PhD Odyssey
Interweaving PhD stories into the Semantic Web

Viet Bach Nguyen

Faculty of Informatics and Statistics
University of Economics, Prague, Czech Republic
nguv03@vse.cz

Abstract. In this article, I present a research project proposal for my PhD study program in Applied Informatics. The main focus of this project is knowledge engineering on the Semantic Web for the academic domain namely involving the creation of ontologies about research and campus activities of PhD students throughout their academic journeys. The ultimate goal then is to create a game-like reference application that would feature PhD story sharing to enable the embroilment of academic stories into the Semantic Web. I also provide a pilot version of the project specification, preliminary concept research, anticipated challenges, outline and approach of my research activities in the future as well as a couple of preliminary results from conducted preparatory experiments.

Keywords: academia · design research · knowledge base · linked data · ontology engineering · PhD story · qualitative research · semantic web

1 Introduction

Doctoral study plays an important role in extending the boundaries of human knowledge. For example, discoveries such as GPS and MP3 technologies would never have happened were it not for postgraduate research [6]. PhD programs offered by universities represent the privilege which prospective students can hold on to and venture on a journey of an extensive understanding of the unknown while having a sense and belief of that their work should bring a meaning to life. Unfortunately, however, the PhD completion rate within the standard study program duration worldwide is not high enough [1].

This article aims to present a major thesis proposal for my dissertation in Applied Informatics. The societal context problem is to help PhD students have a better understanding of their journeys. This work aims to collect and organize knowledge about academia and PhD experiences in a semantic manner to share these experiences among prospective and on-going PhD students to help them by displaying concepts, hints and best practices. By doing so, I believe this work will help students better adapt to their PhD programs. To achieve all that, I propose a major project in which knowledge engineering activities are involved. The first objective is to create a knowledge base specifically about PhD journeys. On top of this knowledge base, a reference game-like application would be built to enhance the passage of PhD students in their journeys.

2 Problem statement and state of the art

Project PhD Odyssey will be devoted to the analysis and creation of ontologies for the academic domain, specifically about PhD journeys. This would allow PhD students to share their stories on the web in a semantic manner. The ultimate objective of this project is to build a knowledge base and a game-like application. The main use case of this application is to educate prospective and new PhD students with the use of an interactive and progressive story-telling game.

The scientific problem to this project lies within the heterogeneity and variety of data sources that this project will work with. These data sources are listed in section 3. Moreover, a significant part of the data is unstructured and text-based. Since they are heterogeneous, it is reasonable and necessary to solve the problem in a semantic way, e.g. mapping common terms to ontologies.

Presumably, this type of research is quite extensive. For that reason, boundaries and goals should be determined during the earliest phase of the project. As mentioned above, the activities and output of my dissertation should be bounded by the academic field of Applied Informatics and might be limited to several countries. Moreover, an early prototype version of the reference application would only feature a simple tagging model for story creation.

As mentioned before, the context problem or the requirement of this project is how to help PhD students along their journeys. The proposed solution to this project is based on Design Research, meaning how to design and implement an artifact structure that could fulfill the requirement. This consists of research artifacts mentioned in section 5.

As part of preparatory research, relevant projects involving biographies and personal information have been found. The most significant ones are *The Catalogus Professorum Lipsiensis* [9] and *BiographySampo: Finnish Life Stories on the Semantic Web* [8]. These projects tackle the problem of how to semantically incorporate personal data such as biographies of very interesting subjects in order to study them. In the first work, the authors deal with the development of a prosopographical knowledge base about the life and work of professors in the 600 years history of University of Leipzig. The second work also focuses on prosopographical research about renowned Finns and its output is a linked open data service that is used to publish these biographies on the Semantic Web.

3 Data and knowledge sources

As part of the initial research, five main types of existing data and knowledge sources have been identified as they could contain information relevant and useful to this project. The suggestion is to explore them individually and, by applying acquired knowledge, create a grounding foundation for this project.

PhD stories. The use case of this project is described as to help students navigate through their PhD journeys by learning from previous experiences of former students. These experiences are often documented and can be studied by means of semantic understanding. Many contemporary students share their

stories online in forms of blog posts as well as video documentation aka vlogs. The suggestion is to study these stories and retrieve metadata to form a specific understanding of the academic field and processes from the point of view of PhD students. Since these stories likely are unstructured and text-based, the approach for knowledge retrieval would be based on a tagging model. In the early stage, any work involving story tagging would be self-sourced. In the long term, however, a model for crowdsourcing would be created to allow PhD students to tag and compose their stories on their own.

