MASARYK
UNIVERSITY

FACULTY OF INFORMATICS

Refining the process of writing theses with Red Hat Czech

Master's Thesis

KRISTIÁN KOTTFER

Brno, Spring 2023

MASARYK UNIVERSITY

FACULTY OF INFORMATICS

Refining the process of writing theses with Red Hat Czech

Master's Thesis

KRISTIÁN KOTTFER

Advisor: RNDr. Martin Ukrop, Ph.D.

Brno, Spring 2023

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Kristián Kottfer

Advisor: RNDr. Martin Ukrop, Ph.D.

Acknowledgements

I am immensely grateful to Martin, my exceptional supervisor, and the entire Red Hat Research team for their unwavering support, invaluable guidance, and genuine encouragement throughout my thesis journey.

To my family and friends, thank you for your patience, love, and support. Your presence has been a constant source of strength, joy and inspiration.

I am deeply grateful to everyone who has played a part in making this thesis a reality. Thank you from the bottom of my heart.

Abstract

This paper examines the thesis program at Red Hat Czech, a long-standing partner of universities in the Czech Republic. By analyzing almost a decade of program data, this study evaluates the program's current state, identifies significant trends, and pinpoints areas for improvement to ensure its long-term sustainability and success. The study also examines the entire thesis lifecycle, proposing enhancements to existing procedures for greater effectiveness. Additionally, it provides insights into student cooperation with Red Hat across different countries, highlighting regional variations. Research uncovered several areas that require improvement in the program, including gender equality, reducing the number of failed theses, and understanding the impact of regional disparities on collaboration. This thesis sets the stage for further research, establishing a basis for ongoing efforts to improve the program's effectiveness and sustainability.

Keywords

red hat, theses, data analysis, process design, data collection, student-industry cooperation

Contents

1 Introduction **1**

2 Context **3**
2.1 About Red Hat and Red Hat Research 3
2.2 Cooperation with Universities 5

3 Overview of Student Cooperation in Different Regions **7**
3.1 Czech Republic 8
3.2 USA 10
3.3 Israel 13

4 Analyzing and Tracking the Thesis Program Data **15**
4.1 Data Collection 15
4.2 Data Cleaning and Future Tracking 18
4.3 Data Interpretation 21
4.3.1 Thesis Production Over Time 22
4.3.2 Grade Analysis 25
4.3.3 Dean's Awards at FI MUNI 26
4.3.4 Analysis of Red Hatters Involved 29
4.3.5 Exploring Other Thesis Characteristics 31

5 Designing a New Process for Final Theses in Red Hat **35**
5.1 Motivation for Red Hat to Support Theses 35
5.2 Problems with the Current Process 37
5.3 Designing the Refined Process 38
5.3.1 Lifecycle of a Thesis Topic 40
5.3.2 Thesis Topics on the Web Page 42

6 Conclusion **45**

A Attachment Instructions **47**

Bibliography **48**

List of Figures

4.1 Thesis Completion Statistics: 2014-2022 23
4.2 Comparing Grades of BUT and MUNI 25
4.3 Variety of Supervisors 29
4.4 Most Common Keywords 32
4.5 Ratio of Bachelor's and Master's Thesis 33
4.6 English Language Adoption in Theses: FI MUNI and RH 34
5.1 Lifecycle of a Thesis Topic: A State Diagram 40
5.2 BPMN Diagram Showing the RH Thesis Process 43

1 Introduction

The processes within large companies are prone to constant changes due to factors such as personnel changes, legacy systems, or external variables. These changes can have a significant impact on the overall efficiency and effectiveness, making it necessary for the company to regularly review and adjust its processes.

Red Hat, as a large corporation with a lot of processes, has been a partner of universities in Brno since 2007, with a key aspect of this collaboration being their cooperation on a final thesis. Despite this, when the partnership first began, Red Hat lacked a specialized team or designated employees dedicated to university cooperation. As a result, processes were established with dependencies on individuals and factors that have since changed, leading to potential inefficiencies and limitations.

To ensure the long-term success of the thesis program, there needs to be a comprehensive review and revision of the processes related to the thesis writing process. This will help to eliminate any potential roadblocks and ensure that the program can continue to operate effectively and sustainably in the future.

The primary objective of this thesis is to conduct a thorough evaluation of the program and lifecycle of theses at Red Hat Czech, with the aim of enhancing and formalizing existing procedures and introducing new ones where necessary. This will involve consolidating historical datasets and augmenting existing data which is required for establishing data analysis and trend tracking.

The thesis will also provide readers with an overview of student cooperation with Red Hat in different regions, in order to deepen their understanding of the unique position of the Czech thesis program within Red Hat. This overview will draw on the necessary context covered in Chapter 2 and findings from a series of interviews conducted with regional managers in different regions where Red Hat Research operates. The findings from these interviews will be presented in Chapter 3.

Moreover, to enhance the sustainability of the program, this thesis will aim to streamline the process of importing data from the university information systems to Red Hat's internal database, with a specific

focus on the two largest university partners in Brno, Masaryk University and the Brno University of Technology. This will involve extensive data analysis to identify areas for improvement, which will be covered in chapter 4. The thesis will also include a comprehensive overview of the program's current health, as well as actionable insights to guide necessary future improvements.

One of the objectives of this thesis will be to refine the process with the goal of reestablishing the open thesis section on the Red Hat Research website. This includes implementing the necessary mechanisms to support the tracking of topics on the web. The website's reopening can increase the number of thesis projects in collaboration with Red Hat and expand the variety of research areas covered. To achieve this, the thesis will involve the creation of internal processes, which will be discussed in chapter 5.

Furthermore, supportive documentation for both new and existing processes must be developed and existing thesis creation processes must be streamlined to enhance the effectiveness of the program and reduce the administrative workload on employees. All these changes must be approved by the company's internal stakeholders and management.

2 Context

This chapter aims to provide the reader with the necessary context for understanding the thesis topic and the proposed process change. In large companies, changing a process can bring risks that negatively affect its efficiency, functionality, or even its existence. To mitigate these risks, a thorough understanding of the process's context is necessary. This is particularly important for a long-standing process with multiple stakeholders, as is the case in this thesis. Examining the historical context of the process and recognizing stakeholders is essential for understanding how the process has evolved to its current form and ensuring its longevity in the future.

In this chapter, the focus will be on Red Hat as a company, with a specific emphasis on the Red Hat Research team, which is responsible for the thesis program. Section 2.1 will provide an overview of the company's general context, including information about its history, mission, and goals. Section 2.2, on the other hand, will focus on RHR's global approach to university cooperation. While this chapter will offer a broad description of the team and its work, readers seeking more detailed information on the team's activities in specific regions and any regional differences should refer to chapter 3.

2.1 About Red Hat and Red Hat Research

Red Hat is a software company that primarily specializes in open-source software development and enterprise solutions. The company was founded in 1993 and as of 2023, Red Hat has roughly 15,000 employees across the world [1]. In 2019, Red Hat was acquired by IBM, but it continues to operate as an independent subsidiary of IBM, maintaining its unique culture and approach to the open-source community. With a focus on collaboration, transparency, and community-driven innovation, Red Hat is committed to delivering high-quality, secure, and flexible technology solutions to businesses and organizations of all sizes.

Red Hat is a company with a deep-rooted commitment to open-source software and actively contributes to the advancement of open-source technologies. The company has earned its success through its

leadership and participation in the open-source community, as well as its contributions to a vast array of open-source projects. Red Hat is widely recognized as a leader in the open-source software industry and its reputation has been built on its commitment to the open-source philosophy.

Among their many products, some of the most notable include Red Hat Enterprise Linux, which is a leading enterprise-class operating system, and Red Hat OpenShift, a container application platform that allows for rapid and scalable application development. They also offer Red Hat Ansible Automation Platform, which is a leading IT automation and management tool that helps organizations automate and streamline their IT operations.

This thesis is being conducted in collaboration with Red Hat Czech, a subsidiary of Red Hat Inc. Red Hat Czech is located in Brno, Czech Republic, and is one of the key players in the open-source software industry in the region. The company operates in Brno since 2006 and has over thousand employees, who specialize in a wide range of open-source technologies.

Red Hat Research (RHR) is a multidisciplinary department dedicated to supporting research across various domains [2]. Rather than conducting the research itself, RHR serves as a catalyst for innovation by connecting Red Hat engineers with professors, researchers, and students, fostering collaborations that bring new ideas and technologies into open-source communities. These collaborations serve as breeding grounds for novel ideas and cutting-edge technologies that enrich open-source communities worldwide. By acting as a robust support mechanism, RHR enables researchers and innovators to leverage Red Hat's expertise in open-source technologies and communities. This empowers them to bring their ideas to life and make a more significant impact.

Operating across multiple global branches, Red Hat Research strategically positions its main branches in the USA, the Czech Republic, and Israel, each with a dedicated focus on specific universities and research domains. Collaboration lies at the heart of RHR's mission, and its branches collaborate closely on international initiatives, including prestigious events such as Research Day [3]. However, RHR also recognizes the importance of regional differences, and each branch carries

its unique responsibilities based on the needs of the local universities and researchers it supports.

Despite these differences, RHR's global approach to supporting research remains consistent. The organization leverages its expertise in open-source technologies to facilitate collaborations and progress across different domains.

2.2 Cooperation with Universities

Cooperation with universities is one of the key responsibilities of Red Hat Research. The nature of this collaboration varies depending on several factors, making each partnership unique. RHR's collaboration with universities is based on a shared commitment to advancing research and innovation. The department works closely with university researchers, professors, and students to identify areas where it can provide the most value.

This section will describe the various forms of university collaboration within the RHR global team. Following this in 3, we will focus on a more detailed explanation of university cooperation in separate regions, the USA, Israel, and the Czech Republic, which is particularly relevant to this thesis. By examining RHR's university cooperation in detail, we can gain insights into the specific ways in which this thesis can contribute to the goals of the regional RHR team, as well as support the broader objectives of the global department.

As previously discussed, RHR is a globally distributed department that is structured into smaller regional teams. Each team has its own unique agenda and goals, which are often tailored to the specific needs and priorities of the region it serves. The purpose of the department is to connect Red Hat engineers with professors, researchers, and students to bring great research ideas into open-source communities, and while there may be some regional differences, this goal remains consistent across all regions. It is important to note that, rather than leading research efforts, Red Hat Research primarily serves as a coordination team that facilitates research. This is achieved through multiple means, including organizing conferences to connect researchers from different communities and providing resources to help bring research ideas into open-source communities.

RHR provides financial support to multiple researchers at universities who are conducting research in various areas such as AI/ML, cybersecurity, cryptography, and more. Additionally, the department facilitates numerous university courses taught by Red Hat engineers at various universities, including Massachusetts Institute of Technology, Boston University, Northeastern University, University of Massachusetts Lowell, as well as MUNI and BUT in Brno. RHR also collaborates with government agencies and various private institutions to promote open-source ideas and communities. It is worth noting that Red Hat has additional university collaborations across Europe, including partnerships with Newcastle University, Karlstad University, and Graz University of Technology, among others. These collaborations, although valuable, are of a smaller scope and therefore will not be examined in detail within this research.

