Knowledge Engineering of PhD Stories: A Preliminary Study

Viet Bach Nguyen
Vojtech Svatek
Marek Dudas
viet.nguyen@vse.cz
svatek@vse.cz
marek.dudas@vse.cz
Prague University of Economics and Business
Prague, Czech Republic

Oscar Corcho
Universidad Politecnica de Madrid
Madrid, Spain
ocorcho@fi.upm.es

ABSTRACT

Support for PhD students and their advisors in decision-making before and along their PhD journeys requires providing them with a deep understanding and knowledge of the life-cycle of a PhD. This means giving them access to a thorough understanding of causal relations between events, decisions, and the possible outcome. This knowledge can be attained primarily from insider stories, study reports, communications threads with advisors and colleagues, interviews, and scholarly databases. However, it is unclear how to give this knowledge a reasonable structure (due to the heterogeneity of concepts and data sources) so that we can use it for decision-making during the PhD journey. In this paper, we explore how to analyze and model PhD stories to uncover and extract causal relationships found within each story to get insights into the co-occurrences and causalities. We analyze these stories with thematic analysis to understand their main points and we use concept maps to create semi-formal graphs of connected events and objects where the relationships are being emphasized from the perspective of cause and effect. Our results at this point are a collection of PhD stories in the form of concept maps, thematic codes, a proposed approach for goal-directed PhD story modeling which we describe in this paper.

CCS CONCEPTS

• **General and reference** → **Biographies**; • **Theory of computation** → **Semantics and reasoning**.

KEYWORDS

causal relation, causality, concept map, knowledge engineering, PhD story modeling, thematic analysis

ACM Reference Format:
Viet Bach Nguyen, Vojtech Svatek, Marek Dudas, and Oscar Corcho. 2021. Knowledge Engineering of PhD Stories: A Preliminary Study. In *Proceedings of the 11th Knowledge Capture Conference (K-CAP '21), December 2–3, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3460210.3493579

K-CAP '21, December 2–3, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8457-5/21/12…$15.00
https://doi.org/10.1145/3460210.3493579

1 INTRODUCTION

A PhD journey is a long-term process of a scholar learning how to conduct research and advance in a research area. Each PhD journey is a unique experience consisting of several stages depending on the study program. Throughout each PhD journey, many decisions are made and many events happen that lead to a (un)favorable outcome. It is, however, a pity that almost 50% of doctoral students do not graduate [14], which also affects the overall PhD research quality in the world. To provide better support and improve PhD research quality and study retention, it is necessary to gather insights into the co-occurrences and causalities in PhD stories and structure the information, e.g., in a knowledge graph, for practical use. This knowledge base populated with entities from the stories would be analyzed to discover recurrent patterns. These patterns would be kept as an additional layer of the knowledge base, and divided to best practice (what should be followed), anti-pattern (what should be avoided), and neutral. In this context, the most profiting users would be the advisors or PhD program managers, who could better understand the issues in the PhD domain. Top-level focus questions guiding the choice of concepts in a PhD story would be possible reasons and drivers for the eventual success or failure and achieving good research results. The main use cases for such a system would be to inform and educate prospective PhD applicants and to support on-going PhD students and their advisors using the discovered patterns and the inference over them. This envisioned system would consist of a knowledge base with search functionalities and a serious game built on top of it, as well as an interface for users to model and annotate their own stories.

To effectively understand the PhD story passage, it is essential to emphasize the causal relationship between events and decisions. In this paper, we begin exploring the possibility of capturing and analyzing PhD stories to extract events and causal relationships from them using thematic analysis coding method [3] and concept maps. We mainly focus on stories from the research field of *Applied Informatics* and have gathered several PhD stories from various types of sources like evaluation reports, blog posts, and interviews for analysis. Our goal is to create semi-structured graphs of events that are important for individual PhD stories and annotate them with abstract concepts. These graphs could then be selectively enriched from structured knowledge graphs with, e.g., data about publications. The result so far is a collection of modeled PhD stories in the form of concept maps, thematic codes from stories, and a proposed PhD story modeling approach based on exploratory experiments.

2 RELATED WORK

We are unaware of any project having the same target and similar technology as ours. However, PhD stories themselves have indeed been studied, but using social science, rather than knowledge engineering techniques. PhD stories as subjects of knowledge modeling have several features shared with slightly different targets addressed in prior research. Also, the story itself bears some similarities to a scientific workflow. On the other hand, being a story of a particular person, it can be viewed as a special kind of partial biography. Furthermore, as the purpose of the story modeling is the capturing of causality, we could align them to causal ontologies. And, in the field of EduTech, serious games are used to train students in higher education to better get acquainted and improve performance.