Ontologies and vocabularies. There are several projects whose tangible outputs are ontologies that are devoted to the academic domain and research. The following ones have some features of interest that could be closely related and useful to this dissertation project. **OLOUD** [4] is an ontology that provides a high-level model covering multiple education-related use cases involving Linked Open University Datasets and applications built on top of these datasets. **VIVO** [2] is an internationally adopted ontology which enhances the discovery of researchers and collaborators across disciplines and organizations. **SWRC** [10] is an ontology in which research communities and relevant related concepts are modeled. It also emphasizes the importance of ontology re-use. **PWO** [5] is an ontology for the description of workflows that is particularly suitable for formalizing typical publishing processes such as the publication of articles in journals. These ontologies will be examined and investigated for prospective re-use and improvement. The goal is to extract knowledge and create semantic connections from a PhD student's perspective.

Open knowledge graphs. Knowledge graphs are also a significant and promising source for analysis as they must incorporate concepts that are relevant to the targeted field of study. Examples of knowledge graphs are DBpedia[1], Wikidata[2], YAGO[3] and Google Knowledge Graph[4].

Data schemes of information systems. Knowledge of higher education can also be retrieved by exploring the structure of information systems that are used in universities and academic institutions. In particular, the Integrated Study Information System used by the University of Economics, Prague consists of a wide range of modules and functionality which aim to support study processes and manage academic data and records. The examination of this system would be included as part of the project's targets for knowledge acquisition based on the analysis of its underlying data schemes and outputs.

Public code lists. Yet another source of interest is the public repository of government open data, e.g. reference data such as code lists, especially the ones managed by the Ministry of education. These code lists are bound to contain specific educational data that can be used for the modeling of ontologies. The

[1] https://wiki.dbpedia.org/
[2] https://www.wikidata.org/
[3] https://www.mpi-inf.mpg.de/yago-naga/yago
[4] https://developers.google.com/knowledge-graph/

first stage of this project, however, will only focus on the local educational system of the Czech Republic which is the country this project originates from. The posterior extension of this work would include foreign educational systems as well. Since each country has its own distinct educational system, there is a possibility of establishing a collaborative work with other PhD students overseas in the mapping of their specific local systems in the future. This collaboration could take place during my internship in a foreign academic institution in forms of fieldwork to collect data and consult with local experts. The plan, therefore, is to investigate whether and which types of educational code lists are available in several different countries.

4 Possible challenges

Partitioning of knowledge. Since data sources are heterogeneous, the inherent challenge lies in the partitioning of knowledge into different types of interpretation, namely ontologies, vocabularies, knowledge graphs and code lists. This is because there are various sources as discussed before. The partitioning process will have to deal with heterogeneous data sources, apply certain rules and produce a condensed and clean data structures concise enough for the usage in the reference applications. To successfully partition the knowledge topologically, the definition and application of partitioning dimensions are required.

Thematic overlap of modules. The knowledge base will be developed using a module-based approach where each module should address one particular problem or topic in order to be comprehensible. Thus the knowledge engineer must, while developing one module, keep in mind that there could be some concepts that are applicable to other modules as well while trying to confine the scope of these modules.

Quality of resources. Throughout the process of studying existing PhD stories and memoirs, one might come to a problem where stories are incomplete or distributed unevenly in terms of continuity which may cause distortion of facts or misinterpretation of the author's original thought.

Regulations. The proposed application system will definitely work with personal data and other sensitive information. This poses a real challenge for the service provider to organize and protect data in a compliant and secure way. PhD stories would be accessible by the public and should not violate any law or regulation. PhD stories can be obtained in two main ways – collection of published information and personal interview. The latter certainly requires the obtainment of consent from the participant. Therefore, there is a difference between passive and active knowledge acquisition when it comes to personal data. A more proactive approach is to use data anonymization techniques.

Knowledge base and transaction data. The game application should provide an interface for the instantiating of data about PhD stories. This would be

classified as transaction data as they are created by the initiative players. There should also be some illustrative artificial data modeled for the pilot version of the game. The challenge here is how to best structure the mapping process of transaction data to the existing knowledge and how to maintain the system in case of necessary change of concept in the knowledge base.

In the scope of my dissertation, however, it is unlikely to tackle all mentioned challenges in a timely fashion. Hence, it should be more reasonable to select and work on the fundamental ones such as partitioning.