Collaborating with academia is usually a very long-term investment that can yield uncertain results. "The innovation process requires frequent and iterative interactions among innovative actors, in which universities and private companies are central." [4] As a company committed to open source and innovation, Red Hat recognizes the value of working with universities and academic institutions. However, it's essential to set the expectations right and handle this unique relationship so that both parties can benefit from this partnership. That's where the Red Hat Research team comes in.

Supporting research involves much more than just providing funding. Handling multiple educational institutions, each with different rules and systems, connecting the right academic people with business people, working with grants, organizing events to attract new talent, and even doing internal marketing to justify the spending to shareholders and employees of the company are all part of the responsibilities of Red Hat Research employees. Despite the significant investment of time and resources, we believe that the results from these collaborations, such as advancements in technology and the inclusion of diverse communities, make it all worth it. Our work with academia helps us to drive innovation and contribute to a broader social impact, which is essential to the mission of our department and Red Hat as a company.

3 Overview of Student Cooperation in Different Regions

As a part of this thesis, I've conducted a series of interviews with managers from several regional branches discussing the regional differences and forms of university cooperation. Interviews were conducted separately with each manager with questions prepared beforehand. The objective of these interviews was to acquire a comprehensive overview of the regional differences within RHR and their impact on university collaborations in each area. The focus of this research was to investigate the regional departments' experience collaborating with universities and students in their region, focusing on the thesis writing process, and any difficulties or prospects encountered. The information collected from this interview will assist in recognizing best practices and identifying potential areas for improvement in our department's collaboration with universities. Additionally, this overview will provide an explanation for our decision not to extend our efforts to export the thesis writing process to other regions outside of the Czech Republic.

The insights gained from the interviews with regional managers highlighted the complexity and diversity of collaboration in different regions. While the findings were intriguing, the limited scope of the interviews only provided a preliminary understanding of this subject. Based on our findings, we recommend conducting further research to fully comprehend the nuances and impacts of regional differences.

The interviews revealed that some issues in each region overlap, indicating a potential for a common solution. However, regional and historical contexts often influence collaboration practices, making it difficult to fully understand and modify them. The diversity in collaboration practices across regions also presents an opportunity to learn from each other and implement best practices from successful collaboration models.

The interviews with regional managers provided valuable insights into the unique position of the Czech thesis program within the larger Red Hat ecosystem. The differences in regional contexts and forms of collaboration highlight the need for a nuanced approach to improv-

ing the program's processes and sustainability. While this chapter only scratches the surface of this complex topic, it has provided a foundation for future research and potential solutions.

Due to the lack of quantitative data and the presence of various variables, making definitive observations is challenging. However, based on anecdotal evidence gathered from interviews, it is apparent that different regions have distinct approaches to supporting research collaborations. In the United States, the emphasis seems to be on supporting concrete projects. In Israel, RHR focuses on establishing long-term collaborations with professors who oversee research and involve students. In the Czech Republic, the primary focus of support lies with the students themselves. While these observations are based on anecdotal evidence and exceptions may exist, they serve as a starting point for future research and discussions on the benefits and drawbacks of each approach. Further exploration of these regional differences is warranted to gain a deeper understanding of their implications.

3.1 Czech Republic

Around 35% of young adults aged 25-34 in the Czech Republic have completed higher education, which is lower compared to the average of 47% in other OECD countries [5]. Despite a lower percentage of young adults completing higher education, the country offers free tuition at public universities and does not mandate military service. Since 2001, the Czech Republic has implemented a three-cycle structure in its higher education system, comprising bachelor's, master's, and doctoral study programs. The typical length of study programs varies across disciplines, with bachelor's programs typically lasting three years, master's programs lasting two years, and Doctoral programs lasting three to four years, depending on the field of study.

Red Hat's partnership with universities in the Czech Republic dates back to 2007, just one year after the regional branch was established in Brno. As one of the first companies in the region to start collaborations with faculties, Red Hat's efforts began from the grassroots level. These efforts started "bottoms up", driven by one of Red Hat's employees who wanted to lead a university course. The initial form of collabora-

tion involved teaching a DevOps course at the Faculty of Informatics of Masaryk University [6]. This led to other forms of involvement, mainly supervising bachelor and master theses by Red Hat employees. The thesis program quickly became one of the cornerstones of our collaboration with academia in the Czech Republic.

At that time cooperation between private businesses and public universities was a rather rare occurrence. However, this collaboration proved to be mutually beneficial for both Red Hat and the universities. Over time, Red Hat's involvement grew significantly. This amount of cooperation led to a need to establish a regional department for university cooperation. The university program in Red Hat CZ started in 2015, five years before the establishment of the global team, Red Hat Research.

Red Hat's positive experience working with students on their theses led to the establishment of an internship program, which deepened the mutually beneficial relationship between the company and the universities. By providing students with valuable work experience, Red Hat not only contributed to their education but also built relationships with promising new candidates. Red Hat's internship program has become a significant part of its workforce in Brno, with over 150 interns contributing to the company's operations in 2023.

Brno, with its two prominent technical universities in close proximity to Red Hat's offices, is an ideal location for fostering the academic community. Our partnership with Masaryk University and the Brno University of Technology serves as a cornerstone of our research program in the Czech Republic. These universities have been longstanding partners, and a significant number of our employees are alumni, establishing a strong bond between our organizations. Moreover, our active involvement in teaching at these universities further strengthens the connection and collaboration between academia and industry.

Red Hat is an official strategic partner of the Faculty of Informatics at Masaryk University and is actively engaged with the faculty on multiple projects [7]. As of 2023, Red Hat engineers are involved in teaching 12 different courses at the university, and the feedback from students regarding their teaching is overwhelmingly positive. Furthermore, in terms of collaborating on final theses, Red Hat supervises the biggest share of industry-led theses, surpassing other industry

partners of the faculty. Red Hat also supports multiple research internships for Ph.D. candidates at the faculty and has its own lab located on the premises of the building. The company's involvement extends to various research projects at the faculty, including the Trusted Computing Ecosystem [8], the Usable Security for Developers [9], and the Mining issued Common Criteria and FIPS140-2 certificates [10]. These projects represent a sample of the completed and ongoing research initiatives that Red Hat supports at the faculty.

In addition to these initiatives, Red Hat actively participates in various events such as Dni SPP or Film Festival FI. Many students have also chosen to intern at Red Hat while studying, and with the increasing trend of mandatory internships as part of the faculty's curriculum, we anticipate a rise in the number of students selecting Red Hat as their destination.

Red Hat Research maintains a formal partnership with the Faculty of Information and Technology at Brno University of Technology, which commenced in 2014. The collaboration has since grown stronger, with RHR holding a silver partnership status with the university in 2023 [11]. As part of the partnership, Red Hat engineers teach two courses at the university, namely Advanced Topics of Linux Administration and Binary Code Analysis. In addition to this, RHR supports a Ph.D. student who is working on Perun: Lightweight Performance Version System [12], and several other graduate and undergraduate students in their research. RHR also backs projects of its employees at the faculty, such as Viktor Malik's Automatically Checking Semantic Equivalence between Versions of Large-Scale C Projects [13]. We are also participating in events organized by the faculty, including Žijeme IT and Excel at FIT.

3.2 USA

The university systems in the USA and the Czech Republic are different in many aspects. These differences have various reasons such as government structure, historical backgrounds, or cultural norms. One of the most significant differences is the cost of a university education. While in the Czech Republic, public universities offer tuition-free education to domestic students, in the US, the cost of education can

be extremely high, and students often need to take out loans to cover the tuition. Additionally, the duration of studies in the two countries differs as well. In the US, a bachelor's degree typically takes four years to complete, and master's degree programs in the US usually take two years. However, the length of the studies in the US may vary depending on the program with many exceptions. In the United States, 51% of 25-34 year-olds had a tertiary qualification in 2021 [14]. The curriculum in the two countries also differs, with US universities offering generally more diverse and flexible course options [15], while Czech universities tend to have more fixed and structured programs.

The US education system is highly complex and includes a vast number of programs, courses, systems, and rules, which can differ significantly from one university to another and even within different programs at the same institution. Furthermore, state-level legislation can also impact the education system's structure and requirements. In fact, with so many variables, it is challenging to try to establish even a semi-unified process for some universities. This presents a significant challenge for Red Hat's efforts to collaborate with universities or to standardize the processes. However, understanding the system's intricacies and adapting to the unique needs and requirements of each institution can lead to successful partnerships and collaborations as we can see from the example of RHR US.

One of the most significant collaborations between Red Hat Research and academia is the Collaboratory with Boston University (BU). This partnership has produced impressive results, including four major upstream contributions in 2020 and 18 ongoing hybrid cloud research projects between Red Hat and BU, each with the potential to contribute to upstream projects [16]. The collaboration involves eight Ph.D. students financed by Red Hat and 11 BU professors directly engaged with Red Hat engineers on research projects. These joint efforts address important areas for Red Hat, such as improvements to the Hybrid Cloud and Linux. The Boston University Collaboratory serves as one of the best examples of the benefits that can be achieved through close collaboration between industry and academia.

The Collaboratory offers a range of opportunities for student participation, including seminars, colloquiums, and courses. Researchers can also seek funding through proposal events in various research areas, with the requirement that projects are open-source and results are

publicly available. The Collaboratory will have a dedicated space at the Hariri Institute for Computing's new Computing and Data Sciences building at BU [17]. The Collaboratory is a component of the Mass Open Cloud Alliance, an initiative housed by the Hariri Institute [18]. It aims to establish an open research cloud that offers affordable and reliable resources to domain researchers. The initiative also encourages collaboration between academic researchers and open-source developers.

In the US, students can typically engage in research supported by RHR through university Collaboratories, internships, or Research Interest Groups (RIG) [19]. Internships in the US typically have shorter durations compared to those in the Czech Republic, with summer internships being particularly common. These internships often focus on specific projects within Red Hat, providing students with hands-on experience in a particular area.

The RIG initially originated as a monthly meeting among Red Hat associates, students, and academic researchers in the Boston area, serving as a platform for sharing ideas and fostering collaboration. The success of the RIG in facilitating idea exchange led to its expansion on a global scale, with the establishment of the Europe RIG in 2021. The RIG offers students a valuable space to learn about Red Hat's research endeavors, pitch their own ideas, and discover projects in which they can actively participate. It serves as a dynamic hub for students to engage with researchers and contribute to ongoing research initiatives.