Social science. Studies on PhD journeys have been active in the social science field commonly based on ethnographic studies. We have stumbled upon some projects that directly address the issues in the PhD life and show the significance of the amount of workload and the relationship between advisors and PhD students [17].

Knowledge modeling. In our prior work [13], we addressed knowledge patterns used by researchers during information foraging, in specific contexts. Unlike our current project, the patterns only involved entity links that are accessible through public resources, and the considered target was that of satisfying a fine-grained information need rather than of undertaking a complex research activity such as a PhD project.

Scientific workflows. Knowledge graphs and ontologies have been a strong focus for research in the Semantic Web field in recent years, namely in the scholarly, researcher, and academic domains. A systematic and provenance- and ontology-based approach is used for semantically explaining conducted scientific experiments with the primary goal of reproducibility by describing the whole story path of these experiments [16]. The primary focus of this project is set on the level of individual research experiments, which can be perceived as a component of the whole PhD journey as they are much more detailed with a focus on projects, rather than on an individual researcher, whose PhD journey may encompass several interrelated projects and numerous experiments. Our modeling of PhD stories primarily aims at extracting and aggregating patterns; even if a real story were taken as an example to be *reproduced*, this reproduction occurs at a rather approximate level. There are, however, entities likely to appear in both kinds of models, e.g., collaborator, artifacts, datasets, tools, experiments, etc.

Biography. Further relevant state-of-the-art projects involving knowledge graphs include the capturing of biographies via systematic knowledge graph building efforts with a social-science flavor [15] [8]. Generally, aspects of biographies are present in common encyclopedic knowledge graphs such as Wikidata and DBpedia, specialized knowledge graphs from publications, or integrative knowledge graphs wrt. the scholarly domain, especially the Open Research Knowledge Graph (ORKG) [10]. In the knowledge graph-based approaches, the biographical information is captured at the level of basic CV facts such as birth, marriage, and university degree enrollment/completion, primarily for important historical persons with sufficient digital imprint (even if reconstructed from historical documents). The common encyclopedic and specialized knowledge graphs provide access to biographical data of ordinary and still living researchers, but with an even smaller repertory of biographical detail types, and rather exhaustively while not considering the varying degree of importance. Therefore, both these approaches understandably ignore fine-grained events that can only be reliably captured by insiders, never mind possible causalities among them.

Causal ontologies. In the ontology engineering world, causality as a topic has appeared mostly in projects that involve medical and healthcare-related studies [18] to capture relationships between diseases and symptoms as well as medical procedures and treatments for, e.g., improving decision support in domain knowledge-based diagnosis [6]. [7]. Other domains of interest are, e.g., construction [19], NLP [11], and fact-checking [4]. Causation as a general ontological concept has also been created in ontologies such as DTO [5], XKOS [2], and BioTop [1]. Still, the causal aspect of knowledge capturing for the scholarly domain has not been addressed. A theory of modeling causality on the conceptual level has been, however, summarized in related work [12], but no formalization has been proposed as of yet. By modeling the PhD stories in a concept-map manner, we can explore the possibilities of finding sample instance data for causal relationships in the scholarly domain.

EduTech and serious games. In practice, efforts in improving the study retention can be observed in EduTech, where serious games are being applied in higher education [9]. These games are usually created based on several serious game models and frameworks that provide the appropriate and necessary incentives for users and evaluative measurement techniques.

3 RESEARCH METHODOLOGY, PHD STORY COLLECTION, ANALYSIS AND MODELING

Our approach for tackling this research problem consists of several steps, see Fig. 1: literature survey, data source type identification, data collection, story analysis, story modeling, pattern extraction, and understanding causal relationships.

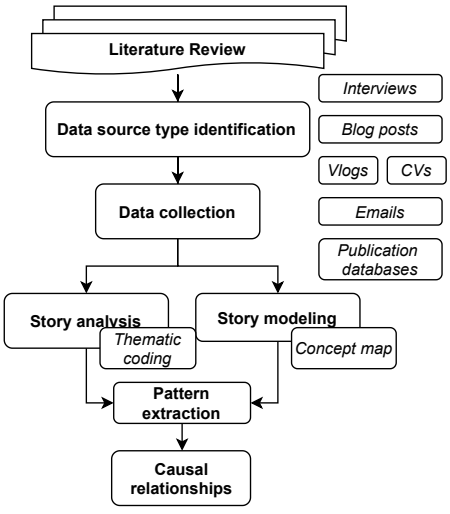

Figure 1: Methodology outline

First, we looked in the recent literature for similar projects featuring topics of our interest, e.g., PhD story, story modeling, biography, causality in ontology engineering. As mentioned in Section 2, we are unaware of any project that focuses on analyzing the PhD narrative to extract causal knowledge for a systematic support solution.

