. In addition to the data representation,it is also important analyzing the querying strategies used

and the aspects that the community have been mostly searching when querying the datasets.
After having a general review, my analysis will focus on the social science and humanities domain. In this case, I need to investigate on the strategies and services adopted in these particular fields. In addition, I will check the available datasets (linked data), and those openly accessible to be interrogated.
Since the general purpose of my research is mining new hidden dynamics in a semantic publishing context, in this first year I expect to work on and consult the state of the art regard the “citation function” research topic, and the last approaches adopted to automatically learn and detect it.
At this phase I will also take in consideration all the questions and demands that the community and industry have mostly worked on through the last years.

Second and Third year

Once I have built a large background on the above mentioned subjects on my first Ph.D year, I will move to the applicative side of this work. At this point I should narrow my focus on the Digital Humanities domain, and start the application of novel methodologies based on what I have learned.
This application oriented phase is based on the researching issues mentioned in the previous section. I expect to have a clearer view on the feasibility of addressing my research questions, after observing and analyzing the results of the evaluations I will conduct to this goal. The outcomes of these evaluation are strictly related to the datasets used, this will help enforce the final results obtained.
It will also be interesting to deepen my research studies with a possible internship abroad, in particular within a team that already works on Science of Science thematics. For instace, the CWTS [15] is an interesting highly qualified place which respects these requirements. One interesting new project named “Open science: Monitoring trends and drivers”, its aim is to further investigate the Open Science research theme synthesizing CWTS research on the current policy drive towards “open science”, and translating the research results as well as theoretical, empirical, and technical expertise to applicable ideas, advice, and technical solutions.

Fourth year

Finally, I will start writing and formalizing the results of my Ph.D research in the final thesis report. In order to have a high visibility on the community it is necessary to build a dynamic interactive interface showing the outcomes that have been reached. Alongside the applicative interface, all the datasets used should be released with open access licenses so as to be available to the community. This will allow me to get further feedback, and can be a good basis for setting up new collaborations with other researchers interested in the topic.
The final evaluation of this work, it’s strictly related to the impact of the use semantic publishing technologies when applied to the social science and humanities domain materials. A formal evaluation, e.g. with questionnaires (also asks for a comparative evaluation with previous strategies), will involve researchers and other users who use social- and humanities-related materials daily, so as to help me to gain additional feedback on the services developed during my Ph.D.

Bibliography

1. Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*.
2. Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, *22*(2), 85-94.
3. Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS computational biology*, *5*(4), e1000361.
4. OpenAIRE, https://www.openaire.eu
5. OpenCitations, http://.opencitations.net
6. SciGraph: Springer Nature, https://scigraph.springernature.com
7. SCAR, http://dasplab.cs.unibo.it/index.php/semantic-coloring-academic-references/
8. Volundarson, Oskar (2016). Semantic Publishing in the Humanities: Enhancing the reader's experience, https://openaccess.leidenuniv.nl/handle/1887/43330
9. (2017), Knowledge engineering models and tools for the digital scholarly publishing of manuscripts,https://iss.unige.ch/research/projects/knowledge-engineering-models-and-tools-digital-scholarly-publishing-manuscripts/
10. Invito a presentare progetti di formazione alla ricerca in attuazione del Piano triennale Alte Competenze per la ricerca, http://formazionelavoro.regione.emilia-romagna.it/entra-in-regione/bandi-regionali/invito-a-presentare-progetti-di-formazione-alla-ricerca-in-attuazione-del-piano-triennale-alte-competenze-per-la-ricerca-il-trasferimento-tecnologico-e-limprenditorialita-approvato-con-deliberazione-dell2019assemblea-legislativa-n-38-del-20-10-2015-por-fs
11. Humanities at Scale (HaS), http://has.dariah.eu/
12. Linked Data for Digital Humanities, http://www.dhoxss.net/linked-data
13. ISIDORE, https://www.dariah.eu/tools-services/tools-and-services/tools/isidore/
14. Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villéger, A., & Attwood, T. K. (2011). Ceci n'est pas un hamburger: modelling and representing the scholarly article. *Learned Publishing*, *24*(3), 207-220.
15. Centre for Science and Technology Studies, https://www.cwts.nl/
16. Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Vespignani, A. (2018). Science of science. Science, 359(6379), eaao0185.
17. Berners-Lee, T., & Hendler, J. (2001). Publishing on the semantic web. Nature, 410(6832), 1023.
18. Peroni, S. (2014). Semantic web technologies and legal scholarly publishing (Vol. 15). Springer.
19. Peroni, S. Automating semantic publishing. Data Science, (Preprint), 1-19.
20. Hamou-Lhadj, A., & Hamdaqa, M. (2009, April). Citation analysis: an approach for facilitating the understanding and the analysis of regulatory compliance documents. In Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on (pp. 278-283). IEEE.
21. Bosch, Thomas, and Benjamin Zapilko. "Semantic Web Applications for the Social Sciences." IASSIST Quarterly 38.4 (2015): 7-7.
22. Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In Proceedings of the 2006 conference on empirical methods in

natural language processing (pp. 103-110). Association for Computational Linguistics.
23. Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social studies of science, 5(1), 86-92.
24. Horrocks, I. (2008). Ontologies and the semantic web. Communications of the ACM, 51(12), 58-67.
25. Chichester, C. Valorizing omics visualization for discovery. Data Science, (Preprint), 1-7.
26. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425.