One notable difference relevant to this thesis is that undergraduate programs in the United States typically do not mandate a final thesis as a compulsory part of their curriculum. While some graduate programs may require a thesis, this requirement is not consistent across all fields of study. Additionally, it is not standard practice for companies to collaborate with students on their theses. As a result, implementing a thesis program in the US could present significant challenges. In contrast, the success of the thesis program in RH CZ is partly due to the cultural significance of writing theses in the Czech Republic. RHR US collaborates with students through various other programs, but not specifically through theses.

3.3 Israel

Israel's education system shares a similar structure with the Czech Republic. In Israel, education is mandatory for children between the ages of 3 and 18, with three years of preschool and nine years of primary and secondary education [20]. After mandatory military service, students can choose between vocational training or pursuing tertiary education. The country has multiple world-renowned universities, both public and private. In Israel, 46% of 25-34 year-olds had a tertiary qualification in 2021 compared to 47% on average across OECD countries [21]. This investment in human capital has contributed significantly to the country's economic growth and development over the years. Moreover, Israel's thriving start-up culture and world-class technical universities, such as Hebrew University and Israel Institute of Technology (Technion), make it a leading country in research and development, with significant contributions to technological advancements.

As a part of our research, we conducted an interview with two research supervisors from Red Hat Research in Israel. In 2019, cooperation with universities in the region began, unfortunately, the COVID-19 pandemic occurred just one year later, posing significant challenges for the team. Despite this, the team established cooperative relationships with five different universities and has made considerable progress in various ongoing projects. One of the notable projects among these collaborations is SpotOS [22], a distributed cloud-based operating system over unreliable resources. This project is being worked on at Technion. The innovative nature of SpotOS has caught the attention of the tech industry, making it a promising achievement for the Red Hat-Technion partnership.

Red Hat Research in Israel has adopted a distinct approach toward research collaborations as compared to our operations in the Czech Republic. An important difference lies in the way collaborations are initiated. In Israel, Red Hat Research seeks to establish collaborations with professors who are then entrusted with the responsibility of leading the research project. These professors often establish practical courses, offering students the opportunity to enroll and actively participate in the research. Red Hat provides the necessary resources, consultations, and exposure. This approach is particularly interesting

as it prioritizes long-term collaborations with academic institutions rather than short-term, project-based collaborations. By engaging with professors, Red Hat Research is able to benefit from their knowledge and experience, which in turn enables us to produce research outcomes that have a greater impact on the industry.

Another noteworthy difference in the Israeli approach to research is their attitude toward student collaboration. While a final thesis is a requirement for completing a master's degree in both countries, students finishing their bachelor's degree are usually not required to do a thesis in Israel. However, despite the potential benefits of cooperating with students on their thesis and improving relations with the university, Red Hat Research in Israel does not engage in such collaboration. The focus remains on establishing collaboration with professors instead of individual students. This approach highlights a clear separation between academia and commercial companies, and it is uncommon for students to do their thesis outside of the university setting. Students who want to work on a research project supported by Red Hat usually enroll in a seminar supervised by a professor associated with Red Hat. Graduate students receive funds from Red Hat distributed by the professor as well as credits from a university.

4 Analyzing and Tracking the Thesis Program Data

This chapter aims to analyze data from three separate internal Red Hat databases containing information about all the successfully defended theses dating back to 2014. The primary goal of this analysis is to provide a quantitative overview of writing processes within the company and to identify current trends in these processes. In addition, this chapter will detail the consolidation of historical datasets that was necessary to achieve a thorough and accurate data analysis. By examining the data in depth, we can gain valuable insights into the evolution of writing processes at Red Hat and identify potential areas for improvement.

This chapter is divided into three sections, each focusing on a specific aspect of our thesis program data analysis. The first section will discuss the data collection process and the design of a new database to better store and manage our data. The second section will cover the data cleaning process and the consolidation of historical datasets, as well as the development of a data tracking script to simplify the tracking process for employees. This section will also highlight some challenges we encountered with the automation of data entry and how we addressed them. We hope that the data tracking script we developed can be useful for other companies collaborating with universities on theses as well. In the third section of this chapter, we will present our findings from the data analysis, highlighting some trends and insights. We will also provide recommendations for future actions that can help improve the effectiveness and efficiency of the thesis program in Red Hat.

4.1 Data Collection

To begin the data analysis, it was necessary to bring together information from different spreadsheets stored on Google Drive. The data was stored in three separate spreadsheets for MUNI, BUT, and other universities. Each spreadsheet had one person responsible for regularly updating the data. However, because of irregularities in the

information systems of universities, there were differences in the types of data and the format of data. Moreover, the manual import of data meant there were a lot of human mistakes. More details about the process of importing the data to the internal RH database from the information system are provided in chapter 4.2.

In order to consolidate the data, we needed to decide which parameters were necessary and feasible to track. We based our decision on two main criteria: the data needed to be easily collectible and at the same time valuable enough for RHR. The format and categories of the final database are also heavily affected by the format of the imports from the information systems of universities. To protect the privacy of Red Hat employees and students, the datasets used in this thesis will not be included. Instead, anonymized graphs and source code will be provided to illustrate the findings of the data analysis. The resulting heading of our final spreadsheet is described in table 4.1. To improve the efficiency of data entry both now and in the future, we have developed a set of scripts outlined in chapter 4.2 that can automate most of the process of data insertion and cleaning. Nevertheless, it is important to note that the "Main Red Hatter", "Link" and "Grade" fields cannot be filled out using these scripts. The "Link" field provides access to the webpage where the rest of the data can be filled out, and determining the primary Red Hatter involved can often be challenging to fully automate.

Table 4.1: Header of the spreadsheet used for tracking final theses

Year	University	Faculty	Type	Student	Supervisor

Consultant	Main Red Hatter	Title	Abstract	Grade	Link

Dean's award	Keywords	Red Hatter	LDAP username

Language	Comment

- **Year** - This field represents the year in which the thesis was successfully defended. We do not distinguish between winter and

summer semesters, and we only take the year of the successful defense as it appears in the university's Information System.

- **University** - This field contains a formalized shortcut for the university where the thesis was defended, such as BUT, MUNI, UPJŠ, etc.
- **Faculty** - This field also contains a formalized shortcut for the faculty where the thesis was defended. In the future, if we collaborate with multiple faculties within the same university, this field will allow us to differentiate between them.

  Type - This field refers to the type of thesis, either bachelor's or master's.
- **Student, Supervisor, Consultant** - These fields contain the full name of the student, supervisor, and consultant involved in the thesis project, with the surname listed first.
- **Main Red Hatter** - This field indicates the most involved Red Hatter in the project, typically the consultant, but not always. This field must be manually filled out.
- **LDAP username** - This field will be used to connect our database with the employee database for future data analysis purposes. LDAP username can be obtained from the name of the most involved Red Hatter.
- **Title, Abstract** - These fields contain the same information as found in the university's Information System.
- **Grade** - This field must be manually filled out and indicates the final grade earned by the student for their thesis project.
- **Link** - This field contains a valid link to the public webpage of the university that contains the thesis.
- **Dean's award** - This field is a simple yes/no indicator of whether the thesis received a dean's award.
- **Keywords** - This field contains a list of keywords associated with the thesis project.

- **Language** - This field indicates the language used in the thesis.
- **Comment** - This field allows for any additional comments or notes to be included about the thesis.

Our final dataset contains almost 400 entries since 2014 with students from multiple universities across the Czech Republic and Slovakia with the overwhelming majority of theses done at Masaryk University and Brno University of Technology. Despite our best efforts, some entries in the dataset lack certain information, and we were unable to collect complete data for periods prior to when the tracking initiative began. We decided to redesign some of the fields we track to match the format and add some new fields that may have interesting statistical significance. Due to the amount of missing data and the newly added fields, we had to start from scratch and reconstruct the database from public university web pages. The new database only uses grade and public website links from the old database, but with this limited information, we were able to create a more comprehensive and accurate dataset. However, in accordance with best practices [23], we have kept the old database stored as a backup.

The resulting dataset provides an overview of the theses successfully defended since 2014 with partial information about failed theses as well. Unfortunately, we were unable to obtain data prior to this timeframe due to the lack of records and limitations in keyword searching. Nonetheless, data from over a decade ago would likely have limited relevance to our research and improvement of the current process.

4.2 Data Cleaning and Future Tracking

The goal of the data-cleaning phase is to ensure the necessary quality of the data. Fortunately, most of the data cleaning problems were single-sourced on an instance level [24]. Although removing these errors can be time-consuming, it is relatively straightforward. Common issues found in our datasets included varying date formats and different representations of the NA symbol, such as empty strings, dashes, and variations of whitespace. These illegal entries were substituted by NA in data analysis software. When cleaning data, other types of issues were outdated link formats, links requiring authentication, and

inconsistent name formats. Fortunately, most of these issues can be easily resolved by implementing simple solutions such as updating links to the current public page with a simplified format and standardizing name formats. It's important to note that data cleaning is an always ongoing process that requires regular attention to ensure data accuracy and reliability.

During the data-cleaning phase, the most common issue we encountered was missing values in the dataset. The most important values that were usually missing were from the "Grade" section in the BUT dataset and from the "Topic" section in the MUNI dataset. However, missing values were far too common in other fields as well. In the case of MUNI, data was usually available on the university website. However, for the BUT dataset, we couldn't easily access the grade information from their information system, which led to incomplete data that had to be filled in manually. Furthermore, since we added new fields to the database, it was needed to update each row with new data. The main goal of this database is to maintain an accurate record of the collaboration on each thesis. Therefore, instead of imputing missing data, we decided to fill out as much missing data as we could and exclude the incomplete data from the rest of the analysis.

Filling in missing data in a database can be a tedious task, especially when it comes to important fields like "Topic", "Abstract", "Keywords", and identifying the primary Red Hatter involved. To streamline the process, I developed a web scraping script that automated the filling in of missing data from MUNI and BUT. However, manually identifying the "Most involved Red Hatter" remains necessary. To tackle this, we reached out to past supervisors who helped us complete this task. The amount of missing data meant that we had to recollect almost all of the data from the information systems of the universities.

Because our data was stored in a private Google Drive in Google Sheets, I chose to write the web scraping script using the Apps Script extension. In addition to its immediate use, this script has the potential to improve future data tracking by reducing the number of errors. "Prevention is better than cure" [25].

Apps Script is a cloud-based JavaScript platform that is free and enables easy manipulation of data in Google Sheets [26]. Apps Script offers a user-friendly interface that enables even inexperienced users to execute the scripts we developed within the web application. This is

especially important for the future of the process, so the tracking can be maintained even by the inexperienced user with enough guidance. Furthermore, Apps Script can enhance the functionality of basic Excel formulas, allowing for more advanced data manipulation.

In this script, I have opted to utilize regexes over other popular web scraping techniques such as HTML DOM or Xpath. I arrived at this decision after weighing the advantages of regexes, such as ease of implementation and low memory consumption, as opposed to other alternatives [27]. However, it is important to acknowledge that regexes can be prone to errors and may not take advantage of domain-specific knowledge [28], which means that other techniques could also be viable. Nonetheless, given that a majority of the data could be easily identified through regular expressions from BibTeX, it was a logical choice to use them in our case. Additionally, regexes are relatively easy to modify in the future if there are any changes on the websites we're scraping. By defining the regular expressions at the beginning of the corresponding function, we can update them quickly, ensuring that the script will remain functional.