For identifying data source types, we have consulted with PhD program advisors and PhD alumni and asked them what kind of information is crucial in their stories and in what way they could be retrieved. The undeniable resources are documented stories in various forms, most often blog posts and video logs. Personal information and major turning points could be obtained from CVs, study information systems, and official study records like annual PhD reports. We also learned that the less accessible, but retrievable, resources are communication threads in the form of emails and personal notes of the participants. Last but not least, also a very promising type of resource, are interviews which we conducted with the help of several PhD students in the Applied Informatics field. The variety of data sources already suggests that the necessary information is very scattered, which is difficult to handle.

In the data collection step, we focused on four main sources: PhD advisors, interviews, publication databases, and documented stories on the web.Since the central subject in each story is the PhD student, we first created a list of candidate names and tried to get as much information as possible through the four mentioned sources for each of them. To extend our collection, we searched over the internet for interesting stories containing good narratives.

This way, we have collected a total of 12 stories, 8 in the form of blog posts and videos, 4 in the form of interviews. We then supplied additional data from emails, annual evaluation reports, and notes of the PhD advisors. Thematic analysis was applied on all stories and concept maps were created for 5 stories.[1]

With the collected data, we proceeded with analyzing and modeling the stories using a proposed method shown in Fig.2. For the analysis part, we used thematic coding technique to find the main points (codes and themes) of the stories. For the modeling part, we rely on concept maps where the story is modeled by a story insider. Both approaches are used partly in parallel and inform one another.

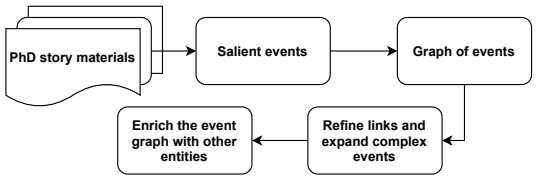

Figure 2: PhD story modeling

During this stage, we have distinguished three main strategies for analyzing a PhD story: 1) *timeline view* – systematically going through the events in their chronological order; 2) *goal-directed* – going rather backward, from the completed and successfully defended thesis, or from the point of failure, to the potentially causing events; and 3) *personal impressions* – going from salient impressions of the student in the given moment, typically based on an interview, and building a cluster of events. These three strategies are described in separate documents listed in our GitHub repository. These approaches provide a solution for how to collect the knowledge from stories with the main goal of providing decision support based on typical causal links that can be used to structure the knowledge.

Our first attempt was based on the timeline view, chronologically going through a story mostly based on annual evaluation reports and publication databases. This led to a richly structured graph with ontology-based flavor, covering entities of many different types residing around the story. Inevitably, the size of the graph grew considerably, which made it hard to manage (even for just two initial years of the story). The second attempt for a different story involved looking at the thesis and final evaluation reports and matching them with the goals and critical factor model consisting of formal requirements for obtaining the PhD title, and then chronologically going through the story, this time using an available detailed story blog post, but only selecting the salient events, primarily in reference to the filled template of critical factors. This produced a graph of only events, including tentative causalities. For the third attempt, we tried to include in the model as much information as possible from the stories while connecting events and decisions with ad-hoc links to preserve the narration details.

We came to the preliminary conclusion that the backbone of each story should be the big and important, salient events throughout a PhD journey. Therefore, the events on the highest conceptual level for the PhD story modeling should correspond to the PhD program passage, such as PhD enrollment, completing subjects of the PhD program, doing internships, research projects, publishing papers, attending conferences, completing state exams, thesis defense.

We looked at the final or most up-to-date PhD materials like theses, reviews, late-stage evaluation reports, etc., and characterize the ways of satisfying the goals and effect of conditions during the study. Next, we went through the story chronology (in a PhD blog, e-mail threads, series and evaluation reports, etc.) and annotate the *salient events* in it while using the findings and additional materials (interviews, if available) to judge the salience. The annotated story was used to build a graph of events linked by broadly understood *led to* links, plus some auxiliary ones, e.g. *part of*. The next step was to refine the links and expand complex events to chains. The last step was to enrich the event graph with other entities primarily such that relate to more than one event, e.g., people, projects, conferences, experiments, artifacts, etc. because parts of the stories could be populated from existing knowledge graphs, especially those of publications and their authors.