5 Anticipated strategy, research approach and goals

To partially visualize the layout and scope of this work, several tasks and goals have been identified as follows:

- Analysis of the mentioned related works about biographies.
- User requirement analysis to determine the use cases of the solution, e.g. asking potential users on what would they expect from the project or what features would they find useful.
- Interviews with potential users while testing concepts and mock-ups of the result application to identify and consolidate the fundamental purpose and goal of the whole project.
- Bibliography research including scientific conferences and journals.
- Research of ontology-related subjects including frameworks, best practices, guidelines and design approaches for utilization in real applications.
- Inspiration from old and contemporary PhD stories which are available online or documented as biographies. Here lies a possibility for application of *thematic analysis* [7] involving practical methods such as examination, sampling, induction, deduction, inference and pattern recording.
- Interpretative ethnographic and qualitative research in forms of participated observation or indirect surveillance. Resources of interest also include various types of documented or live materials.
- Creation of ontology models to capture academic knowledge.
- Creation of a domain-specific knowledge graph containing connected facts and interrelations between concepts. This would acquire and integrate relevant information into an ontology and possibly derive new knowledge (similarly to DBpedia, Wikidata or Google Knowledge Graph, from which there would be some partial re-use of concepts).
- Follow-up questionnaire surveys, experiments and case studies to verify the usefulness of the acquired knowledge.
- Modeling of sample PhD stories as RDF graphs and their evaluation.
- Attempt to perform inference and reasoning from these stories to capture more knowledge data to be used in the reference application. The stories must be sufficiently detailed so this can be achieved.
- Design and implementation of an interactive web interface and application, which would presumably leverage ontology-driven software engineering [3].

- Partial gamification of that application with a simple graphical design, inspired by the Accountant game[5] which aims to educate public officials.
- Build a tagging model for the support of user story creation. This model would enable the semantic annotation of stories based on the common terms presented in the ontologies.

Because the main feature of this project is Design Research, the primary anticipated results of this work would be the following artifacts:

- **Knowledge artifacts** – a **set of ontologies**, **vocabularies**, **code lists** and **knowledge graphs** of PhD stories that incorporate knowledge about scientific research methods and methodologies.
- **Data artifacts** – a **repository of PhD stories** containing specific data that instantiate the ontologies.
- **Software artifacts** – an user interface used to query and navigate through the knowledge base and a standalone **reference game application** that makes use of the knowledge data.

Apart from these, there could also be a derived goal of proposing **possible best practices** based on the process of creating the mentioned artifacts. The presumed results could be useful for knowledge engineering on the Semantic Web, e.g. how to best semantically partition collected data from structured and non-structured sources. Therefore, by solving the societal context problem, there is a potential additional motivation to produce unexpected outputs that could be useful to the research community.

6 Experiments, results and discussion

Two experiments have been conducted so far, each of which respectively reflects the exploration of the first two knowledge sources mentioned in section 3.

The first conducted experiment was to collect sample data from 5 PhD blogs and 2 video posts. While reading and listening, I have noted down 130 specific terms that are relevant to the journey of a PhD student, such as *PhD advisor*, *research topic*, *entrance exam*, etc. while using my own knowledge of the academic domain to determine the relevance. I also looked for generic words and expressions that are specific to the academic domain, such as *student*, *university*, *subject*, etc. Several more interesting words were also featured, e.g. *write-up*, *funding*, *viva voce*, etc. During this experiment, I also tried to understand the meanings of more complex statements and come up with abstract terms to generalize the expressions. For example, some authors write about how they stayed on track by managing themselves in several aspects, essentially talking about *self-management*.

As expected, the blogging styles are all different. Some PhD students try to document their experience on a continual basis, writing posts every week

[5] https://jplusplus.github.io/the-accountant/

or month, basically creating a list of events in chronological order, while other students try to compress the whole journey after graduation into one long post. Content-wise, some students mainly write about issues and difficulties they had to deal with on a daily basis, some students focus on the details of their research projects while other students provide structured retrospects for their journeys.

The second conducted experiment was to explore and examine the VIVO, PWO, OLOUD and SWRC ontologies to see if the acquired knowledge from the first experiment could be included in them. In the VIVO ontology, I have found 19 entities that could cover terms from the first experiment, e.g. classes such as *Campus*, *College*, *Faculty*, *Department*, *Research Proposal*, *Committee*, *Grant*, etc. This ontology proves to be a very relevant resource for semantic re-use.