The user can run the script directly from the Google Sheets UI. Once the script is executed from the menu, a pop-up window prompts the user to input the starting row for the script. The script will then run from the designated starting row, provided there is a functional link in the "Link" column that directs to the official public MUNI or BUT webpage with the required thesis information. The main body of the code is a cycle that iterates over rows in the spreadsheet, starting from the user input. After obtaining a functional link from the row, the script identifies the university and uses appropriate regex patterns needed to extract the required data (each university uses different regexes) from the webpage's source code. The loop skips the rows of universities other than BUT or MUNI since data from these universities need to be filled in manually. The regexes are then matched with the website's source code, and the extracted data is written to the designated cell in the spreadsheet. While simple regex can be used to extract most of the data from the BibTeX section of the webpage used for citation, more complex regex strings are required for categories such as Abstract, Consultant, and Keywords.

During the development of this script, we encountered several difficulties. Firstly, the information system sometimes contained in-

complete or inconsistent data, which had to be taken into account. These inconsistencies were more prevalent than we initially thought especially in the Information system of Masaryk University. For example, sometimes the name of the opponent was missing, and at other times, the language in the citation was not provided. These irregularities made it challenging to find patterns that would work in all cases. To address these issues, we implemented a pop-up window to indicate the row number that requires manual filling. However, this was just a warning, the inconsistency rate on MUNI's webpage was too high - about 1 in 5 theses, making it almost impossible to automatically import the data. With the improved and much more complicated regexes we have implemented, we have managed to bring the inconsistency rate down to around 1 in 30, and with future improvements, we believe we can bring this number even lower.

Secondly, we faced difficulties in identifying the appropriate regex for specific fields due to variations across faculties. Even though both universities use LaTex citations on their web pages, they use different formats. We had to use separate functions and regexes for each university, which unfortunately resulted in an increased level of complexity and repetition in the code. This made it more difficult to refactor the code. Thirdly, the BUT webpage lacks the consultant field, which usually displays the name of the main Red Hatter involved. The public page also does not indicate that it is in cooperation with Red Hat, and we had to retrieve the data from the old database. On the public webpage, you also can't find information about the dean's awards. Although we expect these issues to be resolved in the future, for now, we had to work with the available data.

4.3 Data Interpretation

With our data now properly cleaned, we can now move on to the most intriguing part: data analysis. Our main objective was to identify trends and provide a quantitative overview of the thesis writing process. In this section, we analyzed a dataset containing only successfully defended theses, for which we had complete information. This was done to ensure the quality of the data analysis. Out of the 394 theses collected, only 342 met this criteria and were used in this section. The

remaining entries were mostly theses that lacked some critical data, were still in progress, or were unsuccessful.

There are several ways to analyze the data, and in this section, we will present some of our key findings. Firstly, we will examine the number of theses completed in cooperation with Red Hat over the years. Then, we will analyze grades at different faculties and examine the reasons for variations at different universities. Additionally, we will provide an overview of some thesis characteristics, such as the most common areas, languages, and types of thesis. We will also explore the variety of supervisors involved in the thesis writing process over time and how we can potentially improve this number.

There is still a lot more to explore in our future work with the extensive amount of data we have collected. Our plan is to connect the database to our internal LDAP system to gain deeper insights into how our thesis program affects recruitment, Red Hat employees, and research collaborations. The findings presented in this chapter can provide a great starting point for further analysis and could potentially inspire other companies looking to improve their thesis writing process. Additionally, we will draw attention to some of the current challenges with the program and suggest potential solutions to further enhance this program.

4.3.1 Thesis Production Over Time

In this section, we will analyze the production of theses in collaboration with Red Hat Czech over time. The first graph we will present shows the total number of theses completed each year (see Figure 4.1).

The graph clearly illustrates the pattern of theses completed in collaboration with Red Hat Czech over time. The x-axis represents the timeline, while the y-axis shows the cumulative number of theses completed across all universities. The graph displays an upward trend from the beginning until 2017, with a large dip during 2016. However, since 2018, there has been a consistent decrease, with the steepest decline occurring in 2021. Nevertheless, the year 2022 saw a notable increase, although the downward trend may continue in 2023. The graph also provides information on the universities where the theses were conducted. We have specifically highlighted the two

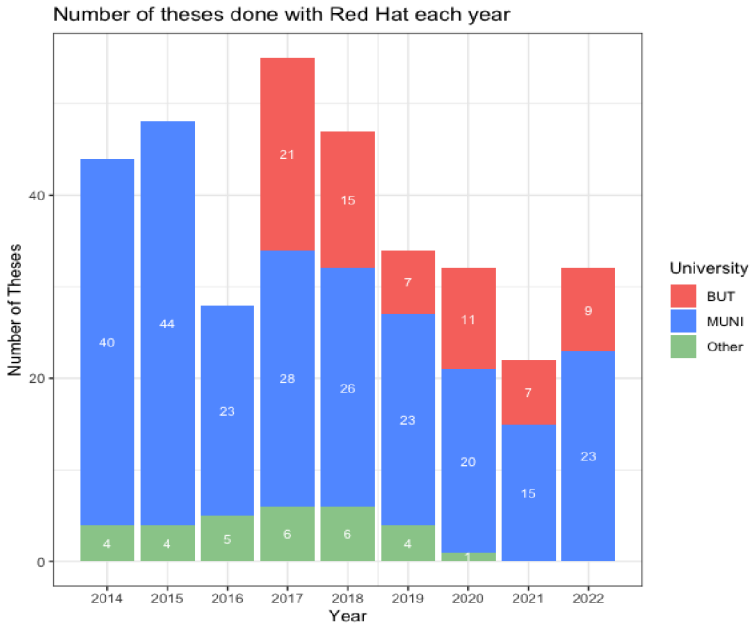

Figure 4.1: Thesis Completion Statistics: 2014-2022

largest universities, BUT and MUNI, as the number of theses from other universities is relatively small.

Firstly, we note that the number of theses completed in cooperation with Red Hat is much higher at MUNI than at BUT. This could be attributed to the fact that these universities have different institutional approaches toward collaboration. While FI MUNI encourages collaboration on theses BUT places greater emphasis on maintaining a clearer separation between academia and industry. This is evident in their respective approaches, where FI MUNI wants each partner to collaborate on at least a certain amount of theses, while BUT limits the number of theses partners can collaborate on, based on the level of partnership. Additionally, it is important to acknowledge that the history of university collaboration with Red Hat began at MUNI, which could also explain the higher number of theses completed in cooperation with Red Hat at this university. Regrettably, the dataset we have available does not contain information on the collaboration between Red Hat and BUT before 2017, and we were unable to collect

the missing data for that period. The tracking of this collaboration started in 2017, which makes it challenging to obtain historical data. Similarly, gathering data for the collaboration with Masaryk University before 2014 also encountered similar difficulties. Therefore, while there have been collaborations with BUT in 2016 and possibly earlier, the exact extent and number of collaborations during that period remain uncertain due to the lack of data.

The rise in the number of theses at the beginning of the graph can be attributed to the novelty of the student-company collaboration concept during that time, which generated considerable interest among students. Red Hat's partnership with the university across multiple courses, combined with its reputation as a prominent industry brand, made thesis cooperation with the company an attractive option for many students. Additionally, there was relatively low competition from other companies at that time. However, it's worth noting that we lack data from before the winter semester of 2013, as many theses were completed without tracking and there was no system of tags and filters in the Information System of Masaryk University. While the lack of data from BUT may explain the uncertainty regarding collaborations during that period, it does not fully account for the drop in the number of theses at MUNI.

The declining number of theses can be attributed to various factors, both internal and external. External factors include increased competition from other companies and the growing trend of industries collaborating with students on final theses. It was anticipated that these factors would have an impact on the number of theses submitted, and it would be unrealistic to expect Red Hat to maintain such a high proportion of theses with increased competition. Internally, Red Hat has faced challenges due to missing processes and a decline in the variety of available supervisors, as depicted in figure 4.3. The COVID-19 pandemic has also had a significant impact on the program, resulting in a steep drop in the number of theses submitted in 2021. However, after the initial disruptions caused by the pandemic, the program appears to be recovering and moving back on track.

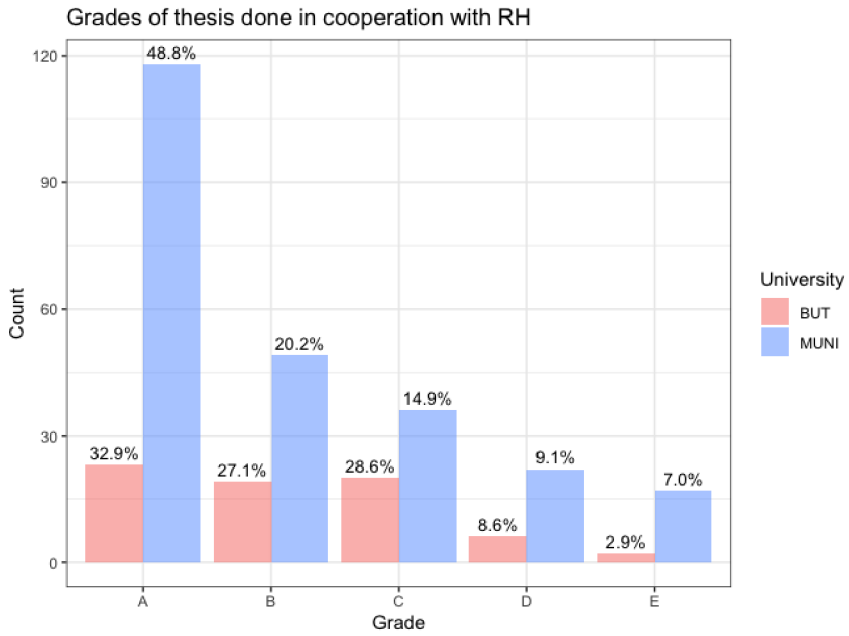

Figure 4.2: Comparing Grades of BUT and MUNI

4.3.2 Grade Analysis

One approach to exploring the data further is to conduct a grade analysis. Figure 4.2 presents a bar chart that displays the distribution of grades ranging from A to E on the x-axis and the frequency of each grade on the y-axis. The chart depicts the grades received by students who completed their theses in collaboration with Red Hat at the Faculty of Informatics of Masaryk University and at the Brno University of Technology.

Our initial observation is that the grades of defended theses at MUNI tend to be better than those at BUT. This difference could be due to variations in the average grade distribution among the faculties or potential issues in our own processes. Unfortunately, we were unable to obtain statistics for the average grade from BUT, which is something we will address in our future work.