For modeling stories, we used ContextMinds web app[2] which allows for lightweight graphical modeling and the possibility of subsequent export of the whole graph to RDF. While the modeling is lightweight compared to knowledge graphs, which are tightly associated with formal ontologies, it is still possible to transform the models (maps) to a formal (RDF) representation relatively easily. The starting point is the concrete stories rather than the domain as such. Thus, the concepts will primarily be ontological particulars while the ontological classes can be approximated via tags, e.g., event, creative work, researcher, experiment, organization, etc.[3].

[1] https://app.contextminds.com/?m=L2J7e, https://app.contextminds.com/?m=1nW6E, https://app.contextminds.com/?m=d0wgg, https://app.contextminds.com/?m=NJb7g, https://app.contextminds.com/?m=YBE7j

[2] https://www.contextminds.com

[3] https://app.contextminds.com/?m=4K7Ml

The analytical part of our research primarily involves thematic analysis, a relatively lightweight qualitative research method frequently applied on knowledge-rich texts such as interviews. This technique suggests highlighting sections of transcripts and textual materials with *codes* to describe the content by going through and highlighting everything that we selectively perceive as relevant to our questions or potentially interesting from the story. The data are then grouped using these codes which allows us to gain a condensed overview of the main points that recur throughout the data. Codes are then reviewed and combined to identify *themes* or *types* that help us understand the meaning of the content and identify concepts and terms in the domain of interest. Our primary goal for analyzing PhD stories using this method is to provide ideas for tags and concepts to be then suggested in graphical modeling. The most significant result of this analysis is a list of keywords and their types found in the stories, see Tab. 1.

Table 1: Thematic coding results examples[4]

Codes				
Term	Teaching	Paper	Advisor	Experiment
Journal	Meeting	Survey	Consultation	...
Types				
Achievement	Document	Event	Institution	...

4 CONCLUSION AND FUTURE WORK

In this paper, we have presented our first effort towards exploring and creating a method for collecting, analysis and modeling of PhD stories in order to give structure to the knowledge found within them, which we hope will be useful in the development of a knowledge-based information system that can provide knowledge support for PhD students and their advisors. After literature review and identifying data source types, we have used thematic coding and the proposed story modeling approach to analyze the collected data and build PhD stories in the forms of concept maps. The results of this preliminary study are published on GitHub[5] including analyzed stories, coding results, and modeled stories.

The proposed approach has limitations: the stories analysis and modeling are not easily replicated as all steps must be repeated for new stories, apart from thematic coding, where codes can be reused. We intend to apply crowd-sourcing by the PhD students and graduates themselves for the approach to scale. For this, we would presumably need a dedicated story editor or annotator for submitting stories in a semi-structured form.

Future work will include creating a system for capturing stories as concept maps by stakeholders (students, graduates, and advisors), and adding support for pre-population of concept maps from existing resources like structured knowledge graphs. We also intend to bring in an extensive usage of a concept reuse functionality in ContextMinds to allow to cluster the stories. We plan to conduct a systematic use of qualitative research methods to build an insightful model that would guide further quantitative experiments and rule mining from co-occurrences of events in PhD stories. Lastly, we plan to prepare a requirement analysis for a knowledge-based search application over PhD concepts and a serious game that simulate the course of the study (primarily for early-phase students and people considering the pros and cons of enrolling to a PhD program in Applied Informatics).

[4]Full table: https://github.com/nvbach91/phd-odyssey/tree/master/thematic-analysis
[5]https://github.com/nvbach91/phd-odyssey

ACKNOWLEDGMENTS

This research was supported by the project IGA VŠE № F4/56/2021.