In the PWO ontology, however, I have found no entities that are related to the collected terms. This is because the PWO ontology is focused only on the publishing process. I have found 8 entities in OLOUD, e.g. *Course*, *Degree*, *Study programme*, *PhD*, and 9 entities in SWRC, e.g. *Student*, *Conference*, *Article*, *Department*, *Topic*. Table 1 shows the coverage of collected terms in all four mentioned ontologies. The rightmost column tells how many entities were not exactly the same as collected terms, but are semantically very closely related.

Table 1. Terms coverage in related ontologies

Ontology	**Terms covered**	**Covered exactly**	**Covered approximately**
VIVO	19	6	13
PWO	0	0	0
OLOUD	8	7	1
SWRC	9	8	1

As a tangible output, I have also tried to create a proof-of-concept ontology that describes 71 of the acquired terms from the first experiment. This ontology serves as an experimental result and can be found online[6]. Please note that this ontology is nowhere near perfection and I am aware of its shortcomings. I anticipate my approach can achieve more significant results in the future as I keep on studying more PhD stories and exploring more knowledge sources, and then changing the ontology model as needed step by step.

7 Conclusion

In this paper, I introduced the context problem to my dissertation thesis and I provided a descriptive insight into the proposed project. Next, I showed an important part of my conducted research for data and knowledge sources and I mentioned the anticipated challenges to be considered. For the solution, I provided an outline of the anticipated strategy, research approaches and goals to

[6] https://github.com/nvbach91/phd-odyssey

be achieved. Last but not least, I provided details and results of my first two conducted experiments which follow the guidelines I laid out before.

Once again, the ultimate objective of this dissertation project is to build an accessible game-flavored knowledge-based system, providing users with an enjoyable process of learning about campus activities, research processes and research methods from the perspective of former PhD students. I believe my work will help open the doors and create opportunities for prospective and ongoing PhD students to do a better job upon their pilgrimage to push the scientific research community and the entire humanity forward.

Acknowledgments. I would like to express my sincerest gratitude towards Prof. Vojtěch Svátek for supporting me in my study as my PhD advisor. *This research has been partially supported by the CSF under 18-23964S.*

References

1. Bednall, T.C.: PhD completion: An evidence-based guide for students, supervisors and universities @ONLINE (July 2018, accessed September 20, 2018), http://theconversation.com/phd-completion-an-evidence-based-guide-for-students-supervisors-and-universities-99650
2. Börner, K., Conlon, M., Corson-Rikert, J., Ding, Y.: VIVO: A Semantic Approach to Scholarly Networking and Discovery. Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers (2012)
3. de Cesare, S., Gailly, F., Holland, G., Lycett, M., Partridge, C.: Ontology-driven software engineering 2010. In: Companion to the 25th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, SPLASH/OOPSLA 2010, October 17-21, 2010, Reno/Tahoe, Nevada, USA. pp. 279–280 (2010)
4. Fleiner, R., Szász, B., Micsik, A.: OLOUD – An ontology for linked open university data. Acta Polytechnica Hungarica **14**(4), 63–82 (2017)
5. Gangemi, A., Peroni, S., Shotton, D.M., Vitali, F.: The publishing workflow ontology (PWO). Semantic Web **8**(5), 703–718 (2017)
6. Gray, A.: These countries have the most doctoral graduates @ONLINE (February 2017, accessed September 20, 2018), https://www.weforum.org/agenda/2017/02/countries-with-most-doctoral-graduates/
7. Guest, G., MacQueen, K.M., Namey, E.E.: Applied Thematic Analysis. Thousand Oaks, CA: SAGE Publications Inc. (2012)
8. Hyvönen, E., Leskinen, P., Tamper, M., Tuominen, J., Keravuori, K.: Semantic national biography of Finland. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), Helsinki, Finland, March (2018)
9. Riechert, T., Morgenstern, U., Auer, S., Tramp, S., Martin, M.: Knowledge engineering for historians on the example of the catalogus professorum lipsiensis. In: International Semantic Web Conference. pp. 225–240. Springer (2010)
10. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D.: The SWRC ontology – Semantic web for research communities. In: Progress in Artificial Intelligence, 12th Portuguese Conference on Artificial Intelligence, EPIA 2005, Covilhã, Portugal, December 5-8, 2005, Proceedings. pp. 218–231 (2005)