However, when we compare the grades of MUNI theses completed with and without Red Hat's cooperation, we find that the ones done

in collaboration with Red Hat tend to have a similar average grade to those completed within the Faculty of Informatics at MUNI, but with a higher proportion of A grades. Analyzing the data for theses completed with and without Red Hat's involvement, we observe a greater prevalence of A grades in the theses completed in collaboration with Red Hat at Masaryk University. This indicates that partnering with Red Hat can have a positive impact on the overall quality of the final thesis, potentially leading to improved academic performance.

Additionally, we have noticed that the number of failed theses at BUT is relatively high, with 8 instances occurring in the past two years out of a total of 24 theses. However, these failed theses were not included in the graph due to the lack of data on failed theses at MUNI for comparison. It is worth noting that our preliminary research suggests that a similar problem may exist at other universities as well, although on a smaller scale.

Our current system faces challenges in tracking failed theses as we only receive notifications about the outcomes of thesis defenses, which means we may overlook cases where students abandon their topics before the defense. We are hopeful that our new system will enable us to track these cases more effectively, but it will require additional efforts.

In order to draw definitive conclusions about the phenomenon of failed theses, it is necessary to gather more data and conduct further research beyond the scope of this thesis. Based on our current hypothesis, we speculate that stricter deadlines at BUT, which are set much earlier in the semester compared to MUNI, may contribute to a higher number of failures. The deadlines for the final thesis description at BUT are more than 6 months earlier, and thesis defense takes place only once per year, unlike the twice per year defense system at MUNI. However, we currently lack sufficient data to confirm this hypothesis or draw definite conclusions. Addressing this issue is a priority for us, as it has a negative impact on the current program and warrants further investigation.

4.3.3 Dean's Awards at FI MUNI

We have a database of 242 theses completed in collaboration with Red Hat since 2014 on FI MUNI, with 29 (11.98%) of them receiving the

dean's award. Additionally, we have data on 4174 theses completed at FI MUNI, of which 258 (6.18%) received the award. To test our hypothesis, we performed a two-sample z-test for proportions to determine whether there is a significant difference between the proportion of theses that receive awards in our company's database and the larger dataset from FI MUNI. Our hypothesis is that collaborating with Red Hat on final theses can increase the chances of receiving dean's awards compared to theses completed without our involvement at Masaryk University.

We formulate the following hypotheses:

H_0: The proportion of theses receiving awards in the company group is equal to the proportion of theses receiving awards in the university group, that is $p_1 = p_2$.

H_A: The proportion of theses receiving awards in the company group is not equal to the proportion of theses receiving awards in the university group, that is $p_1 \neq p_2$.

Here, p_1 represents the proportion of theses that received awards in the company group, and p_2 represents the proportion of theses that received awards in the university group. We used a significance level of 0.05 in this test.

After checking all the prerequisites for the test and calculating the test statistic, we can determine the p-value associated with the test by consulting a standard normal distribution table or statistical software. Using a two-tailed test with $\alpha = 0.05$, we obtained a p-value less than 0.0001, indicating strong evidence against the null hypothesis. As a result, we can reject the null hypothesis and conclude that there is a significant difference between the proportion of theses that receive awards in our company's database compared to the proportion of theses that receive awards in the larger dataset from FI MUNI.

The data indicates that we have a higher ratio of dean's awards than the standard on the faculty. Specifically, theses with our involvement receive more dean's awards compared to those completed without our involvement at FI MUNI. This finding has practical implications for students who aim to enhance the quality of their thesis or research by partnering with Red Hat. The fact that our thesis projects are linked

to exceptional academic results, promotes the company's brand and reputation at the university

However, it is important to remember that correlation does not imply causation, and the higher ratio of dean's awards for theses with Red Hat's involvement does not mean that Red Hat's involvement is the sole reason for the awards. There are other factors, such as the quality of the students, supervisors, or their field of research, which may also play a role. However, the data does suggest that Red Hat's involvement in thesis projects is associated with a higher likelihood of receiving dean's awards, and this can be seen as evidence that Red Hat's resources and mentorship contribute to the success of these projects.

The criteria for receiving a dean's award may vary, but typically such theses demonstrate originality, creativity, and a high level of scholarship. They showcase advanced research skills, methodology, and the ability to analyze and synthesize complex information. The role of the thesis advisor or opponent also cannot be overlooked, as they play a crucial role in guiding the research process. Additionally, it's important to recognize that a thesis receiving a dean's award doesn't need to be purely research-focused, as it can also involve high-quality implementation work. Nonetheless, we believe this finding means that theses with Red Hat collaboration can be of higher quality than the already high standard on the faculty.

By collaborating with Red Hat, students receive resources and mentorship, and we build relationships with potential future employees who demonstrate exceptional skills and potential in their fields. This can lead to future collaborations and partnerships, furthering research in cooperation with Red Hat at the faculty. Overall, the p-test provides strong evidence that our efforts in leading thesis projects are positively impacting the academic success of students and the company's reputation at the university. In summary, while the data does not prove a direct causation between Red Hat's involvement and the higher ratio of dean's awards, it does suggest a correlation and provides evidence to support the potential benefits of collaborating with Red Hat on thesis projects.

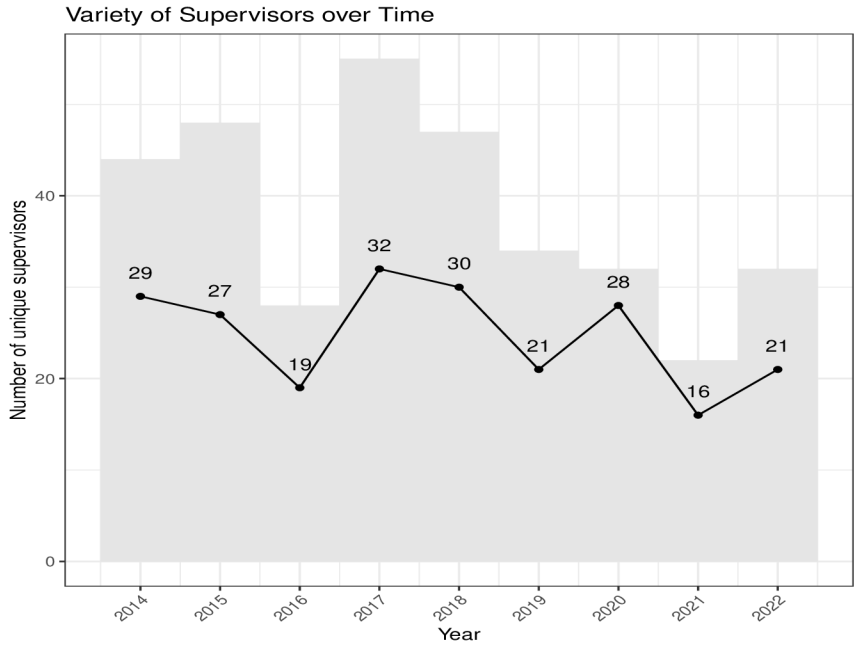

Figure 4.3: Variety of Supervisors

4.3.4 Analysis of Red Hatters Involved

The initial hypothesis of this research was that the decreasing number of theses each year was due to a decline in the number of unique supervisors from Red Hat. However, upon analyzing the data, this assumption proved to be incorrect. The graph in Figure 4.3 shows that while the number of unique supervisors per year varies significantly, there is no clear negative trend. In fact, when considering the overall decrease in the number of theses, the ratio of unique supervisors to theses is actually slightly increasing.

This finding suggests that Red Hat's involvement in thesis supervision remains steady, despite the overall decline in the number of theses being produced. This could be indicative of a shift in focus towards more targeted research areas, resulting in fewer but more impactful theses. Alternatively, it may reflect a more selective approach to thesis supervision, with Red Hat only supporting the highest-quality research projects.

As part of this chapter, we also collected information on the most involved Red Hat employees in the thesis program. We believe that these individuals should be recognized for their continued work with students, and we will share this information with their managers. However, the fact that only a few individuals are shouldering a significant portion of the theses may pose a risk to the future of the program if these employees leave Red Hat. This is another reason why it's important to diversify the pool of available supervisors. Additionally, we should explore if it's worth having consultants serve as formal leaders for the thesis program or if it's more beneficial for Red Hat employees to stay in a consulting role and leave formal supervision on faculty members who are not employed by Red Hat.

Overall, these findings have highlighted the need for a more strategic and targeted approach to our thesis writing program in order to maximize its impact and align with our research and priorities. While the lower number of theses can be beneficial if the quality is maintained, the lack of new supervisors is a concern. Often, only engineers experienced in this program lead theses, and it's difficult to attract new engineers. However, a positive aspect we discovered is that we have engaged over 150 unique Red Hatters as thesis consultants. This finding underscores the program's success in engaging a significant number of Red Hatters with academia.

As we move forward, we aim to conduct a more comprehensive analysis of our thesis writing program by integrating our database with LDAP. This will allow us to delve deeper into how the program contributes to the retention of former students as future Red Hat employees. We also plan to investigate which teams are more actively involved in the program and the seniority levels of our supervisors. Unfortunately, this data is confidential and will not be disclosed publicly.

By analyzing the data, we hope to identify successful strategies to engage even more employees in the program, thereby reducing the burden on our current supervisors and improving the overall quality and variety of research topics. To achieve this, we plan to implement supporting projects such as internal marketing within the company, benefits for supervisors, and extensive training. With these efforts, we hope to increase the involvement of our employees in the thesis writing program and ensure its continued success.

Additionally, during the analysis of the most involved Red Hatters in our thesis program, we discovered that there is a significant lack of gender diversity. Out of the nearly 400 theses defended at several universities over the last nine years, not a single woman has been marked as the main Red Hatter involved. This is a serious concern and highlights the ongoing gender equality issues within our field that need to be addressed. While we actively support several organizations focused on involving women in IT [29], we must also focus on efforts to diversify our own thesis program. RHR is committed to identifying and addressing any underlying issues that may be preventing female engineers from becoming more involved in the program. We can start by better marketing our thesis program to female engineers and actively seeking out experienced female engineers to become supervisors. By taking action to diversify our pool of available supervisors, we can ensure that our thesis program continues to provide valuable learning opportunities for students while promoting gender equality in our field.

4.3.5 Exploring Other Thesis Characteristics

This section should provide an overview on other various characteristics of theses done in Red Hat. There are some interesting similarities and patterns in these theses and we believe that understanding and leveraging these common signs can prove useful in the future.

The graph in figure 4.4 shows the most common keywords used to describe the thesis in information systems. The data shown is from MUNI and BUT, as we lack sufficient data from other universities. The graph shows that the keyword "java" is the most frequently used, with a considerable lead over other keywords, followed by "red hat." It is noteworthy that "red hat" is not the top keyword, despite the fact that all the theses in this study involve collaboration with Red Hat. This raises an interesting observation regarding the use of keywords to describe content. It prompts a question as to whether "red hat" should be used as a keyword in most cases when the research is not directly focused on the company itself. This finding emphasizes a significant issue discovered while working with keywords, which is a lack of consistency in keywords.