REFERENCES

[1] E. Beisswanger, S. Schulz, Stenzhorn. H., and U. Hahn. 2008. BioTop: An upper domain ontology for the life sciences – A description of its current structure, contents and interfaces to OBO ontologies. *Appl. Ontology* 3, 4 (2008), 205–212.
[2] D.W. Gillman, F. Cotton, and Y. Jaques. 2013. XKOS: Extending SKOS for Describing Statistical Classifications. In *Proceedings of the 1st International Workshop on Semantic Statistics co-located with 13th International Semantic Web Conference, SemStats@ISWC 2013, Sydney, Australia October 11th, 2013 (CEUR Workshop Proceedings, Vol. 1549)*. CEUR-WS.org. http://ceur-ws.org/Vol-1549/article-03.pdf
[3] G. Guest, K. M. MacQueen, and E. E. Namey. 2012. *Applied Thematic Analysis*. Thousand Oaks, CA: SAGE Publications Inc. 3–20 pages.
[4] M. Guévremont and A. Hammad. 2021. Ontology for Linking Delay Claims with 4D Simulation to Analyze Effects-Causes and Responsibilities. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction* 13, 4 (2021), 04521024. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000489
[5] A. Hamdan, M. Bonduel, and R.J. Scherer. 2019. An ontological model for the representation of damage to constructions. *CEUR Workshop Proceedings* 2389, 6, 64–77. http://ceur-ws.org/Vol-2389/05paper.pdf
[6] L. Hong, H. Xu, and Shi. X. 2021. Constructing Ontology of Brain Areas and Autism to Support Domain Knowledge Exploration and Discovery. *International Journal of Computational Intelligence Systems* 14 (2021), 834–846. Issue 1.
[7] H. Hu and L. Kerschberg. 2020. Capturing Causal Knowledge from Authoritative Medical Ontologies. In *44th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2020, Madrid, Spain, July 13-17, 2020*. IEEE, 1373–1378. https://doi.org/10.1109/COMPSAC48688.2020.00-64
[8] E. Hyvönen, P. Leskinen, M. Tamper, J. Tuominen, and K. Keravuori. 2018. Semantic national biography of Finland. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), Helsinki, Finland.*
[9] Ebru Ince. 2018. *Educational Games in Higher Education*. https://doi.org/10.5772/intechopen.71017
[10] M.Y. Jaradeh, A. Oelen, K.E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and Sören Auer. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*. ACM, 243–246.
[11] P. Khalpada and S. Garg. 2021. Simple Automated Narrative Generator (SANG). In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 0909–0915. https://doi.org/10.1109/CCWC51732.2021.9375954
[12] R. Mizoguchi. 2020. Causation: Revisited. In *Proceedings of the Joint Ontology Workshops co-located with the Bolzano Summer of Knowledge (BOSK 2020), Virtual & Bozen-Bolzano, Italy, August 31st to October 7th, 2020 (CEUR Workshop Proceedings, Vol. 2708)*. CEUR-WS.org. http://ceur-ws.org/Vol-2708/foust7.pdf
[13] V.B. Nguyen, V. Svátek, G. Rabby, and O. Corcho. 2020. Ontologies Supporting Research-Related Information Foraging Using Knowledge Graphs: Literature Survey and Holistic Model Mapping. In *Knowledge Engineering and Knowledge Management – 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings*. Springer, 88–103.
[14] Council of Graduate Schools. 2008. Ph.D. Completion & Attrition: Analysis of Baseline Demographic Data from Ph.D. Completion Project. (2008).
[15] T. Riechert, U. Morgenstern, S. Auer, S. Tramp, and M. Martin. 2010. Knowledge engineering for historians on the example of the Catalogus Professorum Lipsiensis. In *International Semantic Web Conference*. 225–240.
[16] S. Samuel, K. Groeneveld, F. Taubert, D. Walther, T. Kache, T. Langenstück, B. König-Ries, H. Bücker, and C. Biskup. 2018. The Story of an Experiment: A Provenance-based Semantic Approach towards Research Reproducibility.
[17] E. van Rooij, M. Fokkens-Bruinsma, and E. Jansen. 2021. Factors that influence PhD candidates' success: the importance of PhD project characteristics. *Studies in Continuing Education* 43, 1 (2021), 48–67.
[18] M.S. Yin, M. Pomarlan, P. Haddawy, M.R. Tabassam, C. Chaimanakarn, N. Srimaneekarn, and S. Hassan. 2020. Automated Extraction of Causal Relations from Text for Teaching Surgical Concepts. In *8th IEEE International Conference on Healthcare Informatics, ICHI 2020, Oldenburg, Germany, November 30 - December 3, 2020*. IEEE, 1–3. https://doi.org/10.1109/ICHI48887.2020.9374310
[19] J. Zhang and J. Du. 2021. Ontology-based Defect Causation Analysis for Urban Tunnel Maintenance. In *2021 11th International Conference on Information Science and Technology (ICIST)*. 297–303.