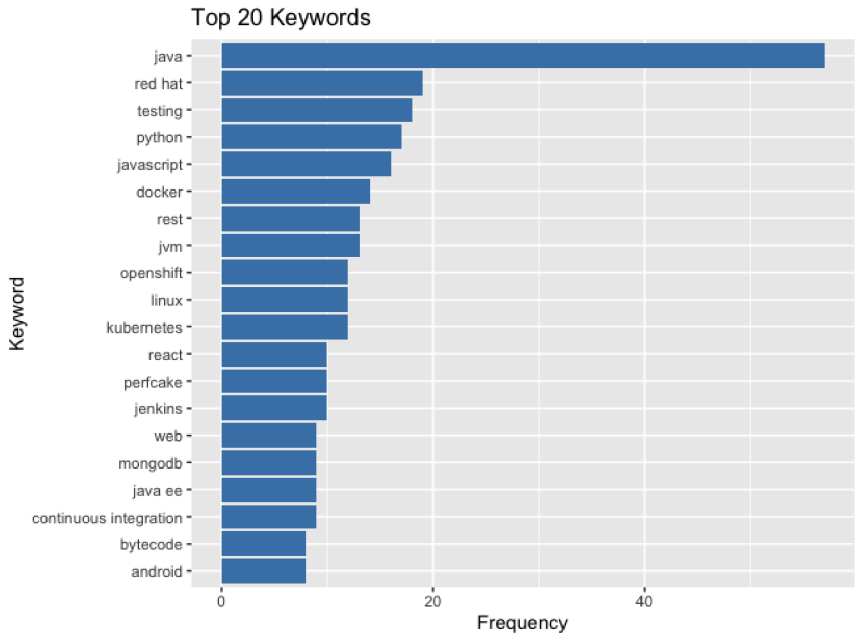

Figure 4.4: Most Common Keywords

In theory, keywords are meant to help users in searching and filtering through the database of theses. Unfortunately, currently, they lack consistency. This inconsistency can be observed on two levels. Firstly, in terms of format, keywords have inconsistent naming for similar things and often use shortcuts, and inconsistent lower/upper case letters, which makes data analysis and filtering challenging. In our study, we had to parse keywords and conduct additional data cleaning to obtain reasonable outputs. Secondly, keywords frequently lack descriptive power and are either overlooked or entered incorrectly. This issue not only hinders proper data analysis but also results in the loss of valuable input.

After cleaning keywords we observe that the most common keywords contain the usual technologies that we use in Red Hat. While this outcome was expected, it's intriguing to see how these trends evolve over time. When analyzing the data we noticed that the use of the most popular keyword Java is decreasing over time. It still remains one of the most popular keywords, however since 2018 it has been

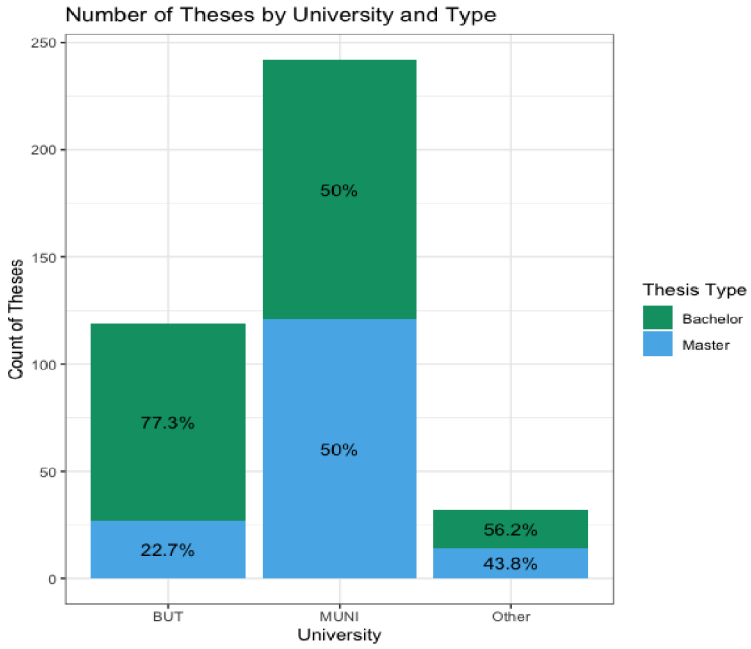

Figure 4.5: Ratio of Bachelor's and Master's Thesis

either in second place or tied for first. We also observed the emergence of other keywords such as Jenkins, web application, Kubernetes, and Docker, which are quickly gaining popularity but tend to fall off rapidly.

In figure 4.5, we can observe the number of bachelor's and master's theses completed over the years. As expected, the number of bachelor's theses is higher overall. However, it is surprising that the number of bachelor's and master's theses is somewhat comparable at Masaryk University. This is quite unexpected as a majority of interns in Red Hat Czech are undergraduate students, and thus, we anticipated the number of bachelor's theses to be much higher than the master's. There could be several reasons for this, including the fact that RH engineers are more inclined towards larger research projects typically associated with master's theses. In the future, we probably want to increase the number of master's theses as these students have a higher potential to produce exceptional research and even become potential employees.

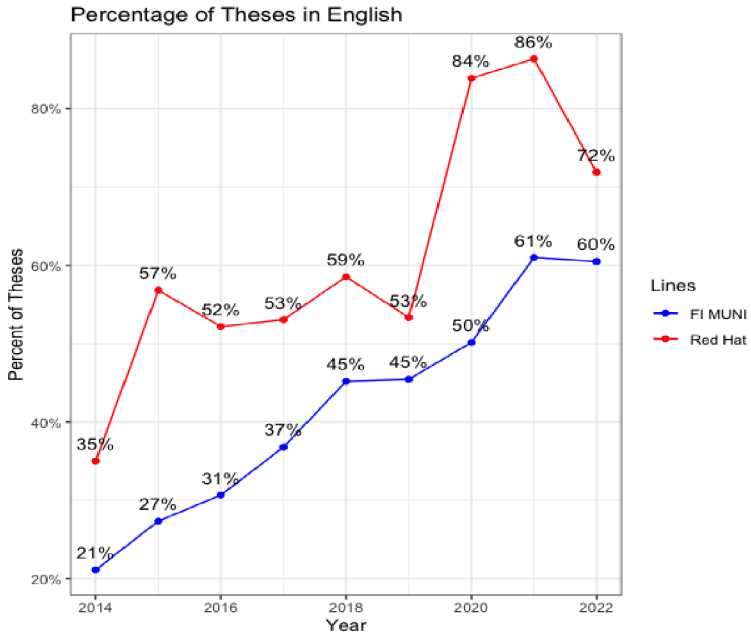

Figure 4.6: English Language Adoption in Theses: FI MUNI and RH

We conducted an additional analysis to examine the number of theses written in English. As research done in English is more accessible to a wider community, it is desirable to have a high number of theses written in English. This effort is also evidently supported by FI MUNI, which provides special scholarships for students writing in English [30]. To gather this data, we requested information from the faculty and compared it to our own database. The results of this analysis are presented in Figure 4.6.

As anticipated, the ratio of theses written in English exceeded that of the faculty. In 2021, nearly 86% of theses were written in English, marking a significant peak. We also observed a growing trend towards the use of English at the faculty. These findings suggest that there is a growing preference for English in the academic community and call for Red Hat to consider further efforts to publish even more theses in English. This could increase the accessibility of our research to a wider audience and contribute to the global dissemination of our knowledge.

5 Designing a New Process for Final Theses in Red Hat

As the thesis program has grown, the existing process for writing theses has become increasingly complex and outdated. As a result, it becomes essential to revisit and refine the existing processes to retain efficiency and ensure the program's continued success. This chapter aims to address this challenge by presenting a refined process for writing theses in cooperation with Red Hat.

The chapter starts by providing an overview of the motivation behind supporting the thesis program, as discussed in Section 5.1. It then delves into an in-depth examination of the current process, highlighting its challenges and issues, in Section 5.2. The final section, Section 5.3, presents a detailed design of the refined process. This includes the incorporation of diagrams and use cases, aiming to enhance the efficiency and effectiveness of the thesis program at Red Hat Czech.

5.1 Motivation for Red Hat to Support Theses

Red Hat is a company that is committed to innovation and collaboration with its communities. As part of this commitment, Red Hat Czech has established a successful thesis program that allows students to develop innovative projects while contributing to the community by providing experience and consultations. There are several other key motivations that drive Red Hat to support theses, including knowledge sharing, talent development, marketing, and community engagement. This section will explore these motivations in detail, providing insights into why Red Hat Czech values the thesis program as a critical aspect of its commitment to open source and innovation. We hope that this overview will serve as a compelling case for the continuation of the thesis program and contribute to the ongoing success of Red Hat's commitment to collaboration with academic institutions.

The OECD reports that the industry conducts about two-thirds of research and development in science and technology, with universities contributing around 20% and the government accounting for approximately 10% [31]. These figures emphasize the importance of

collaboration with academia. In the United States, there is a growing trend of support for universities and their research endeavors. For example, the University of Washington has set a goal to double the number of startup companies originating from their research within the next three years [32]. This initiative reflects the increasing focus on translating academic research into practical applications and promoting entrepreneurship.

However, Red Hat's motivation for supporting theses is not solely based on the desire to translate academic research into practical applications. Rather, it is driven by a deep commitment to collaboration, innovation, and the open-source community. The thesis program serves as a means to advance these goals.

One of the key motivations for Red Hat to support theses is knowledge sharing. Collaboration with academia is seen as vital for advancing innovation. Through the thesis program, Red Hat grants students access to its expertise, technology, and resources, enabling them to undertake innovative projects and gain practical experience with open-source software. In return, Red Hat benefits from the fresh ideas, diverse perspectives, and valuable skills that students bring to the company, thereby fostering innovation and driving growth.

Another significant motivation for Red Hat is talent development. The company recognizes the importance of attracting and nurturing talent in the industry. Supporting theses offers Red Hat an opportunity to identify and cultivate promising students. Through collaboration on their theses, Red Hat can provide mentorship, guidance, and training to help students develop the necessary skills and knowledge to thrive as professionals in the technology field. In return, Red Hat benefits from a pool of skilled individuals who are familiar with the company's culture, values, and technology. This motivation will be further reinforced by future work, where we plan to connect the dataset we created with the employee database, providing empirical evidence of talent development within Red Hat.

Thirdly, the thesis program enhances the company's marketing efforts by raising its visibility and promoting its brand. By supporting students in their theses, RH gains exposure on faculties, effectively demonstrating its dedication to open source, innovation, and collaboration. Success stories and project outcomes can be shared through marketing channels, bolstering Red Hat's reputation and generating

interest from prospective researchers or employees. The findings presented in Chapter 4 highlight the higher ratio of dean's awards received by our thesis participants. In this way, the thesis program plays a crucial role in reinforcing Red Hat's marketing presence and positioning within academic faculties.

Finally, Red Hat is motivated to support theses because of its commitment to community engagement. Red Hat recognizes the significance of fostering a robust open-source community, and by backing theses, the company can connect with like-minded students who share this commitment. This helps spread the values of open source and promotes collaboration. In return, Red Hat benefits from increased visibility, recognition, and support as a leading figure in the open-source community. Moreover, the thesis program serves as a way for RH engineers to engage with the academic community, with over 150 unique engineers who already participated as consultants on thesis projects.

5.2 Problems with the Current Process

As mentioned in section 2, Red Hat has been supervising theses for over a decade, and at its peak, the company supervised nearly one-third of all theses conducted with industry partners at FI MUNI [6]. However, as other companies started adopting similar practices, this ratio gradually declined. In this chapter, I will delve into Red Hat's present policy for final theses, with a specific emphasis on providing context regarding the current state of the process as of 2023. It is important to understand the context and shortcomings of the existing system when designing an improved process.

At the start of Red Hat's thesis process over a decade ago, RH adopted the practice of dividing the role of the thesis supervisor into two distinct roles: the formal supervisor and the technical supervisor (consultant). While the formal supervisor must be someone from the university, the technical supervisor can be a regular employee from Red Hat. The formal supervisor is responsible for ensuring that the formal requirements of the thesis are met, while the technical supervisor primarily focuses on collaborating with the student on the technical aspects of the project.

In 2023, the most common way for students to apply for a thesis project at Red Hat was through their team, if they were already employed by RH. Generally, students who were not employed by RH applied through the university system. However, there is an issue with creating a thesis within a Red Hat team, as it often bypasses the Red Hat Research team. RHR can provide valuable support in formulating formal requirements and assisting students throughout the process. Additionally, it can be challenging to find a suitable leader for the project within the RH team. Red Hat engineers may have limited experience in leading thesis projects, which can result in difficulties when defending the thesis or ensuring that it meets all the formal requirements.

To address these issues, several guides were created to assist engineers interested in leading a thesis. Unfortunately, these guides are often outdated, contain only partial information, or are difficult to find, as they are often on legacy platforms. Furthermore, engineers have often been unaware of the opportunities available through Red Hat Research, resulting in limited outreach to potential thesis leaders. Most importantly, we had trouble getting new engineers interested in leading theses. Often only engineers who had academic experience and got accustomed to leading theses were most common. This meant that the variety of engineers slowly became stale and we had trouble attracting new consultants. Additionally, engineers didn't really have a platform to come up with their own idea for a thesis, because they didn't even know they might lead one. This became more of an issue as Red Hat's team grew and cooperation became more complex. As a result, there has been a lack of diversity in the range of thesis topics. Despite our efforts to publish open topics on the Red Hat Research webpage, we had to discontinue the thesis section due to a lack of established processes and a designated process owner.

5.3 Designing the Refined Process

To develop a more effective and efficient process, we must first consider the different categories of students who will be involved in the process. These students can be distinguished based on several factors.

The first factor is whether the student is employed by Red Hat. This distinction affects various aspects, including the length of the project, access to RH resources, and even the perceived value of the research for Red Hat.

1. **Students employed by Red Hat**

   These students have access to Red Hat's resources and typically conduct their thesis research within their team. These students are usually interns or existing employees, and their research is typically focused on the technologies they are using in their work. Although these projects can be arguably more immediately beneficial for Red Hat, there is a drawback that thesis writing can be time-consuming for both the consultant and the student, compared to working on the project without writing a thesis. However, these types of theses often serve as useful, "nice to have" projects that the team may not have the immediate capacity to undertake themselves.

2. **Students not employed by Red Hat**

   Students in this category usually discover their thesis topics through the open thesis offerings provided by the university system. Although the topics are often related to common Red Hat technologies, they don't necessarily have to be directly related to the team's current work. Additionally, while internships may be linked to the thesis in some cases, this is not always the norm.

There is a second distinction among students based on how they obtain their thesis topic.

1. **Students who apply to a pre-made topic** These students either apply through the university's IS or Red Hat's website. Typically, the exact wording of the formal requirements of the thesis topic may change while the student works on the project.

2. **Students who don't have a topic** This category typically applies to interns working at Red Hat. In this case, they often come up with a topic collaboratively with a senior member of their team. While they should consult with RHR about the draft topic, it is unfortunately common for them to bypass this.

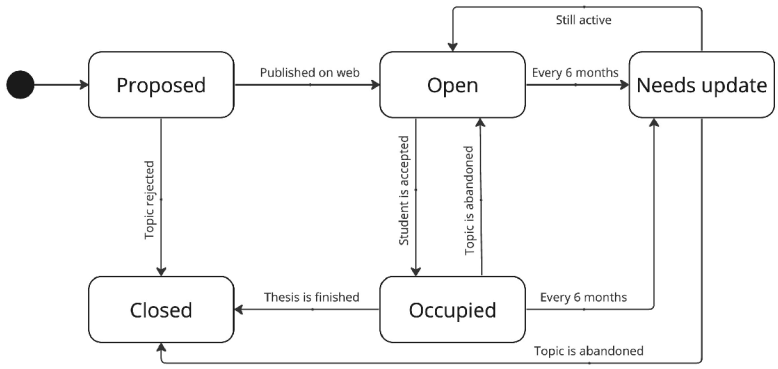

Figure 5.1: Lifecycle of a Thesis Topic: A State Diagram

3. **Students who have their own topic** This is the rarest case. In such situations, the topics must be evaluated individually, and students should contact RHR for assistance. We can direct them to potential consultants or help them refine their topic if necessary.

Now that we have identified the different categories of students involved in the thesis process, we can begin to design a refined process that takes into account the unique needs of each category.

5.3.1 Lifecycle of a Thesis Topic

Managing a thesis topic is a complex process that requires careful attention to its current state. To streamline this process, the lifecycle of a thesis topic can be broken down into several states, each with its own significance and role in the overall management of the topic. These states include Proposed, Open, Occupied, Closed, and Needs Update.

- **Proposed:** The topic was proposed and is not yet approved by the RHR or published on the webpage. This state usually occurs when an engineer fills out the topic proposal form.

- **Open:** The topic is published on the webpage and there is no student involved yet. Every topic on the website should be in this state.
- **Occupied:** The topic is not on the website and the student is working on it. When the student starts to work on the thesis, set the state to occupied and remove it from the webpage.
- **Closed:** The topic is not on the website and the topic was abandoned. This can happen when the student finishes the thesis or the employee decides to abandon the project (leaves the company, is not interested anymore, etc.).
- **Needs Update:** This state occurs automatically every 6 months. It serves to ensure that the website is up to date. When the state occurs, the process owner should check the state of the thesis with the main Red Hatter involved and update the state and the topic accordingly.

The state diagram in Figure 5.1 illustrates how the different states of a thesis topic can transition into each other. This diagram provides a visual representation of the lifecycle of a thesis topic and can be a useful reference for managing topics and tracking their progress. By providing a clear reference of how topics move through the different states, this diagram should help streamline the process and ensure that topics are being appropriately tracked and monitored. Additionally, this diagram is included in the internal guide we have created for managing the thesis topic lifecycle, providing a comprehensive overview for all involved in the process in the future.

For a more detailed description of the entire thesis management process, readers can refer to Figure 5.2. This figure is a BPMN diagram covering the entire lifecycle, starting when the RH engineer wants to lead a thesis. While BPMN diagrams can be complicated to read, they provide significant added value in terms of "human sense-making and communication" [33]. By using this diagram, the process users can understand the process in the same way, ensuring a consistent understanding and effective communication.

It is important to note that the diagram in Figure 5.2 won't be used in guides for employees, as it may be too complex for practical

use. However, it can be used as a reference for process owners and stakeholders who need to understand the bigger picture. Additionally, the diagram follows standards and best practices set by the Object Management Group (OMG) for BPMN 2.0, which is an ISO standard [34]. By adhering to these standards, the diagram provides a reliable and consistent representation of the thesis management process.

5.3.2 Thesis Topics on the Web Page

The absence of up-to-date and comprehensive open thesis topics on the RH website is a major issue with the current process. While there used to be a section for open topics, it was often outdated and difficult to manage, resulting in confusion for students seeking information on potential thesis projects. Because of this, RHR decided to remove the section altogether, leading to difficulties in directing students to a centralized resource for available topics. At faculty events, students frequently asked about available thesis topics, but due to the lack of published topics on the website, we were unable to provide them with a centralized resource. This resulted in missed opportunities for potential research collaborations.

Additionally, the section on the website that previously listed the defended thesis projects completed in collaboration with RH took up a lot of space, prioritizing the large number of theses done in the Czech Republic over projects from other regions. This meant that projects from other regions did not receive enough visibility and even exceptional research projects were often lost in the amount of thesis we do. As a result, RHR decided to hide this section as well.

Moving forward, it is important for RH to have an up-to-date and comprehensive list of open thesis topics on the website to provide students with a centralized resource for finding potential projects. The refined process should enable us to keep track of current open projects and publish them on the website while removing outdated topics in a timely manner.

To ensure that the website is kept up to date, we have implemented a system that involves tracking all the thesis topics using a spreadsheet, which contains comprehensive information about each topic and their current status, as outlined in section 5.3.1. This process is overseen by a designated process owner, who receives automatic calendar and email

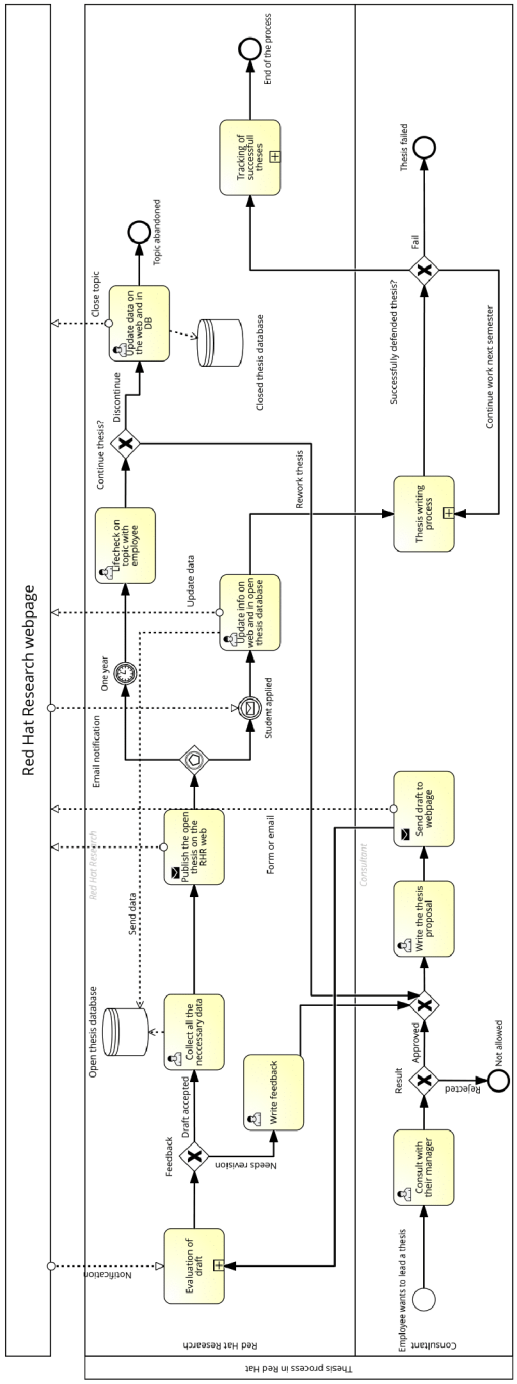

Figure 5.2: BPMN Diagram Showing the RH Thesis Process

reminders and has access to comprehensive guides to help manage the process effectively.

The spreadsheet is programmed to automatically mark topics that require an update every six months, and send reminders to the process owner via email and calendar notifications. All the necessary information and guides for managing the process, including tracking current open topics and managing successfully defended ones, are available directly from the spreadsheet. With this system in place, we are confident that we can avoid the mistakes of the past and keep the topics on the website up to date.

We plan to implement a change in our approach to uploading successfully defended theses to the website. Instead of uploading all of them, we will only handpick a few to showcase our best work. This change is aimed at preventing information overload and ensuring that exceptional research projects get the attention they deserve. We will still maintain an internal database of all the theses, but we will be selective about what we publish on the website.

6 Conclusion

The main objective of this thesis was to analyze and refine the thesis writing program at Red Hat Czech. To achieve this, we began by gaining a deeper understanding of the role of the thesis program within the company, as well as examining various forms of student cooperation in Israel, the USA, and the Czech Republic. While the thesis program has a long-standing tradition at Red Hat Czech, our interviews with managers from Israel and the USA provided unique insight into its distinct position.

As part of our research, we attempted to consolidate the historical datasets to perform data analysis. However, we encountered numerous challenges with manual data cleaning and had to devise an alternative solution. This led us to implement a new tracking process that includes a web scraping script to extract thesis information from the web pages of our two biggest university partners. This script not only helped us with past data but also streamlines future tracking efforts. By semi-automating the process, we ensure the accuracy and correctness of the data format.

With almost a decade's worth of thesis data, we collected over 400 entries and found several statistically significant trends. Firstly, we confirmed a steady decline in the number of theses per year. However, we were pleasantly surprised to see that the variety of unique consultants each year remained at a similar level. Our analysis of grades showed that our theses receive grades above the average on the Faculty of Informatics of Masaryk University. However, we also observed that our theses receive worse grades at BUT compared to FI MUNI. This could potentially be caused by a different average grade at BUT, but we do not currently have the data to confirm this. It is an area that we plan to explore in future research.

Furthermore, our data analysis showed interesting findings in the variety of topics we covered by examining keywords. We also identified several interesting findings that we could use in marketing, such as the higher number of dean's awards and the number of theses written in English.

Looking towards the future, one of the most significant works we have ahead of us is connecting the thesis database with the employee

database. During our work on this thesis, we prepared the groundwork for connecting our database with the employee database using LDAP. This connection will enable us to collect more information, such as the rates of how many students who do a thesis become employees, the seniority of employees involved in the program, or from which teams they come from. Such advanced analysis could prove useful in justifying the thesis program for different stakeholders in Red Hat and ensuring the program's continuity.

Finally, this thesis aimed to refine various aspects of the current thesis program at Red Hat Czech. Due to its age, the program had become quite complex with several outdated processes. We believe that our work will help to modernize and streamline the program, allowing us to fully utilize the webpage. Thanks to the revision of our processes, we can now track successfully defended theses as well as open topics, which will help us to better manage and improve the program in the future.

A Attachment Instructions

In the attachement of this thesis, you will find a JavaScript file that contains the data scraping script used for importing data from the universities. Please note that while the file is in JavaScript format, you need to use the script in the Apps Script together with a valid Google Sheet with the correct format. Apps Script utilizes .gs files, so you can either import the JavaScript file or create a new .gs file in Apps Script and paste the code.

For the convenience of readers, I have prepared an example Google Sheet with the scripts already set up. You can directly observe the script from there by going to Extensions > Apps Script. In the Read Me section, you will find a link to the guide. I recommend reading it before attempting to run the script. If you encounter any issues or require permissions to run the script, please feel free to contact me.

Feel free to make a copy of the Google Sheet along with the scripts and use it for your own purposes. You can access the sheet through the following link: Example GSheet

Bibliography

1. *Red Hat* [online]. Red Hat [visited on 2023]. Available from: https://www.redhat.com.
2. *Red Hat Research* [online]. Red Hat [visited on 2023]. Available from: https://research.redhat.com/.
3. *Research Day* [online]. Red Hat Research [visited on 2023]. Available from: https://research.redhat.com/red-hat-research-day-europe-2022/.
4. MORISSON, Arnault; PATTINSON, Marc. *University-Industry Collaboration* [online]. 2020. [visited on 2023]. Available from: https://www.omg.org/spec/BPMN/2.0/PDF.
5. *Czech Republic - Overview of the education system (EAG 2022)* [online]. [visited on 2023]. Available from: https://gpseducation.oecd.org/CountryProfile?plotter=h5&primaryCountry=CZE&treshold=5&topic=EO.
6. HAXHIDAUTIOVA, Nora. *Internal research*. 2022.
7. *SPP - O Red Hat* [online]. FI MUNI [visited on 2023]. Available from: https://www.fi.muni.cz/for-partners/redhat.html.cs.
8. *Trusted Computing Ecosystem* [online]. [visited on 2023]. Available from: https://research.redhat.com/blog/research_project/trusted-computing-ecosystem/.
9. *Usable Security for Developers* [online]. [visited on 2023]. Available from: https://research.redhat.com/blog/research_project/usable-security-for-developers/.
10. *Mining issued Common Criteria and FIPS140-2 certificates* [online]. [visited on 2023]. Available from: https://research.redhat.com/blog/research_project/security-certificates-chain-analysis/.
11. *BUT - Red Hat Czech s.r.o.* [online]. [visited on 2023]. Available from: https://www.fit.vut.cz/cooperation/partner-detail/1073/.cs.

12. *https://github.com/tfiedor/perun* [online]. 2022. [visited on 2023]. GitHub repository.

13. MALÍK, Viktor; VOJNAR, Tomáš. Automatically Checking Semantic Equivalence between Versions of Large-Scale C Projects. In: 2021, pp. 329–339.

14. *United States - Overview of the education system* (*EAG 2022*) [online]. [visited on 2023]. Available from: `https://gpseducation.oecd.org/CountryProfile?plotter=h5&primaryCountry=USA&treshold=5&topic=EO`.

15. ANTHONY, Amanda. A review of the US higher education system: Its structure, funding, quality and the future. *Journal of Education and Practice*. [N.d.]. Available also from: `http://www.repozytorium.uni.wroc.pl/Content/63656/21_Anthony_Amanda.pdf`.

16. *Collaboratory Report Year 3* (*2019-2020*). 2020. Internal presentation.

17. *Grant Agreement for Collaborative Research and Education between Red Hat and Boston University*. 2021. Internal contract.

18. *Red Hat Collaboratory at Boston University* [online]. [visited on 2023]. Available from: `https://www.bu.edu/rhcollab/`.

19. *Research Interest Groups* (*RIGs*) [online]. [visited on 2023]. Available from: `https://research.redhat.com/rigs/`.

20. *Education system Israel: described and compared with the Dutch system* [online]. Nuffic, 2013 [visited on 2023]. Available from: `https://www.nuffic.nl/sites/default/files/2020-08/education-system-israel.pdf`.

21. *Israel - Overview of the education system* (*EAG 2022*) [online]. [visited on 2023]. Available from: `https://gpseducation.oecd.org/CountryProfile?primaryCountry=ISR&treshold=10&topic=EO`.

22. *SpotOS – a distributed cloud-based operating system over unreliable resources* [online]. Red Hat Research [visited on 2023]. Available from: `https://research.redhat.com/blog/research_project/spotos-a-distributed-cloud-based-operating-system/`.

23. OSBORNE, Jason W. *Best practices in data cleaning* [online]. 2013. [visited on 2023]. Available from: https://doi.org/10.4135/9781452269948.

24. RAHM, Erhard; DO, Hong Hai. Data Cleaning: Problems and Current Approaches [online]. [N.d.] [visited on 2023]. Available from: https://cs.brown.edu/courses/cs227/archives/2017/papers/data-cleaning-IEEE.pdf#page=5.

25. CHAPMAN, Arthur D. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data* [online]. 2005. [visited on 2023]. Available from: https://assets.ctfassets.net/uo17ejk9rkwj/46SfGRfOesU0IagMMAOIkk/1c03ea3e21fcd9025cc800d786890e72/Principles_20and_20Methods_20of_20Data_20Cleaning_20-_20ENGLISH.pdf.

26. *AppScript documentation* [online]. Google [visited on 2023]. Available from: https://developers.google.com/apps-script.

27. GUNAWAN, Rohmat; RAHMATULLOH, Alam; DARMAWAN, Irfan; FIRDAUS, Firman. Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath. In: [online]. Atlantis Press, 2019/03, pp. 283–287 [visited on 2023]. Available from: https://doi.org/10.2991/icoiese-18.2019.50.

28. ERWIG, Martin; GOPINATH, Rahul. Explanations for Regular Expressions. In: 2012, vol. 7212, pp. 394–408.

29. HRUŠOVSKÝ, Matej. *Gender diversity on the rise among engineers* [online]. 2021. [visited on 2023]. Available from: https://research.redhat.com/blog/article/gender-diversity-on-the-rise-among-engineers/.

30. *Zpráva o stipendiích za rok 2022* [online]. MUNI, 2020 [visited on 2023]. Available from: https://www.fi.muni.cz/files/studijni/zprava-o-stipendiich-za-rok-2022.pdf.

31. *Release of Main Science and Technology Indicators – Latest Estimates of R&D Investment in OECD and Major Economies*. OECD, 2017.

32. YOUNG, Michael. *Industry, academia and government collaboration* [online]. [visited on 2023]. Available from: https://www.uschamberfoundation.org/sites/default/files/article/foundation/IndustryAcademia.pdf.

33. AAGESEN, Gustav; KROGSTIE, John. Analysis and Design of Business Processes Using BPMN. In: 2010.

34. *Business Process Model and Notation* [online]. OMG, 2010 [visited on 2023]. Available from: `https://www.omg.org/spec/BPMN/2.0/PDF`.